Analyzing Feature Generation for Value-Function Approximation

Ronald Parr*, Christopher Painter-Wakefield*, Lihong Li**, Michael Littman**
*Duke University, **Rutgers University

Abstract

We analyze a simple, Bellman-error-based approach to generating features, or basis functions, for value-function approximation in MDPs and reinforcement learning. We show that this procedure generates orthogonal basis functions that provably tighten approximation error bounds. We also prove sufficient conditions for a basis function that approximates the Bellman error to improve the value-function. Our work is the first rigorous analysis on the effect of new basis functions on value-function accuracy.

1 Introduction

We extend recent efforts in feature discovery for value-function approximation [3; 2]. We consider approaches based upon the Bellman error [2] in the context of linear value-function approximation. Specifically, we consider a general family of approaches that iteratively add basis functions to a linear approximation architecture in a manner where each new basis function is derived from the Bellman error of the previous set of basis functions. We call these Bellman Error Basis Functions (BEBFs).

Our main theoretical contribution is to show that BEBFs form an orthonormal basis with guaranteed improvement in approximation quality at each iteration. Since the Bellman error can be a quite complicated function that may not be any easier to represent than the true value-function, we consider the use of a Bellman error approximator to represent the new basis function. Our work is distinguished form earlier, similarly motivated work [2], in that we prove a general result showing that the approximation quality can still improve even if there is significant error in our estimate of the Bellman error.

2 Formal Framework and Notation

Due to space limitations, we assume that the reader is familiar with value-function approximation for MDPs and provide a very terse description of our notation. Our theory so far focuses on the case where the policy is fixed. Given a state s_i, the probability of a transition to a state s_j, is given by P_{ij} with expected reward of $R[s_i]$. We use P and R to stand for the matrix and column-vector representation of these terms. $V[s_i]$ is the expected total γ-discounted reward for state s_i. We define the Bellman operator T on value-functions as $(TV)[s_i] = R[s_i] + \gamma \sum_j P_{ij} V[s_j]$. It is a contraction in the weighted L_2 norm [5]: $\|V\|_\rho = \sqrt{\sum_{i=1}^{n} V[s_i]^2 \rho[s_i]}$, where ρ is the stationary distribution of P. Unless otherwise indicated, we will use $\|\cdot\|$ for $\|\cdot\|_\rho$.

A linear value-function approximator represents the value-function as a linear combination of features or basis functions: $\hat{V} = \sum_{i=1}^{k} w_i \phi_i$, where $\Phi = \{\phi_1 \ldots \phi_k\}$ is a set of linearly independent basis functions of the state, and $\mathbf{w} = \{w_1 \ldots w_k\}$ is a set of scalar weights. For a set of weights $\mathbf{w}$ expressed as a column vector, $\hat{V} = \Phi\mathbf{w}$.

Methods for finding reasonable $\mathbf{w}$ given Φ and a set of samples include linear TD [4], LSTD [1] and LSPE [6]. We refer to this family of methods as *linear fixed point* methods because they all solve for the same fixed point: $\hat{V} = \Phi\mathbf{w} = \Pi_\rho(R + \gamma P \Phi \mathbf{w})$, where Π_ρ is an operator that is the ρ-weighted L_2 projection into the span of Φ, that is, if $\Delta = diag(\rho)$, $\Pi_\rho = \Phi(\Phi^T\Delta^T\Delta\Phi)^{-1}\Phi^T\Delta^T\Delta$. We will use Π as shorthand for Π_ρ unless otherwise indicated. The closest point (in $\|\cdot\|_\rho$) in the span of Φ to V^* is ΠV^*, but the linear fixed point methods are not guaranteed to find this point. However, the distance from $\hat{V}$ to V^* can be bounded in terms of the distance from ΠV^* to V^* [5]:

$\|V^* - \hat{V}\| \leq \frac{1}{\sqrt{1-\kappa^2}}\|V^* - \Pi V^*\|$. The effective contraction rate κ arises from the combination of the Bellman operator, T, with contraction rate γ, and the L_2 projection. For our purposes, we conservatively assume $\kappa = \gamma$.

3 Feature Generation

We address the following question, also considered by others [3; 2]: Given a set of basis functions $\phi_1 \dots \phi_k$ and a linear fixed-point solution $\hat{V}$, what is a good ϕ_{k+1} to add to the basis?

The Bellman error is an intuitively appealing approach to expanding the basis since it is, loosely speaking, pointing towards V^*. We say that ϕ_{k+1} is a Bellman Error Basis Function (BEBF) for $\hat{V} = \Phi\mathbf{w}$ if $\phi_{k+1} = T\hat{V} - \hat{V}$. Constructing $\Phi' = [\Phi, \phi_{k+1}]$ (concatenating column vector ϕ_{k+1} to design matrix Φ) ensures that $T\hat{V}$ is in the span of Φ' (trivially by picking new weights $w'_i = w_i$ for $1 \leq i \leq k$, and $w'_{k+1} = 1$). While this formulation ensures that $\hat{V}$ can be represented, it leaves many open questions such as: (1) How does increasing the expressive power to include $T\hat{V}$ affect the fixed-point error bound?; (2) How does performance degrade if, due to the difficulty of representing ϕ_{k+1} exactly, $\widehat{\phi_{k+1}} \approx \phi_{k+1}$ is used instead? Our preliminary results:

Lemma 3.1 *Let $\hat{V}$ be a linear fixed-point solution using the basis $\Phi = \{\phi_1 \dots \phi_k\}$, then the BEBF $\phi' = T\hat{V} - \hat{V}$ is orthogonal to the span of Φ.*

Corollary 3.2 *A sequence of normalized BEBFs $\phi_1 \dots \phi_k$ forms an orthonormal basis.*

Corollary 3.3 *For a system with n states, V^* can be represented exactly using a sequence of no more than n BEBFs.*

Theorem 3.4 *Let $\hat{V}$ be the linear fixed-point solution using a sequence of normalized BEBFs $\phi_1 \dots \phi_k$. If $\|V^* - \hat{V}\| - \|V^* - T\hat{V}\| = x$, then for new BEBF ϕ_{k+1}, with $\Phi' = [\Phi, \phi_{k+1}]$, and corresponding Π', $\|V^* - \Pi V^*\| - \|V^* - \Pi' V^*\| \geq x$.*

This states that the bound tightening from the new basis function is *at least* as strong as value-iteration. In practice, we may be forced to use an approximate representation error for a BEBF [2]. For $\widehat{\phi_{k+1}} \approx \phi_{k+1}$, we can state some qualitative results. The first is that expanding the basis in the general direction of V^* ensures progress:

Lemma 3.5 *If $\widehat{\phi_{k+1}}$ is not orthogonal to $V^* - \hat{V}$, then there exists a β such that $\|V^* - (\hat{V} + \beta\widehat{\phi_{k+1}})\| < \|V^* - \hat{V}\|$. Moreover, if $\widehat{\phi_{k+1}}$ is not in the span of Φ, then for $\Pi' = \Pi \cup \widehat{\phi_{k+1}}$, $\|V^* - \Pi' V^*\| < \|V^* - \Pi V^*\|$.*

This lemma is encouraging, but the ease or difficulty in obtaining a $\widehat{\phi_{k+1}}$ that points towards V^* may not be obvious since the true direction of V^* typically isn't known until the problem is solved exactly. The angle between $\widehat{\phi_{k+1}}$ and ϕ_{k+1} provides a weaker, sufficient (though not necessary) condition for ensuring progress:

Theorem 3.6 *If (1) the angle between ϕ^{k+1} and $\widehat{\phi_{k+1}}$ is less than $cos^{-1}(\gamma)$ radians and (2) $\hat{V} \neq V^*$, then there exists a β such that $\|\hat{V} + \beta\widehat{\phi_{k+1}}\| < \|V^* - \hat{V}\|$. Moreover, if conditions (1) and (2) hold and $\widehat{\phi_{k+1}}$ is not in the span of Φ, then for $\Phi' = \Phi \cup \widehat{\phi_{k+1}}$, and corresponding Π', $\|V^* - \Pi' V^*\| < \|V^* - \Pi V^*\|$.*

References

[1] S. Bradtke and A. Barto. Linear least-squares algorithms for temporal difference learning. *Machine learning*, 2(1):33–58, January 1996.

[2] P. Keller, S. Mannor, and D. Precup. Automatic basis function construction for approximate dynamic programming and reinforcement learning. In *Proc. ICML*, 2006.

[3] S. Mahadevan and M. Maggioni. Value function approximation using diffusion wavelets and Laplacian eigenfunctions. In *Advances in Neural Information Processing Systems 19*, Cambridge, MA, 2006. MIT Press.

[4] R. S. Sutton. Learning to predict by the method of temporal differences. *Machine Learning*, 3(1):9–44, 1988.

[5] B. Van Roy. *Learning and Value Function Approximation in Complex Decision Processes*. PhD thesis, Massachusetts Institute of Technology, May 1998.

[6] H. Yu and D. Bertsekas. Convergence results for some temporal difference methods based on least squares. Technical Report LIDS-2697, Laboratory for Information and Decision Systems, MIT, 2006.