14th Triennial World Congress, Beijing, P.R. China

CONVERGENCE ANALYSIS OF DISCRETIZATION PROCEDURE IN Q-LEARNING

Guofei Jiang Hui-Qi Gao Cang-Pu Wu

Department of Automatic Control, Beijing Institute of Technology
P.O.Box 327, Beijing 100081, P.R.China
email : wucpu@sun.ihep.ac.cn

George Cybenko

Thayer School of Engineering, Dartmouth College
Hanover, NH 03755, USA
email : george.cybenko@dartmouth.edu

Abstract: Q-Learning is a direct reinforcement learning algorithm for solving stochastic control problems with incomplete information. Discretization of the state and decision spaces is required when Q-Learning is used to solve stochastic optimal control problems with the state and decision spaces which both are continua. In this paper it is shown that under certain compactness and Lipschitz continuity assumptions, the optimal solution obtained with Q-Learning converges almost surely to the optimal solution obtain with the continuous dynamic programming algorithm as the maximal discetization grids approach to zero. *Copyright © 1999 IFAC*

Keywords: dynamic programming, discretization, stochastic control, machine learning, optimization, convergence analysis, Markov decision problems.

1. INTRODUCTION

Since Watkins proposed Q-Learning and proved its convergence in 1989, it has become one of the most widely used reinforcement learning algorithm. However, at present Q-Learning is mainly applied to solve Markov decision problems (MDPs) with states and decisions of finite number. Few empirical or theoretical results about Q-Learning's applications in MDPs with continuous states and decisions have been reported. But in practice the states and decisions of many stochastic controlled systems are often continuous. Discretization of the state and decision spaces is required when Q-Learning is used to solve the problems in these situations. This paper extends the results on convergence of discretization procedure in stochastic dynamic programming (Bertsekas, 1975; Whitt, 1978; Chow and Tsitsiklis, 1991) to those in Q-Learning. Under certain compactness and Lipschitz continuity assumptions, it is shown that the optimal solution obtained with Q-Learning converges almost surely to the optimal solution obtained with the continuous dynamic programming algorithm as the maximal discetization grid approach to zero.

2. MARKOV DECISION PROBLEMS AND Q-LEARNING

Before Q-Learning is introduced, a discrete-time Markov decision model is first described: At each time stage k, the current state x_k is observed and a decision a_k is selected from a finite set of $A(x_k)$. After a_k is performed, the system goes to a next state y_k with some probability $P_{x_k y_k}(a_k)$. Associated with this state transition, an immediate

reward r_k is gained. The object of the controller is to find an optimal policy that maximizes $E\left\{\sum_{i=0}^{\infty}\gamma^i r_{k+i}\right\}$, where γ is the discount factor.

With respect to an arbitrary policy π, Q values are defined as:

$$Q^{\pi}(x,a)=R(x,a)+\gamma\sum_{y}P_{xy}(a)V^{\pi}(y) \qquad (1)$$

Where

$$R(x,a)=E\{r|x,a\},V^{\pi}(y)=\max_{a\in A(y)}Q^{\pi}(y,a).$$

The object in Q-Learning is to estimate the Q values for an optimal policy when the state transition probabilities and reward functions are unknown a priori . To simplify the notation, let $Q^*(x,a)=Q^{\pi^*}(x,a)$, here π^* is the optimal policy. Q Learning works as follows: At each time step k, the controller observes the system state x_k and selects an action a_k from $A(x_k)$. After a_k is executed, the controller receives an immediate reward r_k while the System state transfers to y_k. Then Q_{k-1} is adjusted as follows:

If $(x,a)=(x_k,a_k)$,

$$\begin{aligned}&Q_k(x,a)\\&=(1-\alpha_k)Q_{k-1}(x,a)+\alpha_k[r_k+\gamma V_{k-1}(y_k)];\\&\forall(x,a)\neq(x_k,a_k),\\&Q_k(x,a)=Q_{k-1}(x,a).\end{aligned} \qquad (2)$$

Where $\alpha_k(0\le\alpha_k<1)$ is the learning rate, $V_{k-1}(y)=\max_{b\in A(y)}Q_{k-1}(y,b)$.

Watkins (1989) proved the following convergence theorem for Q-Learning:

Theorem 1: If the following conditions A.1-A.2 are satisfied, $Q_k(x,a)$ in Equation (2) converges to $Q^*(x,a)$ with probability one as $k\to\infty$, $\forall x,a$.

A.1: $\sum_{k=1}^{\infty}\alpha_k(x,a)=\infty,\ \sum_{k=1}^{\infty}[\alpha_k(x,a)]^2<\infty$

A.2: Rewards r_k are bounded, i.e. $|r_k|\le\Re$. $\Re$ is a positive constant.

3. DISCRETIZATION PROCEDURE

For convenience, at first consider an N-stage stochastic optimal control problem.

1. State transitions are determined by the following equation :

$$x_{k+1}=f(x_k,u_k,w_k)\quad k=0,1,2..,\text{N-1} \qquad (3)$$

Where state x_k belongs to a state set S which is defined to be a subset of a metric space bestowed with a metric $\|x\|$. Decision u_k belongs to the constrained decision set $U(x_k)$. Define a decision set $U=\bigcup_{x\in S}U(x)$ which is a subset of a metric space bestowed with a metric $\|u\|$. Stochastic noise w_k belongs to the set W and its distribution may depend on x_k and u_k.

2. The object of the controller is to find an optimal policy to maximize the total reward function :

$$\underset{w_0,\cdots,w_{N-1}}{E}\left\{\sum_{k=0}^{N-1}\gamma^k g(x_k,u_k,w_k)\right\} \qquad (4)$$

Where γ is the discount factor, $g(x_k,u_k,w_k)$ is the reward at stage k that is a real number and is bounded by $\Re$, $k=$ N-1,..,1,0.

The continuous dynamic programming algorithm for solving the above stochastic optimal control problem is given by the following: $\forall x\in S,u\in U(x)$,

$$\overline{Q}_N(x,u)=0\ ; \qquad (5)$$

$$\begin{aligned}\overline{Q}_k(x,u)=\underset{w}{E}\{&g(x,u,w)\\&+\gamma\overline{V}_{k+1}[f(x,u,w)]|x,u\}\ ;\end{aligned} \qquad (6)$$

$$\overline{V}_{k+1}(x)=\sup_{u\in U(x)}\overline{Q}_{k+1}(x,u)\ . \qquad (7)$$

where $k=$ N-1,..,1,0.

Now the state set S is partitioned into n disjointed subsets $S_1,S_2,\cdots,S_n$ with $S=\cup_{i=1}^{n}S_i$ and a point $x^i\in S_i$ is selected from each subset S_i, then a finite aggregated state set $G=\{x^1,x^2,\cdots,x^n\}$ is formed. In the same way U is partitioned into p disjointed subsets $U_1,U_2,\cdots,U_p$ with $U=\cup_{j=1}^{p}U_j$, and a point $u^j\in U_j$ is selected from each subset U_j to obtain a finite aggregated decision set $H=\{u^1,u^2,\cdots,u^p\}$. Assume that $U(x^i)\cap H\neq\phi,\forall i=1,2,...,n$, where $U(x^i)$ is the constrained decision set at state x^i, ϕ denotes the empty set . Then the above continuous dynamic programming algorithm (5)-(7) can be approximated with the following discrete dynamic programming algorithm.

$$\overline{Q}_N(x,u)=0\ ,\ \forall x\in S_i,u\in U(x)\ ; \qquad (8)$$

$$\tilde{Q}_N(x,u)=\tilde{Q}_N(x^i,\tilde{u}_i),\quad i=1,\cdots,n\ ; \tag{9}$$

$\forall x\in G, u\in U(x)\cap H$,

$$\tilde{Q}_k(x,u)=\underset{w}{E}\{g(x,u,w)+\gamma\tilde{V}_{k+1}[f(x,u,w)]|x,u\}; \tag{10}$$

$\forall x\in S_i, u\in U(x)$,

$$\tilde{Q}_k(x,u)=\tilde{Q}_k(x^i,\tilde{u}_i),\quad i=1,\cdots,n\ ; \tag{11}$$

$$\tilde{V}_{k+1}(x)=\max_{u\in U(x)\cap H}\tilde{Q}_{k+1}(x,u)\ ,\ \forall x\in G\ ; \tag{12}$$

$$\tilde{V}_{k+1}(x)=\tilde{V}_{k+1}(x^i)\ ,\ \forall x\in S_i, i=1,\cdots,n\ . \tag{13}$$

where $\tilde{u}_i=\arg\min_{u'\in U(x^i)\cap H}\|u-u'\|$, k = N-1,...,1,0 .

Before the convergence theorem is proved, the following assumptions concerning certain compactness and Lipschitz continuity are introduced.

B.1 : Assume that the state set S , the constrained decision set $U(x)$ and set $U=\bigcup_{x\in S}U(x)$ are compact , and that $\forall x,x'\in S$ and $u\in U(x)$, there exists $u'\in U(x')$ which satisfies the following inequality:

$$\|u-u'\|\le\overline{F}\|x-x'\| \tag{14}$$

where $\overline{F}$ is a positive constant .

B.2 : Assume that $\forall x,x'\in S, u,u'\in U$ and $w\in W$, the functions f,g satisfy the following Lipschitz continuity conditions:

$$\|f(x,u,w)-f(x',u',w)\|\le\overline{L}(\|x-x'\|+\|u-u'\|) \tag{15}$$

$$|g(x,u,w)-g(x',u',w)|\le\overline{M}(\|x-x'\|+\|u-u'\|) \tag{16}$$

where $\overline{L},\overline{M}$ are positive constants .

B.3 : With a given stochastic noise w 's probabilty measure space (W,σ,P) , for $\forall x\in S$ and $u\in U$, define a measurable and integrable function $h(w|x,u)$ on the set W which has parameters x and u . Now Define a norm $\|h(w|x,u)\|\equiv\int_W|h(w|x,u)|dw$ and assume that $\forall x,x'\in S$ and $u,u'\in U$, the function $p(w|x,u)$ satisfies the following Lipschitz continuity condition:

$$\|p(w|x,u)-p(w|x',u')\|\le\overline{O}(\|x-x'\|+\|u-u'\|) \tag{17}$$

where $\overline{O}$ is a positive constant .

4. CONVERGENCE ANALYSIS

Theorem 2: If the assumptions B.1-B.3 hold , then the Q-value function $\overline{Q}_k(x,u)$ determined by Equation (5)-(7) satisfies the following inequality: $\forall x,x'\in S$ and $u,u'\in U$, K=N-1,...,1,0,

$$|\overline{Q}_k(x,u)-\overline{Q}_k(x',u')|\le A_k(\|x-x'\|+\|u-u'\|) \tag{18}$$

where A_k (k=N-1,...,1,0) are positive constants .

Proof: For k=N-1 , from Equation (5)-(7) it can be shown that for each $x,x'\in S$ and $u,u'\in U$,

$$\begin{aligned}&|\overline{Q}_{N-1}(x,u)-\overline{Q}_{N-1}(x',u')|\\&\le\left|\int_W g(x,u,w)p(w|x,u)dw-\int_W g(x,u,w)p(w|x',u')dw\right|\\&+\left|\int_W g(x,u,w)p(w|x',u')dw-\int_W g(x',u',w)p(w|x',u')dw\right|\\&\le\sup\{g(x,u,w)|x\in S,u\in U,w\in W\}\cdot\int_W|p(w|x,u)-p(w|x',u')|dw\\&+\int_W|g(x,u,w)-g(x',u',w)|p(w|x',u')dw\end{aligned}$$

Because of the assumption B.3 , then

$$\begin{aligned}&\int_W|p(w|x,u)-p(w|x',u')|dw\\&=\|p(w|x,u)-p(w|x',u')\|\\&\le\overline{O}(\|x-x'\|+\|u-u'\|)\end{aligned}$$

Thus the following inequality holds

$$\begin{aligned}&|\overline{Q}_{N-1}(x,u)-\overline{Q}_{N-1}(x',u')|\\&\le A_{N-1}(\|x-x'\|+\|u-u'\|)\end{aligned} \tag{19}$$

where $A_{N-1}=B_{N-1}\overline{O}+\overline{M}$,

$B_{N-1}=\sup\{g(x,u,w)|x\in S,u\in U,w\in W\}$.

From Equation (7) , then

$$\begin{aligned}&\overline{V}_{N-1}(x)-\overline{V}_{N-1}(x')\\&\le\left|\max_{u\in U(x)}\overline{Q}_{N-1}(x,u)-\max_{u\in U(x')}\overline{Q}_{N-1}(x',u)\right|\end{aligned}$$

Because $U(x)$ is a compact set , there exists $v\in U(x)$ which satisfies $\overline{Q}_{N-1}(x,v)=\max_{u\in U(x)}\overline{Q}_{N-1}(x,u)$. According to the assumption B.1 , there also exists $v'\in U(x')$ such that

$\|v - v'\| \le \overline{F}\|x - x'\|$. It follows from $\overline{Q}_{N-1}(x',v') \le \max_{u \in U(x')} \overline{Q}_{N-1}(x',u)$ that

$$\overline{V}_{N-1}(x) - \overline{V}_{N-1}(x') \le \left|\overline{Q}_{N-1}(x,v) - \overline{Q}_{N-1}(x',v')\right|$$
$$\le A_{N-1}\left(1 + \overline{F}\right)\|x - x'\|$$

Because of the symmetry of x and x', similarly

$$\overline{V}_{N-1}(x') - \overline{V}_{N-1}(x) \le A_{N-1}\left(1 + \overline{F}\right)\|x - x'\|$$

Then $\left|\overline{V}_{N-1}(x) - \overline{V}_{N-1}(x')\right|$

$$\le A_{N-1}\left(1 + \overline{F}\right)\|x - x'\| \qquad (20)$$

Now consider the stage k=N-2 and the following inequality holds :

$$\left|\overline{Q}_{N-2}(x,u) - \overline{Q}_{N-2}(x',u')\right|$$
$$\le \left|\int_W g(x,u,w)p(w|x,u)dw - \int_W g(x',u',w)p(w|x',u')dw\right|$$
$$+ \gamma\left|\int_W \overline{V}_{N-1}(f(x,u,w))p(w|x,u)dw - \int_W \overline{V}_{N-1}(f(x',u',w))p(w|x',u')dw\right|$$
$$\le A_{N-2}\left(\|x - x'\| + \|u - u'\|\right) \qquad (21)$$

where $A_{N-2} = B_{N-2}\overline{O} + \overline{M} + \gamma A_{N-1}\left(1 + \overline{F}\right)\overline{L}$,

$$B_{N-2} = \sup\left\{g(x,u,w) \middle| x \in S, u \in U, w \in W\right\}$$
$$+ \gamma \sup\left\{\overline{V}_{N-1}(f(x,u,w)) \middle| x \in S, u \in U, w \in W\right\}$$

Thus for k=N-2 the Inequality (18) is proved . In the same way it can be proved that the Inequalities (18) hold for every k= N-3,...,1,0 . **Q.E.D.**

Now define

$$d_s = \max_{i=1,\cdots,n} \sup_{x \in S_i} \|x - x^i\| , \qquad (22)$$

$$d_u = \max_{i=1,\cdots,n} \sup_{x \in S_i} \max_{u \in U(x)} \min_{u' \in U(x^i) \cap H} \|u - u'\| . \qquad (23)$$

Theorem 3 : If the assumptions B.1-B.3 hold , the following inequalities hold for every $x \in S$, $u \in U(x)$ and k=N-1,...,1,0 ,

$$\left|\overline{Q}_k(x,u) - \tilde{Q}_k(x,u)\right| \le \beta_k(d_s + d_u) \qquad (24)$$

where $\beta_k (k = 0,1,\cdots,N-1)$ are positive constants . $\overline{Q}_k(x,u), \tilde{Q}_k(x,u)$ are separately given in Equation (5)-(7) and (8)-(13) .

Proof : At stage k=N-1 , for every $x \in G$ and $u \in U(x) \cap H$, it is straightforward to show by Equation (5)-(7) and (8)-(13) that $\overline{Q}_{N-1}(x,u) = \tilde{Q}_{N-1}(x,u)$. For every and $u \in U(x)$ and $x \in S_i (i = 1,\cdots,n)$, then

$$\left|\overline{Q}_{N-1}(x,u) - \tilde{Q}_{N-1}(x,u)\right|$$
$$\le A_{N-1}(d_s + d_u) = \beta_{N-1}(d_s + d_u) \qquad (25)$$

where $\beta_{N-1} = A_{N-1}$, $\tilde{u}_i = \arg\min_{u' \in U(x^i) \cap H} \|u - u'\|$.

According to Equation (7) and (12)-(13) , for every $x \in S_i (i = 1,\cdots,n)$,

$$\left|\overline{V}_{N-1}(x) - \tilde{V}_{N-1}(x)\right|$$
$$= \left|\max_{u \in U(x)} \overline{Q}_{N-1}(x,u) - \max_{u \in U(x^i) \cap H} \tilde{Q}_{N-1}(x^i,u)\right|$$

By Inequality (20) , it is shown that

$$\left|\max_{u \in U(x)} \overline{Q}_{N-1}(x,u) - \max_{u \in U(x^i)} \overline{Q}_{N-1}(x^i,u)\right|$$
$$\le A_{N-1}\left(1 + \overline{F}\right)\|x - x^i\|$$

By Inequality (25) , the following inequality holds

$$\left|\max_{u \in U(x^i)} \overline{Q}_{N-1}(x^i,u) - \max_{u \in U(x^i)} \tilde{Q}_{N-1}(x^i,u)\right|$$
$$\le \max_{u \in U(x^i)} \left|\overline{Q}_{N-1}(x^i,u) - \tilde{Q}_{N-1}(x^i,u)\right|$$
$$\le \beta_{N-1}(d_s + d_u)$$

Then $\left|\overline{V}_{N-1}(x) - \tilde{V}_{N-1}(x)\right|$

$$\le \left(A_{N-1} + A_{N-1}\overline{F} + \beta_{N-1}\right)(d_s + d_u) \qquad (26)$$

Now consider the stage k=N-2 . For each $x^i \in G$ and $u^j \in U(x^i) \cap H$ $(i = 1,\cdots,n;\ j = 1,\cdots,p)$, by Equation (5)-(7) and (8)-(13) ,

$$\left|\overline{Q}_{N-2}(x^i,u^j) - \tilde{Q}_{N-2}(x^i,u^j)\right|$$
$$\le \gamma\left(A_{N-1} + A_{N-1}\overline{F} + \beta_{N-1}\right)(d_s + d_u) \qquad (27)$$

For every $x \in S_i (i = 1,\cdots,n)$ and $u \in U(x)$,

$$\left|\overline{Q}_{N-2}(x,u) - \tilde{Q}_{N-2}(x,u)\right|$$
$$\le \left|\overline{Q}_{N-2}(x,u) - \overline{Q}_{N-2}(x^i,\tilde{u}_i)\right|$$
$$+ \left|\overline{Q}_{N-2}(x^i,\tilde{u}_i) - \tilde{Q}_{N-2}(x^i,\tilde{u}_i)\right|$$

where $\tilde{u}_i = \arg\min_{u' \in U(x^i) \cap H} \|u - u'\|$.

By theorem 2 , the following inequlity holds

$$\left|\overline{Q}_{N-2}(x,u) - \overline{Q}_{N-2}(x^i,\tilde{u}_i)\right| \le A_{N-2}(d_s + d_u)$$

By Inequality (27), then

$$\left|\overline{Q}_{N-2}(x^i,\tilde{u}_i) - \tilde{Q}_{N-2}(x^i,\tilde{u}_i)\right|$$
$$\le \gamma\left(A_{N-1} + A_{N-1}\overline{F} + \beta_{N-1}\right)(d_s + d_u)$$

Thus for every $x \in S$ and $u \in U(x)$,

$$\left|\overline{Q}_{N-2}(x,u) - \tilde{Q}_{N-2}(x,u)\right| \le \left(A_{N-2} + \gamma A_{N-1} + \gamma A_{N-1}\overline{F} + \gamma\beta_{N-1}\right)(d_s + d_a) = \beta_{N-2}(d_s + d_a) \quad (28)$$

where $\beta_{N-2} = A_{N-2} + \gamma A_{N-1} + \gamma A_{N-1}\overline{F} + \gamma\beta_{N-1}$.

Thus for k=N-2 the Inequality (24) is proved and similarly it can be proved for every k. **Q.E.D.**

Now consider the infinite-horizon stochastic optimal control problems. In fact by replacing the stage index N in Equation (3)-(13) with ∞, an infinite-horizon stochastic optimal control problem and the related dynamic programming algorithm can be obtained. According to the convergence theory of stochastic dynamic programming (Bertseks,1978), the solution sequence obtained from the continuous dynamic programming algorithm expressed by Equation (5)-(7) converges to the following optimal solution as $N \to \infty$.

$$\overline{Q}^*(x,u) = \overline{Q}_\infty(x,u) = \lim_{k\to\infty}\overline{Q}_k(x,u) \quad (29)$$

Similarly the solution sequence obtained from the discrete dynamic programming algorithm expressed by Equation (8)-(13) converges to the following optimal solution as $N \to \infty$.

$$\tilde{Q}^*(x,u) = \tilde{Q}_\infty(x,u) = \lim_{k\to\infty}\tilde{Q}_k(x,u) \quad (30)$$

Theorem 4 : If the assumptions B.1-B.3 hold, for every $x \in S$ and $u \in U(x)$, the following equation holds:

$$\lim_{(d_s+d_a)\to 0}\left|\overline{Q}^*(x,u) - \tilde{Q}^*(x,u)\right| = 0 \quad (31)$$

where $\overline{Q}^*(x,u), \tilde{Q}^*(x,u)$ are separately given by Equation (29) and (30).

Proof : By Equation (29), then

$$\left|\overline{Q}^*(x,u) - \overline{Q}_k(x,u)\right| = \left|\overline{Q}_\infty(x,u) - \overline{Q}_k(x,u)\right| \le \left|\sum_{i=k+1}^{\infty}\gamma^i\Re\right| = \frac{\gamma^{k+1}\Re}{1-\gamma} \quad (32)$$

where $\Re$ is the bound of the reward function. Similarly by Equation (30),

$$\left|\tilde{Q}^*(x,u) - \tilde{Q}_k(x,u)\right| = \left|\tilde{Q}_\infty(x,u) - \tilde{Q}_k(x,u)\right| \le \left|\sum_{i=k+1}^{\infty}\gamma^i\Re\right| = \frac{\gamma^{k+1}\Re}{1-\gamma} \quad (33)$$

Further by Inequalities (32), (33) and Theorem 3, the following inequality holds for $\forall k \in \{0,1,2,\cdots\}$,

$$\left|\overline{Q}^*(x,u) - \tilde{Q}^*(x,u)\right| \le \frac{2\gamma^{k+1}\Re}{1-\gamma} + \beta_k(d_s + d_a) \quad (34)$$

where $\beta_k\,(k = 0,1,\cdots,N-1)$ are positive constants.

It follows that $\forall\varepsilon > 0$, there exists a $\delta > 0$ such that if $(d_s + d_a) < \delta$, then

$$\left|\overline{Q}^*(x,u) - \tilde{Q}^*(x,u)\right| < \varepsilon \quad (35)$$

which implies that the Equation (31) holds. **Q.E.D.**

In section 3, it is shown that after the discretization procedure, an MDP with states and decisions of finite number can be formulated, which is associated with the original continuous stochastic optimal control problem. States x belongs to the finite aggregated state set $G = \{x^1, x^2, \cdots, x^n\}$ and decisions u belongs to the constrained and aggregated decision set $U(x) \cap H$ with $H = \{u^1, u^2, \cdots, u^p\}$. If the model of this MDP is unknown (i.e. the distribution of the stochastic noise w is unknown), Q-Learning can be used to solve this MDP in the following way:

$$Q_0(x,u) = 0, \quad \forall x \in S, u \in U(x); \quad (36)$$

$$Q_k(x,u) = (1-\alpha_k)Q_{k-1}(x,u) + \alpha_k\left[r_k + \gamma V_{k-1}(y)\right], \quad \forall x \in G, u \in U(x) \cap H; \quad (37)$$

$$Q_k(x,u) = Q_k(x^i, \tilde{u}_i), \quad \forall x \in S_i, i = 1,\cdots,n, u \in U(x); \quad (38)$$

$$V_{k-1}(x) = \max_{u \in U(x)\cap H} Q_{k-1}(x,u), \quad \forall x \in G; \quad (39)$$

$$V_{k-1}(x) = V_{k-1}(x^i), \quad \forall x \in S_i, i = 1,\cdots,n. \quad (40)$$

where $\tilde{u}_i = \arg\min_{x' \in U(x^i)\cap H}\|u - u'\|$, k=1,2,3,....

Theorem 5: If the following conditions C.1-C.2 are satisfied:

C.1 : The original continuous stochastic optimal problem satisfies the assumptions B.1-B.3.

C.2 : The conditions A.1-A.2 are satisfied when $Q_k(x,u)$ is updated by Equation (37).

then after the above discretization procedure, for every $x \in S$ and $u \in U(x)$, $Q_k(x,u)$ converges to the optimal solution $\overline{Q}^*(x,u)$ as $(d_s + d_a) \to 0$ and $k \to \infty$, i.e.

$$\lim_{\substack{k\to\infty \\ (d_s+d_a)\to 0}} P\left\{\left|Q_k(x,u) - \overline{Q}^*(x,u)\right| = 0\right\} = 1 \quad (41)$$

where $Q_k(x,u), \overline{Q}^*(x,u)$ are separately given by Equation (36)-(40) and (29).

Proof : If the assumptions B.1-B.3 hold, then according to theorem 4, the following conclusion holds : $\forall\varepsilon_1\,(0 < \varepsilon_1 < \varepsilon)$, there exists a $\delta_1 > 0$ such that if $(d_s + d_a) < \delta_1$, then the following

inequality holds for every $x \in S$ and $u \in U(x)$,

$$\left|\overline{Q}^*(x,u) - \tilde{Q}^*(x,u)\right| < \varepsilon_1 \qquad (42)$$

where $\tilde{Q}^*(x,u)$ is given by Equation (30).

After the discretization procedure, an MDP with states and decisions of finite number is formulated, which is associated with the original continuous stochastic optimal control problem. If the model of this MDP is unknown, Q-Learning can be used to solve this problem. According to theorem 1, if the conditions A.1-A.2 are satisfied and $k \to \infty$, then $Q_k(x,u)$ in Equation (37) converges to $\tilde{Q}^*(x,u)$ with probability one for every $x \in G$ and $u \in U(x) \cap H$. Further by Equation (8)-(13) and (36)-(40), it can be shown that for every $x \in S_i (i = 1,2,\cdots,n)$ and $u \in U(x)$, $Q_k(x,u)$ also converges to $\tilde{Q}^*(x,u)$ with probability one. Thus the following conclusion can be obtained: if the conditions A.1-A.2 are satisfied, $\forall \varepsilon_2 (0 < \varepsilon_2 < \varepsilon - \varepsilon_1)$, there exists a N_1 such that if $k > N_1$, then the following equation holds for every $x \in S$ and $u \in U(x)$,

$$P\left\{\left|Q_k(x,u) - \tilde{Q}^*(x,u)\right| < \varepsilon_2\right\} = 1 \qquad (43)$$

It follows from Inequalities (42), (43) that for an arbitrary $\varepsilon > 0$, there exist $N = N_1$ and $\delta = \delta_1 > 0$ such that if $(d_r + d_u) < \delta$ and $k > N$, then

$$P\left\{\left|Q_k(x,u) - \overline{Q}^*(x,u)\right| < \varepsilon\right\}$$
$$\geq P\left\{\left|Q_k(x,u) - \tilde{Q}^*(x,u)\right| < \varepsilon - \varepsilon_1\right\}$$
$$\geq P\left\{\left|Q_k(x,u) - \tilde{Q}^*(x,u)\right| < \varepsilon_2\right\} = 1$$

Therefore $P\left\{\left|Q_k(x,u) - \overline{Q}^*(x,u)\right| < \varepsilon\right\} = 1$, which implies that the Equation (41) holds. **Q.E.D.**

5. DISCUSSIONS

This paper is motivated by the desire to use Q-Learning to solve MDPs with state and decision spaces which both are continua. Before closing the paper, it should be noted that Theorem 5 only holds for off-line Q-Learning. Off-line Q-Learning can be viewed as an off-line asynchronous dynamic programming that is unique in not requiring direct access to the state-transition probabilities of the decision problem. On-line(real-time Q-Learning) obtains the sequence of quadruples (x_k, a_k, y_k, r_k) from the real system while off-line Q-Learning obtains the sequence from the simulated system model. An excellent discussion of these two methods can be found in (Barto, *et al.*, 1995).

In off-line Q-Learning, the immediate reward and subsequent state for every aggregated state-action pairs of the finite sets G and H can be determined from the simulated system model and at each time step an action is tried on an aggregated state. Then a new MDP can be formulated on the finite aggregated state and decision sets and the procedure of Q-learning can be applied to this MDP. In the online case, at each stage k the quadruple (x_k, a_k, y_k, r_k) is observed and mapped from the real system (the MDP with state and decision spaces of continua) and an action can only be limited to be tried on the current real system state (but not the aggregated state). Now the original MDPs with continuous state and decision spaces can be viewed as an external decision problem and the associated reinforcement learning problem with aggregated states can be viewed as an internal decision problem. Obviously the internal decision problem may be non-Markovian because of perceptual aliasing, e.g. $r(x^i, u^j)$ may be equal to every element of the set $\{r(x,u) | x \in S_i, u \in U(x) \cap U_j\}$. The perceptual aliasing problem occurs because one internal state (e.g. x^i) represents multiple external states (e.g. $\forall x \in S_i$). So it is not clear whether the on-line Q-Learning algorithm can converge when it is applied to this non-Markov decision problem.

ACKNOWLEDGEMENTS

The work of this paper is supported by National Natural Science Foundation of China.

REFERENCES

Barto, A.G., S.J.Bradtke and S.P. Singh(1995). Learning to act using real-time dynamic programming, *Artificial Intelligence*, **vol. 72**, pp. 81-138.

Bertsekas, D.P.(1975). Convergence of discretization procedure in dynamic programming. *IEEE Trans. on A.C.*, **vol. 20**, pp. 415-419.

Bertsekas,D.P.(1978). *Stochastic Optimal Control : the Discrete Time Case*, Academic Press, New York, NY.

Chow,C.S.and J.N.Tsitsiklis(1991). An optimal one-way multigrid algorithm for discrete-time stochastic control, *IEEE Trans. on A.C.*, **vol. 36**, pp. 898-914.

Watkins,C.J.C.H(1989). *Learning from Delayed Rewards*, Ph.D. Dissertation, Cambridge University, UK.

Whitt, W.(1978). Approximations of dynamic programs I, *Mathematics of Operation Research*, **vol. 3**, pp. 231-243.