A Comparison of Policy Iteration Methods for Solving Continuous-State, Infinite-Horizon MDPs using Random, Quasi-Random and Deterministic Discretizations

John Rust, *Yale University*

For presentation at CESG September, 1997

Goal: to compare the performance of alternative random and deterministic *discrete approximation methods* for solving infinite horizon, continuous-state MDPs.

Approach: I compare the accuracy of different methods for discretizing MDPs by computing the difference between the exact and approximate solutions for two "test problems" with closed-form solutions: 1) a replacement problem and 2) a job search problem.

Methods Compared: Each discretization generates a finite subset of points in the state space. *Policy iteration* is used to compute an exact solution to the embedded finite state MDP defined on this grid. I consider four different discretizations or *grids:*

1. **Uniform grids**
2. **Quadrature grids**
3. **Random grids**
4. **"Low Discrepancy" grids**

Outline of Talk:

1. Introduction
2. Theoretical Bounds on Rates of Convergence
3. Numerical Results for 1-Dimensional Test Problems
4. Numerical Results for 2-Dimensional Test Problems

Conclusions:

1. Randomization does indeed break the curse of dimensionality: performing much better than expected in low-dimensional problems.
2. Deterministic "low discrepancy" methods work even better: they break the curse of dimensionality (for fixed MDP problems) and converge even faster than methods based on "random" grids.
3. Performance of all methods seriously degrade when there are large discontinuities in the integrand. In these cases all methods require large, dense grids to avoid the *needle in the haystack problem.*
4. In 'worst case" test problems with sufficiently large discontinuities the curse of dimensionality can reappear: the "severity" of the discontinuities can increase exponentially fast as the problem dimension d increases.

Discrete Approximation of Bellman Operators

1. *Bellman operator:*

$$\Gamma(V)(s) \equiv \max_{a \in A(s)} [u(s,a) + \beta \int V(s')p(s'|s,a)ds']. \tag{0.1}$$

2. *Discretized Bellman operator:*

$$\Gamma_N(V)(s) \equiv \max_{a \in A(s)} \left[u(s,a) + \beta \sum_{i=1}^{N} V(s_i) p_N(s_i|s,a) \right], \tag{0.2}$$

where: p_N is a discrete probability distribution over a finite *grid* $\{s_1, \ldots, s_N\}$ in S, where $S = [0,1]^d$ for simplicity. In that case $\Gamma : B \to B$ where B is the (infinite-dimensional) Banach space of all bounded, measurable functions from $[0,1]^d \to R$.

Approximation Strategy: Compute $V_N = \Gamma_N(V_N)$, where V_N is the value function for the embedded discrete MDP with state space $\{s_1, \ldots, s_N\}$. Thus $V_N \in R^N$. Extend V_N to a function defined over all of S using some approximation/interpolation method. Let $V_N \in B$ also denote the extended value function: then the maximum error over S, $\|V - V_N\|$, satisfies the following inequality:

$$\|V - V_N\| \leq \frac{\|\Gamma(V) - \Gamma_N(V)\|}{(1-\beta)}. \tag{0.3}$$

Policy Iteration (Infinite-Dimensional Version)

1. **Policy Valuation Step:** given an initial guess of policy α compute V_α, the value function implied by policy α:

$$V_\alpha(s) = u(s, \alpha(s)) + \beta E_\alpha V_\alpha(s),$$

where E_α is the *Markov operator* corresponding to α:

$$E_\alpha V(s) = \int V(s')p(s'|s, \alpha(s))ds'.$$

There is a unique solution to the linear operator equation defining V_α (Fredholm integral equation of the second kind):

$$V_\alpha = u_\alpha + \beta E_\alpha V = (I - \beta E_\alpha)^{-1} u_\alpha,$$

where $(I - \beta E_\alpha)$ exists and has the following geometric series or *Neumann series* representation:

$$(I - \beta E_\alpha)^{-1} = \sum_{t=0}^{\infty} [\beta E_\alpha]^t.$$

2. **Policy Improvement Step:** Compute improved policy α' using V_α:

$$\alpha'(s) = \underset{a \in A(s)}{argmax}[u(s, a) + \beta \int V_\alpha(s')p(s'|s, a)ds'].$$

Policy Iteration (Finite-Dimensional Version)

1. **Policy Valuation Step:** given an initial guess of policy α compute $V_{\alpha,N}$, the value function (in R^N) implied by policy α:

$$V_{\alpha,N}(s) = u(s, \alpha(s)) + \beta E_{\alpha,N} V_{\alpha,N}(s),$$

where $E_{\alpha,N}$ is the *discrete Markov operator* corresponding to α:

$$E_{\alpha,N} V(s) = \sum_{i=1}^{N} V(s_i) p_N(s_i | s, \alpha(s)).$$

There is a unique solution to the linear system of equations defining $V_{\alpha,N} \in R^N$

$$V_{\alpha,N} = u_\alpha + \beta E_{\alpha,N} V_{\alpha,N} = (I - \beta E_{\alpha,N})^{-1} u_\alpha,$$

where $(I - \beta E_{\alpha,n})$ exists and has the following geometric series representation:

$$(I - \beta E_{\alpha,N})^{-1} = \sum_{t=0}^{\infty} [\beta E_{\alpha,N}]^t.$$

where I is the $N \times N$ identity matrix and $E_{\alpha,N}$ is the $N \times N$ Markov transition matrix with (i, j) entry given by:

$$E_{\alpha,N}[i, j] = [p_N(s_i | s_j, \alpha(s_j))].$$

2. **Policy Improvement Step:** Compute improved policy α' using V_α:

$$\alpha'(s) = \underset{a \in A(s)}{argmax} [u(s, a) + \beta \sum_{i=1}^{N} V_{\alpha,N}(s_i) p_N(s_i | s, a)].$$

Strategies for extending $V_N \in R^N$ to a $V_N \in B$:

1. **Use your favorite approximation method.** There are a vast number of different methods including including *series regressions, kernel regressions,* and *spline approximations.* A spline is a local-polynomial approximation defined over a set of *knot points* that define a *triangulation* of S. In this paper I use a 1-spline or *linear interpolation* to extend V_N from R^N to B. This yields approximations of the form:

$$V_N(s) = \sum_{i=1}^{N} \mu_i V_N(s_i) \text{ where } s = \sum_{i=1}^{N} \mu_i s_i \text{ and } \sum_{i=1}^{N} \mu_i = 1,\ \mu_i \geq 0, \tag{0.4}$$

   where $V_N(s_i)$ denotes the i^{th} component of the vector V_N.

2. **Use Nyström's Method.** This is a *self-approximating* method whenever $p_N(s_i|s,a)$ is defined for all $s \in S$ and $a \in A$. In that case it is easy to see that Γ_N in equation (0.2) has a dual interpretation:

   a. Γ_N is a contraction mapping on R^N

   b. Γ_N is a contraction mapping on B.

Strategy for deriving error bounds: Inequality (0.3) gives us error bounds for $||V - V_N||$ in terms of an error bound for $||\Gamma_N(V) - \Gamma(V)||$. Need to derive bound for latter.

1. If you an auxiliary approximation method to extend V_N from R^N to B, then can also extend $\Gamma_N(V)$, but *the extension of* V_N *from* R^N *to* B *need not coincide with the fixed point of the extension of* Γ_N *from* R^N *to* B. In this case to derive an error bound for $||V - V_N||$ on B we need to derive an error bound for $||V - V_N||$ on R^N, where V and $\Gamma(V)$ now denote the *restriction* of V and $\Gamma(V)$ in B to their values on the embedded MDP at the grid points $\{s_1, \ldots, s_N\}$. The error bound for the chosen approximation method must account for *noisy information* on the function being approximated. In this case V_N is treated as a noisy estimate of V of the form $V_N = V + \eta_N$ where η_N is the $N \times 1$ vector of approximation errors.

2. Nyström's Method yields a much more direct bound since the self-approximation property leads directly to bounds on $||\Gamma_N(V) - \Gamma(V)||$ where both Γ_N and Γ are treated as contraction mappings on B. We can derive on bound on $||\Gamma_N(V) - \Gamma(V)||$ via error bounds for discrete approximation of *Markov operators.*

Discrete Approximation of Markov Operators

1. *Markov operator:*

$$E_\alpha V(s) = \int V(s')p(s'|s,\alpha(s))ds'. \tag{0.5}$$

2. *Discretized Markov operator:*

$$E_{\alpha,N}V(s) = \sum_{i=1}^{N} V(s_i)p_N(s_i|s,\alpha(s)) \tag{0.6}$$

Bounding $\|V - V_N\|$ in terms of error bounds for Markov Operators

$$\|V - V_N\| \leq \|\Gamma(V) - \Gamma_N(V)\| \leq \frac{max_{a\in A}\,\|E_a(V) - E_{a,N}(V)\|}{(1-\beta)}. \tag{0.7}$$

To derive bounds on $\|E_a(V) - E_{a,N}(V)\|$ we need to specify p_N:

1. **Uniform grids:** Use *exact integration* of partition elements containing grid points. For example in 1-dimensional case we divide $S = [0,1]$ into N equal intervals of length $1/N$ with grid points located at the middle of each interval:

$$\begin{aligned} s_i &= \frac{1}{2N} + \frac{(i-1)}{N} \quad i = 1, \ldots, N \\ p_N(s_i|s,a) &= \int_{\frac{(i-1)}{N}}^{\frac{i}{N}} p(s'|s,a)ds' \quad s \in \left(\frac{(i-1)}{N}, \frac{i}{N}\right], \end{aligned} \tag{0.8}$$

2. **Quadrature grids:** Use *quadrature weights and abscissae* to construct p_N:

$$p_N(s_k|s,a) = \frac{w_k p(s_k|s,a)}{\Sigma_{i=1}^N w_i p(s_i|s,a)}, \tag{0.9}$$

3. **Random and Low Discrepancy grids:** Use grid points and equal weights $w_k = 1$ to construct p_N:

$$p_N(s_k|s,a) = \frac{p(s_k|s,a)}{\Sigma_{i=1}^N p(s_i|s,a)}, \tag{0.10}$$

Strategies for bounding $\|E_a(V) - E_{a,N}(V)\|$

Uniform grids: Use fact that $E_{a,N}$ can be represented as:

$$E_{a,N}(V) = \frac{1}{N}\sum_{i=1}^{N} V(s_i)p_N(s_i|s,a) = \int \hat{V}_N(s')p(s'|s,a)ds', \quad (0.11)$$

where $\hat{V}_N$ is a step function approximation to V satisfying $\hat{V}_N(s) = V(s_i)$ whenever s is in the partition element with center at grid point s_i. Then we have:

$$\|E_a(V) - E_{a,N}(V)\| \leq \int \|V - \hat{V}_N\| p(s'|s,a)ds' \leq \|V - \hat{V}_N\|, \quad (0.12)$$

If V is Lipschitz continuous with Lipschitz bound K then it is straightforward to show that

$$\|V - \hat{V}_N\| \leq \frac{K\sqrt{d}}{N^{\frac{1}{d}}}. \quad (0.13)$$

This implies:

Theorem 2.2: *Let V_N be a 0-spline or step function approximation to V computed from an embedded N-state MDP defined on a uniform grid in $[0,1]^d$, i.e. $V_N(s) = V_N(s_i)$ if s is in the i^{th} partition element of $[0,1]^d$ centered at s_i. Then we have:*

$$\|V - V_N\| = O\left(\frac{1}{N^{\frac{1}{d}}}\right). \quad (0.14)$$

Quadrature grids: Use *Haber's Inequality* for multidimensional quadrature to show:

Theorem 2.3 *Let V_N be an approximate value function computed from a product rule quadrature grid using weights which are products of the corresponding unidimensional quadrature weights and transition probability p_N given in (0.9). Then we have:*

$$\|V - V_N\| = O\left(\frac{1}{N^{\frac{1}{d}}}\right). \tag{0.15}$$

Random grids: Use *Maximal Inequality for empirical processes* due to Pollard (1989) to show:

Theorem 3.3 (Rust 1997) *Let p be Lipschitz continuous in its first argument with Lipschitz bound K_p. Then:*

$$E\left\{\|E_a(V) - \tilde{E}_{a,N}\|\right\} \le \sqrt{\frac{\pi}{2}}\frac{[1 + d\sqrt{\pi}C]K_p\|V\|}{\sqrt{N}}, \tag{0.16}$$

where K_p is a Lipschitz bound on $p(s'|s,a)$, and C is a universal constant independent of the problem or dimension.

Low Discrepancy grids: Use *Koksma-Hlawka inequality* to show:

Theorem 2.1: *Under the conditions of Lemma 2.1, it follows that the approximate value function V_N will be a consistent estimator of V (i.e. $\|V - V_N\| \to 0$ as $N \to \infty$), if the grid $\{s_1, \ldots, s_N\}$ is uniformly distributed. Furthermore we have:*

$$\|V - V_N\| = O\left(D_N^*(\{s_N\})\right). \tag{0.17}$$

where D_N^* is the *discrepancy:*

$$D_N^*(s_1, \ldots, s_N) \equiv \underset{B \in \mathcal{B}}{sup} \left|\lambda_N(B) - \lambda(B)\right|. \tag{0.18}$$

Definition: *A* **Low Discrepancy Grid** $\{s_1, \ldots, s_N\}$ *is any sequence of points in* $[0, 1]^d$ *whose discrepancy satisfies:*

$$D_N^*(s_1, \ldots, s_N) = O\left(\frac{(log\, N)^d}{N}\right). \tag{0.19}$$

Comment: A number of different sequences have discrepancies that satisfy the bound in (0.19), including the *Hammersley, Halton, Lapeyre-Pagés, Faure* and *Sobol'* sequences. Many of these sequences have deep connections to number theory and it is beyond the scope of this paper to go into details about how these sequences are generated and why they satisfy the bound (0.19).

Theorem (Roth) *The discrepancy satisfies the lower bound:*

$$D_N^*(s_1, \ldots, s_N) = \Omega\left(\frac{(log\, N)^{(d-1)/2}}{N}\right). \tag{0.20}$$

Note: When $d = 1$ it is easy to show that uniform grid has minimal discrepancy with $D_N^*(s_1, \ldots, s_N) = 1/2N$.

Koksma-Hlawka Inequality

$$\left|\frac{1}{N}\sum_{i=1}^{N} f(s_i) - \int f(s)\lambda(ds)\right| \leq \mathcal{V}(f)D_N^*(s_1, \ldots, s_N), \tag{0.21}$$

where $\mathcal{V}(f)$ is the *variation* of f (see p. 67 of Bouleau and Lépingle 1994 for a definition).

Definition: Any sequence $\{s_N\}$ that satisfies $D_N^*(\{s_N\}) \to 0$ as $N \to \infty$ is said to be *uniformly distributed.*

Comment: The Koksma-Hlawka inequality implies that if the sequence $\{s_N\}$ is uniformly distributed, the $\lambda_N \underset{d}{\Longrightarrow} \lambda$, i.e. the empirical CDF for $\{s_N\}$ converges in distribution to Lebesgue measure on $[0,1]^d$.

It follows immediately that if f is any bounded function which is continuous except on a set of Lebesgue measure 0 in $[0,1]^d$ that

$$\underset{N\to\infty}{lim} \frac{1}{N} \sum_{i=1}^{N} f(s_i) = \int f(s)ds. \tag{0.22}$$

Comment: For smoother functions, say if $f : [0,1]^d \to R$ is in C^d (i.e. is d-times continuously differentiable), then one can derive a more explicit formula for the variation of f:

$$\mathcal{V}(f) = \left\| \frac{\partial^d f}{\partial s_1 \cdots \partial s_d} \right\|_1 = \int \cdots \int \left| \frac{\partial^d f(s_1, \ldots, s_d)}{\partial s_1 \cdots \partial s_d} \right| ds_1 \cdots ds_d. \tag{0.23}$$

Summary of Theoretical Rates of Convergence:

$$\|V - V_N\| = \begin{cases} O\left(\frac{1}{N^{\frac{1}{d}}}\right) & \text{for uniform grids} \\ O\left(\frac{1}{N^{\frac{1}{d}}}\right) & \text{for quadrature grids} \\ O\left(\frac{1}{\sqrt{N}}\right) & \text{for random grids} \\ O\left(\frac{log(N)^d}{N}\right) & \text{for low discrepancy grids.} \end{cases} \tag{0.24}$$

Test Problems Used in Numerical Comparisons

1. Optimal Replacement Problem Consider a durable whose state $s_t \in R_+$ can be interpreted as a measure of the accumulated utilization (such as the odometer reading on a car). Thus $s_t = 0$ denotes a brand new durable good. At each time t there are two possible decisions {keep,replace} corresponding to the binary constraint set $A(s) = \{0, 1\}$ where $a_t = 1$ corresponds to selling the existing durable for scrap price $\underline{P}$ and replacing it with a new durable at cost $\overline{P}$.

$$p(ds_{t+1}|s_t, a_t) = \begin{cases} 1 - exp\{-\lambda(ds_{t+1} - s_t)\} & \text{if } a_t = 0 \text{ and } s_{t+1} \geq s_t \\ 1 - exp\{-\lambda(ds_{t+1} - 0)\} & \text{if } a_t = 1 \text{ and } s_{t+1} \geq 0 \\ 0 & \textit{otherwise.} \end{cases} \tag{0.25}$$

$$u(s_t, a_t) = \begin{cases} -\, c(s_t) & \text{if } a_t = 0 \\ -\,[\,\overline{P} - \underline{P}\,] - c(0) & \text{if } a_t = 1. \end{cases} \tag{0.26}$$

Bellman's equation takes the form:

$$\begin{aligned} V(s) = max[\, &-\, c(s) + \beta \int_s^\infty V(s')\lambda\, exp\{-\lambda(s' - s)\}ds', \\ &-\,[\,\overline{P} - \underline{P}\,] - c(0) + \beta \int_0^\infty V(s')\lambda\, exp\{-\lambda(s')\}ds']. \end{aligned} \tag{0.27}$$

Differentiate Bellman's equation to derive a differential equation for V. The differential equation is a *free boundary value problem* since the boundary condition:

$$V(\gamma) = [\,\overline{P} - \underline{P}\,] + V(0) = -c(\gamma) + \beta V(\gamma) = \frac{-c(\gamma)}{1-\beta}, \tag{0.28}$$

is determined endogenously. Rust showed this differential equation has the following closed-form solution:

$$V(s) = max[\frac{-c(\gamma)}{1-\beta}, \frac{-c(\gamma)}{1-\beta} + \int_s^\gamma \frac{c'(y)}{1-\beta}[1 - \beta e^{-\lambda(1-\beta)(y-s)}]dy], \tag{0.29}$$

where γ is the unique solution to:

$$[\overline{P} - \underline{P}] = \int_0^{\gamma} \frac{c'(y)}{1-\beta}[1 - \beta e^{-\lambda(1-\beta)y}]dy. \tag{0.30}$$

It follows that the optimal decision rule is given by:

$$\alpha(s) = \begin{cases} 0 & \text{if } s \in [0, \gamma] \\ 1 & \text{if } s > \gamma. \end{cases} \tag{0.31}$$

Comment: Notice that the transition density $p(s'|s, a)$ has a discontinuity at $s' = s$ when $s > 0$: this discontinuity arises from an "irreversibility" condition that except for replacement (which constitutes complete regeneration) the durable can only deteriorate with age.

2. Optimal Search Problem Consider an unemployed worker who receives one job offer each period while searching, a random draw from a density $f(w)$. If the worker accepts a wage offer w, the search problem terminates and the worker is assumed to be employed forever after at wage w. If the worker reject a wage offer w the worker receives an unemployment benefit of u and continuous searching in the next period. Bellman's equation for this problem is give by:

$$V(w) = max\left[\frac{w}{(1-\beta)}, u + \beta \int_0^\infty V(w')f(w')dw'\right]. \tag{0.32}$$

As is well known, the optimal search strategy involves accepting any offer greater than the *reservation wage* $\overline{w}$ given by:

$$\overline{w} = \frac{(1-\beta)\left[u + \beta \int_{\overline{w}}^\infty w' f(w')dw'\right]}{1 - F(\overline{w})}. \tag{0.33}$$

Given $\overline{w}$, the value function is given by

$$V(w) = \begin{cases} \dfrac{w}{(1-\beta)} & \text{if } w \geq \overline{w} \\ \dfrac{\overline{w}}{(1-\beta)} & \text{if } w < \overline{w}. \end{cases} \tag{0.34}$$

Comment: Notice that the transition probability density $p(s'|s,a)$ for the search problem does not exist when the decision is made to stop searching: in this case $p(s'|s,a)$ is a unit probability mass (also known as a dirac delta function) on the accepted wage offer w.

Generating Multidimensional Test Problems Consider a multidimensional MDP problem with the following additive-separable structure:

$$u(\mathbf{s}, \mathbf{a}) = \sum_{i=1}^{N} u_i(s_i, a_i) \tag{0.35}$$

$$p(\mathbf{s'}|\mathbf{s}, \mathbf{a}) = \prod_{i=1}^{N} p_i(s_i'|s_i, a_i) \tag{0.36}$$

$$A(\mathbf{s}) = A_1(s_1) \times A_2(s_2) \cdots A_N(s_N) \tag{0.37}$$

Comment: Each s_i can be thought of as representing the state of "task" i. Equation (0.36) states that task states evolve independently. For example, the task of "driving to work by automobile" may be separable from task of "debugging a computer program at work", and the task of "deciding where to go for lunch", and so forth.

Multidimensional Bellman Equation:

$$V(\mathbf{s}) = \max_{\mathbf{a} \in A(\mathbf{s})} \left[u(\mathbf{s}, \mathbf{a}) + \beta \int_{\mathbf{s'}} V(\mathbf{s'}) p(\mathbf{ds'}|\mathbf{s}, \mathbf{a}) \right]. \tag{0.38}$$

Lemma 3.1: If $u(\mathbf{s}, \mathbf{a})$ has the additive structure given in equation (0.35) and $p(\mathbf{s'}|\mathbf{s}, \mathbf{a})$ satisfies the independence decomposition in (0.36), and the action sets $A(\mathbf{s})$ have the product structure given in (0.37), then the value function $V(\mathbf{s})$ has an additive decomposition given by:

$$V(\mathbf{s}) = \sum_{i=1}^{N} V_i(s_i). \tag{0.39}$$

where each V_i is given by:

$$V_i(s_i) = \max_{a_i \in A_i(s_i)} \left[u_i(s_i, a_i) + \beta \int_{s_i'} V_i(s_i') p_i(ds_i'|s_i, a_i) \right]. \tag{0.40}$$