Proceedings of the
American Control Conference
Baltimore, Maryland • June 1994

FP12 - 4:40

Adaptive linear quadratic control using policy iteration

Steven J. Bradtke
Computer Science Department
University of Massachusetts
Amherst, MA 01003
(413) 545-1596
bradtke@cs.umass.edu

B. Erik Ydstie
Department of Chemical Engineeering
Carnegie Mellon University
Pittsburgh, PA 15213
(412) 268-2235
ydstie+@andrew.cmu.edu

Andrew G. Barto
Computer Science Department
University of Massachusetts
Amherst, MA 01003
(413) 545-2109
barto@cs.umass.edu

Abstract

In this paper we present stability and convergence results for Dynamic Programming-based reinforcement learning applied to Linear Quadratic Regulation (LQR). The specific algorithm we analyze is based on Q-learning and it is proven to converge to the optimal controller provided that the underlying system is controllable and a particular signal vector is persistently excited. This is the first convergence result for DP-based reinforcement learning algorithms for a continuous problem.

1. Introduction

In many practical applications a stabilizing feedback control for the system may be known. In this paper we discuss the problem of how to improve this controller and, under certain circumstances, make it converge to the optimal. The approach we take can be classified as direct optimal adaptive control and it is motivated by recent research on reinforcement learning which uses the principles of Dynamic Programming (DP). DP-based reinforcement learning algorithms include Sutton's Temporal Differences methods [6], Watkins' Q-learning [7], and Werbos' Heuristic Dynamic Programming [9]. Our approach is closely related to Q-learning. We apply the method to the Linear Qudratic Regulator (LQR) problem and we show that it converges to the optimal cost if the system is controllable and a particular signal vector is persistently excited. This is the first convergence result for DP-based reinforcement learning algorithms for a continuous problem. Previous results are limited to finite-state systems, with either lookup-table or linear function approximators [6, 8].

The optimal control for an LQR problem is easily found [1] *if accurate models of the system and cost functions are available.* The problem we address is how to define an adaptive policy that converges to the optimal control *without* access to such models.

Despite the paucity of theoretical results, applications of DP based algorithms have shown promise in application to continuous state problems [5]. This paper takes a first step toward providing a theoretical grounding for continuous problems.

2. Problem Statement

Consider the discrete-time, multivariable system

$$x_{t+1} = f(x_t, u_t) = Ax_t + Bu_t \tag{1}$$

with feedback control

$$u_t = Ux_t.$$

U is chosen so that the matrix $A + BU$ has all of its eigenvalues strictly within the unit circle.

Associated with this system we assign a one step cost:

$$c_t = c(x_t, u_t) = x_t'Ex_t + u_t'Fu_t \tag{2}$$

This work was supported by the Air Force Office of Scientific Research, Bolling AFB, under Grant AFOSR-89-0526 and by the National Science Foundation under Grant ECS-8912623.

where E is a symmetric positive semidefinite matrix and F is a symmetric positive definite matrix. The *total cost* of a state x_t under the control policy U, $V_U(x_t)$, is defined as the discounted sum of all costs that will be incurred by using U from time t onward, *i.e.*, $V_U(x_t) = \sum_{i=0}^{\infty} \gamma^i c_{t+i}$, where $0 \leq \gamma \leq 1$ is the discount factor. V_U is a quadratic function [1] and therefore can be expressed as

$$V_U(x_t) = x_t' K_U x_t, \qquad (3)$$

where K_U is the symmetric *cost matrix* for policy U. U^* denotes the policy which is optimal in the sense that the total dicounted cost of every state is minimized. K^* represents the cost matrix for U^*.

3. Q-functions and Policy improvement

Watkins [7] defined the Q-function for a stable control policy U as

$$Q_U(x,u) = c(x,u) + \gamma V_U(f(x,u)). \qquad (4)$$

The value $Q_U(x,u)$ is the sum of the one step cost incurred by taking action u from state x, plus the total cost that would accrue if the fixed policy U were followed from the state $f(x,u)$ and all subsequent states. u need not be the action specified by the given control policy for the state x. $Q_U(x,u)$ is defined for all states x and *all* admissible control signals u. The function Q_U can also be defined recursively as

$$Q_U(x_t,u_t) = c(x_t,u_t) + \gamma Q_U(x_{t+1}, Ux_{t+1}). \qquad (5)$$

For an LQR problem the Q-function can be computed explicitly as

$$\begin{aligned} Q_U(x,u) &= \left[\begin{array}{c} x, u \end{array}\right] \left[\begin{array}{cc} H_{U(11)} & H_{U(12)} \\ H_{U(21)} & H_{U(22)} \end{array}\right] \left[\begin{array}{c} x, u \end{array}\right]' \quad (6) \\ &= \left[\begin{array}{c} x, u \end{array}\right]' H_U \left[\begin{array}{c} x, u \end{array}\right], \qquad (7) \end{aligned}$$

where $[x,u]$ is the column vector concatenation of x and u, H_U is a symmetric positive definite matrix, and

$$\begin{aligned} H_{U(11)} &= E + \gamma A' K_U A, & H_{U(12)} &= \gamma A' K_U B, \\ H_{U(21)} &= \gamma B' K_U A, & H_{U(22)} &= F + \gamma B' K_U B. \end{aligned}$$

The submatrix $H_{U(22)}$ is symmetric positive definite.

Given the policy U_k and the value function V_U, we can find an improved policy, U_{k+1}, by defining U_{k+1} as

$$U_{k+1}x = \operatorname*{argmin}_u \left[c(x,u) + \gamma V_U(f(x,u))\right].$$

But equation (4) tells us that this can be rewritten as

$$U_{k+1}x = \operatorname*{argmin}_u Q_U(x,u).$$

We can find the minimizing u by taking the partial derivative of $Q_U(x,u)$ with respect to u, setting that to zero, and solving for u. This yields

$$u = \underbrace{-\gamma \left(F + \gamma B' K_{U_k} B\right)^{-1} B' K_{U_k} A}_{U_{k+1}} x.$$

Since the new policy U_{k+1} does not depend on x, it is the minimizing policy for all x. Using (6), U_{k+1} can be written as

$$U_{k+1} = -H^{-1}_{U_k(22)} H_{U_k(21)}.$$

The feedback policy U_{k+1} is per definition a stabilizing policy – it has no higher cost than U_k. A new Q-function can then be assigned to this policy and the policy improvement procedure can be repeated *ad infinitum.*

Earlier work by Kleinman [4] and Bertsekas [1] showed that policy iteration will converge for LQR problems. However, their algorithms required exact knowledge of the system model (equation 1) and the one-step cost function (equation 2). The analysis presented in this paper shows how policy iteration can be performed *without* that knowledge. Knowledge of the sequence of functions Q_{U_k} is sufficient.

4. Direct Estimation of Q-functions

We use Recursive Least Squares (RLS) to directly estimate the function Q_U. It is not necessary to identify either the system model or the one-step cost function separately. First, define the "overbar" function for vectors so that $\overline{x}$ is the vector whose elements are all of the quadratic basis functions over the elements of x, *i.e.*,

$$\overline{x} = \left[x_1^2, \ldots, x_1 x_n, x_2^2, \ldots, x_2 x_n, \ldots, x_n^2\right]'.$$

Next, define the function Θ for square matrices. $\Theta(K)$ is the vector whose elements are the n diagonal entries of K and the $n(n+1)/2 - n$ distinct sums $(K_{ij} + K_{ji})$. The elements of $\overline{x}$ and $\Theta(K)$ are ordered so that $x'Kx = \overline{x}'\Theta(K)$. The original matrix K can be retrieved from $\Theta(K)$ if K is symmetric. If K is not symmetric, then we retrieve the symmetric matrix $\frac{1}{2}(K + K')$, which defines the same quadratic function as K. We can now write

$$Q_U(x,u) = \left[\begin{array}{c} x, u \end{array}\right]' H_U \left[\begin{array}{c} x, u \end{array}\right] = \overline{\left[\begin{array}{c} x, u \end{array}\right]}' \Theta(H_U).$$

Finally, we rearrange equation (5) to yield

$$\begin{aligned} c(x_t,u_t) &= Q_U(x_t,u_t) - \gamma Q_U(x_{t+1}, Ux_{t+1}) \\ &= \overline{\left[\begin{array}{c} x_t, u_t \end{array}\right]}' \Theta(H_U) - \gamma \overline{[x_{t+1}, Ux_{t+1}]}' \Theta(H_U) \\ &= \phi_t' \theta_U, \end{aligned}$$

where $\phi_t = \left\{\overline{[x_t, u_t]} - \gamma\overline{[x_{t+1}, Ux_{t+1}]}\right\}$, and $\theta_{\text{U}} = \Theta(H_{\text{U}})$.

RLS can now be used to estimate θ_{U}. The recurrence relations for RLS are given by

$$e_k(i) = c_t - \phi_t'\hat{\theta}_k(i-1) \tag{8a}$$

$$\hat{\theta}_k(i) = \hat{\theta}_k(i-1) + \frac{P_k(i-1)\phi_t e_k(i)}{1+\phi_t' P_k(i-1)\phi_t} \tag{8b}$$

$$P_k(i) = P_k(i-1) - \frac{P_k(i-1)\phi_t\phi_t' P_k(i-1)}{1+\phi_t' P_k(i-1)\phi_t} \tag{8c}$$

$$P_k(0) = P_0. \tag{8d}$$

$P_0 = \beta I$ for some large positive constant β. $\theta_k = \Theta(H_{\text{U}_k})$ is the true parameter vector for the function Q_{U_k}. $\hat{\theta}_k(i)$ is the i^{th} estimate of θ_k. The subscript t and the index i are both incremented at each time step. The reason for the distinction between t and i will be made clear in the next section.

Goodwin and Sin [3] show that this algorithm converges asymptotically to the true parameters if θ_k is fixed and ϕ_t satisfies the persistent excitation condition

$$\epsilon_0 I \le \frac{1}{N}\sum_{i=1}^{N} \phi_{t-i}\phi_{t-i}' \le \bar{\epsilon}_0 I, \tag{9}$$

for all $t \ge N_0$ and $N \ge N_0$, where $\epsilon_0 \le \bar{\epsilon}_0$, and N_0 is a positive number.

5. Adaptive Policy Iteration for LQR

The policy improvement process based on Q-functions (Section 3) and the ability to directly estimate H_{U} (Section 4) are the two key elements of the adaptive policy iteration algorithm that is the focus of this paper. Figure 1 gives an outline of the algorithm. The index i used in equations (8) counts the number of time steps since the beginning of the estimation interval.

Since the k^{th} policy improvement step is based on an *estimate* of $\Theta(H_{\text{U}_k})$, it is not clear *a priori* that the sequence U_k will converge to the optimal policy U^*, or even that each of the U_k's is guaranteed to be stabilizing. The convergence proofs of Kleinman [4] and Bertsekas [1] require exact knowledge of the system and take no account of estimation error. Theorem 1 establishes that the adaptive policy iteration algorithm presented above does indeed converge, under certain conditions, to the optimal controller.

```
Initialize parameters: $\hat{\theta}_1(0)$.
$t = 0$
for $k = 0$ to $\infty$ {
    Initialize RLS: $P_k(0) = P_0$.
    for $i = 1$ to N {
        $u_t = U_k x_t + e_t$, where $e_t$ is the "exploration" component of the control signal.
        Apply $u_t$, resulting in state $x_{t+1}$.
        Update $\hat{\theta}_k(i)$ using RLS (8).
        $t = t + 1$.
    }
    Find the matrix $\hat{H}_k$ corresponding to $\hat{\theta}_k$.
    Policy improvement: $U_{k+1} = -\hat{H}_{\text{k}(22)}^{-1}\hat{H}_{\text{k}(21)}$.
    Initialize parameters: $\hat{\theta}_{k+1}(0) = \hat{\theta}_k$.
}
```

Figure 1: The Q-function based policy iteration algorithm. It starts with the system in some initial state x_0 and with some stabilizing controller U_0. k is the number of policy iteration steps. t is the total number of time steps. i is the number of time steps since the last policy change.

Theorem 1: (Convergence of adaptive policy iteration). *Suppose that $\{A, B\}$ is a controllable pair, that U_0 is a stabilizing control, and that the vector $\phi(t)$ is persistently excited according to inequality (9). Then there exists an estimation interval $N < \infty$ so that the adaptive policy iteration mechanism described above generates a sequence $\{U_k, k = 1, 2, 3, ...\}$ of stabilizing controls, converging so that*

$$\lim_{k\to\infty} ||U_k - U^*|| = 0,$$

where U^* is the optimal feedback control matrix.

Proof: In order to prove this we need a few intermediate results concerning the policy iteration scheme and RLS estimation. These preliminary results are summarized below and the proofs are given in [2]. First, define the function

$$\sigma(U_k) = \text{trace}(K_{\text{U}_k}). \tag{10}$$

Lemma 1. *If $\{A, B\}$ is controllable, U_1 is stabilizing with associated cost matrix K_1 and U_2 is the result of one policy improvement step from U_1, i.e. $U_2 = -\gamma(F + \gamma B'K_1B)^{-1}B'K_1A$, then*

$$\Delta||U_1 - U_2||^2 \le \sigma(U_1) - \sigma(U_2) \le \delta||U_1 - U_2||^2,$$

where

$$0 < \Delta = \underline{\sigma}(F) \le \delta = \\ \text{trace}\,(F + \gamma B'K_1B)||\sum_{i=0}^{\infty}\gamma^{(i/2)}(A + BU_2)^i||^2,$$

and $\underline{\sigma}(\cdot)$ denotes the minimum singular value of a matrix.

Lemma 2. *If ϕ_t is persistently excited as given by inequaliy (9) and $N \ge N_0$, then we have*

$$||\theta_k - \hat{\theta}_k|| \le \epsilon_N(||\theta_k - \theta_{k-1}|| + ||\theta_{k-1} - \hat{\theta}_{k-1}||),$$

where $\epsilon_N = \frac{1}{\epsilon_0 N p_0}$ *and* p_0 *is the minimum singular value of* P_0.

Define a scalar "Lyapunov" function candidate

$$s_k = \sigma(U_{k-1}) + ||\theta_{k-2} - \hat{\theta}_{k-2}|| \qquad (11)$$

and suppose that

$$s_i \leq \bar{s}_0 < \infty \qquad \text{for all} \qquad 0 \leq i \leq k \qquad (12)$$

for some upper bound $\bar{s}_0$. From this it follows that U_{k-1} is stabilizing in the sense that

$$\sigma(U_{k-1}) \leq \bar{s}_0 \qquad (13)$$

and that the parameter estimation error is bounded so that

$$||\theta_{k-2} - \hat{\theta}_{k-2}|| \leq \bar{s}_0. \qquad (14)$$

It also follows that that the control resulting from a policy update using accurate parameters, U_k^*, is stabilizing and that $\sigma(U_k^*) \leq \bar{s}_0$. From continuity of the optimal policy update it then follows that for every $\delta > 0$ there exists $\epsilon_\delta > 0$ so that

$$|\sigma(U) - \sigma(U_k^*)| \leq \delta||U_k^* - U|| \qquad (15)$$

for all $||U_k^* - U|| \leq \epsilon_\delta$. This implies that control laws in a sufficiently small neighborhood around the optimal are stabilizing as well.

We will show that $s_{k+1} \leq s_k$ provided that the estimation interval N, is chosen to be long enough.

Define

$$v_k = ||\theta_{k-1} - \hat{\theta}_{k-1}||,$$

and we get from Lemma 2 that for all k

$$v_k \leq \epsilon_N(v_{k-1} + ||\theta_{k-1} - \theta_{k-2}||), \qquad (16)$$

where $\lim_{N\to\infty} \epsilon_N = 0$. Now from the inductive hypothesis (assumption (12)) we have

$$v_{k-1} \leq \bar{s}_0 \qquad \text{and} \qquad ||\theta_{k-2} - \theta_{k-3}|| \leq \kappa_1, \qquad (17)$$

where κ_1 is a constant. By application of (16) we then get

$$v_k \leq \epsilon_N(\bar{s}_0 + \kappa_1). \qquad (18)$$

It follows that $v_k = ||\theta_{k-1} - \hat{\theta}_{k-1}||$ can be made arbitrarily small be choosing the estimation interval N long enough.

U_k^* is defined to be the result from applying one step of policy iteration using accurate parameter values, *i.e.*

$$U_k^* = -H_{k\text{-}1(22)}^{-1} H_{k\text{-}1(21)}, \qquad (19)$$

whereas U_k is the feedback law which results from applying the estimated parameters, *i.e.*

$$U_k = -\hat{H}_{k\text{-}1(22)}^{-1} \hat{H}_{k\text{-}1(21)}. \qquad (20)$$

The matrix inverse is guaranteed to exist when the estimation interval is long enough. From equations (19) and (20) we now have

$$U_k - U_k^* = -\hat{H}_{k\text{-}1(22)}^{-1} \hat{H}_{k\text{-}1(21)} + H_{k\text{-}1(22)}^{-1} H_{k\text{-}1(21)}.$$

Hence

$$\begin{aligned} U_k - U_k^* &= H_{k\text{-}1(22)}^{-1}(H_{k\text{-}1(21)} - \hat{H}_{k\text{-}1(21)}) + \\ &\quad (H_{k\text{-}1(22)}^{-1} - \hat{H}_{k\text{-}1(21)}^{-1})\hat{H}_{k\text{-}1(21)} \\ &= H_{k\text{-}1(22)}^{-1}\{(H_{k\text{-}1(21)} - \hat{H}_{k\text{-}1(21)}) + \\ &\quad (\hat{H}_{k\text{-}1(22)} - H_{k\text{-}1(22)})\hat{H}_{k\text{-}1(22)}^{-1}\hat{H}_{k\text{-}1(21)}\}. \end{aligned}$$

From the definition of θ we have

$$||\hat{H}_{k\text{-}1(22)} - H_{k\text{-}1(22)}|| \leq ||\theta_{k-1} - \hat{\theta}_{k-1}|| \qquad \text{and}$$
$$||\hat{H}_{k\text{-}1(22)}|| \leq ||\hat{\theta}_{k-1}||.$$

It follows that we have

$$||U_k - U_k^*|| \leq \bar{\kappa}_0(1 + ||\hat{\theta}_{k-1}||) \cdot ||\theta_{k-1} - \hat{\theta}_{k-1}||,$$

where $\bar{\kappa}_0$ is a positive constant, provided that N is sufficiently large. Since the estimated parameters are bounded it follows that there exists another constant κ_0 so that

$$||U_k - U_k^*|| \leq \kappa_0||\theta_{k-1} - \hat{\theta}_{k-1}|| = \kappa_0 v_k. \qquad (21)$$

It follows from equation (18) that we have

$$||U_k - U_k^*|| \leq \epsilon_N \kappa_0(\bar{s}_0 + \kappa_1). \qquad (22)$$

It then follows from (15) that

$$|\sigma(U_k) - \sigma(U_k^*)| \leq \delta||U_k^* - U_k||$$

for all N such that $\epsilon_N\kappa_0(\bar{s}_0 + \kappa_1) \leq \epsilon_\delta$. This implies that U_k is stabilizing if N is large enough and that there exists an integer N_1 and an associated constant $\bar{\delta}$, so that

$$|\sigma(U_k) - \sigma(U_{k-1})| \leq \bar{\delta}||U_k - U_{k-1}|| \quad \text{for all} \quad N \geq N_1.$$

In other words, if the estimation interval is long enough, then the difference between two consecutive costs is bounded by the difference between two consecutive controls. We use the definition of the parameter estimation vector to write this as

$$||\theta_{k-1} - \theta_{k-2}|| \leq \delta_1||U_k - U_{k-1}||^2 \quad \text{for all} \quad N \geq N_1, \qquad (23)$$

where δ_1 is a constant. We now re-write (23) as

$$||\theta_{k-1} - \theta_{k-2}|| \leq 2\delta_1(||U_k^* - U_{k-1}||^2 + ||U_k^* - U_k||^2).$$

From inequality (21) and the definition of v_k, we then get

$$||\theta_{k-1} - \theta_{k-2}|| \leq 2\delta_1(w_k^2 + \kappa_0 v_k), \qquad (24)$$

where

$$w_k = ||U_k^* - U_{k-1}||.$$

By combining equations (16) and (24) we then get

$$v_k \leq \epsilon_N(v_{k-1} + 2\delta_1(w_k^2 + \kappa_0 v_k)),$$

which we re-write as

$$v_k \leq \epsilon_N \mu_N (v_{k-1} + 2\delta_1 w_k^2), \tag{25}$$

where

$$\mu_N = (1 - 2\delta_1 \kappa_0 \epsilon_N)^{-1}.$$

According to the assumption we can choose N large enough so that $0 < \mu_N < \infty$. This gives a recursion for v_k. The critical point to notice is that v_k has a strong stability property when the estimation interval is long. The parameter $\epsilon_N \mu$ is then small since ϵ_N converges uniformly to 0 and μ towards 1.

We now develop the recursion for $\sigma(U_k)$. First we have

$$\sigma(U_k) - \sigma(U_{k-1}) = \sigma(U_k^*) - \sigma(U_{k-1}) + \sigma(U_k) - \sigma(U_k^*). \tag{26}$$

From equation (26) and Lemma 1, using (15) again, it follows that we can choose the update interval so that we have a constant δ_2 so that

$$\sigma(U_k) - \sigma(U_{k-1}) \leq -\Delta ||U_k^* - U_{k-1}||^2 + \delta_2 ||U_k^* - U_k||^2.$$

Using equation (21) we then get

$$\begin{aligned} &\sigma(U_k) - \sigma(U_{k-1}) \\ &\leq -\Delta ||U_k^* - U_{k-1}||^2 + \delta_2 \kappa_0 ||\theta_{k-1} - \hat{\theta}_{k-1}|| \\ &\leq -\Delta w_k^2 + \delta_2 \kappa_0 v_k. \end{aligned}$$

By using equation (25) and the recursion for v_k we then have

$$\sigma(U_k) - \sigma(U_{k-1}) \leq -\Delta w_k^2 + \delta_1 \kappa_0 \epsilon_N \mu_N (v_{k-1} + 2\delta_2 w_k^2). \tag{27}$$

Equations (25) and (27) together define the system

$$\begin{bmatrix} v_k \\ \sigma(U_k) \end{bmatrix} = \begin{bmatrix} \epsilon_N \mu_N & 0 \\ \delta_2 \kappa_0 \epsilon_N \mu_N & 1 \end{bmatrix} \begin{bmatrix} v_{k-1} \\ \sigma(U_{k-1}) \end{bmatrix} + \begin{bmatrix} 2\epsilon_N \mu_N \delta_2 \\ -\Delta + 2\delta_2 \kappa_0 \epsilon_N \mu_N \end{bmatrix} w_k^2.$$

In order to study this system we defined the function

$$s_k = \sigma(U_{k-1}) + v_{k-1}.$$

From the above we then have

$$\begin{aligned} s_{k+1} = s_k + (-1 + \epsilon_N \mu_N (1 + \delta_2 \kappa_0)) v_{k-1} + \\ (-\Delta + 2\epsilon_N \mu_N \delta_2 (1 + \kappa_0)) w_k^2. \end{aligned}$$

It now suffices to choose N so that ϵ is small enough to give

$$\begin{aligned} 1 - \epsilon_N \mu_N (1 + \delta_2 \kappa_0) = \epsilon_1 > 0 \\ \Delta - 2\epsilon_N \mu_N \delta_2 (1 + \kappa_0) = \epsilon_2 > 0. \end{aligned}$$

We then get

$$s_{k+1} = s_k - \epsilon_1 v_{k-1} - \epsilon_2 w_k^2 \leq s_k.$$

From this we conclude that $s_{k+1} \leq s_k$ and using induction we finally have

$$\epsilon_1 \sum_{k=1}^{\infty} v_k \leq \bar{s}_0 \qquad \text{and} \qquad \epsilon_2 \sum_{k=1}^{\infty} w_k^2 \leq \bar{s}_0.$$

The result now follows since U_0 is stabilizing.

6. Conclusions

In this paper we take a first step toward extending the theory of DP-based reinforcement learning to continuous domains. We concentrate on the problem of Linear Quadratic Regulation. We describe a policy iteration algorithm for LQR problems that is proven to converge to the optimal policy. In contrast to standard methods of policy iteration, it does not require a system model. It only requires a suitably accurate estimate of H_{U_k}. This is the first theoretical result of which we are aware that proves convergence of a DP-based reinforcement learning algorithm in a continuous domain.

References

[1] D. P. Bertsekas. *Dynamic Programming: Deterministic and Stochastic Models.* Prentice Hall, Englewood Cliffs, NJ, 1987.

[2] S. J. Bradtke. *Incremental Dynamic Programming for On-Line Adaptive Optimal Control.* PhD thesis, University of Massachusetts at Amherst, May, 1994.

[3] G. C. Goodwin and K. S. Sin. *Adaptive filtering prediction and control.* Prentice-Hall, Inc., Englewood Cliffs, New Jersey, 1984.

[4] D. L. Kleinman. On an iterative technique for Riccati equation computations. *IEEE Transactions on Automatic Control,* pages 114–115, February 1968.

[5] D. A. Sofge and D. A. White. Neural network based process optimization and control. In *Proceedings of the 29th Conference on Decision and Control,* Honolulu, Hawaii, December 1990.

[6] R. S. Sutton. Learning to predict by the method of temporal differences. *Machine Learning,* 3:9–44, 1988.

[7] C. J. C. H. Watkins. *Learning from Delayed Rewards.* PhD thesis, Cambridge University, Cambridge, England, 1989.

[8] C. J. C. H. Watkins and P. Dayan. Q-learning. *Machine Learning,* 8(3/4):257–277, May 1992.

[9] P. J. Werbos. Building and understanding adaptive systems: A statistical/numerical approach to factory automation and brain research. *IEEE Transactions on Systems, Man, and Cybernetics,* 17(1):7–20, 1987.