Proceedings of the
American Control Conference
Baltimore, Maryland • June 1994

Solving Bellman's Equation by the Method of Continuity

Michael Duff
Department of Computer Science
University of Massachusetts
Amherst, MA 01003
413–545–1596 duff@cs.umass.edu

Abstract

It is known [2] that policy iteration can be identified with Newton's method (and value iteration with successive approximation) for solving Bellman's optimality equation, which for Markov Decision problems takes the form: for $i = 1, N$

$$F(v(i)) = \max_{u \in \mathcal{U}} \left[\bar{r}_u(i) + \alpha \sum_{j=1}^{N} P_u(i,j) v(j) \right] - v(i) = 0. \quad (1)$$

One is naturally led to consider what new computational methods might be suggested by adopting alternative root-finding procedures. This paper summarizes an investigation of the consequences of adopting one particular root-finding scheme called the *method of continuity* (also known as the *method of imbedding* or *homotopy*).

Root-finding procedure	*Algorithmic conseqences for Bellman's equation*
Successive Approximation	Value Iteration
Newton-Kantorovitch	Policy Improvement
Continuation/Homotopy	?

1. Markov Decision Problems

Consider a system whose dynamics are described by a finite state Markov chain with transition matrix P, and suppose that at each time step, in addition to making a transition from state $x_t = i$ to $x_{t+1} = j$ with probability p_{ij}, the system produces a randomly determined reward, r_{t+1}, whose expected value is $\bar{r}(i)$. The *evaluation function*, v, maps states to their expected, infinite-horizon discounted returns:

$$v(i) = E\left\{ \sum_{t=0}^{\infty} \alpha^t r_{t+1} | x_0 = i \right\}.$$

In addition, for each state, suppose that there are a number of actions, $u \in \mathcal{U}$, from which to choose, and that choosing an action determines the transition probabilities and reward distribution associated with that state. A *policy* is a mapping of states to actions, and one may think of P above as the transition matrix associated with some particular policy. The *Markov Decision Problem* is to find a policy that optimizes the value function over all states. The *optimal* value function satisfies Bellman's optimality equation (1), and the optimal policy is greedy with respect to it.

2. The Method of Continuity and an Imbedding for Bellman's Equation

Suppose that the nonlinear equation to be solved is $F(x) = 0$. In the method of continuity [1], one constructs a function $G(x, \lambda)$, where λ is a real parameter, such that when $\lambda = \lambda_1$, $G(x, \lambda_1) = F(x)$, and when $\lambda = \lambda_0$, $G(x, \lambda_0)$ is an equation that is easily solved. The intention is that λ provide a *continuous* transition between the simpler, *base problem*, and the desired, *target problem*. Consider

$$G(x, \lambda) = 0, \quad (2)$$

and suppose we have a solution x_0 at λ_0; *i.e.*, $G(x_0, \lambda_0) = 0$. Our hope is that (1) implicitly defines a function $x(\lambda)$, $\lambda_0 \leq \lambda \leq \lambda_1$. $x(\lambda)$ is said to be the "branch" of (1) passing through (x_0, λ_0), and what we would like to do is follow the branch from (x_0, λ_0) to $x(\lambda_1)$, the solution of target problem The Implicit Function Theorem states that if ∇G is continuous and $\frac{\partial G}{\partial x}(x_0, \lambda_0) \neq 0$, then there will exist, for a small λ-neighborhood about λ_0, a branch through (x_0, λ_0) having the functional form $x = x(\lambda)$. In the method of continuity, we extend $x(\lambda)$ by solving a sequence of initial value problems obtained by differentiating (2) with respect to λ:

$$\frac{\partial G(x,\lambda)}{\partial x} \frac{\partial x}{\partial \lambda} + \frac{\partial G(x,\lambda)}{\partial \lambda} = 0.$$

If $\frac{\partial G}{\partial x} \neq 0$, then

$$\frac{\partial x}{\partial \lambda} = -\left[\frac{\partial G(x,\lambda)}{\partial x}\right]^{-1} \frac{\partial G(x,\lambda)}{\partial \lambda}. \qquad (3)$$

At $\lambda = \lambda_0$, $x = x_0$, so the right-hand-side of (3) is known, and therefore so is $\frac{\partial x(\lambda)}{\partial \lambda}|_{\lambda=\lambda_0}$. For $\Delta\lambda$ small,

$$x(\lambda_0 + \Delta\lambda) \cong x(\lambda_0) + \frac{\partial x(\lambda)}{\partial \lambda}|_{\lambda=\lambda_0}\Delta\lambda. \qquad (4)$$

We can use (3) again to obtain $\frac{\partial x(\lambda_0+\Delta\lambda)}{\partial \lambda}$ and (4) to find $x(\lambda_0 + 2\lambda)$, *etc.* In this way, we hope to construct the desired branch away from the initial point, signifying the solution to the base problem, and step along the branch toward the solution, $x(\lambda_1)$, to the target problem.

Recalling Bellman's optimality equation (1), consider the function derived by replacing the "max" operator by a variant of the L_p-norm, the "gereralized mean:"

$$G(v(i),\lambda) = \left(\frac{1}{M}\sum_{u\in\mathcal{U}}\left|\bar{r}_u(i) + \alpha\sum_{j=1}^{N} P_u(i,j)v(j)\right|^{\lambda}\right)^{\frac{1}{\lambda}} - v(i) = 0, \qquad (5)$$

and note that the target problem corresponds to $\lambda = \infty$. The entries for $\frac{\partial G(v,\lambda)}{\partial v}$ and $\frac{\partial G(v,\lambda)}{\partial \lambda}$ are

$$\frac{\partial G(i)}{\partial v(j)} = \frac{\alpha}{\beta}\left(\frac{\beta}{M}\right)^{\frac{1}{\lambda}}\sum_{k=1}^{M} P(i,j,k)Q(i,k)^{\lambda-1} - \delta_{ij}$$

$$\frac{\partial G(i)}{\partial \lambda} = \frac{1}{\lambda}\left(\frac{\beta}{M}\right)^{\frac{1}{\lambda}}\left[\frac{1}{\lambda}\log\frac{M}{\beta} + \frac{1}{\beta}\sum_{k=1}^{M} Q(i,k)^{\lambda}\log Q(i,k)\right]$$

where we have assumed that the set of possible states has size N, the set of possible actions has size M, and that

$$Q(i,k) \equiv \bar{r}(i,k) + \alpha\sum_{l=1}^{N} P(i,l,k)v(l)$$

$$\beta \equiv \sum_{k=1}^{M} Q(i,k)^{\lambda}.$$

The base problem, taking $\lambda = 1$ in equation (5) and assuming nonegative rewards, is to solve $(MI - \alpha\mathcal{P})v = \mathcal{R}$, where $\mathcal{P} \equiv \sum_u P_u$ and $\mathcal{R} \equiv \sum_u \bar{r}_u$. The existence of a solution for every $\alpha, 0 \leq \alpha < 1$ is guaranteed by utilizing results from the Perron-Frobenius theory for irreducible matrices, particularly the "Subinvariance Theorem" of [3].

Example: Consider the following Markov-decision problem: $N = M = 2 \qquad \alpha = .9$

$$P_{u_1} = \begin{bmatrix} .9 & .1 \\ .1 & .9 \end{bmatrix} \qquad \bar{r}_{u_1} = \begin{bmatrix} 1.1 \\ 1.9 \end{bmatrix}$$

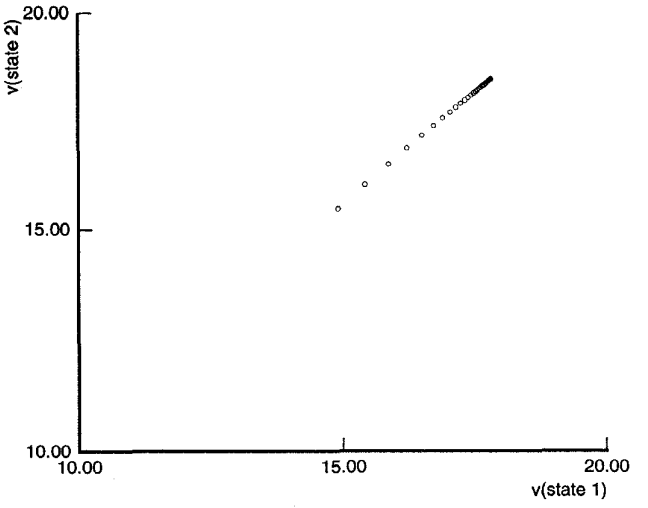

Figure 1: Branch constructed for the example.

$$P_{u_2} = \begin{bmatrix} .5 & .5 \\ .6 & .4 \end{bmatrix} \qquad \bar{r}_{u_2} = \begin{bmatrix} 1.5 \\ 1.4 \end{bmatrix}$$

The results of applying the method of continuity to this problem are shown in Figure 1, where the branch of solutions constructed by the method is shown projected onto the value function plane.

3. Discussion

Another look at Equation 4 suggests that the method of continuity may be viewed as solving the ordinary differential equation, $\frac{dx}{d\lambda} = f(x,\lambda)$ by Euler's method. One might consider conducting experiments with different DE-solvers in search of one requiring a reduced number of right-hand side evaluations while maintaining accuracy. As it is, at each step, $O(MN^2)$ multiplications are required to compute all the entries of $\frac{\partial G}{\partial v}$ and $\frac{\partial G}{\partial \lambda}$, and solving the linear system, $\frac{\partial G}{\partial x}\frac{\partial x}{\partial \lambda} = -\frac{\partial G}{\partial \lambda}$, requires $O(N^3)$ operations. So the complexity of the method as a whole is $O(N^3)$ per step, which is similar to policy iteration, which must also solve a system of linear equations at each step. For further developments and the details, contact the author.

References

[1] J. M. Ortega and W. Rheinboldt *Iterative Solution of Nonlinear Equations in Several Variables*, 1970.

[2] M. L. Puterman and S. L. Brumelle, "The Analytic Theory of Policy Iteration,"in *Dynamic Programming and Its Applications*, M.L. Puterman ed., 1978.

[3] E. Seneta, *Non-Negative Matrices*, 1973.