Proceedings of the 2006 American Control Conference
Minneapolis, Minnesota, USA, June 14-16, 2006

Suboptimality Bounds in Stochastic Control: A Queueing Example

Randy Cogill[1] Sanjay Lall[2]

Abstract

In this paper we consider Markov decision processes with average cost criteria, and discuss an approach for characterizing the performance loss associated with using a suboptimal control policy.

Because there are often difficulties associated with computing and implementing optimal control policies, heuristic control policies are often used in practice. For such a policy, we would like to be able to compute guaranteed bounds on its performance, specifically its performance relative to an optimal policy. In other words, our goal is to produce a systematic approach for evaluating how far a specific policy is from optimality.

This approach is demonstrated on a simple queuing system with a single server and multiple job classes. We use the general methods developed in the first part of the paper to show that for any non-idling policy, suboptimality of the resulting average queue length is bounded by a factor which only involves service rates.

1 Introduction

Computation of optimal control policies for Markov decision processes is often intractable, particularly for systems with infinite or very large state spaces. If the general form of the optimal control policy is known, it still may not be implementable due to certain difficulties. These include high computational costs required to evaluate the control action at each time step, requirements that complete state information is known at each time step, or the fact that control decisions are based on a complete statistical description of the system. As a result, suboptimal heuristic control policies are often used in practice. In this paper, we consider Markov decision processes with general state spaces and an average per-period cost objective. The main contribution of this paper is an approach for characterizing the performance loss associated with using a suboptimal control policy. Our goal is to produce a systematic approach for evaluating how far from optimality are the costs incurred using a specific policy. The methods described in this paper are used to provide a worst-case ratio between the cost incurred by a specific suboptimal policy and the cost incurred by an optimal policy.

The general methods are then demonstrated on a simple queuing model. We consider the problem of controlling a queue with a single server and multiple job classes. For this problem, the optimal policy is well known, but its implementation requires knowledge of the exact service rates for each job class. On the other hand, performance analysis of some easily implementable policies, such as FIFO, is generally recognized to be quite difficult. We apply the general methodology discussed in the first part of the paper to show that for any non-idling policy, suboptimality of the resulting queue occupancy is bounded by a factor which only involves service rates. This bound supports the intuition that, if service rates for different classes are reasonably close, then service discipline shouldn't have much effect on queue length, regardless of arrival rates.

1.1 Prior Work

The main contributions of this paper are bounds on average per-period cost for general state space Markov decision processes, and the application of these bounds to the example problem of controlling a multiclass queue. Extensive work has been done previously both on bounding costs in Markov chains, and in control and analysis of multiclass queues.

For finite state Markov chains, bounds similar to those in Section 2 of this paper originally appeared in [9], where they were used for the purpose of proving convergence of a value iteration algorithm for average cost Markov decision processes. Similar bounds appeared later in [10], again for the finite state case. For general state spaces, the bounds in Section 2 are closely related to Lyapunov theorems for Markov chains. For example, a similar upper bound can be found in Theorem 14.2.2 of [8]. One drawback of the standard Lyapunov theorems is that, for systems with positive unbounded costs, they are typically only useful for producing upper bounds. The bounds presented in this paper can be thought of a gen-

[1]Department of Electrical Engineering, Stanford University, Stanford, CA 94305, U.S.A. Email: rcogill@stanford.edu
[2]Department of Aeronautics and Astronautics, Stanford University, Stanford CA 94305-4035, U.S.A. Email lall@stanford.edu
[1]The first author was partially supported by a Stanford Graduate Fellowship.
[1,2]Partially supported by the Stanford URI *Architectures for Secure and Robust Distributed Infrastructures*, AFOSR DoD award number 49620-01-1-0365.

eralization of both the finite state bounds and the Lyapunov bounds, with the particularly attractive feature that they can produce useful upper and lower bounds for systems with unbounded costs.

The type of queueing system that we consider as an example has been extensively studied, primarily in the continuous-time case. The optimal control policy for this system is well known, and is a special case of the $c\mu$ rule [5]. In the discrete-time case, optimality of the $c\mu$ rule was established in [2] and [4]. In [1], the discrete time model is considered and it is shown that the region of achievable average queue lengths for all policies is a polyhedron. It is worth noting that it is possible to obtain the bounds of Theorem 4 from this polyhedral characterization, however, we believe that the approach taken in this paper results in a much simpler proof.

1.2 Preliminaries

In this paper we consider discrete-time Markov decision processes. The systems considered have a general state space $\mathcal{X}$ which is measurable with respect to some given σ-field $B(\mathcal{X})$, and a finite set of actions $\mathcal{U}$ available at each time step. Taking action $u \in \mathcal{U}$ when in state $x \in \mathcal{X}$ incurs a cost $r(x,u)$. After taking action u in state x, the system state evolves according to the probability

$$p(\mathcal{S}|x,u) = \mathbf{Pr}\{X(t+1) \in \mathcal{S} | X(t) = x, U(t) = u\},$$

where $\mathcal{S} \in B(\mathcal{X})$.

We consider the performance of such systems under *static state-feedback* policies. A static state-feedback policy $\mu : \mathcal{X} \to \mathcal{U}$ is a rule which chooses the action in each time step based on the current system state. Under a static state-feedback policy μ, the random process describing state evolution is a time-homogeneous Markov chain X^μ with transition probability $p(\mathcal{S}|x,\mu(x))$. Throughout this paper, we will occasionally simply use the word *policy* when referring to a static state-feedback policy.

In this paper, we evaluate performance of a system under a particular policy μ in terms of the *average per-period cost*

$$J_\mu = \lim_{t\to\infty} \frac{1}{t+1} \sum_{k=0}^{t} E\left[r\big(X^\mu(k), U^\mu(k)\big) | X^\mu(0)\right],$$

where $U^\mu(t) = \mu(X^\mu(t))$. An optimal static state-feedback policy is one which achieves the minimum average per-period cost of all such policies.

Throughout this paper, we assume all functions from $\mathcal{X}$ to $\mathbb{R}$ are measurable with respect to $B(\mathcal{X})$ and the Borel σ-field $B(\mathbb{R})$. Also, to simplify notation, we will occasionally write conditional expectations as

$$E[h(X(t+1)) \mid x,\ u],$$

where it understood that we mean

$$E[h(X(t+1))|X(t) = x, U(t) = u].$$

2 Bounds on Average Per-Period Cost

In this section we give a method for determining bounds on the average per-period cost incurred by a Markov decision process. We will first show how bounds can be computed for Markov chains without control (or with a given state-feedback control). This approach will then be extended to provide a lower bound on the average per-period cost incurred by any static state-feedback policy. By determining an upper bound on the cost incurred by a given policy and a lower bound on the cost incurred by any policy, we can quantify how far from optimal the given policy is.

2.1 Markov Chains Without Control

Theorem 1, which is the main result of this paper, is used to establish upper and lower bounds on the average cost incurred by Markov chains with general measurable state spaces.

Theorem 1. *For any $h : \mathcal{X} \to \mathbb{R}$, let*

$$\beta_u = \sup_{x\in\mathcal{X}} \{r(x) + E[h(X(t+1))|X(t) = x] - h(x)\}$$

and

$$\beta_l = \inf_{x\in\mathcal{X}} \{r(x) + E[h(X(t+1))|X(t) = x] - h(x)\}.$$

If

$$\sup_{x\in\mathcal{X}} \left\{E[h(X(t+1))^2|X(t) = x] - h(x)^2\right\} < \infty,$$

then

$$\beta_l \le \lim_{t\to\infty} \frac{1}{t+1} \sum_{k=0}^{t} E[r(X(k))|X(0)] \le \beta_u$$

for all $X(0) \in \mathcal{X}$.

Proof. Let

$$\Delta_1(x) = E[h(X(t+1))|X(t) = x] - h(x).$$

The definition of β_u implies

$$\begin{aligned}(t+1)\beta_u &\ge E\left[\sum_{k=0}^{t}\Big(r(X(k)) + \Delta_1(X(k))\Big)\,\middle|\, X(0)\right] \\ &= \sum_{k=0}^{t} E[r(X(k))|X(0)] + \\ &\quad E[h(X(t+1))|X(0)] - h(X(0)).\end{aligned}$$

Therefore,

$$\frac{1}{t+1}\sum_{k=0}^{t} E[r(X(k))|X(0)] \leq \beta_u + \frac{1}{t+1}\Big(h(X(0)) - E[h(X(t+1))|X(0)]\Big).$$

Similarly, we can establish the inequality

$$\frac{1}{t+1}\sum_{k=0}^{t} E[r(X(k))|X(0)] \geq \beta_l + \frac{1}{t+1}\Big(h(X(0)) - E[h(X(t+1))|X(0)]\Big).$$

To complete the theorem, we must show that

$$\lim_{t\to\infty} \frac{1}{t+1} E[h(X(t+1))|X(0)] = 0$$

for all $X(0) \in \mathcal{X}$. Let

$$\Delta_2(x) = E[h(X(t+1))^2|X(t) = x] - h(x)^2.$$

If

$$\sup_{x\in\mathcal{X}} \{\Delta_2(x)\} = M < \infty,$$

then

$$\begin{aligned}(t+1)M &\geq E\left[\sum_{k=0}^{t} \Delta_2(X(k)) \,\middle|\, X(0)\right] \\ &= E[h(X(t+1))^2|X(0)] - h(X(0))^2.\end{aligned}$$

Therefore,

$$\begin{aligned}|h(X(0))| + \sqrt{(t+1)M} &\geq \sqrt{h(X(0))^2 + (t+1)M} \\ &\geq \sqrt{E[h(X(t+1))^2|X(0)]} \\ &\geq E\big[|h(X(t+1))| \;\big|\; X(0)\big],\end{aligned}$$

where the last inequality follows from the concavity of the square root and Jensen's inequality. Finally,

$$\lim_{t\to\infty} \frac{1}{t+1} E\big[|h(X(t+1))| \;\big|\; X(0)\big] \leq \lim_{t\to\infty} \frac{1}{t+1}\left(|h(X(0))| + \sqrt{(t+1)M}\right) = 0,$$

implying that $\lim_{t\to\infty} \frac{1}{t+1} E[h(X(t+1))|X(0)] = 0$ for all $X(0) \in \mathcal{X}$. ■

2.2 Markov Chains With Control

Theorem 1 is used to establish upper and lower bounds on the average per-period cost incurred by an irreducible, positive recurrent Markov chain. Since our ultimate goal is to bound the performance gap between a given policy and a policy which achieves minimum cost, this result is now extended to provide a lower bound on the average per-period cost incurred by *any* static state-feedback policy.

Theorem 2. *For any $h : \mathcal{X} \to \mathbb{R}$, let*

$$\beta_l = \inf_{x\in\mathcal{X}, u\in\mathcal{U}} \{r(x,u) + E[h(X(t+1))|x,u] - h(x)\}.$$

Any static state-feedback control policy $\mu : \mathcal{X} \to \mathcal{U}$ such that

$$\sup_{x\in\mathcal{X}} \big\{E[h(X(t+1))^2|x,\mu(x)] - h(x)^2\big\} < \infty$$

has average per-period cost satisfying

$$\beta_l \leq \lim_{t\to\infty} \frac{1}{t+1}\sum_{k=0}^{t} E[r\big(X(k),\mu(X(k))\big)|X(0)]$$

for all $X(0)$.

Proof.

$$\begin{aligned}\beta_l &= \inf_{x\in\mathcal{X}, u\in\mathcal{U}} \{r(x,u) + E[h(X(t+1))|x,u] - h(x)\} \\ &\leq \inf_{x\in\mathcal{X}} \{r(x,\mu(x)) + E[h(X(t+1))|x,\mu(x)] - h(x)\}\end{aligned}$$

The remainder of the proof simply requires application of Theorem 1. ■

Note that in order to use Theorem 2 to prove a lower bound on the performance of an optimal policy, we must show that the expected drift in h^2 is bounded under an optimal policy. This is often accomplished, as in the example of the next section, by showing that the expected drift in h^2 is bounded for all stable policies.

Summary: We will end this section with a summary of how each of these theorems are used:

1. For a given policy $\widehat{\mu}$, Theorem 1 is used to determine an upper bound β_u on the average per-period cost incurred by this policy.
2. Theorem 2 is used to determine a lower bound β_l on the average per-period cost incurred by an optimal state-feedback policy.
3. We can then use β_u/β_l as a bound on the worst-case cost ratio between $\widehat{\mu}$ and an optimal policy.

3 Example: Control of a Multiclass Queue

Here we will apply the methods of the previous section to the problem of controlling a multiclass queue. Our

goal is to show that for any reasonable control policy, the average queue length is within a fixed bound of the optimal average queue length.

A multiclass queue is simply a queue where each job may be a member of one of N distinct job classes. What differentiates job classes is that jobs of certain classes may arrive more frequently than jobs of other classes, and jobs of certain classes may require longer service times than jobs of other classes. For queues with jobs of a single class, the order in which arriving jobs are served has no effect on certain quantities, such as the average queue length. Therefore, a simple control policy such as first in-first out (FIFO) results in the same average queue lengths as a more complex control policy. However, for systems with job dependent service times, the order in which arriving jobs are serviced does have an effect on average queue length. For a system with a finite number of job classes and memoryless arrival and service time distributions, the policy which minimizes average queue length is well known [2, 4]. In the optimal policy, job classes are prioritized according to average service time, and jobs with shorter average service times are served before jobs with longer average service times. Also, if a low priority job is in service, it is temporarily put aside if a job of higher priority arrives. In other words, this policy assumes that it possible to distinguish between job classes, requires knowledge of service statistics for each job class, and allows to possibility of preempting jobs in service. When it is impractical to implement such a control policy, we need to resort to a simpler suboptimal control policy. This raises the question, "If I use a suboptimal control policy, how much longer than optimal may the average queue length be?" We consider the set of all *non-idling* policies, which are policies that always serve jobs as long as there are jobs in the queue. We will show that the queue lengths resulting from any non-idling policy are bounded by a factor involving only the service rates. The bound we obtain gives a way of quantifying the intuitive notion that the control policy has little effect on average queue length if service rates for all job classes are close.

We consider a discrete time model of this queueing system. In time slot t, $A_i(t) \in \{0,1\}$ jobs of class i arrive in the queue. We assume that at most one job arrives in the queue in each time slot; that is $\sum_i A_i(t) \leq 1$ for all t. We also assume that for all $t' \neq t$, the arrival vectors $A(t)$ and $A(t')$ are independent and identically distributed. We denote $\lambda_i = E[A_i(t)]$, where this expectation is independent of t. We let $X_i(t)$ denote the number of class i jobs in the queue at time t, and let $X(t) = (X_1(t), \ldots, X_N(t))$. We define the control sequence U_i such that $U_i(t) = 1$ if a class i job is being serviced in time slot t, and $U_i(t) = 0$ otherwise. If we are servicing a job of class i in a time slot, then service is completed with probability σ_i. This probability is independent of service history, resulting in geometrically distributed service times. Finally, we let $D_i(t) = U_i(t)I(X_i(t))B_i$, where B_i is a Bernoulli random variable with $E[B_i] = \sigma_i$ and I denotes the indicator function defined as

$$I(x) = \begin{cases} 0 & \text{if } x = 0 \\ 1 & \text{otherwise} \end{cases}.$$

The random variable $D_i(t)$ indicates the number of jobs of class i successfully served in time slot t. The queue length dynamics evolve according to

$$X_i(t+1) = X_i(t) + A_i(t) - D_i(t)$$

for each $i \in \{1, \ldots, N\}$.

It is clear that the problem of choosing how to serve job classes in order to minimize average queue length is a Markov decision process with average per-period cost criteria. For the model we consider, the state space is $\mathcal{X} = \mathbb{Z}_+^N$, the set of possible queue occupancies for each job class. The action space is

$$\mathcal{U} = \left\{ u \in \{0,1\}^N \,\middle|\, \sum_{i=1}^N u_i = 1 \right\}.$$

Here the cost incurred in time slot t is $r(X(t)) = \sum_{i=1}^N X_i(t)$, the total number of jobs in the queue. Let $\mu : \mathcal{X} \to \mathcal{U}$ be a state-feedback control policy, and let X^μ be the queue length process under this policy. Then

$$J_\mu = \lim_{t\to\infty} \frac{1}{t+1} \sum_{k=0}^{t} E[r(X^\mu(k))|X^\mu(0)]$$

is the average queue length under policy μ. As mentioned before, we will consider the class of non-idling policies. A policy μ is said to be non-idling if, for any $x \neq 0$, $\mu(x) = u_i$ implies $x_i > 0$.

We would now like to consider the effect of control policy on average queue length. Under certain conditions, bounded queue lengths may not exist for any non-idling policy (i.e., the system is not stabilizable). Therefore, we must identify and restrict our attention to the cases where the system is stabilizable. The following lemma identifies the cases in which the system cannot be stabilized by a non-idling policy. This result is standard (see, for example, [6]), so the proof is omitted due to space constraints. However, we would like to point out that the methods of the previous section could be used to establish this result.

Lemma 3. *If $\sum_{i=1}^N \frac{\lambda_i}{\sigma_i} > 1$, then there is no non-idling policy with bounded average queue length.*

It turns out that, when $\sum_{i=1}^N \frac{\lambda_i}{\sigma_i} < 1$, any non-idling policy achieves bounded average queue lengths. The following result relates the average queue length of an arbitrary non-idling policy to the average queue length of an optimal policy.

Theorem 4. *Let μ_{NI} be an arbitrary non-idling policy, and let μ_{OPT} be an optimal non-idling policy. If $\sum_{i=1}^{N} \frac{\lambda_i}{\sigma_i} < 1$, then $J_{\mu_{NI}}$ and $J_{\mu_{OPT}}$ are finite and satisfy*

$$\frac{J_{\mu_{NI}}}{J_{\mu_{OPT}}} \leq \frac{\max_i\{\sigma_i\}}{\min_i\{\sigma_i\}}.$$

Proof. To prove an upper bound on $J_{\mu_{\rm NI}}$, we use

$$h_u(x) = K_1 \left(\left(\sum_{i=1}^{N} \frac{x_i}{\sigma_i} \right)^2 + K_2 \sum_{i=1}^{N} \frac{x_i}{\sigma_i} \right),$$

where

$$\begin{aligned} K_1 &= \frac{\max_i\{\sigma_i\}}{2\left(1 - \sum_{i=1}^{N} \frac{\lambda_i}{\sigma_i}\right)} \\ K_2 &= \left(1 - 2\sum_{i=1}^{N} \frac{\lambda_i}{\sigma_i}\right). \end{aligned}$$

Let

$$\Delta_u(x) = E[h_u(X(t+1)) \mid X(t) = x] - h_u(x).$$

Using $X_i(t+1) = X_i(t) + A_i(t) - D_i(t)$, after some algebra one can obtain

$$r(x) + \Delta_u(x) = \sum_{i=1}^{N} \left(1 - \frac{\max_j\{\sigma_j\}}{\sigma_i}\right) x_i + \beta \max_j\{\sigma_j\},$$

where

$$\beta = \frac{\sum_{i=1}^{N} \left(1 + \frac{1}{\sigma_i}\right) \frac{\lambda_i}{\sigma_i} - 2\left(\sum_{i=1}^{N} \frac{\lambda_i}{\sigma_i}\right)^2}{2\left(1 - \sum_{i=1}^{N} \frac{\lambda_i}{\sigma_i}\right)}.$$

To prove a lower bound on $J_{\mu_{\rm OPT}}$, we use

$$h_l(x) = K_3 \left(\left(\sum_{i=1}^{N} \frac{x_i}{\sigma_i} \right)^2 + K_2 \sum_{i=1}^{N} \frac{x_i}{\sigma_i} \right)$$

where

$$K_3 = \frac{\min_i\{\sigma_i\}}{2\left(1 - \sum_{i=1}^{N} \frac{\lambda_i}{\sigma_i}\right)}$$

and K_2 is defined as before. Let

$$\Delta_l(x,u) = E[h_l(X(t+1)) \mid X(t) = x, U(t) = u] - h_l(x).$$

The choice of u which minimizes $r(x) + \Delta_l(x,u)$ has $E\left[\sum_{i=1}^{N} \frac{D_i(t)}{\sigma_i} \middle| X(t) = x, U(t) = u\right] = 1$ whenever $x \neq 0$. Therefore,

$$\min_{u \in \mathcal{U}}\{r(x) + \Delta_l(x,u)\} = \sum_{i=1}^{N} \left(1 - \frac{\min_j\{\sigma_j\}}{\sigma_i}\right) x_i + \beta \min_j\{\sigma_j\}.$$

To complete this proof, we need to show that

$$\sup_{x \in \mathcal{X}} \left\{ E[h_u(X(t+1))^2 \mid x, \mu(x)] - h_u(x)^2 \right\} < \infty$$

and

$$\sup_{x \in \mathcal{X}} \left\{ E[h_l(X(t+1))^2 \mid x, \mu(x)] - h_l(x)^2 \right\} < \infty$$

for all non-idling policies μ. Both h_u and h_l are of the form

$$h(y) = K(y^2 + K_2 y),$$

where $y = \sum_{i=1}^{N} \frac{x_i}{\sigma_i}$. Squaring h obtains

$$h(y)^2 = K^2(y^4 + 2K_2 y^3 + K_2^2 y^2).$$

For any non-idling policy, the expected drift

$$E[h(X(t+1))^2 \mid x, \mu(x)] - h(x)^2$$

is a third order polynomial with the third order term equal to

$$4K^2 \left(\sum_{i=1}^{N} \frac{\lambda_i}{\sigma_i} - 1 \right) \left(\sum_{i=1}^{N} \frac{x_i}{\sigma_i} \right)^3.$$

Whenever the system is stabilizable, this is negative for all $x \neq 0$. Hence, the expected drift in h^2 is bounded above for all non-idling policies. This implies that the bounds β_u and β_l are valid and

$$\frac{J_{\mu_{\rm NI}}}{J_{\mu_{\rm OPT}}} \leq \frac{\beta_u}{\beta_l} = \frac{\max_i\{\sigma_i\}}{\min_i\{\sigma_i\}}.$$

■

We would like to point out that this bound is *not* obtained by simply upper and lower bounding the average queue length of the multiclass queue by the queue lengths of queues which serve all job classes at the minimum and maximum service rates, respectively. While such an approach would provide upper and lower bounds, the gap between these bounds can be made arbitrarily large for given service rates. In fact, we can show that the bound produced in Theorem 4 is tight.

The proof of tightness of the bound is left out due to space constraints. Essentially, this proof considers a system with two job classes and compares the performance of the two policies which give strict priority to one class. By computing the average queue lengths under these policies, it is shown that arrival statistics can be chosen so that the the bound in Theorem 4 can be approached arbitrarily closely.

It is also worth noting that the proof of Theorem 4 does not require any knowledge of the optimal policy. The lower bound on optimal performance is determined via a much simpler argument than would be required to derive and evaluate the optimal policy. Also, it may

appear at first sight that the upper bound in the proof of Theorem 4 may not apply to FIFO, since FIFO is not a state-feedback policy under the current definition of the system state. However, it is possible to create a countable state space under which FIFO is a state-feedback policy. Moreover, r and h_u can be interpreted in this new state space and an identical upper bound can be derived. A similar procedure could be carried out to verify that the upper bound is also valid for non-preemptive policies.

4 Conclusion

In this paper, we presented a method which can be applied to obtain bounds on costs in Markov decision processes with general state spaces and average per-period cost criteria. This method was applied to the problem of control of a multiclass queue to establish a bound on the ratio between the average queue length achieved by any practical policy and the average queue length achieved by an optimal policy.

References

[1] E. Altman and A. Schwartz. Optimal priority assignment: a time sharing approach. *IEEE Trans. Automatic Control*, 34:1089–1102, 1989.

[2] J.S. Baras, A.J. Dorsey, and A.M. Makowski. K competing queues with geometric service requirements and linear costs: the μc rule is always optimal. *Systems and Control Letters*, 6:173–180, 1985.

[3] D. Bertsimas, D. Gamarnik, and J.N. Tsitsiklis. Performance of multiclass Markovian queueing networks via piecewise linear lyapunov functions. *Ann. Applied Probability*, 11(4):1384–1428, 2001.

[4] C. Buyukkoc, P. Varaiya, and J. Walrand. The $c\mu$ rule revisited. *Advances in Applied Probability*, 17:237–238, 1985.

[5] D.R. Cox and W.L. Smith. *Queues*. John Wiley, New York, 1961.

[6] L. Kleinrock. *Queueing Systems. Volume II: Computer Applications*. John Wiley, New York, 1976.

[7] S. Kumar and P.R. Kumar. Performance bounds for queueing networks and scheduling policies. *IEEE Trans. Automatic Control*, 39(8):1600–1611, 1994.

[8] S. Meyn and R. Tweedie. *Markov Chains and Stochastic Stability*. Springer-Verlag, 1993.

[9] A.R. Odoni. On finding the maximal gain for Markov decision processes. *Operations Research*, 17:857–860, 1969.

[10] P.J. Schweitzer and A. Seidmann. Generalized polynomial approximations in Markovian decision processes. *J. Mathematical Analysis and Applications*, 110:568–582, 1985.