June 2007 (Revised November 2007) **Report LIDS - 2754**

Solution of Large Systems of Equations Using Approximate Dynamic Programming Methods

Dimitri P. Bertsekas[1] and Huizhen Yu[2]

Abstract

We consider fixed point equations, and approximation of the solution by projection on a low-dimensional subspace. We propose stochastic iterative algorithms, based on simulation, which converge to the approximate solution and are suitable for large-dimensional problems. We focus primarily on general linear systems and propose extensions of recent approximate dynamic programming methods, based on the use of temporal differences, which solve a projected form of Bellman's equation by using simulation-based approximations to this equation, or by using a projected value iteration method.

Contents

[1] Dimitri Bertsekas is with the Dept. of Electr. Engineering and Comp. Science, M.I.T., Cambridge, Mass., 02139.

[2] Huizhen Yu is with the Helsinki Institute for Information Technology, Univ. of Helsinki, Finland. Her research was supported in part by the IST Programme of the European Community, under the PASCAL Network of Excellence, IST-2002-506778.

1. INTRODUCTION

In this paper we focus primarily on large systems of linear equations of the form

$$x = Ax + b, \tag{1.1}$$

where A is an $n \times n$ matrix and b is a column vector in the n-dimensional space $\Re^n$. Equivalently, we want to find a fixed point of the mapping T, given by

$$T(x) = Ax + b.$$

We propose methods to approximate the fixed point within a subspace spanned by a relatively small number of basis functions.

The motivation for our work comes from recent advances in the field of dynamic programming (DP), where large systems of equations of the form (1.1) appear in the context of evaluation of the cost of a stationary policy in a Markovian decision problem. In this DP context, we are given an n-state Markov chain with transition probability matrix P, which evolves for an infinite number of discrete time periods, and a cost vector $g \in \Re^n$, whose components g_i represent the costs of being at the corresponding states $i = 1, \ldots, n$, for a single time period. The problem is to evaluate the total cost vector

$$x = \sum_{t=0}^{\infty} \alpha^t P^t g,$$

where $\alpha \in (0, 1)$ is a discount factor, and the components x_i represent the total expected α-discounted cost over an infinite number of time periods, starting from the corresponding states $i = 1, \ldots, n$. It is well known that x is the unique solution of the equation,

$$x = \alpha P x + g,$$

and furthermore, x can also be computed iteratively by the Jacobi method $x_{t+1} = \alpha P x_t + g$ (also known as value iteration in the context of DP), since the mapping $x \mapsto \alpha P x + g$ is a contraction with respect to the sup norm; see textbooks on DP, such as for example Bertsekas [Ber07], or Puterman [Put94].

We focus on the case where n is very large, and it may be worth (even imperative) to consider a low-dimensional approximation of a solution within a subspace

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

where the columns of the $n \times s$ matrix Φ are basis functions that are linearly independent. This type of approximation approach has been the subject of much recent research in approximate DP, where several methods have been proposed and substantial computational experience has been accumulated. The most popular of these methods use projection with respect to the weighted Euclidean norm given by

$$\|x\|_\xi = \sqrt{\sum_{i=1}^{n} \xi_i x_i^2},$$

where $\xi \in \Re^n$ is a probability distribution with positive components. We denote by Π the projection operation onto S with respect to this norm (while Π depends on ξ, we do not show the dependence, since the associated vector ξ will always be clear from the context). The aforementioned methods for approximating the solution of the DP equation $x = \alpha Px + g$ aim to solve the equation

$$\Phi r = \Pi(\alpha P \Phi r + b)$$

with ξ being the invariant distribution of the transition probability matrix P (which is assumed irreducible; i.e., has a single recurrent class and no transient states). The more general methods for approximating a fixed point of $T(x) = Ax + b$, proposed in this paper, aim to solve the equation

$$\Phi r = \Pi T(\Phi r) = \Pi(A\Phi r + b), \tag{1.2}$$

where the projection norm $\|\cdot\|_\xi$ is determined in part by the structure of A in a way to induce some desired property. We view Π as a matrix and *we implicitly assume throughout that $I - \Pi A$ is invertible*. Thus, for a given ξ, Eq. (1.2) has a unique solution Φr^*, which together with the linear independence assumption on the columns of Φ, implies a unique solution r^*.

To evaluate the distance between Φr^* and a fixed point x^* of T, we write

$$x^* - \Phi r^* = x^* - \Pi x^* + \Pi x^* - \Phi r^* = x^* - \Pi x^* + \Pi T x^* - \Pi T \Phi r^* = x^* - \Pi x^* + \Pi A(x^* - \Phi r^*), \tag{1.3}$$

from which

$$x^* - \Phi r^* = (I - \Pi A)^{-1}(x^* - \Pi x^*).$$

Thus, we have for any norm $\|\cdot\|$ and fixed point x^* of T

$$\|x^* - \Phi r^*\| \leq \big\|(I - \Pi A)^{-1}\big\| \, \|x^* - \Pi x^*\|, \tag{1.4}$$

and the approximation error $\|x^* - \Phi r^*\|$ is proportional to the distance of the solution x^* from the approximation subspace. If ΠT is a contraction mapping of modulus $\alpha \in (0,1)$ with respect to $\|\cdot\|$, from Eq. (1.3), we have

$$\|x^* - \Phi r^*\| \leq \|x^* - \Pi x^*\| + \|\Pi T(x^*) - \Pi T(\Phi r^*)\| \leq \|x^* - \Pi x^*\| + \alpha \|x^* - \Phi r^*\|,$$

so that

$$\|x^* - \Phi r^*\| \leq \frac{1}{1-\alpha}\|x^* - \Pi x^*\|. \tag{1.5}$$

A slightly better bound is obtained when ΠT is a contraction mapping of modulus $\alpha \in [0,1)$ with respect to a Euclidean norm (e.g., $\|\cdot\|_\xi$). Then, using the Pythagorean Theorem, we have

$$\begin{aligned} \|x^* - \Phi r^*\|^2 &= \|x^* - \Pi x^*\|^2 + \|\Pi x^* - \Phi r^*\|^2 \\ &= \|x^* - \Pi x^*\|^2 + \|\Pi T(x^*) - \Pi T(\Phi r^*)\|^2 \\ &\leq \|x^* - \Pi x^*\|^2 + \alpha^2 \|x^* - \Phi r^*\|^2 \end{aligned}$$

from which we obtain

$$\|x^* - \Phi r^*\|^2 \leq \frac{1}{1-\alpha^2}\|x^* - \Pi x^*\|^2. \tag{1.6}$$

We note that in the case where ΠT is a contraction mapping with respect to some norm, there are some additional algorithmic approaches for approximation. In particular, we may consider a Jacobi/fixed point iteration, restricted within S, which involves projection of the iterates onto S:

$$\Phi r_{t+1} = \Pi T(\Phi r_t), \qquad t = 0, 1, \ldots. \tag{1.7}$$

In the context of DP this is known as *projected value iteration* (see [Ber07]). It converges to r^*, but is unwieldy when n is very large, because the vector $T(\Phi r_t)$ has dimension n.

The preceding observations suggest that it is desirable that ΠT be a contraction with respect to some norm. However, this is a complicated issue because *for a given ξ, ΠT need not be a contraction with respect to a given norm, even if T is a contraction with respect to that norm.* It is thus important to choose ξ, and the associated projection Π, in special ways that guarantee that ΠT is a contraction. This question will be discussed in Section 3.

In this paper, we introduce simulation-based algorithms for solving the equation $\Phi r = \Pi T(\Phi r)$. *The key desirable property of these algorithms is that they involve low-dimensional matrices and vectors, so they do not require n-dimensional calculations.* We consider two types of methods:

(a) *Equation approximation methods*, whereby r^* is approximated by $\hat{r}$, the solution of a linear system of the form

$$\Phi\hat{r} = \hat{\Pi}\hat{T}(\Phi\hat{r}), \tag{1.8}$$

where $\hat{\Pi}$ and $\hat{T}$ are simulation-based approximations to Π and T, respectively. As the number of simulation samples increases, $\hat{r}$ converges to r^*.

(b) *Approximate Jacobi methods*, which [without explicit calculation of $T(\Phi r_t)$] can be written in the form

$$\Phi r_{t+1} = \Pi T(\Phi r_t) + \epsilon_t, \qquad t = 0, 1, \ldots, \tag{1.9}$$

where ϵ_t is a simulation-induced error that diminishes to 0 as the number of simulation samples increases. Similar to the methods in (a), they do not require n-dimensional calculations, but apply only when ΠT is a contraction with respect to some norm. Then, since ϵ_t converges to 0, asymptotically iteration (1.9) becomes the projected Jacobi iteration (1.7), and r_t converges to r^*. We will also interpret later iteration (1.9) as a single iteration of an algorithm for solving the system (1.8).

Within the DP context, the approximation methods in (a) above have been proposed by Bradtke and Barto [BrB96], and by Boyan [Boy02] (see also analysis by Nedić and Bertsekas [NeB03]), and are known as *least squares temporal differences* (LSTD) methods. The approximate Jacobi methods in (b) above have been proposed by Bertsekas and Ioffe [BeI96] (see also analysis by Nedić and Bertsekas [NeB03], Bertsekas, Borkar, and Nedić [BBN04], Yu and Bertsekas [YuB06], and Bertsekas [Ber07]), and are known as *least squares policy evaluation* (LSPE) methods. An earlier, but computationally less effective method, is TD(λ), which was first proposed by Sutton [Sut88] and was instrumental in launching a substantial body of research on approximate DP in the 1990s (see Bertsekas and Tsitsiklis [BeT96], Sutton and Barto [SuB98], and Tsitsiklis and Van Roy [TsV97], [TsV99a] for discussion, extensions, and analysis of this method). Within the specialized approximate DP context, LSTD, LSPE, and TD(λ) offer some distinct advantages, which make them suitable for the approximate solution of problems involving Markov chains of very large dimension (in a case study where LSPE was used to evaluate the expected score of a game playing strategy, a Markov chain with more than 2^{200} states was involved; see [BeI96], and [BeT96], Section 8.3). These advantages are:

(1) The vector x need not be stored at any time. Furthermore, inner products involving the rows of A need not be computed; this can be critically important if some of the rows are not sparse.

(2) There is a projection norm such that the matrix ΠA is a contraction, so the bounds (1.6), (1.5), and other related bounds apply.

(3) The vector ξ of the projection norm need not be known explicitly. Instead, the values of the components of ξ are naturally incorporated within the simulation as relative frequencies of occurrence of the corresponding states (ξ is the invariant distribution vector of the associated Markov chain).

These advantages, particularly the first, make the simulation-based approach an attractive (possibly the only) option in problems so large that traditional approaches are prohibitively expensive in terms of time and storage. Indeed, even if ΠA were a contraction and a suitable vector ξ were explicitly known, the solution of the equation $\Phi r = \Pi T(\Phi r)$ with traditional non-simulation methods, would still require the formation of n-dimensional inner products, and would be prohibitive for n large enough. Of course, the simulation-based methods are also affected by a large n, and may in fact require more computation than traditional methods to obtain a highly accurate approximation to Φr^*. However, when nearly exact solution is prohibitively expensive, but solution accuracy is not critical, simulation-based methods may be able to produce adequate approximations to Φr^* much faster than traditional methods. An additional advantage of our methods is that they are far better suited for parallel computation than traditional methods, because the associated simulation is easily parallelizable.

The present paper extends the approximate DP methods just discussed to the case where A does not have the character of a stochastic matrix; just invertibility of $I - \Pi A$ is assumed. An important difficulty in the nonDP context considered here is that there may be no natural choice of ξ (and associated Markov chain to be used in the simulation process) such that ΠT is a contraction. Nonetheless, we show that all of the advantages (1)-(3) of LSTD, LSPE, and TD(λ) within the DP context are preserved under certain conditions, the most prominent of which is

$$|a_{ij}| \le p_{ij}, \qquad \forall\ i, j = 1, \ldots, n, \tag{1.10}$$

where a_{ij} are the components of A and p_{ij} are the transition probabilities of a Markov chain, which is used for simulation. In this case, again ξ is an invariant distribution of the chain and need not be known a priori. This is shown in Section 3, where some examples, including the important special case of a weakly diagonally dominant system, are also discussed.

When the condition $|a_{ij}| \le p_{ij}$, for all i, j, does not hold, the selection of the Markov chain used for simulation and the associated vector ξ used in the projection operation may be somewhat ad hoc. Furthermore, if ΠA is not a contraction, the approximate Jacobi methods are not valid and the bound (1.5) does not apply. Instead, the bound of Eq. (1.4) applies. If $I - \Pi A$ is nearly singular, this latter bound is poor, and the associated Eq. (1.3) suggests potential difficulties. Still, the equation approximation methods of Section 2 are valid, and maintain some important characteristics, namely that the vector x need not be stored at any time, and inner products involving the rows of A need not be computed.

We note that LSTD and LSPE are in fact entire classes of methods, parameterized with a scalar $\lambda \in [0, 1)$. They are called LSTD(λ) and LSPE(λ), respectively, and they use the parameter λ similar to the method of TD(λ). A value $\lambda > 0$ corresponds to approximating, in place of $x = T(x)$, the equation $x = T^{(\lambda)}(x)$, where $T^{(\lambda)}$ is the mapping

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

Note that the fixed points of T are also fixed points of $T^{(\lambda)}$, and that $T^{(\lambda)}$ coincides with T for $\lambda = 0$. However, it can be seen that when T is a contraction with respect to some norm with modulus $\alpha \in [0, 1)$,

$T^{(\lambda)}$ is a contraction with the more favorable modulus

$$\alpha_\lambda = (1-\lambda)\sum_{k=0}^{\infty} \lambda^k \alpha^{k+1} = \frac{\alpha(1-\lambda)}{1-\alpha\lambda}.$$

Thus, when approximating the equation $x = T^{(\lambda)}(x)$, rather than $x = T(x)$, the error bounds (1.6)-(1.5) become more favorable; in fact $\alpha_\lambda \to 0$ as $\lambda \to 1$, so asymptotically, from Eq. (1.5), we obtain optimal approximation: $\|x^* - \Phi r^*\| = \|x^* - \Pi x^*\|$. Furthermore, $T^{(\lambda)}$ and $\Pi T^{(\lambda)}$ are arbitrarily close to 0, if λ is sufficiently close to 1, so they can become contractions with respect to any norm. An important characteristic of our methods is that under certain conditions [typically Eq. (1.10)], they can be straightforwardly applied to the equation $x = T^{(\lambda)}(x)$, while this is much harder with traditional methods (see Section 5). However, while the error bounds improve as λ is increased towards 1, the simulation required to solve the equation $\Phi r = \Pi T^{(\lambda)}(\Phi r)$ becomes more time-consuming because the associated simulation samples become more "noisy." This "accuracy-noise" tradeoff is widely recognized in the approximate DP literature (see e.g., the textbook [Ber07] and the references quoted there).

In this paper, we focus on describing the methods, making the connection with their DP antecedents, proving some basic results relating to contraction properties of ΠA, and providing some examples of interesting special cases. A more detailed delineation of important relevant classes of problems, and the associated formulations, projection norm selections, and other related issues, are beyond the scope of the present paper, and are a subject for further research.

The paper is organized as follows. In Section 2, we formulate the simulation framework that underlies our methods, and we discuss the equation approximation approach of (a) above. We also briefly consider an alternative approach, which is based on minimization of the equation error norm

$$\|\Phi r - A\Phi r - b\|_\xi.$$

In Section 3, we discuss the selection of Markov chains, together with some related examples and special cases, for which various contraction properties can be shown. In Section 4, we develop the approximate Jacobi methods in (b) above, and we discuss their relation with the equation approximation methods in (a) above. In Section 5, we develop multistep analogs of the methods of Section 2 and 4, in the spirit of the LSTD(λ), LSPE(λ), and TD(λ) methods of approximate DP. The methods of Sections 2-5 assume that the rows of Φ are either explicitly known or can be easily generated when needed. In Section 6, we discuss special methods that use basis functions of the form $A^m g$, $m \geq 0$, where g is some vector the components of which can be exactly computed. These methods bear similarity to Krylov subspace methods (see e.g. Saad [Saa03]), but suffer from the potential difficulty that the rows of $A^m g$ may be hard to compute. We discuss variants of our methods of Sections 2-5 where the rows of $A^m g$ are approximated by simulation of a single sample. These variants are new even in the context of approximate DP ($A = \alpha P$), where generating appropriate basis vectors for cost function approximation is a currently prominent research issue. Finally, in Section 7, we discuss some extensions of our methods that apply to certain nonlinear fixed point problems. We generalize approximate DP methods proposed for optimal stopping problems (see Tsitsiklis and Van Roy [TsV99b], Choi and Van Roy [ChV06], and Yu and Bertsekas [YuB07]), and a method for approximating the dominant eigenvalue and corresponding eigenvector of a nonnegative matrix, first proposed by Basu, Bhatatacharyya, and Borkar [BBB6].

Regarding notation, throughout the paper, vectors are considered to be column vectors, and a prime denotes transposition. We generally use subscripts to indicate the scalar components of various vectors and matrices. Vector and matrix inequalities are to be interpreted componentwise. For example, for two matrices A, B, the inequality $A \leq B$ means that $A_{ij} \leq B_{ij}$ for all i and j.

2. EQUATION APPROXIMATION METHODS

In this section, we discuss the construction of simulation-based approximations to the projected equation $\Phi r = \Pi(A\Phi r + b)$. This methodology descends from the LSTD methods of approximate DP, referred to in Section 1. Let us assume that the positive distribution vector ξ is given. By the definition of projection with respect to $\|\cdot\|_\xi$, the unique solution r^* of this equation satisfies

$$r^* = \arg\min_r \sum_{i=1}^n \xi_i \left(\phi(i)'r - \sum_{j=1}^n a_{ij}\phi(j)'r^* - b_i \right)^2,$$

where $\phi(i)'$ denotes the ith row of the matrix Φ. By setting the gradient of the minimized expression above to 0, we have

$$\sum_{i=1}^n \xi_i \phi(i) \left(\phi(i)'r^* - \sum_{j=1}^n a_{ij}\phi(j)'r^* - b_i \right) = 0.$$

We thus obtain the following equivalent form of the projected equation $\Phi r = \Pi(A\Phi r + b)$:

$$\sum_{i=1}^n \xi_i \phi(i) \left(\phi(i) - \sum_{j=1}^n a_{ij}\phi(j) \right)' r^* = \sum_{i=1}^n \xi_i \phi(i) b_i. \tag{2.1}$$

The key idea of our methodology can be simply explained by focusing on the two expected values with respect to ξ, which appear in the left and right sides of the above equation: *we approximate these two expected values by simulation-obtained sample averages*. For this, we need to generate a sequence of indices i according to the probabilities ξ_i. A convenient way to do so is by using a Markov chain that has states $1, \ldots, n$ and has ξ as an invariant distribution. State i of the chain will be associated with the index of the component x_i of x. We leave open for the time being the method to select the transition probabilities p_{ij} of the chain, except for the requirement

$$p_{ij} > 0 \qquad \text{if} \qquad a_{ij} \neq 0, \tag{2.2}$$

as well as the implicit requirement that the chain should have no transient states (since otherwise it would not have an invariant distribution with all components positive). In Section 3, we will discuss favorable methods for constructing p_{ij} from a_{ij}, which lead to bounds for the error $\|x^* - \Phi r^*\|_\xi$ and contraction properties for ΠT.

In the most basic form of our methods, we generate a sequence of states $\{i_0, i_1, \ldots\}$, and a sequence of transitions $\big\{(i_0, j_0), (i_1, j_1), \ldots\big\}$ of the chain. We use any probabilistic mechanism for this, subject to the following two requirements (cf. Fig. 2.1):

(1) The sequence $\{i_0, i_1, \ldots\}$ is generated according to the distribution ξ, which defines the projection norm $\|\cdot\|_\xi$, in the sense that with probability 1,

$$\lim_{t\to\infty} \frac{\sum_{k=0}^t \delta(i_k = i)}{t+1} = \xi_i, \qquad i = 1, \ldots, n, \tag{2.3}$$

where $\delta(\cdot)$ denotes the indicator function [$\delta(E) = 1$ if the event E has occurred and $\delta(E) = 0$ otherwise].

(2) The sequence $\big\{(i_0,j_0),(i_1,j_1),\ldots\big\}$ is generated according to the transition probabilities p_{ij} of the Markov chain, in the sense that with probability 1,

$$\lim_{t\to\infty}\frac{\sum_{k=0}^t \delta(i_k=i,j_k=j)}{\sum_{k=0}^t \delta(i_k=i)}=p_{ij},\qquad i,j=1,\ldots,n. \tag{2.4}$$

At time t, we form the linear equation

$$\sum_{k=0}^t \phi(i_k)\left(\phi(i_k)-\frac{a_{i_kj_k}}{p_{i_kj_k}}\phi(j_k)\right)'r=\sum_{k=0}^t \phi(i_k)b_{i_k}. \tag{2.5}$$

We claim that this is a valid approximation to Eq. (2.1), the equivalent form of the projected equation. The idea is to view

$\phi(i_k)\phi(i_k)'$ as a sample whose steady-state expected value is the matrix $\sum_{i=1}^n \xi_i\phi(i)\phi(i)'$,

$\frac{a_{i_kj_k}}{p_{i_kj_k}}\phi(i_k)\phi(j_k)'$ as a sample whose steady-state expected value is the matrix $\sum_{i=1}^n \xi_i\phi(i)\sum_{j=1}^n a_{ij}\phi(j)'$,

$\phi(i_k)b_{i_k}$ as a sample whose steady-state expected value is the vector $\sum_{i=1}^n \xi_i\phi(i)b_i$.

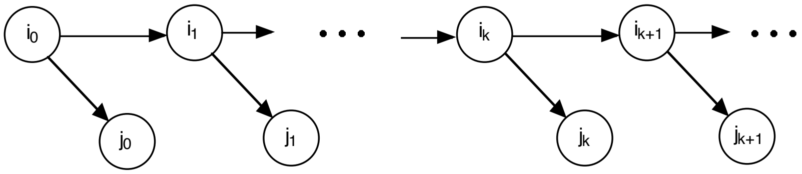

Figure 2.1. One method to generate states according to the distribution ξ [cf. Eq. (2.3)] for the case of an irreducible Markov chain is to simulate a single infinitely long sample trajectory $\{i_0,i_1,\ldots\}$ of the chain. The transition sequence $\{(i_0,j_0),(i_1,j_1),\ldots\}$ is simultaneously generated according to the transition probabilities p_{ij}. It is possible that $j_k=i_{k+1}$, but this is not necessary.

Indeed, by counting the number of times a state occurs and collecting terms, we can write Eq. (2.5) as

$$\sum_{i=1}^n \hat\xi_{i,t}\phi(i)\left(\phi(i)-\sum_{j=1}^n \hat p_{ij,t}\frac{a_{ij}}{p_{ij}}\phi(j)\right)'r=\sum_{i=1}^n \hat\xi_{i,t}\phi(i)b_i, \tag{2.6}$$

where

$$\hat\xi_{i,t}=\frac{\sum_{k=0}^t\delta(i_k=i)}{t+1},\qquad \hat p_{ij,t}=\frac{\sum_{k=0}^t\delta(i_k=i,j_k=j)}{\sum_{k=0}^t\delta(i_k=i)}.$$

In view of the assumption

$$\hat\xi_{i,t}\to\xi_i,\qquad \hat p_{ij,t}\to p_{ij},\qquad i,j=1,\ldots,n,$$

[cf. Eqs. (2.3) and (2.4)], by comparing Eqs. (2.1) and (2.6), we see that they asymptotically coincide. Since the solution r^* of the system (2.1) exists and is unique, the same is true for the system (2.6) for all t sufficiently large. Thus, with probability 1,

$$\hat{r}_t \to r^*,$$

where $\hat{r}_t$ is the solution of the system (2.5).

A comparison of Eqs. (2.1) and (2.6) indicates some considerations for selecting the Markov chain (other than that it should have no transient states). It can be seen that “important” (e.g., large) components a_{ij} should be simulated more often (p_{ij}: large).† In particular, if (i,j) is such that $a_{ij}=0$, there is an incentive to choose $p_{ij}=0$, since corresponding transitions (i,j) are “wasted” in that they do not contribute to improvement of the approximation of Eq. (2.1) by Eq. (2.6). This suggests that the structure of the Markov chain should match in some sense the structure of the matrix A. This point will be further discussed in Section 3.

Note that there is a lot of flexibility for generating the state sequence $\{i_0,i_1,\ldots\}$ and the transition sequence $\big\{(i_0,j_0),(i_1,j_1),\ldots\big\}$ to satisfy Eqs. (2.3) and (2.4). One possibility is to start at some state i_0 and generate a single infinitely long trajectory of the Markov chain, in which case $j_k=i_{k+1}$ for all k. Then the requirements (2.3) and (2.4) will be satisfied if the Markov chain is irreducible, in which case ξ will be the unique invariant distribution of the chain and will have positive components; this is an important special case in approximate DP applications (see the references given in Section 1), and will be discussed further in the next section.

An alternative possibility is to generate multiple infinitely long trajectories of the chain, starting at several different states, and for each trajectory use $j_k=i_{k+1}$ for all k. This will work even if the chain has multiple recurrent classes, as long as there are no transient states and at least one trajectory is started from within each recurrent class. Again ξ will be an invariant distribution of the chain, and need not be known explicitly. Note that using multiple trajectories may be interesting even if there is a single recurrent class, for at least two reasons:

(1) The generation of trajectories may be parallelized among multiple processors, with significant speedup resulting.

† For a simplified analysis, note that the variance of each coefficient $\hat{p}_{ij,t}\frac{a_{ij}}{p_{ij}}$ appearing in Eq. (2.6) can be calculated to be

$$V_{ij,t}=\gamma_t p_{ij}(1-p_{ij})\frac{a_{ij}^2}{p_{ij}^2}=\frac{\gamma_t a_{ij}^2}{p_{ij}}-\gamma_t a_{ij}^2,$$

where γ_t is the expected value of $1/\sum_{k=0}^t\delta(i_k=i)$, assuming the initial state i_0 is distributed according to ξ. [To see this, note that $\hat{p}_{ij,t}$ is the average of Bernoulli random variables whose mean and variance are p_{ij} and $p_{ij}(1-p_{ij})$, respectively, and whose number is the random variable $\sum_{k=0}^t\delta(i_k=i)$.] For a given i, let us consider the problem of finding p_{ij}, $j=1,\ldots,n$, that minimize $\sum_{j=1}^n V_{ij,t}$ subject to the constraints $p_{ij}=0$ if and only if $a_{ij}=0$, and $\sum_{j=1}^n p_{ij}=1$. By introducing a Lagrange multiplier ν for the constraint $\sum_{j=1}^n p_{ij}=1$, and forming and minimizing the corresponding Lagrangian, we see that the optimal solution satisfies $\frac{a_{ij}^2}{p_{ij}^2}=\frac{\nu}{\gamma_t}$, implying that p_{ij} should be chosen proportional to $|a_{ij}|$ (indeed this is standard practice in approximate DP, and is consistent with the principles of importance sampling [Liu01]). This analysis, however, does not take into account the fact that the choice of p_{ij} also affects the steady-state probabilities ξ_i, and through them the variance of both sides of Eq. (2.6). In order to optimize more meaningfully the choice of p_{ij}, this relation must be taken into account, as well as the dependence of the variance of the solution of Eq. (2.6) on other terms, such as the vectors $\phi(i)$ and b.

(b) The empirical frequencies of occurrence of the states may approach the invariant probabilities more quickly; this is particularly so for large and "stiff" Markov chains.

An important observation is that it is not important to explicitly know ξ; it is just necessary to know the Markov chain and to be able to simulate its transitions. The choice of the Markov chain is significant, however, since it determines ξ, and hence it determines whether ΠT is a contraction and the error bounds (1.5) or (1.6) of Section 1 apply. Furthermore, the Markov chain should be chosen so that it has no transient states (so that $\xi_i > 0$ for all i) and it can be conveniently simulated. Sometimes the choice of the Markov chain is evident (this is true for example in DP). Some other cases where there are favorable choices of the Markov chain will be discussed in the next section.

A Variant With Independently Simulated Transitions

The preceding algorithm requires that the sequence of transitions is generated according to transition probabilities p_{ij}. However, this is not necessary, although it saves some computation in the case where $j_k = i_{k+1}$. In particular, we may generate the state sequence $\{i_0, i_1, \ldots\}$ in the same way as before using the transition probabilities p_{ij} [cf. Eq. (2.3)], but generate the transition sequence $\{(i_0, j_0), (i_1, j_1), \ldots\}$ according to other transition probabilities $\tilde{p}_{ij}$ that satisfy

$$\tilde{p}_{ij} > 0 \qquad \text{if} \qquad a_{ij} \neq 0,$$

and

$$\lim_{t\to\infty} \frac{\sum_{k=0}^{t} \delta(i_k = i, j_k = j)}{\sum_{k=0}^{t} \delta(i_k = i)} = \tilde{p}_{ij}, \qquad i, j = 1, \ldots, n, \tag{2.7}$$

with probability 1. At time t, we obtain $\hat{r}_t$ as the solution of the linear equation

$$\sum_{k=0}^{t} \phi(i_k) \left(\phi(i_k) - \frac{a_{i_k j_k}}{\tilde{p}_{i_k j_k}} \phi(j_k) \right)' r = \sum_{k=0}^{t} \phi(i_k) b_{i_k}. \tag{2.8}$$

The preceding analysis carries through and shows that $\hat{r}_t \to r^*$ with probability 1.

A Variant Without Simulated Transitions

Let us note an alternative approximation, which requires the generation of a sequence $\{i_0, i_1, \ldots\}$ according to the distribution ξ [cf. Eq. (2.3)], but does not require the sequence of transitions $\{(i_0, j_0), (i_1, j_1), \ldots\}$. In this approach, at time t, we form the linear equation

$$\sum_{k=0}^{t} \phi(i_k) \left(\phi(i_k) - \sum_{j=1}^{n} a_{i_k j} \phi(j) \right)' r = \sum_{k=0}^{t} \phi(i_k) b_{i_k}. \tag{2.9}$$

Clearly, the solution of this system exists for t sufficiently large (with probability 1), and similar to the preceding analysis, it can be shown to converge to r^*. A potential difficulty with this method is that the summation over j in Eq. (2.9) may be very time-consuming (proportional to n). On the other hand, if each row of A has a relatively small number of easily computable nonzero components, the approach based on solution of Eq. (2.9) may be competitive or superior to the approach based on Eq. (2.5), because it involves less simulation "noise."

Note that in cases where a favorable choice of ξ is explicitly known and the formation of the sums $\sum_{j=1}^{n} a_{i_k j}\phi(j)$ in Eq. (2.9) are not prohibitively expensive, one need not use a Markov chain, but rather just generate a sequence $\{i_0, i_1, \ldots\}$ according to the probabilities ξ_i, and then replace Eq. (2.9) with

$$\sum_{i=1}^{n} \delta\big(\{i_k = i \text{ for some } k \in [0,t]\}\big)\xi_i\phi(i)\left(\phi(i) - \sum_{j=1}^{n} a_{ij}\phi(j)\right)' r = \sum_{i=1}^{n} \delta\big(\{i_k = i \text{ for some } k \in [0,t]\}\big)\xi_i\phi(i)b_i. \tag{2.10}$$

The solution of Eq. (2.10) converges with probability 1 to r^* as $t \to \infty$, and the rate of convergence will likely be faster than the one of Eq. (2.9). In particular, the terms corresponding to different states in Eq. (2.10) are weighted with the same weights as in the true projected equation (2.1).

There is also a method that is intermediate between the one based on Eq. (2.9) and the one based on Eq. (2.5). In this method, we partition the set of indices $\{1, \ldots, n\}$ into "blocks" $J_1, \ldots, J_M$, i.e.,

$$\{1, \ldots, n\} = J_1 \cup \cdots \cup J_M,$$

and we write

$$\sum_{j=1}^{n} a_{ij}\phi(j) = \sum_{m=1}^{M} \sum_{j\in J_m} a_{ij}\phi(j).$$

Then, when at state i, instead of sampling the columns j of A with probabilities p_{ij}, we sample the blocks J_m with some probabilities $\tilde{p}_{im}$, and instead of Eq. (2.9) or (2.5), we solve the equation

$$\sum_{k=0}^{t} \phi(i_k)\left(\phi(i_k) - \sum_{j\in J_{m_k}} \frac{a_{i_k j}}{\tilde{p}_{i_k m_k}}\phi(j)\right)' r = \sum_{k=0}^{t} \phi(i_k)b_{i_k},$$

where J_{m_k} is the block sampled at state i_k.

Variants With Noisy Samples of the Problem Data

In further variants of the preceding iterations, zero mean noise with appropriate independence properties may be added to $a_{i_k j}$ and b_{i_k}. For example, Eq. (2.9) may be replaced by

$$\sum_{k=0}^{t} \phi(i_k)\left(\phi(i_k) - \sum_{j=1}^{n}\big(a_{i_k j} + \zeta_k(j)\big)\phi(j)\right)' r = \sum_{k=0}^{t} \phi(i_k)(b_{i_k} + \theta_k), \tag{2.11}$$

where for each j, $\zeta_k(j)$ is a sequence of random variables such that, with probability 1,

$$\lim_{t\to\infty} \frac{\sum_{k=0}^{t} \delta(i_k = i)\zeta_k(j)}{t+1} = 0, \qquad \forall\ i, j = 1, \ldots, n, \tag{2.12}$$

and θ_k is a sequence of random variables such that, with probability 1,

$$\lim_{t\to\infty} \frac{\sum_{k=0}^{t} \delta(i_k = i)\theta_k}{t+1} = 0, \qquad \forall\ i = 1, \ldots, n. \tag{2.13}$$

This variant can be used in situations where the components a_{ij} and b_i represent the expected values of random variables whose samples can be conveniently simulated with additive "noises" $\zeta_k(j)$ and θ_k, respectively, such that Eqs. (2.12) and (2.13) hold with probability 1.

There are also other variants where $a_{i_kj_k}$ and b_{i_k} are expected values, which are replaced in the earlier formulas by suitable weighted samples. For example, if b_i has the form

$$b_i = \sum_{j=1}^n q_{ij}c(i,j),$$

where $c(i,j)$ are given scalars and q_{ij} are transition probabilities, we may replace b_{i_k} in Eq. (2.5) by

$$\frac{q_{i_kj_k}}{p_{i_k,j_k}}c(i_k,j_k).$$

An Alternative: Minimizing the Equation Error Norm

Let us finally consider an alternative approach for approximate solution of the equation $x = T(x)$, based on finding a vector r that minimizes†

$$\|\Phi r - T(\Phi r)\|_\xi^2,$$

or

$$\sum_{i=1}^n \xi_i \left(\phi(i)'r - \sum_{j=1}^n a_{ij}\phi(j)'r - b_i\right)^2.$$

In the DP literature, this is known as the *Bellman equation error approach*. We assume that the matrix $(I-A)\Phi$ has rank s, which guarantees that the vector r^* that minimizes the weighted sum of squared errors is unique. A detailed comparison of this approach and the earlier approach based on solving the projected equation is beyond the scope of the present paper. However, the simulation-based solution methods and the

† Error bounds similar to the ones of Eq. (1.4) and (1.5) can be developed for this approach, assuming that $I - A$ is invertible and x^* is the unique solution. In particular, let $\tilde r$ minimize $\|\Phi r - T(\Phi r)\|_\xi^2$. Then

$$x^* - \Phi\tilde r = x^* - T(\Phi\tilde r) + T(\Phi\tilde r) - \Phi\tilde r = Tx^* - T(\Phi\tilde r) + T(\Phi\tilde r) - \Phi\tilde r = A(x^* - \Phi\tilde r) + T(\Phi\tilde r) - \Phi\tilde r,$$

so that

$$x^* - \Phi\tilde r = (I-A)^{-1}\big(T(\Phi\tilde r) - \Phi\tilde r\big).$$

Thus, we obtain

$$\begin{aligned}\|x^* - \Phi\tilde r\|_\xi &\le \big\|(I-A)^{-1}\big\|_\xi \|\Phi\tilde r - T(\Phi\tilde r)\|_\xi \\ &\le \big\|(I-A)^{-1}\big\|_\xi \big\|\Pi x^* - T(\Pi x^*)\big\|_\xi \\ &= \big\|(I-A)^{-1}\big\|_\xi \big\|\Pi x^* - x^* + Tx^* - T(\Pi x^*)\big\|_\xi \\ &= \big\|(I-A)^{-1}\big\|_\xi \big\|(I-A)(\Pi x^* - x^*)\big\|_\xi \\ &\le \big\|(I-A)^{-1}\big\|_\xi \|I-A\|_\xi \|x^* - \Pi x^*\|_\xi,\end{aligned}$$

where the second inequality holds because $\tilde r$ minimizes $\|\Phi r - T(\Phi r)\|_\xi^2$. In the case where T is a contraction mapping with respect to the norm $\|\cdot\|_\xi$, with modulus $\alpha \in (0,1)$, a similar calculation yields

$$\|x^* - \Phi\tilde r\|_\xi \le \frac{1+\alpha}{1-\alpha}\|x^* - \Pi x^*\|_\xi.$$

analysis of the two approaches are quite similar, so the alternative equation error-based approach is worth mentioning here.

The optimal solution r^* satisfies the corresponding necessary optimality condition

$$\sum_{i=1}^{n} \xi_i \left(\phi(i) - \sum_{j=1}^{n} a_{ij}\phi(j) \right) \left(\phi(i) - \sum_{j=1}^{n} a_{ij}\phi(j) \right)' r^* = \sum_{i=1}^{n} \xi_i \left(\phi(i) - \sum_{j=1}^{n} a_{ij}\phi(j) \right) b_i. \tag{2.14}$$

A simulation-based approximation to this equation, which requires the formation of row sums as in Eq. (2.9), is the linear equation

$$\sum_{k=0}^{t} \left(\phi(i_k) - \sum_{j=1}^{n} a_{i_k j}\phi(j) \right) \left(\phi(i_k) - \sum_{j=1}^{n} a_{i_k j}\phi(j) \right)' r = \sum_{k=0}^{t} \left(\phi(i_k) - \sum_{j=1}^{n} a_{i_k j}\phi(j) \right) b_{i_k}. \tag{2.15}$$

Similar to our earlier analysis, it can be seen that the solution to this equation approaches r^*, the solution to Eq. (2.14), as $t \to \infty$.

To obtain a simulation-based approximation to Eq. (2.14), without requiring the calculation of row sums of the form $\sum_{j=1}^{n} a_{ij}\phi(j)$, we introduce an additional sequence of transitions $\{(i_0, j_0'), (i_1, j_1'), \ldots\}$, which is generated according to the transition probabilities p_{ij} of the Markov chain, and is also "independent" of the sequence $\{(i_0, j_0), (i_1, j_1), \ldots\}$ in the sense that with probability 1,

$$\lim_{t\to\infty} \frac{\sum_{k=0}^{t} \delta(i_k = i, j_k = j)}{\sum_{k=0}^{t} \delta(i_k = i)} = \lim_{t\to\infty} \frac{\sum_{k=0}^{t} \delta(i_k = i, j_k' = j)}{\sum_{k=0}^{t} \delta(i_k = i)} = p_{ij}, \qquad i, j = 1, \ldots, n, \tag{2.16}$$

and

$$\lim_{t\to\infty} \frac{\sum_{k=0}^{t} \delta(i_k = i, j_k = j, j_k' = j')}{\sum_{k=0}^{t} \delta(i_k = i)} = p_{ij}p_{ij'}, \qquad i, j, j' = 1, \ldots, n; \tag{2.17}$$

(see Fig. 2.2). At time t, we form the linear equation

$$\sum_{k=0}^{t} \left(\phi(i_k) - \frac{a_{i_k j_k}}{p_{i_k j_k}}\phi(j_k) \right) \left(\phi(i_k) - \frac{a_{i_k j_k'}}{p_{i_k j_k'}}\phi(j_k') \right)' r = \sum_{k=0}^{t} \left(\phi(i_k) - \frac{a_{i_k j_k}}{p_{i_k j_k}}\phi(j_k) \right) b_{i_k}. \tag{2.18}$$

Similar to our earlier analysis, it can be seen that this is a valid approximation to Eq. (2.14). In what follows, we will focus on the solution of the equation $\Phi r = \Pi T(\Phi r)$, but some of the analysis of the next section is relevant to the simulation-based minimization of $\|\Phi r - T(\Phi r)\|_\xi^2$, using Eq. (2.15) or Eq. (2.18).

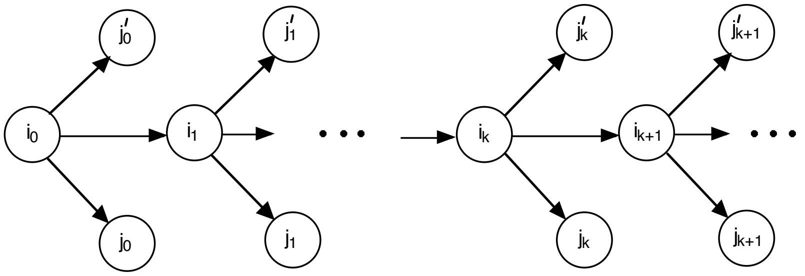

Figure 2.2. A possible simulation mechanism for minimizing the equation error norm [cf. Eq. (2.18)]. We generate a sequence of states $\{i_0, i_1, \ldots\}$ according to the distribution ξ, by simulating a single infinitely long sample trajectory of the chain. Simultaneously, we generate two independent sequences of transitions, $\{(i_0, j_0), (i_1, j_1), \ldots\}$ and $\{(i_0, j_0'), (i_1, j_1'), \ldots\}$, according to the transition probabilities p_{ij}, so that Eqs. (2.16) and (2.17) are satisfied.

Let us finally note that the equation error approach can be generalized to yield a simulation-based method for solving the general linear least squares problem

$$\min_r \sum_{i=1}^n \xi_i \left(c_i - \sum_{j=1}^m q_{ij}\phi(j)'r \right)^2,$$

where q_{ij} are the components of an $n \times m$ matrix Q, and c_i are the components of a vector $c \in \Re^n$. In particular, one may write the corresponding optimality condition [cf. Eq. (2.14)] and then approximate it by simulation [cf. Eq. (2.18)].

3. MARKOV CHAIN CONSTRUCTION

The methods of the preceding section leave open the issue of selection of the distribution vector ξ and the transition probabilities p_{ij}. The present section gives some results, which can be used to prove interesting properties such as contraction of ΠT with associated error bounds. These results are illustrated through their applications in some interesting classes of problems.

We denote by P the matrix whose components are the transition probabilities p_{ij}. Thus, ξ is an invariant distribution of P if $\xi' = \xi' P$. We also denote by $|A|$ the matrix whose components are

$$|A|_{ij} = |a_{ij}|, \qquad i, j = 1, \ldots, n.$$

Generally, it seems hard to guarantee that ΠT is a contraction mapping, unless $|A| \leq P$. The following propositions assume this condition, as well as additional conditions that guarantee that ΠT is a contraction, so that the error bounds of Eqs. (1.5) and (1.6) apply.

Proposition 1: Assume that there are no transient states corresponding to P, that ξ is an invariant distribution of P, and that one of the following two conditions holds:

(1) For some scalar $\alpha \in (0, 1)$, we have $|A| \leq \alpha P$.

(2) We have $|A| \leq P$, and there exists an index $\bar{i}$ such that

$$|a_{\bar{i}j}| < p_{\bar{i}j}, \qquad \forall\ j = 1, \ldots, n.$$

Then T and ΠT are contraction mappings with respect to the norm $\|\cdot\|_\xi$, and under condition (1), the modulus of contraction is α.

Proof: Since Π is nonexpansive with respect to $\|\cdot\|_\xi$, it will suffice to show that A is a contraction with respect to $\|\cdot\|_\xi$.

Assume condition(1). Then for all $z \in \Re^n$ with $z \neq 0$, we have (considering 0/0 to be 0)

$$\|Az\|_\xi^2 = \sum_{i=1}^n \xi_i \left(\sum_{j=1}^n a_{ij} z_j \right)^2$$

$$
\begin{aligned}
&= \sum_{i=1}^{n} \xi_i \left(\sum_{j=1}^{n} p_{ij} \frac{a_{ij}}{p_{ij}} z_j \right)^2 \\
&\le \sum_{i=1}^{n} \xi_i \sum_{j=1}^{n} p_{ij} \left(\frac{a_{ij}}{p_{ij}} z_j \right)^2 \\
&\le \alpha^2 \sum_{i=1}^{n} \xi_i \sum_{j=1}^{n} p_{ij} z_j^2 \\
&= \alpha^2 \sum_{j=1}^{n} \sum_{i=1}^{n} \xi_i p_{ij} z_j^2 \\
&= \alpha^2 \sum_{j=1}^{n} \xi_j z_j^2 \\
&= \alpha^2 \|z\|_\xi^2,
\end{aligned}
$$

where the first inequality follows from the convexity of the quadratic function, the second inequality follows from the assumption $|A| \le \alpha P$, and the next to last equality follows from the property $\sum_{i=1}^{n} \xi_i p_{ij} = \xi_j$ of the invariant distribution. Thus, A is a contraction with respect to $\| \cdot \|_\xi$ with modulus α.

The proof under condition (2) is similar, by using the preceding calculation with $\alpha = 1$ and with the second inequality replaced by a strict inequality. We thus obtain $\|Az\|_\xi < \|z\|_\xi$ for all $z \ne 0$, and it follows that A is a contraction with respect to $\| \cdot \|_\xi$, with modulus $\max_{\|z\|_\xi \le 1} \|Az\|_\xi$. **Q.E.D.**

Proposition 1 can be used in the case where

$$\sum_{j=1}^{n} |a_{ij}| \le 1, \qquad \forall\ i = 1, \ldots, n, \tag{3.1}$$

with strict inequality for at least one index i, as a potential means to construct a transition probability matrix P for which ΠT is a contraction with respect to $\| \cdot \|_\xi$. In particular, let

$$P = |A| + (e - |A|e)q', \tag{3.2}$$

where q is a probability distribution vector with positive components, and e is the unit vector that has all components equal to 1. We note that $|A| \le P$ and since $q'e = 1$, we have

$$Pe = |A|e + (e - |A|e)q'e = |A|e + (e - |A|e) = e,$$

so that P is a transition probability matrix. To interpret P as given by Eq. (3.2), consider an additional artificial restart state 0, and the following transition mechanism: at state i, we either go to state j (probability $|a_{ij}|$), or we go to an artificial "restart" state (probability $1 - \sum_{m=1}^{n} |a_{im}|$) from which we immediately move to state j with probability q_j, the jth component of q. Note the following generalization of Eq. (3.2), which uses different restart probabilities at each state i:

$$P = |A| + diag(e - |A|e)Q,$$

where Q is a transition probability matrix, and $diag(e - |A|e)$ is the diagonal matrix with $1 - \sum_{m=1}^{n} |a_{im}|$, $i = 1, \ldots, n$, on the diagonal, so the row sum deficit of row i is distributed to the columns j according to fractions q_{ij}.

As mentioned in Section 2, intuition suggests that the structure of P should be close to the structure of $|A|$, in the sense that we should aim to choose $p_{ij} = 0$ whenever $a_{ij} = 0$. The reason is that a transition (i, j) for which $a_{ij} = 0$ does not contribute to improvement of the approximation of Eq. (2.1) by Eq. (2.6). There is therefore an incentive to use as few positive restart probabilities as possible, consistently with the requirement that P should have no transient states, in order to minimize the number of pairs (i, j) for which $p_{ij} > 0$ but $a_{ij} = 0$. The following proposition becomes useful within this context, as it allows the use of restart probabilities that are not necessarily all positive.

Proposition 2: Assume that P is irreducible, ξ is its invariant distribution, $|A| \leq P$, and there exists an index $\bar{\imath}$ such that

$$\sum_{j=1}^{n} |a_{\bar{\imath}j}| < 1.$$

Then ΠT is a contraction with respect to some norm.

Proof: It will suffice to show that the eigenvalues of ΠA lie strictly within the unit circle.† Let $\bar{P}$ be the matrix which is identical to P except for the $\bar{\imath}$th row which is identical to the $\bar{\imath}$th row of $|A|$. From the irreducibility of P, it follows that for any $i_1 \neq \bar{\imath}$ it is possible to find a sequence of nonzero components $\bar{P}_{i_1 i_2}, \ldots, \bar{P}_{i_{k-1} i_k}, \bar{P}_{i_k \bar{\imath}}$ that "lead" from i_1 to $\bar{\imath}$. Using a well-known result, we have $\bar{P}^t \to 0$. Since $|A| \leq \bar{P}$, we also have $|A|^t \to 0$, and hence also $A^t \to 0$ (since $|A^t| \leq |A|^t$). Thus, all eigenvalues of A are strictly within the unit circle. Next, by essentially repeating the proof of Prop. 1, we see that

$$\|\Pi A z\|_\xi \leq \|z\|_\xi, \qquad \forall\ z \in \Re^n,$$

so the eigenvalues of ΠA cannot lie outside the unit circle.

Assume to arrive at a contradiction that ν is an eigenvalue of ΠA with $|\nu| = 1$, and let ζ be a corresponding eigenvector. We claim that $A\zeta$ must have both real and imaginary components in the subspace S. If this were not so, we would have $A\zeta \neq \Pi A \zeta$, so that

$$\|A\zeta\|_\xi > \|\Pi A \zeta\|_\xi = \|\nu\zeta\|_\xi = |\nu|\ \|\zeta\|_\xi = \|\zeta\|_\xi,$$

which contradicts the fact $\|Az\|_\xi \leq \|z\|_\xi$ for all z, shown earlier. Thus, the real and imaginary components of $A\zeta$ are in S, which implies that $A\zeta = \Pi A \zeta = \nu\zeta$, so that ν is an eigenvalue of A. This is a contradiction because $|\nu| = 1$, while the eigenvalues of A are strictly within the unit circle. **Q.E.D.**

† We use here the fact that if a square matrix has eigenvalues strictly within the unit circle, then there exists a norm with respect to which the linear mapping defined by the matrix is a contraction. Also in the following argument, the projection Πz of a complex vector z is obtained by separately projecting the real and the imaginary components of z on S. The projection norm for a complex vector $x + iy$ is defined by

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

Note that under the assumptions of Prop. 2, T and ΠT need not be contractions with respect to the norm $\|\cdot\|_\xi$. For a counterexample, take $a_{i,i+1} = 1$ for $i = 1, \ldots, n-1$, and $a_{n,1} = 1/2$, with every other entry of A equal to 0. Take also $p_{i,i+1} = 1$ for $i = 1, \ldots, n-1$, and $p_{n,1} = 1$, with every other entry of P equal to 0, so $\xi_i = 1/n$ for all i. Then for $z = (0, 1, \ldots, 1)'$ we have $Az = (1, \ldots, 1, 0)'$ and $\|Az\|_\xi = \|z\|_\xi$, so A is not a contraction with respect to $\|\cdot\|_\xi$. Taking S to be the entire space $\Re^n$, we see that the same is true for ΠA.

The next proposition uses different assumptions than Props. 1 and 2, and applies to cases where there is no special index $\bar{i}$ such that $\sum_{j=1}^n |a_{\bar{i}j}| < 1$. In fact A may itself be a transition probability matrix, so that $I - A$ need not be invertible, and the original system may have multiple solutions; see the subsequent Example 3. The proposition suggests the use of a damped version of the T mapping in various methods, and is closely connected to a result on approximate DP methods for average cost problems ([YuB06], Prop. 3).

Proposition 3: Assume that there are no transient states corresponding to P, that ξ is an invariant distribution of P, and that $|A| \leq P$. Assume further that $I - \Pi A$ is invertible. Then the mapping ΠT_γ, where

$$T_\gamma = (1-\gamma)I + \gamma T,$$

is a contraction with respect to $\|\cdot\|_\xi$ for all $\gamma \in (0,1)$.

Proof: A slight modification of the argument of the proof of Prop. 1 shows that the condition $|A| \leq P$, implies that A is nonexpansive with respect to the norm $\|\cdot\|_\xi$. Furthermore, since $I - \Pi A$ is invertible, we have $z \neq \Pi A z$ for all $z \neq 0$. Hence for all $\gamma \in (0,1)$,

$$\|(1-\gamma)z + \gamma \Pi A z\|_\xi < (1-\gamma)\|z\|_\xi + \gamma\|\Pi A z\|_\xi \leq (1-\gamma)\|z\|_\xi + \gamma\|z\|_\xi = \|z\|_\xi, \quad \forall\ z \in \Re^n, \tag{3.3}$$

where the strict inequality follows from the strict convexity of the norm, and the weak inequality follows from the nonexpansiveness of ΠA. If we define

$$\rho_\gamma = \sup\big\{\|(1-\gamma)z + \gamma\Pi A z\|_\xi \mid \|z\| \leq 1\big\},$$

and note that the supremum above is attained by Weierstrass' Theorem, we see that Eq. (3.3) yields $\rho_\gamma < 1$ and

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

From the definition of T_γ, we have for all $x, y \in \Re^n$,

$$\Pi T_\gamma x - \Pi T_\gamma y = \Pi T_\gamma (x-y) = (1-\gamma)\Pi(x-y) + \gamma \Pi A(x-y) = (1-\gamma)\Pi(x-y) + \gamma\Pi\big(\Pi A(x-y)\big),$$

so defining $z = x - y$, and using the preceding two relations and the nonexpansiveness of Π, we obtain

$$\|\Pi T_\gamma x - \Pi T_\gamma y\|_\xi = \|(1-\gamma)\Pi z + \gamma\Pi(\Pi A z)\|_\xi \leq \|(1-\gamma)z + \gamma\Pi A z\|_\xi \leq \rho_\gamma\|z\|_\xi = \rho_\gamma\|x-y\|_\xi,$$

for all $x, y \in \Re^n$. **Q.E.D.**

Note that the mappings ΠT_γ and ΠT have the same fixed points, so under the assumptions of Prop. 3, there is a unique fixed point Φr^* of ΠT. However, if T has a nonempty linear manifold of fixed points, there arises the question of how close Φr^* is to this manifold. It may be possible to address this issue in specialized contexts; in particular, it has been addressed in [TsV99a] in the context of average cost DP problems (cf. the subsequent Example 3).

We now discuss examples of choices of ξ and P in some interesting special cases.

Example 1: (Discounted DP)

As mentioned in Section 1, Bellman's equation for the cost vector of a stationary policy in an n-state discounted Markovian decision problem has the form

$$x = \alpha P x + g,$$

where g is the vector of single-stage costs associated with the n states, P is the transition probability matrix of the associated Markov chain, and $\alpha \in (0, 1)$ is the discount factor. If P is an irreducible Markov chain, and ξ is chosen to be its unique invariant distribution, the equation approximation method based on Eq. (2.5) yields a popular policy evaluation method known as LSTD(0) (see the references given in Section 1). It generates a single infinitely long trajectory $\{i_0, i_1, \ldots\}$ of the Markov chain, and at time t, it solves the equation

$$\sum_{k=0}^{t} \phi(i_k)\big(\phi(i_k) - \alpha\phi(i_{k+1})\big)' r = \sum_{k=0}^{t} \phi(i_k) g_{i_k}. \tag{3.4}$$

The solution r_t converges to r^* with probability 1. Furthermore, since condition (1) of Prop. 1 is satisfied, it follows that ΠT is a contraction with respect to $\|\cdot\|_\xi$, the error bound (1.6) holds, and the Jacobi/fixed point method (1.7) applies. These results are well-known in the approximate DP literature (see the references given in Section 1).

Example 2: (Stochastic Shortest Path Problems)

Consider the equation $x = Ax + b$, for the case where A is a substochastic matrix ($a_{ij} \geq 0$ for all i, j and $\sum_{j=1}^{n} a_{ij} \leq 1$ for all i). This is Bellman's equation for the cost vector of a stationary policy of a DP problem of the stochastic shortest path type, involving a Markov chain with states $1, \ldots, n$, plus an additional absorbing state denoted 0 (see e.g., [Ber07]). Here a_{ij} is the transition probability from state $i \neq 0$ to state $j \neq 0$, and $1 - \sum_{j=1}^{n} a_{ij}$ is the transition probability from state $i \neq 0$ to the absorbing state 0. If the stationary policy is proper in the sense that from any state $i \neq 0$ there exists a path of positive probability transitions from i to the absorbing state 0, the matrix $P = |A| + (e - |A|e)q'$ [cf. Eq. (3.2)] is irreducible, provided q has positive components. As a result, the conditions of Prop. 1 under condition (2) are satisfied, and T and ΠT are contractions with respect to $\|\cdot\|_\xi$. It is also possible to use a restart probability vector q whose components are not all positive, as long as P is irreducible, in which case Prop. 2 applies.

Example 3: (Average Cost DP Problems)

Consider the equation

$$x = Px + b,$$

where P is an irreducible transition probability matrix, with invariant distribution ξ. This is related to Bellman's equation for the differential cost vector of a stationary policy of an average cost DP problem involving a Markov chain with transition probability matrix P. Then, if the unit vector e is not contained in the subspace S spanned by the basis functions, it can be shown that the matrix $I - \Pi P$ is invertible (see Tsitsiklis and Van Roy [TsV99a], who also give a related error bound). As a result, Prop. 3 applies and shows that the mapping $(1-\gamma)I + \gamma P$, is a contraction with respect to $\|\cdot\|_\xi$ for all $\gamma \in (0, 1)$. The corresponding equation approximation approach and approximate Jacobi method are discussed in [YuB06].

Example 4: (Weakly Diagonally Dominant Systems)

Consider the solution of the system

$$Cx = d,$$

where $d \in \Re^n$ and C is an $n \times n$ matrix that is weakly diagonally dominant, i.e., its components satisfy

$$c_{ii} \neq 0, \qquad \sum_{j \neq i} |c_{ij}| \leq |c_{ii}|, \qquad i = 1, \ldots, n. \tag{3.5}$$

By dividing the ith row by c_{ii}, we obtain the equivalent system $x = Ax + b$, where the components of A and b are

$$a_{ij} = \begin{cases} 0 & \text{if } i = j, \\ -\frac{c_{ij}}{c_{ii}} & \text{if } i \neq j, \end{cases} \qquad b_i = \frac{d_i}{c_{ii}}, \qquad i = 1, \ldots, n.$$

Then, from Eq. (3.5), we have

$$\sum_{j=1}^{n} |a_{ij}| = \sum_{j \neq i} \frac{|c_{ij}|}{|c_{ii}|} \leq 1, \qquad i = 1, \ldots, n,$$

so Props. 1-3 may be used under the appropriate conditions. In particular, if the matrix P given by Eq. (3.2) has no transient states and there exists an index $\bar{i}$ such that $\sum_{j=1}^{n} |a_{\bar{i}j}| < 1$, Prop. 1 applies and shows that ΠT is a contraction with respect to $\| \cdot \|_\xi$.

Alternatively, instead of Eq. (3.5), assume the somewhat more restrictive condition

$$|1 - c_{ii}| + \sum_{j \neq i} |c_{ij}| \leq 1, \qquad i = 1, \ldots, n, \tag{3.6}$$

and consider the equivalent system $x = Ax + b$, where

$$A = I - C, \qquad b = d.$$

Then, from Eq. (3.6), we have

$$\sum_{j=1}^{n} |a_{ij}| = |1 - c_{ii}| + \sum_{j \neq i} |c_{ij}| \leq 1, \qquad i = 1, \ldots, n,$$

so again Prop. 1-3 apply under appropriate conditions.

Example 5: (Discretized Poisson's Equation)

Diagonally dominant linear systems arise in many contexts, including discretized partial differential equations, finite element methods, and economics applications. As an example, consider a discretized version of Poisson's equation over a two-dimensional square grid of N^2 points with fixed boundary conditions, which has the form

$$x_{i,j} = \frac{1}{4}(x_{i+1,j} + x_{i-1,j} + x_{i,j+1} + x_{i,j-1}) + g_{i,j}, \qquad i, j = 1, \ldots, N,$$

where $g_{i,j}$ are given scalars, and by convention $x_{N+1,j} = x_{0,j} = x_{i,N+1} = x_{i,0} = 0$. A subset of the points (i, j) in the square grid are "boundary points," where $x_{i,j}$ is fixed and given. The problem is to compute the values $x_{i,j}$ at the remaining points, which are referred to as "interior points." Thus, we have one equation for each interior grid point. Clearly, this is a special case of Examples 2 and 4, with the row components of A corresponding to (i, j) being $1/4$ for each neighboring interior point of (i, j), and 0 otherwise. If from any interior point it is possible to arrive to some boundary point through a path of adjacent interior points, then clearly based on the graph-geometric structure of the problem, one can construct an irreducible P satisfying $|A| \leq P$.

We finally address a question that was left largely unresolved by the preceding analysis and examples. Given a matrix A with $\sum_{j=1}^n |a_{ij}| \leq 1$ for all i, is it possible to find P such that $|A| \leq P$ and the corresponding Markov chain has no transient states, or even more, is irreducible? Clearly, there are cases where every P with $|A| \leq P$ is irreducible (e.g., when $|A|$ is itself irreducible), and there are cases where there is no such P (e.g., when $|A|$ is the transition probability matrix of a chain with some transient states). To analyze this issue, consider the set

$$\bar{I} = \left\{ i \;\middle|\; \sum_{j=1}^n |a_{ij}| < 1 \right\},$$

and assume that $\bar{I}$ is nonempty (otherwise we must have $P = |A|$). Let $\tilde{I}$ be the set of i such that there exists a sequence of nonzero components $a_{ij_1}, a_{j_1j_2}, \ldots, a_{j_m\bar{i}}$ such that $\bar{i} \in \bar{I}$, and let $\hat{I} = \{i \mid i \notin \bar{I} \cup \tilde{I}\}$ (we allow here the possibility that $\tilde{I}$ or $\hat{I}$ may be empty). Note that the square submatrix of $|A|$ corresponding to $\hat{I}$ is a transition probability matrix, and that we have $a_{ij} = 0$ for all $i \in \hat{I}$ and $j \notin \hat{I}$.

We can now choose $P = |A| + (e - |A|e)q'$ with $q_j > 0$ if and only if $j \in \bar{I} \cup \tilde{I}$ [cf. Eq. (3.2)] to create a positive transition probability from every $i \in \bar{I}$ to every $j \in \bar{I} \cup \tilde{I}$. In this way, the submatrix corresponding to $\bar{I} \cup \tilde{I}$ will be an irreducible transition probability matrix, while any states that are transient in the Markov chain corresponding to $\hat{I}$ will remain transient after the modification. It can thus be seen that *there exists P with $|A| \leq P$ and no transient states if and only if the Markov chain corresponding to $\hat{I}$ has no transient states*. Furthermore, *there exists an irreducible P with $|A| \leq P$ if and only if $\hat{I}$ is empty*. Note that the preceding construction of P requires explicit knowledge of $\hat{I}$, which may be unavailable. Furthermore, when $\hat{I}$ is nonempty, such a P will have multiple recurrent classes, even if it can be found. In this case, multiple simulation trajectories are necessary, with at least one starting within each recurrent class, as discussed in Section 2. This underscores the importance of being able to use an irreducible P.

4. APPROXIMATE JACOBI METHODS

We will now focus on the iteration

$$\Phi r_{t+1} = \Pi T(\Phi r_t), \qquad t = 0, 1, \ldots, \tag{4.1}$$

[cf. Eq. (1.7)], which we refer to as *projected Jacobi* method (PJ for short). We assume throughout this section that ΠT *is a contraction with respect to some norm*, and note that Props. 1-3 provide tools for verifying that this is so. Our simulation-based approximation to the PJ iteration descends from the LSPE methods of approximate DP, referred to in Section 1.

By expressing the projection as a least squares minimization, we can write the PJ iteration (4.1) as

$$r_{t+1} = \arg\min_{r \in \Re^s} \big\| \Phi r - T(\Phi r_t) \big\|_\xi^2,$$

or equivalently

$$r_{t+1} = \arg\min_{r \in \Re^s} \sum_{i=1}^n \xi_i \left(\phi(i)'r - \sum_{j=1}^n a_{ij}\phi(j)'r_t - b_i \right)^2. \tag{4.2}$$

By setting the gradient of the cost function above to 0 and using a straightforward calculation, we have

$$r_{t+1} = \left(\sum_{i=1}^{n} \xi_i \, \phi(i)\phi(i)' \right)^{-1} \sum_{i=1}^{n} \xi_i \, \phi(i) \left(\sum_{j=1}^{n} a_{ij}\phi(j)' r_t + b_i \right). \tag{4.3}$$

Similar to the equation approximation methods of Section 2, we observe that this iteration involves two expected values with respect to the distribution ξ, and we approximate these expected values by sample averages. Thus, we approximate iteration (4.3) with

$$r_{t+1} = \left(\sum_{k=0}^{t} \phi(i_k)\phi(i_k)' \right)^{-1} \sum_{k=0}^{t} \phi(i_k) \left(\frac{a_{i_k j_k}}{p_{i_k j_k}} \phi(j_k)' r_t + b_{i_k} \right), \tag{4.4}$$

which we refer to as the *approximate projected Jacobi* method (APJ for short). Here again $\{i_0, i_1, \ldots\}$ is a state sequence, and $\{(i_0, j_0), (i_1, j_1), \ldots\}$ is a transition sequence satisfying Eqs. (2.3) and (2.4) with probability 1

As in Section 2, we connect the PJ iteration (4.3) with the APJ iteration (4.4) by viewing

$$\phi(i_k)\phi(i_k)' \quad \text{as a sample whose steady-state expected value is the matrix} \quad \sum_{i=1}^{n} \xi_i \phi(i)\phi(i)',$$

$$\frac{a_{i_k j_k}}{p_{i_k j_k}} \phi(i_k)\phi(j_k)' \quad \text{as a sample whose steady-state expected value is the matrix} \quad \sum_{i=1}^{n} \xi_i \phi(i) \sum_{j=1}^{n} a_{ij}\phi(j)',$$

$$\phi(i_k) b_{i_k} \quad \text{as a sample whose steady-state expected value is the vector} \quad \sum_{i=1}^{n} \xi_i \phi(i) b_i.$$

In particular, we write Eq. (4.4) as

$$r_{t+1} = \left(\sum_{i=1}^{n} \hat{\xi}_{i,t} \, \phi(i)\phi(i)' \right)^{-1} \sum_{i=1}^{n} \hat{\xi}_{i,t} \, \phi(i) \left(\sum_{j=1}^{n} \hat{p}_{ij,t} \frac{a_{ij}}{p_{ij}} \phi(j)' r_t + b_i \right), \tag{4.5}$$

where $\hat{\xi}_{i,t}$ and $\hat{p}_{ij,t}$ are defined by

$$\hat{\xi}_{i,t} = \frac{\sum_{k=0}^{t} \delta(i_k = i)}{t+1}, \qquad \hat{p}_{ij,t} = \frac{\sum_{k=0}^{t} \delta(i_k = i, j_k = j)}{\sum_{k=0}^{t} \delta(i_k = i)}.$$

We then note that by Eqs. (2.3) and (2.4), $\hat{\xi}_{i,t}$ and $\hat{p}_{ij,t}$ converge (with probability 1) to ξ_i and p_{ij}, respectively, so by comparing Eqs. (4.3) and (4.5), we see that they asymptotically coincide. Since Eq. (4.3) is a contracting fixed point iteration that converges to r^*, it follows with a simple argument that the same is true for iteration (4.5) (with probability 1).

To streamline and efficiently implement the APJ iteration (4.4), we introduce the matrices

$$B_t = \sum_{k=0}^{t} \phi(i_k)\phi(i_k)', \qquad C_t = \sum_{k=0}^{t} \phi(i_k) \left(\frac{a_{i_k j_k}}{p_{i_k j_k}} \phi(j_k) - \phi(i_k) \right)',$$

and the vector

$$d_t = \sum_{k=0}^{t} \phi(i_k) b_{i_k}.$$

We then write Eq. (4.4) compactly as

$$r_{t+1} = r_t + B_t^{-1}(C_t r_t + d_t), \tag{4.6}$$

and also note that B_t, C_t, and d_t can be efficiently updated using the formulas

$$B_t = B_{t-1} + \phi(i_t)\phi(i_t)', \qquad C_t = C_{t-1} + \phi(i_t)\left(\frac{a_{i_t j_t}}{p_{i_t j_t}}\phi(j_t) - \phi(i_t)\right)', \tag{4.7}$$

$$d_t = d_{t-1} + \phi(i_t)b_{i_t}. \tag{4.8}$$

Let us also observe that Eq. (2.5), the first equation approximation method of Section 2, can be written compactly as

$$C_t r + d_t = 0. \tag{4.9}$$

We can use this formula to establish a connection between the equation approximation and APJ approaches. In particular, suppose that we truncate the state and transition sequences after t transitions, but continue the APJ iteration with B_t, C_t, and d_t held fixed, i.e., consider the iteration

$$r_{m+1} = r_m + B_t^{-1}(C_t r_m + d_t), \qquad m = t, t+1, \ldots. \tag{4.10}$$

Then, since APJ approximates PJ and ΠT is assumed to be a contraction, it follows that with probability 1, for sufficiently large t, the matrix $I + B_t^{-1}C_t$ will be a contraction and iteration (4.10) will converge, by necessity to the solution $\hat{r}_t$ of Eq. (4.9). The conclusion is that, for large t, *the APJ iteration (4.6) can be viewed as a single/first iteration of the algorithm (4.10) that solves the approximate projected equation (4.9).*

Another issue of interest is the rate of convergence of the difference $r_t - \hat{r}_t$ of the results of the two approaches. An interesting result within the appropriate DP context is that under some natural assumptions, $r_t - \hat{r}_t$ converges to 0 faster than the error differences $r_t - r^*$ and $\hat{r}_t - r^*$ (see [BBN04], [YuB06]). Within the more general context of the present paper, a similar analysis is possible, but is outside the scope of the present paper and will be reported elsewhere.

We finally note a variant of the APJ iteration (4.4) that does not require simulated transitions, similar to iteration (2.9). This variant has the form [cf. Eq. (4.4)]

$$r_{t+1} = \left(\sum_{k=0}^{t}\phi(i_k)\phi(i_k)'\right)^{-1}\sum_{k=0}^{t}\phi(i_k)\left(\sum_{j=1}^{n}a_{i_k j}\phi(j)'r_t + b_{i_k}\right),$$

and may be suitable for sparse problems where only few components of each row of A are nonzero.

5. MULTISTEP VERSIONS

An interesting possibility within our context is to replace T with another mapping that has the same fixed points, such as T^l with $l > 1$, or $T^{(\lambda)}$ given by

$$T^{(\lambda)} = (1-\lambda)\sum_{l=0}^{\infty}\lambda^l T^{l+1},$$

where $\lambda \in (0,1)$. In connection with $T^{(\lambda)}$, we assume throughout this section that the eigenvalues of λA lie strictly within the unit circle, so that the preceding infinite series is convergent. The replacement of T by something like T^l or $T^{(\lambda)}$ is seldom considered in traditional fixed point methods, because either the gain in rate of convergence is offset by increased overhead per iteration, or the implementation becomes cumbersome, or both. However, in the context of our simulation-based methods, this replacement is possible, and in fact has a long history in DP approximation methods, as mentioned in Section 1.

As motivation, note that if T is a contraction, the modulus of contraction may be enhanced through the use of T^l or $T^{(\lambda)}$. In particular, if $\alpha \in [0,1)$ is the modulus of contraction of T, the modulus of contraction of T^l is α^l, while the modulus of contraction of $T^{(\lambda)}$ is

$$\alpha^{(\lambda)} = (1-\lambda)\sum_{k=0}^{\infty} \lambda^l \alpha^{l+1} = \frac{\alpha(1-\lambda)}{1-\alpha\lambda}.$$

Furthermore, for all $\lambda \in (0,1)$, $\alpha^l < \alpha$ and $\alpha^{(\lambda)} < \alpha$. Thus the error bounds (1.6) or (1.5) are enhanced. Moreover, there are circumstances where $T^{(\lambda)}$ is a contraction, while T is not, as we will demonstrate shortly (see the following Prop. 4).

To gain some understanding into the properties of $T^{(\lambda)}$, let us write it as

$$T^{(\lambda)}(x) = A^{(\lambda)}x + b^{(\lambda)},$$

where from the equations $T^{l+1}(x) = A^{l+1}x + \sum_{m=0}^{l} A^m b$ and $T^{(\lambda)} = (1-\lambda)\sum_{l=0}^{\infty} \lambda^l T^{l+1}$, we have

$$A^{(\lambda)} = (1-\lambda)\sum_{l=0}^{\infty} \lambda^l A^{l+1}, \qquad b^{(\lambda)} = (1-\lambda)\sum_{l=0}^{\infty} \lambda^l \sum_{m=0}^{l} A^m b = (1-\lambda)\sum_{l=0}^{\infty} A^l b \sum_{m=l}^{\infty} \lambda^m = \sum_{l=0}^{\infty} \lambda^l A^l b. \quad (5.1)$$

The following proposition provides some interesting properties of $A^{(\lambda)}$, which in turn determine contraction and other properties of $T^{(\lambda)}$. We denote by $a_{ij}^{(\lambda)}$ the components of $A^{(\lambda)}$, and by $\sigma(A)$ and $\sigma\big(A^{(\lambda)}\big)$ the spectral radius of A and $A^{(\lambda)}$, respectively. Note that $\sigma(M) \leq \|M\|$, where for any $n \times n$ matrix M and norm $\|\cdot\|$ of $\Re^n$, we denote by $\|M\|$ the corresponding matrix norm of M: $\|M\| = \max_{\|z\|\leq 1} \|Mz\|$. Note also that for a transition probability matrix P, having as an invariant distribution a vector ξ with positive components, we have $\sigma(P) = \|P\|_\xi = 1$. This is well-known, and can be shown with a simple modification of the proof of Prop. 1.

Proposition 4: Assume that $I - A$ is invertible and $\sigma(A) \leq 1$.

(a) For $\lambda \in (0,1)$, $\sigma(A^{(\lambda)})$ decreases monotonically as λ increases, and we have $\sigma(A^{(\lambda)}) < 1$ and $\lim_{\lambda\to 1} \sigma\big(A^{(\lambda)}\big) = 0$.

(b) Assume further that $|A| \leq P$, where P is a transition probability matrix with invariant distribution a vector ξ with positive components. Then,

$$|A^{(\lambda)}| \leq P^{(\lambda)}, \qquad \|A^{(\lambda)}\|_\xi \leq \|P^{(\lambda)}\|_\xi = 1, \qquad \forall\, \lambda \in [0,1),$$

where $P^{(\lambda)} = (1-\lambda)\sum_{l=0}^{\infty} \lambda^l P^{l+1}$. Furthermore, for all $\lambda \in (0,1)$ the eigenvalues of $\Pi A^{(\lambda)}$ lie strictly within the unit circle, where Π denotes projection on S with respect to $\|\cdot\|_\xi$.

Proof: (a) From Eq. (5.1), we see that the eigenvalues of $A^{(\lambda)}$ have the form

$$(1-\lambda)\sum_{l=0}^{\infty}\lambda^l\beta^{l+1}=\frac{\beta(1-\lambda)}{1-\beta\lambda},$$

where β is an eigenvalue of A. To show that $\sigma(A^{(\lambda)})<1$ we must show that $|\beta|(1-\lambda)<|1-\beta\lambda|$, for all complex numbers β with $|\beta|\leq 1$ but $\beta\neq 1$. This can be proved with a triangle geometry argument on the two-dimensional circle; see Fig. 5.1 (an algebraic proof is also quite straightforward). The monotonic decrease of $\sigma(A^{(\lambda)})$ and the fact $\lim_{\lambda\to 1}\sigma\big(A^{(\lambda)}\big)=0$ are also evident from the figure.

(b) To see that $|A^{(\lambda)}|\leq P^{(\lambda)}$, note that for all $l>1$, the components of $|A|^l$ are no greater than the corresponding components of P^l, since they can be written as products of corresponding components of $|A|$ and P, and by assumption, we have $|a_{ij}|\leq p_{ij}$ for all $i,j=1,\ldots,n$. We have $\|P^{(\lambda)}\|_\xi=1$ because $P^{(\lambda)}$ is a transition probability matrix and ξ is an invariant distribution of $P^{(\lambda)}$. The inequality $\|A^{(\lambda)}\|_\xi\leq\|P^{(\lambda)}\|_\xi$ follows by a simple modification of the proof of Prop. 1.

Since $\|A^{(\lambda)}\|_\xi\leq\|P^{(\lambda)}\|_\xi=1$ and Π is nonexpansive with respect to $\|\cdot\|_\xi$, it follows that $\|\Pi A^{(\lambda)}\|_\xi\leq 1$, so all eigenvalues of $\Pi A^{(\lambda)}$ lie within the unit circle. Furthermore, by Lemma 1 of [YuB06], all eigenvalues ν of $\Pi A^{(\lambda)}$ with $|\nu|=1$ must also be eigenvalues of $A^{(\lambda)}$. Since by part (a) we have $\sigma(A^{(\lambda)})<1$ for $\lambda>0$, there are no such eigenvalues. **Q.E.D.**

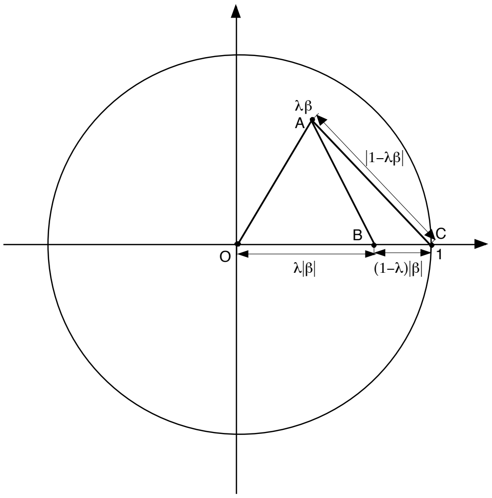

Figure 5.1. Proof of the inequality $(1-\lambda)|\beta|<|1-\beta\lambda|$. We consider the complex plane, an eigenvalue β of A, and the complex number $\lambda\beta$. We note that $|\beta|\leq 1$, $\beta\neq 1$, and $\lambda|\beta|<1$ by assumption. If $\beta=0$, the desired inequality holds, and if β is in the left-hand side of the plane or the vertical axis, we have

$$(1-\lambda)|\beta|<1\leq|1-\beta\lambda|,$$

so it is sufficient to consider the case where $\beta\neq 0$ and the real part of β is nonnegative, as shown in the figure. We consider the isoskeles triangle OAB, and note that the angles bordering the side AB are less than 90 degrees. It follows that the angle ABC is greater than 90 degrees, so the side AC of the triangle ABC is strictly larger than the side BC. This is equivalent to the desired result.

Note from Prop. 4(a) that $T^{(\lambda)}$ is a contraction for $\lambda>0$ even if T is not, provided that $I-A$ is invertible and $\sigma(A) \leq 1$. This is not true for T^l, $l > 1$, since the eigenvalues of A^l are β^l where β is an eigenvalue of A, so that $\sigma(A^l) = 1$ if $\sigma(A) = 1$.† Furthermore, under the assumptions of Prop. 4(b), $\Pi T^{(\lambda)}$ is a contraction for $\lambda > 0$. This suggests an advantage for using $\lambda > 0$, and the fact $\lim_{\lambda \to 1} \sigma\big(A^{(\lambda)}\big) = 0$ also suggests an advantage for using λ close to 1. However, as we will discuss later, there is also a disadvantage in our simulation-based methods for using λ close to 1, because of increased simulation noise.

We will now develop the formulas for various multistep methods. The detailed verification of these formulas is somewhat tedious, and will only be sketched. The key idea is that the ith component $(A^m g)(i)$ of a vector of the form $A^m g$, where $g \in \Re^n$, can be computed by generating a sequence of states $\{i_0, i_1, \ldots\}$ of a Markov chain with invariant distribution ξ and transition probabilities p_{ij}, and by forming the average of $w_{k,m} g_{i_{k+m}}$ over all indices k such that $i_k = i$, where

$$w_{k,m} = \begin{cases} \frac{a_{i_k i_{k+1}}}{p_{i_k i_{k+1}}} \frac{a_{i_{k+1} i_{k+2}}}{p_{i_{k+1} i_{k+2}}} \cdots \frac{a_{i_{k+m-1} i_{k+m}}}{p_{i_{k+m-1} i_{k+m}}} & \text{if } m \geq 1, \\ 1 & \text{if } m = 0. \end{cases} \tag{5.2}$$

In short

$$(A^m g)(i) \approx \frac{\sum_{k=0}^{t} \delta(i_k = i) w_{k,m} g_{i_{k+m}}}{\sum_{k=0}^{t} \delta(i_k = i)}. \tag{5.3}$$

The justification is that, assuming

$$\lim_{t \to \infty} \frac{\sum_{k=0}^{t} \delta(i_k = i, i_{k+1} = j_1, \ldots, i_{k+m} = j_m)}{\sum_{k=0}^{t} \delta(i_k = i)} = p_{i j_1} p_{j_1 j_2} \cdots p_{j_{m-1} j_m}, \tag{5.4}$$

the limit of the right-hand side of Eq. (5.3) can be written as

$$\begin{aligned} \lim_{t \to \infty} \frac{\sum_{k=0}^{t} \delta(i_k = i) w_{k,m} g_{i_{k+m}}}{\sum_{k=0}^{t} \delta(i_k = i)} &= \lim_{t \to \infty} \frac{\sum_{k=0}^{t} \sum_{j_1=1}^{n} \cdots \sum_{j_m=1}^{n} \delta(i_k = i, i_{k+1} = j_1, \ldots, i_{k+m} = j_m) w_{k,m} g_{i_{k+m}}}{\sum_{k=0}^{t} \delta(i_k = i)} \\ &= \sum_{j_1=1}^{n} \cdots \sum_{j_m=1}^{n} \lim_{t \to \infty} \frac{\sum_{k=0}^{t} \delta(i_k = i, i_{k+1} = j_1, \ldots, i_{k+m} = j_m)}{\sum_{k=0}^{t} \delta(i_k = i)} w_{k,m} g_{i_{k+m}} \\ &= \sum_{j_1=1}^{n} \cdots \sum_{j_m=1}^{n} a_{i j_1} a_{j_1 j_2} \cdots a_{j_{m-1} j_m} g_{j_m} \\ &= (A^m g)(i), \end{aligned}$$

where the third equality follows using Eqs. (5.2) and (5.4). By using the approximation formula (5.3), it is possible to construct complex simulation-based approximations to formulas that involve powers of A. The subsequent multistep methods in this section and the basis construction methods of Section 6 rely on this idea.

l-Step Methods

Let us now develop simulation methods based on the equation $\Phi r = \Pi T^l(\Phi r)$, corresponding to T^l with $l > 1$. Consider the projected Jacobi iteration

$$\Phi r_{t+1} = \Pi T^l(\Phi r_t) = \Pi \left(A^l \Phi r_t + \sum_{m=0}^{l-1} A^m b \right).$$

† What is happening here is that individual powers of eigenvalues of A may lie on the unit circle, but taking a mixture/linear combination of many different powers of an eigenvalue β of A to form the corresponding eigenvalue $(1 - \lambda) \sum_{l=0}^{\infty} \lambda^l \beta^{l+1}$ of $A^{(\lambda)}$, results in a complex number that lies in the interior of the unit circle.

Equivalently,

$$
\begin{aligned}
r_{t+1} &= \arg\min_r \sum_{i=1}^n \xi_i \left(\phi(i)'r - \big(T^l(\Phi r_t)\big)(i)\right)^2 \\
&= \arg\min_r \sum_{i=1}^n \xi_i \left(\phi(i)'r - (A^l\Phi)(i)r_t - \left(\sum_{m=0}^{l-1} A^m b\right)(i)\right)^2,
\end{aligned}
$$

where $(A^l\Phi)(i)$ is the ith row of the matrix $A^l\Phi$, and $\big(T^l(\Phi r_t)\big)(i)$ and $\left(\sum_{m=0}^{l-1} A^m b\right)(i)$ are the ith components of the vectors $T^l(\Phi r_t)$ and $\sum_{m=0}^{l-1} A^m b$, respectively. By solving for the minimum over r, we finally obtain

$$
r_{t+1} = \left(\sum_{i=1}^n \xi_i \, \phi(i)\phi(i)'\right)^{-1} \sum_{i=1}^n \xi_i \, \phi(i) \left((A^l\Phi)(i)r_t + \left(\sum_{m=0}^{l-1} A^m b\right)(i)\right). \tag{5.5}
$$

We propose the following approximation to this iteration:

$$
r_{t+1} = \left(\sum_{k=0}^t \phi(i_k)\phi(i_k)'\right)^{-1} \sum_{k=0}^t \phi(i_k) \left(w_{k,l}\phi(i_{k+l})'r_t + \sum_{m=0}^{l-1} w_{k,m} b_{i_{k+m}}\right) \tag{5.6}
$$

where $w_{k,m}$ is given by Eq. (5.2). Its validity is based on Eq. (5.3), and on the fact that each index i is sampled with probability ξ_i, and each transition (i,j) is generated with probability p_{ij}, as part of the infinitely long sequence of states $\{i_0, i_1, \ldots\}$ of a Markov chain whose invariant distribution is ξ and transition probabilities are p_{ij}. In particular, similar to Section 4, we view

$$
\phi(i_k)\phi(i_k)' \quad \text{as a sample whose steady-state expected value is the matrix} \quad \sum_{i=1}^n \xi_i \phi(i)\phi(i)',
$$

$$
w_{k,l}\phi(i_k)\phi(i_{k+l})' \quad \text{as a sample whose steady-state expected value is the matrix} \quad \sum_{i=1}^n \xi_i \phi(i)(A^l\Phi)(i),
$$

$$
w_{k,m}\phi(i_k)b_{i_{k+m}} \quad \text{as a sample whose steady-state expected value is the vector} \quad \sum_{i=1}^n \xi_i \phi(i)(A^m b)(i).
$$

As in Section 4, we can express the l-step APJ iteration (5.6) in compact form. In particular, we can write Eq. (5.6) as

$$
r_{t+1} = r_t + B_t^{-1}(C_t r_t + h_t), \tag{5.7}
$$

where

$$
C_t = \sum_{k=0}^t \phi(i_k)\big(w_{k,l}\phi(i_{k+l}) - \phi(i_k)\big)', \qquad B_t = \sum_{k=0}^t \phi(i_k)\phi(i_k)',
$$

$$
h_t = \sum_{k=0}^t \phi(i_k) \sum_{m=0}^{l-1} w_{k,m} b_{i_{k+m}}.
$$

Note that to calculate C_t and h_t, it is necessary to generate the future l states $i_{t+1}, \ldots, i_{t+l}$. Note also that C_t, B_t, and h_t can be efficiently updated via

$$
C_t = C_{t-1} + \phi(t_t)\big(w_{t,l}\phi(i_{t+l}) - \phi(i_t)\big)', \qquad B_t = B_{t-1} + \phi(i_t)\phi(i_t)',
$$

$$
h_t = h_{t-1} + \phi(i_t)z_t,
$$

where z_t is given by

$$z_t = \sum_{m=0}^{l-1} w_{t,m} b_{i_{t+m}},$$

and can be updated by

$$z_t = \frac{z_{t-1} - b_{i_{t-1}}}{w_{t-1,1}} + w_{t,l-1} b_{i_{t+l-1}}.$$

We can also write the l-step APJ iteration (5.6) in an alternative form by using the scalars

$$d_t(i_{k+m}) = b_{i_{k+m}} + \frac{a_{i_{k+m} i_{k+m+1}}}{p_{i_{k+m} i_{k+m+1}}} \phi(i_{k+m+1})' r_t - \phi(i_{k+m})' r_t, \qquad t \geq 0,\ m \geq 0, \tag{5.8}$$

which we call *temporal differences*, their recognized name within the specialized context of approximate DP (see the references given in Section 1). In particular, we can rewrite the last parenthesized expression in iteration (5.6) in terms of temporal differences, and obtain

$$r_{t+1} = \left(\sum_{k=0}^{t} \phi(i_k)\phi(i_k)' \right)^{-1} \sum_{k=0}^{t} \phi(i_k) \left(\phi(i_k)' r_t + \sum_{m=0}^{l-1} w_{k,m} d_t(i_{k+m}) \right),$$

or

$$r_{t+1} = r_t + \left(\sum_{k=0}^{t} \phi(i_k)\phi(i_k)' \right)^{-1} \sum_{k=0}^{t} \phi(i_k) \sum_{m=0}^{l-1} w_{k,m} d_t(i_{k+m}). \tag{5.9}$$

To see this, note that the sum over m in Eq. (5.6) can be expressed, using the definition (5.8), as

$$\sum_{m=0}^{l-1} w_{k,m} \left(b_{i_{k+m}} + \frac{a_{i_{k+m} i_{k+m+1}}}{p_{i_{k+m} i_{k+m+1}}} \phi(i_{k+m+1})' r_t - \phi(i_{k+m})' r_t \right),$$

which by canceling and collecting terms, using the definition (5.2) of $w_{k,m}$, reduces to

$$w_{k,l} \phi(i_{k+l})' r_t - \phi(i_k)' r_t + \sum_{m=0}^{l-1} w_{k,m} b_{i_{k+m}}.$$

Substituting this expression in Eq. (5.9), we obtain Eq. (5.6). Using the temporal differences formula (5.9) does not seem to result in any advantage over using the formula (5.6). In approximate DP, algorithms are usually stated in terms of temporal differences.

An important observation is that compared to the case $l = 1$ [cf. Eqs. (4.4) and (2.1)], the term $w_{k,l}$ multiplying $\phi(i_{k+l})'$ and the terms $w_{k,m}$ multiplying $b_{i_{k+m}}$ in Eq. (5.6) tend to increase the variance of the samples used in the approximations as l increases. This is a generic tradeoff in the multistep methods of this section: by using equations involving greater dependence on more distant steps (larger values of l or λ) we improve the modulus of contraction, but we degrade the quality of the simulation through greater variance of the associated samples.

The preceding analysis also yields an equation approximation method corresponding to the APJ iteration (5.7). It has the form

$$C_t r + h_t = 0,$$

and we have $\hat{r}_t \rightarrow r^*$ with probability 1, where $\hat{r}_t$ is a solution to this equation.

λ-Methods

We will now develop simulation methods based on the equation $\Phi r = \Pi T^{(\lambda)}(\Phi r)$ for $\lambda \in (0,1)$. We will express these methods using convenient recursive formulas that use temporal differences (this is standard in approximate DP). However, we note that there are several alternative recursive formulas, of nearly equal effectiveness, which do not involve temporal differences. We first express $T^{(\lambda)}$ in terms of temporal difference-like terms:†

$$T^{(\lambda)}(x) = x + \sum_{m=0}^{\infty} \lambda^m (A^m b + A^{m+1}x - A^m x).$$

Using the above expression, we write the projected Jacobi iteration as

$$\Phi r_{t+1} = \Pi T^{(\lambda)}(\Phi r_t) = \Pi \left(\Phi r_t + \sum_{m=0}^{\infty} \lambda^m (A^m b + A^{m+1}\Phi r_t - A^m \Phi r_t) \right),$$

or equivalently

$$r_{t+1} = \arg\min_r \sum_{i=1}^{n} \xi_i \left(\phi(i)'r - \phi(i)'r_t - \sum_{m=0}^{\infty} \lambda^m \big((A^m b)(i) + (A^{m+1}\Phi)(i) r_t - (A^m \Phi)(i) r_t \big) \right)^2,$$

where $(A^k\Phi)(i)$ denotes the ith row of the matrix $A^k\Phi$, and $(A^l b)(i)$ denotes the ith component of the vector $A^l b$, respectively. By solving for the minimum over r, we can write this iteration as

$$r_{t+1} = r_t + \left(\sum_{i=1}^{n} \xi_i \, \phi(i)\phi(i)' \right)^{-1} \sum_{i=1}^{n} \xi_i \, \phi(i) \left(\sum_{m=0}^{\infty} \lambda^m \big((A^m b)(i) + (A^{m+1}\Phi)(i) r_t - (A^m \Phi)(i) r_t \big) \right). \qquad (5.10)$$

We approximate this iteration by

$$r_{t+1} = r_t + \left(\sum_{k=0}^{t} \phi(i_k)\phi(i_k)' \right)^{-1} \sum_{k=0}^{t} \phi(i_k) \sum_{m=k}^{t} \lambda^{m-k} w_{k,m-k} d_t(i_m), \qquad (5.11)$$

where $d_t(i_m)$ are the temporal differences

$$d_t(i_m) = b_{i_m} + \frac{a_{i_m i_{m+1}}}{p_{i_m i_{m+1}}} \phi(i_{m+1})' r_t - \phi(i_m)' r_t, \qquad t \geq 0, \; m \geq 0. \qquad (5.12)$$

† This can be seen from the following calculation [cf. Eq. (5.1)]:

$$\begin{aligned} T^{(\lambda)}(x) &= \sum_{l=0}^{\infty} (1-\lambda)\lambda^l (A^{l+1}x + A^l b + A^{l-1}b + \cdots + b) \\ &= x + (1-\lambda) \sum_{l=0}^{\infty} \lambda^l \sum_{m=0}^{l} (A^m b + A^{m+1}x - A^m x) \\ &= x + (1-\lambda) \sum_{m=0}^{\infty} \left(\sum_{l=m}^{\infty} \lambda^l \right) (A^m b + A^{m+1}x - A^m x) \\ &= x + \sum_{m=0}^{\infty} \lambda^m (A^m b + A^{m+1}x - A^m x). \end{aligned}$$

Similar to earlier cases, the basis for this is to replace the two expected values in the right-hand side of Eq. (5.10) with averages of samples corresponding to the states i_k, $k = 0, 1, \ldots$. In particular, we view

$$\phi(i_k)\phi(i_k)' \quad \text{as a sample whose steady-state expected value is} \quad \sum_{i=1}^{n} \xi_i \phi(i)\phi(i)',$$

$$\phi(i_k)\sum_{m=k}^{t} \lambda^{m-k} w_{k,m-k} d_t(i_m) \quad \text{as a sample whose steady-state expected value is approximately}$$

$$\sum_{i=1}^{n} \xi_i\, \phi(i) \sum_{m=0}^{\infty} \lambda^m \big((A^m b)(i) + (A^{m+1}\Phi)(i) r_t - (A^m\Phi)(i) r_t\big).$$

Note that the summation of the second sample above is truncated at time t, but is a good approximation when k is much smaller than t and also when λ is small. Proofs that $r_t \to r^*$ with probability 1 have been given for special cases arising in approximate DP (see [NeB03], [BBN04], and [YuB06]), but they are beyond the scope of the present paper.

By using the temporal difference formula (5.12), we can write iteration (5.11) in compact form as

$$r_{t+1} = r_t + B_t^{-1}\left(C_t r_t + h_t\right), \tag{5.13}$$

where

$$B_t = \sum_{k=0}^{t} \phi(i_k)\phi(i_k)', \tag{5.14}$$

$$C_t = \sum_{k=0}^{t} \phi(i_k) \sum_{m=k}^{t} \lambda^{m-k} w_{k,m-k} \left(\frac{a_{i_m i_{m+1}}}{p_{i_m i_{m+1}}}\phi(i_{m+1}) - \phi(i_m)\right)', \tag{5.15}$$

$$h_t = \sum_{k=0}^{t} \phi(i_k) \sum_{m=k}^{t} \lambda^{m-k} w_{k,m-k} b_{i_m}. \tag{5.16}$$

We now introduce the auxiliary vector

$$z_k = \sum_{m=0}^{k} \lambda^{k-m} w_{m,k-m} \phi(i_m), \tag{5.17}$$

and we will show that C_t and h_t can be written as

$$C_t = \sum_{k=0}^{t} z_k \left(\frac{a_{i_k i_{k+1}}}{p_{i_k i_{k+1}}}\phi(i_{k+1}) - \phi(i_k)\right)', \tag{5.18}$$

$$h_t = \sum_{k=0}^{t} z_k b_{i_k}. \tag{5.19}$$

Thus, the quantities B_t, C_t, h_t, and z_t can be efficiently updated with the recursive formulas:

$$B_t = B_{t-1} + \phi(i_t)\phi(i_t)', \qquad C_t = C_{t-1} + z_t \left(\frac{a_{i_t i_{t+1}}}{p_{i_t i_{t+1}}}\phi(i_{t+1}) - \phi(i_t)\right)',$$

$$h_t = h_{t-1} + z_t b_{i_t}, \qquad z_t = \lambda \frac{a_{i_{t-1} i_t}}{p_{i_{t-1} i_t}} z_{t-1} + \phi(i_t).$$

Indeed, we write Eq. (5.15) as

$$\begin{aligned}
C_t &= \sum_{k=0}^{t}\sum_{m=k}^{t} \lambda^{m-k} w_{k,m-k}\phi(i_k)\left(\frac{a_{i_m i_{m+1}}}{p_{i_m i_{m+1}}}\phi(i_{m+1})-\phi(i_m)\right)' \\
&= \sum_{m=0}^{t}\sum_{k=0}^{m} \lambda^{m-k} w_{k,m-k}\phi(i_k)\left(\frac{a_{i_m i_{m+1}}}{p_{i_m i_{m+1}}}\phi(i_{m+1})-\phi(i_m)\right)' \\
&= \sum_{k=0}^{t}\sum_{m=0}^{k} \lambda^{k-m} w_{m,k-m}\phi(i_m)\left(\frac{a_{i_k i_{k+1}}}{p_{i_k i_{k+1}}}\phi(i_{k+1})-\phi(i_k)\right)' \\
&= \sum_{k=0}^{t} z_k\left(\frac{a_{i_k i_{k+1}}}{p_{i_k i_{k+1}}}\phi(i_{k+1})-\phi(i_k)\right)',
\end{aligned}$$

thus proving Eq. (5.18). Similarly,

$$\begin{aligned}
h_t &= \sum_{k=0}^{t}\sum_{m=k}^{t} \lambda^{m-k} w_{k,m-k}\phi(i_k)b_{i_m} \\
&= \sum_{m=0}^{t}\sum_{k=0}^{m} \lambda^{m-k} w_{k,m-k}\phi(i_k)b_{i_m} \\
&= \sum_{k=0}^{t}\sum_{m=0}^{k} \lambda^{k-m} w_{m,k-m}\phi(i_m)b_{i_k} \\
&= \sum_{k=0}^{t} z_k b_{i_k},
\end{aligned}$$

thus proving Eq. (5.19).

We finally note that an equation approximation method corresponding to the APJ iteration (5.13) has the form

$$C_t r + h_t = 0.$$

The convergence $\hat{r}_t \to r^*$, where $\hat{r}_t$ is the solution to this equation, has been shown for special cases in approximate DP (see [NeB03]). A convergence proof for the general case is again beyond the scope of the present paper.

A Generalization of TD(λ)

Finally, let us indicate a generalized version of the TD(λ) method of approximate DP (Sutton [Sut88]). It has the form

$$r_{t+1} = r_t + \gamma_t z_t d_t(i_t), \tag{5.20}$$

where γ_t is a diminishing positive scalar stepsize, z_t is given by Eq. (5.17), and $d_t(i_t)$ is the temporal difference given by Eq. (5.12). The analysis of TD(λ) that is most relevant to our work is the one by Tsitsiklis and Van Roy [TsV97]. Much of this analysis generalizes easily. In particular, the essence of the convergence proof of [TsV97] is to write the algorithm as

$$r_{t+1} = r_t + \gamma_t(Cr_t + h) + \gamma_t(V_t r_t + v_t), \qquad t = 0, 1, \ldots,$$

where

$$C = \Phi'\Xi(A^{(\lambda)} - I)\Phi, \tag{5.21}$$

h is a some vector in $\Re^s$, Ξ is the diagonal matrix having ξ_i, $i = 1, \ldots, n$, on the diagonal, and V_t and v_t are random matrices and vectors, respectively, which asymptotically have zero mean. The essence of the convergence proof of Tsitsiklis and Van Roy is that the matrix C is negative definite, in the sense that $r'Cr < 0$ for all $r \neq 0$, so it has eigenvalues with negative real parts, which implies in turn that the matrix $I + \gamma_t C$ has eigenvalues strictly within the unit circle for sufficiently small γ_t. The following is a generalization of this key fact (Lemma 9 of [TsV97]).

Proposition 5: For all $\lambda \in [0, 1)$, if $\Pi T^{(\lambda)}$ is a contraction on S with respect to $\|\cdot\|_\xi$, then the matrix C of Eq. (5.21) is negative definite.

Proof: By the contraction assumption, we have for some $\alpha \in [0, 1)$,

$$\|\Pi A^{(\lambda)}\Phi r\|_\xi \leq \alpha \|\Phi r\|_\xi, \qquad \forall\, r \in \Re^s. \tag{5.22}$$

Also, Π is given in matrix form as $\Pi = \Phi(\Phi'\Xi\Phi)^{-1}\Phi'\Xi$, from which it follows that

$$\Phi'\Xi(I - \Pi) = 0. \tag{5.23}$$

Thus, we have for all $r \neq 0$,

$$\begin{aligned} r'Cr &= r'\Phi'\Xi(A^{(\lambda)} - I)\Phi r \\ &= r'\Phi'\Xi\big((I - \Pi)A^{(\lambda)} + \Pi A^{(\lambda)} - I\big)\Phi r \\ &= r'\Phi'\Xi(\Pi A^{(\lambda)} - I)\Phi r \\ &= r'\Phi'\Xi\Pi A^{(\lambda)}\Phi r - \|\Phi r\|_\xi^2 \\ &\leq \|\Phi r\|_\xi \cdot \|\Pi A^{(\lambda)}\Phi r\|_\xi - \|\Phi r\|_\xi^2 \\ &\leq (\alpha - 1)\|\Phi r\|_\xi^2 \\ &< 0, \end{aligned}$$

where the third equality follows from Eq. (5.23), the first inequality follows from the Cauchy-Schwartz inequality applied with the inner product $< x, y >= x'\Xi y$ that corresponds to the norm $\|\cdot\|_\xi$, and the second inequality follows from Eq. (5.22). **Q.E.D.**

The preceding proposition supports the validity of the algorithm (5.20), and provides a starting point for its analysis. However, the details are beyond the scope of the present paper.

6. USING BASIS FUNCTIONS INVOLVING POWERS OF A

We have assumed in the preceding sections that the columns of Φ, the basis functions, are known, and the rows $\phi(i)'$ of Φ are explicitly available to use in the various simulation-based formulas. We will now discuss a class of basis functions that may not be available, but may be approximated by simulation in the course of our algorithms. Let us first consider basis functions of the form $A^m g$, $m \geq 0$, where g is some vector in $\Re^n$. Such basis functions are implicitly used in the context of Krylov subspace methods; see e.g., Saad [Saa03]. A simple justification is that the fixed point of T has an expansion of the form

$$x^* = \sum_{t=0}^{\infty} A^t b,$$

provided the spectral radius of A is less than 1. Thus the basis functions $b, Ab, \ldots, A^s b$ yield an approximation based on the first $s+1$ terms of the expansion. Also a more general expansion is

$$x^* = \bar{x} + \sum_{t=0}^{\infty} A^t q,$$

where $\bar{x}$ is any vector in $\Re^n$ and q is the residual vector

$$q = T(\bar{x}) - \bar{x} = A\bar{x} + b - \bar{x};$$

this can be seen from the equation $x^* - \bar{x} = A(x^* - \bar{x}) + q$. Thus the basis functions $\bar{x}, q, Aq, \ldots, A^{s-1}q$ yield an approximation based on the first $s+1$ terms of the preceding expansion. Note that we have

$$A^m q = T^{m+1}(\bar{x}) - T^m(\bar{x}), \qquad \forall\ m \geq 0,$$

so the subspace spanned by these basis functions is the subspace spanned by $\bar{x}, T(\bar{x}), \ldots, T^s(\bar{x})$.

Generally, to implement the methods of the preceding sections with basis functions of the form $A^m g$, $m \geq 0$, one would need to generate the ith components $(A^m g)(i)$ for any given i, but these may be hard to calculate. However, one can use instead single sample approximations of $(A^m g)(i)$, and rely on the formula

$$(A^m g)(i) \approx \frac{\sum_{k=0}^{t} \delta(i_k = i) w_{k,m} g_{i_{k+m}}}{\sum_{k=0}^{t} \delta(i_k = i)} \tag{6.1}$$

[cf. Eq. (5.3)]. Thus in principle, to approximate the algorithms of earlier sections using such basis functions, we only need to substitute each occurence of $(A^m g)(i)$ in the vector $\phi(i)$ by a sample $w_{k,m} g_{i_{k+m}}$ generated independently of the "main" Markov chain trajectory.

It is possible to use, in addition to $g, Ag, \ldots, A^s g$, other basis functions, whose components are available with no error, or to use several sets of basis functions of the form $g, Ag, \ldots, A^s g$, corresponding to multiple vectors g. However, for simplicity in what follows in this section, we assume that the only basis functions are $g, Ag, \ldots, A^s g$ for a single given vector g, so the matrix Φ has the form

$$\Phi = \big(g \ Ag \ \cdots \ A^s g \big). \tag{6.2}$$

The ith row of Φ is

$$\phi(i)' = \big(g(i) \ (Ag)(i) \ \cdots \ (A^s g)(i) \big).$$

We will focus on a version of the equation approximation method of Section 2, which uses single sample approximations of these rows. The multistep methods of Section 5 admit similar versions, since the corresponding formulas involve powers of A multiplying vectors, which can be approximated using Eq. (6.1).

We recall [cf. Eq. (2.1)] that the projected equation $\Phi r = \Pi(A\Phi r + b)$ has the form

$$\sum_{i=1}^{n} \xi_i \phi(i) \left(\phi(i) - \sum_{j=1}^{n} a_{ij} \phi(j) \right)' r^* = \sum_{i=1}^{n} \xi_i \phi(i) b_i, \tag{6.3}$$

or equivalently, using Eq. (6.2),

$$\sum_{i=1}^{n} \xi_i \begin{pmatrix} g(i) \\ (Ag)(i) \\ \vdots \\ (A^s g)(i) \end{pmatrix} \Big(g(i) - (Ag)(i) \;\; (Ag)(i) - (A^2 g)(i) \;\; \cdots \;\; (A^s g)(i) - (A^{s+1} g)(i) \Big) r^* = \sum_{i=1}^{n} \xi_i \begin{pmatrix} g(i) \\ (Ag)(i) \\ \vdots \\ (A^s g)(i) \end{pmatrix} b_i, \tag{6.4}$$

To approximate this equation, we generate a sequence of states $\{i_0, i_1, \ldots\}$ of the Markov chain such that with probability 1,

$$\lim_{t\to\infty} \frac{\sum_{k=0}^{t} \delta(i_k = i)}{t+1} = \xi_i, \qquad \lim_{t\to\infty} \frac{\sum_{k=0}^{t} \delta(i_k = i, i_{k+1} = j)}{\sum_{k=0}^{t} \delta(i_k = i)} = p_{ij}, \quad i, j = 1, \ldots, n.$$

For each i_k, we also generate two additional mutually "independent" sequences

$$\big\{(i_k, \hat{i}_{k,1}), (\hat{i}_{k,1}, \hat{i}_{k,2}), \ldots, (\hat{i}_{k,s-1}, \hat{i}_{k,s})\big\}, \qquad \big\{(i_k, \tilde{i}_{k,1}), (\tilde{i}_{k,1}, \tilde{i}_{k,2}), \ldots, (\tilde{i}_{k,s}, \tilde{i}_{k,s+1})\big\}, \tag{6.5}$$

of lengths s and $s+1$, respectively, according to the transition probabilities p_{ij}, which are also "independent" of the sequence $\{i_0, i_1, \ldots\}$. At time t, we form the following linear equation to approximate Eq. (6.4):

$$\sum_{k=0}^{t} \begin{pmatrix} g_{i_k} \\ \hat{w}_{k,1} g_{\hat{i}_{k,1}} \\ \vdots \\ \hat{w}_{k,s} g_{\hat{i}_{k,s}} \end{pmatrix} \Big(g_{i_k} - \tilde{w}_{k,1}\; g_{\tilde{i}_{k+1}} \;\; \tilde{w}_{k,1} g_{\tilde{i}_{k+1}} - \tilde{w}_{k,2} g_{\tilde{i}_{k+2}} \;\; \cdots \;\; \tilde{w}_{k,s} g_{\tilde{i}_{k+s}} - \tilde{w}_{k,s+1} g_{\tilde{i}_{k+s+1}} \Big) r = \sum_{k=0}^{t} \begin{pmatrix} g_{i_k} \\ \hat{w}_{k,1} g_{\hat{i}_{k,1}} \\ \vdots \\ \hat{w}_{k,s} g_{\hat{i}_{k,s}} \end{pmatrix} b_{i_k}, \tag{6.6}$$

where for all m,

$$\hat{w}_{k,m} = \frac{a_{i_k \hat{i}_{k,1}}}{p_{i_k \hat{i}_{k,1}}} \frac{a_{\hat{i}_{k,1} \hat{i}_{k,2}}}{p_{\hat{i}_{k,1} \hat{i}_{k,2}}} \cdots \frac{a_{\hat{i}_{k,m-1} \hat{i}_{k,m}}}{p_{\hat{i}_{k,m-1} \hat{i}_{k,m}}}, \qquad \tilde{w}_{k,m} = \frac{a_{i_k \tilde{i}_{k,1}}}{p_{i_k \tilde{i}_{k,1}}} \frac{a_{\tilde{i}_{k,1} \tilde{i}_{k,2}}}{p_{\tilde{i}_{k,1} \tilde{i}_{k,2}}} \cdots \frac{a_{\tilde{i}_{k,m-1} \tilde{i}_{k,m}}}{p_{\tilde{i}_{k,m-1} \tilde{i}_{k,m}}}; \tag{6.7}$$

[cf. Eq. (5.2)].

It is also possible to save some simulation overhead per time step by replacing the finite sequence $\{(i_k, \tilde{i}_{k,1}), (\tilde{i}_{k,1}, \tilde{i}_{k,2}), \ldots, (\tilde{i}_{k,s}, \tilde{i}_{k,s+1})\}$ with the infinite length sequence $\{(i_0, i_1), (i_1, i_2), \ldots\}$, which is used to generate successively the states i_k. However, in this case the corresponding samples $w_{k,m} g_{k+m}$ will generally be correlated, because they correspond to (possibly overlapping) segments from the same trajectory, likely increasing the variance of the simulation error. The device of using an extra independent sequence per time step, may also be applied to reduce the simulation noise of the l-step methods of Section 5 [cf. Eq. (5.6)], but is harder to use in the context of the λ-methods of that section.

To verify the validity of this approximation, we can use Eq. (6.1), and a similar analysis to the one of Section 2. We omit the straightforward details. It is also possible to construct a corresponding approximate Jacobi method along similar lines.

The preceding methodology can be extended in a few different ways. First, the sequences (6.5) need not be generated according to the same transition probabilities p_{ij} as the main sequence of states $\{i_0, i_1, \ldots\}$. Instead, any transition probabilities $\hat{p}_{ij}$ can be used, as long as the basic condition

$$\hat{p}_{ij} > 0 \qquad \text{if} \qquad a_{ij} \neq 0,$$

is satisfied for all (i, j), and p_{ij} is replaced by $\hat{p}_{ij}$ in the formula (6.7).

A second extension of the preceding methodology is to the case where the rows $\phi(i)'$ of Φ represent expected values with respect to some distribution depending on i, and can be calculated by simulation. In this case the terms $\phi(i)$ and $\phi(i) - \sum_{j=1}^{n} a_{ij}\phi(j)$ in Eq. (6.4) may be replaced by independently generated samples, and the equation approximation formulas may be appropriately adjusted in similar spirit as Eq. (6.6).

A third extension of the preceding methodology is to the approach of minimization of the equation error norm. In particular, one may use of Eq. (2.18) as the basis for a method that involves simulation-based approximation of basis functions. In this approach, in analogy with Eq. (6.6), we generate two additional mutually "independent" sequences [cf. Eq. (6.5)] and we solve the equation

$$\sum_{k=0}^{t} \begin{pmatrix} g_{i_k} - \hat{w}_{k,1} g_{\hat{i}_{k+1}} \\ \hat{w}_{k,1} g_{\hat{i}_{k,1}} - \hat{w}_{k,2} g_{\hat{i}_{k+2}} \\ \vdots \\ \hat{w}_{k,s} g_{\hat{i}_{k,s}} - \hat{w}_{k,s+1} g_{\hat{i}_{k+s+1}} \end{pmatrix} \begin{pmatrix} g_{i_k} - \tilde{w}_{k,1} g_{\tilde{i}_{k+1}} & \tilde{w}_{k,1} g_{\tilde{i}_{k+1}} - \tilde{w}_{k,2} g_{\tilde{i}_{k+2}} & \cdots & \tilde{w}_{k,s} g_{\tilde{i}_{k+s}} - \tilde{w}_{k,s+1} g_{\tilde{i}_{k+s+1}} \end{pmatrix} r$$

$$= \sum_{k=0}^{t} \begin{pmatrix} g_{i_k} - \hat{w}_{k,1} g_{\hat{i}_{k+1}} \\ \hat{w}_{k,1} g_{\hat{i}_{k,1}} - \hat{w}_{k,2} g_{\hat{i}_{k+2}} \\ \vdots \\ \hat{w}_{k,s} g_{\hat{i}_{k,s}} - \hat{w}_{k,s+1} g_{\hat{i}_{k+s+1}} \end{pmatrix} b_{i_k}$$

where $\hat{w}_{k,m}$ and $\tilde{w}_{k,m}$ are given by Eq. (6.7).

We note that constructing basis functions for subspace approximation is an important research issue, and has received considerable attention recently in the approximate DP literature (see, e.g., Keller, Mannor, and Precup [KMP06], Menache, Mannor, and Shimkin [MMS05], Parr et. al. [PPL07], Valenti [Val07]). However, the methods of the present section are new, even within the context of approximate DP, and to our knowledge, they are the first proposals to introduce sampling for basis function approximation directly within the LSTD and LSPE-type methods.

We finally note a generic difficulty associated with the method of this section: even if a solution $r^* = (r_0^*, r_1^*, \ldots, r_s^*)$ of the projected fixed point equation $\Phi r = \Pi T(\Phi r)$ is found, the approximation of the ith component of x^* has the form

$$\phi(i)' r^* = \sum_{m=0}^{s} r_m^* (A^m g)(i),$$

and requires the evaluation of the basis function components $(Ag)(i), \ldots, (A^s g)(i)$. For this, additional computation and simulation is needed, using the approximation formula (6.1).

7. EXTENSION TO SOME NONLINEAR FIXED POINT EQUATIONS

In this section, we briefly discuss how some of the methods of earlier sections can be modified for the case of some nonlinear systems of equations. One potential approach for the general fixed point equation $x = T(x)$, where T is a differentiable mapping, is to use Newton's method to solve the projected equation. In this approach, given r_k, we generate r_{k+1} by using one of the simulation-based methods given earlier to solve a linearized version (at r_k) of the projected equation $\Phi r = \Pi T(\Phi r)$. This is the linear equation

$$\Phi r_{k+1} = \Pi\big(T(\Phi r_k) + \partial T(\Phi r_k)\Phi(r_{k+1} - r_k)\big).$$

where $\partial T(\Phi r_k)$ is the Jacobian matrix of T, evaluated at Φr_k. We do not discuss this approach further, and focus instead on a few special cases involving contraction mappings, where convergence from any starting point r_0 is guaranteed.

Extension of Q-Learning for Optimal Stopping

We first consider a system of the form

$$x = T(x) = Af(x) + b,$$

where $f : \Re^n \mapsto \Re^n$ is a mapping with scalar function components of the form $f(x) = \big(f_1(x_1), \ldots, f_n(x_n)\big)$. We assume that each of the mappings $f_i : \Re \mapsto \Re$ is nonexpansive in the sense that

$$\big|f_i(x_i) - f_i(\bar{x}_i)\big| \leq |x_i - \bar{x}_i|, \qquad \forall\ i = 1, \ldots, n,\ x_i, \bar{x}_i \in \Re. \tag{7.1}$$

This guarantees that T is a contraction mapping with respect to any norm $\|\cdot\|$ with the property

$$\|y\| \leq \|z\| \qquad \text{if}\ \ |y_i| \leq |z_i|,\ \ \forall\ i = 1, \ldots, n,$$

whenever A is a contraction with respect to that norm. Such norms include weighted l_1 and l_∞ norms, the norm $\|\cdot\|_\xi$, as well as any scaled Euclidean norm $\|x\| = \sqrt{x'Qx}$, where Q is a positive definite symmetric matrix with nonnegative components. Under the assumption (7.1), the theory of Section 3 applies and suggests appropriate choices of a Markov chain for simulation.

An interesting special case has been studied in the context of an optimal stopping problem in [TsV99b], which gave a Q-learning algorithm that is similar in spirit to TD(λ), but more complex since it involves a nonlinear mapping. The following example outlines this context.

Example 6: (Optimal Stopping)

Consider the equation

$$x = T(x) = \alpha P f(x) + b,$$

where P is an irreducible transition probability matrix with invariant distribution ξ, $\alpha \in (0, 1)$ is a scalar discount factor, and f is a mapping with components

$$f_i(x_i) = \min\{c_i, x_i\}, \qquad i = 1, \ldots, n,$$

where c_i are some scalars. This is the Q-factor equation corresponding to a discounted optimal stopping problem with states $i = 1, \ldots, n$, and a choice between two actions at each state i: stop at a cost c_i, or continue at a cost b_i and move to state j with probability p_{ij}. The optimal cost starting from state i is $\min\{c_i, x_i^*\}$, where

x^* is the fixed point of T, which is unique because T is a sup-norm contraction, as shown in [TsV99b]. As a special case of Prop. 1, we obtain that ΠT is a contraction with respect to $\| \cdot \|_\xi$, and the associated error bounds apply. Similar results hold in the case where αP is replaced by a matrix A satisfying condition (2) of Prop. 1, or the conditions of Prop. 3. The case where $\sum_{j=1}^n |a_{\bar{i}j}| < 1$ for some index $\bar{i}$, and $0 \le A \le P$, where P is an irreducible transition probability matrix, corresponds to an optimal stopping problem where the stopping state will be reached from all other states with probability 1, even without applying the stopping action. In this case, by Prop. 2, ΠA is a contraction with respect to some norm, and hence $I - \Pi A$ is invertible. Using this fact, it follows by modifying the proof of Prop. 3 that $\Pi\big((1-\gamma)I + \gamma T\big)$ is a contraction with respect to $\| \cdot \|_\xi$.

We will now describe an approximate Jacobi algorithm that extends the method proposed in [YuB07] for the optimal stopping problem of the preceding example. Similar to Section 4, the projected Jacobi iteration

$$\Phi r_{t+1} = \Pi T(\Phi r_t), \qquad t = 0, 1, \ldots,$$

takes the form

$$r_{t+1} = \left(\sum_{i=1}^n \xi_i \, \phi(i)\phi(i)' \right)^{-1} \sum_{i=1}^n \xi_i \, \phi(i) \left(\sum_{j=1}^n a_{ij} f_j\big(\phi(j)'r_t\big) + b_i \right).$$

We approximate this iteration with

$$r_{t+1} = \left(\sum_{k=0}^t \phi(i_k)\phi(i_k)' \right)^{-1} \sum_{k=0}^t \phi(i_k) \left(\frac{a_{i_k j_k}}{p_{i_k j_k}} f_{j_k}\big(\phi(j_k)'r_t\big) + b_{i_k} \right). \tag{7.2}$$

Here, as before, $\{i_0, i_1, \ldots\}$ is a state sequence, and $\{(i_0, j_0), (i_1, j_1), \ldots\}$ is a transition sequence satisfying Eqs. (2.3) and (2.4) with probability 1. The justification of this approximation is very similar to the ones given so far, and will not be discussed further.

A difficulty with iteration (7.2) is that the terms $f_{j_k}\big(\phi(j_k)'r_t\big)$ must be computed for all $k = 0, \ldots, t$, at every step t, thereby resulting in significant overhead. Methods to bypass this difficulty in the case of optimal stopping are given in [YuB07]. These methods can be extended to the more general context of this paper. We finally note that because of the nonlinearity of T, it is hard to implement the equation approximation methods of Section 2, and there are no corresponding versions of the multistep methods of Section 5.

Approximating the Dominant Eigenvalue and Eigenvector of a Matrix

Consider the problem of finding an approximation of the dominant (i.e., real with maximum modulus) eigenvalue of a square $n \times n$ matrix A and an approximation of the corresponding eigenvector within the subspace of basis functions. We assume that such an eigenvalue exists: this can be guaranteed by the Perron-Frobenius Theorem for various cases where A is nonnegative. One possible approach, proposed by Basu, Bhattacharyya, and Borkar [BBB06], is to find instead the dominant eigenvalue $\bar{\lambda}$ of ΠA and an associated eigenvector $\bar{\theta}$. Note that if A is the transpose of an irreducible, aperiodic transition probability matrix Q, then 1 is its dominant eigenvalue and the corresponding eigenvector, suitably normalized, is the invariant distribution of Q.

The method of [BBB06] was proposed in the context of risk-sensitive DP, and is related to the projected value iteration/LSPE methodology discussed in Section 1, but applies to the general case where $A \ge 0$. In the proposal of [BBB06], it is guaranteed that $\Pi A \geq 0$ by a special choice of the basis functions (an aggregation-type matrix Φ, with columns that are nonnegative and have positive components at nonoverlapping positions, so that $\Phi\Phi'$ is diagonal). We will develop a similar method, except that we will just assume that ΠA has a positive dominant eigenvalue $\bar{\lambda}$, with $|\lambda| < \bar{\lambda}$ for all other eigenvalues $\lambda \neq \bar{\lambda}$ of ΠA (rather than require that $A \geq 0$ and $\Pi A \geq 0$). However, we do not know of any conditions, other than the ones of [BBB06], guaranteeing that such $\bar{\lambda}$ exists.

The following version of the power method

$$\Phi r_{t+1} = \frac{\Pi A \Phi r_t}{\|\Phi r_t\|_\xi} \tag{7.3}$$

is known to converge at the rate of a geometric progression to $\bar{\lambda}\bar{\theta}$, where $\bar{\theta}$ is a corresponding eigenvector normalized so that $\|\bar{\theta}\|_\xi = 1$. To obtain a simulation-based approximation of this method, we use a Markov chain with transition probability matrix P and invariant distribution ξ. We generate sequences of states $\{i_0, i_1, \ldots\}$ and transitions $\{(i_0, j_0), (i_1, j_1), \ldots\}$ of the chain, satisfying Eqs. (2.3) and (2.4) with probability 1. We write Eq. (7.3) as

$$r_{t+1} = \frac{\bar{r}_t}{\sqrt{\sum_{i=1}^n \xi_i r_t' \phi(i)\phi(i)' r_t}},$$

where

$$\bar{r}_t = \arg\min_r \sum_{i=1}^n \xi_i \left(\phi(i)'r - \sum_{j=1}^n a_{ij}\phi(j)'r_t \right)^2 = \left(\sum_{i=1}^n \xi_i \phi(i)\phi(i)' \right)^{-1} \sum_{i=1}^n \xi_i \phi(i) \sum_{j=1}^n a_{ij}\phi(j)'r_t.$$

At time t, we approximate this iteration with

$$r_{t+1} = \frac{\left(\sum_{k=0}^t \phi(i_k)\phi(i_k)'\right)^{-1} \sum_{k=0}^t \phi(i_k) \frac{a_{i_k j_k}}{p_{i_k j_k}} \phi(j_k)' r_t}{\sqrt{(t+1)^{-1} r_t' \left(\sum_{k=0}^t \phi(i_k)\phi(i_k)'\right) r_t}}. \tag{7.4}$$

This method is closely related to the method of [BBB06], where a diminishing stepsize was used for a convergence proof. If in their method, the stepsize is taken to be equal to 1, the method (7.4) would be obtained, except for a slight difference in the denominator in Eqs. (7.3) and (7.4).

It is quite clear that $\Phi r_t \to \bar{\lambda}\bar{\theta}$, so that $\|\Phi r_t\|_\xi \to \bar{\lambda}$ with probability 1. Thus it appears that a stepsize equal to 1 is a better practical choice than a diminishing stepsize. We have

$$(t+1)^{-1} r_t' \left(\sum_{k=0}^t \phi(i_k)\phi(i_k)' \right) r_t \to \bar{\lambda}^2$$

with probability 1. This is true for any P and Φ such that ΠA has a dominant eigenvalue. A very important issue is how to choose P and Φ so that $\bar{\lambda}$ and $\bar{\theta}$ are good approximations to the corresponding dominant eigenvalue and eigenvector of A. Generally, the approximations can be arbitrarily poor. As an example, take A to be diagonal with $a_{11} = 1$ and $a_{ii} = \epsilon^i$ for $i \neq 1$ and some $\epsilon \in (0, 1)$, take $\xi = (1/n, \ldots, 1/n)'$, and take S be the subspace that is orthogonal to the dominant eigenvector $(1, 0, \ldots, 0)'$. Then the dominant eigenvalue of A is 1 but the dominant eigenvalue of ΠA is ϵ.

The dominant eigenvalues of A and ΠA will be close to each other if the dominant eigenvector of A is at small distance from S. One way to achieve this is by choosing vectors of the form $A^l x$ to be among the basis functions, where l is sufficiently large and x does not lie in the subspace spanned by the eigenvectors corresponding to the nondominant eigenvalues of A. This is because $\frac{A^l x}{\|A^l x\|_\xi}$ converges to a normalized dominant eigenvector of A as $l \to \infty$. While the components of the basis functions $A^l x$ are unlikely to be explicitly available within our context, the method (7.4) can be suitably modified to use simulation-based approximations, in the spirit of the methods of Section 6.

When A is a symmetric matrix, there is an alternative approach, which is to approximate the maximum eigenvalue/eigenvector of A with the maximum eigenvalue/eigenvector of the matrix

$$\tilde{A} = \Phi' A \Phi = \sum_{i=1}^{n} \sum_{j=1}^{n} a_{ij} \phi(i) \phi(j)'.$$

We can estimate $\tilde{A}$ using an LSTD-type approach, similar to the ones of Section 2. In particular, we form the matrix

$$\hat{A} = \frac{1}{t+1} \sum_{k=0}^{t} \frac{a_{i_k j_k}}{p_{i_k j_k}} \phi(i_k) \phi(j_k)', \tag{7.5}$$

where $\{i_0, i_1, \ldots, i_t\}$ is generated according to a uniform distribution, and $\big\{(i_0, j_0), (i_1, j_1), \ldots, (i_t, j_t)\big\}$ is a sequence of simulated transitions generated according to transition probabilities p_{ij}. We then approximate $\tilde{A}$ with the symmetric matrix $(\hat{A} + \hat{A}')/2$.

If A is positive semidefinite of the form $A = C'C$, with C being an $m \times n$ matrix, and $m << n$, a somewhat different method may be preferable. Again we consider the matrix

$$\tilde{A} = \Phi' C' C \Phi = \sum_{i=1}^{m} \left(\sum_{j=1}^{n} c_{ij} \phi(j) \right) \left(\sum_{j=1}^{n} c_{ij} \phi(j) \right)',$$

and approximate the maximum eigenvalue/eigenvector of A with the maximum eigenvalue/eigenvector of $\tilde{A}$. However, instead of estimating $\tilde{A}$ directly, we form for each $i = 1, \ldots, m$, a simulation-based estimate of

$$v_i = \sum_{j=1}^{n} c_{ij} \phi(j),$$

by using a sequence $\big\{(i, j_0), (i, j_1), \ldots, (i, j_{t_i})\big\}$ generated according to a set of transition probabilities p_{ij}:

$$\tilde{v}_i = \frac{1}{t_i + 1} \sum_{k=0}^{t_i} \frac{c_{ij_k}}{p_{ij_k}} \phi(j_k).$$

We then estimate $\tilde{A}$ by

$$\tilde{A} \approx \sum_{i=1}^{m} \tilde{v}_i \tilde{v}_i'. \tag{7.6}$$

This method may be preferable to the one of Eq. (7.5), because its estimate of $\tilde{A}$ will likely involve less simulation noise [Eq. (7.5) involves averaging of n^2 terms, while Eq. (7.6) involves averaging of n terms for each i, for a total of mn terms]. It is possible to choose p_{ij} using an optimization approach. In particular, it can be shown that for each i, the trace of the covariance of $(c_{ij}/p_{ij})\phi(j)$ when the index j is chosen with sampling probabilities p_{ij}, $j = 1, \ldots, n$, is minimized when p_{ij} is proportional to $|c_{ij}| \, \|\phi(j)\|$. This is consistent with the ideas of importance sampling (see e.g., [Liu01]).

8. CONCLUSIONS

In this paper we have shown how fixed point equations can be solved approximately by projection on a low-dimensional subspace and simulation, thereby generalizing recent methods from the field of approximate DP. We have given error bounds that apply to special types of contraction mappings, most prominently some involving diagonally dominant matrices. However, our methods apply to any linear system of equations whose projected solution is unique. While our simulation-based methods are likely not competitive with other methods for moderately-sized problems, they provide an approach for addressing extremely large problems, because they do not involve any vector operations or storage of size comparable to the original problem dimension. Our methods have been motivated by recent analysis and computational experience in approximate DP. Much remains to be done to apply them and to assess their potential in other fields.

9. REFERENCES

[BBB06] Basu, A., Bhattacharyya, and Borkar, V., 2006. “A Learning Algorithm for Risk-sensitive Cost,” Tech. Report no. 2006/25, Dept. of Math., Indian Institute of Science, Bangalore, India.

[BBN04] Bertsekas, D. P., Borkar, V., and Nedić, A., 2004. “Improved Temporal Difference Methods with Linear Function Approximation,” in Learning and Approximate Dynamic Programming, by J. Si, A. Barto, W. Powell, (Eds.), IEEE Press, N. Y.

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

[Ber07] Bertsekas, D. P., 2007. Dynamic Programming and Optimal Control, Vol. II, 3rd Edition, Athena Scientific, Belmont, MA.

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

[Boy02] Boyan, J. A., 2002. “Technical Update: Least-Squares Temporal Difference Learning,” Machine Learning, Vol. 49, pp. 1-15.

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

[ChV06] Choi, D. S., and Van Roy, B., 2006. “A Generalized Kalman Filter for Fixed Point Approximation and Efficient Temporal-Difference Learning, Discrete Event Dynamic Systems: Theory and Applications, Vol. 16, pp. 207-239.

[KMP06] Keller, P., Mannor, S., and Precup, D., 2006. “Automatic Basis Function Construction for Approximate Dynamic Programming and Reinforcement Learning,” Proceedings of the Twenty-third International Conference on Machine Learning.

[Liu01] Liu, J. S., 2001. Monte Carlo Strategies in Scientific Computing, Springer, N. Y.

[MMS05] Menache, I., Mannor, S., and Shimkin, N., 2005. “Basis Function Adaptation in Temporal Difference Reinforcement Learning,” Annals of Operations Research, Vol. 134.

[NeB03] Nedić, A., and Bertsekas, D. P., 2003. "Least Squares Policy Evaluation Algorithms with Linear Function Approximation," Discrete Event Dynamic Systems: Theory and Applications, Vol. 13, pp. 79-110.

[PPL07] Parr, R., Painter-Wakefield, C, Li, L., Littman, M., 2007. "Analyzing Feature Generation for Value-Function Approximation," Proc. of the 24th International Conference on Machine Learning, Corvallis, OR.

[Put94] Puterman, M. L., 1994. Markov Decision Processes: Discrete Stochastic Dynamic Programming, J. Wiley, N. Y.

[Saa03] Saad, Y., 2003. Iterative Methods for Sparse Linear Systems, 2nd edition, SIAM, Phila., PA.

[SuB98] Sutton, R. S., and Barto, A. G., 1998. Reinforcement Learning, MIT Press, Cambridge, MA.

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

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

[TsV99a] Tsitsiklis, J. N., and Van Roy, B., 1999. "Average Cost Temporal-Difference Learning," Automatica, Vol. 35, pp. 1799-1808.

[TsV99b] Tsitsiklis, J. N., and Van Roy, B., 1999. "Optimal Stopping of Markov Processes: Hilbert Space Theory, Approximation Algorithms, and an Application to Pricing Financial Derivatives," IEEE Transactions on Automatic Control, Vol. 44, pp. 1840-1851.

[Val07] Valenti, M. J., 2007. Approximate Dynamic Programming with Applications in Multi-Agent Systems, Ph.D. Thesis, Dept. of Electrical Engineering and Computer Science, MIT.

[YuB06] Yu, H., and Bertsekas, D. P., 2006. "Convergence Results for Some Temporal Difference Methods Based on Least Squares," Lab. for Information and Decision Systems Report 2697, MIT.

[YuB07] Yu, H., and Bertsekas, D. P., 2007. "A Least Squares Q-Learning Algorithm for Optimal Stopping Problems," Lab. for Information and Decision Systems Report 2731, MIT.