Proceedings of the 28th Conference
on Decision and Control
Tampa, Florida • December 1989

VARIABILITY SENSITIVE MARKOV DECISION PROCESSES*

Melike Baykal-Gursoy
Industrial Engineering Department
Rutgers University
Piscataway, NJ 08855

Keith W. Ross†
Department of Systems
University of Pennsylvania
Philadelphia, PA 19104

Abstract

The time-average Markov Decision Processes with finite state and action spaces are considered. Several definitions of variability are introduced and compared. It is shown that a stationary policy maximizes one of these criteria, namely, the expected long-run average variability. Furthermore, an algorithm is given which produces such an optimal stationary policy.

1 Introduction

We consider a discrete-time Markov Decision Process(MDP) with finite state space and finite action space. Denote R_m for the reward obtained at epoch m. This paper considers the problem of finding a policy $\mathbf{u}$ that maximizes

$$\nu(\mathbf{u}) = E^{\xi}_{\mathbf{u}}[\liminf_{n\to\infty} \frac{1}{n}\sum_{m=1}^{n} h(R_m, \frac{1}{n}\sum_{m=1}^{n} R_m)],$$

for some fixed and given initial state ξ. The variability function $h(.,.)$ compares at epoch m the current reward with the average reward over an interval that includes m. If $h(x,y) = (x-y)^2$, then $\nu(\mathbf{u})$ may be interpreted as the expected time-average variance. If $h(x,y) = x - \lambda(x-y)^2$ for some $\lambda > 0$, then maximizing $\nu(\mathbf{u})$ would correspond to finding a policy $\mathbf{u}$ that has high expected average reward but low expected average variance.

Under mild continuity conditions on the variability function, it shall be shown that there exists a stationary policy that maximizes $\nu(\mathbf{u})$. Moreover, this policy can be located by the following four-step procedure: 1) The state space is decomposed into "strongly communicating classes" and a set of transient states; 2) For each strongly communicating class, a mathematical program with linear constraints and nonlinear objective function is solved; 3) An average reward MDP problem is solved where there is one state for every strongly communicating class; 4) Lastly, an optimal stationary policy is formed by combining the optimal solutions in step 2 with the optimal policy in step 3.

This paper is organized as follows. The notation is given in Section 2. In Section 3 several notions of variability are introduced and compared. The problem of maximizing $\nu(\mathbf{u})$ over all policies is investigated in Section 4.

2 Notation

Let $\mathcal{S}$ and $\mathcal{A}$ denote the finite state and action space, respectively. The underlying sample space for the MDP is $\Omega := \{(x_1, a_1, x_2, a_2, \ldots) : x_n \in \mathcal{S}, a_n \in \mathcal{A} \text{ for all } n = 1, 2, \ldots\}$. The sample space Ω is equipped with the σ-algebra $\mathcal{F}$ generated by the random variables $\{X_1, A_1, X_2, A_2, \ldots\}$.

Let P_{xay}, $x, y \in \mathcal{S}$, $a \in \mathcal{A}$ denote the law of motion for MDP, $\mathbf{f}$ denote a *stationary* policy and $\mathbf{g}$ denote a *pure* or *nonrandomized* policy. Let C and C_S denote the class of all policies and stationary policies, respectively.

Under any stationary policy $\mathbf{f}$, the state process $\{X_m\}$ is a homogeneous Markov chain with transition matrix $P(\mathbf{f})$. A transition matrix $P(\mathbf{f})$ is said to be *unichain* if it has at most one recurrent class plus a (possibly empty) set of transient states. In this case, $\underline{\pi}(\mathbf{f})$ denotes the unique equlibrium vector associated with $P(\mathbf{f})$.

For each $x \in \mathcal{S}$ and $a \in \mathcal{A}$ define the random variables denoting the average state-action frequencies through epoch n as

$$Z_n(x,a) := \frac{1}{n}\sum_{m=1}^{n} \mathbf{1}\{X_m = x, A_m = a\},$$

where $\mathbf{1}\{\Lambda\}$ is the indicator function of set Λ. Let C_0 denote the class of all policies $\mathbf{u}$ such that $\{Z_n(x,a)\}$ converges $P_{\mathbf{u}}$-almost surely($P_{\mathbf{u}}$-a.s.) for all $x \in \mathcal{S}$ and $a \in \mathcal{A}$. Let C_1 denote the class of all policies $\mathbf{u}$ such that $\{E_{\mathbf{u}}[Z_n(x,a)]\}$ converges for all x and a (see e.g. [3]).

3 Notions of Variability

In order to compare $\nu(\mathbf{u})$ with the different notions of variability introduced by other researchers, denote,

$$\phi(\mathbf{u}) := \liminf_{n\to\infty} \frac{1}{n}\sum_{m=1}^{n} E_{\mathbf{u}}[R_m],$$

for the long-run average expected reward where $R_m := r(X_m, A_m)$. Let $h(.,.)$ be a continuous real-valued function defined on $\Delta \times \mathcal{R}$, where $\Delta := \{r(x,a) : x \in \mathcal{S}, a \in \mathcal{A}\}$ and $\mathcal{R}$ is the set of real numbers. Define the average expected variability as

$$v_1(\mathbf{u}) := \liminf_{n\to\infty} \frac{1}{n}\sum_{m=1}^{n} E_{\mathbf{u}}[h(R_m, \phi(\mathbf{u}))].$$

Note that if $h(x,y) = (x-y)^2$, then $v_1(\mathbf{u}) = var(\mathbf{u})$ for all $\mathbf{u} \in C_1$, where

$$var(\mathbf{u}) := \lim_{n\to\infty} \frac{1}{n}\sum_{m=1}^{n} E_{\mathbf{u}}[(R_m - \phi(\mathbf{u}))^2].$$

If $h(x,y) = x - \lambda(x-y)^2$ then $v_1(\mathbf{u}) = \phi(\mathbf{u}) - \lambda var(\mathbf{u})$ for all $\mathbf{u} \in C_1$, which corresponds to the criterion considered by Filar *et al* [4] and Sobel[7].

As an alternative definition to $v_1(\mathbf{u})$, define

$$v_2(\mathbf{u}) := \liminf_{n\to\infty} \frac{1}{n}\sum_{m=1}^{n} E_{\mathbf{u}}[h(R_m, \phi_n(\mathbf{u}))],$$

where

$$\phi_n(\mathbf{u}) := \frac{1}{n} E_{\mathbf{u}}[\sum_{m=1}^{n} r(X_m, A_m)].$$

Similarly, the variability $v_3(\mathbf{u})$ is defined as

$$v_3(\mathbf{u}) := \liminf_{n\to\infty} \frac{1}{n}\sum_{m=1}^{n} E_{\mathbf{u}}[h(R_m, E_{\mathbf{u}}[R_m])].$$

Proposition 1 *If* $\mathbf{u} \in C_1$, *then*

$$v_2(\mathbf{u}) = v_1(\mathbf{u}) = \sum_{x,a} h[r(x,a), \sum_{x,a} r(x,a) z_{xa}(\mathbf{u})] z_{xa}(\mathbf{u}), \qquad (1)$$

where $z_{xa}(\mathbf{u}) := \lim_{n\to\infty} E_{\mathbf{u}}[Z_n(x,a)]$. *Furthermore, if* $\mathbf{u}$ *is such that* $\{P_{\mathbf{u}}(X_m = x, A_m = a); m = 1, 2, \ldots\}$ *converges for all* $x \in \mathcal{S}$, $a \in \mathcal{A}$, *then*

$$v_1(\mathbf{u}) = v_2(\mathbf{u}) = v_3(\mathbf{u}).$$

Also consider,

$$V_1 := \liminf_{n\to\infty} \frac{1}{n}\sum_{m=1}^{n} h(R_m, \liminf_{n\to\infty} \frac{1}{n}\sum_{m=1}^{n} R_m)$$

$$V_2 := \liminf_{n\to\infty} \frac{1}{n}\sum_{m=1}^{n} h(R_m, \frac{1}{n}\sum_{m=1}^{n} R_m).$$

With this notation, the variability criterion $\nu(\mathbf{u})$ of the Introduction becomes $\nu(\mathbf{u}) = E_{\mathbf{u}}[V_2]$.

*Write to the authors for the complete manuscript.

†Partially supported by NSF through grant number NCR-8707620.

CH2642-7/89/0000-1261$1.00 © 1989 IEEE

Proposition 2 *Suppose* $\mathbf{u} \in C_0$. *Then*

$$V_1 = V_2 = \sum_{x,a} h[r(x,a), \sum_{x,a} r(x,a)Z(x,a)]Z(x,a) \qquad (2)$$

holds $P_{\mathbf{u}}$*-a.s. Consequently* $E_{\mathbf{u}}[V_1] = E_{\mathbf{u}}[V_2] = \nu(\mathbf{u})$ *for all* $\mathbf{u} \in C_0$.

Proposition 3 *Let* $\mathbf{f} \in C_S$ *and let* $\mathcal{R}^1(\mathbf{f}), \ldots, \mathcal{R}^q(\mathbf{f})$ *be the recurrent classes induced by* $P(\mathbf{f})$. *Denote* $(\pi_x^i(\mathbf{f}) : x \in \mathcal{R}^i(\mathbf{f}))$ *for the equilibrium probability vector associated with class* i, $i = 1, \ldots, q$. *Further denote*

$$\psi_i = \sum_{x,a} r(x,a)\pi_x^i(\mathbf{f})f_{xa}$$

and $\tau := \min\{m : m \geq 1,\ X_m \in \cup_{i=1}^q \mathcal{R}^i(\mathbf{f})\}$. *Then*

$$v_1(\mathbf{f}) = \sum_{i=1}^q P_{\mathbf{f}}(X_\tau \in \mathcal{R}^i(\mathbf{f})) \sum_{x,a} h[r(x,a), \sum_{i=1}^q P_{\mathbf{f}}(X_\tau \in \mathcal{R}^i(\mathbf{f}))\psi_i] \pi_x^i(\mathbf{f})f_{xa}, \qquad (3)$$

$$\nu(\mathbf{f}) = \sum_{i=1}^q P_{\mathbf{f}}(X_\tau \in \mathcal{R}^i(\mathbf{f})) \sum_{x,a} h[r(x,a), \psi_i]\pi_x^i(\mathbf{f})f_{xa}. \qquad (4)$$

Proposition 4 *Suppose* $h(x,y) = x - \lambda(x-y)^2$ *for some* $\lambda > 0$. *Then,* $\nu(\mathbf{u}) \geq v_1(\mathbf{u})$ *for all* $\mathbf{u} \in C_0$.

4 Optimization Results

To construct an optimal policy $\mathbf{f} \in C_S$, first the state space $\mathcal{S}$ is partitioned into strongly communicating classes $\mathcal{C}^1, \mathcal{C}^2, \ldots, \mathcal{C}^p$ and a set of transient states $\mathcal{T}$ (see e.g.[1], [6]) so that $\sum_{i=1}^p P_{\mathbf{u}}(\Phi_i) = 1$, where $\Phi_i := \{X_n \in \mathcal{C}^i \text{ almost always}\}$. A set of states $\mathcal{C}$ is said to be a *strongly communicating class* i) if $\mathcal{C}$ is a recurrent class for some stationary policy; ii) $\mathcal{C}$ is not a proper subset of some set $\mathcal{D}$ for which (i) holds true.

Next the MDP is restricted to each of the strongly communicating classes. Each restricted MDP corresponds to a mathematical program that involves maximizing a nonlinear function over a simple polytope. Based on the optimal values of the restricted MDPs, an aggregated MDP is constructed. An optimal stationary policy for the original problem is then obtained by combining the optimal policy for the aggregated MDP with the optimal solutions for the restricted MDPs.

4.1 The Restricted MDP

The restricted MDP, MDP-i is obtained for each $i = 1, \ldots, p$ by considering the set $\mathcal{C}^i$ as the state space and for $x \in \mathcal{C}^i$ the set $\mathcal{F}_x = \{a \in \mathcal{A} : P_{xay} = 0 \text{ for all } y \notin \mathcal{C}^i\}$ as the state dependent action space.

For a fixed MDP-i and a fixed initial state $\xi \in \mathcal{C}^i$ the corresponding expected average variability for MDP-i is given by

$$\nu_\xi^i(\mathbf{u}) := E_{\mathbf{u},\xi}^i[\liminf_{n\to\infty} \frac{1}{n}\sum_{m=1}^n h(R_m, \frac{1}{n}\sum_{m=1}^n R_m)].$$

For each MDP-i, consider the following mathematical program with decision variables $\{z_{xa} : x \in \mathcal{C}^i, a \in \mathcal{F}_x\}$:
Program Q^i

$$\begin{aligned} T^i := \ max \ & \sum_{x\in\mathcal{C}^i}\sum_{a\in\mathcal{F}_x} h[r(x,a), \sum_{x\in\mathcal{C}^i}\sum_{a\in\mathcal{F}_x} r(x,a)z_{xa}]z_{xa} \\ s.t. \ & \sum_{x\in\mathcal{C}^i}\sum_{a\in\mathcal{F}_x} (\delta_{xy} - P_{xay})z_{xa} = 0, \ \forall y \in \mathcal{C}^i \\ & \sum_{x\in\mathcal{C}^i}\sum_{a\in\mathcal{F}_x} z_{xa} = 1 \\ & z_{xa} \geq 0. \end{aligned}$$

Theorem 1 *For each* $i = 1, \ldots, p$, *and for all policies* $\mathbf{u} \in \mathcal{C}$ *the following holds:*

$$P_{\mathbf{u}}\{\liminf_{n\to\infty} \frac{1}{n}\sum_{m=1}^n h(R_m, \frac{1}{n}\sum_{m=1}^n R_m) \leq T^i | \Phi_i\} = 1.$$

An algorithm similar to the one given in [5] constructs a stationary policy $\mathbf{f}^i$ for MDP-i.

Theorem 2 *The stationary policy* $\mathbf{f}^i$ *is optimal for MDP-*i, *for all initial states* $\xi \in \mathcal{C}^i$. *Moreover,* $\nu_\xi^i(\mathbf{f}^i) = T^i$ *for all* $\xi \in \mathcal{C}^i$.

4.2 The Aggregated MDP

In the aggregated MDP, there is one state corresponding to each strongly communicating class plus states corresponding to the transient states in $\mathcal{T}$. For each state $i = 1, \ldots, p$, the action θ is available, which keeps the aggregated MDP in state i with probability 1. The actions of the form (x,a) are also available, which, in the original MDP, correspond to a movement to state x and then a selection of action a. The law of motion is defined accordingly.

Let $\beta^i(\mathbf{u})$ denote the average reward given that the initial state is i for the aggregated MDP, i.e.,

$$\beta^i(\mathbf{u}) := E_{\mathbf{u}}^i[\liminf_{n\to\infty} \frac{1}{n}\sum_{m=1}^n \sum_{k=1}^p T^k \mathbf{1}\{X_m = k\}].$$

A policy is optimal for the aggregated MDP if it attains β^i for all $i = 1, \ldots, p+t$, where $\beta^i := \sup_{\mathbf{u}\in C} \beta^i(\mathbf{u})$. An optimal pure policy $\mathbf{g}$ can be found for the aggregated MDP by any one of the standard MDP algorithms (for example policy improvement of linear programming). The optimal pure policy $\mathbf{g}$ in the aggregated MDP will have the following interpretation in the original MDP. If $g(i) = \theta$, then it is best for the original MDP to remain in C_i once having entered that strongly communicating class. On the other hand, if $g(i) = (l,a)$, the original MDP should move to state l and choose action a.

A stationary policy $\mathbf{f}$ can be constructed for the original problem of maximizing $\nu(\mathbf{u})$ over $\mathbf{u} \in C$, such that a) if the state i is recurrent under pure policy $\mathbf{g}$ in the aggregated MDP, then f_{xa} is identical to f_{xa}^i for $x \in \mathcal{C}^i$ and $a \in \mathcal{F}_x$, b) if state x is transient then f_x is identical to $g(x)$, c) if a state i is not recurrent under pure policy $\mathbf{g}$ in the aggregated MDP, then one can find an algorithm which moves the MDP to a recurrent class. Details are given in [2]

Theorem 3 *The stationary policy* $\mathbf{f}$ *constructed by the above procedure is optimal for the original problem of maximizing* $\nu(\mathbf{u})$ *over all* $\mathbf{u} \in C$.

References

[1] J. Bather. Optimal decision procedures in finite Markov chains. Part III: general convex systems. *Adv. in Appl. Prob.*, 5:541–553, 1973.

[2] M. Baykal-Gursoy and K. W. Ross. Variance-penalized Markov decision processes. *submitted.*

[3] C. Derman. *Finite State Markovian Decision Processes.* Academic Press, New York, 1970.

[4] J. Filar, L. Kallenberg, and H. Lee. Variance penalized Markov decision processes. *Math. of Op. Res.*, 14:147–161, 1989.

[5] L. Kallenberg. *Linear Programming and Finite Markovian Control Problems*, volume 148. Mathematical Centre Tracts, Amsterdam, 1983.

[6] K. Ross and R. Varadarajan. The decomposition of time average markov decision processes: Theory, algorithms and applications. *submitted.*

[7] M. Sobel. Mean variance tradeoffs in an undiscounted MDP. Preprint, 1984.