LIDS REPORT 2697
Jun 2006, Revised Aug 2008
To appear in IEEE Transactions on Aut. Control

Convergence Results for Some Temporal Difference Methods Based on Least Squares

Huizhen Yu*
janey.yu@cs.helsinki.fi

Dimitri P. Bertsekas†
dimitrib@mit.edu

Abstract

We consider finite-state Markov decision processes, and prove convergence and rate of convergence results for certain least squares policy evaluation algorithms of the type known as LSPE(λ). These are temporal difference methods for constructing a linear function approximation of the cost function of a stationary policy, within the context of infinite-horizon discounted and average cost dynamic programming. We introduce an average cost method, patterned after the known discounted cost method, and we prove its convergence for a range of constant stepsize choices. We also show that the convergence rate of both the discounted and the average cost methods is optimal within the class of temporal difference methods. Analysis and experiment indicate that our methods are substantially and often dramatically faster than TD(λ), as well as more reliable.

*Huizhen Yu was with the Laboratory for Information and Decision Systems (LIDS), M.I.T., and is presently with the Department of Computer Science and HIIT, Univ. of Helsinki, Finland.

†Dimitri Bertsekas is with the Laboratory for Information and Decision Systems (LIDS), M.I.T., Cambridge, MA 02139.

1 Introduction

We consider finite-state Markov decision processes (MDP) with the discounted and the average cost criteria. We focus on a single stationary policy, and discuss the approximate evaluation of the corresponding cost function (in the discounted case) or bias/differential cost function (in the average cost case). Such evaluation methods are essential for approximate policy iteration, including gradient-descent type of algorithms (e.g., actor-critic algorithms [1]) when parametrized policies are considered. A prominent algorithm for approximating this cost function using a linear combination of basis functions is TD(λ). This is an iterative temporal differences (TD) method, which uses a single infinitely long sample trajectory, and depends on a scalar parameter $\lambda \in [0, 1]$ that controls a tradeoff between accuracy of the approximation and susceptibility to simulation noise. The method was originally proposed for discounted problems by Sutton [2], and analyzed by several authors, including Dayan [3], Gurvits, Lin, and Hanson [4], Pineda [5], Tsitsiklis and Van Roy [6]. An extension to average cost problems and $\lambda \in [0, 1)$ was proposed and analyzed by Tsitsiklis and Van Roy [7, 8] (the case $\lambda = 1$ may lead to divergence and was excluded; it needs a different treatment as given by Marbach and Tsitsiklis [9]).

Alternatively, there are two least squares-based algorithms, which employ the same approximation framework as TD(λ), but use simulation more efficiently. In particular, let us denote by $J = TJ$ a (linear, multiple-step) Bellman equation involving a single policy, and let Π denote projection on a subspace of basis functions with respect to a suitable Euclidean projection norm. Then TD(λ) aims to solve the *projected Bellman equation* $J = \Pi TJ$, with a stochastic approximation (SA) type of iteration. The two least squares-based algorithms solve the same linear equation, but they use simulation to construct directly the low-dimensional quantities defining the equation, instead of only the solution itself, unlike TD(λ). The two algorithms are called the least squares temporal difference algorithm, LSTD(λ), first proposed by Bradtke and Barto [10] for $\lambda = 0$ and generalized by Boyan [11] to $\lambda \in (0, 1]$, and the least squares policy evaluation algorithm, LSPE(λ), first proposed for stochastic shortest path problems by Bertsekas and Ioffe [12]. Roughly speaking, LSPE(λ) differs from LSTD(λ) in that LSPE(λ) can be viewed as a *simulation-based approximation of the value iteration* algorithm, and is essentially a Jacobi method, while LSTD(λ) solves directly at each iteration an approximation of the equation. The differences between LSPE(λ) and LSTD(λ) become more pronounced in the important application context where they are embedded within a policy iteration scheme, as explained in Section 6. Both LSPE(λ) and LSTD(λ) have superior performance to standard TD(λ), as suggested not only by practice but also by theory: it has been shown by Konda [13] that LSTD(λ) has optimal convergence rate, compared to other TD(λ) algorithms, and it will be shown in this paper that LSPE(λ) has the same property. Both algorithms have been applied to approximate policy iteration. In fact, in the original paper [12] (see also the book by Bertsekas and Tsitsiklis [14]), LSPE(λ) was called "λ-policy iteration" and applied in the framework of optimistic policy iteration, a version of the simulation-based approximate policy iteration, to solve the computer game Tetris, which involves a very large state space of approximately 2^{200} states. LSTD(λ) was applied with approximate policy iteration by Lagoudakis and Parr [15]. Both works reported favorable computational results which were not possible by using TD(λ).

In this paper we will focus on the LSPE(λ) algorithm, analyzing its convergence for the average cost case (Section 3), and analyzing its rate of convergence for both the discounted and average cost cases (Section 4). The convergence of LSPE(λ) under the discounted criterion has been analyzed in previous works. In particular, LSPE(λ) uses a parameter $\lambda \in [0, 1]$, similar to other TD methods, and a positive stepsize. For discounted problems, Nedić and Bertsekas [16] proved the convergence of LSPE(λ) with a diminishing stepsize, while Bertsekas, Borkar, and Nedić [17], improving on the analysis of [16], proved the convergence of LSPE(λ) for a range of constant stepsizes including the unit stepsize. Both analysis and experiment have indicated that LSPE(λ) with a constant stepsize has better performance than standard TD(λ) as well as LSPE(λ) with a diminishing stepsize. In this paper, we will focus on the constant stepsize version. There has been no rigorous analysis of

LSPE(λ) in the context of the average cost problem, despite applications of LSPE(λ) with policy gradient in this context [18], and one of the purposes of this paper is to provide such an analysis.

The average cost case requires a somewhat more general treatment than the proof given in [17] for the discounted case. LSPE(λ) is a simulation-based fixed point iteration, the convergence of which relies on the underlying mapping being a contraction. The projected Bellman equation in the average cost case involves sometimes nonexpansive mappings (unlike the discounted case where it involves contraction mappings with known modulus determined in part by the discount factor). Two means for inducing or ensuring the contraction property required by LSPE(λ) are (i) the choice of basis functions and (ii) a constant stepsize. The former, (i), is reflected by a condition given by Tsitsiklis and Van Roy [7] on the basis functions of the average cost TD(λ) algorithm, which is required to ensure that the projected Bellman equation has a unique solution and also induces contraction for the case of $\lambda > 0$, and the case of $\lambda = 0$ and an aperiodic Markov chain, as illustrated in Prop. 2 in Section 3. The latter, (ii), is closely connected to the damping mechanism for turning nonexpansive mappings into contraction mappings (this is to be differentiated from the role of a constant and diminishing stepsizes used in SA algorithms, which is to track a varying system without ensuring convergence of the iterates, in the case of a constant stepsize, and to enforce convergence through averaging the noise, in the case of a diminishing stepsize). Our convergence analysis of a constant stepsize LSPE(λ) will involve both (i) and (ii), and arguments that are technically different and more general than those of [17]. Our analysis also covers the convergence results of [17] for the discounted case, and simplifies proofs in the latter work.

For convergence rate analysis, we will show that in both the discounted and average cost cases, LSPE(λ) with any constant stepsize under which it converges has the same convergence rate as LSTD(λ). In fact, we will show that LSPE(λ) and LSTD(λ) converge to each other at a faster rate than they converge to the common limit. This was conjectured, but not proved, by [17] in the discounted case. Since Konda [13] has shown that LSTD(λ) has optimal asymptotic convergence rate, as mentioned earlier, LSPE(λ) with a constant stepsize shares this optimality property.

Let us mention that the part of the iterations in LSTD(λ) and LSPE(λ) that approximates low-dimensional quantities defining the projected Bellman equation/fixed point mapping can be viewed as a simple SA algorithm, whose convergence under a fixed policy is ensured by the law of large numbers for samples from a certain Markov chain. This connection provides the basis for designing two-time-scale algorithms using LSTD(λ) and LSPE(λ) when the policy is changing. We will highlight this in the context of approximate policy iteration with actor-critic type of policy gradient methods, which are two-time-scale SA algorithms, when we discuss the use of LSTD(λ) and LSPE(λ) as a critic (Section 6).

The paper is organized as follows. In Section 2, after some background on TD with function approximation, we introduce the LSPE(λ) method, we motivate the convergence analysis of Section 3, and we also provide a qualitative comparison to LSTD(λ). In Section 3, we provide convergence results for LSPE(λ) by using a spectral radius analysis. We also introduce a contraction theorem for nonexpansive fixed point iterations involving Euclidean projections, we use this theorem to analyze the contraction properties of the mapping associated with the average cost TD(λ), and to interpret all of our convergence results for $\lambda > 0$, but only some of our results for $\lambda = 0$. In Section 4, we discuss the convergence rate of LSPE(λ) for both the discounted and the average cost cases, and we show that it is identical to that of LSTD(λ). In Section 5, we provide some computational results that are in agreement with the analytical conclusions, and indicate a substantial and often dramatic speed of convergence advantage over TD(λ), even when the latter is enhanced with Polyak-type averaging. Finally, in Section 6, we discuss various extensions, as well as application of the algorithms in the context of approximate policy iteration.

2 Preliminaries: the Average Cost LSPE(λ) and LSTD(λ) Algorithms

We focus on a time-homogeneous finite-state Markov chain whose states are denoted by $1, \ldots, n$. Let P be the state transition probability matrix with entries $P_{ij} = p(X_1 = j | X_0 = i)$, where the random variable X_t is the state at time t. Throughout the paper we operate under the following recurrence assumption (in the last section we discuss the case where this assumption is violated).

Assumption 1. *The states of the Markov chain form a single recurrent class.*

Under the above assumption, the Markov chain has a unique invariant distribution $\pi = [\pi(1), \ldots, \pi(n)]$, which is the unique probability distribution satisfying the system of equations $\pi P = \pi$. We allow the possibility that the chain may be aperiodic or periodic, in which case, with slight abuse of terminology, we say that P is aperiodic or periodic, respectively.

Let $g(i, j)$ be the cost of transition from state i to state j, and let g be the length-n column vector with components the expected state costs $\sum_{j=1}^{n} P_{ij} g(i, j)$, $i = 1, \ldots, n$. It is well known that the average cost starting at state i,

$$\lim_{t \to \infty} \frac{1}{t+1} \sum_{k=0}^{t} E\big[g(X_k, X_{k+1}) \mid X_0 = i\big],$$

is a constant η^* independent of the initial state i, and

$$\eta^* = \pi g.$$

The differential cost function, or bias function, that we aim to approximate, is defined by

$$h(i) = \lim_{t \to \infty} \sum_{k=0}^{t} E\big[g(X_k, X_{k+1}) - \eta^* \mid X_0 = i\big], \quad i = 1, \ldots n,$$

when the Markov chain is aperiodic, and is defined by the Cesaro limit when the Markov chain is periodic: for $i = 1, \ldots n$,

$$h(i) = \lim_{t \to \infty} \frac{1}{t+1} \sum_{m=0}^{t} \sum_{k=0}^{m} E\big[g(X_k, X_{k+1}) - \eta^* \mid X_0 = i\big].$$

It satisfies the average cost dynamic programming equation, which in matrix notation is

$$h = g - \eta^* e + Ph, \tag{1}$$

where e is the length-n column vector of all 1s, and h is treated as a length-n column vector. Under the recurrence Assumption 1, the function h is the unique solution of this equation up to addition of a scalar multiple of e.

2.1 Background of the TD/Function Approximation Approach

In LSPE(λ) and LSTD(λ), like in recursive TD(λ), we use an $n \times s$ matrix Φ to approximate the bias function h with a vector of the form Φr,

$$h \approx \Phi r.$$

In particular, for each state i, we introduce the vector

$$\phi(i)' = \big[\phi_1(i), \ldots, \phi_s(i)\big]$$

which forms the ith row of the matrix Φ. We view these rows as describing attributes or features of the corresponding state i, and we view the columns of Φ as basis functions. We denote by $S \subseteq \Re^n$ the subspace spanned by the basis vectors,

$$S = \{\Phi r \mid r \in \Re^s\}.$$

We adopt throughout our paper for the average cost case the following assumption from [7], which differs from the discounted counterpart in that $e \notin S$.

Assumption 2. *The columns of the matrix* $[\,\Phi \;\; e\,]$ *are linearly independent.*

For every $\lambda \in [0,1)$, all algorithms, LSPE(λ) (as will be shown), LSTD(λ), and TD(λ), compute the same vector r and hence the same approximation of h on the subspace S. This approximation, denoted by Φr^*, is the solution of a fixed point equation parametrized by λ,

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

Here Π is a projection mapping on S, and $T^{(\lambda)}$ is a mapping that has h as a fixed point (unique up to a constant shift); the details of the two mappings will be given below. Both mappings play a central role in the analysis of Tsitsiklis and Van Roy [7] of the TD(λ) algorithm, as well as in our subsequent analysis of LSPE(λ).

We define the mapping $T : \Re^n \mapsto \Re^n$ by

$$TJ = g - \eta^* e + PJ,$$

and view the Bellman equation (1) as the fixed point equation $h = Th$. We consider the multiple-step fixed point equations $h = T^m h, m \geq 1$, and combine them with geometrically decreasing weights that depend on the parameter $\lambda \in [0,1)$, thereby obtaining the fixed point equation:

$$h = T^{(\lambda)} h, \qquad \lambda \in [0,1), \tag{2}$$

where

$$T^{(\lambda)} = (1-\lambda) \sum_{m=0}^{\infty} \lambda^m T^{m+1}. \tag{3}$$

In matrix notation, the mapping $T^{(\lambda)}$ can be written as

$$T^{(\lambda)} J = (1-\lambda) \sum_{m=0}^{\infty} \lambda^m P^{m+1} J + \sum_{m=0}^{\infty} \lambda^m P^m (g - \eta^* e),$$

or more compactly as

$$T^{(\lambda)} J = P^{(\lambda)} J + (I - \lambda P)^{-1} (g - \eta^* e), \tag{4}$$

where the matrix $P^{(\lambda)}$ is defined by

$$P^{(\lambda)} = (1-\lambda) \sum_{m=0}^{\infty} \lambda^m P^{m+1}. \tag{5}$$

Note that $T^{(0)} = T$ and $P^{(0)} = P$ for $\lambda = 0$. When function approximation is used, a positive λ improves approximation accuracy, in the sense that will be explained later.

The projection norm with respect to which Π, the operation of projection on S is defined, is the weighted Euclidean norm specified by the invariant distribution vector π. This choice of norm is important for convergence purposes. (There are other possible choices of norm, which may be important in the context of policy iteration and the issue of exploration [14, 19], but this subject is

beyond the scope of the present paper.) In particular, we denote by $\|\cdot\|_\pi$ the weighted Euclidean norm on $\Re^n$,

$$\|z\|_\pi = \left(\sum_{i=1}^n \pi(i) z_i^2\right)^{1/2}, \qquad \forall\ z = (z_1, \ldots, z_n) \in \Re^n,$$

and define

$$\Pi J = \Phi \hat{r}_J,$$

where

$$\hat{r}_J = \arg\min_{r \in \Re^s} \|\Phi r - J\|_\pi = \arg\min_{r \in \Re^s} \sum_{i=1}^n \pi(i)\big(\phi(i)'r - J(i)\big)^2.$$

In matrix notation, with D being the diagonal matrix $D = diag\big(\pi(1), \ldots, \pi(n)\big)$,

$$\Pi = \Phi(\Phi' D \Phi)^{-1} \Phi' D. \tag{6}$$

By Tsitsiklis and Van Roy [6, Lemma 1],

$$\|P^{(\lambda)}\|_\pi \le 1, \quad \|\Pi P^{(\lambda)}\|_\pi \le 1, \qquad \forall\ \lambda \in [0,1), \tag{7}$$

so $\Pi T^{(\lambda)}, \lambda \in [0,1)$ are nonexpansive mappings with respect to $\|\cdot\|_\pi$; their contraction properties will be discussed later in Section 3.2.

Tsitsiklis and Van Roy [7] show that there is a unique solution Φr^* of the fixed point equation:

$$\Phi r^* = \Pi T^{(\lambda)}(\Phi r^*), \tag{8}$$

to which recursive TD(λ) algorithms converge in the limit. Tsitsiklis and Van Roy [7] also provide an estimate of the error between Πh, the projection of the true bias function, and Φr^*, modulo a constant shift, which indicates that the error diminishes as λ approaches 1. Their analysis was given under Assumptions 1, 2, and the additional assumption that P is aperiodic, but extends to the periodic case as well. Their error analysis supports the use of Φr^* as approximation of h in approximate value iteration or in actor-critic algorithms. (Sharper and more general error bounds for projected equations have been recently derived in our paper [20].)

It will be useful for our purposes to express $\Pi T^{(\lambda)}$ and the solution r^* explicitly in terms of matrices and vectors of dimension s, and to identify fixed point iterations on the subspace S with corresponding iterations on the space of r. Define

$$B = \Phi' D \Phi, \qquad A = \Phi' D (P^{(\lambda)} - I)\Phi, \tag{9}$$

$$b = \Phi' D \sum_{m=0}^{\infty} \lambda^m P^m (g - \eta^* e), \tag{10}$$

where the matrix $P^{(\lambda)}$ is defined by (5), and the vector b can also be written more compactly as

$$b = \Phi' D (I - \lambda P)^{-1}(g - \eta^* e). \tag{11}$$

Using the definitions of Π [cf. (6)] and $T^{(\lambda)}$ [cf. (4)], it is easy to verify that

$$\Pi T^{(\lambda)}(\Phi r) = \Phi B^{-1}(Ar + b) + \Phi r, \tag{12}$$

with the linear term corresponding to

$$\Pi P^{(\lambda)}(\Phi r) = \Phi(I + B^{-1}A)r, \tag{13}$$

and, by the linear independence of columns of Φ,

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

It follows from (12) that the fixed point iteration

$$J_1 = \Pi T^{(\lambda)} J_0$$

on S is identical to the following iteration on $\Re^s$ with $\Phi r_i = J_i, i = 0, 1$:

$$r_1 = r_0 + B^{-1}(Ar_0 + b), \tag{14}$$

and similarly, the damped iteration

$$J_1 = \big((1-\gamma)I + \gamma \Pi T^{(\lambda)}\big) J_0$$

on S is identical to

$$r_1 = r_0 + \gamma B^{-1}(Ar_0 + b). \tag{15}$$

These relations will be used later in our analysis to relate the LSPE(λ) updates on the space of r to the more intuitive approximate value iterations on the subspace S.

2.2 The LSPE(λ) Algorithm

We now introduce the LSPE(λ) algorithm for average cost problems. Let $(x_0, x_1, \ldots)$ be an infinitely long sample trajectory of the Markov chain associated with P, where x_t is the state at time t. Let η_t be the following estimate of the average cost at time t:

$$\eta_t = \frac{1}{t+1} \sum_{k=0}^{t} g(x_k, x_{k+1}),$$

which converges to the average cost η^* with probability 1. We define our algorithm in terms of the solution of a linear least squares problem and the temporal differences

$$d_t(m) = g(x_m, x_{m+1}) - \eta_m + \phi(x_{m+1})' r_t - \phi(x_m)' r_t.$$

In particular, we define $\tilde{r}_t$ by

$$\tilde{r}_t = \arg\min_{r \in \Re^s} \sum_{k=0}^{t} \left(\phi(x_k)' r - \phi(x_k)' r_t - \sum_{m=k}^{t} \lambda^{m-k} d_t(m) \right)^2. \tag{16}$$

The new vector r_{t+1} of LSPE(λ) is obtained by interpolating from the current iterate with a constant stepsize γ:

$$r_{t+1} = r_t + \gamma(\tilde{r}_t - r_t). \tag{17}$$

It is straightforward to verify that the least squares solution is

$$\tilde{r}_t = r_t + \bar{B}_t^{-1}(\bar{A}_t r_t + \bar{b}_t),$$

where

$$\bar{B}_t = \frac{B_t}{t+1}, \qquad \bar{A}_t = \frac{A_t}{t+1}, \qquad \bar{b}_t = \frac{b_t}{t+1},$$

and the matrices B_t, A_t and vector b_t are defined by[1]

$$B_t = \sum_{k=0}^{t} \phi(x_k)\,\phi(x_k)', \qquad A_t = \sum_{k=0}^{t} z_k\big(\phi(x_{k+1})' - \phi(x_k)'\big),$$

$$b_t = \sum_{k=0}^{t} z_k\,\big(g(x_k, x_{k+1}) - \eta_k\big), \qquad z_k = \sum_{m=0}^{k} \lambda^{k-m}\phi(x_m).$$

These matrices and vectors can be computed recursively:

$$\bar{B}_t = \tfrac{t}{t+1}\bar{B}_{t-1} + \tfrac{1}{t+1}\phi(x_t)\,\phi(x_t)', \tag{18}$$

$$\bar{A}_t = \tfrac{t}{t+1}\bar{A}_{t-1} + \tfrac{1}{t+1}z_t\big(\phi(x_{t+1})' - \phi(x_t)'\big), \tag{19}$$

$$\bar{b}_t = \tfrac{t}{t+1}\bar{b}_{t-1} + \tfrac{1}{t+1}z_t\,\big(g(x_t, x_{t+1}) - \eta_t\big), \tag{20}$$

$$z_t = \lambda z_{t-1} + \phi(x_t). \tag{21}$$

The matrices $\bar{B}_t, \bar{A}_t$, and vector $\bar{b}_t$ are convergent. Using the analysis of Tsitsiklis and Van Roy [7, Lemma 4] on average cost TD(λ) algorithms, and Nedić and Bertsekas [16] on discounted LSPE(λ) algorithms, it can be easily shown that with probability 1

$$\bar{B}_t \to B, \qquad \bar{A}_t \to A, \qquad \bar{b}_t \to b,$$

as $t \to \infty$, where A, B, and b are given by (9)-(10).

Our average cost LSPE(λ) algorithm (17) thus uses a constant stepsize γ and updates the vector r_t by

$$r_{t+1} = r_t + \gamma\bar{B}_t^{-1}\left(\bar{A}_t\,r_t + \bar{b}_t\right). \tag{22}$$

In the case where $\gamma = 1$, r_{t+1} is simply the least squares solution of (16). In Section 3 we will derive the range of stepsize γ that guarantees the convergence of LSPE(λ) for various values of λ. For this analysis, as well as for a high-level interpretation of the LSPE(λ) algorithm, we need the preliminaries given in the next subsection.

2.3 LSPE(λ) as Simulation-Based Fixed Point Iteration

We write the LSPE(λ) iteration (22) as a deterministic iteration plus stochastic noise:

$$r_{t+1} = r_t + \gamma B^{-1}(Ar_t + b) + \gamma\big(Z_t r_t + \zeta_t\big), \tag{23}$$

where Z_t and ζ_t are defined by

$$Z_t = \bar{B}_t^{-1}\bar{A}_t - B^{-1}A, \qquad \zeta_t = \bar{B}_t^{-1}\bar{b}_t - B^{-1}b,$$

and they converge to zero with probability 1. Similar to its discounted case counterpart in [17], the convergence analysis of iteration (23) can be reduced to that of its deterministic portion under a spectral radius condition. In particular, (23) is equivalent to

$$r_{t+1} - r^* = \big(I + \gamma B^{-1}A + \gamma Z_t\big)\big(r_t - r^*\big) + \gamma\big(Z_t r^* + \zeta_t\big). \tag{24}$$

When $Z_t \to 0$ and $\zeta_t \to 0$, the stochastic noise term $\gamma(Z_t r^* + \zeta_t)$ diminishes to 0, and the iteration matrix $(I + \gamma B^{-1}A + \gamma Z_t)$ converges to the matrix $(I + \gamma B^{-1}A)$. Thus, convergence hinges on the condition

$$\sigma(I + \gamma B^{-1}A) < 1, \tag{25}$$

where for any square matrix F, $\sigma(F)$ denotes the spectral radius of F (i.e., the maximum of the moduli of the eigenvalues of F). This is shown in the following proposition.

[1] A theoretically slightly better version of the algorithm is to replace the term η_k in b_t by η_t; the resulting updates can be computed recursively as before. The subsequent convergence analysis is not affected by this modification, or any modification in which $\eta_t \to \eta^*$ with probability 1.

Proposition 1. *Assume that Assumptions 1 and 2 and the spectral radius condition (25) hold. Then the average cost LSPE(λ) iteration (22) converges to $r^* = -A^{-1}b$ with probability 1 as $t \to \infty$.*

Proof. The spectral radius condition implies that there exists an induced matrix norm $\|\cdot\|_w$ such that

$$\sigma(I + \gamma B^{-1}A) \leq \|I + \gamma B^{-1}A\|_w < 1. \tag{26}$$

For any sample trajectory such that $Z_t \to 0$, there exists $\bar{t}$ such that for all $t \geq \bar{t}$,

$$\|I + \gamma B^{-1}A + \gamma Z_t\|_w < 1 - \epsilon$$

for some positive ϵ, and consequently, from (24)

$$\|r_{t+1} - r^*\|_w \leq (1-\epsilon)\|r_t - r^*\|_w + \gamma\|Z_t r^* + \zeta_t\|_w.$$

The above relation implies that for all sample trajectories such that both $Z_t \to 0$ and $\zeta_t \to 0$ (so that $\gamma\|Z_t r^* + \zeta_t\|_w \to 0$), we have $r_t - r^* \to 0$. Since the set of these trajectories has probability 1, we have $r_t \to r^*$ with probability 1. □

The preceding proposition implies that for deriving the convergence condition of the constant stepsize LSPE(λ) iteration (23) (e.g., range of stepsize γ), we can focus on the deterministic portion:

$$r_{t+1} = r_t + \gamma B^{-1}(Ar_t + b). \tag{27}$$

This deterministic iteration is equivalent to

$$\Phi r_{t+1} = F_{\gamma,\lambda}(\Phi r_t), \tag{28}$$

where

$$F_{\gamma,\lambda} = (1-\gamma)I + \gamma\Pi T^{(\lambda)}, \tag{29}$$

[cf. (15) and its equivalent iteration]. To exploit this equivalence between (27) and (28), we will associate the spectral radius condition $\sigma(I+\gamma B^{-1}A) < 1$ with the contraction and non-expansiveness of the mapping $F_{\gamma,\lambda}$ on the subspace S.[2] In this connection, we note that the spectral radius $\sigma(I + \gamma B^{-1}A)$ is bounded above by the induced norm of the mapping $F_{\gamma,\lambda}$ restricted to S with respect to any norm, and that the condition $\sigma(I + \gamma B^{-1}A) < 1$ is equivalent to $F_{\gamma,\lambda}$ being a contraction mapping on S for some norm. It is convenient to consider the $\|\cdot\|_\pi$ norm and use the non-expansiveness or contraction property of $F_{\gamma,\lambda}$ to bound the spectral radius $\sigma(I + \gamma B^{-1}A)$, because the properties of $\Pi T^{(\lambda)}$ under this norm are well-known. For example, using the fact

$$\|P^{(\lambda)}\|_\pi \leq 1, \quad \|\Pi P^{(\lambda)}\|_\pi \leq 1, \qquad \forall\ \lambda \in [0,1),$$

we have that the mapping $F_{\gamma,\lambda}$ of (29) is nonexpansive for all $\lambda \in [0,1)$ and $\gamma \in (0,1]$, so

$$\sigma(I + \gamma B^{-1}A) \leq 1, \qquad \forall\ \lambda \in [0,1), \quad \forall\ \gamma \in (0,1]. \tag{30}$$

Thus, to prove that the spectral radius condition $\sigma(I + \gamma B^{-1}A) < 1$ holds for various values of λ and γ, we may follow one of two approaches:

(1) A direct approach, which involves showing that the modulus of each eigenvalue of $I + \gamma B^{-1}A$ is less than 1; this is the approach followed by Bertsekas et al. [17] for the discounted case.

[2] Throughout the paper, we say that a mapping $G : \Re^n \mapsto \Re^n$ is a contraction or is nonexpansive over a set $X \subseteq \Re^n$ if $\|G(x) - G(y)\| \leq \rho\|x - y\|$ for all $x, y \in X$, where $\rho \in (0,1)$ or $\rho = 1$, respectively. The set X and the norm $\|\cdot\|$ will be either clearly implied by the context or specified explicitly.

(2) An indirect approach, which involves showing that the mapping $F_{\gamma,\lambda} = (1-\gamma)I + \gamma\Pi T^{(\lambda)}$ is a contraction with respect to $\|\cdot\|_\pi$.

The first approach provides stronger results and can address exceptional cases that the second approach cannot handle (we will see that one such case is when $\lambda = 0$ and $\gamma = 1$), while the second approach provides insight, and yields results that can be applied to more general contexts of compositions of Euclidean projections and nonexpansive mappings. The second approach also has the merit of simplifying the analysis. As an example, in the discounted case with a discount factor β, because the mapping $T^{(\lambda)}$ (given by the multiple-step Bellman equation for the discounted problem) is a $\|\cdot\|_\pi$-norm contraction with modulus $\rho(\beta,\lambda) = \frac{(1-\lambda)\beta}{1-\lambda\beta} \leq \beta$ for all $\lambda \in [0,1]$, it follows immediately from the second approach that the constant stepsize discounted LSPE(λ) algorithm converges if its stepsize γ lies in the interval $\left(0, \frac{2}{1+\rho(\beta,\lambda)}\right)$. This simplifies parts of the proof given in [17]. For the average cost case, we will give both lines of analysis in Section 3, and the assumption that $e \notin S$ (Assumption 2) will play an important role in both, as we will see.

Note a high-level interpretation of the LSPE(λ) iteration, based on (23): With γ chosen in the convergence range of the algorithm (given in Section 3), the LSPE(λ) iteration can be viewed as a *contracting* (possibly damped) approximate value iteration plus asymptotically diminishing stochastic noise ϵ_t [cf. (23), (27) and (28)],

$$\Phi r_{t+1} = F_{\gamma,\lambda}(\Phi r_t) + \epsilon_t.$$

2.4 The LSTD(λ) Algorithm

A different least squares TD algorithm, the average cost LSTD(λ) method, calculates at time t

$$\hat{r}_{t+1} = -\bar{A}_t^{-1}\,\bar{b}_t. \tag{31}$$

For large enough t the iterates are well-defined[3] and converge to $r^* = -A^{-1}b$. Thus LSTD(λ) estimates by simulation two quantities defining the solution to which TD(λ) converges. We see that the rationales behind LSPE(λ) and LSTD(λ) are quite different: the former approximates the fixed point iteration $\Phi r_{t+1} = F_{\gamma,\lambda}(\Phi r_t)$ [or when $\gamma = 1$, the iteration $\Phi r_{t+1} = \Pi T^{(\lambda)}(\Phi r_t)$] by introducing asymptotically diminishing simulation noise in its right-hand side, while the latter solves at each iteration an increasingly accurate simulation-based approximation to the equation $\Phi r = \Pi T^{(\lambda)}(\Phi r)$.

Note that LSTD(λ) differs from LSPE(λ) in an important respect: it does not use an initial guess r_0 and hence cannot take advantage of any knowledge about the value of r^*. This can make a difference in the context of policy iteration, where many policies are successively evaluated, often using relatively few simulation samples, as discussed in Section 6.

Some insight into the connection of LSPE(λ) and LSTD(λ) can be obtained by verifying that the LSTD(λ) estimate $\hat{r}_{t+1}$ is also the unique vector $\hat{r}$ satisfying

$$\hat{r} = \arg\min_{r\in\Re^s} \sum_{k=0}^{t} \left(\phi(x_k)'r - \phi(x_k)'\hat{r} - \sum_{m=k}^{t} \lambda^{k-m}\hat{d}(m) \right)^2, \tag{32}$$

where

$$\hat{d}(m) = g(x_m, x_{m+1}) - \eta_m + \phi(x_{m+1})'\hat{r} - \phi(x_m)'\hat{r}.$$

Note that finding $\hat{r}$ that satisfies (32) is not a least squares problem, because the expression in the right-hand side of (32) involves $\hat{r}$. Yet, the similarity with the least squares problem solved by

[3]The inverse $\bar{A}_t^{-1}$ exists for t sufficiently large. The reason is that $\bar{A}_t$ converges with probability 1 to the matrix $A = \Phi' D(P^{(\lambda)} - I)\Phi$, which is negative definite (in the sense $r'Ar < 0$ for all $r \neq 0$) and hence invertible (see the proof of Lemma 7 of [7]).

LSPE(λ) [cf. (16)] is evident. Empirically, the two methods also produce similar iterates. Indeed, it can be verified from (22) and (31) that the difference of the iterates produced by the two methods satisfies the following recursion:

$$r_{t+1} - \hat{r}_{t+1} = \left(I + \gamma \bar{B}_t^{-1} \bar{A}_t\right)(r_t - \hat{r}_t) + \left(I + \gamma \bar{B}_t^{-1} \bar{A}_t\right)(\hat{r}_t - \hat{r}_{t+1}). \tag{33}$$

In Section 4 we will use this recursion and the spectral radius result $\sigma(I + \gamma B^{-1}A) < 1$ of Section 3 to establish one of our main results, namely that the difference $r_t - \hat{r}_t$ converges to 0 faster than r_t and $\hat{r}_t$ converge to their limit r^*.

3 Convergence of Average Cost LSPE(λ) with a Constant Stepsize

In this section, we will analyze the convergence of the constant stepsize average cost LSPE(λ) algorithm under Assumptions 1 and 2. We will derive conditions guaranteeing that $\sigma(I+\gamma B^{-1}A) < 1$, and hence guaranteeing that LSPE(λ) converges, as per Prop. 1. In particular, the convergent stepsize range for LSPE(λ) will be shown to contain the interval $(0, 1]$ for $\lambda \in (0, 1)$, the interval $(0, 1)$ for $\lambda = 0$, and the interval $(0, 1]$ for $\lambda = 0$ and an aperiodic Markov chain (Prop. 2). We will then provide an analysis of the contraction property of the mapping $F_{\lambda,\gamma}$ underlying LSPE(λ) with respect to the $\|\cdot\|_\pi$ norm, which yields as a byproduct an alternative line of convergence proof, as discussed in Section 2.3.

For both lines of analysis, our approach will be to investigate the properties of the stochastic matrix $P^{(\lambda)}$, the approximation subspace S and its relation to the eigenspace of $P^{(\lambda)}$, and the composition of projection Π on S with $P^{(\lambda)}$, and to then, for the spectral radius-based analysis, pass the results to the s-dimensional matrix $I+\gamma B^{-1}A$ using equivalence relations discussed in Section 2.

3.1 Convergence Analysis Based on Spectral Radius

We start with a general result relating to the spectral radius of certain matrices that involve projections. In the proof we will need an extension of a Euclidean norm to the space $\mathcal{C}^n$ of n-tuples of complex numbers. For any Euclidean norm $\|\cdot\|$ in $\Re^n$ (a norm of the form $\|x\| = \sqrt{x'Qx}$, where Q is a positive definite symmetric matrix), the norm of a complex number $x + iy \in \mathcal{C}^n$ is defined by

$$\|x + iy\| = \sqrt{\|x\|^2 + \|y\|^2}.$$

For a set $X \subseteq \Re^n$, we denote by $X + iX$ the set of complex numbers $\{x + iy \mid x \in X, y \in X\}$. We also use the fact that for a projection matrix Π that projects a real vector to a subspace of $\Re^n$, the complex vector Πz has as its real and imaginary parts the projections of the corresponding real and imaginary parts of z, respectively.

Lemma 1. *Let S be a subspace of $\Re^n$ and let C be an $n \times n$ real matrix, such that for some Euclidean norm $\|\cdot\|$ we have $\|C\| \leq 1$. Denote by Π the projection matrix which projects a real vector onto S with respect to this norm. Let ν be a complex number with $|\nu| = 1$, and let ξ be a vector in $\mathcal{C}^n$. Then ν is an eigenvalue of ΠC with corresponding eigenvector ξ if and only if ν is an eigenvalue of C with corresponding eigenvector ξ, and $\nu\xi \in S + iS$.*

Proof. Assume that $\Pi C\xi = \nu\xi$. We claim that $C\xi \in S + iS$; if this were not so, we would have

$$\|C\xi\| > \|\Pi C\xi\| = \|\nu\xi\| = |\nu|\,\|\xi\| = \|\xi\|,$$

which contradicts the assumption $\|C\| \leq 1$. Thus, $C\xi \in S + iS$, which implies that $C\xi = \Pi C\xi = \nu\xi$, and $\nu\xi \in S + iS$. Conversely, if $C\xi = \nu\xi$ and $\nu\xi \in S + iS$, we have $\Pi C\xi = \nu\xi$. □

We now specialize the preceding lemma to obtain a necessary and sufficient condition for the spectral radius condition (25) to hold.

Lemma 2. *Let ν be a complex number with $|\nu| = 1$ and let z be a nonzero vector in $\mathcal{C}^s$. Then under Assumption 2, ν is an eigenvalue of $I + B^{-1}A$ and z is a corresponding eigenvector if and only if ν is an eigenvalue of $P^{(\lambda)}$ and Φz is a corresponding eigenvector.*

Proof. We apply Lemma 1 for the special case where $C = P^{(\lambda)}$, S is the subspace spanned by the columns of Φ, and the Euclidean norm is $\|\cdot\|_\pi$. We have $\|P^{(\lambda)}\|_\pi = 1$ [cf. (7)]. Since

$$\Pi P^{(\lambda)} \Phi = \Phi(I + B^{-1}A)$$

[cf. (13)], and by Assumption 2, Φ has linearly independent columns, we have that (ν, z) is an eigenvalue/eigenvector pair of $I + B^{-1}A$ if and only if $(\nu, \Phi z)$ is an eigenvalue/eigenvector pair of $\Pi P^{(\lambda)}$, which by Lemma 1, for a complex number ν with $|\nu| = 1$, holds if and only if ν is an eigenvalue of $P^{(\lambda)}$ and Φz is a corresponding eigenvector. □

We now apply the preceding lemma to prove the convergence of LSPE(λ).

Proposition 2. *Under Assumptions 1 and 2, we have*

$$\sigma(I + \gamma B^{-1}A) < 1,$$

and hence the average cost LSPE(λ) iteration (22) with constant stepsize γ converges to r^* with probability 1 as $t \to \infty$, for any one of the following cases:

(i) $\lambda \in (0, 1)$ *and* $\gamma \in (0, 1]$*;*

(ii) $\lambda = 0$, $\gamma \in (0, 1]$, *and P is aperiodic;*

(iii) $\lambda = 0$, $\gamma \in (0, 1]$, *P is periodic, and all its eigenvectors that correspond to some eigenvalue ν with $\nu \neq 1$ and $|\nu| = 1$, do not lie in the subspace $S = \{\Phi r \mid r \in \Re^s\}$;*

(iv) $\lambda = 0$, $\gamma \in (0, 1)$.

Proof. We first note that by (30), we have $\sigma(I + \gamma B^{-1}A) \leq 1$, so we must show that $I + \gamma B^{-1}A$ has no eigenvalue with modulus 1.

In cases (i)-(iii), we show that there is no eigenvalue ν of $P^{(\lambda)}$ that has modulus 1 and an eigenvector of the form Φz, and then use Lemma 2 to conclude that $\sigma(I + B^{-1}A) < 1$. This also implies that $\sigma(I + \gamma B^{-1}A) < 1$ for all $\gamma \in (0, 1)$, since $I + \gamma B^{-1}A = (1 - \gamma)I + \gamma(I + B^{-1}A)$.

Indeed, in both cases (i) and (ii), $P^{(\lambda)}$ is aperiodic [in case (i), all entries of $P^{(\lambda)}$ are positive, so it is aperiodic, while in case (ii), $P^{(0)}$ is equal to P, which is aperiodic by assumption]. Thus, the only eigenvalue of $P^{(\lambda)}$ with unit modulus is $\nu = 1$, and its eigenvectors are the scalar multiples of e, which are not of the form Φz by Assumption 2.

In case (iii), a similar argument applies, using the hypothesis.

Finally, consider case (iv). By Lemma 2, an eigenvalue ν of $I + B^{-1}A$ with $|\nu| = 1$ is an eigenvalue of P with eigenvectors of the form Φz. Hence we cannot have $\nu = 1$, since the corresponding eigenvectors of P are the scalar multiples of e, which cannot be of the form Φz by Assumption 2. Therefore, the convex combinations $(1 - \gamma) + \gamma\nu$, $\gamma \in (0, 1)$, lie in the interior of the unit circle for all eigenvalues ν of $I + B^{-1}A$, showing that $\sigma(I + \gamma B^{-1}A) < 1$ for $\gamma \in (0, 1)$. □

Remark 1. We give an example showing that when $\lambda = 0$ and P is periodic, the matrix $I + B^{-1}A$ can have spectral radius equal to 1, if the assumption in case (iii) of Prop. 2 is not satisfied. Let

$$P = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}, \qquad \Phi = \begin{bmatrix} 1 \\ -1 \end{bmatrix}.$$

For any $r \in \Re$, using (9), we have

$$(I + B^{-1}A)r = (\Phi' D \Phi)^{-1} \Phi' D P \Phi r = -r,$$

so $\sigma(I + B^{-1}A) = 1$. Here the eigenvectors corresponding to the eigenvalue -1 of P are the nonzero multiples of $(1, -1)'$, and belong to S.

Remark 2. Our analysis can be extended to show the convergence of LSPE(λ) with a time varying stepsize γ_t, where γ_t for all t lies in a closed interval contained in the range of stepsizes given by Prop. 2. This follows from combining the spectral radius result of Prop. 2 with a refinement in the proof argument of Prop. 1. In particular, the refinement is to assert that for all γ in the closed interval given above, we can choose a common norm $\|\cdot\|_w$ in the proof of Prop. 1. This in turn follows from explicitly constructing such a norm using the Jordan form of the matrix $I + \gamma B^{-1}A$ (for a related reference, see e.g., Ortega and Rheinboldt [21], p. 44).

3.2 Contraction Property of $F_{\gamma,\lambda}$ with respect to $\|\cdot\|_\pi$

For the set of pairs (λ, γ) given in the preceding spectral radius analysis (Prop. 2), $F_{\gamma,\lambda}$ of (28) is a contraction mapping with respect to some, albeit unknown, norm. We will now refine this characterization of $F_{\gamma,\lambda}$ by deriving the pairs (λ, γ) for which $F_{\gamma,\lambda}$ is a contraction with respect to the norm $\|\cdot\|_\pi$ (see the subsequent Prop. 4). These values form a subset of the former set; alternatively, as discussed in Section 2.3, one can follow this line of analysis to assert the convergence of LSPE(λ) for the respective smaller set of stepsize choices (the case $\lambda = 0$ turns out to be exceptional).

First, we prove the following proposition, which can be applied to the convergence analysis of general iterations involving the composition of a nonexpansive linear mapping and a projection on a subspace. The analysis generalizes some proof arguments used in the error analysis in [7], part of which is essentially also based on the contraction property.

Proposition 3. *Let S be a subspace of $\Re^n$ and let $H : \Re^n \mapsto \Re^n$ be a linear mapping,*

$$H(x) = Cx + d,$$

where C is an $n \times n$ matrix and d is a vector in $\Re^n$. Let $\|\cdot\|$ be a Euclidean norm with respect to which H is nonexpansive, and let Π denote projection onto S with respect to that norm.

(a) ΠH has a unique fixed point if and only if either 1 is not an eigenvalue of C, or else the eigenvectors corresponding to the eigenvalue 1 do not belong to S.

(b) If ΠH has a unique fixed point, then for all $\gamma \in (0, 1)$, the mapping

$$G_\gamma = (1 - \gamma)I + \gamma \Pi H$$

is a contraction, i.e., for some scalar $\rho_\gamma \in (0, 1)$, we have

$$\|G_\gamma x - G_\gamma y\| \leq \rho_\gamma \|x - y\|, \qquad \forall\ x, y \in \Re^n.$$

Proof. (a) The linear mapping ΠH has a unique fixed point if and only if 1 is not an eigenvalue of ΠC. By Lemma 1, 1 is an eigenvalue of ΠC if and only if 1 is an eigenvalue of C with the corresponding eigenvectors in $S + iS$, from which part (a) follows.

(b) Since ΠH has a unique fixed point, we have $z \neq \Pi C z$ for all $z \neq 0$. Hence, $\forall\ z \in \Re^n$, either $\Pi C z \notin \{cz | c \in \Re\}$, or $\Pi C z = cz$ for some scalar $c \in [-1, 1)$ due to the nonexpansiveness of ΠC. In the first case we have

$$\begin{aligned} \|(1 - \gamma)z + \gamma \Pi C z\| &< (1 - \gamma)\|z\| + \gamma \|\Pi C z\| \\ &\leq (1 - \gamma)\|z\| + \gamma \|z\| = \|z\|, \end{aligned} \tag{34}$$

where the strict inequality follows from the strict convexity of the norm, and the weak inequality follows from the non-expansiveness of ΠC. In the second case, (34) follows easily. If we define $\rho_\gamma = \sup\{\|(1-\gamma)z + \gamma\Pi Cz\| \mid \|z\| \leq 1\}$, and note that the supremum above is attained by Weierstrass' Theorem, we see that (34) yields $\rho_\gamma < 1$ and

$$\|(1-\gamma)z + \gamma\Pi Cz\| \leq \rho_\gamma \|z\|, \qquad \forall\ z \in \Re^n.$$

By letting $z = x - y$, with $x, y \in \Re^n$, and by using the definition of G_γ, part (b) follows. □

We can now derive the pairs (λ, γ) for which the mapping $F_{\gamma,\lambda}$ underlying the LSPE(λ) iteration is a $\|\cdot\|_\pi$-norm contraction.

Proposition 4. *Under Assumptions 1 and 2, the mapping*

$$F_{\gamma,\lambda} = (1-\gamma)I + \gamma\Pi T^{(\lambda)}$$

is a contraction with respect to $\|\cdot\|_\pi$ for either one of the following cases:

(i) $\lambda \in (0,1)$ *and* $\gamma \in (0,1]$,

(ii) $\lambda = 0$ *and* $\gamma \in (0,1)$.

Proof. For $\gamma \in (0,1)$, we apply Prop. 3, with H equal to $T^{(\lambda)}$, C equal to the stochastic matrix $P^{(\lambda)}$, and S equal to the subspace spanned by the columns of Φ. The mapping $\Pi T^{(\lambda)}$ has a unique fixed point, the vector Φr^*, as shown by Tsitsiklis and Van Roy [7] [this can also be shown simply by using Prop. 3 (a)]. Thus, the result follows from Prop. 3 (b).

Consider now the remaining case, $\gamma = 1$ and $\lambda \in (0,1)$. Then $T^{(\lambda)}$ is a linear mapping involving the matrix $P^{(\lambda)}$ [cf. (4)]. Since $\lambda > 0$ and all states form a single recurrent class, all entries of $P^{(\lambda)}$ are positive [cf. (5)]. Thus $P^{(\lambda)}$ can be expressed as a convex combination

$$P^{(\lambda)} = (1-\alpha)I + \alpha\bar{P}$$

for some $\alpha \in (0,1)$, where $\bar{P}$ is a stochastic matrix with positive entries. We make the following observations:

(i) $\bar{P}$ corresponds to a nonexpansive mapping with respect to the norm $\|\cdot\|_\pi$. The reason is that π is an invariant distribution of $\bar{P}$, i.e., $\pi = \pi\bar{P}$, [as can be verified by using the relation $\pi = \pi P^{(\lambda)}$]. Thus, we have $\|\bar{P}z\|_\pi \leq \|z\|_\pi$ for all $z \in \Re^n$ [6, Lemma 1], implying that $\bar{P}$ has the non-expansiveness property mentioned.

(ii) Since $\bar{P}$ has all positive entries, the states of the Markov chain corresponding to $\bar{P}$ form a single recurrent class. Hence the eigenvectors of $\bar{P}$ corresponding to the eigenvalue 1 are the nonzero scalar multiples of e, which by Assumption 2, do not belong to the subspace S.

It follows from Prop. 3 (with $\bar{P}$ in place of C, and α in place of γ) that $\Pi P^{(\lambda)}$ is a contraction with respect to the norm $\|\cdot\|_\pi$, which implies that $\Pi T^{(\lambda)}$ is also a contraction. □

Remark 3. As Prop. 2 and Prop. 4 suggest, if P is aperiodic, $\Pi T^{(0)}$ may not be a contraction on the subspace S with respect to the norm $\|\cdot\|_\pi$, while it is a contraction on S with respect to another norm. As an example, let

$$P = \begin{bmatrix} 0 & 1 & 0 \\ 1/2 & 0 & 1/2 \\ 1/2 & 0 & 1/2 \end{bmatrix}, \qquad \Phi = \begin{bmatrix} 0 & 1 \\ 1 & 0 \\ 0 & 0 \end{bmatrix},$$

$$g = \begin{bmatrix} 0 & 0 & 0 \end{bmatrix}', \qquad \eta^* = 0,$$

and note that P is aperiodic. Then $\pi = (1/3, 1/3, 1/3)$, so the norm $\|\cdot\|_\pi$ coincides with a scaled version of the standard Euclidean norm. Let Φ_1 and Φ_2 denote the columns of Φ. For $\lambda = 0$,

$$\Pi T^{(0)} \Phi_1 - \Pi T^{(0)}(0) = \Pi P \Phi_1 = \Phi_2.$$

Since $\|\Phi_1\|_\pi = \|\Phi_2\|_\pi$, $\Pi T^{(0)}$ is not a contraction on S with respect to $\|\cdot\|_\pi$. However, according to Prop. 2 (ii), we have $\sigma(I + B^{-1}A) < 1$, which implies that $\Pi T^{(0)}$ is a contraction on S with respect to a different norm.

4 Rate of Convergence of LSPE(λ)

In this section we prove that LSPE(λ) has the same asymptotic convergence rate as LSTD(λ), for any constant stepsize γ under which LSPE(λ) converges. The proof applies to both the discounted and average cost cases and for all values of λ for which convergence has been proved ($\lambda \in [0, 1]$ for the discounted case and $\lambda \in [0, 1)$ for the average cost case).

For both discounted[4] and average cost cases, the LSPE(λ) updates can be expressed as

$$r_{t+1} = r_t + \gamma \bar{B}_t^{-1} \left(\bar{A}_t \, r_t + \bar{b}_t \right),$$

while the LSTD(λ) updates can be expressed as

$$\hat{r}_{t+1} = -\bar{A}_t^{-1} \, \bar{b}_t.$$

Informally, it has been observed in [17] that r_t became close to and "tracked" $\hat{r}_t$ well before the convergence to r^* took place - see also the experiments in Section 5. The explanation of this phenomenon given in [17] is a two-time-scale type of argument: when t is large, $\bar{A}_t$, $\bar{B}_t$ and $\bar{b}_t$ change slowly so that they are essentially "frozen" at certain values, and r_t then "converges" to the unique fixed point of the linear system

$$r = r + \gamma \bar{B}_t^{-1} \left(\bar{A}_t \, r + \bar{b}_t \right),$$

which is $-\bar{A}_t^{-1} \, \bar{b}_t$, the value of $\hat{r}_t$ of LSTD(λ).

In what follows, we will make the above argument more precise, by first showing that the distance between LSPE(λ) and LSTD(λ) iterates shrinks at the order of $O(1/t)$ (Prop. 5). We will then appeal to the results of Konda [13], which show that the LSTD(λ) iterates converge to their limit at the

[4] For the β-discounted criterion and $\lambda \in [0, 1]$, the update rules of LSPE(λ) and LSTD(λ) are given by (22) and (31), respectively, with the corresponding matrices

$$B_t = \sum_{k=0}^{t} \phi(x_k) \, \phi(x_k)', \qquad A_t = \sum_{k=0}^{t} z_k \left(\beta \phi(x_{k+1})' - \phi(x_k)' \right),$$

$$b_t = \sum_{k=0}^{t} z_k \, g(x_k, x_{k+1}), \qquad z_k = \sum_{m=0}^{k} (\beta\lambda)^{k-m} \phi(x_m),$$

(see [17]); and the stepsize of LSPE(λ) is chosen in the range $\left(0, \frac{2}{1+\rho(\beta,\lambda)}\right)$, where $\rho(\beta, \lambda) = \frac{(1-\lambda)\beta}{1-\lambda\beta}$ (cf. [17, Prop. 3.1] and also our discussion in Section 2.3). The matrix $\bar{A}_t$ and vector $\bar{b}_t$ converge to A and b, respectively, with

$$A = \Phi' D (P^{(\beta,\lambda)} - I) \Phi, \qquad b = \Phi' D (I - \lambda\beta P)^{-1} g,$$

where

$$P^{(\beta,\lambda)} = (1 - \lambda) \sum_{m=0}^{\infty} \lambda^m (\beta P)^{m+1}, \quad \lambda \in [0, 1],$$

(see [16]). LSPE(λ) and LSTD(λ) converge to the same limit $r^* = -A^{-1}b$. Alternatively, one may approximate relative cost differences, similar to the average cost case and to the discussion in [8]; the resulting iterates may have lower variance. Our analysis can be easily applied to such algorithm variants.

order of $O(1/\sqrt{t})$. It then follows that LSPE(λ) and LSTD(λ) converge to each other at a faster time scale than to the common limit; the asymptotic convergence rate of LSPE(λ) also follows as a consequence (Prop. 6).

For the results of this section, we assume the conditions that ensure the convergence of LSPE(λ) and LSTD(λ) algorithms. In particular, we assume the following conditions:

Condition 1. *For the average cost case, Assumptions 1 and 2 hold, and in addition, for LSPE(λ), the stepsize γ is chosen as in Prop. 2; and for the β-discounted case, Assumption 1 holds, the columns of Φ are linearly independent, and in addition, for LSPE(λ), the stepsize γ is in the range $\left(0, \frac{2}{1+\rho(\beta,\lambda)}\right)$, where $\rho(\beta,\lambda) = \frac{(1-\lambda)\beta}{1-\lambda\beta}$ (cf. [17]).*

The difference between the LSPE(λ) and LSTD(λ) updates can be written as [cf. (33)]

$$r_{t+1} - \hat{r}_{t+1} = \left(I + \gamma \bar{B}_t^{-1} \bar{A}_t\right)(r_t - \hat{r}_t) + \left(I + \gamma \bar{B}_t^{-1} \bar{A}_t\right)(\hat{r}_t - \hat{r}_{t+1}). \tag{35}$$

The norm of the difference term $\hat{r}_t - \hat{r}_{t+1}$ of the LSTD(λ) iterates in the right-hand side above is of the order $O(1/t)$, as shown in the next lemma. To simplify the description, in what follows, we say a sample path is *convergent* if it is such that $\bar{A}_t$, $\bar{B}_t$, and $\bar{b}_t$ converge to A, B, and b, respectively. (All such paths form a set of probability 1, on which both LSTD(λ) and LSPE(λ) converge to $r^* = -A^{-1}b$.)

Lemma 3. *Let Condition 1 hold and consider a convergent sample path. Then for each norm $\|\cdot\|$, there exists a constant C such that for all t sufficiently large,*

$$\|\hat{r}_{t+1} - \hat{r}_t\| \le \frac{C}{t}.$$

Proof. This is a straightforward verification. By definition of the LSTD(λ) updates, we have

$$\begin{aligned}\|\hat{r}_{t+1} - \hat{r}_t\| &= \left\|\bar{A}_t^{-1}\bar{b}_t - \bar{A}_{t-1}^{-1}\bar{b}_{t-1}\right\| \\ &\le \left\|\bar{A}_t^{-1} - \bar{A}_{t-1}^{-1}\right\| \left\|\bar{b}_t\right\| + \left\|\bar{A}_{t-1}^{-1}\right\| \left\|\bar{b}_t - \bar{b}_{t-1}\right\|. \end{aligned} \tag{36}$$

Since $\left\|\bar{b}_t\right\| \to \|b\|$ and $\left\|\bar{A}_{t-1}^{-1}\right\| \to \left\|A^{-1}\right\|$, we have $\left\|\bar{b}_t\right\| \le C_1, \left\|\bar{A}_{t-1}^{-1}\right\| \le C_2$ for some constants C_1 and C_2 and for all t sufficiently large. Thus we only need to bound the terms $\left\|\bar{A}_t^{-1} - \bar{A}_{t-1}^{-1}\right\|$ and $\left\|\bar{b}_t - \bar{b}_{t-1}\right\|$ by $\frac{C}{t}$ for some constant C. By the definition of $\bar{b}_t$, it can be seen that for t sufficiently large, for the average cost case,

$$\left\|\bar{b}_t - \bar{b}_{t-1}\right\| \le \frac{1}{t}\left\|z_t(g(x_t) - \eta_t)\right\| + \frac{1}{t}\|\bar{b}_t\| \le \frac{C}{t}$$

for some constant C, (since z_t, η_t, and $\bar{b}_t$ are bounded for all t), and similarly, the relation holds for the discounted case (the difference being without the term η_t). By the definition of $\bar{A}_t$,

$$\begin{aligned}\left\|\bar{A}_t^{-1} - \bar{A}_{t-1}^{-1}\right\| &= \left\|(t+1)A_t^{-1} - tA_{t-1}^{-1}\right\| \\ &= \left\|A_t^{-1} + t\left(A_t^{-1} - A_{t-1}^{-1}\right)\right\| \\ &\le \left\|A_t^{-1}\right\| + t\left\|A_t^{-1} - A_{t-1}^{-1}\right\|.\end{aligned}$$

Applying the Sherman-Morisson formula for matrix inversion to A_t^{-1} in the second term of the last expression, it can be seen that for $\beta \in [0,1]$,

$$\begin{aligned}\left\|\bar{A}_t^{-1} - \bar{A}_{t-1}^{-1}\right\| \le& \left\|A_t^{-1}\right\| + t\left\|\frac{A_{t-1}^{-1} z_t \left(\beta\phi(x_{t+1})' - \phi(x_t)'\right) A_{t-1}^{-1}}{1 + \left(\beta\phi(x_{t+1})' - \phi(x_t)'\right) A_{t-1}^{-1} z_t}\right\| \\ =& \frac{1}{t+1}\left\|\bar{A}_t^{-1}\right\| + \left\|\frac{\bar{A}_{t-1}^{-1} z_t \left(\beta\phi(x_{t+1})' - \phi(x_t)'\right) \bar{A}_{t-1}^{-1}}{t + \left(\beta\phi(x_{t+1})' - \phi(x_t)'\right) \bar{A}_{t-1}^{-1} z_t}\right\| \\ \le& \frac{C_2}{t} + \frac{C_3}{t}\end{aligned}$$

for some constant C_3 and t sufficiently large. Combine these relations with (36) and the claim follows. □

The next result provides the rate at which the LSPE(λ) and LSTD(λ) iterates converge to each other.

Proposition 5. *Under Condition 1, the sequence of random variables $t(r_t - \hat{r}_t)$ is bounded with probability* 1.

Proof. Consider a convergent sample path. Since $\sigma(I + \gamma B^{-1}A) < 1$ (as proved in [17, Prop. 3.1] for the discounted case and in our Prop. 2 of Section 3 for the average cost case), we may assume that there exist a scalar $\rho \in (0,1)$ and a norm $\|\cdot\|_w$ such that

$$\|I + \gamma \bar{B}_t^{-1}\bar{A}_t\|_w \leq \rho$$

for all t sufficiently large. From (35), we see that

$$\|r_{t+1} - \hat{r}_{t+1}\|_w \leq \left\|I + \gamma \bar{B}_t^{-1}\bar{A}_t\right\|_w \|r_t - \hat{r}_t\|_w + \left\|I + \gamma \bar{B}_t^{-1}\bar{A}_t\right\|_w \|\hat{r}_t - \hat{r}_{t+1}\|_w.$$

Thus, using also Lemma 3 with the norm being $\|\cdot\|_w$, we obtain

$$\|r_{t+1} - \hat{r}_{t+1}\|_w \leq \rho \|r_t - \hat{r}_t\|_w + \frac{\rho C}{t},$$

for all t sufficiently large. This relation can be written as

$$\zeta_{t+1} \leq \frac{t+1}{t}\rho\,\zeta_t + \frac{t+1}{t}\rho\,C, \tag{37}$$

where

$$\zeta_t = t\,\|r_t - \hat{r}_t\|_w.$$

Let $\bar{t}$ be such that $\bar{\rho} < 1$, where $\bar{\rho} = \frac{\bar{t}+1}{\bar{t}}\rho$. Then, for all $t \geq \bar{t}$, from the relation $\zeta_{t+1} \leq \bar{\rho}\,\zeta_t + \bar{\rho}\,C$ [cf. (37)], we have

$$\zeta_t \leq \bar{\rho}^{(t-\bar{t})}\,\zeta_{\bar{t}} + \frac{C\bar{\rho}(1-\bar{\rho}^{(t-\bar{t})})}{1-\bar{\rho}} \leq \zeta_{\bar{t}} + \frac{C\bar{\rho}}{1-\bar{\rho}}.$$

Thus the sequence ζ_t is bounded, which implies the desired result. □

Note that Prop. 5 implies that the sequence of random variables $t^\alpha(r_t - \hat{r}_t)$ converges to zero with probability 1 as $t \to \infty$ for any $\alpha < 1$. Using this implication, we now show that LSPE(λ) has the same convergence rate as LSTD(λ), assuming that LSTD(λ) converges to its limit with error that is normally distributed, in accordance with the central limit theorem (as shown by Konda [13]). We denote by $\mathcal{N}(0,\Sigma)$ a vector-valued Gaussian random variable with zero mean and covariance matrix Σ.

Proposition 6. *Let Condition 1 hold. Suppose that the sequence of random variables $\sqrt{t}\,(\hat{r}_t - r^*)$ of LSTD(λ) converges in distribution to $\mathcal{N}(0,\Sigma_0)$ as $t \to \infty$. Then for any given initial r_0, the sequence of random variables $\sqrt{t}\,(r_t - r^*)$ of LSPE(λ) converges in distribution to $\mathcal{N}(0,\Sigma_0)$ as $t \to \infty$.*

Proof. Using the definition of LSPE(λ) and LSTD(λ) [cf. (22) and (31)], it can be verified that

$$\sqrt{t+1}(r_{t+1} - r^*) = \sqrt{t+1}\left(I + \gamma \bar{B}_t^{-1}\bar{A}_t\right)(r_t - \hat{r}_{t+1}) + \sqrt{t+1}(\hat{r}_{t+1} - r^*),$$

and thus it suffices to show that $\sqrt{t+1}\left(I + \gamma \bar{B}_t^{-1}\bar{A}_t\right)(r_t - \hat{r}_{t+1}) \to 0$ with probability 1. (Here we have used the following fact: if X_n converges to X in distribution and Z_n converges to 0 with

probability 1, then $X_n + Z_n$ converges to X in distribution. See e.g., Duflo [22, Properties 2.1.2 (3) and (4)], p. 40.)

Consider a sample path for which both LSTD(λ) and LSPE(λ) converge. Choose a norm $\|\cdot\|$. When t is sufficiently large, we have $\|I + \gamma \bar{B}_t^{-1} \bar{A}_t\| \leq C$ for some constant C, so that

$$\left\|\sqrt{t+1}\left(I + \gamma \bar{B}_t^{-1} \bar{A}_t\right)(r_t - \hat{r}_{t+1})\right\| \leq C\sqrt{t+1}\left(\|r_t - \hat{r}_t\| + \|\hat{r}_t - \hat{r}_{t+1}\|\right).$$

Since $\|\hat{r}_t - \hat{r}_{t+1}\| \leq \frac{C'}{t}$ for some constant C' (Lemma 3), the second term $C\sqrt{t+1}\|\hat{r}_t - \hat{r}_{t+1}\|$ converges to 0. By Prop. 5, the first term, $C\sqrt{t+1}\|r_t - \hat{r}_t\|$, also converges to 0. The proof is thus complete. □

Remark 4. A convergence rate analysis of LSTD(λ) and TD(λ) is provided by Konda [13, Chapter 6]. (In this analysis, the estimate η_t for the average cost case is fixed to be η^* in both LSTD(λ) and TD(λ) for simplicity.) Konda shows [13, Theorem 6.3] that the covariance matrix Σ_0 in the preceding proposition is given by $\Sigma_0 = A^{-1}\Gamma(A')^{-1}$, where Γ is the covariance matrix of the Gaussian distribution to which $\sqrt{t}\,(\bar{A}_t r^* + \bar{b}_t)$ converges in distribution. As Konda also shows [13, Theorem 6.1], LSTD(λ) has the asymptotically optimal convergence rate compared to other recursive TD(λ) algorithms (the ones analyzed in [6] and [7]), whose updates $\tilde{r}_t$ have the form

$$\tilde{r}_{t+1} = \tilde{r}_t + \gamma_t z_t d_t,$$

where

$$d_t = g(x_t, x_{t+1}) - \eta_t + (\phi(x_{t+1})' - \phi(x_t)')\tilde{r}_t$$

for the average cost case, and

$$d_t = g(x_t, x_{t+1}) + (\beta\phi(x_{t+1})' - \phi(x_t)')\tilde{r}_t$$

for the β-discounted case. The convergence rate of LSTD(λ) is asymptotically optimal in the following sense. Suppose that $\gamma_t^{-1/2}(\tilde{r}_t - r^*)$ converges in distribution to $\mathcal{N}(0, \Sigma)$, (which can be shown under common assumptions – see [13, Theorem 6.1] – for analyzing asymptotic Gaussian approximations for iterative methods), and also suppose that the limit $\bar{\gamma} = \lim_{t\to\infty}(\gamma_{t+1}^{-1} - \gamma_t^{-1})$ is well defined. Then, the covariance matrix Σ of the limiting Gaussian distribution is such that $\Sigma - \bar{\gamma}\,\Sigma_0$ is positive semidefinite. (In particular, this means that if $\gamma_t = \frac{1}{ct}$, where c is a constant scalar, then $\bar{\gamma} = c$ and $\sqrt{t}(\tilde{r}_t - r^*)$ converges in distribution to $\mathcal{N}(0, \frac{\Sigma}{c})$, where $\frac{\Sigma}{c} - \Sigma_0$ is positive semidefinite.)

Remark 5. We have proved that LSPE(λ) with *any* constant stepsize (under which LSPE(λ) converges) has the same asymptotic optimal convergence rate as LSTD(λ), i.e., the convergence rate of LSPE(λ) does not depend on the constant stepsize. Essentially, the LSPE(λ) iterate r_t tracks the LSTD(λ) iterate $\hat{r}_t$ at the rate of $O(t)$ regardless of the value of the stepsize (see Prop. 5 and its proof), while the LSTD(λ) update converges to r^* at the slower rate of $O(\sqrt{t})$. This explains why the constant stepsize does not affect the asymptotic convergence rate of LSPE(λ). On the other hand, the stepsize γ affects the spectral radius of the matrix $(I + \gamma B^{-1}A)$ and the corresponding scalar ρ (see the proof of Prop. 5), and therefore also the (geometric) rate at which $\|r_t - \hat{r}_t\|_w$, the distance between the LSPE(λ) and LSTD(λ) iterates, converges to 0. This can also be observed from the computational results of the next section.

Remark 6. Similar to the argument in Remark 2, our convergence rate results Props. 5 and 6 extend to LSPE(λ) with a time varying stepsize γ_t, where γ_t for all t lies in a closed interval contained in the range of stepsizes given by Condition 1. This can be seen by noticing that the norm $\|\cdot\|_w$ in the proof of Prop. 5 can be chosen to be the same for all γ in the above closed interval.

5 Computational Experiments

The following experiments on three examples show that

- LSPE(λ) and LSTD(λ) converge to each other faster than to the common limit, and
- the algorithm of recursive TD(λ) with Polyak averaging, which theoretically also has asymptotically optimal convergence rate (cf. Konda [13]), does not seem to scale well with the problem size.

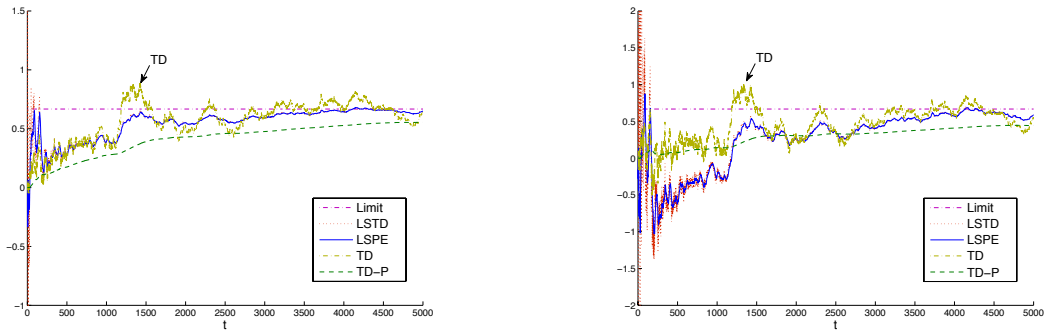
Here are a few details of the three algorithms used in experiments. We use pseudoinverse for matrix inversions in LSPE(λ) and LSTD(λ) at the beginning stages, when matrices tend to be singular. The stepsize γ in LSPE(λ) is taken to be 1, except when noted. Recursive TD(λ) algorithms tend to diverge during early stages, so we truncate the components of their updates $\tilde{r}_t$ to be within the range $[-1000, 1000]$. The TD(λ) algorithm with Polyak averaging, works as follows. The stepsizes γ_t of TD(λ) are taken to be an order of magnitude greater than $1/t$, $\gamma_t = 1/t^{0.8}$ in our experiments. The updates $\tilde{r}_t$ of TD(λ) are then averaged over time to have $\frac{1}{t+1}\sum_{i=0}^{t}\tilde{r}_i$ as the updates of the Polyak averaging algorithm. (For a general reference on Polyak averaging, see e.g., Kushner and Yin [23].)

In all the following figures, the horizontal axes index the time in the LSPE(λ), LSTD(λ), and TD(λ) iterations, which use the same single sample trajectory.

Example 1. This is a 2-state toy example. The parameters are:

$$P = \begin{bmatrix} 0.2 & 0.8 \\ 0.7 & 0.3 \end{bmatrix}, \quad g(1,j) = 1, \quad g(2,j) = 2, \quad j = 1, 2.$$

We use one basis function: $\Phi = \begin{bmatrix} 1 & 2 \end{bmatrix}'$. The updates of LSPE($\lambda$), LSTD($\lambda$), TD($\lambda$), and TD($\lambda$) with Polyak averaging are thus one dimensional scalars. The results are given in Fig. 1. □

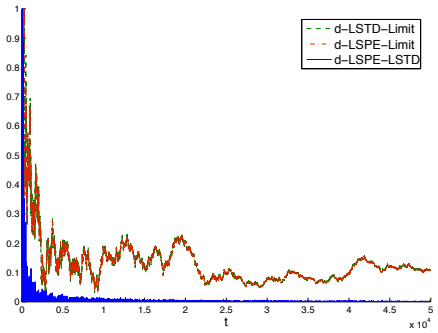


(a) $\lambda = 0.7$ (b) $\lambda = 0.9$

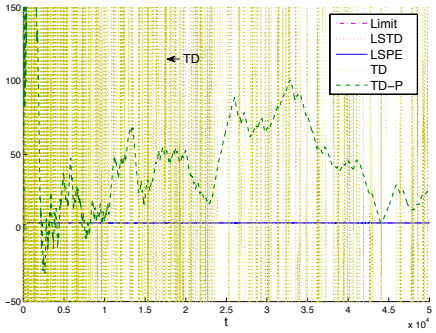
Figure 1: Computational results obtained for Example 1. Graphs of updates of average cost LSPE(λ), LSTD(λ), TD(λ), and TD(λ) with Polyak averaging (TD-P) using the same single trajectory and for different values of λ. At the scale used, LSPE(λ) and LSTD(λ) almost coincide with each other. The behavior of TD(λ) with Polyak averaging conforms with the theoretical analysis in this case.

Example 2. This example is a randomly generated fast-mixing Markov chain with 100 states indexed by 1 to 100. The state transition probability matrix is

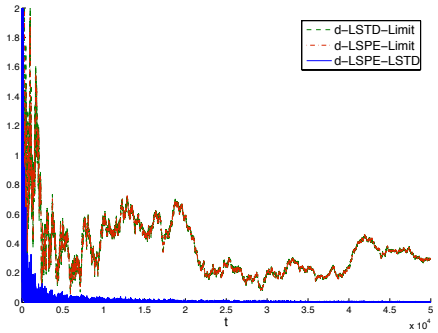
$$P = 0.1I + 0.9R,$$



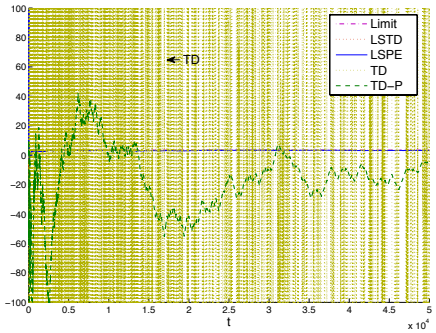
(a) Distance plots for $\lambda = 0.7$



(b) Updates for $\lambda = 0.7$



(c) Distance plots for $\lambda = 0.9$



(d) Updates for $\lambda = 0.9$

Figure 2: Computational results obtained for Example 2. Graphs of distances and updates of the TD algorithms using the same single trajectory and for different values of λ. Only the parts within the range of the vertical axis are shown. (a) and (c): Distances between LSPE(λ) and LSTD(λ) (d-LSPE-LSTD), between LSPE(λ) and the limit (d-LSPE-Limit), and between LSTD(λ) and the limit (d-LSTD-Limit). LSPE(λ) and LSTD(λ) are at all times much closer to each other than to the limit. (b) and (d): Graphs of one of the components of the updates of LSPE(λ), LSTD(λ), TD(λ), and TD(λ) with Polyak averaging (TD-P). We were not able to get TD(λ) to converge in this case.

where I is the identity matrix, and R is a random stochastic matrix with mutually independent rows which are uniformly distributed in the space of probability distributions over the state space. The per-stage costs are

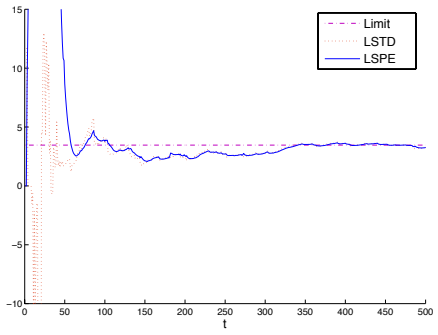
$$g(i,j) = \begin{cases} rand, & i < 90, \\ i/30 + rand, & i = 90, \ldots, 100, \end{cases}$$

where *rand* denotes a random number uniform in $[0, 1]$ and independently generated for each i. We use 3 basis functions in the average cost case.

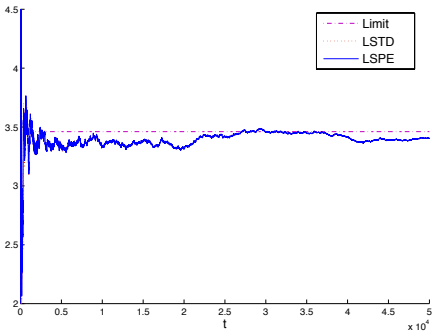
Even though the chain mixes rapidly, because of the cost structure, it is not an easy case for the recursive TD(λ) algorithm. The results are given in Figs. 2 and 3. □

Example 3. This example is a 100-state Markov chain that has a random walk structure and a slow mixing rate relative to the previous example. Using $P(j|i)$ as a shorthand for $p(X_1 = j \mid X_0 = i)$, we let the state transition probabilities be

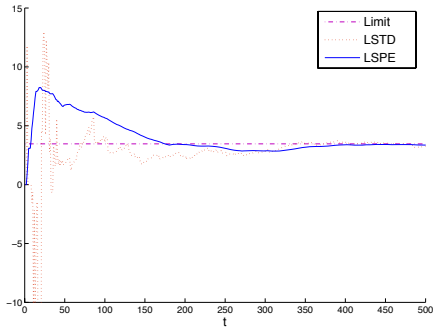
$$P(i|i) = 0.1, \quad P(i+1|i) = P(i-1|i) = 0.45, \quad i = 2, \ldots, 99,$$



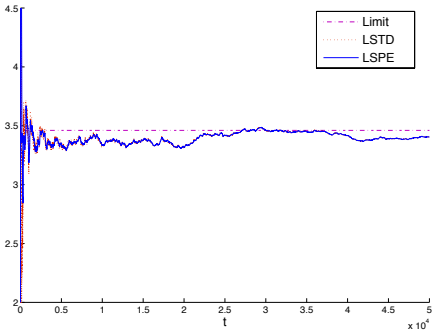
(a) $\lambda = 0.7, \gamma = 1$, Initial iterations



(b) $\lambda = 0.7, \gamma = 1$



(c) $\lambda = 0.7, \gamma = 0.1$, initial iterations



(d) $\lambda = 0.7, \gamma = 0.1$

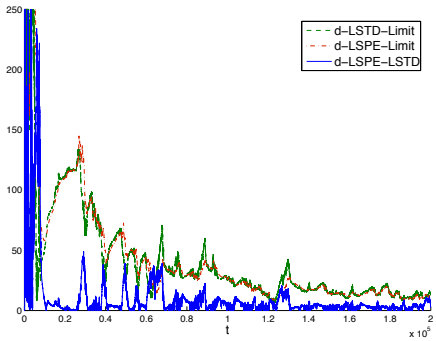
Figure 3: Comparison LSTD(λ) and LSPE(λ) with different constant stepsizes γ for Example 2. Plotted are one of the components of the updates of LSPE(λ) and LSTD(λ).

$P(1|1) = 0.1$, $P(2|1) = 0.9$, $P(100|100) = 0.1$, and $P(99|100) = 0.9$. The per-stage costs are the same as in Example 2, and so are the basis functions. The results are given in Figs. 4 and 5. □

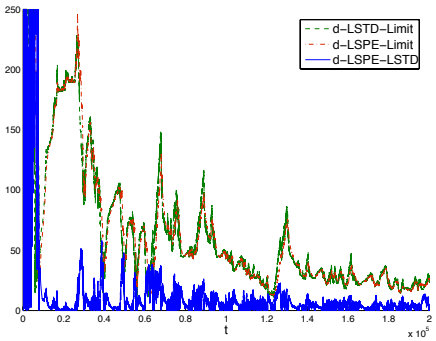
6 Extensions to Multiple Policies and Policy Iteration

In this section, we discuss various uses and extensions of LSPE(λ) for the more general MDP problem that involves optimization over multiple policies (as opposed to just a single policy as we have assumed so far). The main difficulty here is that when function approximation is introduced, the contraction properties that are inherent in the single policy evaluation case are lost. In particular, the corresponding projected Bellman equation (which is now nonlinear) may have multiple fixed points or none at all (see De Farias and Van Roy [24]). As a result the development of LSPE-type algorithms with solid convergence properties becomes very difficult.

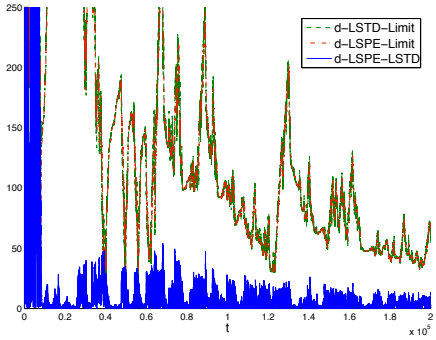
However, there is one important class of MDP for which the aforementioned difficulties largely disappear, because the corresponding (nonlinear) projected Bellman equation involves a contraction mapping under certain conditions. This is the class of discounted optimal stopping problems, for which Tsitsiklis and Van Roy [25] have shown the contraction property and analyzed the application of TD(0). It can be shown that LSPE(0) can also be applied to such problems, and its convergence properties can be analyzed using appropriate extensions of the methods of the present paper. Note that the deterministic portion of the iteration here involves a nonlinear contraction mapping. Because of this nonlinearity, the least squares problem corresponding to LSTD(0) is not easy to solve and thus LSTD(λ) is not easy to apply. This analysis is reported elsewhere (see Yu



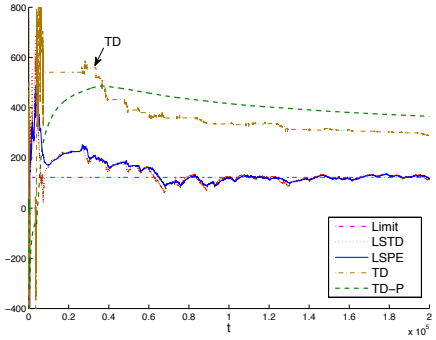
(a) Distance plots for $\lambda = 0.7$



(b) Distance plots for $\lambda = 0.9$



(c) Distance plots for $\lambda = 0.98$



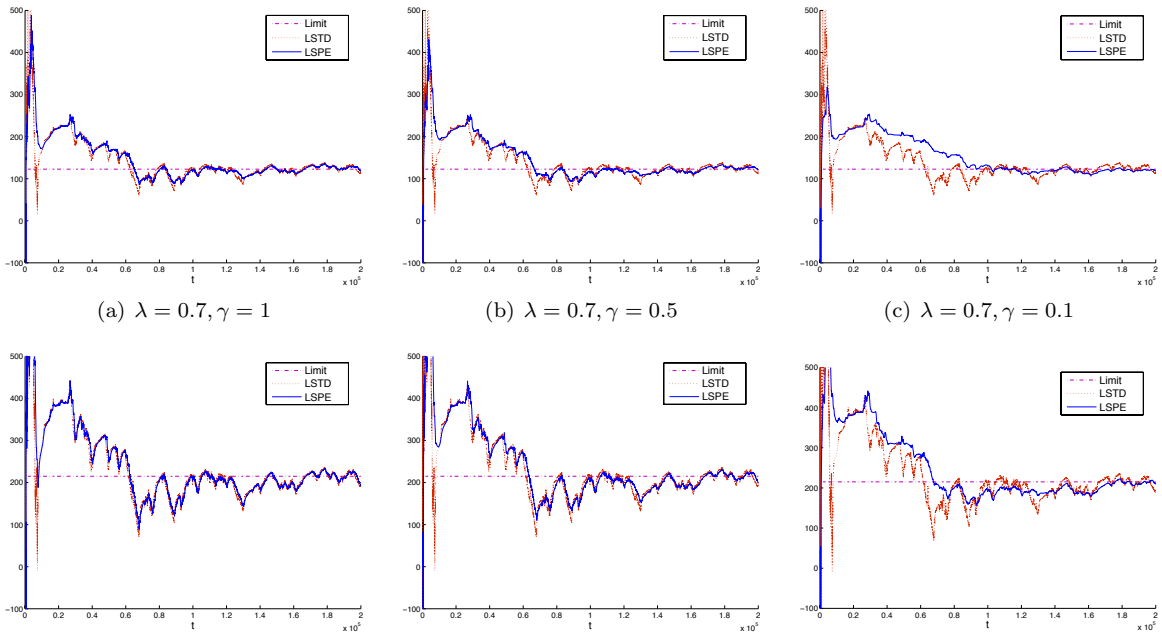
(d) Updates for $\lambda = 0.7$

Figure 4: Computational results obtained for Example 3. Graphs of distances and updates of the TD algorithms using the same single trajectory and for different values of λ. Only the parts within the range of the vertical axis are shown. (a), (b) and (c): Distances between LSPE(λ) and LSTD(λ) (d-LSPE-LSTD), between LSPE(λ) and the limit (d-LSPE-Limit), and between LSTD(λ) and the limit (d-LSTD-Limit). LSPE(λ) and LSTD(λ) are closer to each other than to the limit for most of the time. (d): Graphs of one of the components of the updates of LSPE(λ), LSTD(λ), TD(λ), and TD(λ) with Polyak averaging (TD-P). The convergence of the recursive TD(λ) (hence also that of the Polyak averaging) is much slower than LSPE(λ) and LSTD(λ) in this case.

and Bertsekas [26, 27]).

Let us now consider the use of LSPE(λ) and LSTD(λ) in the context of approximate policy iteration. Here, multiple policies are generated, each obtained by policy improvement using the approximate cost function or Q-function of the preceding policy, which in turn may be obtained by using simulation and LSPE(λ) or LSTD(λ). This context is central in approximate DP, and has been discussed extensively in various sources, such as the books by Bertsekas and Tsitsiklis [14], and Sutton and Barto [19]. Lagoudakis and Parr [15] discuss LSTD(λ) in the context of approximate policy iteration and discounted problems, and report favorable computational results. The use of LSPE(λ) in the context of approximate policy iteration was proposed in the original paper by Bertsekas and Ioffe [12], under the name λ-policy iteration, and favorable results were reported in the context of a challenging tetris training problem, which could not be solved using TD(λ).

Generally, one may distinguish between two types of policy iteration: (1) *regular* where each policy evaluation is done with a long simulation in order to achieve the maximum feasible policy evaluation accuracy before switching to a new policy via policy improvement, and (2) *optimistic* where each policy evaluation is done inaccurately, using a few simulation samples (sometimes only



(a) $\lambda = 0.7, \gamma = 1$ (b) $\lambda = 0.7, \gamma = 0.5$ (c) $\lambda = 0.7, \gamma = 0.1$

(d) $\lambda = 0.9, \gamma = 1$ (e) $\lambda = 0.9, \gamma = 0.5$ (f) $\lambda = 0.9, \gamma = 0.1$

Figure 5: Comparison of LSTD(λ) and LSPE(λ) with different constant stepsizes γ for Example 3. Plotted are one of the components of the updates of LSPE(λ) and LSTD(λ).

one), before switching to a new policy. The tradeoffs between these two variants are discussed extensively in the literature, with experience tending to favor the optimistic variants. However, the behavior of approximate policy iteration is extremely complicated, as explained for example in Bertsekas and Tsitsiklis [14, Section 6.4], so there is no clear understanding of the circumstances that favor the regular or optimistic versions.

Given our convergence rate analysis, it appears that LSPE(λ) and LSTD(λ) should perform comparably when used for regular policy iteration, since they have an identical asymptotic convergence rate. However, for optimistic policy iteration, the asymptotic convergence rate is not relevant, and the ability to make fast initial progress is most important. Within this context, upon change of a policy, LSPE(λ) may rely on the current iterate r_t for stability, but LSTD(λ) in its pure form may be difficult to stabilize (think of LSTD(λ) within an optimistic policy iteration framework that changes policy after each sample). It is thus interesting to investigate the circumstances in which one method may be having an advantage over the other.

An alternative to the above use of approximate policy iteration in the case of multiple policies is a policy gradient method. Let us outline the use of LSTD(λ) and LSPE(λ) algorithms in the policy gradient method of the actor-critic type, as considered by Konda and Tsitsiklis [1], and Konda [13]. This discussion will also clarify the relation between LSTD(λ)/LSPE(λ) and SA algorithms. Actor-critic algorithms are two-time-scale SA algorithms in which the actor part refers to stochastic gradient descent iterations on the space of policy parameters at the slow time-scale, while the critic part is to estimate/track at the fast time-scale the cost function of the current policy, which can then be used in the actor part for estimating the gradient. Konda and Tsitsiklis [1], and Konda [13] have analyzed this type of algorithms with the critic implemented using TD(λ). When we implement the critic using least squares methods such as LSPE(λ) and LSTD(λ), at the fast time-scale, we track directly the mapping which defines the projected Bellman equation associated

with the current policy. This is to be contrasted with the TD(λ)-critic in which we only track the solution of the projected Bellman equation without estimating the mapping/equation itself.

To make our point more concrete, we consider here the average cost criterion. (Other cost criteria are similar.) We consider randomized policies parametrized by a d-dimensional vector θ, and we view the state-action pairs as the joint state variables. The basis functions, the projected Bellman equation and its solution, as well as the Bellman equation, now depend on θ. We will use subscripts to indicate this dependence. Under certain differentiability conditions, the gradient of the average cost $\nabla\eta^*(\theta)$ can be expressed as (see e.g., Konda and Tsitsiklis [1], Konda [13, Chapter 2.3])

$$\nabla\eta^*(\theta) = Q'_\theta D_\theta \Psi_\theta = (\Pi_\theta Q_\theta)' D_\theta \Psi_\theta,$$

where Q_θ is the Q-factor, or equivalently, the bias function of the MDP on the joint state-action space, D_θ is as before the diagonal matrix with the invariant distribution π_θ of the Markov chain on its diagonal, Ψ_θ is an $n\times d$ matrix whose columns consist of a certain set of basis functions determined by θ, and Π_θ is the projection on a certain subspace $\mathrm{col}(\Phi_\theta)$ such that $\mathrm{col}(\Phi_\theta) \supseteq \mathrm{col}(\Psi_\theta)$. We consider one variant of the actor-critic algorithm, (the idea that follows applies similarly to other variants), in which the critic approximates the projection $\Pi_\theta Q_\theta$ by $\Phi_\theta r^*_\theta$, the solution of the projected Bellman equation $J = \Pi_\theta T^{(\lambda)}_\theta J$, and then uses it to approximate the gradient:

$$\nabla\eta^*(\theta) \approx (\Phi_\theta r^*_\theta)' D_\theta \Psi_\theta.$$

This is biased estimation, with the bias diminishing as λ tends to 1 or as the subspace $\mathrm{col}(\Phi_\theta)$ is enlarged.

When the critic is implemented using LSTD(λ) or LSPE(λ), the actor part has the form of a stochastic gradient descent iteration, as with the TD(λ)-critic:

$$\theta_t = \theta_{t-1} - \alpha_t s_t, \tag{38}$$

where α_t is a stepsize and s_t is an estimate of $\nabla\eta^*(\theta_t)$, while gradient estimation can be done as follows. Let $(x_0, x_1, \ldots, x_t, \ldots)$ be a single infinitely long simulation trajectory with x_t being the state-action at time t. Omitting the explicit dependence on θ of various quantities such as g and ϕ for notational simplicity, we define iterations

$$\eta_t = (1-\delta_t)\eta_{t-1} + \delta_t g(x_t, x_{t+1}), \tag{39}$$
$$q_t = (1-\delta_t) q_{t-1} + \delta_t \phi(x_t)\psi(x_t)', \tag{40}$$

and

$$z_t = \lambda z_{t-1} + \phi(x_t), \tag{41}$$
$$\bar{b}_t = (1-\delta_t)\bar{b}_{t-1} + \delta_t z_t\big(g(x_t, x_{t+1}) - \eta_t\big) \tag{42}$$
$$\bar{B}_t = (1-\delta_t)\bar{B}_{t-1} + \delta_t \phi(x_t)\,\phi(x_t)', \tag{43}$$
$$\bar{A}_t = (1-\delta_t)\bar{A}_{t-1} + \delta_t z_t\big(\phi(x_{t+1})' - \phi(x_t)'\big), \tag{44}$$

[cf. (18)-(21) for LSPE(λ) under a single policy]. In the above, δ_t is a stepsize that satisfies the standard conditions $\sum_{t=0}^\infty \delta_t = \infty, \sum_{t=0}^\infty \delta_t^2 < \infty$, as well as the additional eventually non-increasing condition: $\delta_t \le \delta_{t-1}$ for t sufficiently large. Furthermore, the stepsizes α_t and δ_t satisfy $\sum_{t=0}^\infty \alpha_t = \infty, \sum_{t=0}^\infty \alpha_t^2 < \infty$, and

$$\lim_{t\to\infty} \frac{\alpha_t}{\delta_t} \to 0,$$

which makes θ_t evolve at a slower time-scale than the iterates (39)-(40) and (42)-(44), which use δ_t as the stepsize. Possible choices of such sequences are $\alpha_t = \frac{1}{1+t\ln t}$ and $\delta_t = \frac{1}{t+1}$, or $\alpha_t = \frac{1}{t+1}$ and

$\delta_t = t^{-c}$ with $c \in (1/2, 1)$. The latter is indeed preferred, as it makes the estimates depend "less" on the data from the remote past. We let r_t be updated either by LSTD(λ) or LSPE(λ) with a constant stepsize $\gamma \in (0, 1]$ as given in the present paper, i.e.,

$$r_{t+1} = -\bar{A}_t^{-1}\bar{b}_t, \quad \text{or} \quad r_{t+1} = r_t + \gamma \bar{B}_t^{-1}\big(\bar{A}_t^{-1} r_t + \bar{b}_t\big).$$

Then, under standard conditions (which involve the boundedness of θ_t and s_t, the smoothness of $\eta^*(\theta)$, Ψ_θ, and Φ_θ), viewing z_t as part of the Markov process (x_t, x_{t+1}, z_t), one can apply the results of Borkar [28] and [29, Chapter 6] to show that θ_t can be viewed as "quasi-static" for the iterates in (39)-(40) and (42)-(44). In particular, the latter iterates track the respective quantities associated with θ_t:

$$\begin{aligned} &\eta_t \approx \eta^*(\theta_t), \quad q_t \approx \Phi'_{\theta_t} D_{\theta_t} \Psi_{\theta_t}, \quad \bar{b}_t \approx b_{\theta_t}, \\ &\bar{B}_t \approx B_{\theta_t}, \quad \bar{A}_t \approx A_{\theta_t}, \end{aligned}$$

with the differences between the two sides asymptotically diminishing as $t \to \infty$. In the above, note particularly that $b_{\theta_t}, B_{\theta_t}, A_{\theta_t}$ together with Φ_{θ_t} define the projected Bellman equation and its associated mapping $\Pi_{\theta_t} T^{(\lambda)}_{\theta_t}$ at θ_t, therefore the iterates $\bar{b}_t, \bar{B}_t, \bar{A}_t$ track the projected Bellman equation/mapping associated with θ_t. From this one can further show (under a uniform contraction condition such as $\sup_t \|(1-\gamma)I + \gamma \Pi_{\theta_t} T^{(\lambda)}_{\theta_t}\|_{\pi_{\theta_t}} < 1$ in the case of LSPE(λ)) that r_t tracks $r^*_{\theta_t}$:

$$r_t \approx r^*_{\theta_t},$$

and hence $r'_t q_t$ tracks the approximating gradient:

$$r'_t q_t \approx (\Phi_{\theta_t} r^*_{\theta_t})' D_{\theta_t} \Psi_{\theta_t},$$

with asymptotically diminishing differences. In the actor's iteration (38), one may let $s_t = r'_t q_t$ or let s_t be a bounded version of $r'_t q_t$. The limiting behavior of θ_t can then be analyzed following standard methods.

7 Concluding Remarks

In this paper, we introduced an average cost version of the LSPE(λ) algorithm, and we proved its convergence for any $\lambda \in (0, 1)$ and any constant stepsize $\gamma \in (0, 1]$, as well as for $\lambda = 0$ and $\gamma \in (0, 1)$. We then proved the optimal convergence rate of LSPE(λ) with a constant stepsize for both the discounted and average cost cases. The analysis and computational experiments also show that LSPE(λ) and LSTD(λ) converge to each other at a faster scale than they converge to the common limit.

Our algorithm and analysis apply not only to a single infinitely long trajectory, but also to multiple infinitely long simulation trajectories. In particular, assuming k trajectories, denoted by $\{x_{j,0}, x_{j,1}, \ldots\}$, $j = 1, \ldots, k$, the least squares problem for LSPE(λ) can be formulated as the minimization of $\sum_{j=1}^k \alpha_j f_{j,t}(r)$, where $f_{j,t}(r)$ is the least squares objective function for the j-th trajectory at time t as in the case of a single trajectory, and α_j is a positive weight on the j-th trajectory, with $\sum_{j=1}^k \alpha_j = 1$. Asymptotically, the algorithm will be speeded up by a factor k at the expense of k times more computation per iteration, so in terms of running time for the same level of error to convergence, the algorithm will be essentially unaffected. On the other hand, we expect that the transient behavior of the algorithm would be significantly improved, especially when the Markov chain has a slow mixing rate. This conjecture, however, is not supported by a quantitative analysis at present.

When the states of the Markov chain form multiple recurrent classes $C_1, \ldots, C_m$, (assuming there are no transient states), it is essential to use multiple simulation trajectories, in order to construct an

approximate cost function that reflects the costs of starting points from different recurrent classes. While there is no unique invariant distribution, the one that relates to our algorithm using multiple trajectories, is $\pi(i) = \pi_k(i) \sum_{\{j|x_{j,0} \in C_k\}} \alpha_j$, $i \in C_k$, where π_k is the unique invariant distribution on the set C_k. Our earlier analysis can be adapted to show for the average cost case that the constant stepsize LSPE(λ) algorithm converges if the basis functions and the eigenvectors of the transition matrix P corresponding to the eigenvalue 1 are linearly independent. The approximate cost function may be combined with the average costs of the recurrent classes (computed separately for each trajectory) to design proper approximate policy iteration schemes in the multi-chain context.

We finally note that in recent work [30], we have extended the linear function approximation framework to the approximate solution of general linear equations (not necessarily related to MDP). Some of the analysis of the present paper is applicable to this more general linear equation context, particularly in connection to rate of convergence and to compositions of projection and nonexpansive mappings.

Acknowledgment

We thank Prof. John Tsitsiklis for helpful discussions. And we thank two reviewers for their suggestions that improved this manuscript. This work was started while Huizhen Yu was a Ph.D. student of the Laboratory for Information and Decision Systems (LIDS), M.I.T., and was finished while she was at the University of Helsinki and supported by the postdoctoral research program. This work was also supported by NSF Grants ECS-0218328 and ECCS-0801549.

References

[1] V. R. Konda and J. N. Tsitsiklis, "Actor-critic algorithms," *SIAM J. Control Optim.*, vol. 42, no. 4, pp. 1143–1166, 2003.

[2] R. S. Sutton, "Learning to predict by the methods of temporal differences," *Machine Learning*, vol. 3, pp. 9–44, 1988.

[3] P. D. Dayan, "The convergence of TD(λ) for general λ," *Machine Learning*, vol. 8, pp. 341–362, 1992.

[4] L. Gurvits, L. J. Lin, and S. J. Hanson, "Incremental learning of evaluation functions for absorbing Markov chains: New methods and theorems," 1994, preprint.

[5] F. Pineda, "Mean-field analysis for batched TD(λ)," *Neural Computation*, pp. 1403–1419, 1997.

[6] J. N. Tsitsiklis and B. Van Roy, "An analysis of temporal-difference learning with function approximation," *IEEE Trans. Automat. Contr.*, vol. 42, no. 5, pp. 674–690, 1997.

[7] ——, "Average cost temporal-difference learning," *Automatica*, vol. 35, no. 11, pp. 1799–1808, 1999.

[8] ——, "On average versus discounted reward temporal-difference learning," *Machine Learning*, vol. 49, pp. 179–191, 2002.

[9] P. Marbach and J. N. Tsitsiklis, "Simulation-based optimization of Markov reward processes," *IEEE Trans. Automatic Control*, vol. 46, no. 2, pp. 191–209, 2001.

[10] S. J. Bradtke and A. G. Barto, "Linear least-squares algorithms for temporal difference learning," *Machine Learning*, vol. 22, no. 2, pp. 33–57, 1996.

[11] J. A. Boyan, "Least-squares temporal difference learning," in *Proc. The 16th Int. Conf. Machine Learning*, 1999.

[12] D. P. Bertsekas and S. Ioffe, "Temporal differences-based policy iteration and applications in neuro-dynamic programming," MIT, LIDS Tech. Report LIDS-P-2349, 1996.

[13] V. R. Konda, "Actor-critic algorithms," Ph.D. dissertation, Dept. Comput. Sci. Elect. Eng., MIT, Cambridge, MA, 2002.

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

[15] M. G. Lagoudakis and R. Parr, "Least-squares policy iteration," *J. Machine Learning Res.*, vol. 4, pp. 1107–1149, 2003.

[16] A. Nedić and D. P. Bertsekas, "Least squares policy evaluation algorithms with linear function approximation," *Discrete Event Dynamic Systems: Theory and Applications*, vol. 13, pp. 79–110, 2003.

[17] D. P. Bertsekas, V. S. Borkar, and A. Nedić, "Improved temporal difference methods with linear function approximation," MIT, LIDS Tech. Report 2573, 2003, also appears in "Learning and Approximate Dynamic Programming," by A. Barto, W. Powell, J. Si, (Eds.), IEEE Press, 2004.

[18] H. Yu, "A function approximation approach to estimation of policy gradient for POMDP with structured polices," in *Proc. The 21st Conf. Uncertainty in Artificial Intelligence*, 2005.

[19] R. S. Sutton and A. G. Barto, *Reinforcement Learning*. Cambridge, MA: MIT Press, 1998.

[20] H. Yu and D. P. Bertsekas, "New error bounds for approximations from projected linear equations," *Math. Oper. Res.*, 2008, submitted for publication; also as Univ. Helsinki, Technical Report C-2008-43.

[21] J. M. Ortega and W. C. Rheinboldt, *Iterative Solution of Nonlinear Equations in Several Variables*. New York: Academic Press, 1970.

[22] M. Duflo, *Random Iterative Models*. Berlin: Springer-Verlag, 1997.

[23] H. J. Kushner and G. G. Yin, *Stochastic Approximation and Recursive Algorithms and Applications*, 2nd ed. New York: Springer-Verlag, 2003.

[24] D. P. de Farias and B. Van Roy, "On the existence of fixed points for approximate value iteration and temporal-difference learning," *J. Optim. Theory Appl.*, vol. 105, no. 3, 2000.

[25] J. N. Tsitsiklis and B. Van Roy, "Optimal stopping of Markov processes: Hilbert space theory, approximation algorithms, and an application to pricing financial derivatives," *IEEE Trans. Automat. Contr.*, vol. 44, pp. 1840–1851, 1999.

[26] H. Yu and D. P. Bertsekas, "A least squares Q-learning algorithm for optimal stopping problems," MIT, LIDS Tech. Report 2731, 2006.

[27] ——, "Q-learning algorithms for optimal stopping based on least squares," in *Proc. The European Control Conf.*, 2007.

[28] V. S. Borkar, "Stochastic approximation with 'controlled Markov' noise," *Systems Control Lett.*, vol. 55, pp. 139–145, 2006.

[29] ——, *Stochastic Approximation: A Dynamic Viewpoint*. New Delhi: Hindustan Book Agency, 2008.

[30] D. P. Bertsekas and H. Yu, "Projected equation methods for approximate solution of large linear systems," *J. Comput. Sci. Appl. Math.*, 2008, to be published.