PERGAMON

Automatica 35 (1999) 1799–1808

www.elsevier.com/locate/automatica

Brief paper

Average cost temporal-difference learning[☆]

John N. Tsitsiklis*, Benjamin Van Roy

Laboratory for Information and Decision Systems, Room 35-209, 77 Massachusetts Avenue, Massachusetts Institute of Technology, Cambridge, MA 02139-4307, USA

Received 23 March 1997; received in final form 12 February 1999

Abstract

We propose a variant of temporal-difference learning that approximates average and differential costs of an irreducible aperiodic Markov chain. Approximations are comprised of linear combinations of fixed basis functions whose weights are incrementally updated during a single endless trajectory of the Markov chain. We present a proof of convergence (with probability 1) and a characterization of the limit of convergence. We also provide a bound on the resulting approximation error that exhibits an interesting dependence on the "mixing time" of the Markov chain. The results parallel previous work by the authors, involving approximations of discounted cost-to-go. © 1999 Elsevier Science Ltd. All rights reserved.

Keywords: Dynamic programming; Learning; Average cost; Reinforcement learning; Neuro-dynamic programming; Approximation; Temporal differences

1. Introduction

Temporal-difference (TD) learning, as proposed by Sutton (1988), is an algorithm for approximating the cost-to-go function of a Markov chain (the expected future cost, as a function of the initial state) by a linear combination of a given collection of basis functions, on the basis of simulation or observation of the process. Such approximations are used primarily in approximate policy iteration methods for large-scale Markov decision problems, when the size of the state space is too large to allow exact computation of the cost-to-go function (Bertsekas & Tsitsiklis, 1996).

A comprehensive convergence analysis for the case of discounted Markov chains has been provided by the authors (Tsitsiklis & Van Roy, 1997). A simplified version of that work, together with extensions to the case of undiscounted absorbing Markov chains, is presented in (Bertsekas & Tsitsiklis, 1996). Related analyses are given by (Sutton, 1988; Dayan, 1992; Gurvits, Lin & Hansen, 1994), and (Pineda, 1996). The purpose of the present paper is to propose and analyze a variant of TD learning that is suitable for approximating differential cost functions of undiscounted Markov chains (i.e., solutions to Poisson's equation). The results parallel those available for the discounted case: we have convergence (with probability 1), a characterization of the limit, and graceful bounds on the resulting approximation error. Furthermore, a relationship between error bounds and mixing properties of the Markov chain are identified.

We note that the simulation-based (or reinforcement learning) methods pertinent to the average cost formulation that have been studied in the literature generally involve look-up table representations, which store and update one value per state in the state space; see (Mahadevan, 1996) for a survey of relevant experimental work and (Abounadi, 1998) for a theoretical treatment. In the context of approximations, the common practice is to use a discounted formulation as a proxy for an average cost problem. (The discount factor is usually set very close to unity, which can lead to numerical difficulties.) Our results show that this practice is unnecessary, as has already been illustrated in a successful application to

[☆] This research was supported by the NSF under grant DMI-9625489 and the AFOSR under grant F49620-95-1-0219. This paper was not presented at any IFAC meeting. This paper was recommended for publication in revised form by Associate Editor Arun Bagchi under the direction of Editor T. Basar.

* Corresponding author. Tel.: 001-617-253-6175; fax: 001-617-258-7336.

E-mail addresses: jnt@mit.edu (J.N. Tsitsiklis), bvr@stanford.edu (B. Van Roy)

PII: S0005-1098(99)00099-0

a large-scale problem (Marbach, Mihatsch & Tsitsiklis, 1998).

2. Average cost temporal-difference learning

We consider a Markov chain $\{i_t \mid t = 0, 1, \ldots\}$ on a finite state space $S = \{1, \ldots, n\}$, with transition probability matrix P.

Assumption 1. *The Markov chain corresponding to P is irreducible and aperiodic.*

It follows that the Markov chain has a unique invariant probability distribution π, that satisfies $\pi' P = \pi'$, with $\pi(i) > 0$ for all i. Let $E_0[\cdot]$ denote expectation with respect to this distribution.

Let $g(i)$ be a cost per stage associated with state i. We define the average cost by $\mu^* = E_0[g(i_t)]$, and a differential-cost function as any function $J: S \mapsto \Re$ satisfying Poisson's equation, which takes the form

$$J = g - \mu^* e + PJ,$$

where $e \in \Re^n$ is the vector with each component equal to 1, and J and g are viewed as vectors in $\Re^n$. Under Assumption 1, it is known that differential cost functions exist and the set of all differential cost functions takes the form $\{J^* + ce \mid c \in \Re\}$, for some function J^* satisfying $\pi' J^* = 0$ (see, e.g., Gallager, 1996). We will refer to J^* as the *basic* differential cost function, and it is known that, under Assumption 1, this function is given by

$$J^* = \sum_{t=0}^{\infty} P^t (g - \mu^* e). \tag{1}$$

We consider approximations to differential cost functions using a function of the form

$$\tilde{J}(i, r) = \sum_{k=1}^{K} r(k) \phi_k(i).$$

Here, $r = (r(1), \ldots, r(K))'$ is a tunable parameter vector and each ϕ_k is a basis function defined on the state space S to be viewed as a vector of dimension n.

It is convenient to define a vector-valued function $\phi: S \mapsto \Re^K$, by letting $\phi(i) = (\phi_1(i), \ldots, \phi_K(i))'$. With this notation, the approximation can also be written in the form $\tilde{J}(i,r) = r'\phi(i)$ or $\tilde{J}(r) = \Phi r$, where Φ is an $n \times K$ matrix whose kth column is equal to ϕ_k.

Assumption 2. (a) *The basis functions $\{\phi_k \mid k = 1, \ldots, K\}$ are linearly independent. In particular, $K \leq n$ and Φ has full rank.*
(b) *For every $r \in \Re^K$, $\Phi r \neq e$.*

Suppose that we observe a sequence of states i_t generated according to the transition probability matrix P. Given that at a time t, the parameter vector r has been set to some value r_t, and we have an approximation μ_t to the average cost μ^*, we define the temporal difference d_t corresponding to the transition from i_t to i_{t+1} by

$$d_t = g(i_t) - \mu_t + \tilde{J}(i_{t+1}, r_t) - \tilde{J}(i_t, r_t). \tag{2}$$

The TD(λ) algorithm that we will be studying updates r_t and μ_t according to

$$\mu_{t+1} = (1 - \eta_t)\mu_t + \eta_t g(i_t),$$

and

$$r_{t+1} = r_t + \gamma_t d_t \sum_{k=0}^{t} \lambda^{t-k} \phi(i_k), \tag{3}$$

where γ_t and η_t are scalar step sizes and λ is a parameter in [0,1). It is convenient to define a sequence of *eligibility vectors* z_t (of dimension K) by

$$z_t = \sum_{k=0}^{t} \lambda^{t-k} \phi(i_k). \tag{4}$$

With this new notation, the parameter updates are given by

$$r_{t+1} = r_t + \gamma_t d_t z_t,$$

and the eligibility vectors can be updated according to

$$z_{t+1} = \lambda z_t + \phi(i_{t+1}),$$

initialized with $z_{-1} = 0$.

Assumption 3. (a) *The sequence γ_t is positive, deterministic, and satisfies $\sum_{t=0}^{\infty} \gamma_t = \infty$ and $\sum_{t=0}^{\infty} \gamma_t^2 < \infty$.*
(b) *There exists a positive scalar c such that the sequence η_t satisfies $\eta_t = c\gamma_t$, for all t.*

3. Convergence result

We begin with some notation that helps to streamline the formal statement of results, as well as the analysis.

Recall that $\pi(1), \ldots, \pi(n)$ denote the steady-state probabilities for the process i_t. We define an $n \times n$ diagonal matrix D with diagonal entries $\pi(1), \ldots, \pi(n)$. It is easy to see that $\langle x, y \rangle_D = x' D y$ defines an inner product space with norm $\| \cdot \|_D = \sqrt{\langle \cdot, \cdot \rangle_D}$. To interpret this norm, note that for every $J: S \mapsto \Re$, we have

$$\|J\|_D^2 = E_0[J(i_t)^2].$$

We say that two vectors $J, \bar{J}$ are D-orthogonal if $J' D \bar{J} = 0$. We will also use $\| \cdot \|$, without a subscript, to denote the Euclidean norm on vectors or the Euclidean-induced norm on matrices. (That is, for any matrix A, we have $\|A\| = \max_{\|x\|=1} \|Ax\|$.)

We define a projection matrix Π that projects onto the subspace spanned by the basis functions. In particular, we let $\Pi = \Phi(\Phi' D\Phi)^{-1}\Phi' D$. For any $J \in \Re^n$, we then have

$$\Pi J = \operatorname*{argmin}_{\bar{J} \in \{\Phi r \mid r \in \Re^K\}} \|J - \bar{J}\|_D.$$

For any $\lambda \in [0,1)$, we define an operator $T^{(\lambda)}: \Re^n \mapsto \Re^n$ by

$$T^{(\lambda)}J = (1-\lambda)\sum_{m=0}^{\infty}\lambda^m\left(\sum_{t=0}^{m}P^t(g-\mu^* e) + P^{m+1}J\right).$$

To interpret $T^{(\lambda)}$ in a meaningful manner, note that, for each m, the term

$$\sum_{t=0}^{m}P^t(g-\mu^* e) + P^{m+1}J$$

is an approximation to the basic differential cost function where the summation in Eq. (1) is truncated after m terms, and the remainder of the summation is approximated by $P^{m+1}J$. In fact, the remainder of the summation is exactly equal to $P^{m+1}J^*$, so $P^{m+1}J$ is a reasonable approximation when J^* is unknown and J is its estimate. The function $T^{(\lambda)}J$ is therefore a geometrically weighted average of approximations to the differential cost function.

Our convergence result follows.

Theorem 1. *Under Assumptions* 1–3, *the following hold*:

(a) *For any* $\lambda \in [0,1)$, *the average cost TD*(λ) *algorithm*, *as defined in Section* 2, *converges with probability* 1.
(b) *The limit of the sequence* μ_t *is the average cost* μ^*.
(c) *The limit* r^* *of the sequence* r_t *is the unique solution of the equation*

$$\Pi T^{(\lambda)}(\Phi r^*) = \Phi r^*.$$

3.1. Preliminaries

In order to represent the algorithm in a compact form, we construct a process $X_t = (i_t, i_{t+1}, z_t)$, where z_t is the eligibility vector defined by Eq. (4). It is easy to see that X_t is a Markov process. In particular, z_{t+1} and i_{t+1} are deterministic functions of X_t, and the distribution of i_{t+2} only depends on i_{t+1}. Note that at each time t, the random vector X_t, together with the current values of μ_t and r_t, provides all necessary information for computing μ_{t+1} and r_{t+1}.

So that we can think of the TD(λ) algorithm as adapting only a single vector, we introduce a sequence $\theta_t \in \Re^{K+1}$ with components $\theta_t(1) = \mu_t$ and $\theta_t(i) = r_t(i-1)$ for $i \in \{2,\ldots,n+1\}$, or using more compact notation,

$$\theta_t = \begin{bmatrix}\mu_t\\ r_t\end{bmatrix}.$$

The TD(λ) updates can be rewritten as

$$\theta_{t+1} = \theta_t + \gamma_t(A(X_t)\theta_t + b(X_t)), \qquad (5)$$

where for any $X = (i,j,z)$, we have

$$A(X) = \begin{bmatrix}-c & 0\cdots 0\\ -z & z(\phi'(j)-\phi'(i))\end{bmatrix},$$

$$b(X) = \begin{bmatrix}cg(i)\\ zg(i)\end{bmatrix},$$

and c is the constant in Assumption 3(b).

As we will show later, $A(X_t)$ and $b(X_t)$ have well-defined "steady-state" expectations, which we denote by A and b. General results concerning stochastic approximation algorithms can be used to show that the asymptotic behavior of the sequence generated by Eq. (5) mimics that of an ordinary differential equation:

$$\dot{\theta}_t = A\theta_t + b.$$

Our analysis can be broken down into two parts. The first establishes that the relevant ordinary differential equation converges (we will show that the matrix A is stable). The second involves the application of a result from stochastic approximation theory to show that the algorithm delivers similar behavior.

3.2. Lemmas

We start with an easy consequence of Jensen's inequality, which is central to our analysis; see Lemma 1 in (Tsitsiklis & Van Roy, 1997) for a proof.

Lemma 1. *Under Assumption* 1, *for all* $J \in \Re^n$,

$$\|PJ\|_D \le \|J\|_D.$$

Furthermore, unless J is proportional to e, we have $PJ \neq J$.

Under Assumption 1, the matrix $P^{(\lambda)}$ defined below is an irreducible and aperiodic stochastic matrix, and Lemma 2 follows from Lemma 1.

Lemma 2. *Let* $P^{(\lambda)} = (1-\lambda)\sum_{m=0}^{\infty}\lambda^m P^{m+1}$. *Then, under Assumption* 1, *for any* $\lambda \in [0,1)$ *and* $J \in \Re^n$,

$$\|P^{(\lambda)}J\|_D \le \|J\|_D.$$

Furthermore, unless J is proportional to e, we have $P^{(\lambda)}J \neq J$.

We now establish that the set of fixed points of $T^{(\lambda)}$ is the set of differential cost functions.

Lemma 3. *Under Assumption* 1, *for any* $\lambda \in [0,1)$, *we have*

$$T^{(\lambda)}J = J \quad \textit{if and only if } J \in \{J^* + ce \mid c \in \Re\}.$$

Proof. Suppose that $J = J^* + ce$, for some scalar c. Then,

$$\begin{aligned}
T^{(\lambda)}J &= (1-\lambda)\sum_{m=0}^{\infty}\lambda^m\left(\sum_{t=0}^{m}P^t(g-\mu^*e)+P^{m+1}(J^*+ce)\right)\\
&= (1-\lambda)\sum_{m=0}^{\infty}\lambda^m\left(\sum_{t=0}^{m}P^t(g-\mu^*e)+P^{m+1}J^*\right)+ce\\
&= (1-\lambda)\sum_{m=0}^{\infty}\lambda^m\left(\sum_{t=0}^{m}P^t(g-\mu^*e)\right.\\
&\qquad \left.+P^{m+1}\sum_{t=0}^{\infty}P^t(g-\mu^*e)\right)+ce\\
&= (1-\lambda)\sum_{m=0}^{\infty}\lambda^m\sum_{t=0}^{\infty}P^t(g-\mu^*e)+ce\\
&= J^*+ce\\
&= J.
\end{aligned}$$

On the other hand, suppose that J is not of the form $J^* + ce$. Then,

$$\begin{aligned}
T^{(\lambda)}J &= T^{(\lambda)}J^* + P^{(\lambda)}(J-J^*)\\
&= J^* + P^{(\lambda)}(J-J^*)\\
&\neq J^* + (J-J^*)\\
&= J,
\end{aligned}$$

where the inequality follows from Lemma 2. □

We next set out to characterize the "steady-state" expectations of $A(X_t)$ and $b(X_t)$. While this can be done by taking limits of expectations as t goes to infinity, it is simpler to characterize expectations of a process that is already in steady state. We therefore make a short digression to construct a stationary version of X_t.

We proceed as follows. Let $\{i_t \mid -\infty < t < \infty\}$ be a Markov chain that evolves according to the transition probability matrix P and is in steady state, in the sense that $\Pr(i_t = i) = \pi(i)$ for all i and all t. Given any sample path of this Markov chain, we define

$$z_t = \sum_{\tau=-\infty}^{t}\lambda^{t-\tau}\phi(i_\tau). \tag{6}$$

Note that z_t is constructed by taking the stationary process $\phi(i_t)$, whose magnitude is bounded by a constant, and passing it through an exponentially stable linear time invariant filter. The output z_t of this filter is stationary and its magnitude is bounded by a constant (the same constant applies to all sample paths). With z_t so constructed, we let $X_t = (i_t, i_{t+1}, z_t)$ and note that this is a Markov process with the same transition probabilities as the process constructed in Section 3.2. Furthermore, the state space of this process, which we will denote by $\mathscr{S}$, is bounded. We can now identify $E_0[\cdot]$ with the expectation with respect to the invariant distribution of this process.

We now characterize the steady-state expectation of several expressions of interest. We omit the proof, because it follows the same steps as the proof of Lemma 7 in (Tsitsiklis & Van Roy, 1997).

Lemma 4. *Under Assumption* 1, *the following relations hold*:

(a) $E_0[z_t\phi'(i_t)] = \sum_{m=0}^{\infty}\lambda^m\Phi' DP^m\Phi$,

(b) $E_0[z_t\phi'(i_{t+1})] = \sum_{m=0}^{\infty}\lambda^m\Phi' DP^{m+1}\Phi$,

(c) $E_0[z_t] = \dfrac{1}{1-\lambda}\Phi' De.$

(d) $E_0[z_t g(i_t)] = \sum_{m=0}^{\infty}\lambda^m\Phi' DP^m g.$

The following lemma characterizes the steady-state expectations $E_0[A(X_t)]$ and $E_0[b(X_t)]$ of the terms in Eq. (5), which we will denote by A and b.

Lemma 5. *Under Assumption* 1, *the steady-state expectations $A = E_0[A(X_t)]$ and $b = E_0[b(X_t)]$ are given by*

$$A = \begin{bmatrix} -c & 0\cdots 0\\ -\frac{1}{1-\lambda}\Phi' De & \Phi' D(P^{(\lambda)}-I)\Phi\end{bmatrix},$$

and

$$b = \begin{bmatrix} c\mu^*\\ \Phi' D(1-\lambda)\sum_{m=0}^{\infty}\lambda^m\sum_{t=0}^{m}P^t g\end{bmatrix}.$$

Proof. Using Lemma 4, and the relation

$$\sum_{m=0}^{\infty}(\lambda P)^m = (1-\lambda)\sum_{m=0}^{\infty}\lambda^m\sum_{t=0}^{m}P^t,$$

we have

$$\begin{aligned}
E_0[z_t(\phi'(i_{t+1})-\phi(i_t))] &= \Phi' D\sum_{m=0}^{\infty}(\lambda P)^m(P-I)\Phi\\
&= \Phi' D\left((1-\lambda)\sum_{m=0}^{\infty}\lambda^m P^{m+1}-I\right)\Phi\\
&= \Phi' D(P^{(\lambda)}-I)\Phi.
\end{aligned}$$

Since A is given by

$$A = \begin{bmatrix} -c & 0\cdots 0\\ -E_0[z_t] & E_0[z_t(\phi'(i_{t+1})-\phi'(i_t))]\end{bmatrix},$$

this establishes the desired characterization of A. As for the case of b, using Lemma 4, we have

$$\begin{aligned}
E_0[z_t g(i_t)] &= \sum_{m=0}^{\infty}\lambda^m\Phi' DP^m g\\
&= \Phi' D(1-\lambda)\sum_{m=0}^{\infty}\lambda^m\sum_{t=0}^{m}P^t g.
\end{aligned}$$

Combining this with the fact that

$$b = \begin{bmatrix} cE_0[g(i_t)] \\ E_0[z_t g(i_t)] \end{bmatrix},$$

completes the proof. □

The following lemma establishes that the expectations of $A(X_t)$ and $b(X_t)$ converge to their steady-state values at a geometric rate. The proof makes use of the geometric convergence of finite state Markov chains to steady state. It is similar to the proof of a corresponding lemma in Tsitsiklis and Van Roy (1997), and is omitted.

Lemma 6. *Under Assumption* 1, *there exist scalars C and* $\rho \in (0,1)$ *such that for any* $X_0 \in \mathscr{S}$ *and* $t \geq 0$, *we have*

$$\|E[A(X_t)|X_0] - A\| \leq C\rho^t,$$

and

$$\|E[b(X_t)|X_0] - b\| \leq C\rho^t.$$

We say that a square matrix M is negative definite if $x'Mx < 0$ for every $x \neq 0$, even if M is not symmetric. The matrix A is not necessarily negative definite, but becomes negative definite under an appropriate coordinate scaling.

Lemma 7. *Under Assumptions* 1 *and* 2, *there exists a diagonal matrix L with positive diagonal entries*, *such that the matrix LA is negative definite.*

Proof. Let J be a nonconstant function on the state space. Since the Markov chain $\{i_t\}$ is irreducible, $J(i_t)$ is not a constant function of time, which implies that

$$\begin{aligned} 0 &< \tfrac{1}{2}E_0[(J(i_{t+1}) - J(i_t))^2] \\ &= E_0[J(i_t)^2] - E_0[J(i_{t+1})J(i_t)] = J'DJ - J'DPJ \\ &= J'D(I-P)J. \end{aligned}$$

For any $r \neq 0$, $J = \Phi r$ is a nonconstant vector, because of Assumption 2. Thus, $r'\Phi'D(P-I)J\Phi r > 0$ for every $r \neq 0$, which shows that the matrix $\Phi'D(P-I)\Phi$ is negative definite. The same argument works for the matrix $\Phi'D(P^{(\lambda)} - I)\Phi$, because $P^{(\lambda)}$ is also an irreducible and aperiodic stochastic matrix with the same invariant distribution.

Let L be a diagonal matrix with the first diagonal entry equal to some scalar $\ell > 0$ and every other diagonal entry equal to one. Using the special form of the matrix A (see Lemma 5) and the just established negative definiteness of the lower diagonal block of A, it is a matter of simple algebra to verify that LA becomes negative definite, when ℓ is chosen sufficiently large. □

3.3. A result on stochastic approximation

To establish convergence of TD(λ) based on the steady-state dynamics, we rely on results from stochastic approximation theory. The following Theorem (Proposition 4.8 from p. 174 of Bertsekas & Tsitsiklis (1996)) is a special case of a very general result (Theorem 17 on p. 239 of Benveniste, Metivier & Priouret (1990)), and it provides the basis for a corollary that will suit our needs.

Theorem 2. *Consider an iterative algorithm of the form*

$$\theta_{t+1} = \theta_t + \gamma_t(A(X_t)\theta_t + b(X_t)),$$

where:

(a) *The step sizes* γ_t *are positive*, *deterministic*, *and satisfy* $\sum_{t=0}^{\infty}\gamma_t = \infty$ *and* $\sum_{t=0}^{\infty}\gamma_t^2 < \infty$.
(b) *The Markov process* X_t, *which evolves in a state space* $\mathscr{S}$, *has an invariant* (*steady state*) *distribution. Let* $E_0[\cdot]$ *stand for the expectation with respect to this invariant distribution.*
(c) *The matrix A defined by* $A = E_0[A(X_t)]$ *is negative definite.*
(d) *There exists a constant C such that* $\|A(X)\| \leq C$ *and* $\|b(X)\| \leq C$, *for all* $X \in \mathscr{S}$.
(e) *There exist scalars C and* $\rho \in (0,1)$ *such that*

$$\|E[A(X_t)|X_0 = X] - A\| \leq C\rho^t, \quad \forall t \geq 0,\ X \in \mathscr{S},$$

and

$$\|E[b(X_t)|X_0 = X] - b\| \leq C\rho^t, \quad \forall t \geq 0,\ X \in \mathscr{S},$$

where $b = E_0[b(X_t)]$.

Then, θ_t *converges to* θ^*, *with probability* 1, *where* θ^* *is the unique vector that satisfies* $A\theta^* + b = 0$.

Consider the change of coordinates $\tilde{\theta}_t = L^{1/2}\theta_t$. If we rewrite the algorithm in terms of $\tilde{\theta}$, the matrix A gets replaced by $L^{1/2}AL^{-1/2}$. If LA is negative definite, so is $L^{1/2}AL^{-1/2}$, and Theorem 2 implies the following.

Corollary 1. *The conclusions of Theorem* 2 *remain valid if Condition* (c) *is replaced by the following condition*:

(c′) *Let the matrix A be defined by* $A = E_0[A(X_t)]$. *There exists a diagonal matrix L with positive diagonal entries such that LA is negative definite.*

3.4. Proof of Theorem 1

The various lemmas given in Section 3.2 establish that the conditions of Corollary 1 are satisfied by the TD(λ) algorithm. Hence the algorithm converges (with probability 1) to a limit θ^* that satisfies

$$A\theta^* + b = 0.$$

Invoking Lemma 5, we recall that $b(1) = c\mu^*$, and observe that $(A\theta^*)(1) = -c\theta^*(1)$. We therefore have $\theta^*(1) = \mu^*$, i.e., the sequence μ_t converges to μ^*. Let the vector $r^* \in \Re^n$ be given by $r^* = (\theta^*(2), \ldots, \theta^*(n+1))'$. Then, using Lemmas 3 and 5, the relation $1/(1-\lambda) = (1-\lambda)\sum_{m=0}^{\infty}\lambda^m(m+1)$, and the equation $A\theta^* + b = 0$, we obtain

$$-\Phi' D(1-\lambda)\sum_{m=0}^{\infty}\lambda^m \sum_{t=0}^{m} P^t g = \Phi' D(P^{(\lambda)} - I)\Phi r^* - \frac{\mu^*}{1-\lambda}\Phi' De,$$

$$-\Phi' D(1-\lambda)\sum_{m=0}^{\infty}\lambda^m \sum_{t=0}^{m} P^t g = \Phi' D(P^{(\lambda)} - I)\Phi r^*$$

$$-\Phi' D(1-\lambda)\sum_{m=0}^{\infty}\lambda^m \sum_{t=0}^{m} \mu^* e,$$

$$\Phi' D\Phi r^* = \Phi' D\left(P^{(\lambda)}\Phi r^* + (1-\lambda)\sum_{m=0}^{\infty}\lambda^m \sum_{t=0}^{m} P^t(g - \mu^* e)\right),$$

$$\Phi' D\Phi r^* = \Phi' D T^{(\lambda)}(\Phi r^*),$$

$$\Phi(\Phi' D\Phi)^{-1}\Phi' D\Phi r^* = \Phi(\Phi' D\Phi)^{-1}\Phi' D T^{(\lambda)}(\Phi r^*),$$

$$\Phi r^* = \Pi T^{(\lambda)}(\Phi r^*).$$

This completes the proof. □

4. Approximation error

In this section, we propose a definition of approximation error, study a few of its properties, and derive error bounds.

4.1. A definition of error

In our analysis of discounted cost TD(λ) (Tsitsiklis & Van Roy, 1997), we employed the error metric $\|\Phi r^* - J^*\|_D$, where J^* was the cost-to-go function for a discounted Markov chain. This formulation enabled the development of a graceful error bound. In the context of average cost problems, one is usually content with an approximation of *any* differential cost function J, not necessarily the basic one J^*. And it is possible that there exists a parameter vector $\bar{r}$ such that $\|\Phi\bar{r} - J\|_D$ is very small for some differential cost function J, while $\|\Phi r - J^*\|_D$ is large for all r. For this reason, we will define the approximation error as the infimum of the weighted Euclidean distance from the set of all differential cost functions:

$$\inf_{J \in \{J^* + ce \mid c \in \Re\}} \|\Phi r^* - J\|_D = \inf_{c \in \Re} \|\Phi r^* - (J^* + ce)\|_D.$$

In addition to catering intuitive appeal, this definition will lead to a graceful error bound.

We now derive an alternative characterization of the error metric above. Any vector $J \in \Re^n$ can be decomposed into a component $\mathscr{P}J$ that is D-orthogonal to e, and a component $(I - \mathscr{P})J$ that is a multiple of e, where $\mathscr{P}$ is the projection matrix defined by

$$\mathscr{P} = I - ee'D.$$

It is easily checked that

$$\mathscr{P} = I - e\pi' = I - \lim_{t\to\infty} P^t.$$

This implies that P and $\mathscr{P}$ commute (i.e., $P\mathscr{P} = \mathscr{P}P$). By definition of J^*, we have

$$e'DJ^* = \pi' J^* = 0.$$

It follows that $\mathscr{P}J^* = J^*$. Since the minimum distance of the vector $\Phi r^* - J^*$ from the subspace $\{ce \mid c \in \Re\}$ is equal to the magnitude of the projection onto the orthogonal complement of the subspace, we have

$$\inf_{c\in\Re} \|\Phi r^* - (J^* + ce)\|_D = \|\mathscr{P}\Phi r^* - J^*\|_D.$$

4.2. A decomposition of basis functions

The projection introduced in the previous subsection can be applied to each basis function ϕ_k to obtain the function $\mathscr{P}\phi_k$, which is D-orthogonal to e. In this subsection, we show that replacing each ϕ_k by $\mathscr{P}\phi_k$ does not change the limit to which TD(λ) converges or the resulting approximation error.

Recall that TD(λ) converges to the unique solution r^* of the equation $\Pi T^{(\lambda)}(\Phi r^*) = \Phi r^*$. Let

$$\bar{\Phi} = \mathscr{P}\Phi,$$

and note that $\bar{\Phi}$ replaces Φ, if each basis functions ϕ_k is replaced by $\mathscr{P}\phi_k$. If $r \neq 0$ and $\mathscr{P}\Phi r = 0$, then Φr must be a multiple of e, which is impossible by Assumption 2. Thus, $\bar{\Phi}$ also satisfies Assumption 2. When the basis functions $\mathscr{P}\phi_1, \ldots, \mathscr{P}\phi_K$ are employed, TD(λ) converges to a vector $\bar{r}$ that satisfies

$$\bar{\Pi} T^{(\lambda)}(\bar{\Phi}\bar{r}) = \bar{\Phi}\bar{r},$$

where the matrix $\bar{\Pi}$ is defined by

$$\bar{\Pi} = \bar{\Phi}(\bar{\Phi}' D\bar{\Phi})^{-1}\bar{\Phi}' D.$$

We will now show that $r^* = \bar{r}$.

Using the definition of $T^{(\lambda)}$ and the property $e'DP = \pi' P = \pi'$, it is easily verified that for any r,

$$e'D(T^{(\lambda)}(\Phi r) - \Phi r) = 0.$$

By the fixed point equation $\Pi T^{(\lambda)}(\Phi r^*) = \Phi r^*$, we also have

$$\phi_k' D(T^{(\lambda)}(\Phi r^*) - \Phi r^*) = 0,$$

for each basis function ϕ_k. It follows that for any projected basis function $\bar{\phi}_k = \mathscr{P}\phi_k$, there is a scalar c such that

$$\begin{aligned}\bar{\phi}_k' D(T^{(\lambda)}(\Phi r^*) - \Phi r^*) &= (\phi_k + ce)' D(T^{(\lambda)}(\Phi r^*) - \Phi r^*)\\ &= 0.\end{aligned}$$

The fact that

$$T^{(\lambda)}(\bar{\Phi} r^*) = T^{(\lambda)}(\Phi r^* + \hat{c}e) = T^{(\lambda)}(\Phi r^*) + \hat{c}e,$$

for some constant $\hat{c}$, then leads to the conclusion that

$$\bar{\phi}_k' D(T^{(\lambda)}(\bar{\Phi} r^*) - \bar{\Phi} r^*) = \bar{\phi}_k' D(T^{(\lambda)}(\Phi r^*) - \Phi r^*) = 0.$$

Hence, $\bar{\Pi} T^{(\lambda)}(\bar{\Phi} r^*) = \bar{\Phi} r^*$ and $r^* = \bar{r}$.

4.3. Mixing factor

In the next subsection, we will provide a bound on the error associated with the limiting weight vector r^*. Central to the development of this bound will be a "mixing factor", that reflects the speed with which steady state is reached.

Let J be some function defined on the state space. Mixing can be viewed as an assumption that $E[J(i_t) \mid i_0]$ converges to $E_0[J(i_t)]$ at the rate of α^t, where α is a "mixing factor". Since our definition of the approximation error factors out constant offsets, and since $(I - \mathscr{P})J$ is aligned with e, we can focus on $E[(\mathscr{P}J)(i_t) \mid i_0]$. Thus, one possible assumption could be that $E[(\mathscr{P}J)(i_t) \mid i_0]$ decreases like α^t, for all functions J. In terms of the transition probability matrix P, this would be captured by an assumption that $\|\mathscr{P}P\|_D \leq \alpha$.

For the purposes of our error bounds, we do not need every possible function J to converge rapidly to steady state. Rather, it suffices to consider only those functions that are representable by our approximation architecture, i.e., linear combinations of the basis functions ϕ_k. We can capture this effect by projecting, using the projection matrix $\bar{\Pi}$, and placing an assumption on the induced norm $\|\bar{\Pi}\mathscr{P}P\|_D$, which is actually the same as $\|\bar{\Pi}P\|_D$ since $\bar{\Pi}\mathscr{P} = \bar{\Pi}$ (this follows from the fact that $\bar{\Pi}$ projects onto a subspace of the range onto which $\mathscr{P}$ projects).

Finally, it turns out that an even weaker assumption will do, using the following idea. Given any $\delta \in (0,1)$, we define an auxiliary Markov chain with a transition matrix $P_\delta = I + \delta(P - I)$ and a cost function $g_\delta = \delta g$. The basic differential cost function for this Markov chain remains unchanged. This is because

$$\begin{aligned}&\delta g - \delta\mu^* e + (I + \delta(P - I))J^*\\ &\quad = \delta(g - \mu^* e + PJ^*) + (1 - \delta)J^* = J^*.\end{aligned}$$

Similarly, it is easy to show that TD(0) generates the same limit of convergence for this auxiliary Markov chain as it did for the original one. In this spirit, we can consider $\|\bar{\Pi}P_\delta\|_D$ as the relevant mixing factor. Furthermore, since there is freedom in choosing δ, we can obtain the tightest possible bound by minimizing over all possible choices of δ.

For the more general case of $\lambda \in [0,1)$, the pertinent mixing time is that of the stochastic matrix $P^{(\lambda)} = (1 - \lambda)\sum_{m=0}^{\infty} \lambda^m P^{m+1}$. (Note that $P^{(0)} = P$, which brings us back to our previous discussion concerning the case of $\lambda = 0$.) Similar to the context of TD(0), we define $P_\delta^{(\lambda)} = I + \delta(P^{(\lambda)} - I)$, and we define a scalar α_λ for each $\lambda \in [0,1)$ by

$$\alpha_\lambda = \inf_{\delta > 0} \|\bar{\Pi} P_\delta^{(\lambda)}\|_D.$$

(Note that here we also allow $\delta \geq 1$, even though the motivation in the preceding paragraph does not apply.) This mixing factor will be used to establish our error bound.

4.4. The error bound

We now state a theorem that provides a bound on approximation error. A proof is provided in the next subsection.

Theorem 3. *Let Assumptions* 1 *and* 2 *hold. For each* $\lambda \in [0,1)$, *let* $r_\lambda^* \in \Re^K$ *be the vector satisfying*

$$\Phi r_\lambda^* = \Pi T^{(\lambda)}(\Phi r_\lambda^*).$$

Then:

(a) *For each* $\lambda \in [0,1)$, *the mixing factor* α_λ *is in* $[0,1)$ *and* $\lim_{\lambda \uparrow 1} \alpha_\lambda = 0$.

(b) *The following bound holds*:

$$\|\mathscr{P}\Phi r_\lambda^* - J^*\|_D \leq \frac{1}{\sqrt{1 - \alpha_\lambda^2}} \inf_{r \in \Re^K} \|\mathscr{P}\Phi r - J^*\|_D.$$

Note that the bound is a multiple of

$$\inf_{r \in \Re^K} \|\mathscr{P}\Phi r - J^*\|_D,$$

which is the minimal error possible given the fixed set of basis functions. This term becomes zero if there exists a parameter vector r and a scalar c for which $\Phi r = J^* + ce$, that is, if our "approximation architecture" is capable of representing exactly some differential cost function.

The term $1/\sqrt{1 - \alpha_\lambda^2}$ decreases as α_λ decreases. Hence, the term is guaranteed to approach its optimal value of 1 as λ approaches 1. This suggests that larger values of λ may lead to lower approximation error.

4.5. Proof of Theorem 3

We begin by establishing part (a) of the theorem. Since α_λ is the infimum of a set of nonnegative reals, $\alpha_\lambda \geq 0$. From Lemma 2, we have $\|P^{(\lambda)}\|_D \leq 1$ and $P^{(\lambda)}J \neq J$ if J is not proportional to e. It follows that for any $\delta \in (0,1)$ and any J that is not proportional to e, we have

$$\|\mathscr{P}P_\delta^{(\lambda)}J\|_D \leq \|P_\delta^{(\lambda)}J\|_D = \|\delta P^{(\lambda)}J + (1-\delta)J\|_D < \|J\|_D.$$

(The first inequality uses the nonexpansive property of projections. The last one holds because J and $P^{(\lambda)}J$ are distinct elements of the ball $\{\bar{J} \mid \|\bar{J}\|_D \leq \|J\|_D\}$, so their strictly convex combination must lie in the interior.) Note that $\|\mathscr{P}P_\delta^{(\lambda)}J\|_D$ is a continuous function of J and that the set $\{J \mid \|J\|_D \leq 1\}$ is compact. It follows from Weierstrass' theorem that for any $\delta \in (0,1)$, $\|\mathscr{P}P_\delta^{(\lambda)}\|_D < 1$. Since $\bar{\Pi} = \bar{\Pi}\mathscr{P}$, we then have

$$\begin{aligned}\alpha_\lambda &= \inf_{\delta > 0} \|\bar{\Pi}P_\delta^{(\lambda)}\|_D \\ &\leq \inf_{\delta > 0} \|\mathscr{P}P_\delta^{(\lambda)}\|_D \leq \inf_{\delta \in (0,1)} \|\mathscr{P}P_\delta^{(\lambda)}\|_D < 1.\end{aligned}$$

As for the limit as λ approaches 1, we have

$$\begin{aligned}\lim_{\lambda \uparrow 1} \alpha_\lambda &= \lim_{\lambda \uparrow 1} \inf_{\delta > 0} \|\bar{\Pi}P_\delta^{(\lambda)}\|_D \\ &\leq \lim_{\lambda \uparrow 1} \|\bar{\Pi}P^{(\lambda)}\|_D \leq \lim_{\lambda \uparrow 1} \|\mathscr{P}P^{(\lambda)}\|_D.\end{aligned}$$

Assumption 1 implies that

$$\lim_{t \to \infty} \|\mathscr{P}P^t\|_D = 0.$$

It follows that

$$\lim_{\lambda \uparrow 1} \|\mathscr{P}P^{(\lambda)}\|_D = \lim_{\lambda \uparrow 1} \left\|(1-\lambda)\sum_{t=0}^{\infty} \lambda^t \mathscr{P}P^{t+1}\right\|_D = 0.$$

This completes the proof for part (a). □

Let $T_\delta^{(\lambda)} = (1-\delta)I + \delta T^{(\lambda)}$. It is easy to see that $T_\delta^{(\lambda)}J^* = J^*$ and $\bar{\Pi}T_\delta^{(\lambda)}(\bar{\Phi}r_\lambda^*) = \bar{\Phi}r_\lambda^*$. For any nonnegative scalar δ, we have

$$\begin{aligned}\|\mathscr{P}\Phi r_\lambda^* - J^*\|_D^2 &= \|\bar{\Phi}r_\lambda^* - J^*\|_D^2 \\ &= \|\bar{\Pi}T_\delta^{(\lambda)}(\bar{\Phi}r_\lambda^*) - T_\delta^{(\lambda)}J^*\|_D^2 \\ &\leq \|\bar{\Pi}T_\delta^{(\lambda)}(\bar{\Phi}r_\lambda^*) - \bar{\Pi}T_\delta^{(\lambda)}J^*\|_D^2 \\ &\quad + \|T_\delta^{(\lambda)}J^* - \bar{\Pi}T_\delta^{(\lambda)}J^*\|_D^2 \\ &= \|\bar{\Pi}P_\delta^{(\lambda)}(\bar{\Phi}r_\lambda^*) - \bar{\Pi}P_\delta^{(\lambda)}J^*\|_D^2 \\ &\quad + \|J^* - \bar{\Pi}J^*\|_D^2 \\ &\leq \|\bar{\Pi}P_\delta^{(\lambda)}\|_D^2 \|\bar{\Phi}r_\lambda^* - J^*\|_D^2 + \|J^* - \bar{\Pi}J^*\|_D^2.\end{aligned}$$

Since δ is an arbitrary nonnegative scalar, we have

$$\|\mathscr{P}\Phi r_\lambda^* - J^*\|_D^2 \leq \alpha_\lambda^2 \|\bar{\Phi}r_\lambda^* - J^*\|_D^2 + \|J^* - \bar{\Pi}J^*\|_D^2,$$

and it follows that

$$\|\mathscr{P}\Phi r_\lambda^* - J^*\|_D \leq \frac{1}{\sqrt{1-\alpha_\lambda^2}}\|J^* - \bar{\Pi}J^*\|_D.$$

Since

$$\|J^* - \bar{\Pi}J^*\|_D = \inf_r \|\mathscr{P}\Phi r - J^*\|_D,$$

this completes the proof for part (b). □

5. Using a fixed average cost estimate

In this section, we introduce and study a variant that employs a fixed estimate μ of the average cost, in place of μ_t. In particular, the parameter vector r_t is updated according to the same rule (Eq. (3)), but the definition of the temporal difference (2) is changed to

$$d_t = (g(i_t) - \mu) + \phi'(i_{t+1})r_t - \phi'(i_t)r_t.$$

Our analysis involves an additional mixing factor defined by

$$\beta_\lambda = \inf_{\delta \in [0,1]} \|\Pi P_\delta^{(\lambda)}\|_D,$$

which is similar to α_λ but involves a projection onto the range of Φ instead of $\bar{\Phi}$.

Theorem 4. *Under Assumptions* 1–3, *for any* $\lambda \in [0,1)$, *the following hold*:

(a) *The TD(λ) algorithm with a fixed average cost estimate, as defined above, converges with probability* 1.

(b) *The limit of convergence* $\bar{r}_\lambda$ *is the unique solution of the equation*

$$\Pi T^{(\lambda)}(\Phi\bar{r}_\lambda) + \frac{\mu^* - \mu}{1-\lambda}\Pi e = \Phi\bar{r}_\lambda.$$

(c) *For any* $\lambda \in [0,1)$, *the mixing factor* β_λ *is in* $[0,1)$, *and* $\lim_{\lambda \uparrow 1} \beta_\lambda = 0$.

(d) *The limit of convergence* $\bar{r}_\lambda$ *satisfies*

$$\begin{aligned}\|\mathscr{P}\Phi\bar{r}_\lambda - J^*\|_D &\leq \frac{1}{\sqrt{1-\alpha_\lambda^2}} \inf_{r \in \Re^K} \|\mathscr{P}\Phi r - J^*\|_D \\ &\quad + \frac{|\mu^* - \mu|}{(1-\beta_\lambda)(1-\lambda)}\|\Pi e\|_D,\end{aligned}$$

where α_λ *and* $\mathscr{P}$ *are defined as in Section* 5.

There are two somewhat unrelated terms involved in the bound of Theorem 4. The first term is equal to the error bound of Theorem 3, and can be viewed as error brought about by the choice of basis functions. The second term is proportional to the error in the average cost estimate. The term is also proportional to $\|\Pi e\|_D$, which is zero if the space spanned by the basis functions is D-orthogonal to e. The dependence on λ and β_λ is a little more complicated. If either λ or β_λ approaches one, the coefficient approaches infinity. In contrast to the discussion in the preceding section, we now have a situation where values of λ close to 1 cease to be preferable.

Proof of Theorem 4. We omit the proof of parts (a)–(c) because it is very similar to the proof of Theorems 1 and 3(a). As in the previous sections, we let r^*_λ denote the unique vector satisfying $\Phi r^*_\lambda = \Pi T^{(\lambda)}(\Phi r^*_\lambda)$. For any $\delta \in [0,1]$, we have

$$\begin{aligned}
&\|\Phi \bar{r}_\lambda - \Phi r^*_\lambda\|_D \\
&\quad = \|(1-\delta)\Phi \bar{r}_\lambda + \delta \Phi \bar{r}_\lambda - (1-\delta)\Phi r^*_\lambda - \delta \Phi r^*_\lambda\|_D \\
&\quad = \left\|(1-\delta)\Pi\Phi \bar{r}_\lambda + \delta\left(\Pi T^{(\lambda)}(\Phi \bar{r}_\lambda) + \frac{\mu^* - \mu}{1-\lambda}\Pi e\right)\right. \\
&\qquad \times \| - (1-\delta)\Pi\Phi r^*_\lambda - \delta \Pi T^{(\lambda)}(\Phi r^*_\lambda)\|_D \\
&\quad = \left\|\Pi T^{(\lambda)}_\delta(\Phi \bar{r}_\lambda) + \frac{\delta(\mu^* - \mu)}{1-\lambda}\Pi e - \Pi T^{(\lambda)}_\delta(\Phi r^*_\lambda)\right\|_D \\
&\quad \le \|\Pi T^{(\lambda)}_\delta(\Phi \bar{r}_\lambda) - \Pi T^{(\lambda)}_\delta(\Phi r^*_\lambda)\|_D + \frac{|\mu^* - \mu|}{1-\lambda}\|\Pi e\|_D \\
&\quad = \|\Pi P^{(\lambda)}_\delta(\Phi \bar{r}_\lambda) - \Pi P^{(\lambda)}_\delta(\Phi r^*_\lambda)\|_D + \frac{|\mu^* - \mu|}{1-\lambda}\|\Pi e\|_D \\
&\quad \le \|\Pi P^{(\lambda)}_\delta\|_D \|\Phi \bar{r}_\lambda - \Phi r^*_\lambda\|_D + \frac{|\mu^* - \mu|}{1-\lambda}\|\Pi e\|_D.
\end{aligned}$$

Since δ is an arbitrary scalar in [0,1], we have

$$\|\Phi \bar{r}_\lambda - \Phi r^*_\lambda\|_D \le \beta_\lambda \|\Phi \bar{r}_\lambda - \Phi r^*_\lambda\|_D + \frac{|\mu^* - \mu|}{1-\lambda}\|\Pi e\|_D,$$

and it follows that

$$\|\Phi \bar{r}_\lambda - \Phi r^*_\lambda\|_D \le \frac{|\mu^* - \mu|}{(1-\beta_\lambda)(1-\lambda)}\|\Pi e\|_D.$$

The desired bound then follows from Theorem 3 and the triangle inequality. □

6. Conclusion

We have proposed a variant of temporal-difference learning that is suitable for approximating differential cost functions, and we have established the convergence of this algorithm when applied to finite state irreducible aperiodic Markov chains. In addition, we have provided bounds on the distance of the limiting function Φr^*_λ from the space of differential cost functions. These bounds involve the expression $\inf_r \|\mathscr{P}\Phi r - J^*\|_D$, which is natural because no approximation could have error smaller than this expression (when the error is measured in terms of $\|\mathscr{P}(\cdot)\|_D$).

It is interesting to note that even if a given Markov chain takes a long time to reach steady state, the mixing factor α_λ may be small due to the choice of basis functions. In particular, the expected future value $E[\phi_k(i_t) \mid i_0]$ of a basis function may converge rapidly even though $E[J(i_t) \mid i_0]$ converges slowly for some other function J. This may partially explain why small values of λ seem to lead to good approximations even with Markov chains that converge to steady state rather slowly.

On the technical side, we mention a few straightforward extensions to our results.

1. With some additional technical assumptions, the proof of Theorem 1 can be extended to the case of infinite state Markov chains where approximations are generated using unbounded basis functions. This extension has been omitted for the sake of brevity, but largely involves arguments of the same type as in (Tsitsiklis & Van Roy, 1997).
2. The linear independence of the basis functions ϕ_k is not essential. In the linearly dependent case, some components of z_t and r_t become linear combinations of the other components and can be simply eliminated, which takes us back to the linearly independent case.
3. Finally, if Assumption 2(b) is removed, then our line of analysis can be used to show that $\mathscr{P}\Phi r_t$ still converges, but $(I - \mathscr{P}\Phi r_t)$ is aligned to e and need not converge.

Acknowledgements

We thank Peter Dayan for originally suggesting that ideas from Tsitsiklis and Van Roy (1997) might be applicable to the average cost setting.

References

Abounadi, J. (1998). Stochastic approximation for non-expansive maps: Application to Q-learning algorithms. Ph.D. Thesis, Department of Electrical Engineering, Massachusetts Institute of Technology, Cambridge, MA.

Bertsekas, D. P., & Tsitsiklis, J. N. (1996). *Neuro-dynamic programming*. Belmont, MA: Athena Scientific.

Benveniste, A., Metivier, M., & Priouret, P. (1990). *Adaptive algorithms and stochastic approximations*. Berlin: Springer.

Dayan, P. D. (1992). The convergence of TD(λ) for general λ. *Machine Learning*, *8*, 341–362.

Gallager, R. G. (1996). *Discrete stochastic processes*: Boston, MA.

Gurvits, L., Lin, L. J., & Hanson, S. J. (1994). Incremental learning of evaluation functions for absorbing Markov chains: New methods and theorems, preprint.

Mahadevan, S. (1996). Average reward reinforcement learning: Foundations, algorithms, and empirical results. *Machine Learning*, *22*, 1–38.

Marbach, P., Mihatsch, O., & Tsitsiklis, J. N. (1998). Call admission control and routing in integrated service networks using reinforcement learning. In *Proceedings of the 1998 IEEE CDC*, Tampa, FL.

Pineda, F. (1996). Mean-field analysis for batched TD(λ). Unpublished.

Sutton, R. S. (1988). Learning to predict by the method of temporal differences. *Machine Learning*, *3*, 9–44.

Tsitsiklis, J. N., & Van Roy, B. (1997). An Analysis of Temporal-Difference Learning with Function Approximation. *IEEE Transactions on Automatic Control*, *42*(5), 674–690.

John N. Tsitsiklis was born in Thessaloniki, Greece, in 1958. He received the B.S. degree in mathematics (1980), and the B.S. (1980), M.S. (1981) and Ph.D. (1984) degrees in electrical engineering, all from the Massachusetts Insitute of Technology, Cambridge, Massachusetts.

During the academic year 1983–1984, he was an acting assistant professor of Electrical Engineering at Stanford University, Stanford, California. Since 1984, he has been with the Massachusetts Institute of Technology, where he is currently Professor of Electrical Engineering and Computer Science. He has sered as acting co-director of the MIT Laboratory for Information and Decision Systems (Spring 1996 and 1997). He has also been a visitor with the Dept. of EECS at the University of California at Berkeley, and the Institute for Computer Science in Iraklion, Greece. His research interests are in the fields of systems, optimization, control, and operations research. He has coauthored about 80 journal papers in these areas.

He is a coauthor of "Parallel and Distributed Computation: Numerical Methods" (with D. Bertsekas, 1989), "Neuro-Dynamic Programming" (with Dimitri Bertsekas, 1996), and "Introduction to Linear Optimization (with Dimitris Bertsimas, 1997). He has been a recipient of an IBM Faculty Development Award by the IEEE Control Systems Society, the M.I.T. Edgerton Faculty Achievement Award (1989), the Bodossakis Foundation Prize (1995), the INFORMS/CSTS prize (1997), and is a Fellow of the IEEE (1999). He was a plenary speaker at the 1992 IEEE Conference on Decision and Control. He is an associate editor of *Applied Mathematics Letters* and has been an associate editor of the *IEEE Transactions on Automatic Control and Automatica*.

Benjamin Van Roy received the SB degree in computer science and engineering and the SM and Ph.D. degrees in electrical engineering and computer science, all from the Massachusetts Institute of Technology. During his time at MIT, he received a Digital Equipment Corporation Scholarship, the George C. Newton (undergraduate laboratory project) Award, the Morris J. Levin Memorial (Master's thesis) Award, and the George M. Sprowls (Ph.D. thesis) Award.

Van Roy is currently Assistant Professor and Terman Fellow in the Department of Engineering-Economic Systems and Operations Research at Stanford University, with courtesy appointments in the Departments of Electrical Engineering and Computer Science. His research interests include the control of complex systems, computational learning, and financial economics.

From 1993 to 1997, Van Roy worked with Unica Technologies. At Unica, he participated in research and development projects, as well as the company's consulting servies. He also co-authored a book, Solving Pattern Recognition Problems (republished by Prentice-Hall as Solving Data Mining Problems Through Pattern Recognition), with several members of Unica's technical staff. During the Summer of 1997, he worked with Equity Derivatives Group at Morgan Stanley.