July 31, 2002 **To appear in J. of DEDS**

LEAST SQUARES POLICY EVALUATION ALGORITHMS WITH LINEAR FUNCTION APPROXIMATION[1]

by

A. Nedić and D. P. Bertsekas[2]

Abstract

We consider policy evaluation algorithms within the context of infinite-horizon dynamic programming problems with discounted cost. We focus on discrete-time dynamic systems with a large number of states, and we discuss two methods, which use simulation, temporal differences, and linear cost function approximation. The first method is a new gradient-like algorithm involving least-squares subproblems and a diminishing stepsize, which is based on the λ-policy iteration method of Bertsekas and Ioffe. The second method is the LSTD(λ) algorithm recently proposed by Boyan, which for $\lambda = 0$ coincides with the linear least-squares temporal-difference algorithm of Bradtke and Barto. At present, there is only a convergence result by Bradtke and Barto for the LSTD(0) algorithm. Here, we strengthen this result by showing the convergence of LSTD(λ), with probability 1, for every $\lambda \in [0, 1]$.

[1] Research supported by NSF under Grant ACI-9873339.

[2] Dept. of Electrical Engineering and Computer Science, M.I.T., Cambridge, MA 02139.

1. 1. INTRODUCTION

In this paper, we analyze methods for approximate evaluation of the cost-to-go function of a fixed stationary policy within the framework of infinite-horizon discounted dynamic programming. We consider a stationary discrete-time dynamic system with states denoted by $1, \ldots, n$. Since the policy is fixed, the system is a stationary Markov chain with transition probabilities denoted by p_{ij}, $i, j = 1, \ldots, n$. When a transition from state i to state j occurs at time t, an immediate cost $\alpha^t g(i,j)$ is incurred, where α is a discount factor with $0 < \alpha < 1$. We want to evaluate the long-term expected cost corresponding to each initial state i, given by

$$J(i) = E\left[\sum_{t=0}^{\infty} \alpha^t g(i_t, i_{t+1}) \;\Big|\; i_0 = i\right], \qquad \forall\, i = 1, \ldots, n,$$

where i_t denotes the state of the system at time t. This problem arises as a subproblem in the policy iteration method and its variations, such as optimistic policy iteration, λ-policy iteration, and optimistic λ-policy iteration (see Bertsekas and Tsitsiklis [BeT96] for an extensive discussion of these methods).

It is possible to calculate exactly the cost-to-go vector J, i.e., the vector whose components are $J(i)$, $i = 1, \ldots, n$, by solving a system of linear equations (see e.g., Bertsekas [Ber01] or Puterman [Put94]). However, in many practical problems where the transition probabilities p_{ij} are not known and/or the number of states is large, an approximation approach based on simulation and cost function approximation may be preferable. A popular method of this type is the TD(λ) algorithm, which has been analyzed in several papers, starting with Sutton [Sut88], and followed by Dayan and Sejnowski [DaS94], Jaakkola, Jordan, and Singh [JJS94], Gurvits, Lin, and Hanson [GLH94], Tsitsiklis and Van Roy [TsV97], and Tadić [Tad01]. An extensive account of the work on TD(λ) can be found in the book by Bertsekas and Tsitsiklis [BeT96].

We will discuss two different algorithms for calculating approximately the cost-to-go vector J. Both algorithms are based on simulation, and use temporal differences and least squares, in conjunction with a linear approximation architecture, whereby the costs $J(i)$ are approximated by a weighted sum of a set of features (the basis of the linear architecture). The first algorithm, called λ-least-squares policy evaluation (λ-LSPE for short), is related to the λ-policy iteration method proposed by Bertsekas and Ioffe [BeI96] (see also Bertsekas and Tsitsiklis [BeT96], Section 2.3.1, and Section 8.3), and may be viewed as simulation-based implementation of that method that uses function approximation. We prove that λ-LSPE is convergent, thus providing the first convergence result for a variant of the λ-policy iteration method that uses function approximation.

The second algorithm of this paper is the LSTD(λ) method proposed by Boyan [Boy02], a generalization of the linear least-squares temporal-difference algorithm due to Bradtke and Barto

[BrB96]. The LSTD(λ) method tries to solve directly the system of equations that characterizes the convergence limit of TD(λ), using all the information available from simulations. At present, the convergence of this method is known only for the case $\lambda = 0$ (Bradtke and Barto [BrB96]). Here, we extend this convergence result by proving convergence for any $\lambda \in [0, 1]$.

The two algorithms of the present paper share some important characteristics, such as the implicit or explicit use of least-squares subproblems, and Kalman filter-like recursive computations. Conceptually, the methods are related to the Gauss-Newton method similar to the relation of the TD(λ) algorithm and the incremental gradient-LMS method for least squares problems (see Bertsekas and Tsitsiklis [BeT96] or Bertsekas [Ber99] for discussions of the Gauss-Newton and incremental gradient methods). Computational experiments by Bertsekas and Ioffe [BeI96] (see also [BeT96]), and by Boyan [Boy02] suggest that the methods of the present paper can perform much better than the TD(λ) algorithm. It is difficult to compare the performance of the two methods of this paper to each other. The first method performs incremental changes of the weights, and consequently can take advantage of a good initial choice of weights, while the second method cannot, at least in the form presented here. Both methods converge to the same limit, which is also the limit to which TD(λ) converges.

The paper is organized as follows. In Section 2, we present the λ-LSPE algorithm, and we state a convergence result under the assumptions used by Tsitsiklis and Van Roy [TsV97] to prove convergence of TD(λ). These assumptions include the use of a diminishing stepsize. It appears, however, that the λ-LSPE method typically converges with a stepsize that is constant and equal to 1. This fact has not been rigorously established, except in the case where $\lambda = 1$, but it can be very important in practice, as it may greatly facilitate the stepsize selection and result in accelerated convergence. In Section 3, we similarly present the LSTD(λ) method and state the corresponding convergence result. In Sections 4 and 5, we prove the convergence results stated in Sections 2 and 3, respectively.

2. 2. THE λ-LSPE METHOD

In this section, we give a new method for approximate policy evaluation, which is motivated by the ideas of Bertsekas and Ioffe [BeI96]. The cost-to-go vector J is approximated by a linear function of the form

$$\tilde{J}(i, r) = \phi(i)'r, \qquad \forall\, i = 1, \ldots, n,$$

where $\phi(i)$ is a K-dimensional feature vector associated with the state i that has components

$\phi_1(i), \ldots, \phi_K(i)$, while r is a weight vector with components $r(1), \ldots, r(K)$. Note that, throughout the paper, we view every vector as a column vector, and that a prime denotes transpose.

We assume that an infinitely long trajectory is generated using a simulator of the system. The trajectory starts with an initial state i_0, which is chosen according to some initial probability distribution. The states $i_0, i_1, \ldots$ that comprise the trajectory are generated according to the transition probabilities p_{ij}. At time t, we have the current weight vector r_t, and as soon as we observe the transition from state i_t to i_{t+1}, we solve the least-squares problem

$$\min_r \sum_{m=0}^{t} \left(\phi(i_m)'r - \phi(i_m)'r_t - \sum_{k=m}^{t} (\alpha\lambda)^{k-m} d_t(i_k, i_{k+1}) \right)^2, \tag{2.1}$$

where

$$d_t(i_k, i_{k+1}) = g(i_k, i_{k+1}) + \big(\alpha\phi(i_{k+1}) - \phi(i_k)\big)'r_t, \qquad \forall\ k, t. \tag{2.2}$$

Let $\hat{r}_t$ be a solution of this problem, i.e.,

$$\hat{r}_t = \arg\min_r \sum_{m=0}^{t} \left(\phi(i_m)'r - \phi(i_m)'r_t - \sum_{k=m}^{t} (\alpha\lambda)^{k-m} d_t(i_k, i_{k+1}) \right)^2. \tag{2.3}$$

We then compute the new weight vector r_{t+1} according to

$$r_{t+1} = r_t + \gamma_t(\hat{r}_t - r_t), \tag{2.4}$$

where r_0 is an initial weight vector and γ_t is a positive deterministic stepsize. Both r_0 and γ_t are chosen independently of the trajectory $\{i_0, i_1, \ldots\}$.

2.1 Interpretations

To interpret the λ-LSPE method (2.3)-(2.4), let us introduce the variables

$$\chi_m(i) = \begin{cases} 1 & \text{if } i = i_m, \\ 0 & \text{otherwise,} \end{cases} \qquad \forall\ m, i.$$

Then, the least-squares solution $\hat{r}_t$ of Eq. (2.3) can be equivalently written as

$$\hat{r}_t = \arg\min_r \sum_{m=0}^{t} \sum_{i=1}^{n} \chi_m(i) \big(\phi(i)'r - \tilde{J}_{mt}(i)\big)^2,$$

where

$$\tilde{J}_{mt}(i) = \phi(i)'r_t + \sum_{k=m}^{t} (\alpha\lambda)^{k-m} d_t(i_k, i_{k+1}), \quad \text{with} \quad i_m = i.$$

Thus, the λ-LSPE method can be viewed as an incremental method for minimizing the cost function

$$\sum_{m=0}^{\infty} \sum_{i=1}^{n} \chi_m(i) \big(\phi(i)'r - \tilde{J}_{mt}(i)\big)^2, \tag{2.5}$$

whereby at each time t, we use the vector $\hat{r}_t$ that minimizes the partial sum

$$\sum_{m=0}^{t}\sum_{i=1}^{n}\chi_m(i)\big(\phi(i)'r-\tilde{J}_{mt}(i)\big)^2.$$

Assuming that the underlying Markov chain is ergodic, we have, with probability 1,

$$\lim_{t\to\infty}\frac{1}{t+1}\sum_{m=0}^{t}\chi_m(i)=\pi(i),\qquad \forall\ i=1,\dots,n,$$

where $\pi(i)$ denotes the steady-state probability of state i. By using this relation, we see that by minimizing the cost function

$$\frac{1}{t+1}\sum_{m=0}^{t}\sum_{i=1}^{n}\chi_m(i)\big(\phi(i)'r-\tilde{J}_{mt}(i)\big)^2,$$

the method attempts to minimize approximately

$$\sum_{i=1}^{n}\pi(i)\left(\phi(i)'r-\phi(i)'r_t-\sum_{k=0}^{\infty}(\alpha\lambda)^k E\big[d_t(i_k,i_{k+1})\mid i_0=i\big]\right)^2.$$

Thus, when $\gamma_t=1$ (in which case $r_{t+1}=\hat{r}_t$) the λ-LSPE method is an approximate version of the iteration

$$J_{t+1}(i)=J_t(i)+\sum_{k=0}^{\infty}(\alpha\lambda)^k E\big[d_t(i_k,i_{k+1})\mid i_0=i\big],\qquad i=1,\dots,n, \tag{2.6}$$

which is the λ-policy iteration method of Bertsekas and Ioffe [BeI96]. This latter method has the convergence property

$$\lim_{t\to\infty}J_t(i)=J(i),\qquad i=1,\dots,n,$$

for an arbitrary initial choice J_0 (see Bertsekas and Tsitsiklis [BeT96], Section 2.3.1).

There are two sources of approximation between the λ-LSPE method (2.3)-(2.4) (with $\gamma_t=1$ for all t) and the λ-policy iteration method (2.6). First, the cost function iterates $J_t(i)$ in Eq. (2.6) are approximated by $\phi(i)'r_t$, and consequently the exact formula (2.6) is approximated by the least squares minimization of Eq. (2.3). Second, the expression

$$\sum_{k=0}^{\infty}(\alpha\lambda)^k E\big[d_t(i_k,i_{k+1})\mid i_0=i\big]$$

[see Eq. (2.6)] is approximated using the finite collection of the finite sums

$$\sum_{k=m}^{t}(\alpha\lambda)^{k-m}d_t(i_k,i_{k+1}),\qquad \forall\ m \text{ such that } i_m=i,$$

obtained from the simulation [see Eq. (2.1)].

For another interpretation, note that when $\lambda = 1$, it can be seen that $\tilde{J}_{mt}(i)$ is the $(t-m)$-stage sample cost starting from state i, i.e.,

$$\tilde{J}_{mt}(i) = \sum_{k=m}^{t} \alpha^{k-m} g(i_k, i_{k+1}) + \alpha^{t+1-m}\phi(i_{t+1})' r_t, \quad \text{with} \quad i_m = i.$$

In this case, if in Eq. (2.4) we use $\gamma_t = 1$ for all t, the method reduces to

$$r_{t+1} = \arg\min_r \sum_{m=0}^{t}\sum_{i=1}^{n} \chi_m(i) \left(\phi(i)' r - \left(\sum_{k=m}^{t} \alpha^{k-m} g(i_k, i_{k+1}) + \alpha^{t+1-m}\phi(i_{t+1})' r_t \right)\right)^2,$$

which bears resemblance to the Kalman filtering algorithm for least-squares parameter identification.

The preceding interpretations suggest that the use of a stepsize that is constant and equal to 1 may work well in practice. Unfortunately, there is no proof of this, except in the case where $\lambda = 1$. As a practical matter, however, one may consider operating the λ-LSPE method by starting with $\gamma_0 = 1$, and by subsequently reducing γ_t very slowly. This type of implementation of the λ-LSPE algorithm appears to have potential for faster and more reliable practical convergence than the TD(λ) algorithm.

The λ-LSPE algorithm also bears some resemblance to the TD(λ) algorithm, which has the form

$$r_{t+1} = r_t + \gamma_t \left(\sum_{m=0}^{t} (\alpha\lambda)^{t-m} \phi(i_m) \right) d_t(i_t, i_{t+1})$$

(cf. Sutton [Sut88]). In particular, if the least-squares problem (2.3) in the least-squares λ-policy evaluation algorithm were to be solved approximately by using a *single* gradient iteration, then the algorithm would take the form

$$r_{t+1} = r_t + \gamma_t \left(\sum_{m=0}^{t} (\alpha\lambda)^{t-m} \phi(i_m) \right) d_t(i_t, i_{t+1}) + \gamma_t \sum_{k=0}^{t-1} \left(\sum_{m=0}^{k} (\alpha\lambda)^{k-m} \phi(i_m) \right) d_t(i_k, i_{k+1}),$$

which resembles the TD(λ) iteration, the difference being the additional last term in the right-hand side above.

2.2 An Efficient Implementation

We now discuss an efficient implementation of the λ-LSPE method, whereby the least-squares problem

$$\min_r \sum_{m=0}^{t} \left(\phi(i_m)' r - \phi(i_m)' r_t - \sum_{k=m}^{t} (\alpha\lambda)^{k-m} d_t(i_k, i_{k+1}) \right)^2$$

to be solved at iteration t is solved quickly by using data obtained from the solution of the preceding least-squares problems. Indeed, by setting the gradient of the quadratic cost function of the above minimization to zero, we see that $\hat{r}_t$ satisfies the equation

$$\left(\sum_{m=0}^{t} \phi(i_m)\phi(i_m)'\right)(\hat{r}_t - r_t) = \sum_{m=0}^{t} \phi(i_m) \sum_{k=m}^{t} (\alpha\lambda)^{k-m} d_t(i_k, i_{k+1}).$$

If the matrix $\sum_{m=0}^{t} \phi(i_m)\phi(i_m)'$ is invertible, we can uniquely determine $\hat{r}_t$. This matrix may not be invertible initially, in which case we can determine $\hat{r}_t$ by solving a perturbed least-squares problem. In particular, we can set

$$\hat{r}_t = \arg\min_r \left\{ \delta\|r - r_t\|^2 + \sum_{m=0}^{t} \left(\phi(i_m)'r - \phi(i_m)'r_t - \sum_{k=m}^{t} (\alpha\lambda)^{k-m} d_t(i_k, i_{k+1}) \right)^2 \right\},$$

where δ is a positive scalar and $\|\cdot\|$ is the standard Euclidean norm. In this case, we have

$$\hat{r}_t = r_t + \left(\delta I + \sum_{m=0}^{t} \phi(i_m)\phi(i_m)' \right)^{-1} \sum_{k=0}^{t} \left(\sum_{m=0}^{k} (\alpha\lambda)^{k-m} \phi(i_m) \right) d_t(i_k, i_{k+1}), \qquad \forall\ t.$$

Regardless of whether $\delta = 0$ or $\delta > 0$ in this formula, by using the definition of the temporal differences $d_t(i_k, i_{k+1})$ [cf. Eq. (2.2)], the update formula (2.4) for r_t can be compactly written as

$$r_{t+1} = r_t + \gamma_t B_t^{-1}(A_t r_t + b_t), \qquad \forall\ t, \tag{2.7}$$

where

$$B_t = \delta I + \sum_{m=0}^{t} \phi(i_m)\phi(i_m)', \qquad A_t = \sum_{k=0}^{t} z_k \big(\alpha\phi(i_{k+1})' - \phi(i_k)'\big), \tag{2.8}$$

$$b_t = \sum_{k=0}^{t} z_k g(i_k, i_{k+1}), \qquad z_k = \sum_{m=0}^{k} (\alpha\lambda)^{k-m} \phi(i_m). \tag{2.9}$$

This form of the method is well suited for practical implementation since the matrices B_t^{-1} and A_t, and the vector b_t can be updated recursively. More specifically, at iteration t, we have the matrices B_t^{-1} and A_t, and the vectors b_t and z_t. At iteration $t+1$, as soon as the transition from i_{t+1} to i_{t+2} takes place, we can compute B_{t+1}^{-1} by applying the Sherman-Morisson formula

$$(H + uv')^{-1} = H^{-1} - \frac{H^{-1}uv'H^{-1}}{1 + v'H^{-1}u},$$

which is valid for any invertible matrix H, and vectors u and v of compatible dimensions (see Golub and Van Loan [GoV96], p. 3). Hence, we have

$$B_{t+1}^{-1} = B_t^{-1} - \frac{B_t^{-1}\phi(i_{t+1})\phi(i_{t+1})'B_t^{-1}}{1 + \phi(i_{t+1})'B_t^{-1}\phi(i_{t+1})}.$$

Then, we compute the vector z_{t+1},

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

which we use to compute A_{t+1} and b_{t+1} as follows

$$A_{t+1} = A_t + z_{t+1}\big(\alpha\phi(i_{t+2})' - \phi(i_{t+1})'\big),$$

$$b_{t+1} = b_t + z_{t+1}g(i_{t+1}, i_{t+2}).$$

It can be seen from these recursive formulas that the iteration (2.7)-(2.9) is not computationally expensive. It is also possible to compute the inverse of B_{t+1} by using the Singular Value Decomposition, which has more overhead per iteration but can have better numerical stability properties. Regarding the memory space required per iteration, note that B_t^{-1} and A_t are matrices of dimension $K \times K$, and z_t and b_t are vectors of dimension K, where K is the number of feature vectors. Thus, the iteration (2.7)-(2.9) can be implemented efficiently.

2.3 Convergence Properties

We now discuss the convergence properties of the λ-LSPE method (2.7)-(2.9). We will use the following assumption.

Assumption 2.1:

(a) The Markov chain has steady-state probabilities $\pi(1), \ldots, \pi(n)$ which are positive, i.e.,

$$\lim_{t\to\infty} P[i_t = j \mid i_0 = i] = \pi(j) > 0, \qquad \forall\ i, j.$$

(b) The matrix Φ given by

$$\Phi = \begin{bmatrix} -\ \phi(1)'\ - \\ \vdots \\ -\ \phi(n)'\ - \end{bmatrix}$$

has full column rank.

For convergence of the λ-LSPE method, the following additional assumption is required for the stepsize γ_t. This assumption is typical for stochastic iterative methods using a diminishing stepsize.

Assumption 2.2: The stepsize γ_t is deterministic and satisfies

$$\sum_{t=0}^{\infty} \gamma_t = \infty, \qquad \sum_{t=0}^{\infty} \gamma_t^2 < \infty.$$

By Assumption 2.1(a), the underlying Markov chain is ergodic, and all states are visited an infinite number of times by the simulated trajectory. By Assumption 2.1(b), the number of features can be smaller than the number of states, which is important for large-scale problems.

Under Assumption 2.1, Tsitsiklis and Van Roy [TsV97] (see also [BeT96]) have considered the linear system of equations

$$Ar + b = 0,$$

where A and b are given by

$$A = \Phi' D(\alpha P - I) \sum_{s=0}^{\infty} (\alpha\lambda P)^s \Phi, \qquad b = \Phi' D \sum_{s=0}^{\infty} (\alpha\lambda P)^s \bar{g}, \tag{2.10}$$

P is the transition probability matrix of the Markov chain, and $\bar{g}$ is the vector with components $\bar{g}(i) = \sum_{j=1}^{n} p_{ij} g(i,j)$. They have shown that A is a negative definite matrix, so that the system $Ar + b = 0$ has a unique solution denoted r^*:

$$r^* = A^{-1} b.$$

Furthermore, r^* satisfies the following error bound

$$\|\Phi r^* - J\|_D \leq \frac{1 - \alpha\lambda}{1 - \alpha} \|\Pi J - J\|_D,$$

where D is the diagonal matrix with diagonal entries $\pi(i)$, $\|\cdot\|_D$ is the norm induced by the matrix D $\big($i.e., $\|x\|_D = \sqrt{x' D x}\big)$, and Π is the matrix given by $\Pi = \Phi(\Phi' D \Phi)^{-1} \Phi' D$. Note that as λ decreases, the error bound deteriorates, which indicates that from the point of view of approximation accuracy it is better to use $\lambda = 1$. Indeed, this is confirmed by examples in Bertsekas [Ber95]. On the other hand, as λ decreases, the detrimental effects of simulation noise seem to be ameliorated, and in practice a value of λ that is less than 1 may be overall preferable.

Under Assumptions 2.1 and 2.2, Tsitsiklis and Van Roy [TsV97] have shown that TD(λ) converges to r^*. The λ-LSPE method will also be shown to converge to r^* under the same assumptions, as stated in the following proposition.

Proposition 2.1: Let Assumptions 2.1 and 2.2 hold, and let the sequence $\{r_t\}$ be generated by the λ-LSPE method (2.7)-(2.9). Then for any $\lambda \in [0, 1]$, the sequence $\{r_t\}$ converges to r^* with probability 1.

3. 3. THE LSTD(λ) METHOD

In this section, we describe the LSTD(λ) method due to Boyan [Boy02]. Similar to the λ-LSPE method, we assume that an infinitely long trajectory is generated, starting at an initial state i_0, which is chosen according to some initial probability distribution. At each time t, we compute the weight vector r_t:

$$r_t = -A_t^{-1} b_t, \tag{3.1}$$

where the matrix A_t and the vector b_t are as in Eqs. (2.8) and (2.9). By writing

$$r_t = -\left(\frac{A_t}{t+1}\right)^{-1} \frac{b_t}{t+1},$$

we see that the LSTD(λ) method attempts to find the solution $r^* = A^{-1}b$ of the system $Ar+b=0$ by separately approximating A and b with their simulation-based approximations $A_t/(t+1)$ and $b_t/(t+1)$, respectively. Note that in the LSTD(λ) method, we only choose the parameter λ and the initial probability distribution for state i_0. Once this is done, we cannot control the method any further. In particular, there is no provision to take advantage of a good initial guess of the vector r.

We now show how to implement the LSTD(λ) method recursively. We first note that the computation of r_t in Eq. (3.1) requires the inverse of A_t, which may not exist initially. To ensure invertibility of A_t for all t, instead of $A_0 = \phi(i_0)\big(\alpha\phi(i_1)' - \phi(i_0)'\big)$, we can use

$$A_0 = \delta I + \phi(i_0)\big(\alpha\phi(i_1)' - \phi(i_0)'\big), \tag{3.2}$$

where δ is a positive scalar. Furthermore, by using the definition of A_t in Eq. (2.8) and by applying the Sherman-Morisson formula, we can recursively compute the inverse of A_t:

$$A_t^{-1} = \Big(A_{t-1} + z_t\big(\alpha\phi(i_{t+1})' - \phi(i_t)'\big)\Big)^{-1} = A_{t-1}^{-1} - \frac{A_{t-1}^{-1} z_t \big(\alpha\phi(i_{t+1})' - \phi(i_t)'\big) A_{t-1}^{-1}}{1 + \big(\alpha\phi(i_{t+1})' - \phi(i_t)'\big) A_{t-1}^{-1} z_t}.$$

This provides an efficient recursive implementation of the method.

The following convergence result applies to both cases where $\delta = 0$ and $\delta > 0$.

Proposition 3.1: Let Assumption 2.1 hold, and let the sequence $\{r_t\}$ be generated by the LSTD(λ) method (3.1). Then for any $\lambda \in [0, 1]$, the sequence $\{r_t\}$ converges to r^* with probability 1.

4. 4. CONVERGENCE PROOF FOR THE λ-LSPE METHOD

In this section, we give a proof of Proposition 2.1. The proof is long, so we break it down into two major steps. The first step is to view the method as a special case of a more general method of the form

$$x_{t+1} = x_t + \gamma_t(h_t + e_t),$$

where h_t is a descent direction of a cost function f, while e_t is a noise term. We establish sufficient conditions for convergence of this method in Proposition 4.1 below. The second step is to show that, for an appropriate choice of h_t, e_t, and f, Proposition 4.1 applies to the λ-LSPE method.

We first state two theorems needed in the proof of Proposition 4.1. Proofs of these theorems can be found in Neveu [Nev75].

Theorem 4.1: (*Supermartingale Convergence Theorem*) Let $\{X_t\}$, $\{Y_t\}$, and $\{Z_t\}$ be sequences of random variables, and let $\{\mathcal{F}_t\}$ be a sequence of sets of random variables such that $\mathcal{F}_t \subset \mathcal{F}_{t+1}$ for all t. Suppose that:

(i) For each t, the random variables X_t, Y_t, and Z_t are nonnegative, and are functions of the random variables in $\mathcal{F}_t$.

(ii) For each t, we have $E[X_{t+1} \mid \mathcal{F}_t] \leq X_t - Y_t + Z_t$.

(iii) There holds $\sum_{t=0}^{\infty} Z_t < \infty$.

Then, with probability 1, the sequence $\{X_t\}$ converges to a nonnegative random variable and $\sum_{t=0}^{\infty} Y_t < \infty$.

Theorem 4.2: (*Martingale Convergence Theorem*) Let $\{X_t\}$ be a sequence of random variables and $\{\mathcal{F}_t\}$ be a sequence of sets of random variables such that $\mathcal{F}_t \subset \mathcal{F}_{t+1}$ for all t. Suppose that:

(i) For each t, the random variable X_t is a function of the random variables in $\mathcal{F}_t$.

(ii) For each t, we have $E[X_{t+1} \mid \mathcal{F}_t] = X_t$.

(iii) There holds $\sup_t E\big[X_t^2\big] < \infty$.

Then, with probability 1, the sequence $\{X_t\}$ converges to a random variable.

The following proposition is an extension of Proposition 4.1 of Bertsekas and Tsitsiklis [BeT96] (see also Proposition 3 of Bertsekas and Tsitsiklis [BeT00]).

Proposition 4.1: Let $f : \Re^n \mapsto \Re$ be a continuously differentiable function, and consider a sequence $\{x_t\}$ generated by the method

$$x_{t+1} = x_t + \gamma_t (h_t + e_t),$$

where γ_t is a positive deterministic stepsize, h_t is a direction, and e_t is a random noise vector. Let $\mathcal{F}_t = \{x_0, x_1, \ldots, x_t\}$ for all t, and assume the following:

(i) The function f satisfies $f(x) \geq 0$ for all x, and has a Lipschitz continuous gradient, i.e., for some positive scalar L,

$$\|\nabla f(x) - \nabla f(\overline{x})\| \leq L\|x - \overline{x}\|, \qquad \forall\ x, \overline{x}.$$

(ii) There exist positive scalars c_1, c_2, and c_3 such that

$$\nabla f(x_t)' E[h_t \mid \mathcal{F}_t] \leq -c_1 \|\nabla f(x_t)\|^2, \qquad \forall\ t,$$

$$\big\|E[e_t \mid \mathcal{F}_t]\big\| \leq c_2 \varepsilon_t \big(1 + \|\nabla f(x_t)\|\big), \qquad \forall\ t,$$

$$E\big[\|h_t + e_t\|^2 \mid \mathcal{F}_t\big] \leq c_3 \big(1 + \|\nabla f(x_t)\|^2\big), \qquad \forall\ t,$$

where ε_t is a positive deterministic scalar.

(iii) The deterministic sequences $\{\gamma_t\}$ and $\{\varepsilon_t\}$ satisfy

$$\sum_{t=0}^{\infty} \gamma_t = \infty, \qquad \sum_{t=0}^{\infty} \gamma_t^2 < \infty, \qquad \sum_{t=0}^{\infty} \gamma_t \varepsilon_t^2 < \infty, \qquad \lim_{t \to \infty} \varepsilon_t = 0.$$

Then, with probability 1:

(a) The sequence $\big\{f(x_t)\big\}$ converges.

(b) The sequence $\big\{\nabla f(x_t)\big\}$ converges to zero.

(c) Every limit point of $\{x_t\}$ is a stationary point of f.

Proof: The idea of the proof is to use a second order Taylor's series expansion to show that $E\big[f(r_{t+1}) \mid \mathcal{F}_t\big] \leq f(r_t) - Y_t + Z_t$, where Y_t and Z_t are nonnegative with $\sum_{t=0}^{\infty} Z_t < \infty$. The convergence of $f(r_t)$ will then follow by the Supermartingale Convergence Theorem. The convergence of $\nabla f(r_t)$ to zero will be established by showing that the excursions of $\nabla f(r_t)$ away from zero are arbitrarily small with probability 1.

Since the gradient of f is Lipschitz continuous, it follows that

$$f(x) \leq f(\overline{x}) + \nabla f(\overline{x})'(x - \overline{x}) + \frac{L}{2}\|x - \overline{x}\|^2, \qquad \forall\ x, \overline{x},$$

(cf. Bertsekas [Ber99], Proposition A.24), so that by using this inequality with $x = x_{t+1} = x_t + \gamma_t(h_t + e_t)$ and $\overline{x} = x_t$, we have

$$f(x_{t+1}) \leq f(x_t) + \gamma_t \nabla f(x_t)'(h_t + e_t) + \frac{L}{2}\gamma_t^2 \|h_t + e_t\|^2, \qquad \forall\ t.$$

By taking the conditional expectation of both sides of this inequality, with respect to $\mathcal{F}_t = \{x_0, x_1, \dots, x_t\}$, and by using assumption (ii), we obtain for all t,

$$\begin{aligned} E\big[f(x_{t+1}) \mid \mathcal{F}_t\big] &\le f(x_t) + \gamma_t \nabla f(x_t)'\big(E[h_t \mid \mathcal{F}_t] + E[e_t \mid \mathcal{F}_t]\big) + \frac{L}{2}\gamma_t^2 E\big[\|h_t + e_t\|^2 \mid \mathcal{F}_t\big] \\ &\le f(x_t) + \gamma_t \nabla f(x_t)' E[h_t \mid \mathcal{F}_t] + \gamma_t \|\nabla f(x_t)\| \, \big\|E[e_t \mid \mathcal{F}_t]\big\| \\ &\quad + \frac{L}{2}\gamma_t^2 E\big[\|h_t + e_t\|^2 \mid \mathcal{F}_t\big] \\ &\le f(x_t) - \gamma_t \left(c_1 - c_2\varepsilon_t - \frac{Lc_3}{2}\gamma_t\right) \|\nabla f(x_t)\|^2 + c_2\gamma_t\varepsilon_t\|\nabla f(x_t)\| + \frac{Lc_3}{2}\gamma_t^2. \end{aligned}$$

Since

$$c_2\gamma_t\varepsilon_t\|\nabla f(x_t)\| = \gamma_t\left(\frac{c_2}{\sqrt{c_1}}\varepsilon_t\right)\sqrt{c_1}\|\nabla f(x_t)\| \le \frac{\gamma_t}{2}\left(\frac{c_2^2}{c_1}\varepsilon_t^2 + c_1\|\nabla f(x_t)\|^2\right),$$

it follows that for all t,

$$\begin{aligned} E\big[f(x_{t+1}) \mid \mathcal{F}_t\big] &\le f(x_t) - \gamma_t\left(\frac{c_1}{2} - c_2\varepsilon_t - \frac{Lc_3}{2}\gamma_t\right)\|\nabla f(x_t)\|^2 + \frac{c_2^2}{2c_1}\gamma_t\varepsilon_t^2 + \frac{Lc_3}{2}\gamma_t^2 \\ &\le f(x_t) - Y_t + Z_t, \end{aligned}$$

where

$$Y_t = \begin{cases} \gamma_t\left(\frac{c_1}{2} - c_2\varepsilon_t - \frac{Lc_3}{2}\gamma_t\right)\|\nabla f(x_t)\|^2 & \text{if } \frac{c_1}{2} \ge c_2\varepsilon_t + \frac{Lc_3}{2}\gamma_t, \\ 0 & \text{otherwise,} \end{cases}$$

$$Z_t = \begin{cases} \frac{c_2^2}{2c_1}\gamma_t\varepsilon_t^2 + \frac{Lc_3}{2}\gamma_t^2 & \text{if } \frac{c_1}{2} \ge c_2\varepsilon_t + \frac{Lc_3}{2}\gamma_t, \\ -\gamma_t\left(\frac{c_1}{2} - c_2\varepsilon_t - \frac{Lc_3}{2}\gamma_t\right)\|\nabla f(x_t)\|^2 + \frac{c_2^2}{2c_1}\gamma_t\varepsilon_t^2 + \frac{Lc_3}{2}\gamma_t^2 & \text{otherwise.} \end{cases}$$

Thus, Y_t and Z_t are nonnegative for all t, and because $\gamma_t \to 0$ and $\varepsilon_t \to 0$, after some finite time, Y_t and Z_t are given by

$$Y_t = \gamma_t\left(\frac{c_1}{2} - c_2\varepsilon_t - \frac{Lc_3}{2}\gamma_t\right)\|\nabla f(x_t)\|^2,$$

$$Z_t = \frac{c_2^2}{2c_1}\gamma_t\varepsilon_t^2 + \frac{Lc_3}{2}\gamma_t^2.$$

Since $\sum_{t=0}^\infty \gamma_t^2 < \infty$ and $\sum_{t=0}^\infty \gamma_t\varepsilon_t^2 < \infty$, it follows that $\sum_{t=0}^\infty Z_t < \infty$, so by the Supermartingale Convergence Theorem, the sequence $\big\{f(x_t)\big\}$ converges with probability 1, showing part (a). By the same theorem, there holds $\sum_{t=0}^\infty Y_t < \infty$ with probability 1, and since after some finite time

$$Y_t = \gamma_t\left(\frac{c_1}{2} - c_2\varepsilon_t - \frac{Lc_3}{2}\gamma_t\right)\|\nabla f(x_t)\|^2 \ge \frac{c_1}{4}\gamma_t\|\nabla f(x_t)\|^2,$$

it follows that, with probability 1,

$$\sum_{t=0}^\infty \gamma_t\|\nabla f(x_t)\|^2 < \infty. \tag{4.1}$$

We next show that $\limsup_{t\to\infty}\|\nabla f(x_t)\| = 0$ with probability 1, by studying the excursions of $\nabla f(x_t)$ away from zero. We fix a positive scalar ϵ and we say that the time interval $\{t, t+1, \dots, \overline{t}\}$ is an *upcrossing interval* (from $\epsilon/2$ to ϵ) if

$$\|\nabla f(x_t)\| \le \frac{\epsilon}{2}, \qquad \|\nabla f(x_{\overline{t}})\| > \epsilon, \qquad \frac{\epsilon}{2} \le \|\nabla f(x_\tau)\| \le \epsilon, \qquad \forall\ \tau,\ t < \tau < \overline{t}.$$

We want to show that almost every sample path has a finite number of upcrossing intervals, and to do so, we first prove a result (Lemma 4.1 below), which we use to bound the effects of the noise e_t within an upcrossing interval.

Let

$$s_t = h_t + e_t, \qquad \overline{s}_t = E[s_t \mid \mathcal{F}_t], \qquad w_t = s_t - \overline{s}_t, \qquad \forall\ t. \tag{4.2}$$

By using the definitions of $\overline{s}_t$ and s_t, and the relation $E\big[\|s_t\|^2 \mid \mathcal{F}_t\big] \le c_3\big(1 + \|\nabla f(x_t)\|^2\big)$ of assumption (ii), since $w_t + \overline{s}_t = s_t$, we have

$$E\big[\|w_t\|^2 \mid \mathcal{F}_t\big] + \|\overline{s}_t\|^2 = E\big[\|s_t\|^2 \mid \mathcal{F}_t\big] \le c_3\big(1 + \|\nabla f(x_t)\|^2\big), \qquad \forall\ t. \tag{4.3}$$

Define

$$u_t = \sum_{\tau=0}^{t-1} \gamma_\tau \nu_\tau w_\tau, \qquad \forall\ t \ge 1,$$

where w_t is as in Eq. (4.2) and

$$\nu_t = \begin{cases} 1 & \text{if } \|\nabla f(x_t)\| \le \epsilon, \\ 0 & \text{otherwise,} \end{cases} \qquad \forall\ t.$$

We have the following lemma.

Lemma 4.1: The vector sequence $\{u_t\}$ converges with probability 1.

Proof: We will show that the vector sequence $\{u_t\}$ is a martingale such that $\sup_t E\big[\|u_t\|^2\big] < \infty$, and then we will apply (componentwise) the Martingale Convergence Theorem. We have

$$E[\gamma_t \nu_t w_t \mid \mathcal{F}_t] = \gamma_t \nu_t E[w_t \mid \mathcal{F}_t] = 0, \qquad \forall\ t,$$

and by using the definition of u_t, we obtain

$$E[u_{t+1} \mid \mathcal{F}_t] = u_t + E[\gamma_t \nu_t w_t \mid \mathcal{F}_t] = u_t, \qquad \forall\ t,$$

showing that the sequence $\{u_t\}$ is a martingale.

If $\nu_t = 0$, then $u_{t+1} = u_t$, implying that

$$E\big[\|u_{t+1}\|^2\big] = E\big[\|u_t\|^2\big].$$

If $\nu_t = 1$, then $\|\nabla f(x_t)\| \le \epsilon$, so by using the relation $E[w_t \mid \mathcal{F}_t] = 0$ and Eq. (4.3), we obtain

$$E\big[\|u_{t+1}\|^2 \mid \mathcal{F}_t\big] = \|u_t\|^2 + 2u_t'\gamma_t E[w_t \mid \mathcal{F}_t] + \gamma_t^2 E\big[\|w_t\|^2 \mid \mathcal{F}_t\big] \le \|u_t\|^2 + c_3\gamma_t^2(1 + \epsilon^2).$$

Hence, $E\big[\|u_{t+1}\|^2\big] \le E\big[\|u_t\|^2\big] + c_3\gamma_t^2(1 + \epsilon^2)$ for all t, and therefore,

$$E\big[\|u_{t+1}\|^2\big] \le c_3(1 + \epsilon^2) \sum_{\tau=0}^{\infty} \gamma_\tau^2 < \infty, \qquad \forall\ t,$$

showing that $\sup_t E\big[\|u_t\|^2\big] < \infty$. We can now apply (componentwise) the Martingale Convergence Theorem to conclude that $\{u_t\}$ converges with probability 1. **Q.E.D.**

Let us now consider a sample path $\mathcal{P}$ such that the sequence $\{u_t\}$ converges. To arrive at a contradiction, assume that the path $\mathcal{P}$ has an infinite number of upcrossing intervals and let $\{t_k, \ldots, \bar{t}_k\}$ be the kth such interval, so that

$$\|\nabla f(x_{t_k})\| \leq \frac{\epsilon}{2}, \qquad \|\nabla f(x_{\bar{t}_k})\| > \epsilon, \qquad \frac{\epsilon}{2} \leq \|\nabla f(x_t)\| \leq \epsilon, \qquad \forall\ t,\ t_k < t < \bar{t}_k, \quad \forall\ k. \tag{4.4}$$

By the definition of ν_t, we have $\nu_t = 1$ for all t such that $t_k \leq t < \bar{t}_k$, and since $\{u_t\}$ converges, it follows that

$$\lim_{k\to\infty} \sum_{t=t_k}^{\bar{t}_k - 1} \gamma_t w_t = 0. \tag{4.5}$$

Using the Lipschitz continuity of ∇f, the relation $x_{t+1} = x_t + \gamma_t(h_t + e_t)$, and Eq. (4.4), we have for all k,

$$\frac{\epsilon}{2} \leq \|\nabla f(x_{\bar{t}_k})\| - \|\nabla f(x_{t_k})\| \leq \|\nabla f(x_{\bar{t}_k}) - \nabla f(x_{t_k})\| \leq L\|x_{\bar{t}_k} - x_{t_k}\| \leq L \left\| \sum_{t=t_k}^{\bar{t}_k - 1} \gamma_t(h_t + e_t) \right\|,$$

which in view of the relation $h_t + e_t = w_t + \bar{s}_t$ [cf. Eq. (4.2)], implies that

$$\frac{\epsilon}{2} \leq L \left\| \sum_{t=t_k}^{\bar{t}_k - 1} \gamma_t w_t \right\| + L \sum_{t=t_k}^{\bar{t}_k - 1} \gamma_t \|\bar{s}_t\|, \qquad \forall\ k.$$

Since $\|\nabla f(x_t)\| < \epsilon$ for all t with $t_k \leq t < \bar{t}_k$ [cf. Eq. (4.4)], in view of the relation (4.3), it follows that $\|\bar{s}_t\|$ is bounded by $\sqrt{c_3(1+\epsilon^2)}$ for all t with $t_k \leq t < \bar{t}_k$. Therefore,

$$\sum_{t=t_k}^{\bar{t}_k - 1} \gamma_t \geq \frac{\epsilon}{2Lc} - \frac{1}{c} \left\| \sum_{t=t_k}^{\bar{t}_k - 1} \gamma_t w_t \right\|, \qquad \forall\ k,$$

where $c = \sqrt{c_3(1+\epsilon^2)}$. By taking the limit inferior as $k \to \infty$ and by using Eq. (4.5), we obtain

$$\liminf_{k\to\infty} \sum_{t=t_k}^{\bar{t}_k - 1} \gamma_t \geq \frac{\epsilon}{2Lc},$$

and since $\gamma_{t_k} \to 0$, it follows that

$$\liminf_{k\to\infty} \sum_{t=t_k+1}^{\bar{t}_k - 1} \gamma_t = \liminf_{k\to\infty} \sum_{t=t_k}^{\bar{t}_k - 1} \gamma_t \geq \frac{\epsilon}{2Lc}.$$

This relation, the inequality $\epsilon/2 \leq \|\nabla f(x_t)\|$ for all t with $t_k < t < \bar{t}_k$ and all k [cf. Eq. (4.4)], and the assumption that the path $\mathcal{P}$ has an infinite number of upcrossing intervals imply that

$$\sum_{t=0}^{\infty} \gamma_t \|\nabla f(x_t)\|^2 \geq \sum_{k=1}^{\infty} \sum_{t=t_k+1}^{\bar{t}_k - 1} \gamma_t \|\nabla f(x_t)\|^2 > \infty,$$

contradicting Eq. (4.1). Hence, the path $\mathcal{P}$ must have a finite number of upcrossing intervals, so that $\|\nabla f(x_t)\|$ can exceed ϵ only a finite number of times, and therefore $\limsup_{t\to\infty} \|\nabla f(x_t)\| \leq \epsilon$. Since ϵ is arbitrary, it follows that $\limsup_{t\to\infty} \|\nabla f(x_t)\| = 0$ for the path $\mathcal{P}$, showing part (b). Finally, if x is a limit point of $\{x_t\}$, then $\nabla f(x)$ is a limit point of $\big\{\nabla f(x_t)\big\}$, and by part (b), $\nabla f(x) = 0$, thus showing part (c) and completing the proof. **Q.E.D.**

We will now need a lemma relating to Markov chains. In what follows, we denote by $\kappa_t(i)$ the number of visits to state i up to time t.

Lemma 4.2: Let Assumption 2.1(a) hold. Then:

(a) With probability 1,

$$\lim_{t\to\infty} \frac{\kappa_t(i)}{t+1} = \pi(i), \qquad \forall\ i = 1, \ldots, n.$$

(b) For some positive scalar C,

$$E\left[\left(\frac{\kappa_t(i)}{t+1} - \pi(i)\right)^2\right] \leq \frac{C}{t+1}, \qquad \forall\ i = 1, \ldots, n, \quad \forall\ t.$$

(c) For some positive scalars c_i,

$$\lim_{t\to\infty} \Big(E\big[\kappa_t(i)\big] - (t+1)\pi(i)\Big) = c_i, \qquad \forall\ i = 1, \ldots, n.$$

For part (a) of the preceding lemma, see Theorem 1 on p. 145 in Gallager [Gal95] or Theorem 4.2.1 in Kemeny and Snell [KeS67]. For part (b), see the proof of Theorem 4.2.1 in Kemeny and Snell [KeS67], while for part (c), see Theorem 4.3.4 in Kemeny and Snell [KeS67].

By using the preceding lemma, we next establish some relations for the matrices B_t and A_t, and vectors b_t, as defined in Eqs. (2.8) and (2.9).

Lemma 4.3: Let Assumption 2.1 hold. Then:

(a) There exist positive scalars C_1, C_2, and C_3 such that with probability 1

$$\left\|\left(\frac{B_t}{t+1}\right)^{-1}\right\| \leq C_1, \qquad \left\|\frac{A_t}{t+1}\right\| \leq C_2, \qquad \left\|\frac{b_t}{t+1}\right\| \leq C_3, \qquad \forall\ t.$$

(b) There exists a positive scalar C_4 such that

$$E\left[\left\|\left(\frac{B_t}{t+1}\right)^{-1} - (\Phi' D\Phi)^{-1}\right\|\right] \leq \frac{C_4}{\sqrt{t+1}}, \qquad \forall\ t.$$

(c) For all t,

$$\sum_{k=0}^{t}\sum_{m=0}^{k}(\alpha\lambda)^{k-m}\phi(i_m)E\big[\alpha\phi(i_{k+1})'-\phi(i_k)'|i_m\big]=\sum_{i=1}^{n}\kappa_t(i)\phi(i)\sum_{s=0}^{\infty}\big[(\alpha P-I)(\alpha\lambda P)^s\Phi\big]_i-V_t,$$

$$V_t=\sum_{m=0}^{t}\phi(i_m)\sum_{k=t+1}^{\infty}\big[(\alpha P-I)(\alpha\lambda P)^{k-m}\Phi\big]_{i_m},$$

$$\sum_{k=0}^{t}\sum_{m=0}^{k}(\alpha\lambda)^{k-m}\phi(i_m)E\big[g(i_k,i_{k+1})\mid i_m\big]=\sum_{i=1}^{n}\kappa_t(i)\phi(i)\sum_{s=0}^{\infty}\big[(\alpha\lambda P)^s\bar{g}\big](i)-v_t,$$

$$v_t=\sum_{m=0}^{t}\phi(i_m)\sum_{k=t+1}^{\infty}\big[(\alpha\lambda P)^{k-m}\bar{g}\big](i_m),$$

where $\big[(\alpha P-I)(\alpha\lambda P)^s\Phi\big]_i$ denotes the ith row vector of the matrix $(\alpha P-I)(\alpha\lambda P)^s\Phi$, $\big[(\alpha\lambda P)^s\bar{g}\big](i)$ denotes the ith component of the vector $(\alpha\lambda P)^s\bar{g}$, and $\bar{g}$ is a vector with components $\bar{g}(i)=\sum_{i=1}^{n}p_{ij}g(i,j)$.

(d) There exist positive scalars C_5 and C_6 such that

$$\left\|\frac{E[A_t]}{t+1}-A\right\|\le\frac{C_5}{t+1},\qquad\left\|\frac{E[b_t]}{t+1}-b\right\|\le\frac{C_6}{t+1},\qquad\forall\ t.$$

Proof: (a) By using $B_t=\delta I+\sum_{m=0}^{t}\phi(i_m)\phi(i_m)'$ [cf. Eq. (2.8)] and by writing

$$\sum_{m=0}^{t}\phi(i_m)\phi(i_m)'=\sum_{i=1}^{n}\kappa_t(i)\phi(i)\phi(i)',$$

we see that

$$B_t=\delta I+\sum_{i=1}^{n}\kappa_t(i)\phi(i)\phi(i)'. \tag{4.6}$$

By Lemma 4.2(a), it follows that $B_t/(t+1)$ converges to $\sum_{i=1}^{n}\pi(i)\phi(i)\phi(i)'$ with probability 1. Thus, the inverse of $B_t/(t+1)$ converges with probability 1, thereby implying that the norm of $\big(B_t/(t+1)\big)^{-1}$ is bounded. The boundedness of $A_t/(t+1)$ and $b_t/(t+1)$ follows from the definitions of A_t and b_t [cf. Eqs. (2.8) and (2.9)], and the relations

$$\big\|\phi(i_k)\big\|\le\max_i\big\|\phi(i)\big\|,\qquad\big|g(i_k,i_{k+1})\big|\le\max_{i,j}\big|g(i,j)\big|,\qquad\forall\ k.$$

(b) It can be seen that, for any two invertible matrices H and R, we have

$$H^{-1}-R^{-1}=H^{-1}(R-H)R^{-1},$$

implying that

$$\|H^{-1}-R^{-1}\|\le\|H^{-1}\|\,\|R-H\|\,\|R^{-1}\|.$$

By using this relation with $H = B_t/(t+1)$ and $R = \Phi' D\Phi$, which is invertible by Assumption 2.1(b), we obtain

$$\left\| \left(\frac{B_t}{t+1} \right)^{-1} - (\Phi' D\Phi)^{-1} \right\| \leq C_1 \left\| \frac{B_t}{t+1} - \Phi' D\Phi \right\| \left\| (\Phi' D\Phi)^{-1} \right\|, \qquad \forall\ t, \tag{4.7}$$

where C_1 is as in part (a). In view of Eq. (4.6) and the fact $\Phi' D\Phi = \sum_{i=1}^n \pi(i)\phi(i)\phi(i)'$, it follows that

$$\begin{aligned} \left\| \frac{B_t}{t+1} - \Phi' D\Phi \right\| &\leq \frac{\delta \|I\|}{t+1} + \left\| \sum_{i=1}^n \frac{\kappa_t(i)}{t+1} \phi(i)\phi(i)' - \sum_{i=1}^n \pi(i)\phi(i)\phi(i)' \right\| \\ &\leq \frac{\delta}{\sqrt{t+1}} + \max_l \|\phi(l)\|^2 \sum_{i=1}^n \left| \frac{\kappa_t(i)}{t+1} - \pi(i) \right|, \qquad \forall\ t, \end{aligned}$$

where we use $1/(t+1) \leq 1/\sqrt{t+1}$ for all t. By taking the expectation of both sides in the preceding inequality, we obtain

$$E\left[\left\| \frac{B_t}{t+1} - \Phi' D\Phi \right\| \right] \leq \frac{\delta}{\sqrt{t+1}} + \max_l \|\phi(l)\|^2 \sum_{i=1}^n E\left[\left| \frac{\kappa_t(i)}{t+1} - \pi(i) \right| \right], \qquad \forall\ t. \tag{4.8}$$

By Lemma 4.2(b), we have

$$E\left[\left(\frac{\kappa_t(i)}{t+1} - \pi(i) \right)^2 \right] \leq \frac{C}{t+1}, \qquad \forall\ t, \quad \forall\ i,$$

and by combining this relation with the inequality $E[|x|] \leq \sqrt{E[x^2]}$, we obtain

$$E\left[\left| \frac{\kappa_t(i)}{t} - \pi(i) \right| \right] \leq \sqrt{\frac{C}{t+1}}, \qquad \forall\ t, \quad \forall\ i.$$

The preceding relation and Eq. (4.8) imply that

$$E\left[\left\| \frac{B_t}{t+1} - \Phi' D\Phi \right\| \right] \leq \frac{\delta}{\sqrt{t+1}} + \max_l \|\phi(l)\|^2 \frac{n\sqrt{C}}{\sqrt{t+1}}, \qquad \forall\ t,$$

which together with Eq. (4.7) yields

$$E\left[\left\| \left(\frac{B_t}{t+1} \right)^{-1} - (\Phi' D\Phi)^{-1} \right\| \right] \leq \frac{C_4}{\sqrt{t+1}}, \qquad \forall\ t.$$

(c) For any m and k with $m \leq k$, we have

$$\begin{aligned} E\big[\alpha\phi(i_{k+1})' - \phi(i_k)' \mid i_m\big] &= \sum_{j=1}^n [P^{k-m}]_{i_m j} \Big(E\big[\alpha\phi(i_{k+1})' - \phi(j)' \mid i_k = j\big] \Big) \\ &= \sum_{j=1}^n [P^{k-m}]_{i_m j} \left(\alpha \sum_{l=1}^n [P]_{jl}\phi(l)' - \phi(j)' \right), \end{aligned}$$

where $[P^{k-m}]_{ij}$ denotes the ijth component of the matrix P^{k-m}. We further have

$$\sum_{j=1}^{n}\sum_{l=1}^{n}[P^{k-m}]_{i_m j}[P]_{jl}\phi(l)' = \sum_{l=1}^{n}[P^{k+1-m}]_{i_m l}\phi(l)' = [P^{k+1-m}\Phi]_{i_m},$$

$$\sum_{j=1}^{n}[P^{k-m}]_{i_m j}\phi(j)' = [P^{k-m}\Phi]_{i_m},$$

where $[P^{k-m}\Phi]_i$ denotes the ith row of the matrix $P^{k-m}\Phi$. Therefore,

$$E\big[\alpha\phi(i_{k+1})' - \phi(i_k)' \mid i_m\big] = [\alpha P^{k+1-m}\Phi]_{i_m} - [P^{k-m}\Phi]_{i_m} = \big[(\alpha P - I)P^{k-m}\Phi\big]_{i_m},$$

implying that for all t,

$$\sum_{k=0}^{t}\sum_{m=0}^{k}(\alpha\lambda)^{k-m}\phi(i_m)E\big[\alpha\phi(i_{k+1})' - \phi(i_k)' \mid i_m\big] = \sum_{k=0}^{t}\sum_{m=0}^{k}(\alpha\lambda)^{k-m}\phi(i_m)\big[(\alpha P - I)P^{k-m}\Phi\big]_{i_m}. \tag{4.9}$$

By exchanging the order of summation, we see that for all t,

$$\sum_{k=0}^{t}\sum_{m=0}^{k}(\alpha\lambda)^{k-m}\phi(i_m)\big[(\alpha P - I)P^{k-m}\Phi\big]_{i_m} = \sum_{m=0}^{t}\phi(i_m)\sum_{k=m}^{t}\big[(\alpha P - I)(\alpha\lambda P)^{k-m}\Phi\big]_{i_m}.$$

Using

$$V_t = \sum_{m=0}^{t}\phi(i_m)\sum_{k=t+1}^{\infty}\big[(\alpha P - I)(\alpha\lambda P)^{k-m}\Phi\big]_{i_m},$$

we further have for all t,

$$\begin{aligned}\sum_{k=0}^{t}\sum_{m=0}^{k}(\alpha\lambda)^{k-m}\phi(i_m)\big[(\alpha P - I)P^{k-m}\Phi\big]_{i_m} &= \sum_{m=0}^{t}\phi(i_m)\sum_{k=m}^{\infty}\big[(\alpha P - I)(\alpha\lambda P)^{k-m}\Phi\big]_{i_m} - V_t\\ &= \sum_{m=0}^{t}\phi(i_m)\sum_{s=0}^{\infty}\big[(\alpha P - I)(\alpha\lambda P)^{s}\Phi\big]_{i_m} - V_t\\ &= \sum_{i=1}^{n}\kappa_t(i)\phi(i)\sum_{s=0}^{\infty}\big[(\alpha P - I)(\alpha\lambda P)^{s}\Phi\big]_{i} - V_t.\end{aligned}$$

This relation and Eq. (4.9) yield for all t,

$$\sum_{k=0}^{t}\sum_{m=0}^{k}(\alpha\lambda)^{k-m}\phi(i_m)E\big[\alpha\phi(i_{k+1})' - \phi(i_k)' \mid i_m\big] = \sum_{i=1}^{n}\kappa_t(i)\phi(i)\sum_{s=0}^{\infty}\big[(\alpha P - I)(\alpha\lambda P)^{s}\Phi\big]_{i} - V_t.$$

Similarly, for any m and k with $m \le k$, we have

$$E\big[g(i_k, i_{k+1}) \mid i_m\big] = \sum_{j=1}^{n}[P^{k-m}]_{i_m j}E\big[g(j, i_{k+1}) \mid i_k = j\big] = \sum_{j=1}^{n}[P^{k-m}]_{i_m j}\bar{g}(j) = [P^{k-m}\bar{g}](i_m),$$

where $\bar{g}$ is a vector with components $\bar{g}(i) = \sum_{i=1}^{n} p_{ij} g(i,j)$, and $[P^{k-m}\bar{g}](i)$ denotes the ith component of the vector $P^{k-m}\bar{g}$. Therefore, for all t,

$$\sum_{k=0}^{t}\sum_{m=0}^{k}(\alpha\lambda)^{k-m}\phi(i_m)E\big[g(i_k,i_{k+1}) \mid i_m\big] = \sum_{k=0}^{t}\sum_{m=0}^{k}(\alpha\lambda)^{k-m}\phi(i_m)[P^{k-m}\bar{g}](i_m), \tag{4.10}$$

and by exchanging the order of summation, we see that for all t,

$$\sum_{k=0}^{t}\sum_{m=0}^{k}(\alpha\lambda)^{k-m}\phi(i_m)[P^{k-m}\bar{g}](i_m) = \sum_{m=0}^{t}\phi(i_m)\sum_{k=m}^{t}\big[(\alpha\lambda P)^{k-m}\bar{g}\big](i_m).$$

By using

$$v_t = \sum_{m=0}^{t}\phi(i_m)\sum_{k=t+1}^{\infty}\big[(\alpha\lambda P)^{k-m}\bar{g}\big](i_m),$$

we obtain for all t,

$$\begin{aligned}\sum_{k=0}^{t}\sum_{m=0}^{k}(\alpha\lambda)^{k-m}\phi(i_m)[P^{k-m}\bar{g}](i_m) &= \sum_{m=0}^{t}\phi(i_m)\sum_{k=m}^{\infty}\big[(\alpha\lambda P)^{k-m}\bar{g}\big](i_m) - v_t\\ &= \sum_{m=0}^{t}\phi(i_m)\sum_{s=0}^{\infty}\big[(\alpha\lambda P)^{s}\bar{g}\big](i_m) - v_t\\ &= \sum_{i=1}^{n}\kappa_t(i)\phi(i)\sum_{s=0}^{\infty}\big[(\alpha\lambda P)^{s}\bar{g}\big](i) - v_t.\end{aligned}$$

This relation and Eq. (4.10) show that

$$\sum_{k=0}^{t}\sum_{m=0}^{k}(\alpha\lambda)^{k-m}\phi(i_m)E\big[g(i_k,i_{k+1}) \mid i_m\big] = \sum_{i=1}^{n}\kappa_t(i)\phi(i)\sum_{s=0}^{\infty}\big[(\alpha\lambda P)^{s}\bar{g}\big](i) - v_t, \qquad \forall\ t.$$

(d) From the definition of A_t [cf. Eq. (2.8)], we have

$$E[A_t] = E\left[\sum_{k=0}^{t}\sum_{m=0}^{k}(\alpha\lambda)^{k-m}\phi(i_m)\big(\alpha\phi(i_{k+1})' - \phi(i_k)'\big)\right], \qquad \forall\ t,$$

and by using the iterated expectation rule and part (c), we obtain

$$\begin{aligned}E[A_t] &= E\left[\sum_{k=0}^{t}\sum_{m=0}^{k}(\alpha\lambda)^{k-m}\phi(i_m)E\big[\alpha\phi(i_{k+1})' - \phi(i_k)' \mid i_m\big]\right]\\ &= \sum_{i=1}^{n}E\big[\kappa_t(i)\big]\phi(i)\sum_{s=0}^{\infty}\big[(\alpha P - I)(\alpha\lambda P)^{s}\Phi\big]_i - E[V_t], \qquad \forall\ t.\end{aligned}$$

By writing the matrix A [cf. Eq. (2.10)] equivalently as

$$A = \sum_{i=1}^{n}\pi(i)\phi(i)\sum_{s=0}^{\infty}\big[(\alpha P - I)(\alpha\lambda P)^{s}\Phi\big]_i,$$

we have

$$\frac{E[A_t]}{t+1} - A = \sum_{i=1}^{n} \left(\frac{E\big[\kappa_t(i)\big]}{t+1} - \pi(i) \right) \phi(i) \sum_{s=0}^{\infty} \big[(\alpha P - I)(\alpha\lambda P)^s \Phi\big]_i - \frac{E[V_t]}{t+1}, \qquad \forall\ t.$$

From the definition of V_t in part (c), we can see that $\|V_t\|$ is bounded, while by Lemma 4.2(c), we have that $\big|E[\kappa_t(i)]/(t+1) - \pi(i)\big|$ is bounded by a constant multiple of $1/(t+1)$ for all t and i. Hence, $\big\|E[A_t]/(t+1) - A\big\|$ is bounded by a constant multiple of $1/(t+1)$ for all t.

Similarly, for b_t as defined in Eq. (2.8), we have

$$E[b_t] = E\left[\sum_{k=0}^{t} \sum_{m=0}^{k} (\alpha\lambda)^{k-m} \phi(i_m) g(i_k, i_{k+1}) \right], \qquad \forall\ t,$$

and by using the iterated expectation rule and part (c), we obtain

$$\begin{aligned} E[b_t] &= E\left[\sum_{k=0}^{t} \sum_{m=0}^{k} (\alpha\lambda)^{k-m} \phi(i_m) E\big[g(i_k, i_{k+1}) \mid i_m\big] \right] \\ &= \sum_{i=1}^{n} E\big[\kappa_t(i)\big] \phi(i) \sum_{s=0}^{\infty} \big[(\alpha\lambda P)^s \bar{g}\big](i) - E[v_t], \qquad \forall\ t. \end{aligned}$$

Since we can write the vector b [cf. Eq. (2.10)] equivalently as

$$b = \sum_{i=1}^{n} \pi(i) \phi(i) \sum_{s=0}^{\infty} \big[(\alpha\lambda P)^s \bar{g}\big](i),$$

it follows that

$$\frac{E[b_t]}{t+1} - b = \sum_{i=1}^{n} \left(\frac{E\big[\kappa_t(i)\big]}{t+1} - \pi(i) \right) \phi(i) \sum_{s=0}^{\infty} \big[(\alpha\lambda P)^s \bar{g}\big](i) - \frac{E[v_t]}{t+1}, \qquad \forall\ t.$$

Using the definition of v_t in part (c), we can see that $\|v_t\|$ is bounded, while by Lemma 4.2(c), we have that $\big|E[\kappa_t(i)]/(t+1) - \pi(i)\big|$ is bounded by a constant multiple of $1/(t+1)$ for all t and i. Therefore, $\big\|E[b_t]/(t+1) - b\big\|$ is bounded by a constant multiple of $1/(t+1)$ for all t. **Q.E.D.**

We are now ready to prove Proposition 2.1. In the proof, we use Proposition 4.1 and Lemma 4.3. We also use the negative definiteness of the matrix A [cf. Eq. (2.10)] , which was shown by Tsitsiklis and Van Roy [TsV97].

Proof of Proposition 2.1: The idea of the proof is to write $r_{t+1} = r_t + \gamma_t(h_t + e_t)$ with appropriately chosen vectors h_t and e_t, and to show that Proposition 4.1 applies for a suitable choice of function f, which will imply the desired convergence of $\{r_t\}$.

Let $\lambda \in [0,1]$ be arbitrary. We can rewrite the iteration $r_{t+1} = r_t + \gamma_t B_t^{-1}(A_t r_t + b_t)$ as

$$r_{t+1} = r_t + \gamma_t(h_t + e_t),$$

where

$$h_t = (\Phi' D\Phi)^{-1}(Ar_t + b),$$

$$e_t = B_t^{-1}(A_t r_t + b_t) - (\Phi' D\Phi)^{-1}(Ar_t + b),$$

with A and b as in Eq. (2.10). Let

$$f(r) = \frac{1}{2}(r - r^*)'\Phi' D\Phi(r - r^*), \qquad \forall\ r, \tag{4.11}$$

where r^* is the solution of the system $Ar^* + b = 0$. Then, clearly, f is nonnegative and has a Lipschitz continuous gradient, so that assumption (i) of Proposition 4.1 is satisfied.

We next show that assumption (ii) of Proposition 4.1 holds. Let $\mathcal{F}_t = \{r_0, r_1, \ldots, r_t\}$ for all t. Then, by using the definition of h_t and the relation $b = -Ar^*$, we have

$$\nabla f(r_t)' E[h_t \mid \mathcal{F}_t] = \nabla f(r_t)'(\Phi' D\Phi)^{-1} A(r_t - r^*), \qquad \forall\ t.$$

By writing $A = A(\Phi' D\Phi)^{-1}\Phi' D\Phi$ and by using $\Phi' D\Phi(r_t - r^*) = \nabla f(r_t)$, we obtain

$$\nabla f(r_t)' E[h_t \mid \mathcal{F}_t] = \nabla f(r_t)'(\Phi' D\Phi)^{-1} A(\Phi' D\Phi)^{-1}\nabla f(r_t), \qquad \forall\ t.$$

Since the matrix A is negative definite, the matrix $(\Phi' D\Phi)^{-1} A(\Phi' D\Phi)^{-1}$ is also negative definite, thereby implying that for some positive scalar c_1,

$$\nabla f(r_t)' E[h_t \mid \mathcal{F}_t] \le -c_1 \|\nabla f(r_t)\|^2, \qquad \forall\ t. \tag{4.12}$$

We next consider $\big\|E[e_t \mid \mathcal{F}_t]\big\|$. By writing

$$e_t = \Theta_t r_t + \theta_t, \qquad \forall\ t,$$

with

$$\Theta_t = B_t^{-1} A_t - (\Phi' D\Phi)^{-1} A, \qquad \theta_t = B_t^{-1} b_t - (\Phi' D\Phi)^{-1} b, \qquad \forall\ t,$$

we have

$$E[e_t \mid \mathcal{F}_t] = E[\Theta_t] r_t + E[\theta_t], \qquad \forall\ t.$$

We will derive an estimate for $\big\|E[e_t \mid \mathcal{F}_t]\big\|$, by estimating $\big\|E[\Theta_t]\big\|$ and $\big\|E[\theta_t]\big\|$.

We first consider Θ_t. For all t, the matrix Θ_t can be equivalently written as follows

$$\begin{aligned}\Theta_t &= \left(\frac{B_t}{t+1}\right)^{-1} \frac{A_t}{t+1} - (\Phi' D\Phi)^{-1} A \\ &= \left(\left(\frac{B_t}{t+1}\right)^{-1} - (\Phi' D\Phi)^{-1}\right) \frac{A_t}{t+1} + (\Phi' D\Phi)^{-1}\left(\frac{A_t}{t+1} - A\right).\end{aligned}$$

Therefore,

$$\big\|E[\Theta_t]\big\| \leq \left\| E\left[\left(\left(\frac{B_t}{t+1}\right)^{-1} - (\Phi' D\Phi)^{-1} \right) \frac{A_t}{t+1} \right] \right\| + \big\|(\Phi' D\Phi)^{-1}\big\| \left\| \frac{E[A_t]}{t+1} - A \right\|, \qquad \forall\ t. \tag{4.13}$$

By using Lemma 4.3(a), we see that the first term on the right hand side of the preceding relation is finite for all t. Thus, by using Jensen's inequality (see, for example, Ash [Ash72], p. 287), and Lemma 4.3 (a) and (b), we obtain for all t,

$$\begin{aligned} \left\| E\left[\left(\left(\frac{B_t}{t+1}\right)^{-1} - (\Phi' D\Phi)^{-1} \right) \frac{A_t}{t+1} \right] \right\| &\leq E\left[\left\| \left(\frac{B_t}{t+1}\right)^{-1} - (\Phi' D\Phi)^{-1} \right\| \left\| \frac{A_t}{t+1} \right\| \right] \\ &\leq \frac{C_4}{\sqrt{t+1}}\, C_2, \qquad \forall\ t. \end{aligned}$$

Furthermore, by Lemma 4.3 (d), we have

$$\left\| \frac{E[A_t]}{t+1} - A \right\| \leq \frac{C_5}{t+1}.$$

From the preceding two relations and Eq. (4.13), it follows that for some positive constant $\bar{c}_1$,

$$\big\|E[\Theta_t]\big\| \leq \frac{\bar{c}_1}{\sqrt{t+1}}, \qquad \forall\ t.$$

Similar to the preceding analysis, where A_t and A are replaced respectively by b_t and b, we can show that for some positive constant $\bar{c}_2$,

$$\big\|E[\theta_t]\big\| \leq \frac{\bar{c}_2}{\sqrt{t+1}}, \qquad \forall\ t.$$

In view of the relation $E[e_t \mid \mathcal{F}_t] = E[\Theta_t] r_t + E[\theta_t]$, and the preceding estimates for $\big\|E[\Theta_t]\big\|$ and $\big\|E[\theta_t]\big\|$, it follows that

$$\big\|E[e_t \mid \mathcal{F}_t]\big\| \leq \frac{\bar{c}_1}{\sqrt{t+1}} \|r_t\| + \frac{\bar{c}_2}{\sqrt{t+1}}, \qquad \forall\ t.$$

Furthermore, we have

$$\|r_t\| \leq \|r_t - r^*\| + \|r^*\| \leq \big\|(\Phi' D\Phi)^{-1}\big\| \, \big\|\Phi' D\Phi(r_t - r^*)\big\| + \|r^*\| = \big\|(\Phi' D\Phi)^{-1}\big\| \, \|\nabla f(r_t)\| + \|r^*\|, \tag{4.14}$$

implying that for some positive constant c_2,

$$\big\|E[e_t \mid \mathcal{F}_t]\big\| \leq \frac{c_2}{\sqrt{t+1}} \big(1 + \|\nabla f(r_t)\|\big), \qquad \forall\ t. \tag{4.15}$$

Finally, we estimate $E\big[\|h_t + e_t\|^2 \mid \mathcal{F}_t\big]$. From the definitions of h_t and e_t, it follows that

$$h_t + e_t = B_t^{-1}(A_t r_t + b_t) = \left(\frac{B_t}{t+1}\right)^{-1} \left(\frac{A_t}{t+1} r_t + \frac{b_t}{t+1} \right), \qquad \forall\ t.$$

By using the boundedness of $\big(B_t/(t+1)\big)^{-1}$, $A_t/(t+1)$, and $b_t/(t+1)$ [cf. Lemma 4.3(a)], we have that $\|h_t + e_t\|$ is bounded by a constant multiple of $1 + \|r_t\|$ for all t. Then, by Eq. (4.14), it can be seen that for a positive constant c_3,

$$\|h_t + e_t\|^2 \leq c_3\big(1 + \|\nabla f(r_t)\|^2\big), \qquad \forall\ t,$$

implying that

$$E\big[\|h_t + e_t\|^2 \mid \mathcal{F}_t\big] = E\big[\|h_t + e_t\|^2 \mid r_t\big] \leq c_3\big(1 + \|\nabla f(r_t)\|^2\big), \qquad \forall\ t. \tag{4.16}$$

The relations (4.12), (4.15), and (4.16) show that assumption (ii) of Proposition 4.1 holds for the function f as given in Eq. (4.11), the sequence $\{r_t\}$, and the sequence $\{\varepsilon_t\}$ given by $\varepsilon_t = 1/\sqrt{t+1}$.

Furthermore, since $\sum_{t=0}^{\infty} \gamma_t = \infty$ and $\sum_{t=0}^{\infty} \gamma_t^2 < \infty$ (cf. Assumption 2.2), and since

$$\sum_{t=0}^{\infty} \gamma_t \varepsilon_t^2 \leq \frac{1}{2} \sum_{t=0}^{\infty} \big(\gamma_t^2 + \varepsilon_t^4\big) < \infty,$$

we see that γ_t and $\varepsilon_t = 1/\sqrt{t+1}$ satisfy assumption (iii) of Proposition 4.1. Hence, all assumptions of Proposition 4.1 are satisfied, and therefore $\nabla f(r_t) \to 0$ with probability 1, thus implying that $r_t \to r^*$ with probability 1. **Q.E.D.**

5. 5. CONVERGENCE PROOF FOR THE LSTD(λ) METHOD

In this section, we prove Proposition 3.1. As discussed in Section 3, we have

$$r_t = -\left(\frac{A_t}{t+1}\right)^{-1} \frac{b_t}{t+1}, \qquad \forall\ t.$$

The proof idea is to show, by using an appropriate law of large numbers, that the matrix $A_t/(t+1)$ and the vector $b_t/(t+1)$ converge with probability 1 to the matrix A and the vector b [cf. Eq. (2.10)], respectively. The following is a law of large numbers that is suitable for our purposes (see Parzen [Par62], Theorem 2B, p. 420).

Theorem 5.1: **(***Law of Large Numbers***)** Let $\{X_k\}$ be a sequence of jointly distributed random variables with zero mean and uniformly bounded variances, and let Z_t be given by

$$Z_t = \frac{1}{t+1} \sum_{k=0}^{t} X_k, \qquad \forall\ t.$$

If there exist positive scalars $\bar{C}$ and q such that

$$\big|E[X_t Z_t]\big| \leq \frac{\bar{C}}{(t+1)^q}, \qquad \forall\ t,$$

then Z_t converges to zero with probability 1.

We will apply this law (componentwise) to some appropriately chosen sequences of random matrices Y_k and vectors x_k. In particular, we define

$$Y_k = \sum_{m=0}^{k} (\alpha\lambda)^{k-m} U_{mk}, \qquad \forall\ k, \tag{5.1}$$

$$U_{mk} = \phi(i_m)\Big(\alpha\phi(i_{k+1})' - \phi(i_k)' - E\big[\alpha\phi(i_{k+1})' - \phi(i_k)' \mid i_m\big]\Big), \qquad \forall\ m,k,\ m \leq k, \tag{5.2}$$

$$x_k = \sum_{m=0}^{k} (\alpha\lambda)^{k-m} w_{mk}, \qquad \forall\ k, \tag{5.3}$$

$$w_{mk} = \phi(i_m)\Big(g(i_k, i_{k+1}) - E\big[g(i_k, i_{k+1}) \mid i_m\big]\Big), \qquad \forall\ m,k,\ m \leq k. \tag{5.4}$$

The crucial property of the matrix sequence $\{Y_k\}$ and the vector sequence $\{x_k\}$ is that their corresponding averaged sums converge to 0, with probability 1, as shown in the following lemma.

Lemma 5.1: For the sequences $\{Y_k\}$ and $\{x_k\}$ defined by Eqs. (5.1) and (5.3), we have, with probability 1,

$$\lim_{t\to\infty} \frac{1}{t+1} \sum_{k=0}^{t} Y_k = 0, \qquad \lim_{t\to\infty} \frac{1}{t+1} \sum_{k=0}^{t} x_k = 0.$$

Proof: The idea of the proof is to apply the Law of Large Numbers to each component sequence of $\{Y_k\}$ and $\{x_k\}$. We first consider the sequence $\{Y_k\}$. In what follows, we view a matrix as a "big" vector (i.e., a vector obtained from the given matrix by placing its columns into a single big column). Accordingly, we will use the Frobenius matrix norm, i.e., $\|H\|_F = \sqrt{\sum_{\nu,\tau} h_{\nu\tau}^2}$ for a matrix H with components $h_{\nu\tau}$.

Because $E[U_{mk}] = 0$ for all m and k [cf. Eq. (5.2)], it follows by the definition of Y_k [cf. Eq. (5.1)] that

$$E[Y_k] = 0, \qquad \forall\ k. \tag{5.5}$$

Furthermore, since $\big\{\phi(i_k)\big\}$ is bounded, it follows that for some positive scalar $\tilde{C}$,

$$\|U_{mk}\|_F \leq \tilde{C}, \qquad \forall\ m,k,\ m \leq k, \tag{5.6}$$

implying by the definition of Y_k that

$$\|Y_k\|_F \leq \frac{\tilde{C}}{1-\alpha\lambda}, \qquad \forall\ k. \tag{5.7}$$

We next define

$$S_t = \frac{1}{t+1}\sum_{k=0}^{t} Y_k, \qquad \forall\ t,$$

for which we want to show that, for some positive scalar $\bar{C}$, there holds

$$\big\|E[Y_t S_t']\big\|_F \leq \frac{\bar{C}}{t+1}, \qquad \forall\ t.$$

By using the definition of S_t and boundedness of $\{Y_k\}$, we obtain for all $t \geq 2$,

$$\big\|E[Y_t S_t']\big\|_F \leq \frac{1}{t+1}\sum_{k=0}^{t}\big\|E[Y_t Y_k']\big\|_F \leq \frac{1}{t+1}\sum_{k=0}^{t-2}\big\|E[Y_t Y_k']\big\|_F + \frac{\tilde{C}_1}{t+1}, \tag{5.8}$$

where $\tilde{C}_1$ is some positive scalar. Furthermore, by the definition of Y_k [cf. Eq. (5.1)], we have for all $t \geq 2$ and $k \leq t-2$,

$$\begin{aligned} E[Y_t Y_k'] &= E\left[\sum_{s=0}^{t}(\alpha\lambda)^{t-s}U_{st}\sum_{m=0}^{k}(\alpha\lambda)^{k-m}U_{mk}'\right] \\ &= E\left[\sum_{s=0}^{k+1}(\alpha\lambda)^{t-s}U_{st}\sum_{m=0}^{k}(\alpha\lambda)^{k-m}U_{mk}'\right] + E\left[\sum_{s=k+2}^{t}(\alpha\lambda)^{t-s}U_{st}\sum_{m=0}^{k}(\alpha\lambda)^{k-m}U_{mk}'\right]. \end{aligned}$$

Since $\sum_{s=k+2}^{t}(\alpha\lambda)^{t-s}U_{st}$ is a function of $i_{k+2},\ldots,i_{t+1}$ and $\sum_{m=0}^{k}(\alpha\lambda)^{k-m}U_{mk}$ is a function of $i_0,\ldots,i_{k+1}$, it follows that

$$E\left[\sum_{s=k+2}^{t}(\alpha\lambda)^{t-s}U_{st}\sum_{m=0}^{k}(\alpha\lambda)^{k-m}U_{mk}'\right] = 0,$$

implying that for all $t \geq 2$ and $k \leq t-2$,

$$E[Y_t Y_k'] = \sum_{s=0}^{k+1}(\alpha\lambda)^{t-s}\sum_{m=0}^{k}(\alpha\lambda)^{k-m}E[U_{st}U_{mk}'].$$

Using the boundedness of $\{U_{mk}\}$ [cf. Eq. (5.6)], it can be seen that $\big\|E[U_{st}U_{mk}']\big\|_F$ is bounded by some positive scalar $\tilde{C}_2$, so that for all $t \geq 2$ and $k \leq t-2$,

$$\big\|E[Y_t Y_k']\big\|_F \leq \sum_{s=0}^{k+1}(\alpha\lambda)^{t-s}\sum_{m=0}^{k}(\alpha\lambda)^{k-m}\tilde{C}_2 \leq \frac{\tilde{C}_2}{1-\alpha\lambda}\sum_{s=0}^{k+1}(\alpha\lambda)^{t-s} \leq \frac{\tilde{C}_2}{(1-\alpha\lambda)^2}(\alpha\lambda)^{t-k-1}.$$

By substituting the preceding relation in Eq. (5.8), we obtain

$$\begin{aligned} \big\|E[Y_t S_t']\big\|_F &\leq \frac{1}{t+1}\,\frac{\tilde{C}_2}{(1-\alpha\lambda)^2}\sum_{k=0}^{t-2}(\alpha\lambda)^{t-k-1} + \frac{\tilde{C}_1}{t+1} \\ &\leq \frac{1}{t+1}\,\frac{\tilde{C}_2\alpha\lambda}{(1-\alpha\lambda)^3} + \frac{\tilde{C}_1}{t+1}, \qquad \forall\ t \geq 2, \end{aligned}$$

thus implying that for some positive scalar $\bar{C}$, we have

$$\big\|E[Y_t S_t']\big\|_F \le \frac{\bar{C}}{t+1}, \qquad \forall\ t. \tag{5.9}$$

We now fix any $\nu, \tau \in \{1, \ldots, K\}$, and we consider the scalar sequence $\big\{[Y_k]_{\nu\tau}\big\}$, where $[Y]_{\nu\tau}$ denotes the $\nu\tau$th component of a matrix Y. Since $E[Y_k] = 0$ for all k [cf. Eq. (5.5)], it follows that $\big\{[Y_k]_{\nu\tau}\big\}$ is a zero mean scalar sequence. Furthermore, in view of the boundedness of $\{Y_k\}$ [cf. Eq. (5.7)], we have that $\big\{[Y_k]_{\nu\tau}\big\}$ is bounded, implying that $\big\{[Y_k]_{\nu\tau}\big\}$ is a sequence of random variables with uniformly bounded variances. Finally, by our convention of viewing a matrix as a big column vector, the relation (5.9) is equivalent to

$$\sqrt{\sum_{\kappa,\rho}\sum_{s,l}\Big(E\big[[Y_t]_{\kappa\rho}[S_t]_{sl}\big]\Big)^2} \le \frac{\bar{C}}{t+1}, \qquad \forall\ t,$$

which implies that

$$\Big|E\big[[Y_t]_{\nu\tau}[S_t]_{\nu\tau}\big]\Big| \le \frac{\bar{C}}{t+1}, \qquad \forall\ t.$$

Thus, by the Law of Large Numbers (cf. Theorem 5.1), we have $[S_t]_{\nu\tau} \to 0$ with probability 1. Since $\nu, \tau \in \{1, \ldots, K\}$ are arbitrary, it follows that $S_t \to 0$ with probability 1, thus showing that with probability 1,

$$\lim_{t\to\infty} \frac{1}{t+1}\sum_{k=0}^{t} Y_k = 0.$$

A nearly identical proof, with x_k in place of Y_k, shows that with probability 1,

$$\lim_{t\to\infty} \frac{1}{t+1}\sum_{k=0}^{t} x_k = 0.$$

Q.E.D.

We now prove Proposition 3.1.

Proof of Proposition 3.1: We will show that $A_t/(t+1) \to A$ and $b_t/(t+1) \to b$ with probability 1. From Eq. (2.8), we have

$$A_t = \sum_{k=0}^{t}\sum_{m=0}^{k} (\alpha\lambda)^{k-m}\phi(i_m)\big(\alpha\phi(i_{k+1})' - \phi(i_k)'\big), \qquad \forall\ t.$$

In view of the definitions of Y_k and U_{mk} [cf. Eqs. (5.1) and (5.2)], it follows that

$$A_t = \sum_{k=0}^{t} Y_k + \sum_{k=0}^{t}\sum_{m=0}^{k} (\alpha\lambda)^{k-m}\phi(i_m)E\big[\alpha\phi(i_{k+1})' - \phi(i_k)' \mid i_m\big].$$

By Lemma 5.1, we have that $\big(\sum_{k=0}^t Y_k\big)/(t+1) \to 0$ with probability 1, implying that

$$\lim_{t\to\infty} \frac{A_t}{t+1} = \lim_{t\to\infty} \frac{1}{t+1} \sum_{k=0}^{t} \sum_{m=0}^{k} (\alpha\lambda)^{k-m} \phi(i_m) E\big[\alpha\phi(i_{k+1})' - \phi(i_k)' \mid i_m\big], \tag{5.10}$$

with probability 1, provided that the limit on the right-hand side of this relation exists, which we show next by actually computing that limit.

By Lemma 4.3(c), for all t, we have

$$\sum_{k=0}^{t} \sum_{m=0}^{k} (\alpha\lambda)^{k-m} \phi(i_m) E\big[\alpha\phi(i_{k+1})' - \phi(i_k)' \mid i_m\big] = \sum_{i=1}^{n} \kappa_t(i)\phi(i) \sum_{s=0}^{\infty} \big[(\alpha P - I)(\alpha\lambda P)^s \Phi\big]_i - V_t, \tag{5.11}$$

with

$$V_t = \sum_{m=0}^{t} \phi(i_m) \sum_{k=t+1}^{\infty} \big[(\alpha P - I)(\alpha\lambda P)^{k-m}\Phi\big]_{i_m}, \qquad \forall\ t.$$

By Lemma 4.2(a), we have that $\kappa_t(i)/(t+1) \to \pi(i)$ for all i, with probability 1. Moreover, from the definition of V_t, we can see that V_t is bounded, thus implying that $V_t/(t+1) \to 0$ with probability 1. From this, together with Eqs. (5.10) and (5.11), it follows that with probability 1,

$$\lim_{t\to\infty} \frac{A_t}{t+1} = \sum_{i=1}^{n} \pi(i)\phi(i) \sum_{s=0}^{\infty} \big[(\alpha P - I)(\alpha\lambda P)^s \Phi\big]_i = \Phi' D \sum_{s=0}^{\infty} (\alpha P - I)(\alpha\lambda P)^s \Phi = A.$$

Similarly, from Eq. (2.9) we have

$$b_t = \sum_{k=0}^{t} \sum_{m=0}^{k} (\alpha\lambda)^{k-m} \phi(i_m) g(i_k, i_{k+1}), \qquad \forall\ t.$$

Using the definitions of x_k and w_{mk} [cf. Eqs. (5.3) and (5.4)], we can write

$$b_t = \sum_{k=0}^{t} x_k + \sum_{k=0}^{t} \sum_{m=0}^{k} (\alpha\lambda)^{k-m} \phi(i_m) E\big[g(i_k, i_{k+1}) \mid i_m\big].$$

By Lemma 5.1, $\big(\sum_{k=0}^t x_k\big)/(t+1) \to 0$ with probability 1, implying that with probability 1,

$$\lim_{t\to\infty} \frac{b_t}{t+1} = \lim_{t\to\infty} \frac{1}{t+1} \sum_{k=0}^{t} \sum_{m=0}^{k} (\alpha\lambda)^{k-m} \phi(i_m) E\big[g(i_k, i_{k+1}) \mid i_m\big]. \tag{5.12}$$

By Lemma 4.3(c), we have

$$\sum_{k=0}^{t} \sum_{m=0}^{k} (\alpha\lambda)^{k-m} \phi(i_m) E\big[g(i_k, i_{k+1}) \mid i_m\big] = \sum_{i=1}^{n} \kappa_t(i)\phi(i) \sum_{s=0}^{\infty} \big[(\alpha\lambda P)^s \bar{g}\big](i) - v_t, \qquad \forall\ t, \tag{5.13}$$

where

$$v_t = \sum_{m=0}^{t} \phi(i_m) \sum_{k=t+1}^{\infty} \big[(\alpha\lambda P)^{k-m} \bar{g}\big](i_m), \qquad \forall\ t.$$

Furthermore, by Lemma 4.2(a), $\kappa_t(i)/(t+1) \to \pi(i)$ for all i, with probability 1. Using the definition of v_t, we can see that v_t is bounded, so that $v_t/(t+1) \to 0$ with probability 1. From this and Eqs. (5.12) and (5.13), we obtain

$$\lim_{t\to\infty} \frac{b_t}{t+1} = \sum_{i=1}^{n} \pi(i)\phi(i) \sum_{s=0}^{\infty} \big[(\alpha\lambda P)^s \bar{g}\big](i) = \Phi' D \sum_{s=0}^{\infty} (\alpha\lambda P)^s \bar{g} = b,$$

with probability 1. **Q.E.D.**

If we use the initial matrix

$$A_0 = \delta I + \phi(i_0)\big(\alpha\phi(i_1) + \phi(i_0)\big),$$

for some positive scalar δ, then in the preceding analysis A_t is replaced by $\delta I + A_t$, so clearly we have $(\delta I + A_t)/(t+1) \to A$. Thus, the method converges to r^* in the case where $\delta > 0$ as well.

6. REFERENCE

[Ash72] Ash, R. B., Real Analysis and Probability, Academic Press Inc., New York, 1972.

[Ber95] Bertsekas, D. P., “A Counterexample to Temporal Differences Learning,” Neural Computation, Vol. 7, 1995, pp. 270–279.

[BeI96] Bertsekas, D. P., and Ioffe, S., “Temporal Differences-Based Policy Iteration and Applications in Neuro-Dynamic Programming,” Lab. for Info. and Decision Systems Report LIDS-P-2349, MIT, Cambridge, MA, 1996.

[Ber99] Bertsekas, D. P., Nonlinear Programming, 2nd edition, Athena Scientific, Belmont, MA, 1999.

[Ber01] Bertsekas, D. P., Dynamic Programming and Optimal Control, 2nd edition, Athena Scientific, Belmont, MA, 2001.

[BeT96] Bertsekas, D. P., and Tsitsiklis, J. N., Neuro-Dynamic Programming, Athena Scientific, Belmont, MA, 1996.

[BeT00] Bertsekas, D. P., and Tsitsiklis, J. N., “Gradient Convergence in Gradient Methods with Errors,” SIAM J. Optim., 10, 2000, pp. 627–642.

[Boy02] Boyan, J. A., “Technical Update: Least-Squares Temporal Difference Learning,” to appear in Machine Learning, 49, 2002.

[BrB96] Bradtke, S. J., and Barto, A. G., "Linear Least-Squares Algorithms for Temporal Difference Learning," Machine Learning, 22, 1996, pp. 33–57.

[DaS94] Dayan, P., and Sejnowski, T. J., "TD(λ) Converges with Probability 1," Machine Learning, 14, 1994, pp. 295–301.

[Gal95] Gallager, R. G., Discrete Stochastic Processes, Kluwer Academic Publishers, Boston, MA, 1995.

[GLH94] Gurvits, L., Lin, L., and Hanson, S. J., "Incremental Learning of Evaluation Functions for Absorbing Markov Chains: New Methods and Theorems," Working Paper, Siemens Corporate Research, Princeton, NJ, 1994.

[GoV96] Golub, G. H., and Van Loan, C. F., Matrix Computations, 3rd edition, Johns Hopkins University Press, Baltimore, MD, 1996.

[JJS94] Jaakkola, T., Jordan, M. I., and Singh S. P., "On the Convergence of Stochastic Iterative Dynamic Programming Algorithms," Neural Computation, 6, 1994, pp. 1185–1201.

[KeS67] Kemeny, J. G., and Snell, J. L., Finite Markov Chains, Van Nostrand Company, New York, 1967.

[Nev75] Neveu, J., Discrete Parameter Martingales, North-Holland, Amsterdam, 1975.

[Par62] Parzen, E., Modern Probability Theory and Its Applications, John Wiley Inc., New York, 1962.

[Put94] Puterman, M. L., Markovian Decision Problems, John Wiley Inc., New York, 1994.

[Sut88] Sutton, R. S., "Learning to Predict by the Methods of Temporal Differences," Machine Learning, 3, 1988, pp. 9–44.

[Tad01] Tadić, V., "On the Convergence of Temporal-Difference Learning with Linear Function Approximation," Machine Learning, 42, 2001, pp. 241–267.

[TsV97] Tsitsiklis, J. N., and Van Roy, B., "An Analysis of Temporal-Difference Learning with Function Approximation," IEEE Transactions on Automatic Control, 42, 1997, pp. 674–690.