ECE 555 Control of Stochastic Systems Fall 2005

Handout: ***Reinforcement learning***

In this handout we analyse reinforcement learning algorithms for Markov decision processes. The reader is referred to [2, 10] for a general background of the subject and to other references listed below for further details. This handout is based on [5].

Stochastic approximation In lecture on November 29th we considered the general stochastic approximation recursion,

$$\theta(n+1) = \theta(k) + a_n[g(\theta(n)) + \Delta(n+1)], \qquad n \geq 0,\ \theta(0) \in \mathbb{R}^d. \tag{1}$$

Here we provide a summary of the main results from [5].

Associated with the recursion (1) are two O.D.E.s,

$$\frac{d}{dt}x(t) = g(x(t)) \tag{2}$$

$$\frac{d}{dt}x(t) = g_\infty(x(t)), \tag{3}$$

where $g_\infty : \mathbb{R}^d \to \mathbb{R}^d$ is the scaled function, $\lim_{r\to\infty} r^{-1}g(rx) = g_\infty(x)$, $x \in \mathbb{R}^d$. We assumed in lecture that this limit exists, along with some additional properties,

(A1) The function g is Lipschitz, and the limit $g_\infty(x)$ exists for each $x \in \mathbb{R}^d$. Furthermore, the origin in $\mathbb{R}^d$ is an asymptotically stable equilibrium for the O.D.E. (3).

(A2) The sequence $\{\Delta(n) : n \geq 1\}$ is a martingale difference sequence with respect to $\mathcal{F}_n = \sigma(\theta(i), \Delta(i), i \leq n)$. Moreover, for some $\sigma_\Delta^2 < \infty$ and any initial condition $\theta(0) \in \mathbb{R}^d$,

$$\mathsf{E}[\|\Delta(n+1)\|^2 \mid \mathcal{F}_n] \leq \sigma_\Delta^2(1 + \|\theta(n)\|^2), \qquad n \geq 0.$$

The sequence $\{a_n\}$ is deterministic and is assumed to satisfy one of the following two assumptions. Here TS stands for 'tapering stepsize' and BS for 'bounded stepsize'.

(TS) The sequence $\{a_n\}$ satisfies $0 < a_n \leq 1,\ n \geq 0$, and

$$\sum_n a_n = \infty, \qquad \sum_n a_n^2 < \infty.$$

(BS) The sequence $\{a_n\}$ is constant: $a_n \equiv a > 0$ for all n.

Stability of the O.D.E. (3) implies stability of the algorithm:

Theorem 1 *Assume that (A1), (A2) hold. Then, for any initial condition $\theta(0) \in \mathbb{R}^d$,*

(i) *Under (TS),* $\sup_n \|\theta(n)\| < \infty$ *a.s..*

(ii) *Under (BS) there exists $a_0 > 0$, $b_0 < \infty$, such that for any fixed $a \in (0, a_0]$,*

$$\limsup_{n\to\infty} \mathsf{E}[\|\theta(n)\|^2] \leq b_0.$$

□

For the TS model we have convergence when the O.D.E. (2) has a stable equilibrium point:

Theorem 2 *Suppose that (A1), (A2), (TS) hold and that the O.D.E. (2) has a unique globally asymptotically stable equilibrium* θ^**. Then* $\theta(n) \to \theta^*$ *a.s. as* $n \to \infty$ *for any initial condition* $\theta(0) \in \mathbb{R}^d$*.*

We can also obtain bounds for the fixed stepsize algorithm. Let $\boldsymbol{e}$ denote the error sequence,

$$e(n) = \|\theta(n) - \theta^*\|, \qquad n \geq 0.$$

Theorem 3 *Assume that (A1), (A2) and (BS) hold, and suppose that (2) has a globally asymptotically stable equilibrium point* θ^**. Then, for* $a \in (0, a_0]$*, and for every initial condition* $\theta(0) \in \mathbb{R}^d$*,*

(i) *For any* $\varepsilon > 0$*, there exists* $b_1 = b_1(\varepsilon) < \infty$ *such that*

$$\limsup_{n\to\infty} \mathsf{P}(e(n) \geq \varepsilon) \leq b_1 a.$$

(ii) *If* θ^* *is a globally exponentially asymptotically stable equilibrium for the O.D.E. (2), then there exists* $b_2 < \infty$ *such that,*

$$\limsup_{n\to\infty} \mathsf{E}[e(n)^2] \leq b_2 a.$$

□

Suppose that the increments of the model take the form,

$$g(\theta(n)) + \Delta(n+1) = f(\theta(n), N(n+1)), \qquad n \geq 0, \tag{4}$$

where $\boldsymbol{N}$ is an i.i.d. sequence on $\mathbb{R}^q$. In this case, for the BS model, the stochastic process $\boldsymbol{\theta}$ is a (time-homogeneous) Markov chain. Assumptions (5) and (6) below are required to establish ψ-irreducibility:

There exists a $n^* \in \mathbb{R}^q$ *with* $f(\theta^*, n^*) = 0$*, and a continuous density* $p : \mathbb{R}^q \to \mathbb{R}_+$ *satisfying* $p(n^*) > 0$ *and*

$$\mathsf{P}(N(1) \in A) \geq \int_A p(z)dz, \qquad A \in \mathcal{B}(\mathbb{R}^q); \tag{5}$$

The pair of matrices (A, B) *is controllable with*

$$A = \frac{\partial}{\partial x} f(\theta^*, n^*) \quad \text{and} \quad B = \frac{\partial}{\partial n} f(\theta^*, n^*), \tag{6}$$

Under Assumptions (5) and (6) there exists a neighborhood $B(\epsilon)$ of θ^* that is *small* in the sense that there exists a probability measure ν on $\mathbb{R}^d$ and $\delta > 0$ such that

$$P^d(x, A) := \mathsf{P}\{\theta(r) \in A \mid \theta(0) = x\} \geq \delta\nu(A), \qquad x \in B(\epsilon)$$

Stability of the O.D.E. (2) can be used to show that the resolvent satisfies,

$$R(x, B(\epsilon)) := \sum_{k=0}^{\infty} 2^{-k-1} P^k(x, B(\epsilon)) > 0, \qquad x \in \mathbb{R}^d,$$

which is equivalent to ψ-irreducibility [9].

Theorem 4 *Suppose that (A1), (A2), (5), and (6) hold for the Markov model satisfying (4) with $a \in (0, a_0]$. Then we have the following bounds:*

(i) *There exist positive-valued functions A_0 and ε_0 of a, and a constant A_1 independent of a, such that*

$$\mathsf{P}\{e(n) \geq \varepsilon \mid \theta(0) = x\} \leq A_0(a) + A_1(\|x\|^2 + 1)\exp(-\varepsilon_0(a)n), \qquad n \geq 0,\ a \in (0, a_0].$$

The functions satisfy $A_0(a) \leq b_1 a$ and $\varepsilon_0(a) \to 0$ as $a \downarrow 0$.

(ii) *If in addition the O.D.E. (2) is exponentially asymptotically stable, then the stronger bound holds,*

$$\mathsf{E}[e(n)^2 \mid \theta(0) = x] \leq B_0(a) + B_1(\|x\|^2 + 1)\exp(-\epsilon_0(a)n), \qquad n \geq 0,\ a \in (0, a_0],$$

where $B_0(a) \leq b_2 a$, $\varepsilon_0(a) \to 0$ as $a \downarrow 0$, and B_1 is independent of a.

Markov decision processes We now review general theory for Markov decision processes. It is assumed that the state process $\boldsymbol{X} = \{X(t) : t \in \mathbb{Z}_+\}$ takes values in a finite state space $\mathsf{X} = \{1, 2, \cdots, s\}$, and the control sequence $\boldsymbol{U} = \{U(t) : t \in \mathbb{Z}_+\}$ takes values in a finite action space $\mathsf{U} = \{u_0, \cdots, u_r\}$. The controlled transition probabilities are denoted $P_u(i,j)$ for $i, j \in \mathsf{X}, u \in \mathsf{U}$. We are most interested in stationary policies of the form $U(t) = \phi(X(t))$, where the *feedback law* ϕ is a function $\phi\colon \mathsf{X} \to \mathsf{U}$.

Let $c : \mathsf{X} \times \mathsf{U} \to \mathbb{R}$ be the one-step cost function, and consider first the infinite horizon discounted cost control problem of minimizing over all admissible $\boldsymbol{U}$ the total discounted cost

$$h_U(i) = \mathsf{E}\Big[\sum_{t=0}^{\infty} (1+\gamma)^{-t-1} c(X(t), U(t)) \mid X(0) = i\Big],$$

where $\gamma \in (0, \infty)$ is the discount factor. The minimal value function is defined as

$$h^*(i) = \min_U h_U(i),$$

where the minimum is over all admissible control sequences $\boldsymbol{U}$. The function h^* satisfies the dynamic programming equation

$$(1+\gamma)h^*(i) = \min_u \Big[c(i,u) + \sum_j P_u(i,j)h^*(j)\Big], \qquad i \in \mathsf{X},$$

and the optimal control minimizing h is given as the stationary policy defined through the feedback law ϕ^* given as any solution to

$$\phi^*(i) := \arg\min_u \Big[c(i,u) + \sum_j P_u(i,j)h^*(j)\Big], \qquad i \in \mathsf{X}.$$

The *value iteration algorithm* is an iterative procedure to compute the minimal value function. Given an initial function $h_0\colon \mathsf{X} \to \mathbb{R}_+$ one obtains a sequence of functions $\{h_n\}$ through the recursion

$$h_{n+1}(i) = (1+\gamma)^{-1} \min_u \Big[c(i,u) + \sum_j P_u(i,j)h_n(j)\Big], \qquad i \in \mathsf{X},\ n \geq 0. \tag{7}$$

This recursion is convergent for any initialization $h_0 \geq 0$.

The value iteration algorithm is initialized with a function $h_0 \colon \mathsf{X} \to \mathbb{R}_+$. In contrast, the *policy iteration algorithm* is initialized with a feedback law ϕ^0, and generates a sequence of feedback laws $\{\phi^n : n \geq 0\}$. At the nth stage of the algorithm a feedback law ϕ^n is given, and the value function h_n is computed. Interpreted as a column vector in $\mathbb{R}^s$, the vector h_n satisfies the equation

$$((1+\gamma)I - P_n)h_n = c_n \tag{8}$$

where the $s \times s$ matrix P_n is defined by $P_n(i,j) = P_{\phi^n(i)}(i,j)$, $i,j \in \mathsf{X}$, and the column vector c_n is given by $c_n(i) = c(i, \phi^n(i))$, $i \in \mathsf{X}$. Given h_n, the next feedback law ϕ^{n+1} is then computed via

$$\phi^{n+1}(i) = \arg\min_u \Big[c(i,u) + \sum_j P_u(i,j) h_n(j)\Big], \qquad i \in \mathsf{X}. \tag{9}$$

Each step of the policy iteration algorithm is computationally intensive for large state spaces since the computation of h_n requires the inversion of the $s \times s$ matrix $(1+\gamma)I - P_n$ to solve (8). For each n, this can be solved using the 'fixed-policy' version of value iteration,

$$V_{N+1}(i) = (1+\gamma)^{-1}[P_n V_N(i) + c_n], \qquad i \in \mathsf{X},\ N \geq 0, \tag{10}$$

where $V_0 \in \mathbb{R}^s$ is given as an initial condition. Then $V_N \to h_n$, the solution to (8), at a geometric rate as $N \to \infty$.

In the average cost optimization problem one seeks to minimize over all admissible $\boldsymbol{U}$,

$$\eta_U(x) := \limsup_{n \to \infty} \frac{1}{n} \sum_{t=0}^{n-1} \mathsf{E}_x[c(X(t), U(t))]. \tag{11}$$

The policy iteration and value iteration algorithms to solve this optimization problem remain unchanged with a few exceptions. One is that the constant γ must be set equal to zero in equations (7) and (10). Secondly, in the policy iteration algorithm the value function h_n is replaced by a solution to Poisson's equation

$$P_n h_n = h_n - c_n + \eta_n, \tag{12}$$

where η_n is the steady state cost under the policy ϕ^n. The computation of h_n and η_n again involves matrix inversions via

$$\pi_n(I - P_n + ee') = e', \quad \eta_n = \pi_n c_n, \quad (I - P_n + ee')h_n = c_n,$$

where $e \in \mathbb{R}^s$ is the column vector consisting of all ones, and the row vector π_n is the invariant probability for P_n. The introduction of the outer product ensures that the matrix $(I - P_n + ee')$ is invertible, provided that the invariant probability π_n is unique.

$\boldsymbol{Q}$-learning If we define *Q-values* via

$$Q^*(i,u) = c(i,u) + \sum_j P_u(i,j) h^*(j), \qquad i \in \mathsf{X}, u \in \mathsf{U}, \tag{13}$$

then $h^*(i) = \min_u Q^*(i,u)$ and the matrix Q^* satisfies

$$Q^*(i,u) = c(i,u) + (1+\gamma)^{-1} \sum_j P_u(i,j) \min_v Q^*(j,v), \qquad i \in \mathsf{X}, u \in \mathsf{U}.$$

The matrix Q^* can be computed using the equivalent formulation of value iteration,

$$Q_{n+1}(i,u) = c(i,u) + (1+\gamma)^{-1} \sum_j P_u(i,j)\big(\min_v Q_n(j,v)\big), \qquad i \in \mathsf{X}, u \in \mathsf{U}, n \geq 0, \tag{14}$$

where $Q_0 \geq 0$ is arbitrary.

If transition probabilities are unknown so that value iteration is not directly applicable, one may apply a stochastic approximation variant known as the *Q-learning algorithm* of Watkins [11, 12]. This is defined through the recursion

$$Q_{n+1}(i,u) = Q_n(i,u) + a_n\Big[(1+\gamma)^{-1} \min_v Q_n(\Xi_{n+1}(i,u),v) + c(i,u) - Q_n(i,u)\Big], \qquad i \in \mathsf{X}, u \in \mathsf{U},$$

where $\Xi_{n+1}(i,u)$ is an independently simulated X-valued random variable with law $P_u(i,\cdot)$.

Making the appropriate correspondences with the stochastic approximation theory surrounding (1), we have $\theta(n) = Q_n \in \mathbb{R}^{s\times(r+1)}$ and the function $g\colon \mathbb{R}^{s\times(r+1)} \to \mathbb{R}^{s\times(r+1)}$ is defined as follows. Define $F\colon \mathbb{R}^{s\times(r+1)} \to \mathbb{R}^{s\times(r+1)}$ as $F(Q) = [F_{iu}(Q)]_{i,u}$ via,

$$F_{iu}(Q) = (1+\gamma)^{-1} \sum_j P_u(i,j) \min_v Q(j,v) \; + \; c(i,u).$$

Then $g(Q) = F(Q) - Q$ and the associated O.D.E. is

$$\tfrac{d}{dt} Q = F(Q) - Q := g(Q). \tag{15}$$

The map $F : \mathbb{R}^{s\times(r+1)} \to \mathbb{R}^{s\times(r+1)}$ is a contraction w.r.t. the max norm $\|\cdot\|_\infty$,

$$\|F(Q^1) - F(Q^2)\|_\infty \leq (1+\gamma)^{-1}\|Q^1 - Q^2\|_\infty, \qquad Q^1, Q^2 \in \mathbb{R}^{s\times(r+1)}.$$

Consequently, one can show that with $\widetilde{Q} = Q - Q^*$,

$$\tfrac{d}{dt}\|\widetilde{Q}\|_\infty \leq -\gamma(1+\gamma)^{-1}\|\widetilde{Q}\|_\infty,$$

which establishes global asymptotic stability of its unique equilibrium point θ^* [7]. Assumption (A1) holds, with the (i,u)-th component of $g_\infty(Q)$ given by

$$(1+\gamma)^{-1} \sum_j P_u(i,j) \min_v Q(j,v) - Q(i,u), \qquad i \in \mathsf{X}, u \in \mathsf{U}.$$

This also is of the form $g_\infty(Q) = F_\infty(Q) - Q$ where $F_\infty(\,\cdot\,)$ is an $\|\cdot\|_\infty$- contraction, and thus the origin is asymptotically stable for the O.D.E. (3).

We conclude that Theorems 1–4 hold for the Q-learning model.

Adaptive critic algorithm Next we consider the *adaptive critic algorithm*, which may be considered as the reinforcement learning analog of policy iteration. There are several variants of this, one of which, taken from [8], is as follows. The algorithm generates a sequence of approximations to h^* denoted $\{h_n : n \geq 0\}$, interpreted as a sequence of s-dimensional vectors. Simultaneously, it generates a sequence of randomized policies denoted $\{\phi^n\}$.

At each time n the following random variables are constructed independently of the past:

(i) For each $i \in \mathsf{X}$, $\Omega_n(i)$ is a U-valued random variable independently simulated with law $\phi^n(i)$;

(ii) For each $i \in \mathsf{X}$, $u \in \mathsf{U}$, $\Xi_n^a(i,u)$ and $\Xi_n^b(i,u)$ are independent X-valued random variables with law $P_u(i,\,\cdot\,)$.

For $1 \le \ell \le r$ we let e^ℓ is the unit r-vector in the ℓ-th coordinate direction. We let $\Gamma(\,\cdot\,)$ denote the projection onto the simplex $\{x \in \mathbb{R}^r_+ : \sum_i x_i \le 1\}$.

For $i \in \mathsf{X}$ the algorithm is defined by the pair of equations,

$$\begin{aligned} h_{n+1}(i) &= h_n(i) + b_n\big[(1+\gamma)^{-1}[c(i,\Omega_n(i)) + h_n(\Xi_n^a(i,\Omega_n(i)))] - h_n(i)\big], \qquad (16) \\ \widehat{\phi}^{n+1}(i) &= \Gamma\Big\{\widehat{\phi}^n(i) + a_n \sum_{\ell=1}^{r}\Big([c(i,u_0) + h_n(\Xi_n^b(i,u_0))] - [c(i,u_\ell) + h_n(\Xi_n^b(i,u_\ell))]\Big)\mathsf{e}^\ell\Big\}. \qquad (17) \end{aligned}$$

For each i, n, $\phi^n(i) = \phi^n(i,\,\cdot\,)$ is a probability vector on U defined in terms of $\widehat{\phi}^n(i) = [\widehat{\phi}^n(i,1),\dots,\widehat{\phi}^n(i,r)]$ as follows,

$$\phi^n(i,u_\ell) = \begin{cases} \widehat{\phi}^n(i,\ell) & \ell \ne 0; \\ 1 - \sum_{j\ne 0}\widehat{\phi}^n(i,j) & \ell = 0. \end{cases}$$

This is an example of a *two time-scale* algorithm: The sequences $\{a_n\},\{b_n\}$ are assumed to satisfy

$$\lim_{n\to\infty}\frac{a_n}{b_n} = 0,$$

as well as the usual conditions for vanishing gain algorithms,

$$\sum_n a_n = \sum_n b_n = \infty, \quad \sum_n (a_n^2 + b_n^2) < \infty.$$

To see why this is based on policy iteration, recall that policy iteration alternates between two steps: One step solves the linear system of equation (8) to compute the fixed-policy value function corresponding to the current policy. We have seen that solving (8) can be accomplished by performing the fixed-policy version of value iteration given in (10). The first step (16) in the above iteration is indeed the 'learning' or 'simulation-based stochastic approximation' analog of this fixed-policy value iteration. The second step in policy iteration updates the current policy by performing an appropriate minimization. The second iteration (17) is a particular search algorithm for computing this minimum over the simplex of probability measures on U.

The different choices of stepsize schedules for the two iterations (16), (17) induces the 'two time-scale' effect discussed in [6]. Thus the first iteration sees the policy computed by the second as nearly static, thus justifying viewing it as a fixed-policy iteration. In turn, the second sees the first as almost equilibrated, justifying the search sheme for minimization over U.

The boundedness of $\{\widehat{\phi}^n\}$ is guaranteed by the projection $\Gamma(\,\cdot\,)$. For $\{h_n\}$, the fact that $b_n = o(a_n)$ allows one to treat $\widehat{\phi}^n(i)$ as constant, say $\bar{\phi}(i)$ [8]. The appropriate O.D.E. then turns out to be

$$\tfrac{d}{dt}x = F(x) - x := g(x) \qquad (18)$$

where $F : \mathbb{R}^s \to \mathbb{R}^s$ is defined by:

$$F_i(x) = (1+\gamma)^{-1}\sum_\ell \bar{\phi}(i,u_\ell)\Big[\sum_j P_{u_\ell}(i,j)x_j + c(i,u_\ell)\Big], \qquad i \in \mathsf{X}.$$

Once again, $F(\cdot)$ is an $\|\cdot\|_\infty$-contraction and it follows that (18) is globally asymptotically stable. The limiting function $g_\infty(x)$ is again of the form $g_\infty(x) = F_\infty(x) - x$ with $F_\infty(x)$ defined so that its i-th component is

$$(1+\gamma)^{-1} \sum_\ell \bar\phi(i,u_\ell) \sum_j P_{u_\ell}(i,j)x_j.$$

We see that F_∞ is also a $\|\cdot\|_\infty$- contraction and the global asymptoyic stability of the origin for the corresponding limiting O.D.E. follows [7].

Average cost optimal control For the average cost control problem we impose the additional restriction that the chain $\boldsymbol{X}$ has a *unique* invariant probability measure under any stationary policy so that the steady state cost (11) is independent of the initial condition.

For the average cost optimal control problem the Q-learning algorithm is given by the recursion

$$Q_{n+1}(i,u) = Q_n(i,u) + a_n\Big(\min_v Q_n(\Xi_n^a(i,u),v) + c(i,u) - Q_n(i,u) - Q_n(i_0,u_0)\Big),$$

where $i_0 \in \mathsf{X}$, $a_0 \in \mathsf{U}$ are fixed a-priori. The appropriate O.D.E. now is (15) with $F(\cdot)$ redefined as $F_{iu}(Q) = \sum_j P_u(i,j)\min_v Q(j,v) + c(i,u) - Q(i_0,u_0)$. The global asymptotic stability for the unique equilibrium point for this O.D.E. has been established in [1]. Once again this fits our framework with $g_\infty(x) = F_\infty(x) - x$ for F_∞ defined the same way as F, except for the terms $c(\cdot,\cdot)$ which are dropped. We conclude that (A1) and (A2) are satisfied for this version of the Q-learning algorithm.

In [8], three variants of the adaptive critic algorithm for the average cost problem are discussed, differing only in the $\{\widehat{\phi}^n\}$ iteration. The iteration for $\{h_n\}$ is common to all and is given by

$$h_{n+1}(i) = h_n(i) + b_n[c(i,\Omega_n(i)) + h_n(\Xi_n^a(i,\Omega_n,(i))) - h_n(i) - h_n(i_0)], \qquad i \in \mathsf{X}$$

where $i_0 \in \mathsf{X}$ is a prescribed fixed state. This leads to the O.D.E. (18) with F redefined as

$$F_i(x) = \sum_\ell \bar\phi(i,u_\ell)\Big(\sum_j p_{u_\ell}(i,j)x_j + c(i,u_\ell)\Big) - x_{i_0}, \qquad i \in \mathsf{X}.$$

The global asymptotic stability of the unique equilibrium point of this O.D.E. has been established in [3, 4]. Once more, this fits our framework with $g_\infty(x) = F_\infty(x) - x$ for F_∞ defined just like F, but without the $c(\cdot,\cdot)$ terms.

References

[1] J. Abounadi, D. Bertsekas, and V. S. Borkar. Learning algorithms for Markov decision processes with average cost. *SIAM J. Control Optim.*, 40(3):681–698 (electronic), 2001.

[2] D.P. Bertsekas and J. Tsitsiklis. *Neuro-Dynamic Programming.* Atena Scientific, Cambridge, Mass, 1996.

[3] V. S. Borkar. Recursive self-tuning control of finite Markov chains. *Appl. Math. (Warsaw)*, 24(2):169–188, 1996.

[4] V. S. Borkar. Correction to: "Recursive self-tuning control of finite Markov chains". *Appl. Math. (Warsaw)*, 24(3):355, 1997.

[5] V. S. Borkar and S. P. Meyn. The O.D.E. method for convergence of stochastic approximation and reinforcement learning. *SIAM J. Control Optim.*, 38(2):447–469, 2000. (also presented at the *IEEE CDC*, December, 1998).

[6] Vivek S. Borkar. Stochastic approximation with two time scales. *Systems Control Lett.*, 29(5):291–294, 1997.

[7] Vivek S. Borkar and K. Soumyanath. An analog scheme for fixed point computation. I. Theory. *IEEE Trans. Circuits Systems I Fund. Theory Appl.*, 44(4):351–355, 1997.

[8] Vijaymohan R. Konda and Vivek S. Borkar. Actor-critic-type learning algorithms for Markov decision processes. *SIAM J. Control Optim.*, 38(1):94–123 (electronic), 1999.

[9] S. P. Meyn and R. L. Tweedie. *Markov Chains and Stochastic Stability*. Springer-Verlag, London, 1993. online: `http://black.csl.uiuc.edu/~meyn/pages/book.html`.

[10] Richard S. Sutton and Andrew G. Barto. *Reinforcement Learning: An Introduction*. MIT Press (and on-line, http://www.cs.ualberta.ca/%7Esutton/book/ebook/the-book.html), 1998.

[11] J.N. Tsitsiklis. Asynchronous stochastic approximation and Q-learning. *Machine Learning*, 16:185–202, 1994.

[12] Christopher J. C. H. Watkins and Peter Dayan. Q-learning. *Machine Learning*, 8(3-4):279–292, 1992.