Convergence Analysis of On-Policy LSPI for Multi-Dimensional Continuous State and Action-Space MDPs and Extension with Orthogonal Polynomial Approximation

Jun Ma and Warren B. Powell
Department of Operations Research and Financial Engineering
Princeton University, Princeton, NJ 08544

February 23, 2010

Abstract

We propose an online, on-policy least-squares policy iteration (LSPI) algorithm which can be applied to infinite horizon problems with where states and controls are vector-valued and continuous. We do not use special structure such as linear, additive noise, and we assume that the expectation cannot be computed exactly. We use the concept of the post-decision state variable to eliminate the expectation inside the optimization problem. We provide a formal convergence analysis of the algorithm under the assumption that value functions are spanned by finitely many known basis functions. Furthermore, the convergence result extends to the more general case of unknown value function form using orthogonal polynomial approximation.

1 Introduction

Central to the solution of Markov decision processes is Bellman's equation, which is often written in the standard form (Puterman (1994))

$$V_t(x_t) = \max_{u_t \in \mathcal{U}} \{C(x_t, u_t) + \gamma \sum_{x' \in \mathcal{X}} P(x'|x_t, u_t) V_{t+1}(x')\}. \tag{1}$$

If the state variable x_t and decision variable u_t are discrete scalars and the transition matrix P is known, the value function $V_t(x_t)$ can be computed by enumerating all the states backward through time, which is a method often referred to as backward dynamic programming. Moreover, there is a mature and elegant convergence theory supporting algorithms that handle problems with finite state and action spaces and computable expectation (Puterman (1994)). However, discrete representations of the problem often suffer from the well-known "curses of dimensionality" which arise in the presence of multidimensional states, actions and random information. In addition, there are a large number of real world applications with continuous state and action spaces, to which direct application of algorithms developed for discrete problems is not appropriate.

In this paper, we consider the problem of solving stochastic, dynamic programs with continuous (and vector-valued) states and actions, without assumptions such as additive noise and quadratic cost functions. An example of a problem arises in managing a resource such as energy which has to be allocated from different sources (coal, natural gas, wind, solar, biomass) to different types of demand (residential, commercial, transportation), including injections to and withdrawals from storage. The control vector may have hundreds to many thousands of dimensions, and must satisfy constraints including production capacity, amount of energy in storage, transmission constraints and demand. These constraints are expressed using $u_t \in U_t$, where the feasible region U_t evolves stochastically over time (reflecting, for example, randomness in supply from wind and variations in demands and prices). The state vector x_t would capture how much energy is in storage, availability of energy from wind, and demand for different types of energy such as electricity, oil and natural gas. We seek to

determine the allocation u_t using Bellman's equation, given by

$$V(x_t) = \max_{u_t \in \mathcal{U}_\sqcup} \left(c_t u_t + \gamma \mathbb{E}\left[V(x_{t+1})|x_t\right]\right),$$

Here, the expectation is over random variations that include energy from wind and solar, prices and demands.

This paper considers continuous value function approximations to handle high-dimensional and continuous applications. We propose an implementable approximate policy iteration algorithm that uses a linear function approximation architecture to handle infinite-horizon discounted Markov decision processes where state, action and information variables are all continuous vectors (possibly of high dimensionality) and the expectation cannot be computed exactly. The algorithm is an online, on-policy modified version of the least squares policy iteration (LSPI Lagoudakis & Parr (2003)), which applies to value functions of the state (instead of state-action Q-factors) and uses least squares or recursive least squares methods for policy evaluation. We provide a rigorous convergence analysis of the algorithm for linear value function approximation with a finite set of known basis functions and extend the result to the case with unknown value function form using Chebyshev basis functions.

We have a special interest in on-policy algorithms since our interest is in problems with multidimensional (and possibly very high dimensional), continuous states and actions. Off-policy algorithms assume that if we are in state x_t, we will choose an action u_t at random to help determine (along with exogenous noise) the next state x_{t+1} that we will visit. If we let w_t be the exogenous noise (random at time t), then x_{t+1} is determined from $x_{t+1} = f(x_t, u_t, w_t)$ where $f(\cdot)$ is an arbitrary transition function. For high-dimensional states and actions, choosing actions (or states) at random to guide exploration becomes impractical, since we will spend the vast majority of the time sampling states and actions that are simply not important. With on-policy algorithms, the action is determined by what we think is a good policy, guiding our attention to the regions that are more likely to be interesting. It is much harder to prove convergence for on-policy algorithms since we are not allowed to directly control the sampling process.

The rest of the paper is organized as follows. In section 2, we review the literature on

continuous function approximations applied to Markov decision process problems and their asymptotic properties and place our work in line with the others. In section 3, we summarize important mathematical foundations to establish convergence and illustrate the details of a least squares/recursive least squares approximate policy iteration (LS/RLSAPI) algorithm. Section 4 presents the convergence results for policy evaluation. In section 5, by applying the results in section 4, we show almost sure convergence of the exact policy iteration algorithm, which requires exact policy evaluation, to the optimal policy. Section 6 proves convergence in the mean of least squares approximate policy iteration, in which case policies are evaluated approximately. In section 7, we extend the convergence result in section 6 to unknown basis functions using Chebyshev polynomials and provide a detailed algorithm. The last section concludes and discusses future research directions.

2 Literature Review

In this section, we first review the literature on continuous function approximations and related convergence theories. Since the inception of the field of dynamic programming, researchers have devoted considerable efforts to explore heuristic value function approximation approaches (see, for example, Bellman & Dreyfus (1959), Bellman et al. (1963), Reetz (1977), Whitt (1978), Schweitzer & Seidmann (1985)) in order to overcome the curses of dimensionality in large-scale stochastic dynamic programming problems. It was not until the early nineties that convergence theory of continuous function approximation was introduced. The literature on continuous value function approximations (include both linear and non-linear approximation) with convergence theories can be divided into two main categories: 1) continuous function approximation of problems with discrete states, and 2) approximation of problems with continuous states. Convergence results have focused either on demonstrating that the value function approximation accurately evaluates a fixed policy (hence no actual optimization), or the much more difficult challenge of proving that we also find an optimal policy. Research into algorithms for finding optimal policies can be divided into: 1) on-policy algorithms where the next state to be sampled is determined by the behavior of the policy we are trying to optimize, and 2) off-policy algorithms which attempts to estimate the value

of an *estimation policy* while using a *sampling policy* (or behavior policy) to determine which state to visit next. To prove convergence, on-policy algorithms need to impose stability conditions on the underlying system by following the policy (which is evolving as we learn it), while off-policy algorithms need stability condition of the system of the sampling policy and also requires that every action taken under the sampling policy is also taken, at least occasionally, under the estimation policy. The popular idea of Q-learning simplifies the problem by estimating the value of a state-action pair, but at a price of a more complex estimation problem.

The table in figure 1 is a brief summary of the literature ordered by year of publication and categorized by their characteristics such as whether the state and action spaces are discrete or continuous (D/C), whether the contribution/reward function is quadratic or general (Q/G), whether the expectation can be computed exactly (Y/N), whether the problem is deterministic or stochastic (Gaussian noise) (D/S(G)), the type of algorithms including value iteration (VI), fixed policy (FP), exact or approximate policy iteration (EPI/API), approximation techniques such as linear (L) and nonlinear (NL), whether the algorithm is on or off-policy (On/Off) and types of convergence including convergence (Y) for deterministic case or computable expectation and almost sure (a.s.), in probability (i.p.) convergence and probability bound (PB) for stochastic algorithms. Details of the algorithms and their convergence properties are discussed in the following subsections.

2.1 Continuous approximations of discrete problems

Many learning algorithms with different continuous function approximation techniques are proposed to handle large-scale discrete MDP problems. Tsitsiklis & Van Roy (1996) first sets up a rigorous framework combining dynamic programming and compact representations using feature-based (linear approximation) value iteration algorithms. With the assumption of computable expectation, Gordon (1995) proves convergence for fitted value iteration algorithm with function approximations that are contraction or expansion mappings such as k-nearest-neighbor, linear interpolation, some types of splines, and local weighted average (but excluding linear regression and neural network). Gordon (2001) proves a weaker result

	State	Action	Reward	Exp.	Noise	Type	Apr.	Policy	Conv.
Bradtke...(1994)	C	C	Q	N	D	API	L	Off	Y
Baird (1995)	D	D	G	NA	D	FP	L/NL	NA	Y
Gordon (1995)	D	D	G	Y	S	VI	NL	NA	Y
Tsitsiklis...(1996)	D	D	G	Y	S	VI	L	NA	Y
Bradtke...(1996)	D	D	G	N	S	FP	L	On	a.s.
Landelius...(1997)	C	C	Q	N	D	VI/API	L	On	Y
Tsitsiklis...(1997)	D	D	G	Y	S	FP	L	On	a.s.
Boyan (1999)	D	D	G	N	S	FP	L	On	a.s.
Papavassiliou...(1999)	D	NA	G	Y	S	FP	L/NL	NA	Y
Gordon (2001)	D	D	G	N	S	VI	L	Off	a.s.
Tadić (2001)	C	NA	G	N	S	FP	L	On	a.s.
Precup...(2001)	C	NA	G	N	S	FP	L	Off	a.s.
Ormoneit...(2002)	C	D	G	N	S	VI	NL	NA	i.p.
Melo...(2007)	C	D	G	N	S	VI	L	Off	a.s.
Szita (2007)	C	C	Q	N	S(G)	VI	L	Off	a.s.
Antos...(2007)	C	D	G	N	S	API	L/NL	Off	PB
Munos...(2008)	C	D	G	N	S	VI	L/NL	NA	PB
Antos...(2008a)	C	D	G	N	S	API	L/NL	Off	PB
Antos...(2008b)	C	C	G	N	S	API	L/NL	Off	PB
Melo...(2008)	C	D	G	N	S	VI	L	Off	a.s.
Sutton...(2009)	D	D	G	N	S	FP	L	Off	a.s.
Maei...(2010)	D	D	G	N	S	FP	NL	On	a.s.

Figure 1: Table of some continuous function approximation algorithms and related convergence results

that the SARSA(0) and V(0) algorithms (value iteration) with linear approximation converge to a bounded region almost surely.

A variety of temporal difference (TD) learning algorithms are proposed to evaluate value function for a fixed policy, which reduces the problem to a discrete Markov chain. Tsitsiklis & Van Roy (1997) proves almost sure convergence of the on-policy TD learning algorithm with linear approximation, while Bradtke & Barto (1996) and Boyan (1999) present almost sure convergence results of the least squares version (LSTD). Precup et al. (2001) and Sutton et al. (2009) shows almost sure convergence of off-policy TD algorithms for a fixed policy with linear approximation, which use importance sampling and i.i.d. sampling of initial states combined with on-policy transition respectively. Maei et al. (2010) proposes on-policy TD learning algorithm that converges almost surely for non-linear smooth value function approximators, such as neural networks. Assuming the ability of computing the expectation, Papavassiliou & Russell (1999) describes the Bridge algorithm for a fixed policy with any "agnostically learnable" function class other than the class of linear combination of fixed basis functions

and provides approximation error bounds. As another alternative to TD algorithms for a fixed policy, the residual gradient algorithm in Baird (1995), which performs gradient descent on the mean squared Bellman residual, is provably convergent for deterministic system.

2.2 Approximations of continuous problems

Early convergence results of function approximation algorithms directly applied to continuous problems can be found for the problem class of linear quadratic regulation (LQR) where value iteration is applied to quadratic contribution/reward and linear state transition, i.e. Bradtke (1993), Bradtke et al. (1994), Landelius & Knutsson (1997) (deterministic) and Szita (2007) (stochastic with Gaussian noise). Tadić (2001) generalizes and extends the almost sure convergence results of on-policy TD algorithm for a fixed policy in Tsitsiklis & Van Roy (1997) to a broader class of Markov chains with uncountable state space. Ormoneit & Sen (2002) adopts a kernel-based approach to off-line value iteration algorithm for continuous state and finite action space, which is convergent in probability as the size of the off-line training sample size increases to infinity. Under strong technical conditions of basis functions, Melo et al. (2007) proves almost sure convergence of Q-learning used with linear function approximation. Melo et al. (2008) shows almost sure convergence of Q-learning and SARSA algorithms with linear approximation under strong assumptions on the sampling policy.

More recently, a series of papers (Antos et al. (2007), Antos, Szepesvári & Munos (2008), Munos & Szepesvári (2008), Antos, Munos & Szepesvari (2008)) derive finite-sample high probability performance bounds for batch reinforcement learning algorithms that produce a near-optimal policy in polynomial time for MDPs with continuous state space. More specifically, Munos & Szepesvári (2008) considers sampling-based fitted value iteration algorithm for MDP problems with large or possibly infinite state spaces but finite action spaces in an off-line setting where a known generative model of the environment is available to sample any transitions from any initial states for each possible action. In a model-free setting, Antos et al. (2007), Antos, Szepesvári & Munos (2008) propose off-policy fitted policy iteration algorithms that are based on a single trajectory of following some fixed sampling policy to handle problems in continuous state space and finite action space. The running time of the algorithms

depends on the mixing rate of the sample trajectory of following the sampling policy, the controllability properties of the underlying MDP, sample size and the approximation power and capacity of the function approximation method used. Antos, Munos & Szepesvari (2008) makes the first step to extend previous convergence analysis to problems with continuous actions by imposing regularity conditions on the action space.

2.3 Discussion of related works

The most directly related works to this paper include the LSTD algorithm developed in Bradtke & Barto (1996) and the LSPI algorithm in Lagoudakis & Parr (2003), which is motivated by LSTD to combine linear value-function approximation and approximate policy iteration. Lagoudakis & Parr (2003) introduces LSPI as an off-policy approximate policy iteration algorithm for finite MDPs that uses LSTDQ (a modified version of LSTD) to evaluate state-action value function (Q-function) of a fixed policy, and no formal convergence analysis of the algorithm is provided except for a generic deterministic error bound of the approximate policy iteration as in Bertsekas & Tsitsiklis (1996). Replacing LSTD with Bellman residual minimization (BRM) for policy evaluation, fitted policy iteration algorithms in Antos et al. (2007), Antos, Szepesvári & Munos (2008), Antos, Munos & Szepesvari (2008) generalize the off-policy LSPI from linear approximation to a richer set of approximation functions for continuous problems and provide probability bounds that ensures high performance. However, this research on off-line LSPI does not detail implementation of the algorithm such as feature selection, which is crucial for determining the approximation power of the function class and in turn providing convergence guarantees for algorithms using linear architecture. Mahadevan & Maggioni (2007) extends LSPI within the representation policy iteration (RPI) framework but lacks a convergence analysis. By representing a finite sample of state transitions induced by the MDP as an undirected graph, RPI constructs an orthonormal set of basis functions with the graph Laplacian operator.

The work presented in our paper extends the LSTD to the continuous framework and solves the problem of persistency of excitation for parameter convergence in system identification with continuous states by imposing orthonormality assumption on basis functions. In

order to handle continuous state and action spaces of potentially high dimensionality and simplify implementation of both policy evaluation and improvement steps, we propose an online, on-policy version of LSPI that uses linear function approximation and LSTD to evaluate the post-decision value function of following a fixed policy and prove convergence in expectation of the algorithm. Furthermore, we propose a solution to feature selection by automatically constructing a finite set of approximate orthonormal basis functions with Chebyshev polynomials and sampled state transitions and provide a sound convergence analysis of the algorithm.

3 Preliminaries and the algorithm

We consider a class of infinite horizon Markov decision processes with continuous state and action spaces. The following subsections discuss several important preliminary concepts including Markov decision processes, contraction operators, continuous correspondence, continuous-state Markov chain and post-decision state variable. These basics are necessary in the convergence proofs in subsequence sections. The last subsection illustrates the details of an approximate policy iteration algorithm with recursive least squares updating method.

3.1 Markov decision processes

We start with a brief review of Markov decision processes. A Markov decision process is a sequential optimization problem where the goal is to find a policy that maximizes (for our application) the expected infinite-horizon discounted rewards. Let x_t be the state of the system at time t, u_t be a vector-valued continuous decision (control) vector, $\pi : \mathcal{X} \to \mathcal{U}$ be a policy in the stationary deterministic policy space Π, $C(x_t, u_t)$ be a contribution/reward function, and γ be a discount factor in $(0, 1)$.

The system evolves according to the following state transition function

$$x_{t+1} = S^M(x_t, u_t, W_{t+1}), \tag{2}$$

where W_{t+1} represents the exogenous information that arrives during the time interval from t

to $t+1$. The problem is to find the policy that solves

$$\sup_{\pi \in \Pi} \mathbb{E} \left\{ \sum_{t=0}^{\infty} \gamma^t C(x_t, \pi(x_t)) \right\}, \tag{3}$$

where the initial state x_0 is deterministic. Since solving the objective function (3) directly is computationally intractable, Bellman's equation is introduced so that the optimal control can be computed recursively. To handle continuous problems, it is more convenient for our purpose but mathematically equivalent to write Bellman's equation (1) with the expectation form (rather than the traditional use of a one-step transition matrix)

$$V_t(x_t) = \sup_{u_t \in \mathcal{U}} \{C(x_t, u_t) + \gamma \mathbb{E}[V_{t+1}(x_{t+1})|x_t]\}, \tag{4}$$

where $V_t(x_t)$ is the value function representing the value of being in state x_t by following the optimal policy onward and the expectation is taken over the random information variable W_{t+1}. It is worth noting that the contributions in (4) can be stochastic. Then, Bellman's equation becomes

$$V_t(x_t) = \sup_{u_t \in \mathcal{U}} \left\{ \mathbb{E} \left[\hat{C}(x_t, u_t, W_{t+1}) + \gamma V_{t+1}(S^M(x_t, u_t, W_{t+1}))|x_t \right] \right\}. \tag{5}$$

Furthermore, if we have an infinite horizon steady state problem, we can drop the subscript t and write the Bellman's optimality equation as

$$V(x) = \sup_{u \in \mathcal{U}} \left\{ C(x, u) + \gamma \mathbb{E}[V(S^M(x, u, W)] \right\}. \tag{6}$$

Value functions define a partial ordering over policies. That is, $\pi \geq \pi'$ if and only if $V^{\pi}(x) \geq V^{\pi'}(x)$ for all $x \in \mathcal{X}$. Let V^* denote the optimal value function defined as

$$V^*(x) = \sup_{\pi \in \Pi} V^{\pi}(x), \tag{7}$$

for all $x \in \mathcal{X}$. It is well-known that the optimal value function V^* satisfies equation (6).

To solve problems with continuous states, we list the following assumptions regarding MDPs for future reference in later sections.

The state space $\mathcal{X}$, the decision space $\mathcal{U}$ and the outcome space $\mathcal{W}$ are convex, compact and Borel subsets of $\mathbb{R}^m$, $\mathbb{R}^n$ and $\mathbb{R}^l$ respectively.

Assume that the contribution function C, the state transition function S^M and the transition probability density function $Q : \mathcal{X} \times \mathcal{U} \times \mathcal{W} \to \mathbb{R}_+$ are all continuous.

It is worth noting the contribution and transition function are uniformly bounded given assumption 3.1 and 3.1. In turn, the objective function (3) is also bounded since we work with discounted problems. Let $C^b(\mathcal{X})$ denote the space of all bounded continuous functions from $\mathcal{X}$ to $\mathbb{R}$. It is well-known that $C^b(\mathcal{X})$ is a complete metric space.

3.2 Contraction operators

In this section, we describe the contraction operators associated with Markov decision processes. Their contraction property is crucial in the convergence proof of our algorithm.

[Bellman optimality operator] Let M be the Bellman operator such that for all $x \in \mathcal{X}$ and $V \in \mathbf{C}^b(\mathcal{X})$,

$$MV(x) = \sup_{u \in \mathcal{U}} \{C(x,u) + \gamma \int_{\mathcal{W}} Q(x,u,dw) V(S^M(x,u,w))\},$$

where Q is assumed to make M map $\mathbf{C}^b(\mathcal{X})$ into itself.

[Policy evaluation Operator] Let M_π be the operator for a fixed policy π such that for all $x \in \mathcal{X}$

$$M_\pi V(x) = C(x,\pi(x)) + \gamma \int_{\mathcal{W}} Q(x,\pi(x),dw) V(S^M(x,\pi(x),w))$$

where Q and S^M have the same property as in definition 3.2.

There are a few elementary properties of the operators M and M_π (Bertsekas & Shreve (1978)) that will play an important role in the subsequent sections.

Proposition 1 (Monotonicity) *For any $V_1, V_2 \in \mathcal{C}^b(\mathcal{X})$, if $V_1(x) \leq V_2(x)$ for all $x \in \mathcal{X}$,*

then for all $k \in \mathbb{N}$ *and* $x \in \mathcal{X}$

$$\begin{aligned} M^k V_1(x) &\leq M^k V_2(x), \\ M_\pi^k V_1(x) &\leq M_\pi^k V_2(x), \end{aligned}$$

where $M^k V(x) = M\left(M^{k-1}V(x)\right)$.

Proposition 2 (Contraction property and Fixed point) *M and M_π are γ-contractions for $\mathcal{C}^b(\mathcal{X})$ with respect to the supremum norm. Furthermore, M and M_π have their respective unique fixed points. That is, for any $V \in \mathcal{C}^b(\mathcal{X})$, $\lim_{k\to\infty} M^k V = V^*$ where V^* is the unique solution to the equation $V = MV$. Similarly, for any $V \in \mathcal{C}^b(\mathcal{X})$, $\lim_{k\to\infty} M_\pi^k V = V^\pi$ where V^π is the unique solution to the equation $V = M_\pi V$.*

3.3 Continuous correspondence

The correspondence Γ defined formally below describes the feasible set of decisions that ensures the Bellman optimality operator M takes the function space $\mathbf{C}^b(\mathcal{X})$ into itself. A correspondence is said to be compact-valued if the set $\Gamma(x)$ is compact for every $x \in \mathcal{X}$. The following definitions and properties of Γ will be used in section 5 to prove the convergence of policies to the optimal.

[Correspondence] A correspondence $\Gamma : \mathcal{X} \to 2^{\mathcal{U}}$ is a relation that assigns a feasible decision set $\Gamma(x) \subset \mathcal{U}$ for each $x \in \mathcal{X}$.

[Lower and upper hemi-continuity of correspondence] A correspondence $\Gamma : \mathcal{X} \to 2^{\mathcal{U}}$ is lower hemi-continuous (l.h.c.) at x if $\Gamma(x)$ is nonempty and if, for every $u \in \Gamma(x)$ and every sequence $x_n \to x$, there exists $N \geq 1$ and a sequence $\{u_n\}_{n=N}^{\infty}$ such that $u_n \to u$ and $u_n \in \Gamma(x_n)$ for all $n \geq N$.

The correspondence is upper hemi-continuous (u.h.c.) at x if $\Gamma(x)$ is nonempty and if, for every sequence $x_n \to x$ and every sequence $\{u_n\}$ such that $u_n \in \Gamma(x_n)$ for all n, there exists a convergent subsequence of $\{u_n\}$ with limit point $u \in \Gamma(x)$.

[Continuity of correspondence] A correspondence $\Gamma : \mathcal{X} \to 2^{\mathcal{U}}$ is continuous at $x \in \mathcal{X}$ if it is both u.h.c. and l.h.c. at x.

3.4 Markov chains with continuous state space

To work with Markov chains with continuous state space, we present the following definitions of irreducibility, invariant measure, recurrence and positivity that all have familiar counterparts in discrete chains. These properties are related to the stability of a Markov chain, which is of great importance in proving the convergence of value function estimates. In addition, the continuity property of the transition kernel is helpful in defining behavior of chains with desirable topological structure of the state space (Meyn & Tweedie (1993)). Hence, we introduce the concepts of Feller chains, petite sets and T-chains, which will be used later to classify positive Harris chains.

[ψ-Irreducibility for general space chains] For any measure φ, a Markov chain Φ on state space $\mathcal{X}$ is called φ-irreducible if there exists a measure φ on $\mathcal{B}(\mathcal{X})$ such that whenever $\varphi(A) > 0$ for $A \in \mathcal{B}(\mathcal{X})$, we have

$$\mathbb{P}_x\{\Phi \text{ ever enters } A\} > 0, \forall x \in \mathcal{X}$$

where $\mathbb{P}_x$ denotes the conditional probability on the event that the chain starts in state x. Let ψ be the maximal irreducibility measure among such measures. (For the existence of ψ, see proposition 4.2.2 of Meyn & Tweedie (1993).)

[Harris recurrence] The set $A \in \mathcal{B}(\mathcal{X})$ is called Harris recurrent if

$$\mathbb{P}_x\{\Phi \in A \text{ infinitely often}\} = 1, \forall x \in \mathcal{X}.$$

A chain Φ is called Harris (recurrent) if it is ψ-irreducible and every set in

$$\mathcal{B}^+(\mathcal{X}) = \{A \in \mathcal{B}(\mathcal{X}) : \psi(A) > 0\}$$

is Harris recurrent.

[Invariant measure] Let $P(\cdot, \cdot)$ be the transition kernel of a chain Φ on the state space $\mathcal{X}$.

A σ-finite measure μ on $\mathcal{B}(\mathcal{X})$ with the property

$$\mu(A) = \int_{\mathcal{X}} \mu(dx) P(x, A), \forall A \in \mathcal{B}(\mathcal{X})$$

will be called invariant.

[Positive chains] Suppose a ψ-irreducible chain Φ admits an invariant probability measure μ. Then Φ is called a positive chain.

[Weak Feller chains] If a chain Φ has a transition kernel P such that $P(\cdot, O)$ is a lower semi-continuous function for any open set $O \in \mathcal{B}(\mathcal{X})$, then Φ is called a weak Feller chain.

It is worth noting that the weak Feller property is often defined by assuming that the transition kernel P maps the set of all continuous functions $\mathcal{C}(\mathcal{X})$ into itself.

[Petite set] A set $C \in \mathcal{B}(\mathcal{X})$ is called petite if for some non-trivial measure ν on $\mathcal{B}(\mathcal{X})$ and $\delta > 0$,

$$K(x, A) \geq \delta \nu(A), x \in C, A \in \mathcal{B}(\mathcal{X})$$

where K is the resolvent kernel defined by

$$K(x, A) = \sum_{n=0}^{\infty} (\frac{1}{2})^{n+1} P^n(x, A).$$

[T-chains] If every compact set of $\mathcal{B}(\mathcal{X})$ is petite, then Φ is called a T-chain. (For another more detailed definition, see Meyn & Tweedie (1993) chapter 6.)

3.5 Post-decision state variable

Computing the expectation within the max operator M is often intractable when the underlying distribution of the evolution of the stochastic system is unknown or the decision u is a vector. However, we can circumvent the difficulty by introducing the notion of the post-decision state variable (Van Roy et al. (1997), Powell (2007)) or end-of-state (Judd (1998)) or after-state (Sutton & Barto (1998)). Suppose we can break the original transition function

(2) into the two steps

$$x_t^u = S^{M,u}(x_t, u_t), \tag{8}$$

$$x_{t+1} = S^{M,W}(x_t^u, W_{t+1}). \tag{9}$$

We call x_t^u the post-decision state, which is the state immediately after we make a decision. Using our original energy allocation example, we let x_t be supplies of and demands for energy at time t, u_t be how much to use at time t, and W_{t+1} be random changes in energy from wind and solar, changes in prices and changes in demand. Then, comparable to equation (8) and (9) we have

$$\begin{aligned} x_t^u &= x_t - Au_t \\ x_{t+1} &= x_t^u + W_{t+1}. \end{aligned}$$

We denote the post-decision state space by $\mathcal{X}^u$. Accordingly, we define the post-decision value function $V^u : \mathcal{X}^u \to \mathbb{R}$ to be

$$V^u(x_t^u) = \mathbb{E}\{V(x_{t+1})|x_t^u\}, \tag{10}$$

where $V^u(x_t^u)$ represents the value of being in the post decision states x_t^u. Suppose the pre-decision value function $V \in \mathcal{B}(\mathcal{X})$ and $V^u \in \mathcal{B}(\mathcal{X}^u)$. There is a simple relationship between the pre-decision value function $V(x_t)$ and post-decision value function $V^u(x_t^u)$ that is summarized as

$$V(x_t) = \max_{u_t \in \mathcal{U}} \{C(x_t, u_t) + \gamma V^u(x_t^u)\}. \tag{11}$$

By substituting (11) into (10), we have Bellman's equation of post-decision value function

$$V^u(x_t^u) = \mathbb{E}\{\max_{u_{t+1} \in \mathcal{U}} \{C(x_{t+1}, u_{t+1}) + \gamma V^u(x_{t+1}^u)\}|x_t^u\}. \tag{12}$$

We note that it is popular in the reinforcement learning community to approximate Q-factor $Q(x, u)$. In our approach, we only need to approximate $V^u(x^u)$, where the dimensionality of the post-decision state x^u is much lower than the dimensionality of the state-action pair (x, u).

Step 0: Initialization:

Step 0a Set the initial values of the value function parameters $\hat{\theta}_0$.

Step 0b Set the initial policy

$$\pi_1(x) = \arg\max_{u \in \Gamma(x)} \{C(x,u) + \gamma\phi(x^u)^T\hat{\theta}_0\}.$$

Step 0c Set the iteration counter $n = 1$.

Step 1: Do for $n = 1, \ldots, N$,

Step 1a: Set the initial State x_0^n.

Step 2: Do for $m = 0, \ldots, M$:

Step 3: Initialize $\hat{\theta}_{n,m}$ and $\hat{v}_m = 0$.

Step 4: Draw randomly or observe W_{m+1} from the chain.

Step 5: Do the following:

Step 5a: Set $u_m^n = \pi_n(x_m^n)$.

Step 5b: Compute $x_m^{n,\pi} = S^{M,\pi}(x_m^n, u_m^n)$ and $x_{m+1}^n = S^M(x_m^n, u_m^n, W_{m+1})$

Step 5c: Compute $u_{m+1}^n = \pi_n(x_{m+1}^n)$ and $x_{m+1}^{n,\pi} = S^{M,\pi}(x_{m+1}^n, u_{m+1}^n)$

Step 5d: Compute input variable/regressor using the corresponding basis function values: $\phi(x_m^{n,\pi}) - \gamma\phi(x_{m+1}^{n,\pi})$.

Step 6: Do the following:

Step 6a Compute/observe the response variable $\hat{v}_m = C(x_m^{n,\pi}, x_{m+1}^{n,\pi})$

Step 6b Update parameters $\hat{\theta}_{n,m}$ with LS/RLS method that regresses response $\hat{v}_m$ on regressor $\phi(x_m^{n,\pi}) - \gamma\phi(x_{m+1}^{n,\pi})$

Step 7: Update the parameter and the policy:

$$\hat{\theta}_{n+1} = \hat{\theta}_{n,M},$$

$$\pi_{n+1}(x) = \arg\max_{u \in \Gamma(x)} \{C(x,u) + \gamma\phi(x^u)^T\hat{\theta}_n\}.$$

Step 8: Return the policy π_{N+1} and parameters $\hat{\theta}_N$.

Figure 2: Infinite-horizon approximate policy iteration algorithm with recursive least squares method

3.6 Algorithm details

The recursive least squares approximate policy iteration algorithm (RLSAPI) is summarized in Figure 2. The details of the least squares subroutine in step 3 for updating the value function parameters are discussed in section 4. It is worth making a remark on the arg max function at the end of the algorithm. This step is usually a multivariate global optimization problem that computes the policy (exactly or approximately) from the post-decision value function of the

previous inner iteration. The updated policy function feeds back a decision given any fixed input of the state variable. We assume that there is a tie-breaking rule that determines a unique solution to the arg max function for all $f \in \mathcal{C}^b(\mathcal{X})$ such as a nonlinear proximal point algorithm (Luque (1987)). As a result, the returned policies are well-defined single-valued functions. It is worth noting that determining the unique solution to the arg max function may not be an easy job in practice. However, the computational difficulty is significantly reduced if we have special problem structures such as strict concavity and differentiability of the value functions as in the blood management example mentioned in the introduction.

4 Almost sure convergence of policy evaluation

In the RLSAPI algorithm, there are both inner and outer loops, which correspond to policy evaluation and policy improvement respectively. We first analyze the convergence of the policy evaluation step. RLSAPI uses the least squares temporal difference (LSTD) learning algorithm Bradtke & Barto (1996) for evaluating a fixed policy.

For a fixed policy π, the transition steps (8) and (9) become

$$\begin{aligned} x_t^\pi &= S^{M,\pi}(x_t, \pi(x_t)), \\ x_{t+1} &= S^{M,W}(x_t^\pi, W_{t+1}). \end{aligned}$$

As a result, the Markov decision problem can be reduced to a Markov chain for post-decision states. Bellman's equation (12) for the post-decision state becomes

$$V^\pi(x) = \int_{\mathcal{X}^\pi} P(x, dx')(C^\pi(x, x') + \gamma V^\pi(x')), \tag{13}$$

where V^π is the value of following the fixed policy π, $P(\cdot, \cdot)$ is the transition probability function of the chain, $C^\pi(\cdot, \cdot)$ is the stochastic contribution/reward function with $C^\pi(x_t^\pi, x_{t+1}^\pi) = C(x_{t+1}, \pi(x_{t+1}))$ and $\mathcal{X}^\pi$ is the post decision state space by following policy π. It is worth noting that $\mathcal{X}^\pi$ is compact since $\mathcal{X}$ and $\mathcal{U}$ are compact and the state transition function S^M is continuous by assumption 3.1 and 3.1 respectively. In addition, x in (13) is the post-decision state variable and we drop the superscript u for simplicity.

In the algorithm, we choose to work with the post-decision value functions resulting from following a fixed policy π (call it policy value function) because in this way we can avoid computing the expectation of the pre-decision value function directly in policy improvement, which is often hard to evaluate. To make the policy improvement step easier and show convergence of the algorithm, we make the following assumption on the policy value function.

Assume that the policy value function for a fixed policy π is continuous and of the linear form i.e. $V^{\pi}(x|\theta) = \phi(x)^T\theta$ where $\phi(x) = [\cdots, \phi_f(x), \cdots]$ is the vector of basis functions of dimension $F = |\mathcal{F}|$ (number of basis functions) and $f \in \mathcal{F}$ ($\mathcal{F}$ denotes the set of features).

Since we assume that the spanning set of the basis functions is known, so it is enough to just estimate the linear parameters for estimating policy value function. It is worth mentioning that the features for the post-decision value functions may not be the same and may not be easily deduced from those of the pre-decision value function (if we assume a linear structure on pre-decision value function), since there is an additional expectation operator as in equation (11), ie. $\phi(x^u)^T\theta = \int_W Q(x, u, dw)\phi'(S^M(x, u, w))^T\theta'$, where ϕ' stands for the features of the value function defined on pre-decision states. An exception is LQR, in which case the features are identical for both pre- and post-decision value functions.

There are two different approaches to derive convergence results for LSTD with finite state space. Bradtke & Barto (1996) employs a general linear regression setting, while Lagoudakis & Parr (2003) considers the least squares fixed point solution of the Bellman operator projected on the feature space. It looks more convenient to extend the convergence result to the continuous case in the regression setting, since the derivation in Lagoudakis & Parr (2003) is very specific to finite states. However, we believe that the intuitive interpretation of the least squares fixed point approximation approach discussed in Lagoudakis & Parr (2003) remains valid in the continuous case.

We proceed by following the regression setting in Bradtke & Barto (1996). Bellman's equation (13) gives us

$$\phi(x)^T\theta^* = \int_{\mathcal{X}^{\pi}} P(x, dx')[C^{\pi}(x, x') + \gamma\phi(x')^T\theta^*].$$

We can rewrite the fixed point equation as

$$C^{\pi}(x,x') = \left(\phi(x) - \gamma \int_{\mathcal{X}^{\pi}} P(x,dx')\phi(x')\right)^{T} \theta^{*} + C^{\pi}(x,x') - \int_{\mathcal{X}^{\pi}} P(x,dx')C^{\pi}(x,x'),$$

which fits the linear regression setting.

Remark: It is worth noting that $\phi(x)$ and $\phi(x')$ are vectors. The integral is taken componentwise for $\phi(x')$, so it feeds back a vector. Similarly, if we take an integral of a matrix, it is taken componentwise.

Since the transition probability function may be unknown or not computable at iteration m, instead of having the exact input variable, we can only observe an unbiased sample estimate $\phi_m - \gamma\phi_{m+1}$ where ϕ_m is the shorthand notation for $\phi(x_m)$. Therefore, this is an errors-in-variable model (Young (1984)), and the regular linear regression estimates for θ^* is biased. An instrumental variable is used in the regression estimates to eliminate the asymptotic bias. The instrumental variable has to be correlated with the true input variable but uncorrelated with the input error term and the observation error term. A good candidate for the instrumental variable in the algorithm is ϕ. As a result, the m-th estimate of θ^* is

$$\theta_m = \left[\frac{1}{m+1}\sum_{i=0}^{m} \phi_i(\phi_i - \gamma\phi_{i+1})^{T}\right]^{-1} \left[\frac{1}{m+1}\sum_{i=0}^{m} \phi_i C_i\right], \tag{14}$$

where $C_i = C^{\pi}(x_i, x_{i+1})$ is the i-th observation of the contribution.

Convergence of the parameter estimates requires stability of the chain. This can be interpreted as saying a chain finally settles down to a stable regime independent of its initial starting point (Meyn & Tweedie (1993)). Positive Harris chains defined in section 3 meet the stability requirement precisely, and the invariant measure μ describes the stationary distribution of the chain. The following lemma from Meyn & Tweedie (1993) states the well-known strong law of large numbers for positive Harris chains, which is essential in the convergence proof of theorem 1.

Lemma 4.1 (Law of large numbers for positive Harris chains) *If Φ is a positive Harris chain (see definition 3.4 and 3.4) with invariant probability measure μ, then for each*

$f \in L_1(\mathcal{X}, \mathcal{B}(\mathcal{X}), \mu)$,

$$\lim_{n\to\infty} \frac{1}{n+1} \sum_{i=0}^{n} f(x_i) = \int_{\mathcal{X}} \mu(dx) f(x)$$

almost surely.

In short, a positive Harris chain (analogous to an ergodic Markov chain in the discrete case) has the highly desirable property that the Strong Law of Large Numbers (SLLN) holds for the chain independent of the initial state of the chain, so that we can show convergence of parameter estimates of policy value functions. There are some sufficient conditions for positive Harris recurrence such as irreducibility, aperiodicity or "minorization" hypotheses on the transition kernel (Meyn & Tweedie (1993), Nummelin (1984)), but they are hard to verify. Owing to our assumptions on compact $\mathcal{X}$ and $\mathcal{U}$ and continuous transition probability function Q, the following proposition states that the only global requirement for each controlled Markov chain Φ^π being positive Harris is an irreducibility condition of the post-decision state space $\mathcal{X}^\pi$.

Proposition 3 *Under assumption 3.1 and 3.1, suppose the controlled Markov chain Φ^π for a fixed policy π with state space $\mathcal{X}^\pi$ is ψ-irreducible and the support of ψ has non-empty interior. Then, Φ^π is positive Harris.*

Proof:

Φ^π is a weak Feller chain (see definition 3.4), since the transition function is continuous and the transition probability function Q has the Feller property. In addition, the state space $\mathcal{X}^\pi$ of Φ^π is compact. By the irreducibility hypothesis on $\mathcal{X}^\pi$ and theorem 6.2.9 of Meyn & Tweedie (1993), Φ^π is a T-Chain (see definition 3.4). By theorem 6.2.5 of Meyn & Tweedie (1993), every compact set in $\mathcal{B}(\mathcal{X}^\pi)$ is petite (see definition 3.4). As a result, $\mathcal{X}^\pi$ is petite since it is compact. By proposition 9.1.7 of Meyn & Tweedie (1993), Φ^π is Harris recurrent. By theorem 10.4.4 of Meyn & Tweedie (1993), Φ^π admits a unique (up to a constant multiple) invariant measure. Finally, by theorem 10.4.10 of Meyn & Tweedie (1993), Φ^π is positive Harris. ∎

ψ-irreducibility is easier to verify than the standard definition of irreducibility for countable chains, and it is often simple to establish problem-dependent "grossly sufficient" conditions Meyn & Tweedie (1993). Hence, we impose the following assumption on the policy space Π in order to prove later convergence results of our algorithm.

Assume that, for all $\pi \in \Pi$, the MDP is reduced to a ψ-irreducible Markov chain Φ^π with state space $\mathcal{X}^\pi$ and the support of ψ has non-empty interior.

Invertibility of the correlation matrix between the instrumental variable and input variable is another requirement to show convergence of the parameter estimates. For problems with a finite state space, Bradtke & Barto (1996) assumes that the number of linearly independent basis functions is the same as the dimension of the state variable so that the correlation matrix between the input and instrumental variables is invertible. However, this assumption defeats the purpose of using a compact representation of the value function such as a linear approximation with basis functions, since the complexity of computing parameter estimates is the same as estimating a look-up table value function directly. In fact, invertibility only requires that the feature matrix (which is an $n \times m$ matrix, n being the number of states and m the number of features) is of full rank, i.e. the rank of the feature matrix is m for $n > m$. However, we do not have the feature matrix in the continuous case since the number of states is uncountably infinite. The following two lemmas show how the invertibility issue are handled for continuous problems.

When we have linearly independent basis functions, the correlation matrix of input and instrumental variables is not guaranteed to be invertible. However, it is reasonable to assume the non-singularity of the correlation matrix. The set of singular $n \times n$ matrices is a null subset with respect to Lebesgue measure over the field of $\mathbb{R}^{n \times n}$. In other words, if you randomly pick a square matrix over the real numbers, the probability that it is singular is zero. The following lemma states that the situation where the correlation matrix is non-singular is extremely rare in the continuous case and the proof is omitted for the known result.

Lemma 4.2 *Suppose the basis functions $\phi = (\phi_1, \ldots, \phi_F)$ are non-zero and linearly independent μ-almost surely where μ is the invariant probability measure of the Markov chain. Then*

the correlation matrix

$$\int_{\mathcal{X}^\pi} \mu(dx)\phi(x)\left(\phi(x) - \gamma \int_{\mathcal{X}^\pi} P(x, dx')\phi(x')\right)^T$$

is invertible for all but at most finitely many $\gamma \in (0, 1)$.

By imposing stronger assumptions on the basis functions and the discount factor, lemma 4.3 shows how the invertibility of the correlation matrix is guaranteed in the continuous case. The result is used in section 7 where orthonormality of basis functions can be obtained by construction. We further describe a way to modify the parameter estimates of the algorithm so that we can avoid the strong assumption on the discount factor.

Lemma 4.3 (Non-singularity of the correlation matrix) *Suppose we have orthonormal basis functions with respect to the invariant measure* μ *of the Markov chain and* $\lambda < \frac{1}{F}$ *where* F *is the number of basis functions. Then the correlation matrix*

$$\int_{\mathcal{X}^\pi} \mu(dx)\phi(x)\left(\phi(x) - \lambda \int_{\mathcal{X}^\pi} P(x, dx')\phi(x')\right)^T$$

is invertible.

Proof:

For shorthand notation, we write $\int_{\mathcal{X}^\pi} \mu(dx)\phi_i(x) = \mu\phi_i$ and $\int_{\mathcal{X}^\pi} P(x, dx')\phi_i(x') = P\phi_i(x)$. It is worth noting that $\mu\phi_i$ is a constant and $P\phi_i(x)$ is a function of x. Then, we can write the correlation matrix explicitly as the $F \times F$ matrix

$$C = \begin{pmatrix} \mu\phi_1^2 - \lambda\mu\phi_1 P\phi_1 & \mu\phi_1\phi_2 - \lambda\mu\phi_1 P\phi_2 & \cdots & \mu\phi_1\phi_F - \lambda\mu\phi_1 P\phi_F \\ \mu\phi_2\phi_1 - \lambda\mu\phi_2 P\phi_1 & \mu\phi_2^2 - \lambda\mu\phi_2 P\phi_2 & \cdots & \mu\phi_2\phi_F - \lambda\mu\phi_2 P\phi_F \\ \vdots & \vdots & & \vdots \\ \mu\phi_F\phi_1 - \lambda\mu\phi_F P\phi_1 & \cdots & \cdots & \mu\phi_F^2 - \lambda\mu\phi_F P\phi_F \end{pmatrix}.$$

Since $\phi(s)$ is a vector of orthonormal basis functions with respect to the invariant measure μ, we have

$$C = \begin{pmatrix} 1 - \lambda\mu\phi_1 P\phi_1 & -\lambda\mu\phi_1 P\phi_2 & \cdots & -\lambda\mu\phi_1 P\phi_F \\ -\lambda\mu\phi_2 P\phi_1 & 1 - \lambda\mu\phi_2 P\phi_2 & \cdots & -\lambda\mu\phi_2 P\phi_F \\ \vdots & \vdots & & \vdots \\ -\lambda\mu\phi_F P\phi_1 & \cdots & \cdots & 1 - \lambda\mu\phi_F P\phi_F \end{pmatrix}.$$

By applying Jensen's inequality on P, we have that

$$\mu(P\phi_i)^2 \leq \mu P\phi_i^2 = \mu\phi_i^2 = 1.$$

Then, for all $1 \leq i, j \leq F$ we have

$$\mu\phi_i P\phi_j \leq \frac{\mu\phi_i^2 + \mu(P\phi_j)^2}{2} \leq 1.$$

Similarly, we have $\mu\phi_i P\phi_j \geq -1$ for all $1 \leq i, j \leq F$. As a result, $C = I - \lambda A$ where A is a matrix with entries $A_{i,j} \in [-1, 1]$. C is invertible iff $|C| \neq 0$, so it suffices to show that $|A - \frac{1}{\lambda} I| \neq 0$. In other words, we need to show that $\frac{1}{\lambda}$ is not a real eigenvalue of A. Suppose λ is an eigenvalue of A and its corresponding eigenvector is v. Let $\bar{v} = \max_{1 \leq i \leq F}\{|v_i|\}$. Since $A_{i,j} \in [-1, 1]$, $-F\bar{v} \leq \lambda\bar{v} \leq F\bar{v}$. Hence, all the real eigenvalues of A are bounded between $-F$ and F. This implies that C is invertible if $\frac{1}{\lambda} > F$. Equivalently, C is non-singular if $\lambda < \frac{1}{F}$. ∎

In general, the discount factor γ in MDPs may be larger than $\frac{1}{F}$, which can be quite small as the number of basis functions increases. However, we can modify the parameter estimates of the algorithm by collapsing k transitions ($x_0 \rightarrow x_k$) into 1 transition so that we can still use the previous lemma. Since $\gamma \in (0, 1)$, there exists $k \in \mathbb{N}$ such that $\gamma^k < \frac{1}{F}$. The following fixed point equation must be satisfied if we keep substituting Bellman's equation (13) back into itself $k - 1$ times:

$$V^\pi(x_0) = \int_{\mathcal{X}^\pi \times \ldots \times \mathcal{X}^\pi} \prod_{i=0}^{k-1} P(x_i, dx_{i+1}) \left\{ \sum_{i=0}^{k-1} \gamma^i C^\pi(x_i, x_{i+1}) + \gamma^k V^\pi(x_k) \right\}.$$

Hence, we can rewrite the linear model as

$$\begin{aligned} \sum_{i=0}^{k-1} \gamma^i C^\pi(x_i, x_{i+1}) &= (\phi(x_0) - \gamma^k \int_{\mathcal{X}^\pi \times \ldots \times \mathcal{X}^\pi} \prod_{i=0}^{k-1} P(x_i, dx_{i+1})\phi(x_k))^T \theta^* \\ &+ \sum_{i=0}^{k-1} \gamma^i C^\pi(x_i, x_{i+1}) - \int_{\mathcal{X}^\pi \times \ldots \times \mathcal{X}^\pi} \prod_{i=0}^{k-1} P(x_i, dx_{i+1}) \sum_{i=0}^{k-1} \gamma^i C^\pi(x_i, x_{i+1}). \end{aligned}$$

Finally, the parameter estimates become

$$\theta'_m = \left[\frac{1}{m+1} \sum_{i=0}^{m} \phi_i(\phi_i - \gamma^k \phi_{i+k})^T \right]^{-1} \left[\frac{1}{m+1} \sum_{i=0}^{m} \phi_i \sum_{j=0}^{k-1} \gamma^j C_{i+j} \right]. \tag{15}$$

The proof of lemma 4.3 still goes through with the replacement of P with P^k and λ with γ^k. Since we assume the known linear structure of the policy value function, the convergence of policy evaluation is reduced to the convergence of linear parameter estimates, as stated in the following theorem. The proof is omitted since the technique is standard in proving convergence of regression estimates.

Theorem 1 *Under assumptions 3.1, 3.1 and 4, suppose that the controlled Markov chain for a fixed policy $\pi \in \Pi$ with state space $\mathcal{X}^\pi$ has transition kernel $P(x, dy)$ and invariant probability measure μ. Further assume that the policy value function V^π satisfies assumption 4 with known basis functions that are either linearly independent (with invertible correlation matrix) or orthonormal with respect to the invariant measure μ. Then, $\theta_m \to \theta^*$ and $\theta'_m \to \theta^*$ μ-almost surely.*

In practice, estimating θ using least squares, which requires matrix inversion at each iteration, is computationally expensive. Instead, recursive least squares method is used to obtain the well-known updating formulas (Bradtke & Barto (1996)):

$$\epsilon_m = C_m - (\phi_m - \gamma\phi_{m+1})^T \theta_{m-1}, \tag{16}$$

$$B_m = B_{m-1} - \frac{B_{m-1}\phi_m(\phi_m - \gamma\phi_{m+1})^T B_{m-1}}{1 + (\phi_m - \gamma\phi_{m+1})^T B_{m-1}\phi_m}, \tag{17}$$

$$\theta_m = \theta_{m-1} + \frac{\epsilon_m B_{m-1}\phi_m}{1 + (\phi_m - \gamma\phi_{m+1})^T B_{m-1}\phi_m}. \tag{18}$$

In the initialization of θ_0 and B_0, θ_0 can be any finite vector and B_0 is usually chosen to be βI for some small positive constant β. The following corollary states that the RLS estimate of θ^* also converges to the true parameters. The proof only requires simple calculation and is virtually the same as the proof of theorem 1, so it is omitted.

Corollary 4.1 *Suppose we have the same assumptions as in theorem 1. Further assume that $1+(\phi_m-\gamma\phi_{m+1})^T B_{m-1}\phi_m \neq 0$ for all m, then $\theta_m \to \theta^*$ almost surely using recursion formulas (16), (17) and (18).*

5 Almost sure convergence of exact policy iteration

Convergence of the exact policy iteration is well known (see Bertsekas & Shreve (1978)). The primary result is stated as follows:

Proposition 4 *Let $(\pi_n)_{n=0}^{\infty}$ be a sequence of policies generated recursively as follows: given an initial policy π_0, for $n \geq 0$,*

$$\pi_{n+1}(x) = \arg\max_u \{C(x,u) + \gamma \int_{\mathcal{W}} Q(x,u,dw) V^{\pi_n}(S^M(x,u,w))\}.$$

Then $V^{\pi_n} \to V^*$ uniformly where V^* is the optimal value function.

In this section, we fit the above convergence result in the LSPI setting when we assume we can evaluate the expected performance of the policy exactly for all states in an almost sure sense. Owing to the assumptions on the value function and state space, we obtain almost sure convergence not only for the value functions and but also for the policies. The result then sets the theoretical background for the next section where we show convergence in expectation for a form of approximate policy iteration.

We would like to show the almost sure convergence of the algorithm so that we build up the corresponding probability space and related probability measure. For a fixed policy π as in the previous section, we follow the convention of letting the sample space Ω^π be $(\mathcal{X}^\pi)^\infty$, which is the whole history of the Markov chain, and the corresponding σ-algebra $\mathcal{F}^\pi$ be the product Borel σ-algebra on Ω^π. Assume that for each initial state $x_0 \in \mathcal{X}^\pi$ of the chain, there exists a probability measure $\mathbb{P}^\pi_{x_0}$ that governs the probability law of the chain on $(\Omega^\pi, \mathcal{F}^\pi)$. If an event E occurs $\mathbb{P}^\pi_{x_0}$-a.s. for all $x_0 \in \mathcal{X}^\pi$ then we write that E occurs $\mathbb{P}^\pi_*$-a.s. In the case of a positive Harris chain, $\mathbb{P}^\pi_*$ corresponds to the invariant probability measure μ^π of the chain and the convergence results in theorem 1 do not depend on the initial state of the chain. Hence, we have that for a fixed policy π the parameter estimates converge to the true value $\mathbb{P}^\pi_*$-a.s.

For the policy iteration algorithm, we consider the product sample space $\Omega = \Omega^{\pi_0} \times \Omega^{\pi_1} \times \cdots$ and the corresponding product σ-algebra $\mathcal{F} = \mathcal{F}^{\pi_0} \oplus \mathcal{F}^{\pi_1} \oplus \cdots$. It is worth noting that Ω^{π_n} depends on the sample spaces of previous inner iterations, $\Omega^{\pi_0}, \cdots, \Omega^{\pi_{n-1}}$. In other words,

any element $\omega \in \Omega$ has the form

$$\omega = (\omega^{\pi_0}, \omega^{\pi_1}(\omega^{\pi_0}), \omega^{\pi_2}(\omega^{\pi_0}, \omega^{\pi_1}(\omega^{\pi_0})), \cdots).$$

Moreover, we need to point out that, except for π_0, all the policy updates π_n's are random and depend on the sample spaces of previous inner iterations. To reflect the dependence, we let the probability measure for Ω^{π_n} be $\mathbb{P}_*^{\pi_n|\pi_0,\cdots,\pi_{n-1}}$. Then, the probability measure for Ω is $\mathbb{P}_* = \mathbb{P}_*^{\pi_0} \cdot \mathbb{P}_*^{\pi_1|\pi_0} \cdot \mathbb{P}_*^{\pi_2|\pi_0,\pi_1} \cdot \cdots$.

Before getting to the convergence theorem, we present the following preliminaries. Lemma 5.1 proves the convergence of a sequence of monotonic policies to the optimal policy given the convergence of corresponding value functions, and Lemma 5.2 shows that integration preserves certain properties of the integrand such as boundedness and continuity required in the proof of the theorem 2. The proofs are omitted for brevity since the technique is standard (see Stokey et al. (1989)).

Lemma 5.1 *Suppose $\mathcal{X}$ and $\mathcal{U}$ satisfy assumption 3.1 and the correspondence Γ is nonempty, compact and convex-valued, and continuous for each $x \in \mathcal{X}$. Let $f_n : \mathcal{X} \times \mathcal{U} \to \mathbb{R}$ be a sequence of continuous functions for each n. Assume f has the same properties and $f_n \to f$ uniformly. Define π_n and π by $\pi_n(x) = \arg\max_{u \in \Gamma(x)} f_n(x,u)$ and $\pi(x) = \arg\max_{u \in \Gamma(x)} f(x,u)$. Then, $\pi_n \to \pi$ pointwise. If $\mathcal{X}$ is compact, $\pi_n \to \pi$ uniformly.*

Lemma 5.2 *Suppose $\mathcal{X}$,$\mathcal{U}$,$\mathcal{W}$, Q satisfy assumptions 3.1 and 3.1. If $g : \mathcal{X} \times \mathcal{U} \times \mathcal{W} \to \mathbb{R}$ is bounded and continuous, then $Tg(x,u) = \int_{\mathcal{W}} Q(x,u,dw) g(x,u,w)$ is also bounded and continuous.*

Theorem 2 (Almost sure convergence of exact LSPI) *Suppose assumptions 3.1, 3.1, 4 hold and Γ satisfies the same condition as in lemma 5.1. Further assume that for all $\pi \in \Pi$ the policy value function V^π satisfies the same assumption as in theorem 1. Let n be the iteration counter for policy improvement and m_n be the number of samples in iteration n for policy evaluation. Assume $m_n \to \infty$ for all n. As $n \to \infty$, the exact policy iteration algorithm converges. That is $\pi_n \to \pi^*$ uniformly $\mathbb{P}_*$-a.s.*

Proof:

We divide the proof into two parts. (a) Suppose that, by following the fixed policy π_n, we can obtain the exact post-decision value function in the linearly parameterized form of $\phi(x^u)'\theta_n^*$. Then define

$$f_n(x,u) = C(x,u) + \gamma \int_{\mathcal{W}} Q(x,u,dw) V^{\pi_n}(S^M(x,u,w)) = C(x,u) + \gamma \phi(x^u)'\theta_n^*,$$

and

$$f(x,u) = C(x,u) + \gamma \int_{\mathcal{W}} Q(x,u,dw) V^{*}(S^M(x,u,w)).$$

Let $V_n = M^n V^{\pi_0}$ and define

$$\tilde{f}_n(x,u) = C(x,u) + \gamma \int_{\mathcal{W}} Q(x,u,dw) V_n(S^M(x,u,w)).$$

By proposition 4, $V^{\pi_n} \nearrow V^*$ uniformly, so $f_n \nearrow f$ pointwise. $\tilde{f}_n \nearrow f$ uniformly by Stokey et al. (1989). Since policy evaluation and improvement are both exact, by proposition 4, we have $V_n \leq V^{\pi_n}$, so $\tilde{f}_n \leq f_n$. In turn, we conclude that $f_n \nearrow f$ uniformly. Since the contribution function and value functions for fixed policies are continuous, they are uniformly bounded on the compact state space $\mathcal{X}$. By lemma 5.2 we have f_n and f are bounded and continuous. By lemma 5.1, $\pi_n \to \pi^*$ pointwise. Since $\mathcal{X}$ is compact, the convergence is uniform.

(b) By assumption, in the inner loop the system evolves according to a positive Harris chain given a fixed policy. By theorem 1 or corollary 4.1, for a fixed initial policy π_0 (n=0), $\theta_{0,m_0} \to \theta_0^*$ $\mathbb{P}_*^{\pi_0}$-a.s. Denote the almost sure set by $\bar{\Omega}^{\pi_0}$. For any $\omega^{\pi_0} \in \bar{\Omega}^{\pi_0}$, we can obtain the exact post-decision value function $\phi(x^a)'\theta_0^*$ and, in turn, the exact policy update π_1. Similarly, there exists a $\mathbb{P}_*^{\pi_1|\pi_0}$-a.s. set $\bar{\Omega}^{\pi_1}(\bar{\Omega}^{\pi_0})$ such that $\theta_{1,M} \to \theta_1^*$ to obtain the exact post decision value function for π_1 and policy update π_2. If we keep going like this and let $\bar{\Omega} = \bar{\Omega}^{\pi_0} \times \bar{\Omega}^{\pi_1}(\bar{\Omega}^{\pi_0}) \times \cdots$, we have

$$\mathbb{P}_*(\bar{\Omega}) = \mathbb{P}_*^{\pi_0}(\bar{\Omega}^{\pi_0}) \cdot \mathbb{P}_*^{\pi_1|\pi_0}(\bar{\Omega}^{\pi_1}(\bar{\Omega}^{\pi_0})) \cdot \mathbb{P}_*^{\pi_2|\pi_0,\pi_1}(\bar{\Omega}^{\pi_2}(\bar{\Omega}^{\pi_0}, \bar{\Omega}^{\pi_1}(\bar{\Omega}^{\pi_0}))) \cdots = 1.$$

Since on $\bar{\Omega}$ the post decision value functions for π_n's and policy updates are obtained exactly, by part (a) we conclude that $\pi_n \to \pi^*$ uniformly $\mathbb{P}_*$-a.s. ∎

In practice, the condition of having a positive Harris chain in each policy evaluation step may not be guaranteed through the properties of the problem. For example, it might be that the chain is non-irreducible but has Harris decomposition, that is the chain can be decomposed into one transient set and a countable disjoint family of absorbing Harris sets with respective ergodic stationary measures. If one sample realization of the chain is trapped in some absorbing set (i.e. an atom of the chain) whose order is less than the number of basis functions, there is not sufficient support to identify the least squares parameter estimates. Hence, a properly designed exploration step such as adding a random exploration component to the policy function is necessary in an actual implementation of the algorithm.

A weaker condition than positive Harris chain is a chain that admits an invariant probability measure (not necessarily unique) on a rich enough full subset of the state space, which means the subset has measure 1 and suffices to identify the least squares parameter estimates. The condition is rather weak because there is no regularity or even irreducibility assumptions on the chain. If we can initialize the chain according to the invariant probability measure rather than from some fixed state, the chain becomes a stationary process and we can apply the Strong Law of Large numbers for stationary processes (Doob (1953)) to achieve convergence of the parameter estimates. In practice, we need to add a sampling stage to estimate the invariant probability measure in the inner loop of the algorithm and then initialize the chain according to the empirical probability measure obtained.

6 Convergence in the mean of approximate policy iteration

The exact policy iteration algorithm is only conceptual, since the policy evaluation step goes to infinity to achieve convergence. As a result, we introduce the approximate policy iteration algorithm, in which the policy evaluation stops in finite time. In our approximate policy iteration algorithm, the estimated value function of the approximate policy is random, since it depends on the sample path of the chain and also the iteration number of the inner loop. That is to say, for fixed $x \in \mathcal{X}$, $\hat{V}^{\hat{\pi}_n}(x)$ is a random variable. Given the state space being compact and the norm being the sup norm $||\cdot||_\infty$ for continuous functions, the following

theorem proves convergence in mean of the approximate policy iteration algorithm when both policy evaluations and policy updates are performed within error tolerances that converge to 0 according to a certain rate. In other words, the mean of the norm of the difference between the optimal value function and the estimated policy value function using approximate policy iteration converges to 0 if the successive approximations become better and better. The proof is omitted since it follows the same line as the proof of error bounds for approximate policy iteration with discrete and deterministic (or in almost sure sense) value function approximations in Bertsekas & Tsitsiklis (1996).

The only complication involved is the conditional expectation notation of $\mathbb{E}_{\hat{\pi}_n}[\cdot]$, which is defined to be the expectation taken with respect to the random sample path during iteration n by following the approximate policy $\hat{\pi}_n$ conditioning on the known sample paths from previous iterations of 1 through $n-1$. Mathematically, $\mathbb{E}_{\hat{\pi}_n}[\cdot] = \mathbb{E}_{\omega^{\hat{\pi}_n}}[\cdot|\omega^{\hat{\pi}_0}, \cdots, \omega^{\hat{\pi}_{n-1}}]$. It is worth pointing out that $\hat{\pi}_n$ conditioning on the known finite sample paths of previous iterations from 1 through $n-1$ is deterministic during iteration n, and so is the true value function of the approximate policy $V^{\hat{\pi}_n}$. As a result, in iteration n the only randomness comes from $\omega^{\hat{\pi}_n}$, the sample path by following $\hat{\pi}_n$. In other words, in our algorithm the sample paths $\omega^{\hat{\pi}_0}, \cdots, \omega^{\hat{\pi}_{n-1}}$ of previous iterations are known at iteration n, so $\mathbb{E}_{\hat{\pi}_n}[\cdot]$ feeds back a deterministic number.

Theorem 3 (convergence in the mean of approximate policy iteration) *Let $\hat{\pi}_0$, $\hat{\pi}_1$, $\ldots, \hat{\pi}_n$ be the sequence of policies generated by an approximate policy iteration algorithm and let $\hat{V}^{\hat{\pi}_0}, \hat{V}^{\hat{\pi}_1}, \ldots, \hat{V}^{\hat{\pi}_n}$ be the corresponding stochastic approximate value functions. Let $\{\epsilon_n\}$ and $\{\delta_n\}$ be positive scalars that bound the mean errors in approximations to value functions and policies (over all iterations) respectively, that is $\forall n \in \mathbb{N}$,*

$$\mathbb{E}_{\hat{\pi}_n}||\hat{V}^{\hat{\pi}_n} - V^{\hat{\pi}_n}||_\infty \leq \epsilon_n, \tag{19}$$

and

$$\mathbb{E}_{\hat{\pi}_n}||M_{\hat{\pi}_{n+1}}\hat{V}^{\hat{\pi}_n} - M\hat{V}^{\hat{\pi}_n}||_\infty \leq \delta_n. \tag{20}$$

Suppose the sequences $\{\epsilon_n\}$ *and* $\{\delta_n\}$ *converge to 0 and*

$$\lim_{n\to\infty} \sum_{i=0}^{n-1} \gamma^{n-1-i}\epsilon_i = \lim_{n\to\infty} \sum_{i=0}^{n-1} \gamma^{n-1-i}\delta_i = 0,$$

e.g. $\epsilon_i = \delta_i = \gamma^i$. *Then, this sequence eventually produces policies whose performance converges to the optimal performance in the mean:*

$$\lim_{n\to\infty} \mathbb{E}_{\hat{\pi}_n} ||\hat{V}^{\hat{\pi}_n} - V^*||_\infty = 0.$$

The above theorem can be directly applied to show that the LS/RLSAPI algorithm converges in the mean.

Corollary 6.1 (Convergence in the mean of LS/RLSAPI) *Under assumption 3.1, 3.1, and 4, suppose that for all policy* $\pi \in \Pi$ *the policy value function* V^π *satisfies the same assumption as in theorem 1. Then, theorem 3 holds for the LS/RLS approximate policy iteration algorithm.*

Proof:

We only need to check whether the conditions (19) and (20) in theorem 3 governing the error tolerances for evaluating policies are satisfied. If we can make policy updates exactly, we can take $\delta_n = 0$ for all n. If the policy update is done inexactly, that is using an approximate nonlinear proximal point algorithm, we can force the procedure to be within the error tolerance satisfying the condition on δ_n in theorem 3. Then, it suffices to show that the mean errors between the approximate and the true value functions of the approximate policy in each inner loop can be made arbitrarily small.

By the Cauchy-Schwartz inequality and the simple relations (10) and (11) between pre- and post-decision value functions, we have for all $x \in \mathcal{X}$,

$$|\hat{V}^{\hat{\pi}_n}(x) - V^{\hat{\pi}_n}(x)|^2 = \gamma^2 |\phi(x^a)^T(\theta_{n,M} - \theta_n^*)|^2 \leq \gamma^2 ||\phi(x^a)||_2^2 ||\theta_{n,M} - \theta_n^*||_2^2.$$

By assumption 3.1 and 3.1, $\mathcal{X}^\pi$ is compact, which implies that $||\phi(x^a)||_2 \leq c$ for some finite positive constant c for all x^a. Then,

$$||\hat{V}^{\hat{\pi}_n}(x) - V^{\hat{\pi}_n}(x)||_\infty \leq \gamma c||\theta_{n,M} - \theta_n^*||_2.$$

Recall that for fixed M, $\theta_{n,M}$ is a random vector of dimension F. Let $\theta_{n,M,i}$ be the i-th component where $i \in \{1, \ldots, F\}$. We have

$$||\theta_{n,M} - \theta_n^*||_2 = \sqrt{\sum_{i=1}^{F}(\theta_{n,M,i} - \theta_{n,i}^*)^2} \leq \sum_{i=1}^{F}|\theta_{n,M,i} - \theta_{n,i}^*|.$$

Recall that $\theta_m = \left[\frac{1}{m+1}\sum_{i=0}^{m}\phi_i(\phi_i - \gamma\phi_{i+1})'\right]^{-1}\left[\frac{1}{m+1}\sum_{i=0}^{m}\phi_i C_i\right]$. Since the post-decision state space is compact, ϕ is continuous and C is bounded, $\phi_i(\phi_i - \gamma\phi_{i+1})$ and $\phi_i C_i$ are uniformly bounded for all i. As a result, both $\frac{1}{m+1}\sum_{i=0}^{m}\phi_i(\phi_i - \gamma\phi_{i+1})$ and $\frac{1}{m+1}\sum_{i=0}^{m}\phi_i C_i$ are uniformly bounded for all m. Since matrix inversion is continuous, θ_m is uniformly bounded for all m. Hence, for each fixed i and n, $(\theta_{n,M,i})$ is a sequence of uniformly bounded random variables. As a result, $(\theta_{n,M,i})$ is uniformly integrable. By assumption and theorem 4.1, $\theta_{n,M,i} \to \theta_{n,i}^*$ $\mu_{\hat{\pi}_n}$-almost surely as $M \to \infty$. Then, we obtain that $\theta_{n,M,i} \to \theta_{n,i}^*$ in L_1 for all $i \in \{1, \ldots, F\}$. Hence, for any $\epsilon_n > 0$, there exists $M_n \in \mathbb{N}$ such that

$$\sum_{i=1}^{F}\mathbb{E}_{\hat{\pi}_n}|\theta_{n,M_n,i} - \theta_{n,i}^*| \leq \frac{\epsilon_n}{\gamma c}.$$

Then,

$$\mathbb{E}_{\hat{\pi}_n}||\theta_{n,M_n} - \theta_n^*||_2 \leq \sum_{i=1}^{F}\mathbb{E}_{\hat{\pi}_n}|\theta_{n,M_n,i} - \theta_{n,i}^*| \leq \frac{\epsilon_n}{\gamma c}.$$

Therefore, in the inner loop of each iteration we can uniformly bound the mean difference between the approximate value function and the true function of the approximate policy:

$$\mathbb{E}_{\hat{\pi}_n}||\hat{V}^{\hat{\pi}_n} - V^{\hat{\pi}_n}||_\infty \leq \gamma c\mathbb{E}_{\hat{\pi}_n}||\theta_{n,M_n} - \theta_n^*||_2 \leq \epsilon_n.$$

It is worth noting that the subscript n of M implies that it is not fixed but depends on the outer loop iteration counter n, since it is an on-policy algorithm and the chain changes as the policy gets updated.

Hence, we conclude that theorem 3 applies to the LS/RLSAPI algorithm. ∎

Remark: Since the parameter estimates using instrumental variable are unbiased, we have for each i

$$\begin{aligned}\mathbb{E}_{\hat{\pi}_n}|\theta_{n,M_n,i} - \theta^*_{n,i}| &\leq \mathbb{E}_{\hat{\pi}_n}|\theta_{n,M_n,i} - \mathbb{E}_{\hat{\pi}_n}\theta_{n,M_n,i}| + |\mathbb{E}_{\hat{\pi}_n}\theta_{n,M_n,i} - \theta^*_{n,i}| \\ &\leq \sqrt{Var_{\mu^{\hat{\pi}_n}}(\theta_{n,M_n,i})}.\end{aligned}$$

So the mean absolute deviation of a parameter from the true value is bounded by the standard deviation of the parameter estimate. In addition, it is worth pointing out that the variance term approach zero asymptotically and it depends on the unknown true variance of samples. In an actual implementation of the algorithm, we use an estimated variance from samples (or standard error) instead to specify a stopping criterion for the inner loop.

7 Extension with Chebyshev polynomial approximation

To apply the convergence results in section 6, we make the strong assumption that the value function of all policies in the policy space are spanned by a finite set of known basis functions. By assuming that the value function of all policies are in smooth function spaces, we can extend the convergence results to value functions with unknown form. For simplicity of presentation, we restrict ourselves to a 1-dimensional state space, a closed interval $[a, b]$, and only consider the function space $C^k[a, b]$, the set of all continuous functions with up to k-th derivative. The convergence result can be generalized to high-dimensional state spaces.

7.1 Orthogonal polynomials

We first introduce the idea of orthogonal polynomials by defining the inner product with respect to a weighting function w in $C^k[a,b]$ to be

$$\langle f, g\rangle_w = \int_a^b f(x)g(x)w(x)dx.$$

This inner product defines a quadratic semi-norm $||f||_w^2 = \langle f, f\rangle_w$. Let $G^w = \{g_n^w\}_{n=1}^\infty$ be a set of orthogonal basis functions with respect to w in $C^k[a,b]$ and $G_N^w = \{g_n^w\}_{n=1}^N$ be the finite subset of order N. Let $\mathcal{G}^w$ and $\mathcal{G}_N^w$ denote the function spaces spanned by G and G_N respectively. Given any f, the best least-square approximation of f with respect to w onto G_N^w is the solution to the following optimization problem:

$$\inf_{g\in\mathcal{G}_N^w} \int_a^b (f(x) - g(x))^2 w(x)dx,$$

and the solution is $f_N^w = \sum_{n=1}^N \frac{\langle f, g_n^w\rangle_w}{||g_n^w||_w^2} g_n^w$.

7.2 Chebyshev polynomial approximation

We focus on one specific weighting function: the Chebyshev weighting function $c(x) = (1 - (\frac{2x-a-b}{b-a})^2)^{-\frac{1}{2}}$ on $[a,b]$. For example, if we take the interval $[a,b]$ to be $[-1,1]$, the family of Chebyshev polynomials $T = \{\tilde{t}_n\}_{n=0}^\infty$ is defined as $\tilde{t}_n(x) = \cos(n\cos^{-1}x)$ (Judd (1998)). It can also be recursively defined as

$$\begin{aligned}
\tilde{t}_0(x) &= 1, \\
\tilde{t}_1(x) &= x, \\
\tilde{t}_{n+1}(x) &= 2x\tilde{t}_n(x) - \tilde{t}_{n-1}(x), n \geq 1.
\end{aligned}$$

We normalize them by letting $t_0 = \frac{\tilde{t}_0}{\pi}$ and $t_n = \frac{2\tilde{t}_n}{\pi}$ for all $n \geq 1$. Chebyshev polynomial approximators are good for smooth functions because they have the desirable uniform convergence property Judd (1998).

Let μ^π be the invariant measure of the Markov chain of following a fixed policy π. Suppose $V^\pi \in C^k[a,b]$ for all $\pi \in \Pi$. Assume that the invariant probability measure μ^π has a continuous

density function f^π i.e. $\mu^\pi(dx) = f^\pi(x)dx$, and f^π is strictly positive on $[a,b]$ and has up to k-th order derivative. Let $\tilde{V}^\pi = V^\pi\sqrt{\frac{f^\pi}{c}}$ and $\mathcal{C}_N$ denote the function space spanned by the finite orthonormal Chebyshev basis set $T_N = \{t_n\}_{n=0}^N$ on $[a,b]$. We consider the following Chebyshev least square approximation problem,

$$\inf_{\tilde{g}\in\mathcal{C}_N}(\tilde{V}(x) - \tilde{g}(x))^2 c(x)dx.$$

The solution to this problem is the N-th degree Chebyshev least squares approximation

$$\tilde{g}^*(x) = C_N(x) = \sum_{j=0}^N c_j t_j(x)$$

where $c_j = \int_a^b \tilde{V}^\pi(x)t_j(x)c(x)dx$.

Let $T^\pi = \{t_n\sqrt{\frac{c}{f^\pi}}\}_{n=0}^\infty$. It is easy to see that T^π is an orthonormal basis set with respect to f^π in $C^k[a,b]$. Finding the best least squares approximation for the value function of policy π on the basis set $\mathcal{T}_N^\pi$ is

$$\inf_{g\in\mathcal{T}_N^\pi}\int_a^b (V^\pi(x) - g(x))^2 f^\pi(x)dx.$$

It can be verified that the solution is

$$g^*(x) = V_N^\pi(x) = C_N(x)\sqrt{\frac{c(x)}{f^\pi(x)}}.$$

Let $\phi_j(x) = t_j(x)\sqrt{\frac{c(x)}{f^\pi(x)}}$. We have

$$V^\pi(x) = V_N^\pi(x) + r(x) = \sum_{j=0}^N c_j\phi_j(x) + r(x),$$

where $r(x)$ is the residual function orthogonal to $\phi_j(x)$ for all $0 \le j \le N$. Then, Bellman's equation (13) gives us

$$\phi(x)^T\theta^* + r(x) = \int_{\mathcal{X}^\pi} P^\pi(x,dx')[C^\pi(x,x') + \gamma(\phi(x')^T\theta^* + r(x'))],$$

and the linear model becomes

$$C^\pi(x,x') = \left(\phi(x) - \gamma\int_{\mathcal{X}^\pi} P^\pi(x,dx')\phi(x')\right)^T\theta^* + \left(r(x) - \gamma\int_{\mathcal{X}^\pi} P^\pi(x,dx')r(x')\right)$$

$$+\left(C^{\pi}(x,x') - \int_{\mathcal{X}^{\pi}} P^{\pi}(x,dx')C^{\pi}(x,x')\right).$$

If we only regress on $\phi(x) - \gamma \int_{\mathcal{X}^{\pi}} P(x,dx')\phi(x')$, $r(x) - \gamma \int_{\mathcal{X}^{\pi}} P(x,dx')r(x')$ enters the error term, which is not necessarily mean zero and uncorrelated with the input variables. As a result, the parameter estimate of the errors-in-variable model is not asymptotically unbiased anymore. With the same approach as in theorem 1, we can show that the m-th parameter estimate θ_m as in equation (14) converges to $\theta^* + \mathbf{b}$ almost surely where

$$\begin{aligned}
\mathbf{b} &= \left[\int_{\mathcal{X}^{\pi}} \mu^{\pi}(dx)\phi(x)\left(\phi(x) - \gamma\int_{\mathcal{X}^{\pi}} P^{\pi}(x,dx')\phi(x')\right)^T\right]^{-1}\left[\int_{\mathcal{X}^{\pi}} \mu^{\pi}(dx)\phi(x)(r(x) - \gamma\int_{\mathcal{X}^{\pi}} P^{\pi}(x,dx')r(x'))\right] \\
&= \left[\int_{\mathcal{X}^{\pi}} \mu^{\pi}(dx)\phi(x)\left(\phi(x) - \gamma\int_{\mathcal{X}^{\pi}} P^{\pi}(x,dx')\phi(s')\right)^T\right]^{-1}\left[\int_{\mathcal{X}^{\pi}} \mu^{\pi}(dx)\phi(x)\cdot(-\gamma\int_{\mathcal{X}^{\pi}} P^{\pi}(x,dx')r(x'))\right].
\end{aligned}$$

We are now ready to extend the mean convergence result developed in theorem 3 to find an error bound for Chebyshev polynomial approximation. Before presenting the main result, we introduce the following lemma.

Lemma 7.1 *Let C be a $F \times F$ perturbed identity matrix i.e. $C = I - \lambda A$ where A is a matrix with entry $A_{ij} \in [-1,1]$. If $\lambda \leq \frac{1}{2F}$, then*

$$||C^{-1}||_{\infty} = \max_{1\leq i\leq F} \sum_{j=1}^{F} |C_{ij}^{-1}| \leq 2$$

where $||\cdot||_{\infty}$ denotes the maximum absolute row sum norm for matrix.

Proof:

By lemma 4.3, we know that C is invertible given $\lambda \leq \frac{1}{2F}$. It is easy to check that

$$||\lambda A||_{\infty} \leq \lambda F \leq \frac{1}{2}.$$

Since the maximum absolute row sum norm is sub-multiplicative, we have

$$\begin{aligned}
||C^{-1}||_\infty &= ||(I-\lambda A)^{-1}||_\infty \\
&= ||I+\lambda A+\lambda^2 A^2+\lambda^3 A^3+\cdots||_\infty \\
&\leq ||I||_\infty + ||\lambda A||_\infty + ||\lambda^2 A^2||_\infty + ||\lambda^3 A^3||_\infty + \cdots \\
&\leq ||I||_\infty + ||\lambda A||_\infty + ||\lambda A||_\infty^2 + ||\lambda A||_\infty^3 + \cdots \\
&\leq 1+\frac{1}{2}+\frac{1}{4}+\cdots \\
&= 2.
\end{aligned}$$

∎

Theorem 4 (Mean error bound of LS/RLSAPI with exact Chebyshev polynomials) *Suppose assumption 3.1, 3.1 and 4 hold and further assume that, for any policy $\pi \in \Pi$, the value function V^π is in $C^k[a,b]$ and the invariant density function f^π is known and bounded away from 0 on $\mathcal{X}^\pi$. Given a desired approximation error tolerance ϵ, for each n and $V^{\hat{\pi}_n}$ we can construct an approximate policy value function $V_{F_n}^{\hat{\pi}_n} = C_{F_n}\sqrt{\frac{c}{f^{\hat{\pi}_n}}}$ with a finite set of Chebyshev basis functions of order F_n+1 such that $\forall n \in \mathbb{N}$*

$$\mathbb{E}_{\hat{\pi}_n}||\hat{V}_{F_n}^{\hat{\pi}_n} - V^{\hat{\pi}_n}||_\infty \leq \epsilon, \tag{21}$$

where $\hat{V}_{F_n}^{\hat{\pi}_n}$ is the statistical estimate of $V_{F_n}^{\hat{\pi}_n}$ using Chebyshev basis functions.

Furthermore, let δ be the positive scalar that uniformly bound the mean errors in policies (over all iterations), that is ,

$$\mathbb{E}_{\hat{\pi}_n}||M_{\hat{\pi}_{n+1}}\hat{V}_{F_n}^{\hat{\pi}_n} - M\hat{V}_{F_n}^{\hat{\pi}_n}||_\infty \leq \delta. \tag{22}$$

Then, LS/RLSAPI with exact Chebyshev polynomial generates a sequence of policies whose mean performance bound (away from the optimal) satisfies:

$$\limsup_{n\to\infty} \mathbb{E}_{\hat{\pi}_n}||\hat{V}_{F_n}^{\hat{\pi}_n} - V^*||_\infty \leq \frac{\delta+(1+\gamma^2)\epsilon}{(1-\gamma)^2}.$$

Proof:

To prove the first part of the theorem, we consider a general policy π (dropping subscript n) and the case of $k = 1$ (that is, V^π is once differentiable) for simplicity of presentation. Let $\tilde{V}^\pi = V^\pi\sqrt{\frac{f^\pi}{c}}$. By uniform convergence of Chebyshev approximator, there exists $K_1 > 0$ such that

$$||\tilde{V}^\pi - C_F||_\infty \leq K_1\frac{\log F}{F}.$$

Let $K_2 = ||\sqrt{\frac{c}{f^\pi}}||_\infty$. Then,

$$||r||_\infty = ||V^\pi - V_F^\pi||_\infty = ||V^\pi - C_F\sqrt{\frac{c}{f^\pi}}||_\infty \leq K_1K_2\frac{\log F}{F}.$$

Then, $V_F^\pi = C_F\sqrt{\frac{c}{f^\pi}}$ is the approximate policy value function for V^π.

To apply lemma 7.1, we collapse transitions as in section 4 and use parameter estimates as in equation (15). There exists k such that $\gamma^k \leq \frac{1}{2(F+1)}$. We rewrite the asymptotic bias of the parameter estimates in shorthand notation as

$$\mathbf{b} = \left[\mu^\pi\phi(\phi - \gamma^k(P^\pi)^k\phi)^T\right]^{-1}\left[\mu^\pi\phi(-\gamma^k(P^\pi)^k r)\right]. \tag{23}$$

Since we have orthonormal basis functions, the correlation matrix is a non-singular perturbed identity matrix by lemma 4.3 and $\mu^\pi\phi$ is a vector with entries in $[-1, 1]$. For each $i \in 1, \cdots, F+1$, by lemma 7.1 we obtain

$$|b_i| \leq \gamma^k||\left[\mu^\pi\phi(\phi - \gamma^k(P^\pi)^k\phi)^T\right]^{-1}||_\infty||r||_\infty \leq 2\gamma^k||r||_\infty \leq K_1K_2\frac{\log F}{F(F+1)}.$$

As in the proof of corollary 6.1), it is easy to see that, for each i, $\theta_{m,i} \to \theta_i^* + b_i$ in L_1 with respect to μ^π. For each i, there exists $M_i \in \mathbb{N}$ such that

$$\mathbb{E}_\pi|\theta_{M,i} - \theta_i^*| \leq \mathbb{E}_\pi|\theta_{M,i} - \theta_i^* - b_i| + |b_i| \leq 2|b_i|.$$

Hence, there exists $M \in \mathbb{N}$ such that

$$\mathbb{E}_\pi||\theta_M - \theta^*||_2 \leq \sum_{i=1}^{F+1}\mathbb{E}_\pi|\theta_{M,i} - \theta_i^*| \leq 2K_1K_2\frac{\log F}{F}.$$

Let $K_3 = \max_{x^a} ||\phi(x^a)||_2$. Hence, we have

$$\mathbb{E}_\pi ||\hat{V}_F^\pi - V_F^\pi||_\infty \leq \gamma K_3 \mathbb{E}_\pi ||\hat{\theta}_M - \theta^*||_2 \leq 2\gamma K_1 K_2 K_3 \frac{\log F}{F}.$$

Finally, we have

$$\begin{aligned} \mathbb{E}_\pi ||\hat{V}_F^\pi - V^\pi||_\infty &\leq \mathbb{E}_\pi ||\hat{V}_F^\pi - V_F^\pi||_\infty + ||V_F^\pi - V^\pi||_\infty \\ &\leq (2\gamma K_3 + 1) K_1 K_2 \frac{\log F}{F}. \end{aligned}$$

For each $\epsilon > 0$, there exists finite $F \in \mathbb{N}$ such that

$$(2\gamma K_3 + 1) K_1 K_2 \frac{\log F}{F} \leq \epsilon.$$

Hence, for each policy evaluation step n we can construct a finite set of basis functions of order $F_n + 1$ given a desired approximation error tolerance ϵ in the inner loop of the LS/RLSAPI algorithm.

In the second part of the theorem, we have uniform bounds on the mean errors in approximations to value functions and policies (over all iterations) respectively, so the proof is similar to the proof of the error bound for approximate policy iteration in Bertsekas & Tsitsiklis (1996) and theorem 3. The details are omitted for brevity. ∎

One major limitation of the algorithm with exact Chebyshev basis functions is that we need the exact invariant density function f^π to construct a finite set of basis functions. To address this difficulty, we call a procedure that produces approximations of f^π in the implementation of the algorithm. Since we have sequential observations of states from the Markov chain, we can obtain estimates of the invariant density with increasing accuracy and construct approximate basis functions from the estimated density function. By relaxing the assumption of known invariant density function to a known lower bound, the following lemma and theorem provide theoretical support to the asymptotic property of the algorithm using approximate Chebyshev basis functions.

Lemma 7.2 *Suppose Φ is a positive Harris chain with invariant probability measure μ. For $f, f_i \in L_1(\mathcal{X}, \mathcal{B}(\mathcal{X}), \mu)$, assume $(f_i), f$ are uniformly bounded and $f_i \to f$ uniformly. Then*

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} f_i(x_i) = \int_{\mathcal{X}} \mu(dx) f(x)$$

μ-almost surely.

Proof:

Let $\epsilon > 0$. Since $f_i \to f$ uniformly, there exists $N \in \mathbb{N}$ such that for all $x \in \mathcal{X}$ and $i > N$, $|f_i(x) - f(x)| < \frac{\epsilon}{4}$. Since f_i and f are bounded, there exists $B > 0$ such that $|f(x)| \leq B$ and $|f_i(x)| \leq B$ for all $x \in \mathcal{X}$ and $i \in \mathbb{N}$. Hence, there exists M_1 ($> N$) large enough such that for any finite sequence $(x_i)_{i=1}^{N}$ in $\mathcal{X}$, $\frac{1}{M_1} \sum_{i=1}^{N} |f_i(x_i) - f(x_i)| \leq \frac{2BN}{M_1} < \frac{\epsilon}{4}$. Then,

$$\begin{aligned}
\left| \frac{1}{M_1} \sum_{i=1}^{M_1} [f_i(x_i) - f(x_i)] \right| &\leq \frac{1}{M_1} \sum_{i=1}^{M_1} |f_i(x_i) - f(x_i)| \\
&= \frac{1}{M_1} \sum_{i=1}^{N} |f_i(x_i) - f(x_i)| + \frac{1}{M_1} \sum_{i=N+1}^{M_1} |f_i(x_i) - f(x_i)| \\
&< \frac{\epsilon}{4} + \frac{M_1 - N}{M_1} \cdot \frac{\epsilon}{4} \\
&< \frac{\epsilon}{2}.
\end{aligned}$$

By lemma 4.1, there exists $M_2 \in \mathbb{N}$ such that for all $n > M_2$,

$$\left| \frac{1}{n} \sum_{i=1}^{n} f(x_i) - \int_{\mathcal{X}} \mu(dx) f(x) \right| < \frac{\epsilon}{2}.$$

Let $M = \max\{M_1, M_2\}$. We have

$$\begin{aligned}
\left| \frac{1}{M} \sum_{i=1}^{M} f_i(x_i) - \int_{\mathcal{X}} \mu(dx) f(x) \right| &\leq \left| \frac{1}{M} \sum_{i=1}^{M} f_i(x_i) - \frac{1}{M} \sum_{i=1}^{M} f(x_i) \right| + \left| \frac{1}{M} \sum_{i=1}^{M} f(x_i) - \int_{\mathcal{X}} \mu(dx) f(x) \right| \\
&< \epsilon.
\end{aligned}$$

Hence, we conclude that $\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} f_i(x_i) = \int_{\mathcal{X}} \mu(dx) f(x)$ μ-almost surely. ∎

Theorem 5 (Mean error bound of LS/RLSAPI with approximate Chebyshev polynomials) *Under assumption 3.1, 3.1 and 4, suppose the value function V^π is in $C^k[a,b]$ for any policy $\pi \in \Pi$ and there is a procedure producing a sequence of functions (f_i^π) such that f_i^π converges to the invariant density f^π uniformly. Further assume f_i^π and f^π are uniformly bounded with known bounds. We construct the set of approximate basis functions at each iteration i in the inner loop by letting $T_F^{\pi,i} = \{t_n\sqrt{\frac{c}{f_i^\pi}}\}_{n=0}^F$. Then, the error bound developed in theorem 4 holds for the LS/RLS approximate policy iteration algorithm with approximate Chebyshev polynomials.*

Proof:

Let $\epsilon > 0$. First, we use the same approach as in the proof of theorem 4 to determine the order of the basis function set F for $\frac{\epsilon}{2}$. Let the vector of approximate basis functions at iteration i in the inner loop be $\phi^{(i)}(x) = \left[t_0(x)\sqrt{\frac{c(x)}{f_i^\pi(x)}}, \cdots, t_F(x)\sqrt{\frac{c(x)}{f_i^\pi(x)}}\right]^T$. Again, we collapse transitions to find $k \in \mathbb{N}$ such that $\gamma^k \leq \frac{1}{2(F+1)}$. Then, the parameter estimate becomes

$$\theta_m^{(m)} = \left[\frac{1}{m+1}\sum_{i=0}^{m}\phi_i^{(i)}(\phi_i^{(i)} - \gamma^k\phi_{i+k}^{(i)})^T\right]^{-1}\left[\frac{1}{m+1}\sum_{i=0}^{m}\phi_i^{(i)}\sum_{j=0}^{k-1}\gamma^j C_{i+j}\right].$$

Hence, we obtain the m-th approximation for V^π: $\hat{V}_{F,(m)}^\pi = (\phi^{(m)})^T\theta_m^{(m)}$. It is worth noting that it is different from $\hat{V}_F^\pi = \phi^T\theta_m$ in theorem 4 due to the approximation of basis functions.

By lemma 7.2, $\theta_m^{(m)} \to \theta^* + \mathbf{b}$ μ^π-almost surely where b is defined as in (23). Since $\theta_m \to \theta^* + \mathbf{b}$ μ^π-almost surely, there exists $M_1 \in \mathbb{N}$ such that for all $m \geq M_1$,

$$\mathbb{E}_\pi||\phi^T\theta_m^{(m)} - \phi^T\theta_m||_\infty \leq \frac{\epsilon}{4}.$$

Since $\phi^{(m)}$ converges to ϕ uniformly componentwise and $\theta_m^{(m)}$ is uniformly bounded for all m, there exists $M_2 \in \mathbb{N}$ such that for all $m \geq M_2$,

$$\mathbb{E}_\pi||(\phi^{(m)})^T\theta_m^{(m)} - \phi^T\theta_m^{(m)}||_\infty \leq \frac{\epsilon}{4}.$$

Then, for all $m \geq \max\{M_1, M_2\}$ we have

$$\mathbb{E}_\pi||\hat{V}_{F,(m)}^\pi - \hat{V}_F^\pi||_\infty \leq \mathbb{E}_\pi||(\phi^{(m)})^T\theta_m^{(m)} - \phi^T\theta_m^{(m)}||_\infty + \mathbb{E}_\pi||\phi^T\theta_m^{(m)} - \phi^T\theta_m||_\infty \leq \frac{\epsilon}{2}.$$

As shown in theorem 4, there exists $M_3 \in \mathbb{N}$ such that for all $m \geq M_3$,

$$\mathbb{E}_\pi ||\hat{V}_F^\pi - V^\pi||_\infty = \mathbb{E}_\pi ||\phi^T \theta_m - V^\pi||_\infty \leq \frac{\epsilon}{2}.$$

Let $M = \max\{M_1, M_2, M_3\}$. Then,

$$\mathbb{E}_\pi ||\hat{V}_{F,(M)}^\pi - V^\pi||_\infty \leq \mathbb{E}_\pi ||\hat{V}_{F,(M)}^\pi - \hat{V}_F^\pi||_\infty + \mathbb{E}_\pi ||\hat{V}_F^\pi - V^\pi||_\infty \leq \epsilon.$$

The conclusion follows since we can also uniformly bound the mean errors in approximations of value functions for arbitrary $\epsilon > 0$ in the algorithm with approximate Chebyshev polynomials. ∎

One of the key assumptions in theorem 5 is that the density estimates converge uniformly to the true invariant density of the chain. There are a variety of common methods for estimating density functions from a finite data set, including histogram, frequency polygon, kernel, nearest neighbor, orthogonal series, wavelet, spline, and likelihood based procedures (see Scott (1992)). Since the kernel estimates of density function converges uniformly under the assumption that the underlying Markov chain has differentiable transition density and satisfies the strong Doeblin condition (see Kristensen (2008)), in the algorithm we select the kernel approach to achieve the invariant density estimates by letting

$$\hat{f}_n(x) = \frac{1}{nh} \sum_{i=1}^{n} K(\frac{x_i - x}{h}),$$

where K is some kernel function with bandwidth h satisfying some regularity conditions.

Finally, the details of the algorithm using approximate Chebyshev basis functions are summarized in figure 3. The algorithm is a modified version of the LS/RLSAPI in figure 2, and the only technical complications reside in the step of policy evaluation where 1) the algorithm determines the order of the Chebyshev basis set F (based on the criterion discussed in the proof of theorem 4) and the corresponding number of collapsing transitions $k \leq \frac{1}{2F}$, 2) it produces initial estimates of invariant density and corresponding approximate Chebyshev basis functions based on the first k sampled state transitions in the first iteration of policy evaluation (step 2.1) and update the estimates in subsequent iterations with new samples (step 2.2).

Step 0: Initialization:

Step 0.2 Set the initial policy π_0.

Step 0.2: Set the iteration counter $n = 0$.

Step 1: Do for $n = 0, \ldots, N$,

Step 1.1: Set the initial State x_0^n.

Step 1.2: Determine F_n (order of the basis function set) and k_n (number of transition collapse).

Step 2: Do for $m = 0, \cdots, M$,

Step 2.1: If $m = 0$, do the following:

Step 2.1.1: Initialize $\hat{\theta}_{n,0} = 0$ and $\hat{v}_0 = 0$.

Step 2.1.2: Draw randomly or observe $W_1, \cdots, W_{k_n+1}$ from the stochastic process.

Step 2.1.3: Do for $j = 0, \cdots, k_n$:

Step 2.1.3.a: Set $u_j^n = \pi_n(x_j^n)$,

Step 2.1.3.b: Compute $x_j^{n,\pi} = S^{M,\pi}(x_j^n, u_j^n)$ and $x_{j+1}^n = S^M(x_j^{n,\pi}, u_j^n, W_{j+1})$.

Step 2.1.4: Produce the initial kernel density estimate $f_0^{n,\pi}$ from $x_0^{n,\pi}, \cdots, x_{k_n}^{n,\pi}$.

Step 2.1.5: Construct the initial approximate Chebyshev basis functions $\phi^{(0)}$ using $f_0^{n,\pi}$

Step 2.2: If $m = 1, \ldots, M$, do the following:

Step 2.2.1: Draw randomly or observe W_{m+k_n+1} from the process.

Step 2.2.2 Set $u_{m+k_n}^n = \pi_n(x_{m+k_n}^n)$.

Step 2.2.3 Compute $x_{m+k_n}^{n,\pi} = S^{M,\pi}(x_{m+k_n}^n, u_{m+k_n}^n)$ and
$x_{m+k_n+1}^n = S^M(x_{m+k_n}^n, u_{m+k_n}^n, W_{m+k_n+1})$.

Step 2.2.4 Compute $u_{m+k_n+1}^n = \pi_n(x_{m+k_n+1}^n)$ and
$x_{m+k_n+1}^{n,\pi} = S^{M,\pi}(x_{m+k_n+1}^n, u_{m+k_n+1}^n)$.

Step 2.2.5 Update density estimate $f_m^{n,\pi}$ from $f_{m-1}^{n,\pi}$ and $x_{k_n+m}^{n,\pi}$.

Step 2.2.6 Update approximate Chebyshev basis functions $\phi^{(m)}$ using $f_m^{n,\pi}$

Step 2.3: Compute regressor $\phi^{(m)}(x_m^{n,\pi}) - \gamma^{k_n}\phi^{(m)}(x_{m+k_n}^{n,\pi})$.

Step 2.4: Compute response variable $\hat{v}_m = \sum_{j=0}^{k_n-1} \gamma^j C(x_{m+j}^{n,\pi}, x_{m+j+1}^{n,\pi})$.

Step 2.5: Update parameters $\hat{\theta}_{n,m}$ with LS/RLS method

Step 3: Update the parameter and the policy:

$$\hat{\theta}_n = \hat{\theta}_{n,M},$$

$$\pi_{n+1}(x) = \arg\max_{u \in \Gamma(x)} \{C(x,u) + \gamma \phi^{(M)}(x^u)^T \hat{\theta}_n\}.$$

Step 4: Return the policy π_{N+1} and parameters $\hat{\theta}_N$.

Figure 3: Infinite-horizon approximate policy iteration algorithm with approximate Chebyshev basis functions

8 Conclusion

In this paper we propose an online, on-policy least squares approximate policy iteration algorithm with linear approximation for infinite-horizon Markov decision process problems with continuous state and action spaces. Under the assumptions that the stochastic system evolves according to a positive Harris chain for any deterministic stationary policy and the true post-decision value functions of policies are spanned by a finite set of known basis functions, we have shown that the algorithm is convergent in the mean, meaning that the mean error between the approximate policy value function and the optimal value function goes to 0 as successive approximations become more accurate. Furthermore, the convergence analysis is extended to the case when the true value functions are in some smooth function space so that we can construct a finite set of orthonormal basis functions from Chebyshev polynomials. Our next goal would be searching for provably convergent online, on-policy algorithms using non-linear function approximations (parametric or non-parametric) suitable for MDP problems with value functions of unknown form. Other advanced and sophisticated value function updating rules and approximation techniques will be considered, including neural networks, kernel smoothing and local polynomial regression (see Fan & Gijbels (1996)).

Acknowledgement

The first author would like to thank Professor Erhan Cinlar, Ning Hao, Yang Feng, Ke Wan, Jingjin Zhang, and Ping Yu for many inspiring discussions. Thanks to Peter Frazier for the proofreading and helpful comments. This work was supported by Castle Lab at Princeton University and partially supported by AFOSR grant FA9550-06-1-0496.

References

Antos, A., Munos, R. & Szepesvari, C. (2008), 'Fitted Q-iteration in continuous action-space MDPs', *Advances in neural information processing systems* **20**, 9–16.

Antos, A., Szepesvári, C. & Munos, R. (2007), Value-Iteration Based Fitted Policy Iteration:

Learning with a Single Trajectory, *in* 'Approximate Dynamic Programming and Reinforcement Learning, 2007. ADPRL 2007. IEEE International Symposium on', pp. 330–337.

Antos, A., Szepesvári, C. & Munos, R. (2008), 'Learning near-optimal policies with Bellman-residual minimization based fitted policy iteration and a single sample path', *Machine Learning* **71**(1), 89–129.

Baird, L. (1995), 'Residual algorithms: Reinforcement learning with function approximation', *Proceedings of the Twelfth International Conference on Machine Learning* pp. 30–37.

Bellman, R. & Dreyfus, S. (1959), 'Functional Approximations and Dynamic Programming', *Mathematical Tables and Other Aids to Computation* **13**(68), 247–251.

Bellman, R., Kalaba, R. & Kotkin, B. (1963), 'Polynomial approximation— a new computational technique in dynamic programming: Allocation processes', *Mathematics of Computation* **17**(82), 155–161.

Bertsekas, D. & Shreve, S. (1978), *Stochastic Optimal Control: The Discrete-Time Case*, Academic Press, Inc. Orlando, FL, USA.

Bertsekas, D. & Tsitsiklis, J. (1996), *Neuro-Dynamic Programming*, Athena Scientific Belmont, MA.

Boyan, J. (1999), Least-squares temporal difference learning, *in* 'Proceedings of the Sixteenth International Conference on Machine Learning', pp. 49–56.

Bradtke, S. (1993), 'Reinforcement Learning Applied to Linear Quadratic Regulation', *Advances In Neural Information Processing Systems* pp. 295–302.

Bradtke, S. & Barto, A. (1996), 'Linear Least-Squares algorithms for temporal difference learning', *Machine Learning* **22**(1), 33–57.

Bradtke, S., Ydstie, B. & Barto, A. (1994), 'Adaptive linear quadratic control using policy iteration', *American Control Conference, 1994* **3**, 3475–3479.

Doob, J. (1953), *Stochastic Processes*, John Wiley & Sons, New York.

Fan, J. & Gijbels, I. (1996), *Local Polynomial Modelling and Its Applications*, Chapman & Hall/CRC, London.

Gordon, G. (1995), 'Stable function approximation in dynamic programming', *Proceedings of the Twelfth International Conference on Machine Learning* pp. 261–268.

Gordon, G. (2001), 'Reinforcement learning with function approximation converges to a region', *Advances in Neural Information Processing Systems* **13**, 1040–1046.

Judd, K. (1998), *Numerical Methods in Economics*, MIT Press Cambridge, MA.

Kristensen, D. (2008), 'Uniform Convergence Rates of Kernel Estimators with Heterogeneous, Dependent Data', *CREATES Research Paper 2008-37* .

Lagoudakis, M. & Parr, R. (2003), 'Least-Squares Policy Iteration', *Journal of Machine Learning Research* **4**(6), 1107–1149.

Landelius, T. & Knutsson, H. (1997), 'Greedy adaptive critics for LQR problems: Convergence proofs', *Neural Computation* .

Luque, J. (1987), A nonlinear proximal point algorithm, *in* '26th IEEE Conference on Decision and Control', Vol. 26, pp. 816–817.

Maei, H., Szepesvári, C., Bhatnagar, S., Silver, D., Precup, D. & Sutton, R. (2010), 'Convergent Temporal-Difference Learning with Arbitrary Smooth Function Approximation'.

Mahadevan, S. & Maggioni, M. (2007), 'Proto-value Functions: A Laplacian Framework for Learning Representation and Control in Markov Decision Processes', *The Journal of Machine Learning Research* **8**, 2169–2231.

Melo, F., Lisboa, P. & Ribeiro, M. (2007), 'Convergence of Q-learning with linear function approximation', *Proceedings of the European Control Conference 2007* pp. 2671–2678.

Melo, F., Meyn, S. & Ribeiro, M. (2008), An analysis of reinforcement learning with function approximation, *in* 'Proceedings of the 25th international conference on Machine learning', ACM, pp. 664–671.

Meyn, S. & Tweedie, R. (1993), *Markov chains and stochastic stability*, Springer, New York.

Munos, R. & Szepesvári, C. (2008), 'Finite-time bounds for fitted value iteration', *The Journal of Machine Learning Research* **9**, 815–857.

Nummelin, E. (1984), *General Irreducible Markov Chains and Non-Negative Operators*, Cambridge University Press, Cambridge.

Ormoneit, D. & Sen, Ś. (2002), 'Kernel-Based Reinforcement Learning', *Machine Learning* **49**(2), 161–178.

Papavassiliou, V. & Russell, S. (1999), 'Convergence of Reinforcement Learning with General Function Approximators', *Proceedings of the Sixteenth International Joint Conference on Artificial Intelligence* pp. 748–757.

Powell, W. B. (2007), *Approximate Dynamic Programming: Solving the curses of dimensionality*, John Wiley and Sons, New York.

Precup, D., Sutton, R. & Dasgupta, S. (2001), Off-policy temporal-difference learning with function approximation, *in* 'Proceedings of ICML', pp. 417–424.

Puterman, M. L. (1994), *Markov Decision Processes*, John Wiley & Sons, New York.

Reetz, D. (1977), 'Approximate Solutions of a Discounted Markovian Decision Process', *Bonner Mathematische Schriften* **98**, 77–92.

Schweitzer, P. & Seidmann, A. (1985), 'Generalized polynomial approximations in Markovian decision processes', *Journal of mathematical analysis and applications* **110**(2), 568–582.

Scott, D. (1992), *Multivariate Density Estimation: Theory, Practice, and Visualization*, John Wiley and Sons, New York.

Stokey, N., Prescott, E. & Lucas, R. (1989), *Recursive Methods in Economic Dynamics*, Harvard University Press Cambridge, MA.

Sutton, R. & Barto, A. (1998), *Reinforcement Learning: An Introduction*, MIT Press Cambridge, MA.

Sutton, R., Szepesvári, C. & Maei, H. (2009), 'A convergent O (n) algorithm for off-policy temporal-difference learning with linear function approximation', *Advances in Neural Information Processing Systems* **21**.

Szita, I. (2007), *Rewarding Excursions: Extending Reinforcement Learning to Complex Domains*, Eotvos Lorand University, Budapest.

Tadić, V. (2001), 'On the convergence of temporal-difference learning with linear function approximation', *Machine learning* **42**(3), 241–267.

Tsitsiklis, J. & Van Roy, B. (1996), 'Feature-based methods for large scale dynamic programming', *Machine Learning* **22**(1), 59–94.

Tsitsiklis, J. & Van Roy, B. (1997), 'An analysis of temporal-difference learning with function approximation', *IEEE Transactions on Automatic Control* **42**(5), 674–690.

Van Roy, B., Bertsekas, D., Lee, Y. & Tsitsiklis, J. (1997), A Neuro-Dynamic Programming Approach to Retailer Inventory Management, *in* 'Proceedings of the 36th IEEE Conference on Decision and Control, 1997', Vol. 4.

Whitt, W. (1978), 'Approximations of Dynamic Programs, I', *Mathematics of Operations Research* **3**(3), 231–243.

Young, P. (1984), *Recursive estimation and time-series analysis: an introduction*, Springer-Verlag, New York.