July 1995 LIDS-P-2307

A VALUE ITERATION METHOD FOR THE [1] AVERAGE COST DYNAMIC PROGRAMMING PROBLEM

by

Dimitri P. Bertsekas[2]

Abstract

We propose a new value iteration method for the classical average cost Markovian Decision problem, under the assumption that all stationary policies are unichain and furthermore there exists a state that is recurrent under all stationary policies. This method is motivated by a relation between the average cost problem and an associated stochastic shortest path problem.

[1] Research supported by NSF under Grant 9300494-DMI.

[2] Dept. of Electrical Engineering and Computer Science, M.I.T., Cambridge, Mass., 02139.

1. INTRODUCTION

We consider a controlled discrete-time dynamic system with n states, denoted $1, \ldots, n$. At each time, if the state is i, a control u is chosen from a given finite constraint set $U(i)$, and the next state is j with given probability $p_{ij}(u)$. An admissible policy is a sequence of functions from states to controls, $\pi = \{\mu_0, \mu_1, \ldots\}$, where $\mu_k(i) \in U(i)$ for all i and k. The average cost corresponding to π and initial state i is

$$J_\pi(i) = \limsup_{N\to\infty} \frac{1}{N} E\left\{ \sum_{k=0}^{N-1} g\big(x_k, \mu_k(x_k)\big) \,\middle|\, x_0 = i \right\},$$

where x_k is the state at time k, and g is a given cost function. A *stationary policy* is an admissible policy of the form $\pi = \{\mu, \mu, \ldots\}$, and its corresponding cost function is denoted by $J_\mu(i)$. For brevity, we refer to $\{\mu, \mu, \ldots\}$ as the stationary policy μ. We want to solve the classical problem of finding an optimal policy, that is, an admissible policy π such that $J_{\pi^*}(i) = \min_\pi J_\pi(i)$ for all i.

A stationary policy is called *unichain* if it gives rise to a Markov chain with a single recurrent class. Throughout the paper, we assume the following:

Assumption 1: All stationary policies are unichain. Furthermore, state n is recurrent in the Markov chain corresponding to each stationary policy.

It is well known that under Assumption 1, the optimal cost $J^*(i)$ has a common value for all initial states, which is denoted by λ^*,

$$J^*(i) = \lambda^*, \qquad i = 1, \ldots, n.$$

Furthermore, λ^* together with a differential cost vector $h = \big(h(1), \ldots, h(n)\big)$ satisfies Bellman's equation

$$\lambda^* + h(i) = \min_{u\in U(i)} \left[g(i,u) + \sum_{j=1}^{n} p_{ij}(u) h(j) \right], \qquad i = 1, \ldots, n. \tag{1}$$

In addition, a stationary policy μ is optimal if and only if $\mu(i)$ attains the minimum in the above equation for all i. These results can be shown under the assumption that all stationary policies are unichain, without requiring the additional condition that there is a common recurrent state to all stationary policies. However, for the methods of this paper, the existence of a common recurrent state is essential.

Under Assumption 1 we can make an important connection of the average cost problem with an associated stochastic shortest path problem, which has been the basis for a recent textbook

treatment of the average cost problem ([Ber95], Vol. I, Section 7.4). This problem is obtained by leaving unchanged all transition probabilities $p_{ij}(u)$ for $j \neq n$, by setting all transition probabilities $p_{in}(u)$ to 0, and by introducing an artificial cost-free and absorbing termination state t to which we move from each state i with probability $p_{in}(u)$; see Fig. 1. The expected stage cost at state i of the stochastic shortest path problem is $g(i,u) - \lambda$, where λ is a scalar parameter. Let $h_{\mu,\lambda}(i)$ be the cost of stationary policy μ for this stochastic shortest path problem, starting from state i; that is, $h_{\mu,\lambda}(i)$ is the total expected cost incurred starting from state i up to reaching the termination state t. We refer to this problem as λ-SSP. Let $h_\lambda(i) = \min_\mu h_{\mu,\lambda}(i)$ be the corresponding optimal cost of the λ-SSP. Then the following can be shown (see Fig. 2):

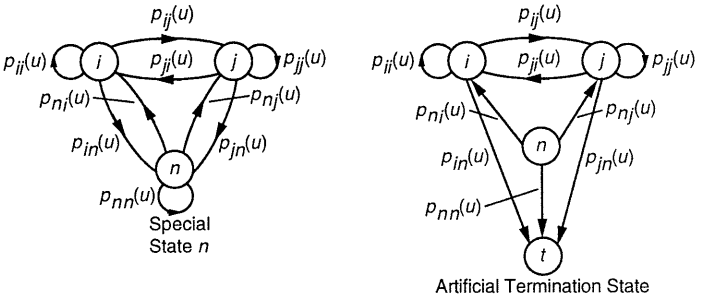

Figure 1. Transition probabilities for an average cost problem and its associated stochastic shortest path problem. The latter problem is obtained by introducing, in addition to $1,\ldots,n$, an artificial termination state t to which we move from each state i with probability $p_{in}(u)$, by setting all transition probabilities $p_{in}(u)$ to 0, and by leaving unchanged all other transition probabilities.

(a) For all μ and λ, we have

$$h_{\mu,\lambda}(i) = h_{\mu,\lambda_\mu}(i) + (\lambda_\mu - \lambda)N_\mu(i), \qquad i = 1,\ldots,n, \tag{2}$$

where $N_\mu(i)$ is the average number of steps required to reach n under μ starting from state i.

(b) The functions

$$h_\lambda(i) = \min_\mu h_{\mu,\lambda}(i), \qquad i = 1,\ldots,n, \tag{3}$$

are concave, monotonically decreasing, and piecewise linear as functions of λ, and

$$h_\lambda(n) = 0 \qquad \text{if and only if} \qquad \lambda = \lambda^*. \tag{4}$$

Furthermore, the vector h_{λ^*} satisfies Bellman's equation (1) together with λ^*.

From Fig. 2, it can be seen that λ^* can be obtained by a one-dimensional search procedure that brackets λ^* within a sequence of nested and diminishing intervals; see [Ber95], Vol. II, Fig.

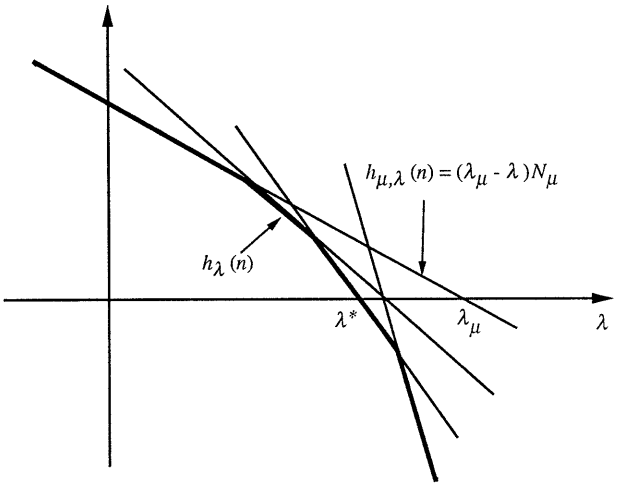

Figure 2. Relation of the costs of stationary policies in the average cost problem and the associated stochastic shortest path problem.

4.5.2. This method requires the (exact) solution of several λ-SSPs, corresponding to several different values of λ. An alternative method, that also requires the exact solution of several λ-SSPs is to update λ by an iteration of the form

$$\lambda^{k+1} = \lambda^k + \gamma^k h_{\lambda^k}(n), \tag{5}$$

where γ^k is a positive stepsize parameter. This iteration is motivated by Fig. 2 where it is seen that $\lambda < \lambda^*$ (or $\lambda > \lambda^*$) if and only if $h_\lambda(n) > 0$ [or $h_\lambda(n) < 0$, respectively]. Indeed, it can be seen from Fig. 2 that the sequence $\{\lambda^k\}$ thus generated converges to λ^* provided the stepsize γ^k is the same for all iterations and does not exceed the threshold value $1/\max_\mu N_\mu(n)$. Such a stepsize is sufficiently small to guarantee that the difference $\lambda - \lambda^*$ does not change sign during the algorithm (5). Note that each λ-SSP can be solved by value iteration, which has the form

$$h^{k+1}(i) = \min_{u \in U(i)} \left[g(i,u) + \sum_{j=1}^{n-1} p_{ij}(u) h^k(j) \right] - \lambda, \qquad i = 1, \ldots, n, \tag{6}$$

with λ kept fixed throughout the value iteration method.

In this paper we propose to change λ during the preceding value iteration process by using an iteration of the form (5), but with $h_{\lambda^k}(n)$ replaced by an approximation, the current value iterate $h^{k+1}(n)$. Such an algorithm may be viewed as a *value iteration algorithm for a slowly varying stochastic shortest path problem.* It has the form

$$h^{k+1}(i) = \min_{u \in U(i)} \left[g(i,u) + \sum_{j=1}^{n-1} p_{ij}(u) h^k(j) \right] - \lambda^k, \qquad i = 1, \ldots, n, \tag{7}$$

$$\lambda^{k+1} = \lambda^k + \gamma^k h^{k+1}(n), \tag{8}$$

where γ^k is a positive stepsize. We prove convergence of this method for the case where γ^k is a sufficiently small constant. Convergence can also be similarly proved for a variety of other stepsize rules.

Our method should be contrasted with the standard relative value iteration method for average cost problems due to [Whi63], which takes the form (see e.g., [Ber95], [Put94])

$$\lambda^{k+1} = \min_{u \in U(n)} \left[g(n,u) + \sum_{j=1}^{n-1} p_{nj}(u) h^k(j) \right], \tag{9}$$

$$h^{k+1}(i) = \min_{u \in U(i)} \left[g(i,u) + \sum_{j=1}^{n-1} p_{ij}(u) h^k(j) \right] - \lambda^{k+1}, \qquad i = 1, \ldots, n. \tag{10}$$

If we use Eq. (7) to write iteration (8) in the equivalent form

$$\lambda^{k+1} = (1-\gamma^k)\lambda^k + \gamma^k \min_{u \in U(n)} \left[g(n,u) + \sum_{j=1}^{n-1} p_{nj}(u) h^k(j) \right],$$

we see that if $\gamma^k = 1$ for all k, the new value iteration (7)-(8) becomes similar to the known value iteration (9)-(10): the updating formulas are the same in both methods, but the order of updating λ is just reversed relatively to the order of updating h. We note that there is also a variant of the standard method (9)-(10) that involves interpolations between h^k and h^{k+1} according to a stepsize parameter (see [Sch71], [Pla77], [Var78], [PBW79], [Put94], [Ber95]). However, the new method does not seem as closely related to this variant. Despite the similarity of the new method with the standard method (9)-(10), the proof of convergence of the latter method does not seem to be applicable to the new method. The line of proof given in the next section is substantially different, and makes essential use of Assumption 1 and the connection with the stochastic shortest path problem. In particular, one can construct examples where Assumption 1 is violated because state n is transient under some stationary policy, and where the new method (7)-(8) does not converge while the known method (9)-(10) converges. It can also be seen that the standard aperiodicity assumption required for convergence of the known method (9)-(10) (see e.g., [Ber95], [Put94]) is not needed for the new method.

A significant improvement in the algorithm, which guarantees that bounded iterates will be generated for any choice of stepsize, is to calculate upper and lower bounds on λ^* from iteration (7) and then modify iteration (8) to project the iterate $\lambda^k + \gamma^k h^k(n)$ on the interval of the bounds. In particular, based on the Odoni bounds [Odo69] for the relative value iteration method, it can be seen that

$$\underline{\beta}^k \leq \lambda^* \leq \overline{\beta}^k,$$

where

$$\underline{\beta}^k = \lambda^k + \min\left[\min_{i\neq n}\left[h^{k+1}(i) - h^k(i)\right],\, h^{k+1}(n)\right], \tag{11}$$

$$\overline{\beta}^k = \lambda^k + \max\left[\max_{i\neq n}\left[h^{k+1}(i) - h^k(i)\right],\, h^{k+1}(n)\right]. \tag{12}$$

Thus we may replace the iteration $\lambda^{k+1} = \lambda^k + \gamma^k h^{k+1}(n)$ [cf. Eq. (8)] by

$$\lambda^{k+1} = \left[\lambda^k + \gamma^k h^{k+1}(n)\right]^+, \tag{13}$$

where $[c]^+$ denotes the projection of a scalar c on the interval

$$\left[\max_{m=0,\ldots,k}\underline{\beta}^m,\ \min_{m=0,\ldots,k}\overline{\beta}^m\right]. \tag{14}$$

We note that the issue of stepsize selection is crucial for the success of our algorithm. In particular, if γ^k is chosen constant but very small, or diminishing at the rate of $1/k$ (as is common in stochastic approximation algorithms), then λ changes slowly relative to h, and the iteration (8) essentially becomes identical to iteration (5) but with a very small stepsize, which leads to slow convergence. On the other hand, if γ^k is too large, λ^k will oscillate and diverge. One may keep the stepsize γ^k constant at a value found by trial and error, but there are some better alternatives. One possibility that has worked reliably and efficiently in our tests is to start with a fairly large γ^k and gradually diminish it if the value $h^k(n)$ changes sign frequently; for example, we may use

$$\gamma^k = \frac{\gamma}{m^k}, \tag{15}$$

where m^k is equal to one plus the number of times that $h(n)$ has changed sign up to iteration k, and γ is the initial stepsize (a positive constant). Our experience has been that it is best to choose the initial stepsize γ in the range $[1, 5]$. Typically, the stepsize is reduced quickly according to Eq. (11) to an appropriate level (which depends on the problem) and then stays constant for the remaining iterations.

The motivation for our method is that value iteration for stochastic shortest path problems involves a contraction. In particular, let us consider the mapping $F : \Re^n \to \Re^n$ with components given by

$$F_i(h) = \min_{u\in U(i)}\left[g(i,u) + \sum_{j=1}^{n-1} p_{ij}(u)h(j)\right], \qquad i = 1,\ldots,n. \tag{16}$$

It is known (see e.g., [Ber89], p. 325, or [Tse90]) that, under Assumption 1, F is a contraction mapping with respect to some weighted sup-norm; that is, for some positive scalars $v_1,\ldots,v_n$, and some scalar $\alpha \in (0,1)$, we have

$$\max_{i=1,\ldots,n}\frac{|F_i(h) - F_i(h')|}{v_i} \leq \alpha \max_{i=1,\ldots,n}\frac{|h(i) - h'(i)|}{v_i} \qquad \forall\ h,\ h' \in \Re^n. \tag{17}$$

Note here that while, there is coupling between the iteration of h as per Eq. (7) and the iteration for λ as per Eq. (8), the latter iteration can be made much slower than the former through the use of the stepsize γ, so that the weighted sup-norm contraction character of the iteration (7) is preserved. By contrast, the standard relative value iteration method (9)-(10) does not involve a weighted sup-norm contraction, and in fact it may not involve a contraction of any kind, unless an additional aperiodicity assumption on the Markov chains corresponding to the stationary policies is imposed. We speculate that the sup-norm contraction structure may be helpful in some situations; for example in Q-learning (stochastic approximation) variants of the method, and when parallel asynchronous variations are considered. In fact, an analysis of Q-learning variants of our method that admit a parallel asynchronous implementation is the subject of a forthcoming report [BBA95].

Regarding the relative performance of the new method and the standard method, it can be seen with simple examples that neither method dominates the other. Suppose for instance that there is only one policy and that the corresponding transition probability matrix is

$$\begin{pmatrix} \epsilon & 1-\epsilon \\ 1-\epsilon & \epsilon \end{pmatrix},$$

where ϵ is a scalar from $[0,1]$. Then both methods (7)-(8) and (9)-(10) become linear iterations, and their rate of convergence is governed by the eigenvalues of the corresponding iteration matrix. The eigenvalues corresponding to the standard relative value iteration (9)-(10) can be shown to be 0 and $1-2\epsilon$, so that the method converges very fast for $\epsilon \sim 1/2$ and slowly for $\epsilon \sim 0$ or $\epsilon \sim 1$. It can also be verified that, for a constant but well-chosen value of γ, the eigenvalue structure of the new value iteration method (7)-(8) is worse than the one for the standard method for $\epsilon \sim 1/2$, more favorable for $\epsilon \sim 0$, and comparably unfavorable for $\epsilon \sim 1$.

Our limited computational experiments also indicate that the new method, when properly implemented with the adaptive stepsize rule (15) and the projection scheme of Eqs. (11)-(14), is at least competitive with the relative value iteration method of Eq. (9)-(10). There are problems where one method outperforms the other and reversely. When the initial stepsize γ in Eq. (15) is equal to 1, the performance of the two methods is similar. However, for larger choices of γ (e.g., $\gamma = 5$) we obtained better performance for the new method. We note, however, that much additional testing is required to reach reliable conclusions in this regard. Both methods can be very slow on unfavorably structured problems. This is to be expected since these methods exhibit similar convergence rate behavior to linear iterations and are subject to ill-conditioning.

2. CONVERGENCE ANALYSIS

We now investigate the convergence of the new value iteration algorithm. For convenience, let us denote by $\|\cdot\|$ the weighted sup-norm with respect to which the contraction property of Eq. (17) holds, that is,

$$\|h\| = \max_{i=1,\ldots,n} \frac{|h(i)|}{v_i}, \qquad \forall\ h \in \Re^n.$$

Let us also normalize the vector v so that its last coordinate is equal to 1; that is

$$v_n = 1.$$

Note that since h_λ is the optimal cost vector of the λ-SSP, we have that h_λ is the unique fixed point of the contraction mapping $F(h) - \lambda e$; that is,

$$h_\lambda = F(h_\lambda) - \lambda e, \qquad \forall\ \lambda \in \Re. \tag{18}$$

By writing for all stationary policies μ, states i, and scalars λ and λ',

$$h_{\mu,\lambda}(i) = h_{\mu,\lambda'}(i) + N_\mu(i)(\lambda' - \lambda),$$

and by using the definition $h_\lambda(i) = \min_\mu h_{\mu,\lambda}(i)$, we obtain the following relation:

$$h_{\lambda'}(i) + \underline{N}(\lambda' - \lambda) \leq h_\lambda(i) \leq h_{\lambda'}(i) + \overline{N}(\lambda' - \lambda), \qquad \forall\ i = 1, \ldots, n, \text{ and } \lambda,\ \lambda' \in \Re, \tag{19}$$

where $\underline{N}$ and $\overline{N}$ are the positive scalars

$$\underline{N} = \min_\mu \min_{i=1,\ldots,n} N_\mu(i), \qquad \overline{N} = \max_\mu \max_{i=1,\ldots,n} N_\mu(i). \tag{20}$$

We can interpret $\underline{N}$ and $\overline{N}$ as uniform lower and upper bounds on the slope of the piecewise linear function $h_\lambda(i)$, viewed as a function of λ (see Fig. 2).

The following is our main result:

Proposition 1: There exists a positive scalar $\overline{\gamma}$ such that if

$$\underline{\gamma} \leq \gamma^k \leq \overline{\gamma} \tag{21}$$

for some positive scalar $\underline{\gamma}$ and all k, the sequence (h^k, λ^k) generated by iteration (7), (8) converges to $(h_{\lambda^*}, \lambda^*)$ at the rate of a geometric progression.

Proof: We will show that there exists a threshold value $\overline{\gamma} > 0$ and a continuous function $c(\gamma)$ with $0 \leq c(\gamma) < 1$ for all $\gamma \in (0, \overline{\gamma}]$ such that for any $B > 0$, the relations

$$\|h^k - h_{\lambda^k}\| \leq B \quad \text{and} \quad |\lambda^k - \lambda^*| \leq \frac{B}{\underline{N}} \tag{22}$$

imply that

$$\|h^{k+1} - h_{\lambda^{k+1}}\| \leq c(\gamma^k)B \quad \text{and} \quad |\lambda^{k+1} - \lambda^*| \leq \frac{c(\gamma^k)B}{\underline{N}}. \tag{23}$$

This implies that for a stepsize sequence satisfying the assumptions of the proposition, the sequence $|\lambda^k - \lambda^*|$ converges to zero at the rate of a geometric progression, and the same is true of the sequence $\|h^k - h_{\lambda^k}\|$. Since, using Eq. (19), we have

$$\|h^k - h_{\lambda^*}\| \leq \|h^k - h_{\lambda^k}\| + \|h_{\lambda^k} - h_{\lambda^*}\| \leq \|h^k - h_{\lambda^k}\| + O\big(|\lambda^k - \lambda^*|\big),$$

we see that $\|h^k - h_{\lambda^*}\|$ also converges to zero at the rate of a geometric progression.

We first show two preliminary relations. We have using Eq. (18),

$$\begin{aligned}\|h_{\lambda^{k+1}} - h_{\lambda^k}\| &= \|F\big(h_{\lambda^{k+1}}\big) - \lambda^{k+1}e - F\big(h_{\lambda^k}\big) + \lambda^k e\| \\ &\leq \|F\big(h_{\lambda^{k+1}}\big) - F\big(h_{\lambda^k}\big)\| + \|(\lambda^{k+1} - \lambda^k)e\| \\ &\leq \alpha\|h_{\lambda^{k+1}} - h_{\lambda^k}\| + |\lambda^{k+1} - \lambda^k|\|e\|.\end{aligned}$$

Thus

$$\|h_{\lambda^{k+1}} - h_{\lambda^k}\| \leq \frac{\|e\|}{1-\alpha}|\lambda^{k+1} - \lambda^k|. \tag{24}$$

Also, by subtracting the relations

$$h^{k+1}(n) = F_n(h^k) - \lambda^k,$$

$$h_{\lambda^k}(n) = F_n(h_{\lambda^k}) - \lambda^k,$$

we have

$$|h^{k+1}(n) - h_{\lambda^k}(n)| = |F_n(h^k) - F_n(h_{\lambda^k})| \leq \alpha\|h^k - h_{\lambda^k}\|. \tag{25}$$

Using this relation and Eq. (19), we obtain

$$|h^{k+1}(n)| \leq |h^{k+1}(n) - h_{\lambda^k}(n)| + |h_{\lambda^k}(n)| \leq \alpha\|h^k - h_{\lambda^k}\| + \overline{N}|\lambda^k - \lambda^*|. \tag{26}$$

We will now derive functions $c_1(\cdot)$ and $c_2(\cdot)$ for which the first relation and the second relation in Eq. (22), respectively, hold. We will then use $c(\gamma) = \max\big[c_1(\gamma), c_2(\gamma)\big]$ in Eq. (22). Regarding the first relation in Eq. (23), we note that

$$\begin{aligned}\|h^{k+1} - h_{\lambda^{k+1}}\| &= \|F(h^k) - \lambda^k e - F\big(h_{\lambda^{k+1}}\big) + \lambda^{k+1}e\| \\ &\leq \|F(h^k) - F\big(h_{\lambda^{k+1}}\big)\| + |\lambda^{k+1} - \lambda^k|\,\|e\| \\ &\leq \alpha\|h^k - h_{\lambda^{k+1}}\| + |\lambda^{k+1} - \lambda^k|\,\|e\| \\ &\leq \alpha\|h^k - h_{\lambda^k}\| + \alpha\|h_{\lambda^k} - h_{\lambda^{k+1}}\| + |\lambda^{k+1} - \lambda^k|\,\|e\|.\end{aligned}$$

Using the above inequality, and Eqs. (22), (24), and (26), we obtain

$$\begin{aligned}
\|h^{k+1} - h_{\lambda^{k+1}}\| &\le \alpha B + \left(\frac{\alpha}{1-\alpha} + 1\right) |\lambda^{k+1} - \lambda^k| \, \|e\| \\
&= \alpha B + \frac{\|e\|\gamma^k}{1-\alpha} |h^{k+1}(n)| \\
&\le \alpha B + \frac{\|e\|\gamma^k}{1-\alpha} \left(\alpha \|h^k - h_{\lambda^k}\| + \overline{N} |\lambda^k - \lambda^*|\right) \\
&\le \alpha B + \frac{\|e\|\gamma^k}{1-\alpha} \left(\alpha B + \frac{\overline{N} B}{\underline{N}}\right) \\
&= c_1(\gamma^k) B,
\end{aligned}$$

where $c_1(\cdot)$ is the function

$$c_1(\gamma) = \alpha + \frac{\gamma \|e\| (\alpha + \overline{N}/\underline{N})}{1-\alpha}.$$

Note that if

$$\gamma < \frac{(1-\alpha)^2}{\|e\|(\alpha + \overline{N}/\underline{N})}$$

we have $c_1(\gamma) < 1$.

We now turn to the second relation in Eq. (23); that is, we show that

$$|\lambda^{k+1} - \lambda^*| \le \frac{c_2(\gamma^k) B}{\underline{N}}$$

for an appropriate continuous function $c_2(\gamma)$. Let $\overline{\lambda}$ and $\tilde{\lambda}$ be the unique scalars such that

$$h_{\overline{\lambda}}(n) = B, \qquad h_{\tilde{\lambda}}(n) = \alpha B, \tag{27}$$

(see Fig. 3). Let also $\hat{\lambda}$ be the midpoint between $\overline{\lambda}$ and $\tilde{\lambda}$:

$$\hat{\lambda} = \frac{\overline{\lambda} + \tilde{\lambda}}{2}. \tag{28}$$

Note that from Eq. (19), we have

$$\frac{(1-\alpha)B}{\overline{N}} \le \tilde{\lambda} - \overline{\lambda} \le \frac{(1-\alpha)B}{\underline{N}} \tag{29}$$

and that

$$\begin{aligned}
\frac{\alpha B}{\overline{N}} \le \lambda^* - \tilde{\lambda} \le \frac{\alpha B}{\underline{N}}, \\
\frac{B}{\overline{N}} \le \lambda^* - \overline{\lambda} \le \frac{B}{\underline{N}}.
\end{aligned}$$

From the last three relations, we also obtain

$$\frac{(1+\alpha)B}{2\overline{N}} \le \lambda^* - \hat{\lambda} \le \frac{(1+\alpha)B}{2\underline{N}}, \tag{30}$$

$$\frac{(1-\alpha)B}{2\overline{N}} \leq \tilde{\lambda} - \hat{\lambda} \leq \frac{(1-\alpha)B}{2\underline{N}}. \tag{31}$$

We assume that $\lambda^k \leq \lambda^*$; the complementary case where $\lambda^k \geq \lambda^*$ is handled similarly. We distinguish between two cases:

(a) $\lambda^k \leq \hat{\lambda}$.

(b) $\hat{\lambda} < \lambda^k \leq \lambda^*$.

In the case where $\lambda^k \leq \hat{\lambda}$, we have using Eqs. (19) and (27)-(29),

$$h_{\lambda^k}(n) \geq h_{\hat{\lambda}}(n) \geq h_{\tilde{\lambda}}(n) + \underline{N}(\tilde{\lambda} - \hat{\lambda}) = \alpha B + \underline{N}(\tilde{\lambda} - \hat{\lambda}) \geq \alpha B + \frac{(1-\alpha)B\underline{N}}{2\overline{N}}. \tag{32}$$

On the other hand, from Eqs. (22) and (25), we have $|h^{k+1}(n) - h_{\lambda^k}(n)| \leq \alpha B$ so that

$$h^{k+1}(n) \geq h_{\lambda^k}(n) - \alpha B. \tag{33}$$

By combining Eqs. (32) and (33), we obtain

$$h^{k+1}(n) \geq \frac{(1-\alpha)B}{2\overline{N}^2}.$$

We now have using the above equation,

$$\lambda^* - \lambda^{k+1} = \lambda^* - \lambda^k - \gamma^k h^{k+1}(n) \leq \frac{B}{\underline{N}} - \frac{\gamma^k(1-\alpha)B\underline{N}}{2\overline{N}} = \frac{B}{\underline{N}^2}\left(1 - \frac{\gamma^k(1-\alpha)\underline{N}}{2\overline{N}}\right), \tag{34}$$

and we also have using Eqs. (25), (22), and (19)

$$\lambda^* - \lambda^{k+1} = \lambda^* - \lambda^k - \gamma^k h^{k+1}(n) \geq \lambda^* - \lambda^k - \gamma^k\big(h_{\lambda_k}(n) + \alpha B\big) \geq (1 - \gamma^k\overline{N})(\lambda^* - \lambda^k) - \gamma^k\alpha B. \tag{35}$$

It can be seen now from Eq. (35) that for $\gamma^k \in (0, 1/\overline{N}]$, we have $\lambda^* - \lambda^{k+1} \geq -\gamma^k\alpha B$, and it follows using also Eq. (34) that

$$|\lambda^* - \lambda^{k+1}| \leq \frac{c_2(\gamma^k)B}{\underline{N}},$$

where $c_2(\cdot)$ is the continuous function

$$c_2(\gamma) = \max\left[1 - \frac{\gamma(1-\alpha)\underline{N}^2}{2\overline{N}},\ \gamma\alpha\underline{N}\right].$$

Since there exists a threshold value $\overline{\gamma} > 0$ such that the continuous function $c_2(\gamma)$ satisfies $0 < c(\gamma) < 1$ for all $\gamma \in (0, \overline{\gamma}]$, the desired relation (23) is proved in the case $\lambda^k \leq \hat{\lambda}$.

In the case where $\hat{\lambda} < \lambda^k \leq \lambda^*$, there are two possibilities:

(1) $h^{k+1}(n) \geq 0$. Then $\lambda^k \leq \lambda^{k+1}$, and by using also Eq. (30), we have

$$\lambda^* \leq \hat{\lambda} + \frac{(1+\alpha)B}{2\underline{N}} \leq \lambda^k + \frac{(1+\alpha)B}{2\underline{N}} \leq \lambda^{k+1} + \frac{(1+\alpha)B}{2\underline{N}}. \tag{36}$$

Furthermore, from Eqs. (22) and (26), we have

$$\lambda^{k+1} = \lambda^k + \gamma^k h^{k+1}(n) \leq \lambda^* + \gamma^k \left(\alpha B + \frac{\overline{N} B}{\underline{N}} \right).$$

Thus, by choosing γ^k sufficiently small, we can guarantee that

$$\lambda^{k+1} \leq \lambda^* + \frac{(1+\alpha)B}{2\underline{N}}. \tag{37}$$

From Eqs. (36) and (37), it follows that for γ^k less than some positive constant, we have

$$|\lambda^{k+1} - \lambda^*| \leq \frac{(1+\alpha)B}{2\underline{N}},$$

proving the second relation in Eq. (23), with $c_2(\gamma) = (1+\alpha)/2$.

(2) $h^{k+1}(n) < 0$. In this case, since from Eqs. (22) and (25) we have

$$h_{\lambda^k}(n) \leq h^{k+1}(n) + \alpha B \leq \alpha B, \tag{38}$$

and since $h_{\tilde{\lambda}}(n) = \alpha B$ and $h_\lambda(n)$ is monotonically decreasing in λ, it follows that $\tilde{\lambda} \leq \lambda^k$. Since $\lambda^k \leq \lambda^*$, we also have $0 \leq h_{\lambda^k}(n) \leq \alpha B$, so that by using Eq. (38) and the fact $h_{\lambda^k}(n) \geq 0$, we obtain $|h^{k+1}(n)| \leq \alpha B$ and

$$|\gamma^k h^{k+1}(n)| \leq \gamma^k \alpha B.$$

By choosing

$$\gamma^k \in \left(0, \frac{1-\alpha}{2\alpha\overline{N}}\right], \tag{39}$$

the above inequality, together with Eq. (31), yields

$$|\gamma^k h^{k+1}(n)| \leq \frac{(1-\alpha)B}{2\overline{N}} \leq \tilde{\lambda} - \hat{\lambda} \leq \lambda^k - \hat{\lambda}.$$

Thus, we have

$$\lambda^{k+1} = \lambda^k + \gamma^k h^{k+1}(n) \geq \hat{\lambda},$$

and from Eq. (30), using also the fact $\lambda^{k+1} \leq \lambda^k \leq \lambda^*$, we obtain for γ^k satisfying Eq. (39),

$$|\lambda^{k+1} - \lambda^*| \leq \frac{(1+\alpha)B}{2\underline{N}},$$

proving the second relation in Eq. (23) for the case $h^{k+1}(n) < 0$ as well.

Thus, Eq. (23) holds with $c(\cdot)$ given by

$$c(\gamma) = \max\left[\alpha + \frac{\gamma\|e\|(\alpha + \overline{N}/\underline{N})}{1-\alpha},\ 1 - \frac{\gamma(1-\alpha)\underline{N}^2}{2\overline{N}},\ \gamma\alpha\underline{N},\ \frac{1+\alpha}{2}\right].$$

Q.E.D.

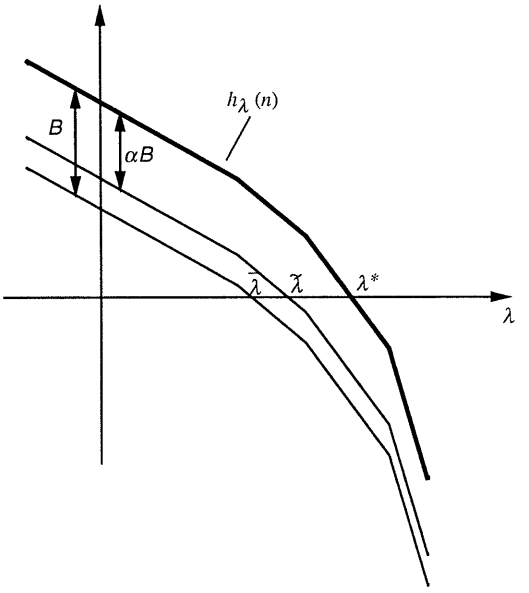

Figure 3. Definition of $\overline{\lambda}$ and $\tilde{\lambda}$ in the proof of Prop. 1.

REFERENCES

[BBA95] Bertsekas, D. P., Borkar, V., and Abounadi, J., 1995. "Q-Learning Algorithms for the Average Cost Markovian Decision Problem," in preparation.

[BeT89] Bertsekas, D. P., and Tsitsiklis, J. N., 1989. Parallel and Distributed Computation: Numerical Methods, Prentice-Hall, Englewood Cliffs, N. J.

[BeT91] Bertsekas, D. P., and Tsitsiklis, J. N., 1991. "An Analysis of Stochastic Shortest Path Problems," Math. Operations Res., Vol. 16, pp. 580-595.

[Ber95] Bertsekas, D. P., 1995. Dynamic Programming and Optimal Control (Vols. I and II), Athena Scientific, Belmont, MA.

[Odo69] Odoni, A. R., 1969. "On Finding the Maximal Gain for Markov Decision Processes," Operations Research, Vol. 17, pp. 857-860.

[PBW79] Popyack, J. L., Brown, R. L., and White, C. C., III, 1969. "Discrete Versions of an Algorithm due to Varaiya," IEEE Trans. Aut. Control, Vol. 24, pp. 503-504.

[Pla77] Platzman, L., 1977. "Improved Conditions for Convergence in Undiscounted Markov Renewal Programming," Operations Research, Vol. 25, pp. 529-533.

[Put94] Puterman, M. L., 1994. Markovian Decision Problems, J. Wiley, N. Y.

[Sch71] Schweitzer, P. J., 1971. "Iterative Solution of the Functional Equations of Undiscounted Markov Renewal Programming," J. Math. Anal. Appl., Vol. 34, pp. 495-501.

[Tse90] Tseng, P., 1990. "Solving H-Horizon, Stationary Markov Decision Problems in Time Proportional to $\log(H)$," Operations Research Letters, Vol. 9, 1990, pp. 287-297.

[Var78] Varaiya, P. P., 1978. "Optimal and Suboptimal Stationary Controls of Markov Chains," IEEE Trans. Automatic Control, Vol. AC-23, pp. 388-394.

[Whi63] White, D. J., 1963. "Dynamic Programming, Markov Chains, and the Method of Successive Approximations," J. Math. Anal. and Appl., Vol. 6, pp. 373-376.