December 2006

Revised: June 2007

A Least Squares Q-Learning Algorithm for Optimal Stopping Problems

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

Dimitri P. Bertsekas[†]
dimitrib@mit.edu

Abstract

We consider the solution of discounted optimal stopping problems using linear function approximation methods. A Q-learning algorithm for such problems, proposed by Tsitsiklis and Van Roy, is based on the method of temporal differences and stochastic approximation. We propose alternative algorithms, which are based on projected value iteration ideas and least squares. We prove the convergence of some of these algorithms and discuss their properties.

[*] Huizhen Yu is with HIIT, University of Helsinki, Finland.

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

1 Introduction

Optimal stopping problems are a special case of Markovian decision problems where the system evolves according to a discrete-time stochastic system equation, until an explicit stopping action is taken. At each state, there are two choices: either to stop and incur a state-dependent stopping cost, or to continue and move to a successor state according to some transition probabilities and incur a state-dependent continuation cost. Once the stopping action is taken, no further costs are incurred. The objective is to minimize the expected value of the total discounted cost. Examples are classical problems, such as search, and sequential hypothesis testing, as well as recent applications in finance and the pricing of derivative financial instruments (see Tsitsiklis and Van Roy [TV99], Barraquand and Martineau [BM95], Longstaff and Schwartz [LS01]).

The problem can be solved in principle by dynamic programming (DP for short), but we are interested in problems with large state spaces where the DP solution is practically infeasible. It is then natural to consider approximate DP techniques where the optimal cost function or the Q-factors of the problem are approximated with a function from a chosen parametric class. Generally, cost function approximation methods are theoretically sound (i.e., are provably convergent) only for the single-policy case, where the cost function of a fixed stationary policy is evaluated. However, for the stopping problem of this paper, Tsitsiklis and Van Roy [TV99] introduced a linear function approximation to the optimal Q-factors, which they prove to be the unique solution of a projected form of Bellman's equation. While in general this equation may not have a solution, this difficulty does not occur in optimal stopping problems thanks to a critical fact: the mapping defining the Q-factors is a contraction mapping with respect to the weighted Euclidean norm corresponding to the steady-state distribution of the associated Markov chain. For textbook analyses, we refer to Bertsekas and Tsitsiklis [BT96], Section 6.8, and Bertsekas [Ber07], Section 6.4.

The algorithm of Tsitsiklis and Van Roy is based on single trajectory simulation, and ideas related to the temporal differences method of Sutton [Sut88], and relies on the contraction property just mentioned. We propose a new algorithm, which is also based on single trajectory simulation and relies on the same contraction property, but uses different algorithmic ideas. It may be viewed as a fixed point iteration for solving the projected Bellman equation, and it relates to the least squares policy evaluation (LSPE) method first proposed by Bertsekas and Ioffe [BI96] and subsequently developed by Nedić and Bertsekas [NB03], Bertsekas, Borkar, and Nedić [NB03], and Yu and Bertsekas [YB06] (see also the books [BT96] and [Ber07]). We prove the convergence of our method for finite-state models, and we discuss some variants.

The paper is organized as follows. In Section 2, we introduce the optimal stopping problem, and we derive the associated contraction properties of the mapping that defines Q-learning. In Section 3, we describe our LSPE-like algorithm, and we prove its convergence. We also discuss the convergence rate of the algorithm, and we provide a comparison with another algorithm that is related to the least squares temporal differences (LSTD) method, proposed by Bradtke and Barto [BB96], and further developed by Boyan [Boy99]. In Section 4, we describe some variants of the algorithm, which involve a reduced computational overhead per iteration. In this section, we also discuss the relation of our algorithms with the recent algorithm by Choi and Van Roy [CV06], which can be used to solve the same optimal stopping problem. In Section 5, we prove the convergence of some of the variants of Section 4. We give two alternative proofs, the first of which uses results from the o.d.e. (ordinary differential equation) line of convergence analysis of stochastic iterative algorithms, and the second of which is a "direct" proof reminiscent of the o.d.e. line of analysis. A computational comparison of our methods with other algorithms for the optimal stopping problem is beyond the scope of the present paper. However, our analysis and the available results using least squares methods (Bradtke and Barto [BB96], Bertsekas and Ioffe [BI96], Boyan [Boy99], Bertsekas, Borkar, and Nedić [BBN03], Choi and Van Roy [CV06]) clearly suggest a superior performance to the algorithm of Tsitsiklis and Van Roy [TV99], and likely an improved convergence rate over the method of Choi and Van Roy [CV06], at the expense of some additional overhead per iteration.

2 Q-Learning for Optimal Stopping Problems

We are given a Markov chain with state space $\{1, \dots, n\}$, described by transition probabilities p_{ij}. We assume that the states form a single recurrent class, so the chain has a steady-state distribution vector $\pi = \big(\pi(1), \dots, \pi(n)\big)$ with $\pi(i) > 0$ for all states i. Given the current state i, we assume that we have two options: to stop and incur a cost $c(i)$, or to continue and incur a cost $g(i,j)$, where j is the next state (there is no control to affect the corresponding transition probabilities). The problem is to minimize the associated α-discounted infinite horizon cost, where $\alpha \in (0,1)$.

For a given state i, we associate a Q-factor with each of the two possible decisions. The Q-factor for the decision to stop is equal to $c(i)$. The Q-factor for the decision to continue is denoted by $Q(i)$. The optimal Q-factor for the decision to continue, denoted by Q^*, relates to the optimal cost function J^* of the stopping problem by

$$Q^*(i) = \sum_{j=1}^{n} p_{ij}\big(g(i,j) + \alpha J^*(j)\big), \qquad i = 1, \dots, n,$$

and

$$J^*(i) = \min\big\{c(i), Q^*(i)\big\}, \qquad i = 1, \dots, n.$$

The value $Q^*(i)$ is equal to the cost of choosing to continue at the initial state i and following an optimal policy afterwards. The function Q^* satisfies Bellman's equation

$$Q^*(i) = \sum_{j=1}^{n} p_{ij}\Big(g(i,j) + \alpha \min\big\{c(j), Q^*(j)\big\}\Big), \qquad i = 1, \dots, n. \tag{1}$$

Once the Q-factors $Q^*(i)$ are calculated, an optimal policy can be implemented by stopping at state i if and only if $c(i) \leq Q^*(i)$.

The Q-learning algorithm (Watkins [Wat89]) is

$$Q(i) := Q(i) + \gamma\big(g(i,j) + \alpha \min\big\{c(j), Q(j)\big\} - Q(i)\big),$$

where i is the state at which we update the Q-factor, j is a successor state, generated randomly according to the transition probabilities p_{ij}, and γ is a small positive stepsize, which diminishes to 0 over time. The convergence of this algorithm is addressed by the general theory of Q-learning (see Watkins and Dayan [WD92], and Tsitsiklis [Tsi94]). However, for problems where the number of states n is large, this algorithm is impractical.

Let us now consider the approximate evaluation of $Q^*(i)$. We introduce the mapping $F : \Re^n \mapsto \Re^n$ given by

$$(FQ)(i) = \sum_{j=1}^{n} p_{ij}\big(g(i,j) + \alpha \min\big\{c(j), Q(j)\big\}\big), \qquad i = 1, \dots, n.$$

We denote by FQ or $F(Q)$ the vector whose components are $(FQ)(i)$, $i = 1, \dots, n$. By Eq. (1), the optimal Q-factor for the choice to continue, Q^*, is a fixed point of F, and it is the unique fixed point because F is a sup-norm contraction mapping.

For the approximation considered here, it turns out to be very important that F is also a Euclidean contraction. Let $\|\cdot\|_\pi$ be the weighted Euclidean norm associated with the steady-state probability vector π, i.e.,

$$\|v\|_\pi^2 = \sum_{i=1}^{n} \pi(i)\big(v(i)\big)^2.$$

It has been shown by Tsitsiklis and Van Roy [TV99] (see also Bertsekas and Tsitsiklis [BT96], Section 6.8.4) that F is a contraction with respect to this norm. For purposes of easy reference, we include the proof.

Lemma 1. *The mapping F is a contraction with respect to $\|\cdot\|_\pi$, with modulus α.*

Proof. For any two vectors Q and $\overline{Q}$, we have

$$\begin{aligned}\big|(FQ)(i) - (F\overline{Q})(i)\big| &\le \alpha \sum_{j=1}^{n} p_{ij}\big|\min\big\{c(j), Q(j)\big\} - \min\big\{c(j), \overline{Q}(j)\big\}\big| \\ &\le \alpha \sum_{j=1}^{n} p_{ij}\big|Q(j) - \overline{Q}(j)\big|,\end{aligned}$$

or, in vector notation,

$$|FQ - F\overline{Q}| \le \alpha P|Q - \overline{Q}|,$$

where $|x|$ denotes a vector whose components are the absolute values of the components of x. Hence,

$$\|FQ - F\overline{Q}\|_\pi \le \alpha\big\|P|Q - \overline{Q}|\,\big\|_\pi \le \alpha\|Q - \overline{Q}\|_\pi,$$

where the last inequality follows from the relation $\|PJ\|_\pi \le \|J\|_\pi$, which holds for every vector J (see Tsitsiklis and Van Roy [TV97] or Bertsekas and Tsitsiklis [BT96], Lemma 6.4). □

We consider Q-factor approximations using a linear approximation architecture

$$\tilde{Q}(i, r) = \phi(i)'r,$$

where $\phi(i)$ is an s-dimensional feature vector associated with state i. (In our notation, all vectors are viewed as column vectors, and prime denotes transposition.) We also write the vector

$$\tilde{Q}_r = \big(\tilde{Q}(1, r), \ldots, \tilde{Q}(n, r)\big)'$$

in the compact form

$$\tilde{Q}_r = \Phi r,$$

where Φ is the $n \times s$ matrix whose rows are $\phi(i)'$, $i = 1, \ldots, n$. We assume that Φ has rank s, and we denote by Π the projection mapping with respect to $\|\cdot\|_\pi$ on the subspace

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

i.e., for all $J \in \Re^n$,

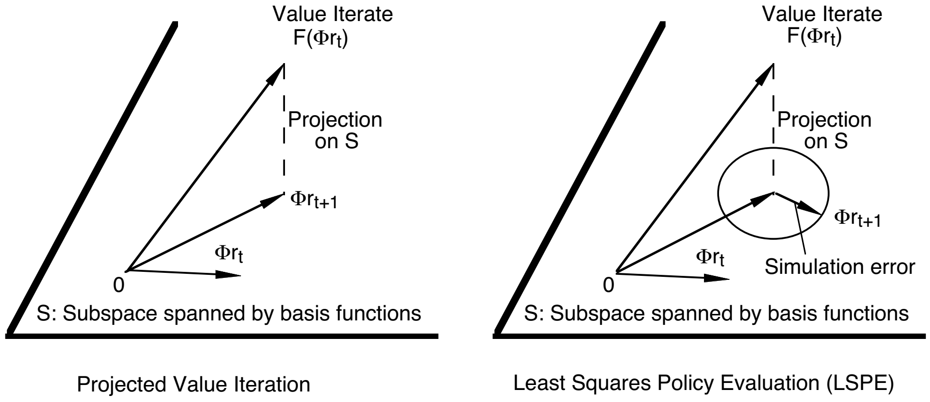
$$\Pi J = \arg\min_{\hat{J} \in S} \|J - \hat{J}\|_\pi.$$

Because F is a contraction with respect to $\|\cdot\|_\pi$ with modulus α, and Π is nonexpansive, the mapping ΠF is a contraction with respect to $\|\cdot\|_\pi$ with modulus α. Therefore, the mapping ΠF has a unique fixed point within the subspace S, which (in view of the rank assumption on Φ) can be uniquely represented as Φr^*. Thus r^* is the unique solution of the equation

$$\Phi r^* = \Pi F(\Phi r^*).$$

Tsitsiklis and Van Roy [TV99] show that the error of this Q-factor approximation can be bounded by

$$\|\Phi r^* - Q^*\|_\pi \le \frac{1}{\sqrt{1 - \alpha^2}}\|\Pi Q^* - Q^*\|_\pi.$$



Projected Value Iteration

Least Squares Policy Evaluation (LSPE)

Figure 1: A conceptual view of projected value iteration and its simulation-based implementation.

Furthermore, if we implement a policy μ that stops at state i if and only if $c(i) \leq \phi(i)'r^*$, then the cost of this policy, denoted by J_μ, satisfies

$$\sum_{i=1}^{n} \pi(i)\big(J_\mu(i) - J^*(i)\big) \leq \frac{2}{(1-\alpha)\sqrt{1-\alpha^2}} \|\Pi Q^* - Q^*\|_\pi.$$

These bounds indicate that if Q^* is close to the subspace S spanned by the basis functions, then the approximate Q-factor and its associated policy will also be close to the optimal.

The contraction property of ΠF suggests the fixed point iteration

$$\Phi r_{t+1} = \Pi F(\Phi r_t),$$

which in the related contexts of policy evaluation for discounted and average cost problems (see [BBN03, YB06, Ber07]) is known as *projected value iteration* [to distinguish it from the value iteration method, which is $Q_{t+1} = F(Q_t)$]; see Fig. 1. This iteration converges to the unique fixed point Φr^* of ΠF, but is not easily implemented because the dimension of the vector $F(\Phi r_t)$ is potentially very large. In the policy evaluation context, a simulation-based implementation of the iteration has been proposed, which does not suffer from this difficulty, because it uses simulation samples of the cost of various states in a least-squares type of parametric approximation of the value iteration method. This algorithm is known as least squares policy evaluation (LSPE), and can be conceptually viewed as taking the form

$$\Phi r_{t+1} = \Pi F(\Phi r_t) + \epsilon_t,$$

where ϵ_t is simulation noise which diminishes to 0 with probability 1 (w.p.1) as $t \to \infty$ (see Fig. 1). The algorithm to be introduced in the next section admits a similar conceptual interpretation, and its analysis has much in common with the analysis given in [BBN03, Ber07] for the case of single-policy evaluation. In fact, if the stopping option was not available [or equivalently if $c(i)$ is so high that it is never optimal to stop], our Q-learning algorithm would coincide with the LSPE algorithm for approximate evaluation of the discounted cost function of a fixed stationary policy. Let us also note that LSPE (like the temporal differences method) is actually a family of methods parameterized by a scalar $\lambda \in [0, 1]$. Our Q-learning algorithm of the next section corresponds to LSPE(0), the case where $\lambda = 0$; we do not have a convenient Q-learning algorithm that parallels LSPE(λ) for $\lambda > 0$.

3 A Least Squares Q-learning Algorithm

3.1 Algorithm

We generate a single[1] infinitely long simulation trajectory $(x_0, x_1, \ldots)$ corresponding to an unstopped system, i.e., using the transition probabilities p_{ij}. Our algorithm starts with an initial guess r_0, and generates a parameter vector sequence $\{r_t\}$. Following the transition (x_t, x_{t+1}), we form the following least squares problem at each time t,

$$\min_{r \in \Re^s} \sum_{k=0}^{t} \Big(\phi(x_k)'r - g(x_k, x_{k+1}) - \alpha \min\big\{c(x_{k+1}), \phi(x_{k+1})'r_t\big\}\Big)^2, \tag{2}$$

whose solution is

$$\hat{r}_{t+1} = \left(\sum_{k=0}^{t} \phi(x_k)\phi(x_k)' \right)^{-1} \sum_{k=0}^{t} \phi(x_k) \Big(g(x_k, x_{k+1}) + \alpha \min\big\{c(x_{k+1}), \phi(x_{k+1})'r_t\big\}\Big). \tag{3}$$

Then we set

$$r_{t+1} = r_t + \gamma(\hat{r}_{t+1} - r_t), \tag{4}$$

where γ is some fixed constant stepsize, whose range will be given later.[2]

This algorithm is related to the LSPE(0) algorithm, which is used for the approximate evaluation of a single stationary policy of a discounted Markovian decision problem, and is analyzed by Bertsekas and Ioffe [BI96], Nedić and Bertsekas [NB03], Bertsekas, Borkar, and Nedić [BBN03], and Yu and Bertsekas [YB06] (see also the recent book by Bertsekas [Ber07], Chapter 6). In particular, if there were no stopping action (or equivalently if the stopping costs are so large that they are inconsequential), then, for $\gamma = 1$, the algorithm (3) becomes

$$r_{t+1} = \left(\sum_{k=0}^{t} \phi(x_k)\phi(x_k)' \right)^{-1} \sum_{k=0}^{t} \phi(x_k) \Big(g(x_k, x_{k+1}) + \alpha \phi(x_{k+1})'r_t \Big), \tag{5}$$

and is identical to the LSPE(0) algorithm for evaluating the policy that never stops. On the other hand, we note that the least squares Q-learning algorithm (3) has much higher computation overhead than the LSPE(0) algorithm (5) for evaluating this policy. In the process of updating r_t via Eq. (3), we can compute the matrix $\Big(\frac{1}{t+1}\sum_{k=0}^{t}\phi(x_k)\phi(x_k)'\Big)^{-1}$ and the vector $\frac{1}{t+1}\sum_{k=0}^{t}\phi(x_k)g(x_k, x_{k+1})$ iteratively and efficiently as in Eq. (5). The terms $\min\big\{c(x_{k+1}), \phi(x_{k+1})'r_t\big\}$, however, need to be recomputed for all the samples x_{k+1}, $k < t$. Intuitively, this computation corresponds to repartitioning the states into those at which to stop and those at which to continue, based on the current approximate Q-factors Φr_t. In Section 4, we will discuss how to reduce this extra overhead.

We will prove that the sequence $\{\Phi r_t\}$ generated by the least squares Q-learning algorithm (3) asymptotically converges to the unique fixed point of ΠF. The idea of the proof is to show that the algorithm can be written as the iteration $\Phi r_{t+1} = \widehat{\Pi}_t \widehat{F}_t(\Phi r_t)$ or its damped version, where $\widehat{\Pi}_t$ and $\widehat{F}_t$ approximate Π and F, respectively, within simulation error that asymptotically diminishes to 0 w.p.1.

[1] Multiple independent infinitely long trajectories can also be used similarly.

[2] We ignore the issues associated with the invertibility of the matrix in Eq. (3). They can be handled, for example, by adding a small positive multiple of the identity to the matrix if it is not invertible.

3.2 Convergence Proof

The iteration (3) can be written equivalently as

$$\hat{r}_{t+1} = \left(\sum_{i=1}^{n} \hat{\pi}_t(i) \phi(i) \phi(i)' \right)^{-1} \sum_{i=1}^{n} \hat{\pi}_t(i) \phi(i) \left(\hat{g}_t(i) + \alpha \sum_{j=1}^{n} \hat{\pi}_t(j|i) \min \left\{ c(j), \phi(j)' r_t \right\} \right),$$

with $\hat{\pi}_t(i)$ and $\hat{\pi}_t(j|i)$ being the empirical frequencies defined by

$$\hat{\pi}_t(i) = \frac{\sum_{k=0}^{t} \delta(x_k = i)}{t+1}, \qquad \hat{\pi}_t(j|i) = \frac{\sum_{k=0}^{t} \delta(x_k = i, x_{k+1} = j)}{\sum_{k=0}^{t} \delta(x_k = i)},$$

where $\delta(\cdot)$ is the indicator function, and $\hat{g}_t$ is the empirical mean of the per-stage costs:

$$\hat{g}_t(i) = \frac{\sum_{k=0}^{t} g(x_k, x_{k+1}) \delta(x_k = i)}{\sum_{k=0}^{t} \delta(x_k = i)}.$$

[In the case where $\sum_{k=0}^{t} \delta(x_k = i) = 0$, we define $\hat{g}_t(i) = 0, \hat{\pi}_t(j|i) = 0$ by convention.] In a more compact notation,

$$\Phi \hat{r}_{t+1} = \widehat{\Pi}_t \widehat{F}_t(\Phi r_t), \tag{6}$$

where the mappings $\widehat{\Pi}_t$ and $\widehat{F}_t$ are simulation-based approximations to Π and F, respectively:

$$\widehat{\Pi}_t = \Phi(\Phi' \widehat{D}_t \Phi)^{-1} \Phi' \widehat{D}_t, \qquad \widehat{F}_t J = \hat{g}_t + \alpha \tilde{P}_t \min \left\{ c, J \right\}, \quad \forall\, J \in \Re^n,$$
$$\widehat{D}_t = \operatorname{diag}\left(\ldots, \hat{\pi}_t(i), \ldots \right), \qquad \left(\tilde{P}_t \right)_{ij} = \hat{\pi}_t(j|i).$$

With a stepsize γ, the least squares Q-learning iteration (4) is written as

$$\Phi r_{t+1} = (1 - \gamma) \Phi r_t + \gamma \widehat{\Pi}_t \widehat{F}_t(\Phi r_t). \tag{7}$$

By ergodicity of the Markov chain, we have w.p.1,

$$\hat{\pi}_t \to \pi, \qquad \tilde{P}_t \to P, \qquad \hat{g}_t \to g, \quad \text{as } t \to \infty,$$

where g denotes the expected per-stage cost vector with $\sum_{j=1}^{n} p_{ij} g(i,j)$ as the i-th component.

For each t, denote the invariant distribution of $\tilde{P}_t$ by $\tilde{\pi}_t$. We now have three distributions, $\pi, \hat{\pi}_t, \tilde{\pi}_t$, which define, respectively, three weighted Euclidean norms, $\| \cdot \|_\pi, \| \cdot \|_{\hat{\pi}_t}, \| \cdot \|_{\tilde{\pi}_t}$. The mappings we consider are non-expansive or contraction mappings with respect to one of these norms. In particular:

- the mapping $\widehat{\Pi}_t$ is non-expansive with respect to $\| \cdot \|_{\hat{\pi}_t}$ (since $\widehat{\Pi}_t$ is projection with respect to $\| \cdot \|_{\hat{\pi}_t}$), and
- the mapping $\widehat{F}_t$ is a contraction, with modulus α, with respect to $\| \cdot \|_{\tilde{\pi}_t}$ (the proof of Lemma 1 can be used to show this).

We have the following facts, each being a consequence of the ones preceding it:

(i) $\hat{\pi}_t, \tilde{\pi}_t \to \pi$ w.p.1, as $t \to \infty$.

(ii) For any $\epsilon > 0$ and a sample trajectory with converging sequences $\hat{\pi}_t, \tilde{\pi}_t$, there exists a time $\bar{t}$ such that for all $t \geq \bar{t}$ and all states i

$$\frac{1}{1+\epsilon} \leq \frac{\hat{\pi}_t(i)}{\pi(i)} \leq 1 + \epsilon, \qquad \frac{1}{1+\epsilon} \leq \frac{\tilde{\pi}_t(i)}{\pi(i)} \leq 1 + \epsilon, \qquad \frac{1}{1+\epsilon} \leq \frac{\tilde{\pi}_t(i)}{\hat{\pi}(i)} \leq 1 + \epsilon.$$

(iii) Under the condition of (ii), for any $J \in \Re^n$, we have

$$\|J\|_\pi \leq (1+\epsilon)\|J\|_{\hat{\pi}_t}, \qquad \|J\|_{\hat{\pi}_t} \leq (1+\epsilon)\|J\|_{\tilde{\pi}_t}, \qquad \|J\|_{\tilde{\pi}_t} \leq (1+\epsilon)\|J\|_\pi,$$

for all t sufficiently large.

Fact (iii) implies the contraction of $\widehat{\Pi}_t \widehat{F}_t$ with respect to $\|\cdot\|_\pi$, as shown in the following lemma.

Lemma 2. *Let $\hat{\alpha} \in (\alpha, 1)$. Then, w.p.1, $\widehat{\Pi}_t \widehat{F}_t$ is a $\|\cdot\|_\pi$-contraction mapping with modulus $\hat{\alpha}$ for all t sufficiently large.*

Proof. Consider a simulation trajectory from the set of probability 1 for which $\tilde{P}_t \to P$ and $\hat{\pi}_t, \tilde{\pi}_t \to \pi$. Fix an $\epsilon > 0$. For any functions J_1 and J_2, using fact (iii) above and the non-expansiveness and contraction properties of $\widehat{\Pi}_t$ and $\widehat{F}_t$, respectively, we have for t sufficiently large,

$$\begin{aligned}
\|\widehat{\Pi}_t \widehat{F}_t J_1 - \widehat{\Pi}_t \widehat{F}_t J_2\|_\pi &\leq (1+\epsilon)\|\widehat{\Pi}_t \widehat{F}_t J_1 - \widehat{\Pi}_t \widehat{F}_t J_2\|_{\hat{\pi}_t} \\
&\leq (1+\epsilon)\|\widehat{F}_t J_1 - \widehat{F}_t J_2\|_{\hat{\pi}_t} \\
&\leq (1+\epsilon)^2\|\widehat{F}_t J_1 - \widehat{F}_t J_2\|_{\tilde{\pi}_t} \\
&\leq (1+\epsilon)^2 \alpha \|J_1 - J_2\|_{\tilde{\pi}_t} \\
&\leq (1+\epsilon)^3 \alpha \|J_1 - J_2\|_\pi.
\end{aligned}$$

Thus, by letting ϵ be such that $(1+\epsilon)^3\alpha < \hat{\alpha} < 1$, we see that $\widehat{\Pi}_t \widehat{F}_t$ is a $\|\cdot\|_\pi$-contraction mapping with modulus $\hat{\alpha}$ for all t sufficiently large. □

Proposition 1. *For any constant stepsize $\gamma \in (0, \frac{2}{1+\alpha})$, r_t converges to r^* w.p.1, as $t \to \infty$.*

Proof. We choose $\bar{t}$ such that for all $t \geq \bar{t}$, the contraction property of Lemma 2 applies. We have for such t,

$$\begin{aligned}
\|\Phi r_{t+1} - \Phi r^*\|_\pi &= \Big\|(1-\gamma)\left(\Phi r_t - \Phi r^*\right) + \gamma\big(\widehat{\Pi}_t \widehat{F}_t(\Phi r_t) - \Pi F(\Phi r^*)\big)\Big\|_\pi \\
&\leq |1-\gamma|\, \|\Phi r_t - \Phi r^*\|_\pi + \gamma\, \|\widehat{\Pi}_t \widehat{F}_t(\Phi r_t) - \widehat{\Pi}_t \widehat{F}_t(\Phi r^*)\|_\pi + \gamma\, \|\widehat{\Pi}_t \widehat{F}_t(\Phi r^*) - \Pi F(\Phi r^*)\|_\pi \\
&\leq \left(|1-\gamma| + \gamma\hat{\alpha}\right) \|\Phi r_t - \Phi r^*\|_\pi + \gamma\epsilon_t, \qquad (8)
\end{aligned}$$

where $\epsilon_t = \|\widehat{\Pi}_t \widehat{F}_t(\Phi r^*) - \Pi F(\Phi r^*)\|_\pi$. Because $\|\widehat{\Pi}_t \widehat{F}_t(\Phi r^*) - \Pi F(\Phi r^*)\|_\pi \to 0$, we have $\epsilon_t \to 0$. Thus, for $\gamma \leq 1$, since

$$(1 - \gamma + \gamma\hat{\alpha}) < 1,$$

it follows that $\Phi r_t \to \Phi r^*$, or equivalently, $r_t \to r^*$, w.p.1. Similarly, based on Eq. (8), in order to have $\|\Phi r_{t+1} - \Phi r^*\|_\pi$ converge to 0 under a stepsize $\gamma > 1$, it is sufficient that $\gamma - 1 + \gamma\hat{\alpha} < 1$, or equivalently,

$$\gamma < \frac{2}{1+\hat{\alpha}}.$$

Hence Φr_t converges to Φr^* for the stepsize $\gamma \in (0, \frac{2}{1+\alpha})$. □

Note that the range of stepsizes for which convergence was shown includes $\gamma = 1$.

Remark 1. The convergence of r_t implies that Φr_t is bounded w.p.1. Using this fact, we can interpret the iteration of the least squares Q-learning algorithm, with the unit stepsize, for instance, as the deterministic fixed point iteration $\Pi F(\Phi r_t)$ plus an asymptotically diminishing stochastic disturbance (see Fig. 1). In particular, the difference between $\Pi F(\Phi r_t)$ and the simulation-based fixed point iteration $\Phi r_{t+1} = \widehat{\Pi}_t \widehat{F}_t(\Phi r_t)$ is

$$\widehat{\Pi}_t \widehat{F}_t(\Phi r_t) - \Pi F(\Phi r_t) = (\widehat{\Pi}_t \hat{g}_t - \Pi g) + \alpha(\widehat{\Pi}_t \tilde{P}_t - \Pi P) \min\{c, \Phi r_t\},$$

and can be bounded by

$$\big\|\widehat{\Pi}_t \widehat{F}_t(\Phi r_t) - \Pi F(\Phi r_t)\big\| \leq \|\widehat{\Pi}_t - \Pi\| \, \|\hat{g}_t\| + \|\Pi\| \, \|\hat{g}_t - g\| + \alpha \|\widehat{\Pi}_t \tilde{P}_t - \Pi P\| \, \big\|\min\{c, \Phi r_t\}\big\|,$$

where $\|\cdot\|$ is any norm. Since Φr_t is bounded w.p.1, the bound on the right-hand side (r.h.s.) can be seen to asymptotically diminish to 0.

Remark 2. A slightly different proof of the convergence of r_t, reminiscent of the argument used later in Section 5, is to interpret the iteration $\Phi r_{t+1} = \widehat{\Pi}_t \widehat{F}_t(\Phi r_t)$ as $\Pi F(\Phi r_t)$ plus a stochastic disturbance whose magnitude is bounded by $\epsilon_t(1 + \|\Phi r_t\|)$, with ϵ_t asymptotically diminishing to 0. (This can be seen from the discussion in the preceding remark.) The convergence of r_t can then be established using the contraction property of ΠF. This will result in a shorter proof. However, the line of proof based on Lemma 2 is more insightful and transparent. Furthermore, Lemma 2 is of independent value, and in particular it will be used in the following convergence rate analysis.

3.3 Comparison to an LSTD Analogue

A natural alternative approach to finding r^* that satisfies $\Phi r^* = \Pi F(\Phi r^*)$ is to replace Π and F with asymptotically convergent approximations. In particular, let $\tilde{r}_{t+1}$ be the solution of $\Phi r = \widehat{\Pi}_t \widehat{F}_t(\Phi r)$, i.e.,

$$\Phi \tilde{r}_{t+1} = \widehat{\Pi}_t \widehat{F}_t(\Phi \tilde{r}_{t+1}), \qquad t = 0, 1, \ldots.$$

With probability 1 the solutions exist for t sufficiently large by Lemma 2. The conceptual algorithm that generates the sequence $\{\tilde{r}_t\}$ may be viewed as the analogue of the LSTD method, proposed by Bradtke and Barto [BB96], and further developed by Boyan [Boy99] (see also the text by Bertsekas [Ber07], Chapter 6). For the optimal stopping problem this is not a viable algorithm because it involves solution of a nonlinear equation. It is introduced here as a vehicle for interpretation of our least squares Q-learning algorithm (2)-(4).

In particular, we note that $\tilde{r}_{t+1}$ is the solution of the equation

$$\tilde{r}_{t+1} = \arg\min_{r \in \Re^s} \sum_{k=0}^{t} \Big(\phi(x_k)' r - g(x_k, x_{k+1}) - \alpha \min\big\{ c(x_{k+1}), \phi(x_{k+1})' \tilde{r}_{t+1} \big\} \Big)^2, \tag{9}$$

so it is the fixed point of the "arg min" mapping in the r.h.s. of the above equation. On the other hand, the least squares Q-learning algorithm (2)-(4), with stepsize $\gamma = 1$, that generates r_{t+1} can be viewed as a single iteration of a fixed point algorithm that aims to find $\tilde{r}_{t+1}$, starting from r_t. This relation can be quantified further. Using an argument similar to the one used in [YB06] for evaluating the optimal asymptotic convergence rate of LSPE, we will show that with any stepsize in the range $(0, \frac{2}{1+\alpha})$, the LSPE-like update Φr_t converges to the LSTD-like update $\Phi \tilde{r}_t$ asymptotically at the rate of $O(t)$ [while we expect both to converge to Φr^* at a slower rate $O(\sqrt{t})$]. In particular, $t\|\Phi r_t - \Phi \tilde{r}_t\|$ is bounded w.p.1, thus $t^\beta(\Phi r_t - \Phi \tilde{r}_t)$ converges to zero w.p.1 for any $\beta < 1$. First we prove the following lemma.

Lemma 3. *(i) $\tilde{r}_t \to r^*$ w.p.1, as $t \to \infty$.*

(ii) $t\|\tilde{r}_{t+1} - \tilde{r}_t\|$ is bounded w.p.1.

Proof. (i) Let $\hat{\alpha} \in (\alpha, 1)$. Similar to the proof of Prop. 1, using Lemma 2, we have that w.p.1, for all t sufficiently large,

$$\begin{aligned}\|\Phi\tilde{r}_{t+1} - \Phi r^*\|_\pi &= \|\widehat{\Pi}_t\widehat{F}_t(\Phi\tilde{r}_{t+1}) - \widehat{\Pi}_t\widehat{F}_t(\Phi r^*) + \widehat{\Pi}_t\widehat{F}_t(\Phi r^*) - \Pi F(\Phi r^*)\|_\pi \\ &\leq \hat{\alpha}\|\Phi\tilde{r}_{t+1} - \Phi r^*\|_\pi + \|\widehat{\Pi}_t\widehat{F}_t(\Phi r^*) - \Pi F(\Phi r^*)\|_\pi.\end{aligned}$$

Thus, w.p.1,

$$\|\Phi\tilde{r}_{t+1} - \Phi r^*\|_\pi \leq \frac{1}{1-\hat{\alpha}}\|\widehat{\Pi}_t\widehat{F}_t(\Phi r^*) - \Pi F(\Phi r^*)\|_\pi \ \rightarrow \ 0,$$

as $t \to \infty$, implying that $\tilde{r}_t$ is bounded and $\tilde{r}_t \to r^*$ as $t \to \infty$.

(ii) By applying Lemma 2, we have that w.p.1, for all t sufficiently large,

$$\begin{aligned}\|\Phi\tilde{r}_{t+1} - \Phi\tilde{r}_t\|_\pi &= \|\widehat{\Pi}_t\widehat{F}_t(\Phi\tilde{r}_{t+1}) - \widehat{\Pi}_t\widehat{F}_t(\Phi\tilde{r}_t) + \widehat{\Pi}_t\widehat{F}_t(\Phi\tilde{r}_t) - \widehat{\Pi}_{t-1}\widehat{F}_{t-1}(\Phi\tilde{r}_t)\|_\pi \\ &\leq \hat{\alpha}\|\Phi\tilde{r}_{t+1} - \Phi\tilde{r}_t\|_\pi + \|\widehat{\Pi}_t\widehat{F}_t(\Phi\tilde{r}_t) - \widehat{\Pi}_{t-1}\widehat{F}_{t-1}(\Phi\tilde{r}_t)\|_\pi,\end{aligned}$$

which implies, by the definition of $\widehat{F}_t$ and $\widehat{F}_{t-1}$, that

$$\|\Phi\tilde{r}_{t+1} - \Phi\tilde{r}_t\|_\pi \leq \frac{1}{1-\hat{\alpha}}\left(\|\widehat{\Pi}_t\hat{g}_t - \widehat{\Pi}_{t-1}\hat{g}_{t-1}\|_\pi + \|\widehat{\Pi}_t\tilde{P}_t - \widehat{\Pi}_{t-1}\tilde{P}_{t-1}\|_\pi\| \min\{c, \Phi\tilde{r}_t\}\|_\pi\right).$$

Evidently (as shown in [YB06]), $\|\widehat{\Pi}_t\hat{g}_t - \widehat{\Pi}_{t-1}\hat{g}_{t-1}\|_\pi$ and $\|\widehat{\Pi}_t\tilde{P}_t - \widehat{\Pi}_{t-1}\tilde{P}_{t-1}\|_\pi$ are bounded by C/t for some constant C and all t sufficiently large. By the first part of our proof, $\Phi\tilde{r}_t$ is bounded. Hence, w.p.1, there exists some sample path-dependent constant C such that for all t sufficiently large,

$$\|\Phi\tilde{r}_{t+1} - \Phi\tilde{r}_t\|_\pi \leq \frac{C}{t}.$$

□

Proposition 2. *For any constant stepsize $\gamma \in (0, \frac{2}{1+\alpha})$, $t(\Phi r_t - \Phi\tilde{r}_t)$ is bounded w.p.1.*

Proof. The proof is similar to that in [YB06]. With the chosen stepsize, by Lemma 2,

$$\begin{aligned}\|\Phi r_{t+1} - \Phi\tilde{r}_{t+1}\|_\pi &= \left\|(1-\gamma)(\Phi r_t - \Phi\tilde{r}_{t+1}) + \gamma\big(\widehat{\Pi}_t\widehat{F}_t(\Phi r_t) - \widehat{\Pi}_t\widehat{F}_t(\Phi\tilde{r}_{t+1})\big)\right\|_\pi \\ &\leq \bar{\alpha}\|\Phi r_t - \Phi\tilde{r}_{t+1}\|_\pi \\ &\leq \bar{\alpha}\|\Phi r_t - \Phi\tilde{r}_t\|_\pi + \bar{\alpha}\|\Phi\tilde{r}_{t+1} - \Phi\tilde{r}_t\|_\pi,\end{aligned}$$

for some $\bar{\alpha} < 1$ and all t sufficiently large. Multiplying both sides by $(t+1)$, using Lemma 3 (ii), and defining $\zeta_t = t\|\Phi r_t - \Phi\tilde{r}_t\|_\pi$, we have that w.p.1 for some constant C', $\beta < 1$ and time $\bar{t}$,

$$\zeta_{t+1} \leq \beta\,\zeta_t + C', \quad \forall\, t \geq \bar{t},$$

which implies that

$$\zeta_t \leq \beta^{t-\bar{t}}\zeta_{\bar{t}} + \frac{C'}{1-\beta}, \quad \forall\, t \geq \bar{t}.$$

Hence $t\|\Phi r_t - \Phi\tilde{r}_t\|_\pi$ is bounded w.p.1. □

4 Variants with Reduced Overhead per Iteration

At each iteration of the least squares Q-learning algorithm (3), (4), while updating r_t, it is necessary to recompute the terms $\min\big\{c(x_{k+1}), \phi(x_{k+1})'r_t\big\}$ for all the samples x_{k+1}, $k < t$. Intuitively, this corresponds to repartitioning the sampled states into those at which to stop and those at which to continue based on the most recent approximate Q-factors Φr_t. In this section we discuss some variants of the algorithm that aim to reduce this computation.

4.1 First Variant

A simple way to reduce the overhead in iteration (3) is to forgo the repartitioning just mentioned. Thus, in this variant we replace the terms $\min\big\{c(x_{k+1}), \phi(x_{k+1})'r_t\big\}$ by $\tilde{q}(x_{k+1}, r_t)$, given by

$$\tilde{q}(x_{k+1}, r_t) = \begin{cases} c(x_{k+1}) & \text{if } k \in K, \\ \phi(x_{k+1})'r_t & \text{if } k \notin K, \end{cases}$$

where $K = \big\{k \mid c(x_{k+1}) \le \phi(x_{k+1})'r_k\big\}$ is the set of states to stop based on the (earlier) approximate Q-factors Φr_k, rather than the (most recent) approximate Q-factors Φr_t. In particular, we replace the term

$$\sum_{k=0}^{t} \phi(x_k) \min\big\{c(x_{k+1}), \phi(x_{k+1})'r_t\big\}$$

in Eq. (3) with

$$\sum_{k=0}^{t} \phi(x_k)\tilde{q}(x_{k+1}, r_t) = \sum_{k\le t,\, k\in K} \phi(x_k)c(x_{k+1}) + \sum_{k\le t,\, k\notin K} \phi(x_k)\phi(x_{k+1})'r_t,$$

which can be efficiently updated at each time t.

Some other similar variants are possible, which employ a limited form of repartitioning the states into those to stop and those to continue. For example, one may repartition only the sampled states within a time window of the m most recent time periods. In particular, in the preceding calculation, instead of the set K, we may use at time t the set

$$K_t = \{k \mid k \in K_{t-1},\, k < t - m\} \cup \big\{k \mid t - m \le k \le t,\, c(x_{k+1}) \le \phi(x_{k+1})'r_t\big\},$$

starting with $K_0 = \{0\}$. Here $m = \infty$ corresponds to the algorithm of the preceding section, while $m = 1$ corresponds to the algorithm of the preceding paragraph. Thus the overhead for repartitioning per iteration is proportional to m, and remains bounded.

An important observation is that in the preceding variations, if r_t converges, then asymptotically the terms $\min\big\{c(x_{k+1}), \phi(x_{k+1})'r_t\big\}$ and $\tilde{q}(x_{k+1}, r_t)$ coincide, and it can be seen that the limit of r_t must satisfy the equation $\Phi r = \Pi F(\Phi r)$, so it must be equal to the unique solution r^*. However, at present we have no proof of convergence of r_t.

4.2 Second Variant

Let us consider another variant, whereby we simply replace the terms $\min\{c(x_{k+1}), \phi(x_{k+1})'r_t\}$ in the least squares problem (2) with $\min\{c(x_{k+1}), \phi(x_{k+1})'r_k\}$. The idea is that for large k and t, these two terms may be close enough to each other, so that convergence may still be maintained. Thus we consider the iteration

$$r_{t+1} = \arg\min_{r\in\Re^s} \sum_{k=0}^{t} \Big(\phi(x_k)'r - g(x_k, x_{k+1}) - \alpha \min\big\{c(x_{k+1}), \phi(x_{k+1})'r_k\big\}\Big)^2. \tag{10}$$

This is a special case of an algorithm due to Choi and Van Roy [CV06], as we will discuss shortly. By carrying out the minimization over r, we can equivalently write Eq. (10) as

$$r_{t+1} = B_{t+1}^{-1} \frac{1}{t+1} \sum_{k=0}^{t} \phi(x_k)\Big(g(x_k, x_{k+1}) + \alpha \min\big\{c(x_{k+1}), \phi(x_{k+1})'r_k\big\}\Big), \tag{11}$$

where we denote

$$B_t = \frac{1}{t} \sum_{k=0}^{t-1} \phi(x_k)\phi(x_k)'.$$

To gain some insight into this iteration, let us rewrite it as follows:

$$\begin{aligned} r_{t+1} &= \frac{1}{t+1} B_{t+1}^{-1} \sum_{k=0}^{t-1} \phi(x_k)\Big(g(x_k, x_{k+1}) + \alpha \min\big\{c(x_{k+1}), \phi(x_{k+1})' r_k\big\}\Big) \\ &\quad + \frac{1}{t+1} B_{t+1}^{-1} \phi(x_t)\Big(g(x_t, x_{t+1}) + \alpha \min\big\{c(x_{t+1}), \phi(x_{t+1})' r_t\big\}\Big) \\ &= \frac{1}{t+1} B_{t+1}^{-1}(tB_t) r_t + \frac{1}{t+1} B_{t+1}^{-1} \phi(x_t)\Big(g(x_t, x_{t+1}) + \alpha \min\big\{c(x_{t+1}), \phi(x_{t+1})' r_t\big\}\Big) \\ &= \frac{1}{t+1} B_{t+1}^{-1}\left(tB_t + \phi(x_t)\phi(x_t)'\right) r_t \\ &\quad + \frac{1}{t+1} B_{t+1}^{-1} \phi(x_t)\Big(g(x_t, x_{t+1}) + \alpha \min\big\{c(x_{t+1}), \phi(x_{t+1})' r_t\big\} - \phi(x_t)' r_t\Big), \end{aligned}$$

and finally

$$r_{t+1} = r_t + \frac{1}{t+1} B_{t+1}^{-1} \phi(x_t)\Big(g(x_t, x_{t+1}) + \alpha \min\big\{c(x_{t+1}), \phi(x_{t+1})' r_t\big\} - \phi(x_t)' r_t\Big). \tag{12}$$

This iteration can be shown to converge to r^*. However, we will show by example that its rate of convergence can be inferior to the least squares Q-learning algorithm [cf. Eqs. (3)-(4)].

Accordingly, we consider another variant that aims to improve the practical (if not the theoretical) rate of convergence of iteration (10) [or equivalently (12)], and is new to our knowledge. In particular, we introduce a time window of size m, and we replace the terms $\min\big\{c(x_{k+1}), \phi(x_{k+1})' r_t\big\}$ in the least squares problem (2) with $\min\big\{c(x_{k+1}), \phi(x_{k+1})' r_{l_{k,t}}\big\}$, where

$$l_{k,t} = \min\{k + m - 1, t\}.$$

In other words, we consider the algorithm

$$r_{t+1} = \arg\min_{r \in \Re^s} \sum_{k=0}^{t} \Big(\phi(x_k)' r - g(x_k, x_{k+1}) - \alpha \min\big\{c(x_{k+1}), \phi(x_{k+1})' r_{l_{k,t}}\big\}\Big)^2. \tag{13}$$

Thus, at time t, the last m terms in the least squares sum are identical to the ones in the corresponding sum for the least squares Q-learning algorithm [cf. Eq. (2)]. The terms $\min\big\{c(x_{k+1}), \phi(x_{k+1})' r_{l_{k,t}}\big\}$ remain constant after m updates (when $l_{k,t}$ reaches the value $k + m - 1$), so they do not need to be updated further.

Note that in the first m iterations, this iteration is identical to the least squares Q-learning algorithm of Section 3 with unit stepsize. An important issue is the size of m. For large m, the algorithm approaches the least squares Q-learning algorithm, while for $m = 1$, it is identical to the earlier variant (10).

4.3 Comparison with Other Algorithms

Let us now consider an algorithm, due to Choi and Van Roy [CV06], and referred to as the *fixed point Kalman filter*. It applies to more general problems, but when specialized to the optimal stopping problem, it takes the form

$$r_{t+1} = r_t + \gamma_t B_{t+1}^{-1} \phi(x_t)\Big(g(x_t, x_{t+1}) + \alpha \min\big\{c(x_{t+1}), \phi(x_{t+1})' r_t\big\} - \phi(x_t)' r_t\Big), \tag{14}$$

where γ_t is a diminishing stepsize. The algorithm is motivated by Kalman filtering ideas and the recursive least squares method in particular. It can also be viewed as a scaled version (with scaling matrix B_{t+1}^{-1}) of the method by Tsitsiklis and Van Roy [TV99], which has the form

$$r_{t+1} = r_t + \gamma_t \phi(x_t)\Big(g(x_t, x_{t+1}) + \alpha \min\{c(x_{t+1}), \phi(x_{t+1})' r_t\} - \phi(x_t)' r_t\Big). \tag{15}$$

Scaling is believed to be instrumental for enhancing the rate of convergence.

It can be seen that when $\gamma_t = 1/(t+1)$, the iterations (12) and (14) coincide. However, the iterations (13) and (14) are different for a window size $m > 1$. As far as we know, the convergence proofs of [TV99] and [CV06] do not extend to iteration (13) or its modification that we will introduce in the next section (in part because of the dependence of r_{t+1} on as many as $t - m$ past iterates through the time window). The following example provides some insight into the behavior of the various algorithms discussed in this paper.

Example 1. This is a somewhat unusual example, which can be viewed as a simple DP model to estimate the mean of a random variable using a sequence of independent samples. It involves a Markov chain with a single state. At each time period, the cost produced at this state is a random variable taking one of n possible values with equal probability.[3] Let g_k be the cost generated at the kth transition. The "stopping cost" is taken to be very high so that the stopping option does not affect the algorithms. We assume that the costs g_k are independent and have zero mean and variance σ^2. The matrix Φ is taken to be the scalar 1, so r^* is equal to the true cost and $r^* = 0$.

Then, the least squares Q-learning algorithm of Section 3 with unit stepsize [cf. Eqs. (3) and (4)] takes the form

$$r_{t+1} = \frac{g_0 + \cdots + g_t}{t+1} + \alpha r_t. \tag{16}$$

The first variant (Section 4.1) also takes this form, regardless of the method used for repartitioning, since the stopping cost is so high that it does not materially affect the calculations. Since iteration (16) coincides with the LSPE(0) method for this example, the corresponding rate of convergence results apply (see Yu and Bertsekas [YB06]). In particular, as $t \to \infty$, $\sqrt{t}\, r_t$ converges in distribution to a Gaussian distribution with mean zero and variance $\sigma^2/(1-\alpha)^2$, so that $E\{r_t^2\}$ converges to 0 at the rate $1/t$, i.e., there is a constant C such that

$$tE\{r_t^2\} \le C, \qquad \forall\ t = 0, 1, \ldots.$$

The second variant [Section 4.2, with time window $m = 1$; cf. Eq. (12)], takes the form

$$r_{t+1} = \frac{g_t}{t+1} + \frac{t+\alpha}{t+1} r_t. \tag{17}$$

The fixed point Kalman filter algorithm [cf. Eq. (14)], and the Tsitsiklis and Van Roy algorithm [cf. Eq. (15)] are identical because the scaling matrix B_{t+1} is the scalar 1 in this example. They take the form

$$r_{t+1} = r_t + \gamma_t(g_t + \alpha r_t - r_t).$$

For a stepsize $\gamma_t = 1/(t+1)$, they are identical to the second variant (17).

We claim that iteration (17) converges more slowly than iteration (16), and that $tE\{r_t^2\} \to \infty$. To this end, we write

$$E\{r_{t+1}^2\} = \left(\frac{t+\alpha}{t+1}\right)^2 E\{r_t^2\} + \frac{\sigma^2}{(t+1)^2}.$$

[3] A more conventional but equivalent example can be obtained by introducing states $1, \ldots, n$, one for each possible value of the cost per stage, and transition probabilities $p_{ij} = 1/n$ for all $i, j = 1, \ldots, n$.

Let $\zeta_t = tE\{r_t^2\}$. Then

$$\zeta_{t+1} = \frac{(t+\alpha)^2}{t(t+1)}\zeta_t + \frac{\sigma^2}{t+1}.$$

From this equation (for $\alpha > 1/2$), we have

$$\zeta_{t+1} \geq \zeta_t + \frac{\sigma^2}{t+1},$$

so ζ_t tends to ∞.

Finally, the variant of Section 4.2 with time window $m > 1$ [cf. Eq. (13)], for $t \geq m$ takes the form

$$r_{t+1} = \frac{g_0 + \cdots + g_t}{t+1} + \alpha\frac{r_{m-1} + r_m + \cdots + r_{t-1} + mr_t}{t+1}, \qquad t \geq m. \tag{18}$$

For $t < m$, it takes the form

$$r_{t+1} = \frac{g_0 + \cdots + g_t}{t+1} + \alpha r_t, \qquad t < m.$$

We may write iteration (18) as

$$r_{t+1} = \frac{g_t}{t+1} + \frac{t+\alpha}{t+1}r_t + \alpha\frac{(m-1)(r_t - r_{t-1})}{t+1}, \qquad t \geq m,$$

and it can be shown again that $t\,E\{r_t^2\} \to \infty$, similar to iteration (17). This suggests that the use of $m > 1$ may affect the practical convegence rate of the algorithm, but is unlikely to affect the theoretical convergence rate.

5 Convergence Analysis for Some Variants

In this section, we prove the convergence of the second variant, iteration (13), with a window-size $m \geq 1$. To simplify notation, we define function h by

$$h(x,r) = \min\big\{c(x), \phi(x)'r\big\},$$

and we write iteration (13) equivalently as

$$r_{t+1} = B_{t+1}^{-1}\frac{1}{t+1}\sum_{k=0}^{t}\phi(x_k)\Big(g(x_k,x_{k+1}) + \alpha h\big(x_{k+1}, r_{l_{k,t}}\big)\Big), \tag{19}$$

where $l_{k,t} = \min\{k+m-1, t\}$.

Proposition 3. *Let r_t be defined by Eq. (19). Then w.p.1, $r_t \to r^*$ as $t \to \infty$.*

In the remainder of this section we provide two alternative proofs. The first proof is based on the o.d.e. (ordinary differential equation) techniques for analyzing stochastic approximation algorithms, and makes use of theorems by Borkar [Bor06] and Borkar and Meyn [BM00], which are also given in the yet unpublished book by Borkar [Bor07] (Chapters 2, 3, and 6). We have adapted the theorems in these sources for our purposes (the subsequent Prop. 4) with the assistance of V. Borkar. This proof is relatively short, but requires familiarity with the intricate methodology of the o.d.e. line of analysis. We have also provided a "direct," somewhat longer proof, which does not rely on references to o.d.e.-related sources, although it is in the same spirit as the o.d.e.-based proof. In an earlier version of this report, we have used the line of argument of the "direct" proof to show the result of Prop. 3 with an additional assumption that guaranteed boundedness of the iterates r_t. We are indebted to V. Borkar who gave us the idea and several suggestions regarding the first proof. These suggestions in turn motivated our modification of the second proof to weaken our boundedness assumption.

5.1 A Proof of Proposition 3 Based on O.D.E. Methods

First, we notice that Eq. (19) implies the following relation,

$$\begin{aligned}(t+1)B_{t+1}r_{t+1} &= tB_t r_t + \phi(x_t)\big(g(x_t,x_{t+1}) + \alpha h(x_{t+1},r_t)\big) \\ &\quad + \sum_{k=0}^{t-1} \alpha\phi(x_k)\big(h(x_{k+1},r_{l_{k,t}}) - h(x_{k+1},r_{l_{k,t-1}})\big) \\ &= \big(tB_t + \phi(x_t)\phi(x_t)'\big)r_t + \phi(x_t)\big(g(x_t,x_{t+1}) + \alpha h(x_{t+1},r_t) - \phi(x_t)'r_t\big) \\ &\quad + \sum_{k=0}^{t-1} \alpha\phi(x_k)\big(h(x_{k+1},r_{l_{k,t}}) - h(x_{k+1},r_{l_{k,t-1}})\big).\end{aligned}$$

Thus iteration (19) is equivalent to

$$\begin{aligned}r_{t+1} =& r_t + \frac{1}{t+1}B_{t+1}^{-1}\phi(x_t)\big(g(x_t,x_{t+1}) + \alpha h\big(x_{t+1},r_t\big) - \phi(x_t)'r_t\big) \\ &+ \frac{1}{t+1}B_{t+1}^{-1}\sum_{k=0}^{t-1} \alpha\phi(x_k)\big(h(x_{k+1},r_{l_{k,t}}) - h(x_{k+1},r_{l_{k,t-1}})\big).\end{aligned} \tag{20}$$

The idea is to reduce iteration (20) to the following form and study its convergence:

$$r_{t+1} = r_t + \frac{1}{t+1}B^{-1}\phi(x_t)\big(g(x_t,x_{t+1}) + \alpha h\big(x_{t+1},r_t\big) - \phi(x_t)'r_t\big) + \frac{1}{t+1}\Delta_t, \tag{21}$$

where $B^{-1} = lim_{t\to\infty}B_t^{-1}$, and Δ_t is a noise sequence. It is worth to point out that the effect of window size $m > 1$ will be neglected in our convergence analysis, and this does not contradict our favoring $m > 1$ to $m = 1$, because in general the asymptotic convergence rate of the iterations with and without the noise term can differ from each other.

We need the following result from the o.d.e. analysis of stochastic approximation, which only requires a rather weak assumption on the noise term.

A General Convergence Result

Consider the iteration

$$r_{t+1} = r_t + \gamma_t\big(H(y_t,r_t) + \Delta_t\big), \tag{22}$$

where γ_t is the stepsize (deterministic or random); $\{y_t\}$ is the state sequence of a Markov process; $H(y,r)$ is a function of (y,r); and Δ_t is the noise sequence. Let the norm of $\Re^s$ be any norm. We assume the following.

Assumption 1. *The function $H(y,r)$ is Lipschitz continuous in r for all y with the same Lipschitz constant. The stepsize γ_t satisfies w.p.1, $\sum_{t=0}^{\infty}\gamma_t = \infty, \sum_{t=0}^{\infty}\gamma_t^2 < \infty$, and $\gamma_t \le \gamma_{t-1}$ for all t sufficiently large. The noise Δ_t satisfies*

$$\|\Delta_t\| \le \epsilon_t(1 + \|r_t\|), \quad w.p.1, \tag{23}$$

where ϵ_t is a scalar sequence that converges to 0 w.p.1, as $t \to \infty$.

The convergence of iteration (22) under Assumption 1 can be analyzed based on the analysis in Borkar [Bor06] on averaging of “Markov noise,” and the stability analysis in Borkar and Meyn [BM00] through the scaling limit o.d.e.

We assume that y_t is the state of a finite-state Markov chain. (The analysis of [BM00, Bor06, Bor07] applies to a much more general setting.) For a function $f(y)$, we denote by $E_0\{f(Y)\}$ the expectation over Y with respect to the invariant distribution of the Markov chain y_t. For a positive scalar b, define

$$H_b(y,r) = \frac{H(y,b\,r)}{b}, \qquad H_\infty(y,r) = \lim_{b\to\infty} \frac{H(y,b\,r)}{b},$$

where the existence of the limit defining $H_\infty(y,r)$ for every (y,r) will be part of our assumption. We refer to

$$\dot{r} = E_0\{H(Y,r)\} \tag{24}$$

as our "basic" o.d.e. We consider also the scaled version of our basic o.d.e.,

$$\dot{r} = E_0\{H_b(Y,r)\}, \tag{25}$$

and the scaling limit o.d.e.,

$$\dot{r} = E_0\{H_\infty(Y,r)\}. \tag{26}$$

Proposition 4. *Suppose that H_∞ exists, and the origin is an asymptotically stable equilibrium point of the scaling limit o.d.e. (26), and the basic o.d.e. (24) has a unique globally asymptotically stable equilibrium point r^*. Suppose also that Assumption 1 holds. Then iteration (22) converges to r^* w.p.1.*

Proof. Assume first that $\sup_t \|r_t\| < \infty$ w.p.1. Then the noise sequence Δ_t satisfies $\lim_{t\to\infty} \Delta_t = 0$ w.p.1. Applying the averaging result in [Bor06] (Corollary 3.1 and its preceding remark on noise, p. 144), we have that r_t converges to r^*, the unique stable equilibrium of the basic o.d.e. $\dot{r} = E_0\{H(Y,r)\}$.

To establish the boundedness of $\|r_t\|$, we use the scaled o.d.e. (25) and the limit o.d.e. (26) together with the averaging argument of [Bor06]. The proof proceeds in three steps.

(i) For any given positive number T, we define time intervals $[t, k_t]$, where $k_t = \min\{k \mid k > t, \sum_{j=t}^{k} \gamma_j \geq T\}$. For every such interval, we consider the scaled sequence

$$\hat{r}_j^{(t)} = \frac{r_j}{b_t}, \quad j \in [t, k_t], \quad \text{where } b_t = \max\{\|r_t\|, 1\}.$$

Then for $j \in [t, k_t)$,

$$\hat{r}_{j+1}^{(t)} = \hat{r}_j^{(t)} + \gamma_j \big(H_{b_t}(y_j, \hat{r}_j^{(t)}) + \widehat{\Delta}_j^{(t)} \big)$$

where $\widehat{\Delta}_j^{(t)} = \frac{\Delta_j}{b_t}$ is the scaled noise and satisfies

$$\|\widehat{\Delta}_j^{(t)}\| \leq \epsilon_j (1 + \|\hat{r}_j^{(t)}\|).$$

Using the Lipschitz continuity of $H(y,\cdot)$ and the discrete Gronwall inequality (Lemma 4.3 of [BM00]), we have that $\hat{r}_j^{(t)}$ is bounded on $[t, k_t]$ with the bound independent of t. Also, as a consequence, the noise satisfies $\|\widehat{\Delta}_j^{(t)}\| \leq \epsilon_j C_T$ with C_T being some constant independent of t. These allow us to apply again the averaging analysis in [Bor06] to $\hat{r}_j^{(t)}, j \in [t, k_t]$ and obtain our convergence result, as we show next.

(ii) Let $x^{(t)}(u)$ be the solution of the scaled o.d.e. $\dot{r} = E_0\{H_{b_t}(Y,r)\}$ at time u with initial condition $x(0) = \hat{r}_t^{(t)}$. (Note that $x^{(t)}(\cdot)$ is a curve on $\Re^s$ whose argument is the natural continuous time.) By applying the analysis in [Bor06] (the proof of Lemma 2.2, Lemma 2.3 itself, and the proof of Lemma 3.1, p. 142-143), we have in particular that

$$\lim_{t\to\infty} \|\hat{r}_{k_t}^{(t)} - x^{(t)}(u_t)\| = 0, \quad w.p.1, \tag{27}$$

where $u_t = \sum_{j=t}^{k_t} \gamma_j$ and $u_t \in [T, T+1)$ for t sufficiently large.

(iii) Notice that $x(0) = \hat{r}_t^{(t)}$ is within the unit ball. Hence, by Lemma 4.4 of [BM00], for suitably chosen T, $x^{(t)}(u_t)$ would be close to the origin 0, had b_t been sufficiently large, because the origin 0 is the globally asymptotically stable equilibrium point of the scaling limit o.d.e. (26). Thus, using Eq. (27) and applying the analysis of [BM00] (Theorem 2.1(i) and its proof in Section 4.1, with Lemma 4.1-4.4 and Lemma 4.6, p. 460-464), we can establish that r_t is bounded w.p.1. □

Expressing the 2nd Variant as Iteration (22)

We will show that our iteration (19), equivalently (20), can be written in the form (21) with the noise Δ_t satisfying Eq. (23). Iteration (21) is a special case of iteration (22) with the following identifications. Let y_t be the pair of states (x_t, x_{t+1}). For $y = (y^1, y^2)$, let $H(y, r)$ be the mapping from $\Re^s$ to $\Re^s$ defined by

$$H(y,r) = B^{-1}\phi(y^1)\big(g(y^1,y^2) + \alpha h\big(y^2,r\big) - \phi(y^1)'r\big).$$

Clearly, $H(y,r)$ is Lipschitz continuous in r uniformly for all y. Let the stepsize γ_t be $\frac{1}{t+1}$.

We consider the associated o.d.e. and verify that they satisfy the corresponding assumptions of Prop. 4. For a function $f(y)$, $E_0\{f(Y)\} = \sum_{i,j}\pi(i)p_{ij}f\big((i,j)\big)$. Consider our "basic" o.d.e. $\dot{r} = E_0\{H(Y,r)\}$ and the mapping associated with its r.h.s.,

$$\Phi E_0\{H(Y,r)\} = \Pi F(\Phi r) - \Phi r.$$

Since $\Pi F(\Phi r)$ is a contraction mapping, its fixed point r^* is the globally asymptotically stable equilibrium point of the basic o.d.e. Consider the scaled o.d.e. $\dot{r} = E_0\{H_b(Y,r)\}$, and the scaling limit o.d.e., $\dot{r} = E_0\{H_\infty(Y,r)\}$. It is easy to see that H_∞ exists, and

$$\Phi E_0\{H_b(Y,r)\} = \Pi\left(g/b + \alpha P\min\left\{c/b, \Phi r\right\}\right) - \Phi r,$$
$$\Phi E_0\{H_\infty(Y,r)\} = \alpha\Pi P\min\{0,\Phi r\} - \Phi r,$$

where g denotes the vector of per-stage costs whose components are defined by $g(i) = \sum_j p_{ij}g(i,j)$. Since by Lemma 1, the mappings $\Pi\left(g/b + \alpha P\min\left\{c/b,\Phi r\right\}\right)$ and $\alpha\Pi P\min\{0,\Phi r\}$ are contractions with modulus α, the scaled o.d.e. (25) and the scaling limit (26) have globally asymptotically stable equilibrium points, which we denote by r^b and r^∞, respectively. We have in particular, $r^b = r^*$ for $b = 1$; $r^\infty = 0$, the original of $\Re^s$; and r^b converges to r^∞ as $b \to \infty$.

We now show that the the noise term Δ_t satisfies Eq. (23), when we reduce iteration (20), i.e.,

$$\begin{aligned} r_{t+1} =& r_t + \frac{1}{t+1}B_{t+1}^{-1}\phi(x_t)\big(g(x_t,x_{t+1}) + \alpha h\big(x_{t+1},r_t\big) - \phi(x_t)'r_t\big) \\ &+ \frac{1}{t+1}B_{t+1}^{-1}\sum_{k=0}^{t-1}\alpha\phi(x_k)\big(h(x_{k+1},r_{l_{k,t}}) - h(x_{k+1},r_{l_{k,t-1}})\big), \end{aligned}$$

to iteration (21), i.e.,

$$r_{t+1} = r_t + \frac{1}{t+1}B^{-1}\phi(x_t)\big(g(x_t,x_{t+1}) + \alpha h\big(x_{t+1},r_t\big) - \phi(x_t)'r_t\big) + \frac{1}{t+1}\Delta_t.$$

To simplify the notation, in the proofs and discussions we will use $o(1)$ to denote a scalar sequence that converges to 0 w.p.1; we can write Eq. (23), for instance, as $\|\Delta_t\| = o(1)(1 + \|r_t\|)$ for short.

Since $B_{t+1}^{-1} \to B^{-1}$ w.p.1, as $t \to \infty$, replacing B_{t+1}^{-1} by B^{-1} in the third term of the r.h.s. of Eq. (20) will only introduce a noise term of magnitude $o(1)(1 + \|r_t\|)$. We aim to show that the second term of the r.h.s. of Eq. (20) can also be treated as a noise term of magnitude $o(1)(1 + \|r_t\|)$.

Since $r_{l_{k,t}}$ and $r_{l_{k,t-1}}$ differ for only the last m values of k, for which $r_{l_{k,t}} = r_t$ and $r_{l_{k,t-1}} = r_{t-1}$, it can be seen that for all t sufficiently large,

$$\frac{1}{t+1}\Big\| B_{t+1}^{-1} \sum_{k=0}^{t-1} \alpha\phi(x_k)\big(h(x_{k+1}, r_{l_{k,t}}) - h(x_{k+1}, r_{l_{k,t-1}})\big)\Big\| \leq \frac{C_1}{t+1}\|r_t - r_{t-1}\| \tag{28}$$

for some constant C_1. Therefore it is sufficient to show that $\|r_t - r_{t-1}\| = o(1)(1 + \|r_t\|)$. This is indicated in the following.

Lemma 4. *For all t sufficiently large and some (path-dependent) constant C,*

$$\|r_{t+1} - r_t\| \leq \frac{C}{t+1}(1 + \|r_t\|), \qquad \|r_{t+1} - r_t\| \leq \frac{C}{t+1}(1 + \|r_{t+1}\|), \quad w.p.1.$$

Proof. From Eqs. (20) and (28), it can be seen that for all $t \geq t_0$,

$$\|r_{t+1} - r_t\| \leq \frac{1}{t+1}\big(C_1\|r_t - r_{t-1}\| + C_2(1 + \|r_t\|)\big) \tag{29}$$

for some sufficiently large t_0 and suitably chosen positive constants C_1 and C_2. Define $K_1 = C_1 + C_2$ and $a(t, K_1) = \frac{1}{1-\frac{K_1}{t+1}}$. Choose $\bar{t} \geq t_0$ such that for all $t \geq \bar{t}$, $a(t, K_1)$ are positive and

$$a(t, K_1)\tfrac{K_1}{t+1} \leq 1, \qquad a(t, K_1)\tfrac{C_1}{t+1} \leq \tfrac{1}{2}. \tag{30}$$

This is possible, since $a(t, K_1)$ tends to 1 and the expressions of the l.h.s. are decreasing to 0 as t increases. For $\bar{t}$, Eq. (29) can be written as, using $\|r_{\bar{t}} - r_{\bar{t}-1}\| \leq \|r_{\bar{t}}\| + \|r_{\bar{t}-1}\|$,

$$\begin{aligned} \|r_{\bar{t}+1} - r_{\bar{t}}\| \leq & \ \frac{1}{\bar{t}+1}(K_1\|r_{\bar{t}}\| + C_1\|r_{\bar{t}-1}\| + C_2) \\ \leq & \ \frac{1}{\bar{t}+1}(K_1\|r_{\bar{t}}\| + K_2) \end{aligned} \tag{31}$$

where K_2 is defined, together with some scalar Δ, such that

$$K_2 = C_1\|r_{\bar{t}-1}\| + C_2 + \Delta, \qquad \tfrac{\Delta}{K_2} \geq \tfrac{1}{2}.$$

Note that in the following argument, K_1 and K_2 are fixed scalars (do not depend on t).

We will show by induction that Eq. (31) holds with $\bar{t}$ replaced by t for all $t \geq \bar{t}$. This will imply the first relation in the lemma with $C = \max\{K_1, K_2\}$. It is sufficient to verify Eq. (31) for $t = \bar{t}+1$, which we now set to do. From Eq. (31) we have

$$\|r_{\bar{t}+1} - r_{\bar{t}}\| \leq \tfrac{K_1}{\bar{t}+1}\big(\|r_{\bar{t}} - r_{\bar{t}+1}\| + \|r_{\bar{t}+1}\|\big) + \tfrac{K_2}{\bar{t}+1}.$$

Subtracting $\frac{K_1}{\bar{t}+1}\|r_{\bar{t}} - r_{\bar{t}+1}\|$ from both sides and multiplying by $a(\bar{t}, K_1)$, we obtain using the definition of $a(\bar{t}, K_1)$,

$$\begin{aligned} \|r_{\bar{t}+1} - r_{\bar{t}}\| \leq & \ a(\bar{t}, K_1)\tfrac{K_1}{\bar{t}+1}\|r_{\bar{t}+1}\| + a(\bar{t}, K_1)\tfrac{K_2}{\bar{t}+1} \\ \leq & \ \|r_{\bar{t}+1}\| + \tfrac{1}{2}\tfrac{K_2}{C_1}, \end{aligned} \tag{32}$$

where the last inequality follows from relations in (30). Substituting into the r.h.s. of (29) for $t = \bar{t}+1$, we have

$$\begin{aligned} \|r_{\bar{t}+2} - r_{\bar{t}+1}\| \leq & \ \frac{1}{\bar{t}+2}\big(C_1\|r_{\bar{t}+1} - r_{\bar{t}}\| + C_2(1 + \|r_{\bar{t}+1}\|)\big) \\ \leq & \ \frac{1}{\bar{t}+2}\big(C_1\|r_{\bar{t}+1}\| + \tfrac{1}{2}K_2 + C_2(1 + \|r_{\bar{t}+1}\|)\big) \\ \leq & \ \frac{1}{\bar{t}+2}\big(K_1\|r_{\bar{t}+1}\| + \tfrac{1}{2}K_2 + C_2\big) \\ \leq & \ \frac{1}{\bar{t}+2}\big(K_1\|r_{\bar{t}+1}\| + K_2\big), \end{aligned}$$

where the last inequality follows from $\frac{1}{2}K_2 + C_2 \leq \Delta + C_2 \leq K_2$ by our definitions of K_2 and Δ. This completes our induction showing the first claim.

The second relation of the lemma follows from the same induction: we have that for all $t \geq \bar{t}$, Eq. (32) holds with $\bar{t}$ replaced by t, i.e.,

$$\|r_{t+1} - r_t\| \leq a(t, K_1)\tfrac{K_1}{t+1}\|r_{t+1}\| + a(t, K_1)\tfrac{K_2}{t+1}, \quad \forall t \geq \bar{t};$$

and since $a(t, K_1) \to 1$ as $t \to \infty$, the claim follows. □

This completes the proof of Prop. 3.

5.2 An Alternative "Direct" Proof of Proposition 3

The method of proof uses a device that is typical of the o.d.e. approach: we define a sequence of times k_j, where the length of the interval $[k_j, k_{j+1}]$ is increasing with j at the rate of a geometric progression. We then analyze the progress of the algorithm using a "different clock" under which the interval $[k_j, k_{j+1}]$ is treated as a unit time that is "long" enough for the contraction property of the algorithm to manifest itself and to allow a convergence proof. In particular, we will argue, roughly speaking, that r_t changes slowly so that during a "considerable" amount of time after t, namely $[t, t + \delta\, t]$ for some small scalar δ, the terms $h\big(x_{k+1}, r_{l_{k,t}}\big)$ are close to $h\big(x_{k+1}, r_t\big)$. This will allow us to view $\Phi r_{t+\delta t}$ as the result of a contracting fixed point iteration applied on Φr_t, plus stochastic noise. Based on this interpretation, we will first show that the subsequence r_k at times $k = (1+\delta)t, (1+\delta)^2 t, \ldots$ comes close to r^* (within a distance that is a decreasing function of δ), and we will then show the convergence of the entire sequence.

To be properly viewed as time indices, the real numbers $\delta\, t$ and $(1+\delta)t$ need to be rounded to the closest integers, but to be concise we will skip this step and treat them as if they were integers (The error introduced can be absorbed in the $o(1)$ factor). Also, in various estimates we will simply denote by $\|\cdot\|$ the Euclidean norm $\|\cdot\|_\pi$ on $\Re^n$, except where noted. For convenience, we define the norm on the space of r by $\|r\| = \|\Phi r\|_\pi$. Thus $\|r\| = \|\Phi r\|$ in our notation.

An Ergodicity Property

In our convergence proof, we will argue that for a sample trajectory, the empirical state frequencies on the segments $[t, t+\delta t)$ approach the steady-state probabilities, as t increases to ∞. This is fairly evident, but for completeness we include the proof.

Lemma 5. *Let δ be a positive number, let i and j be two states such that $p_{ij} > 0$, and let $n_{\delta,t}(i,j)$ be the empirical frequency of a transition (i,j) in the time interval $[t, t + \delta\, t)$. For any $\bar{t} > 0$, let*

$$t_k = (1+\delta)^k \bar{t}, \qquad k = 0, 1, \ldots$$

Then

$$\lim_{k\to\infty} n_{\delta,t_k}(i,j) = \pi(i)\, p_{ij}, \quad w.p.1.$$

Proof. Fix the transition (i,j). By renewal theory and the central limit theorem, for some constant C_1,

$$E\big\{|n_{\delta,t_k}(i,j) - \pi(i)\, p_{ij}|^2\big\} \leq \frac{C_1}{\delta\, t_k}.$$

By Chebyshev's inequality,

$$P\big\{|n_{\delta,t_k}(i,j) - \pi(i)\, p_{ij}| \geq \epsilon_k\big\} \leq \frac{C_1}{\epsilon_k^2 \delta\, t_k}.$$

Let $\epsilon_k = t_k^{-\beta/2}$ for some $\beta \in (0,1)$. Thus

$$\lim_{k\to\infty} \epsilon_k = 0, \qquad \epsilon_k^2\, t_k = t_k^{1-\beta}.$$

Define the events A_k by

$$A_k = \big\{ |n_{\delta,t_k}(i,j) - \pi(i)\, p_{ij}| \geq \epsilon_k \big\}.$$

Then for some constants C_2, C_3,

$$\sum_{k=1}^{\infty} P\{A_k\} \leq C_2 \sum_{k=1}^{\infty} t_k^{-(1-\beta)} = C_3 \sum_{k=1}^{\infty} (1+\delta)^{-k(1-\beta)} < \infty,$$

so by the Borel-Cantelli Lemma, w.p.1, only finitely many events A_k occur, which proves our claim. □

Note that although Lemma 5 is stated for one fixed δ and one fixed $\bar{t}$, the conclusion holds for a countable number of δ and $\bar{t}$ simultaneously. Thus it is valid to choose in the following proof any $\bar{t}$, and any δ arbitrarily small, say, from the sequence 2^{-j}. For conciseness, we will not mention this explicitly again.

Estimates Within Trajectory Segments of the Form $[t, t+\delta\, t]$

Consider a sample trajectory from a set of probability 1 for which the assertions in Lemmas 4 and 5 (for a countable number of δ) hold. We can thus omit the statement "w.p.1" in the following analysis. Fix δ, which can be an arbitrarily small positive scalar.

Lemma 6. *For t sufficiently large and $k \in [t, t+\delta\, t]$*

$$|\phi(i)' r_k - \phi(i)' r_t| \leq \theta_\delta (1 + \|r_t\|), \qquad i = 1, \ldots, n, \tag{33}$$

where θ_δ is a scalar independent of t, and $\theta_\delta \to 0$ as $\delta \to 0$.

Proof. For $m \geq 1$, using $r_{t+m} = r_t + \sum_{j=1}^m (r_{t+j} - r_{t+j-1})$ and using Lemma 4 to bound $\|r_{t+j} - r_{t+j-1}\|$, we can bound $r_{t+m} - r_t$ for t sufficiently large by

$$\begin{aligned}
\|r_{t+m} - r_t\| &\leq \sum_{j=1}^m \tfrac{C}{t+j}(1 + \|r_{t+j-1}\|) \\
&\leq \sum_{j=1}^m \tfrac{C}{t+j}(1 + \|r_t\| + \|r_{t+j-1} - r_t\|) \\
&= \left(\sum_{j=1}^m \tfrac{C}{t+j} \right)(1 + \|r_t\|) + \sum_{j=1}^m \tfrac{C}{t+j} \|r_{t+j-1} - r_t\|
\end{aligned}$$

for some constant C, where the second inequality follows from the triangle inequality. By using the version of the discrete Gronwall inequality given as Lemma 4.3(i) of [BM00], we have for $m \leq \delta\, t$,

$$\|r_{t+m} - r_t\| \leq \tfrac{mC}{t}(1 + \|r_t\|) e^{\sum_{j=1}^m \frac{C}{t+j}} \leq \delta C e^{\delta C}(1 + \|r_t\|).$$

By the equivalence of norms, there exists $C' > 0$ such that for all r, $\|\Phi r\|_\infty \leq C' \|\Phi r\|_\pi = C' \|r\|$. The claim then follows by choosing $\theta_\delta = C' \delta C e^{\delta C}$, which converges to 0 as $\delta \to 0$. □

Since $l_{k,t+\delta t-1} \in [t, t+\delta\, t)$ for $k \in [t, t+\delta\, t)$, inequality (33) implies that for all $k \in [t, t+\delta\, t)$,

$$\big|h\big(x_{k+1}, r_{l_{k,t+\delta t-1}}\big) - h\big(x_{k+1}, r_t\big)\big| \leq \theta_\delta(1 + \|r_t\|). \tag{34}$$

Using this, we now write $\Phi r_{t+\delta t}$ in terms of Φr_t, $\Pi F(\Phi r_t)$, and residual terms.

Lemma 7. *For t sufficiently large,*

$$\Phi r_{t+\delta t} = \frac{1}{1+\delta}\Phi r_t + \frac{\delta}{1+\delta}\Pi F(\Phi r_t) + \Delta_{\delta,t}, \tag{35}$$

where $\Delta_{\delta,t}$ satisfies

$$\|\Delta_{\delta,t}\| \leq \left(\frac{\delta\theta_\delta}{1+\delta} + \epsilon_t\right)(1 + \|\Phi r_t\|),$$

where θ_δ is a scalar independent of t such that $\theta_\delta \to 0$ as $\delta \to 0$, and ϵ_t is a scalar sequence that converges to 0 as $t \to \infty$.

Proof. Assume t is sufficiently large so that $\delta t > m$. We have by definition

$$\begin{aligned}\Phi r_{t+\delta t} = {}& \Phi B_{t+\delta t}^{-1}\frac{1}{t+\delta\, t}\sum_{k=0}^{t-1}\phi(x_k)\big(g(x_k, x_{k+1}) + \alpha h(x_{k+1}, r_{k+m-1})\big) \\ & + \Phi B_{t+\delta t}^{-1}\frac{1}{t+\delta\, t}\sum_{k=t}^{t+\delta t-1}\phi(x_k)\big(g(x_k, x_{k+1}) + \alpha h(x_{k+1}, r_{l_{k,t+\delta t-1}})\big).\end{aligned} \tag{36}$$

We approximate separately the two terms on the r.h.s. Using an expression of r_t similar to Eq. (19), we write the first term as

$$\Phi B_{t+\delta t}^{-1}\frac{1}{t+\delta\, t}(tB_t r_t) + \frac{1}{t+\delta\, t}\Delta = \frac{1}{1+\delta}\Phi r_t + \epsilon_{\delta,t}^1, \tag{37}$$

where Δ accounts for replacing $r_{l_{k,t+\delta t}} = r_{k+m-1}$ with $r_{l_{k,t-1}}$:

$$\Delta = \Phi B_{t+\delta t}^{-1}\sum_{k=0}^{t-1}\alpha\phi(x_k)\big(h(x_{k+1}, r_{k+m-1}) - h(x_{k+1}, r_{l_{k,t-1}})\big),$$

and

$$\epsilon_{\delta,t}^1 = \frac{1}{1+\delta}\Phi\left(B_{t+\delta t}^{-1}B_t - I\right)r_t + \frac{1}{t+\delta\, t}\Delta.$$

As $t \to \infty$, the differences $\|r_{k+m-1} - r_{t-1}\|, k \in [t-m+1, t-1]$ are bounded by a multiple of $(1 + \|r_t\|)$ that diminishes to 0 (by Lemma 4 and the proof of Lemma 6), hence $\|\Delta\| \leq C(1 + \|r_t\|)$, for some constant C and t sufficiently large. Also, $B_{t+\delta t}^{-1}B_t \to I$, as $t \to \infty$ (by Lemma 5). Hence, we have

$$\|\epsilon_{\delta,t}^1\| = o(1)(1 + \|r_t\|). \tag{38}$$

We write the second term of Eq. (36) as

$$\frac{\delta\, t}{(1+\delta)\, t}\Phi B_{t+\delta t}^{-1}\tilde{B}_{\delta t}\tilde{B}_{\delta t}^{-1}\frac{1}{\delta\, t}\sum_{k=t}^{t+\delta t-1}\phi(x_k)\big(g(x_k, x_{k+1}) + \alpha h(x_{k+1}, r_t)\big) + \epsilon_{\delta,t}^2, \tag{39}$$

where

$$\tilde{B}_{\delta t} = \frac{1}{\delta\, t}\sum_{k=t}^{t+\delta t-1}\phi(x_k)\phi(x_k)',$$

and $\epsilon^2_{\delta,t}$ accounts for the residual in substituting r_t for $r_{l_{k,t+\delta t-1}}$ in $h(x_{k+1}, r_{l_{k,t+\delta t-1}})$ for $k \in [t, t+\delta\, t)$. Thus $\epsilon^2_{\delta,t}$ satisfies, by Eq. (34),

$$\|\epsilon^2_{\delta,t}\| \leq \tfrac{\delta}{1+\delta}\theta_\delta(1 + \|r_t\|), \tag{40}$$

for some positive scalar θ_δ (independent of t) converges to 0 as $\delta \to 0$.

We further approximate the first term of Eq. (39) as follows. Using Lemma 5, we have that as $t \to \infty$, within the segment $[t, t+\delta t)$, the term

$$\Phi \tilde{B}^{-1}_{\delta t} \frac{1}{\delta\, t} \sum_{k=t}^{t+\delta t-1} \phi(x_k)\big(g(x_k, x_{k+1}) + \alpha h(x_{k+1}, r_t)\big)$$

converges to $\Pi F(\Phi r_t)$. Using Lemma 5, we also have that $B^{-1}_{t+\delta t}\tilde{B}_{\delta t} \to I$ as $t \to \infty$. Thus we can write the first term of Eq. (39) as

$$\frac{\delta}{1+\delta}\Pi F(\Phi r_t) + \epsilon^3_{\delta,t}, \tag{41}$$

where $\epsilon^3_{\delta,t}$ accounts for the residual in this approximation, and

$$\|\epsilon^3_{\delta,t}\| = o(1)(1 + \|r_t\|). \tag{42}$$

Putting Eqs. (37)-(42) together, we can write $\Phi r_{t+\delta t}$ as

$$\Phi r_{t+\delta t} = \frac{1}{1+\delta}\Phi r_t + \frac{\delta}{1+\delta}\Pi F(\Phi r_t) + \big(\epsilon^1_{\delta,t} + \epsilon^2_{\delta,t} + \epsilon^3_{\delta,t}\big) \tag{43}$$

where

$$\|\epsilon^1_{\delta,t} + \epsilon^2_{\delta,t} + \epsilon^3_{\delta,t}\| \leq \left(\frac{\delta\theta_\delta}{1+\delta} + o(1)\right)(1 + \|\Phi r_t\|).$$

The claim thus follows. □

A Contraction Argument

Fix a scalar $\bar{\alpha} \in (\alpha, 1)$. Let θ_δ and ϵ_t be as in the preceding lemma. Let δ be such that $\alpha + \theta_\delta < \bar{\alpha}$, and choose β_δ such that

$$\frac{1}{1+\delta} + (\alpha + \theta_\delta)\frac{\delta}{1+\delta} < \beta_\delta < \frac{1}{1+\delta} + \bar{\alpha}\frac{\delta}{1+\delta}.$$

Using the contraction property $\|\Pi F(\Phi r_t) - \Phi r^*\| \leq \alpha\|\Phi r_t - \Phi r^*\|$, the preceding lemma, and the triangle inequality $\|\Phi r_t\| \leq \|\Phi r_t - \Phi r^*\| + \|\Phi r^*\|$, we have

$$\begin{aligned}\|\Phi r_{t+\delta t} - \Phi r^*\| \leq\ & \left(\frac{1}{1+\delta} + \alpha\frac{\delta}{1+\delta}\right)\|\Phi r_t - \Phi r^*\| + \left(\theta_\delta \frac{\delta}{1+\delta} + \epsilon_t\right)(1 + \|\Phi r_t\|) \\ \leq\ & \left(\frac{1}{1+\delta} + (\alpha + \theta_\delta)\frac{\delta}{1+\delta} + \epsilon_t\right)\|\Phi r_t - \Phi r^*\| + \left(\theta_\delta \frac{\delta}{1+\delta} + \epsilon_t\right)(1 + \|\Phi r^*\|).\end{aligned} \tag{44}$$

For any arbitrarily small positive scalar ϵ, we have $\epsilon_t(\|\Phi r^*\| + 1) < \epsilon$ and $\alpha + \theta_\delta + \epsilon_t < \bar{\alpha}$ for t sufficiently large (since $\epsilon_t = o(1)$). Hence, inequality (44) implies that for any ϵ, and for some fixed $\bar{t}$ independent of ϵ, the sequence $\big\{\Phi r_{k_j} \mid k_j = (1+\delta)^j\,\bar{t}\big\}$ satisfies

$$\limsup_{j\to\infty} \|\Phi r_{k_j} - \Phi r^*\| \leq \frac{\epsilon}{1-\beta_\delta} + \frac{\frac{\delta}{1+\delta}\theta_\delta}{1-\beta_\delta}(1 + \|\Phi r^*\|). \tag{45}$$

Since $1-\beta_\delta > (1-\bar{\alpha})\frac{\delta}{1+\delta}$, we have

$$\frac{\frac{\delta}{1+\delta}}{1-\beta_\delta} \leq \frac{1}{1-\bar{\alpha}},$$

and hence

$$\limsup_{j\to\infty} \|\Phi r_{k_j} - \Phi r^*\| \leq \frac{\epsilon}{1-\beta_\delta} + \frac{\theta_\delta}{1-\bar{\alpha}}(1+\|\Phi r^*\|). \tag{46}$$

Since ϵ is arbitrary, letting $\theta'_\delta = \frac{\theta_\delta}{1-\bar{\alpha}}$, we have

$$\limsup_{j\to\infty} \|\Phi r_{k_j} - \Phi r^*\| \leq \theta'_\delta(1+\|\Phi r^*\|). \tag{47}$$

In other words, for all δ sufficiently small, there exists a corresponding subsequence of Φr_t "converging" to the $\theta'_\delta(1+\|\Phi r^*\|)$-sphere centered at Φr^*.

We will now establish the convergence of the entire sequence r_t. When j is sufficiently large, for $t \in [k_j, k_{j+1})$, the difference $\|\Phi r_t - \Phi r_{k_j}\|$ is at most $\bar{\theta}_\delta(1+\|r_{k_j}\|)$ for some positive $\bar{\theta}_\delta$ that diminishes to 0 as $\delta \to 0$ (the proof of Lemma 6). Combining this with Eq. (47), we obtain

$$\begin{aligned}\limsup_{t\to\infty} \|\Phi r_t - \Phi r^*\| &\leq \limsup_{j\to\infty} \bar{\theta}_\delta(1+\|\Phi r_{k_j}\|) + \limsup_{j\to\infty}\|\Phi r_{k_j} - \Phi r^*\| \\ &\leq \limsup_{j\to\infty} \bar{\theta}_\delta(1+\|\Phi r_{k_j} - \Phi r^*\| + \|\Phi r^*\|) + \theta'_\delta(1+\|\Phi r^*\|) \\ &\leq (\bar{\theta}_\delta + \bar{\theta}_\delta\theta'_\delta + \theta'_\delta)(1+\|\Phi r^*\|).\end{aligned}$$

Since δ, and consequently $\bar{\theta}_\delta$ and θ'_δ, can be chosen arbitrarily small, we conclude that the sequence r_t converges to r^*. This completes the proof of Prop. 3.

6 Conclusions

In this paper, we have proposed new Q-learning algorithms for the approximate cost evaluation of optimal stopping problems, using least squares ideas that are central in the LSPE method for policy cost evaluation with linear function approximation. We have aimed to provide alternative, faster algorithms than those of Tsitsiklis and Van Roy [TV99], and Choi and Van Roy [CV06]. The distinctive feature of optimal stopping problems is the underlying mapping F, which is a contraction with respect to the projection norm $\|\cdot\|_\pi$ (cf. Lemma 1). Our convergence proofs made strong use of this property.

It is possible to consider the extension of our algorithms to general finite-spaces discounted problems. An essential requirement for the validity of such extended algorithms is that the associated mapping is a contraction with respect to some Euclidean norm. Under this quite restrictive assumption, it is possible to show certain convergence results. In particular, Choi and Van Roy [CV06] have shown the convergence of an algorithm that generalizes the second variant of Section 4 for the case $m = 1$. It is also possible to extend this variant for the case where $m > 1$ and prove a corresponding convergence result.

Acknowledgment

We are grateful to Prof. Vivek Borkar for suggesting how to strengthen our convergence analysis of the second variant of Section 4 by removing the boundedness assumption of the iterates, and for guiding us to the alternative proof using his o.d.e. line of analysis. We thank Prof. Ben Van Roy for helpful comments. Huizhen Yu is supported in part by the IST Programme of the European Community, under the PASCAL Network of Excellence, IST-2002-506778, and by the postdoctoral research program of Univ. of Helsinki.

References

[BB96] S. J. Bradtke and A. G. Barto, *Linear least-squares algorithms for temporal difference learning*, Machine Learning **22** (1996), no. 2, 33–57.

[BBN03] D. P. Bertsekas, V. S. Borkar, and A. Nedić, *Improved temporal difference methods with linear function approximation*, LIDS Tech. Report 2573, MIT, 2003, also appears in "Learning and Approximate Dynamic Programming," by A. Barto, W. Powell, J. Si, (Eds.), IEEE Press, 2004.

[Ber07] D. P. Bertsekas, *Dynamic programming and optimal control, Vol. II*, 3rd ed., Athena Scientific, Belmont, MA, 2007.

[BI96] D. P. Bertsekas and S. Ioffe, *Temporal differences-based policy iteration and applications in neuro-dynamic programming*, LIDS Tech. Report LIDS-P-2349, MIT, 1996.

[BM95] J. Barraquand and D. Martineau, *Numerical valuation of high dimensional multivariate American securities*, Journal of Financial and Quantitative Analysis **30** (1995), 383–405.

[BM00] V. S. Borkar and S. P. Meyn, *The o.d.e. method for convergence of stochastic approximation and reinforcement learning*, SIAM J. Control Optim. **38** (2000), 447–469.

[Bor06] V. S. Borkar, *Stochastic approximation with 'controlled Markov' noise*, Systems Control Lett. **55** (2006), 139–145.

[Bor07] ______, *Stochastic approximation: A dynamic viewpoint*, 2007, Book Preprint.

[Boy99] J. A. Boyan, *Least-squares temporal difference learning*, Proc. The 16th Int. Conf. Machine Learning, 1999.

[BT96] D. P. Bertsekas and J. N. Tsitsiklis, *Neuro-dynamic programming*, Athena Scientific, Belmont, MA, 1996.

[CV06] D. S. Choi and B. Van Roy, *A generalized Kalman filter for fixed point approximation and efficient temporal-difference learning*, Discrete Event Dyn. Syst. **16** (2006), no. 2, 207–239.

[LS01] F. A. Longstaff and E. S. Schwartz, *Valuing American options by simulation: A simple least-squares approach*, Review of Financial Studies **14** (2001), 113–147.

[NB03] A. Nedić and D. P. Bertsekas, *Least squares policy evaluation algorithms with linear function approximation*, Discrete Event Dyn. Syst. **13** (2003), 79–110.

[Sut88] R. S. Sutton, *Learning to predict by the methods of temporal differences*, Machine Learning **3** (1988), 9–44.

[Tsi94] J. N. Tsitsiklis, *Asynchronous stochastic approximation and Q-learning*, Machine Learning **16** (1994), 185–202.

[TV97] J. N. Tsitsiklis and B. Van Roy, *An analysis of temporal-difference learning with function approximation*, IEEE Trans. Automat. Contr. **42** (1997), no. 5, 674–690.

[TV99] ______, *Optimal stopping of Markov processes: Hilbert space theory, approximation algorithms, and an application to pricing financial derivatives*, IEEE Trans. Automat. Contr. **44** (1999), 1840–1851.

[Wat89] C. J. C. H. Watkins, *Learning from delayed rewards*, Doctoral dissertation, University of Cambridge, Cambridge, United Kingdom, 1989.

[WD92] C. J. C. H. Watkins and P. Dayan, *Q-learning*, Machine Learning **8** (1992), 279–292.

[YB06] H. Yu and D. P. Bertsekas, *Convergence results for some temporal difference methods based on least squares*, LIDS Tech. Report 2697, MIT, 2006.