

# New Error Bounds for Approximations from Projected Linear Equations

Huizhen Yu<sup>1</sup> and Dimitri P. Bertsekas<sup>2</sup>

<sup>1</sup> Helsinki Institute for Information Technology (HIIT)  
University of Helsinki, Finland  
`janey.yu@cs.helsinki.fi`

<sup>2</sup> Laboratory for Information and Decision Systems (LIDS)  
M.I.T., Cambridge, MA 02139, USA  
`dimitrib@mit.edu`

**Abstract.** We consider linear fixed point equations and their approximations by projection on a low dimensional subspace. We derive new bounds on the approximation error of the solution, which are expressed in terms of low dimensional matrices and can be computed by simulation. When the fixed point mapping is a contraction, as is typically the case in Markovian decision processes (MDP), one of our bounds is always sharper than the standard worst case bounds, and another one is often sharper. Our bounds also apply to the non-contraction case, including policy evaluation in MDP with nonstandard projections that enhance exploration. There are no error bounds currently available for this case to our knowledge.

## 1 Introduction

For a given  $n \times n$  matrix  $A$  and vector  $b \in \mathbb{R}^n$ , let  $x^*$  and  $\bar{x}$  be solutions of the two linear fixed point equations,

$$x = Ax + b, \quad x = \Pi(Ax + b), \quad (1)$$

respectively, where  $\Pi$  denotes projection on a  $k$ -dimensional subspace  $S$  with respect to certain weighted Euclidean norm  $\|\cdot\|_{\xi}$ . We assume that  $x^*$  and  $\bar{x}$  exist, and that the matrix  $I - \Pi A$  is invertible so that  $\bar{x}$  is unique.

Implicit here is the assumption that  $n$  is very large, so that  $n$ -dimensional vector-matrix operations are practically impossible, while  $k \ll n$ . Our objective in solving the projected equation  $x = \Pi(Ax + b)$  is to approximate the solution of the original equation  $x = Ax + b$  using  $k$ -dimensional computations and storage. This approach is common in MDP, where  $A$  is a stochastic or substochastic matrix, and simulation-based approximate policy evaluation methods, based on temporal differences (TD), have been successfully used (see e.g., [8, 10, 2, 9]). In our recent paper [3], we have extended these methods to the case where  $A$  is an arbitrary matrix, subject only to the restriction that  $I - \Pi A$  is invertible.

In the MDP context, where  $\Pi A$  is usually a contraction, there are two commonly used error bounds that compare the norms of  $x^* - \bar{x}$  and  $x^* - \Pi x^*$ . The

first bound (see e.g., [2, 10]) holds if  $\|IIA\| = \alpha < 1$  with respect to some norm  $\|\cdot\|$ , and has the form

$$\|x^* - \bar{x}\| \leq \frac{1}{1 - \alpha} \|x^* - IIx^*\|. \quad (2)$$

The second bound (see e.g., [11, 1]) holds in the usual case where  $IIA$  is a contraction with respect to the Euclidean norm  $\|\cdot\|_\xi$ , with  $\xi$  being the invariant distribution of the Markov chain underlying the problem, i.e.,  $\|IIA\|_\xi = \alpha < 1$ . It is derived using the Pythagorean theorem  $\|x^* - \bar{x}\|_\xi^2 = \|x^* - IIx^*\|_\xi^2 + \|\bar{x} - IIx^*\|_\xi^2$ , and it is much sharper than the first bound:

$$\|x^* - \bar{x}\|_\xi \leq \frac{1}{\sqrt{1 - \alpha^2}} \|x^* - IIx^*\|_\xi. \quad (3)$$

The bounds (2), (3) are determined by the modulus of contraction  $\alpha$ , and apply only when  $IIA$  is a contraction mapping. We develop in this paper new error bounds, which are sharper when  $IIA$  is a contraction, including important MDP cases, and also apply when  $IIA$  is not a contraction.

Our starting point is the observation that the two terms involved in the bounds (2) and (3) satisfy the following equation with or without contraction assumptions:<sup>3</sup>

$$x^* - \bar{x} = (I - IIA)^{-1}(x^* - IIx^*). \quad (4)$$

We may view the bounds (2), (3) as relaxed versions of this equation. In particular, we may obtain the bound (2) by writing  $(I - IIA)^{-1} = I + IIA + \dots$ , and by upper-bounding each term in the expansion separately:  $\|(IIA)^n\| \leq \alpha^n$ . We may obtain the bound (3) by writing

$$(I - IIA)^{-1} = I + IIA(I - IIA)^{-1}, \quad (5)$$

and by upper-bounding the norm of  $IIA(I - IIA)^{-1}(x^* - IIx^*)$  by  $\alpha\|x^* - \bar{x}\|_\xi$  and rearranging terms.<sup>4</sup> We will develop a different bounding approach, so that  $\alpha$  will not be in the denominator of the bound. To this end, we will express  $(I - IIA)^{-1}$  in the form

$$(I - IIA)^{-1} = I + (I - IIA)^{-1}IIA, \quad (6)$$

<sup>3</sup> This can be seen by subtracting  $\bar{x} = \Pi(A\bar{x} + b)$  from  $IIx^* = \Pi(Ax^* + b)$  to obtain  $IIx^* - \bar{x} = \Pi A(x^* - \bar{x})$ ,  $\Rightarrow (IIx^* - x^*) + (x^* - \bar{x}) = \Pi A(x^* - \bar{x})$ ,  $\Rightarrow$  (4).

<sup>4</sup> From Eqs. (4)-(5) and the orthogonality of  $(x^* - IIx^*)$  to the subspace  $S$ , we have

$$\begin{aligned} \|x^* - \bar{x}\|_\xi^2 &= \|x^* - IIx^*\|_\xi^2 + \|IIA(I - IIA)^{-1}(x^* - IIx^*)\|_\xi^2 \\ &= \|x^* - IIx^*\|_\xi^2 + \|IIA(x^* - \bar{x})\|_\xi^2 \leq \|x^* - IIx^*\|_\xi^2 + \alpha^2 \|x^* - \bar{x}\|_\xi^2. \end{aligned}$$

and aim at bounding the term  $(I - \Pi A)^{-1} \Pi A(x^* - \Pi x^*)$  *directly* (this term is in fact  $\Pi x^* - \bar{x}$ , the bias of  $\bar{x}$  from  $\Pi x^*$ ). In doing so, we will obtain bounds that not only can be sharper than the preceding bounds for the contraction case, but also carry over to the non-contraction case.

We will derive two bounds, which involve the spectral radii of small-size matrices, and provide a “data/problem-dependent” error analysis, in contrast to the fixed error bounds (2), (3); see Theorems 1 and 2. The bounds are *independent* of the parametrization of the subspace  $S$ , and can be computed with low-dimensional operations and simulation, if this is desirable. One of the bounds is sharper than the other, but involves more complex computations. We also have some additional bounds that provide insight into the character of the approximation error, but are qualitative in nature; they are given in an extended version of this paper [12].

Our bounds have the general form

$$\|x^* - \bar{x}\|_\xi \leq B(A, \xi, S) \|x^* - \Pi x^*\|_\xi, \quad (7)$$

where  $B(A, \xi, S)$  is a constant that depends on  $A$ ,  $\xi$ , and  $S$  (but not on  $b$ ). Like the bounds (2), (3), we may view  $\|x^* - \Pi x^*\|_\xi$  as the *baseline error*, i.e., the minimum error in estimating  $x^*$  by a vector in the approximation subspace  $S$ . We may view  $B(A, \xi, S)$  as an upper bound to the *amplification ratio*,  $\|x^* - \bar{x}\|_\xi / \|x^* - \Pi x^*\|_\xi$ , which is due to solving the projected equation  $x = \Pi(Ax + b)$  instead of projecting  $x^*$  on  $S$  (or equivalently, view  $\sqrt{B^2(A, \xi, S) - 1}$  as an upper bound to the “bias-to-distance” ratio  $\|\bar{x} - \Pi x^*\|_\xi / \|x^* - \Pi x^*\|_\xi$ ).

We present our main results in the next section. In Section 3, we address the application of the new error bounds to the approximate policy evaluation in MDP and to the far more general problem of approximate solution of large systems of linear equations. Due to space limitation, proofs and additional related analysis are omitted; they can be found in the expanded version of the present paper [12].

## 2 Main Results

We first introduce the main theorems and explain the underlying ideas. Let  $\Phi$  be an  $n \times k$  matrix whose columns form a basis of  $S$ . Let  $\Xi$  be a diagonal matrix with  $\xi$  on the diagonal. Define  $k \times k$  matrices  $B$ ,  $M$ , and  $F$  by

$$B = \Phi' \Xi \Phi, \quad M = \Phi' \Xi A \Phi, \quad F = (I - B^{-1} M)^{-1} \quad (8)$$

(we will show later that the inverse in the definition of  $F$  exists). Notice that the projection matrix  $\Pi$  can be expressed as  $\Pi = \Phi(\Phi' \Xi \Phi)^{-1} \Phi' \Xi = \Phi B^{-1} \Phi' \Xi$ . For a square matrix  $L$ , let  $\sigma(L)$  denote the spectral radius of  $L$ .

**Theorem 1.** *The approximation error  $x^* - \bar{x}$  satisfies*

$$\|x^* - \bar{x}\|_\xi \leq \sqrt{1 + \sigma(G_1) \|A\|_\xi^2} \|x^* - \Pi x^*\|_\xi, \quad (9)$$

where  $G_1$  is the  $k \times k$  matrix

$$G_1 = B^{-1}F'BF. \quad (10)$$

Furthermore,  $\sigma(G_1) = \|(I - \Pi A)^{-1}\Pi\|_\xi^2$ , so the bound (9) is invariant to the choice of basis vectors of  $S$  (i.e.,  $\Phi$ ).

The idea in deriving Theorem 1 is to combine Eqs. (4)-(5) with the bound

$$\|(I - \Pi A)^{-1}\Pi A(x^* - \Pi x^*)\|_\xi \leq \|(I - \Pi A)^{-1}\Pi\|_\xi \|A\|_\xi \|x^* - \Pi x^*\|_\xi,$$

and to show that  $\|(I - \Pi A)^{-1}\Pi\|_\xi^2 = \sigma(G_1)$ . An important fact, to be demonstrated later, is that  $G_1$  can be obtained by simulation, using low dimensional calculations.

While the bound of Theorem 1 can be conveniently computed, it is less sharp than the bound of the subsequent Theorem 2, and under certain circumstances less sharp than the bound (3). In Theorem 1,  $\|A\|_\xi$  is needed, and this can be a drawback, particularly for the non-contraction case. In Theorem 2,  $\|A\|_\xi$  is no longer needed;  $A$  is absorbed into the matrix to be estimated. Furthermore, Theorem 2 takes into account that  $x^* - \Pi x^*$  is perpendicular to the subspace  $S$ ; this considerably sharpens the bound. On the other hand, the sharpened bound of Theorem 2 involves a  $k \times k$  matrix  $R$  (defined below) in addition to  $B$  and  $M$ , which may not be straightforward to estimate in some cases, as will be commented later.

**Theorem 2.** *The approximation error  $x^* - \bar{x}$  satisfies*

$$\|x^* - \bar{x}\|_\xi \leq \sqrt{1 + \sigma(G_2)} \|x^* - \Pi x^*\|_\xi, \quad (11)$$

where  $G_2$  is the  $k \times k$  matrix

$$G_2 = B^{-1}F'BF B^{-1}(R - MB^{-1}M'), \quad (12)$$

and  $R$  is the  $k \times k$  matrix

$$R = \Phi' \Xi A \Xi^{-1} A' \Xi \Phi.$$

Furthermore,  $\sigma(G_2) = \|(I - \Pi A)^{-1}\Pi A(I - \Pi)\|_\xi^2$ , so the bound (11) is invariant to the choice of basis vectors of  $S$  (i.e.,  $\Phi$ ).

The idea in deriving Theorem 2 is to combine Eqs. (4)-(5) with the bound

$$\begin{aligned} \|(I - \Pi A)^{-1}\Pi A(x^* - \Pi x^*)\|_\xi &= \|(I - \Pi A)^{-1}\Pi A(I - \Pi)(x^* - \Pi x^*)\|_\xi \\ &\leq \|(I - \Pi A)^{-1}\Pi A(I - \Pi)\|_\xi \|x^* - \Pi x^*\|_\xi, \end{aligned}$$

and to show that  $\|(I - \Pi A)^{-1}\Pi A(I - \Pi)\|_\xi^2 = \sigma(G_2)$ . Incorporating the matrix  $I - \Pi$  in the definition of  $G_2$  is crucial for improving the bound of Theorem 1.

Estimating the matrix  $R$ , although not always as straightforward as estimating  $B$  and  $M$ , can be done for a number of applications. A primary exception is when  $A$  itself is an infinite sum of powers of matrices, which is the case of the TD( $\lambda$ ) method with  $\lambda > 0$ . We will address these issues in Section 2.3.

## 2.1 Key Arguments for Proofs

Due to space limitation, we omit the proofs and only point out the main proof arguments. We shall need two technical lemmas. The first lemma introduces an expression of the matrix  $(I - \Pi A)^{-1}$  that will be used to derive our error bounds. The second lemma establishes the relation between the norm of an  $n \times n$  matrix that is a product of  $n \times k$  and  $k \times n$  matrices, and the spectral radius of a certain product of  $k \times k$  matrices.

**Lemma 1.** *The matrix  $I - \Pi A$  is invertible if and only if the inverse  $(I - B^{-1}M)^{-1}$  defining  $F$  exists. When  $I - \Pi A$  is invertible,  $(I - \Pi A)^{-1}$  maps  $S$  onto  $S$ , and furthermore,*

$$(I - \Pi A)^{-1} = I + (I - \Pi A)^{-1} \Pi A = I + \Phi F B^{-1} \Phi' \Xi A. \quad (13)$$

Note that since  $B$  and  $M$  are low-dimensional matrices, the first part of Lemma 1 is useful for verifying the existence of the inverse of  $I - \Pi A$  using the data.

**Lemma 2.** *Let  $H$  and  $D$  be an  $n \times k$  and  $k \times n$  matrix, respectively. Let  $\|\cdot\|$  denote the standard (unweighted) Euclidean norm. Then,*

$$\|HD\|_{\xi}^2 = \|\Xi^{1/2} H D \Xi^{-1/2}\|^2 = \sigma((H' \Xi H)(D \Xi^{-1} D')). \quad (14)$$

Theorems 1 and 2 can now be proved by combining Lemmas 1 and 2, the relation (4), and the proof ideas described immediately after the statements of the theorems.

## 2.2 Comparison of Error Bounds

The error bounds of Theorems 1 and 2 apply to the general case where  $\Pi A$  is not necessarily a contraction mapping, while the worst case error bounds (2) and (3) only apply when  $\Pi A$  is a contraction. We will thus compare them for the contraction case. Nevertheless, our discussion will illuminate the strengths and weaknesses of the new bounds for both contraction and non-contraction cases.

First it can be shown that the error bound of Theorem 2 is always the sharpest.

**Proposition 1.** *Assume that  $\|\Pi A\|_{\xi} \leq \alpha < 1$ . Then, the error bound of Theorem 2 is always no worse than the error bound (3), i.e.,  $1 + \sigma(G_2) \leq 1/(1 - \alpha^2)$ , where  $G_2$  is given by Eq. (12).*

It can also be shown that the error bound of Theorem 2 is tight in the sense that there is a worst case choice of  $b$  for which the bound holds with equality.

Let us compare now the error bound of Theorem 1 with the bounds (2) and (3) from the worst case viewpoint. Since Theorem 1 is effectively equivalent to

$$\|(I - \Pi A)^{-1} \Pi A(x^* - \Pi x^*)\|_{\xi} \leq \|(I - \Pi A)^{-1} \Pi\|_{\xi} \|A\|_{\xi} \|x^* - \Pi x^*\|_{\xi},$$

we see that the bound of Theorem 1 is never worse than the bound (2), because we have bounded the norm of the matrix  $(I - \Pi A)^{-1} \Pi$  as a whole, instead of bounding each term in its expansion separately as in the case in the bound (2). However, the bound of Theorem 1 can be degraded by two over-relaxations:

- (i) The residual vector  $x^* - \Pi x^*$  is special, in that it satisfies  $\Pi(x^* - \Pi x^*) = 0$ , but the bound does not use this fact.
- (ii) When  $\Pi A$  is zero or near zero, the bound cannot fully utilize this fact.

The effect of (i) can be quite significant when  $A$  has a dominant real eigenvalue  $\beta$  with an eigenvector  $x$  that lies in the approximation subspace  $S$ . In such a case, the bound reduces essentially to the bound (2), since

$$\|(I - \Pi A)^{-1} \Pi x\|_{\xi} = \frac{1}{1-\beta} \|x\|_{\xi}. \quad (15)$$

This happens because the analysis has not taken into account that the residual vector  $(x^* - \Pi x^*)$  cannot be an eigenvector that is contained in  $S$ .

The relaxation related to (ii) may not look obvious in the current analysis; it does, however, in an alternative equivalent form of the analysis, by noticing that

$$(I - \Pi A)^{-1} \Pi A = \Pi A + \Pi A (I - \Pi A)^{-1} \Pi A, \quad (16)$$

and the norm of the matrix on the right has been bounded by  $\|\Pi + \Pi A (I - \Pi A)^{-1} \Pi\|_{\xi} \|A\|_{\xi}$  in Theorem 1. When  $\Pi A = 0$  the matrix of Eq. (16) is zero but its bound is not, because the matrices  $\Pi$  and  $A$  are split in the bounding procedure. Accordingly, the spectral radius  $\sigma(G_1)$  becomes  $\|\Pi\|_{\xi}^2 = 1$ . Similarly, over-relaxation occurs when  $\Pi A$  is not zero but is near zero.<sup>5</sup>

The two shortcomings of the bound of Theorem 1 arise in the MDP applications that we will discuss, as well as in non-contraction cases. On the other hand, there are cases where Theorem 1 provides sharper bounds than the fixed error bound (3), and cases where Theorem 1 gives computable bounds while the bound (3) is qualitative (for example, when the modulus of contraction of  $\Pi A$  is unknown). We also mention that the expanded version of this paper [12] contains additional analysis, which in part addresses the shortcomings just discussed.

The advantage that the bound of Theorem 1 holds over the one of Theorem 2 is that it is rather easy to compute: the matrices  $B$  and  $M$  define the solution  $\bar{x}$ , so the bound is obtained together with the approximating solution without extra computation overhead. By contrast, the bound of Theorem 2 involves the matrix  $R$ , which can be hard to estimate for certain applications.

### 2.3 Estimating the Low Dimensional Matrices in the Bounds

We consider estimating the  $k \times k$  matrices involved in the bounds by simulation, and we focus on estimating the matrix  $R$  in Theorem 2:

$$R = \Phi' \Xi A \Xi^{-1} A' \Xi \Phi.$$

<sup>5</sup> In practice, when using the bound of Theorem 1, one may check if  $\Pi A$  is near zero by checking if  $M$  is.

Other cases do not seem to need explanations: the estimation of  $B$  and  $M$  using simulation has been well explained in the literature (see e.g., [4, 7, 3]); and if instead of using simulation, products of  $k \times n$  and  $n \times n$  matrices can be computed directly, then the calculation of  $R$  may be done directly with common matrix algebra.

First, let us note that when the matrix  $\Phi$  actually used in the simulation does not have full rank, Theorems 1 and 2 imply that the bounds can be computed by using the pseudo-inverse of  $B$ , neglecting zero eigenvalues (a tolerance level/threshold needs to be determined, of course, in the simulation context).

Without loss of generality, in this subsection, we assume that  $\sum_{i=1}^n \xi_i = 1$  so that  $\xi$  can be viewed as a distribution. In practice, we never need to normalize  $\xi$  as the normalization constant will be canceled in the product defining the matrices  $G_1$  and  $G_2$ . Let  $\phi(i)'$  denote the  $i$ -th row of  $\Phi$ . Our methods for estimating  $R$  are based on a common procedure: we first express  $R$  as a summation of  $k \times k$  matrices, e.g.,

$$R = \sum_{i, \hat{j}} (a_{ji} a_{\hat{j}i}) \cdot \frac{\xi_j \xi_{\hat{j}}}{\xi_i} \cdot \phi(j) \phi(\hat{j})',$$

and guided by this expression, we generate samples and choose proper weights for them, so that each term in the summation is matched by a weighted long-run average of respective samples.

We will give four examples that apply to different contexts, depending on whether the entries of  $\xi$  and  $A$  in the preceding formula for  $R$  are explicitly known or not, with two main applications in our mind:

- (i) *General linear equations* in which we know explicitly the entries of  $A$ , and we may want to choose a particular projection norm, for instance, the standard Euclidean norm (all entries of  $\xi$  being equal). The procedure of Example 1 and its slight variant in Example 2 refer primarily to this case.
- (ii) *Markov decision processes* in which we do not know  $A$ , but we can generate samples by simulation of a certain Markov chain underlying the problem. Examples 3 and 4 are mostly relevant to this case, including in particular, evaluating the cost or  $Q$ -factors of a policy using TD(0)-like algorithms, with and without exploration enhancements. (We refer to our paper [3] for some algorithms involving exploration, where the simulation procedures of Examples 3 and 4 may apply.)

*Example 1.* Both  $\xi$  and  $A$  are known explicitly. We express  $R$  as the summation given above and generate a sequence of triple indices  $(i_t, j_t, \hat{j}_t)$  as follows. We generate the sequence  $(i_0, i_1, \dots)$  so that its empirical distribution converges to  $\xi$ . At  $i_t$ , we generate two mutually independent transitions  $(i_t, j_t)$  and  $(i_t, \hat{j}_t)$  according to a certain transition probability matrix  $P$  with  $p_{ij} \neq 0$  whenever  $a_{ji} \neq 0$ . We then define  $R_t$  by

$$R_t = \frac{1}{t+1} \sum_{m=0}^t \left( \frac{a_{j_m i_m}}{p_{i_m j_m}} \cdot \frac{a_{\hat{j}_m i_m}}{p_{i_m \hat{j}_m}} \right) \cdot \frac{\xi_{j_m} \xi_{\hat{j}_m}}{\xi_{i_m}^2} \cdot \phi(j_m) \phi(\hat{j}_m)',$$

where  $t$  is a suitably large number, and approximate  $R$  by the symmetrized matrix  $(R_t + R'_t)/2$ . Note that in the special case where  $\Xi = \frac{1}{n}I$ , the indices  $i_t$  can be generated independently with the uniform distribution,  $R$  reduces to  $\frac{1}{n}\Phi'AA'\Phi$ , and the ratio  $\frac{\xi_{j_m}\xi_{j_m}}{\xi_{i_m}^2}$  in  $R_t$  reduces to 1.  $\square$

*Example 2.* The weight vector  $\xi$  is not known explicitly, but  $A$  is; moreover, a sequence  $(i_0, i_1, \dots)$  can be generated so that its empirical distribution converges to  $\xi$ . For example,  $\xi$  may be the unique invariant distribution of a Markov chain, which is used to generate the sequence  $(i_0, i_1, \dots)$ . In this case, we can keep tracking the empirical distribution  $\hat{\xi}_t$  of the sequence  $i_t$  up to time  $t$ . We then apply the same sampling and estimation schemes as in Example 1, replacing the ratio  $\frac{\xi_{j_m}\xi_{j_m}}{\xi_{i_m}^2}$  in  $R_t$  by  $\frac{\hat{\xi}_{t,j_m}\hat{\xi}_{t,j_m}}{\hat{\xi}_{t,i_m}^2}$ .  $\square$

*Example 3.* Both  $\xi$  and  $A$  are not known explicitly; moreover, the ratios  $\beta_{ij} = a_{ij}/p_{ij}$  are known for a certain transition matrix  $P$  with  $p_{ij} \neq 0$  whenever  $a_{ij} \neq 0$ , and  $\xi$  is the unique invariant distribution of the Markov chain associated with  $P$ . While  $P$  is not explicitly known, it is assumed that a simulator is available that can generate transitions according to  $P$ .

To estimate  $R$ , we first express it as

$$R = \sum_{i,l,j} (\beta_{il}\beta_{jl}) \cdot \left( \xi_i p_{il} \cdot \frac{p_{il}\xi_j}{\xi_i} \right) \cdot \phi(i)\phi(j)' .$$

Noticing that  $\frac{p_{il}\xi_j}{\xi_i}$  equals the steady-state conditional probability  $P(X_{t-1} = j \mid X_t = l)$  for the Markov chain  $X_t$ , we thus generate a sequence of pairs of indices  $(i_t, j_t)$  as follows. Let  $(i_0, i_1, \dots)$  be a trajectory of the Markov chain. At  $i_{t+1} = l$ , we generate, using the uniform distribution, one sample  $(j, l)$  from the set of past transitions to  $l$ ,  $\{(i_{t_k-1}, i_{t_k}) \mid i_{t_k} = l, t_k \leq t+1\}$ , and we let  $j_t = j$ . (Indeed, this will also work if we simply let  $j_t = i_{t_{k-1}}$  where  $t_k$  is the most recent time prior to  $t+1$  that  $i_{t_k} = l$ .) It can be seen that the conditional probability of  $j_t$  given  $i_{t+1}$  converges asymptotically to  $\frac{p_{j_t i_{t+1}} \xi_{j_t}}{\xi_{i_{t+1}}}$ . We then define  $R_t$  by

$$R_t = \frac{1}{t+1} \sum_{m=0}^t (\beta_{i_m i_{m+1}} \beta_{j_m i_{m+1}}) \cdot \phi(i_m)\phi(j_m)' ,$$

and we approximate  $R$  by the symmetrized matrix  $(R_t + R'_t)/2$ .

If the Markov chain is reversible, i.e.,  $\xi_j p_{jl} = \xi_l p_{lj}$  for all  $j, l$ , then the method can be substantially simplified. We can omit the procedure of generating  $j_t$  and simply set  $j_m = i_{m+2}$  in  $R_t$ , because if we do so, the proper weight for the sample is  $\frac{\xi_{j_m} p_{j_m i_{m+1}}}{\xi_{i_{m+1}} p_{i_{m+1} j_m}} = 1$ .  $\square$

*Example 4.* The weight vector  $\xi$  is known explicitly, but  $A$  is not; moreover, the ratios  $\beta_{ij} = a_{ij}/p_{ij}$  are known for a certain transition matrix  $P$  with  $p_{ij} \neq 0$  whenever  $a_{ij} \neq 0$ . Here,  $\xi$  need not be the invariant distribution of  $P$ .

We can deal with this case by combining partially the schemes in Examples 2 and 3. We express  $R$  and generate a sequence of pairs of indices  $(i_t, j_t)$  as in Example 3. We keep tracking the empirical distribution  $\kappa_t$  of the sequence  $i_t$  up to time  $t$ , to approximate the invariant distribution of  $P$ . We weight samples properly to define  $R_t$ :

$$R_t = \frac{1}{t+1} \sum_{m=0}^t (\beta_{i_m i_{m+1}} \beta_{j_m i_{m+1}}) \cdot \left( \frac{\xi_{i_m} \xi_{j_m}}{\xi_{i_{m+1}}} \cdot \frac{\kappa_{t, i_{m+1}}}{\kappa_{t, i_m} \kappa_{t, j_m}} \right) \cdot \phi(i_m) \phi(j_m)',$$

and we approximate  $R$  by the symmetrized matrix  $(R_t + R_t')/2$ .

If the Markov chain associated with  $P$  is reversible, then there is simplification, similar to that in Example 3. We simply set  $j_t = i_{t+2}$  and

$$R_t = \frac{1}{t+1} \sum_{m=0}^t (\beta_{i_m i_{m+1}} \beta_{i_{m+2} i_{m+1}}) \cdot \left( \frac{\xi_{i_m} \xi_{i_{m+2}}}{\xi_{i_{m+1}}} \cdot \frac{\kappa_{t, i_{m+1}}}{\kappa_{t, i_m} \kappa_{t, i_{m+2}}} \right) \cdot \phi(i_m) \phi(i_{m+2})',$$

because the extra term needed for weighting the sample properly is  $\frac{\kappa_{t, j_m} P_{j_m i_{m+1}}}{\kappa_{t, i_{m+1}} P_{i_{m+1} j_m}}$ , which converges to 1 as  $m \rightarrow \infty$ .  $\square$

A main source of difficulty in the estimation of  $R$  in MDP, as Examples 3 and 4 illustrate, is the unknown matrix  $A$  and the need of samples of “backward” transitions from a common state/index. Simulating backward transitions according to the steady-state conditional distribution is in general not easy. Consistently, as Example 1 illustrates, the estimation of  $R$  is quite simple when backward transitions can be easily generated, such as when  $A$  is known. A second source of difficulty in the estimation of  $R$ , as Examples 2-4 illustrate, is the memory demand. In particular, in order to either generate backward transitions or to weight samples properly, we must keep track of the past history of the simulation (except in the case of Example 3 and a reversible Markov chain).

Another drawback of the procedures given in Examples 1-4 is that they do not adapt easily to the case where  $A$  itself is a summation of infinitely many matrices, as in TD( $\lambda$ ) with  $\lambda > 0$ .

### 3 Applications

We consider two applications of Theorems 1 and 2. The first one is cost function approximation in MDP with TD-type methods. This includes single policy evaluation with discounted and undiscounted cost criteria. The second application is approximately solving large general systems of linear equations. We also illustrate with figures various issues discussed in Section 2.2 on the comparison of the bounds. We note that for TD( $\lambda$ ) with  $\lambda > 0$ , we do not yet have an efficient simulation-based method for estimating the bound of Theorem 2; we have calculated the bound using common matrix algebra, and we plot it just for comparison.

### 3.1 Cost Function Approximation for MDP

For policy evaluation in MDP,  $x^*$  is the cost function of the policy to be evaluated. Let  $P$  be the transition matrix of the Markov chain induced by the policy. The original linear equation that we want to solve is the Bellman equation, or optimality equation, satisfied by  $x^*$ . It takes the form

$$x^* = b + \alpha P x^* ,$$

where  $b$  is the per-stage cost vector, and  $\alpha \in [0, 1]$  is the discount factor:  $\alpha \in [0, 1)$  corresponds to the discounted cost criterion, while  $\alpha = 1$  corresponds to either the total cost criterion or the average cost criterion (in the latter case  $b$  is the per-stage cost minus the average cost). For simplicity of discussion, we assume that the Markov chain is irreducible.

With the TD( $\lambda$ ) method, we solve a projected form of the multistep Bellman equation  $x = \Pi b + \Pi A x$ , where  $A$  is defined for a pair of values  $(\alpha, \lambda)$  by

$$A = P^{(\alpha, \lambda)} \stackrel{def}{=} (1 - \lambda) \sum_{l=0}^{\infty} \lambda^l (\alpha P)^{l+1}$$

with either  $\alpha \in [0, 1), \lambda \in [0, 1]$ , or  $\alpha = 1, \lambda \in [0, 1)$ . Notice that the case  $\lambda = 0$  corresponds to  $A = \alpha P$ .

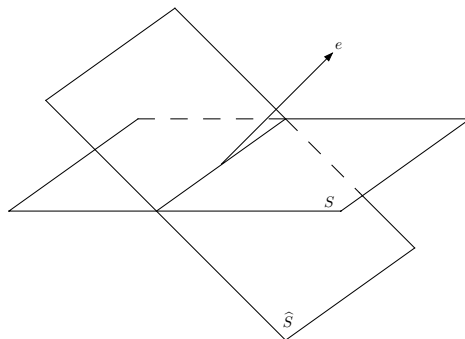
**Discounted Problems.** Consider the discounted case:  $\alpha < 1$ . For  $\lambda \in [0, 1]$ , with  $\xi$  being the invariant distribution of the Markov chain, the modulus of contraction of  $P^{(\alpha, \lambda)}$  with respect to  $\|\cdot\|_{\xi}$  is

$$\|P^{(\alpha, \lambda)}\|_{\xi} = \frac{(1 - \lambda)\alpha}{1 - \lambda\alpha} .$$

Let  $e$  denote the constant vector of all ones. Like  $P$ , the matrix  $P^{(\alpha, \lambda)}$  has  $e$  as an eigenvector associated with the dominant eigenvalue  $\frac{(1-\lambda)\alpha}{1-\lambda\alpha}$ .

If the approximation subspace  $S$  contains or nearly contains  $e$ , the bound of Theorem 1 can degrade to the worst case error bound given by (2), as remarked in Section 2.2. In such a case, in order to have a sharper bound for the approximation of  $\Pi x^*$ , we can estimate separately the projection of  $x^*$  on  $e$  and the projection of  $x^*$  on another subspace  $\hat{S} = (S \oplus e) \cap e^{\perp}$ , which is the orthogonal complement of  $e$  in  $S \oplus e$  (see Figure 1), and redefine  $\bar{x}$  as the sum of the two estimates. It can be shown that when the first projection can be estimated with no bias, the error bound for the second projection carries over to the combined estimate  $\bar{x}$ .<sup>6</sup> Fortunately, for MDP, the projection of  $x^*$  on  $e$  can be calculated

<sup>6</sup> For a subspace  $V$ , let  $\Pi_V$  denote the projection on  $V$ . Let  $V$  and  $W$  be two orthogonal subspaces with  $\Pi_V x^*$  known. Since  $x^* - \Pi_V x^*$  satisfies the linear equation  $x = Ax + \tilde{b}$  with  $\tilde{b} = b + A\Pi_V x^* - \Pi_V x^*$ , to obtain an estimate of  $\Pi_W x^* = \Pi_W(x^* - \Pi_V x^*)$ , we can solve the projected equation  $x = \Pi_W A x + \Pi_W \tilde{b}$ . (In the above MDP case,  $V$  is an eigenspace of  $A$ , so  $\tilde{b}$  can be replaced by  $b$ .) Denote the solution by  $\bar{x}_w$ . Then, error bounds for  $\bar{x}_w$  that are of the form  $\|(x^* - \Pi_V x^*) - \bar{x}_w\|_{\xi} \leq L\|(x^* - \Pi_V x^*) - \Pi_W(x^* - \Pi_V x^*)\|_{\xi}$ , are equivalent to error bounds  $\|x^* - \bar{x}\|_{\xi} \leq L\|x^* - \Pi_{V \oplus W} x^*\|_{\xi}$  with  $\bar{x} = \Pi_V x^* + \bar{x}_w$ .



**Fig. 1.** Illustration of  $\widehat{S}$ , the orthogonal complement of  $e$  in  $S \oplus e$ , i.e.,  $\widehat{S} = (S \oplus e) \cap e^\perp$ .

asymptotically exactly through simulation.<sup>7</sup> In addition, basis vectors of  $\widehat{S}$  can also be generated from  $\Phi$  by using simulation (see eg., [5]), along with the approximation of the matrices  $B$  and  $M$  and without incurring much computation overhead. Figure 2 illustrates the error bounds, and shows how the use of  $\widehat{S}$  may improve them. It can be observed that the bound of Theorem 2 has consistently performed best, as indicated by the analysis.

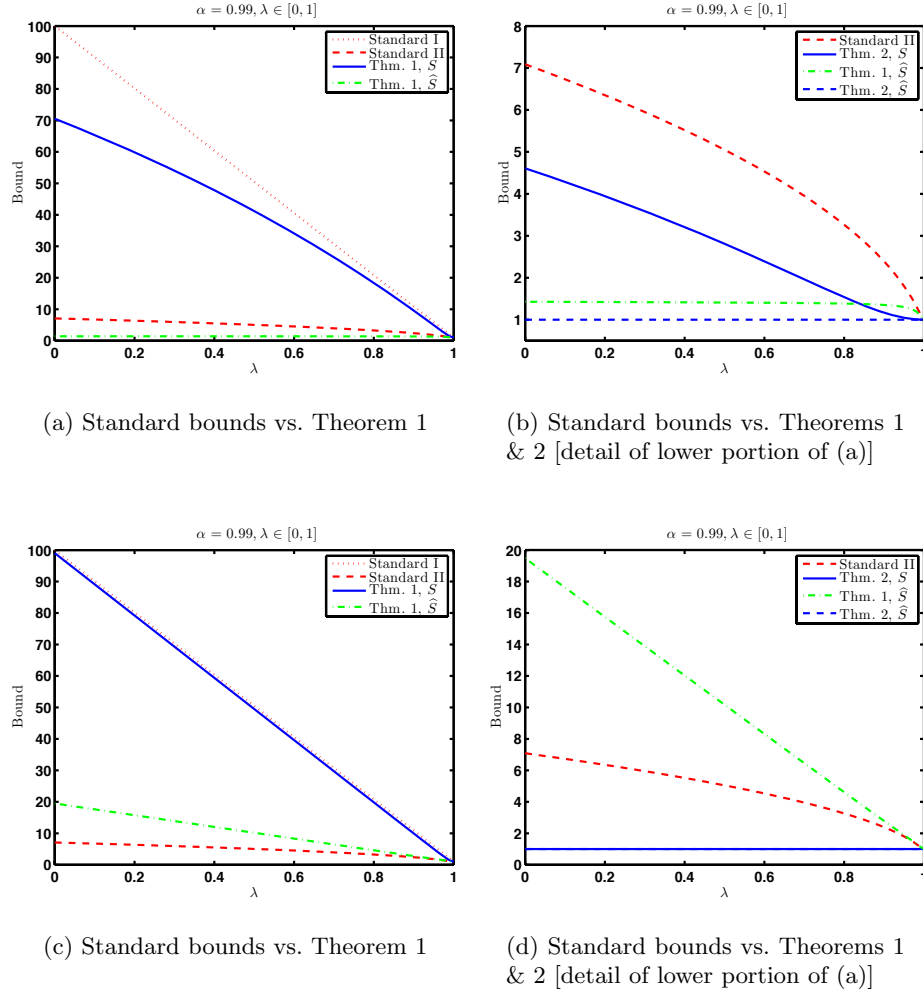
Figure 3 compares the bounds for the case where the projection norm is the standard unweighted Euclidean norm. The standard bounds and the bound of Theorem 1 need the value  $\|A\|$ , while the bound of Theorem 2 does not. For comparison of these bounds, we compute  $\|P\|$  using the knowledge of  $P$ , bound  $\|A\|$  by  $\frac{(1-\lambda)\|\alpha P\|}{1-\lambda\|\alpha P\|}$ , and plug the latter in the standard bounds and the bound of Theorem 1. The value  $\|\alpha P\|$ , which corresponds to  $\|A\|$  for  $\lambda = 0$ , is shown in the titles of Figure 3. With the norm being different from  $\|\cdot\|_\xi$ , the mapping  $\Pi A$  is not necessarily a contraction for small values of  $\lambda$ , even though in this example it is.

Note that the availability of computable error bounds for non-contraction mappings facilitates the design of policy evaluation algorithms with improved exploration. In particular, we can use the LSTD algorithm [4] to evaluate the cost or the  $Q$ -factor of a policy using special sampling methods that enhance exploration, and use the bound of Theorem 1 to estimate the corresponding amplification ratio.<sup>8</sup> Alternatively, we may use the bound of Theorem 2 in conjunction with TD(0)-type algorithms. Examples 3 and 4 show how to estimate the matrix  $R$  in cases where the projection norm is determined by an exploration policy,

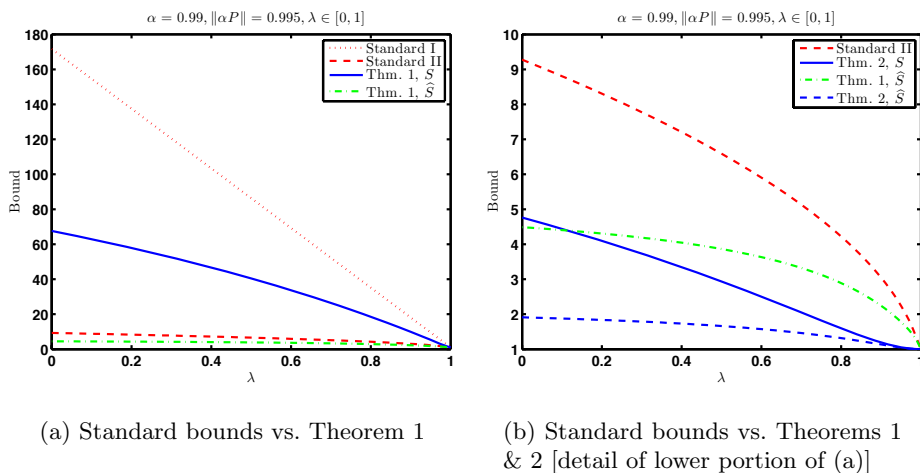
<sup>7</sup> It can be seen that the projection of  $x^*$  on  $e$  equals

$$\xi' x^* = \xi' b + \xi' P^{(\alpha, \lambda)} x^* = \xi' b + \frac{(1-\lambda)\alpha}{1-\lambda\alpha} \xi' x^*, \quad \Rightarrow \quad \xi' x^* = \frac{1-\lambda\alpha}{1-\alpha} \xi' b.$$

<sup>8</sup> When  $\Pi A$  is not necessarily a contraction, a bound on  $\|A\|_\xi$  is needed to apply Theorem 1. There are also algorithms involving exploration and maintaining the contraction property of  $\Pi A$ , for which we refer to our paper [3].



**Fig. 2.** Comparison of error bounds as functions of  $\lambda$  for two discounted problems with randomly generated Markov chains. The dimension parameters are  $n = 200, k = 50$ , and the weights  $\xi$  in the projection norm is the invariant distribution. Standard I and II refer to the worst case bounds (2) and (3), respectively. The Markov chain is the same in (a) and (b), and in (c) and (d). In (c) and (d), the Markov chain has a “noisy” block structure with two blocks, thus  $P$  has a relatively large subdominant eigenvalue;  $S$  is chosen to contain  $e$  and a vector close to an eigenvector associated with that subdominant eigenvalue. The subspace  $\hat{S}$  is derived from  $S$  by orthogonalization, as shown in Figure 1.



**Fig. 3.** Comparison of error bounds for discounted problems. The setup is the same as that for Figure 2, except that the projection norm is the standard Euclidean norm. The Markov chain has a “noisy” block structure. The subspace  $S$  is chosen randomly.

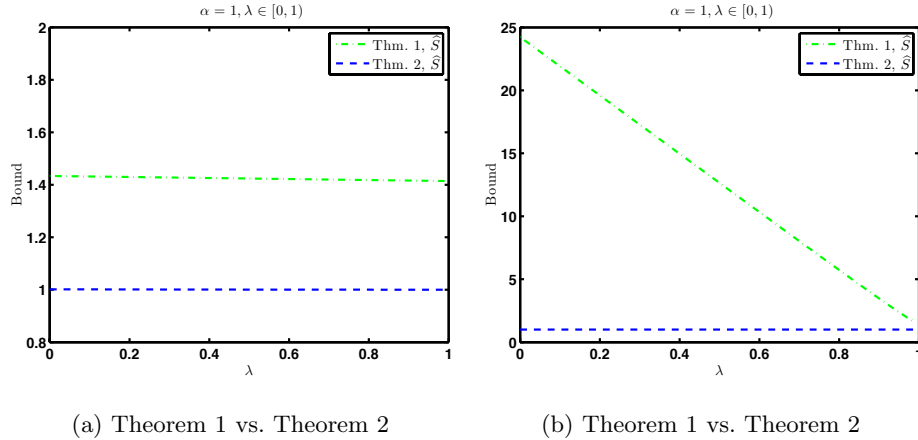
and where the projection norm is given explicitly with the desirable weights, respectively.

**Average Cost and Stochastic Shortest Path (SSP) Problems.** In the average cost case (similarly for SSP),  $x^*$  is the differential cost or bias vector and it is orthogonal to  $e$ . Let us assume that  $S$  is orthogonal to  $e$ , to simplify the discussion. The error bound corresponding to the bound (3), and given by Tsitsiklis and Van Roy [11] is

$$\|x^* - \bar{x}\|_{\xi} \leq \frac{1}{\sqrt{1 - \alpha_{\lambda}^2}} \|x^* - \Pi x^*\|_{\xi},$$

where  $\alpha_{\lambda} < 1$  and  $\alpha_{\lambda} \rightarrow 0$  as  $\lambda \rightarrow 1$ . Here,  $\alpha_{\lambda}$  can be viewed as the modulus of contraction of some mapping that is a damped version of  $\Pi A$ , while  $\alpha_{\lambda} \rightarrow 0$  reflects the fact that the matrix  $\Pi A$  converges to the zero matrix (as  $A$  converges to  $e\xi'$ ) as  $\lambda \rightarrow 1$ . Note that the factor in the bound converges to 1, as  $\lambda \rightarrow 1$ . This bound is qualitative, as usually the value of  $\alpha_{\lambda}$  is unknown.

Figure 4 shows the bounds of Theorems 1 and 2. Notice that as  $\lambda \rightarrow 1$ , the bound of Theorem 1 converges to  $\sqrt{2}$  instead of 1. This is due to the over-relaxation in the analysis for the case where  $\Pi A$  is near zero, as remarked in Section 2.2. Notice also in Figure 4(b) that the bound of Theorem 1 is affected by the relation of  $S$  to the eigenspace of  $A$  associated with eigenvalues that are close to 1, similar to the discounted case. By contrast, the bound of Theorem 2 performs well.



**Fig. 4.** Comparison of error bounds for average cost problems with randomly generated Markov chains. The setup is the same as that for Figure 2. In (b), the Markov chain has a “noisy” block structure, and  $S$  is chosen as in Figure 2(c).

### 3.2 Large General Systems of Linear Equations

For solving large general systems of linear equations using the projected equation approach [3], the bound of Theorem 2 can be computed in a straightforward way (except in the case of  $\text{TD}(\lambda)$  with  $\lambda > 0$ ), as shown in Examples 1 and 2. Theorem 2 is not only much sharper than Theorem 1 for this case, but also more convenient, because it does not require the knowledge of  $\|A\|_\xi$ . Note that we can write linear equations of the form  $Lx = q$  as  $x = Ax + b$ , with  $A = I + cL$  and  $b = -cq$  for any scalar  $c$ , and we can choose  $c$  to optimize the corresponding error bound.

## 4 Discussion

We have considered the projected equation approximation approach, and we have presented new data-dependent computable error bounds that hold for both contraction and non-contraction mappings. Their applicability for non-contraction mappings is not only useful for approximating solutions of general linear equations, but is also useful in the context of MDP for designing exploration mechanisms. Furthermore, in the context of MDP, these bounds can be used in performance bounds for approximate policy iteration, such as [6].

One potential use of our bounds is to suggest changes in the projected equation in order to reduce the amplification ratio. For example, extensive computational experience with  $\text{TD}(\lambda)$  methods suggests that the simulation noise tends to increase as  $\lambda$  increases, so there is motivation to use small values of  $\lambda$  as long as the amplification ratio is close to 1. Unfortunately, the bounds (2), (3)

are too conservative to provide useful information about the amplification ratio, and our bounds can provide quantitative guidance as well as valuable insight in this regard. Furthermore, our bounds can be similarly used in the general non-contraction context, in conjunction with simulation-based TD( $\lambda$ )-like algorithms that have been developed in our recent paper [3]. There may be other potential uses of our bounds, for example in suggesting changes to the choice of approximation subspace, thereby affecting both the baseline error and the amplification ratio, but this is a subject for future research.

**Acknowledgment.** Huizhen Yu is supported in part by Academy of Finland grant 118653 (ALGODAN) and by the IST Programme of the European Community, PASCAL Network of Excellence, IST-2002-506778.

## References

1. D. P. Bertsekas. *Dynamic Programming and Optimal Control*, volume II. Athena Scientific, Belmont, MA, third edition, 2007.
2. D. P. Bertsekas and J. N. Tsitsiklis. *Neuro-Dynamic Programming*. Athena Scientific, Belmont, MA, 1996.
3. D. P. Bertsekas and H. Yu. Projected equation methods for approximate solution of large linear systems. *J. Computational and Applied Mathematics*, 2008. to appear.
4. J. A. Boyan. Least-squares temporal difference learning. In *Proc. The 16th Int. Conf. Machine Learning*, 1999.
5. V. R. Konda. *Actor-Critic Algorithms*. PhD thesis, MIT, Cambridge, MA, 2002.
6. R. Munos. Error bounds for approximate policy iteration. In *Proc. The 20th Int. Conf. Machine Learning*, 2003.
7. A. Nedić and D. P. Bertsekas. Least squares policy evaluation algorithms with linear function approximation. *Discrete Event Dyn. Syst.*, 13:79–110, 2003.
8. R. S. Sutton. Learning to predict by the methods of temporal differences. *Machine Learning*, 3:9–44, 1988.
9. R. S. Sutton and A. G. Barto. *Reinforcement Learning*. MIT Press, Cambridge, MA, 1998.
10. J. N. Tsitsiklis and B. Van Roy. An analysis of temporal-difference learning with function approximation. *IEEE Trans. Automat. Contr.*, 42(5):674–690, 1997.
11. J. N. Tsitsiklis and B. Van Roy. Average cost temporal-difference learning. *Automatica*, 35(11):1799–1808, 1999.
12. H. Yu and D. P. Bertsekas. New error bounds for approximations from projected linear equations. Technical Report C-2008-43, University of Helsinki, 2008.