Control and Cybernetics

vol. **35** (2006) No. 3

Dynamic programming in constrained Markov decision processes

by

A.B. Piunovskiy

Department of Mathematical Sciences
M & O Building, The University of Liverpool
Liverpool, L69 7ZL, UK
e-mail: piunov@liverpool.ac.uk

Abstract: We consider a discounted Markov Decision Process (MDP) supplemented with the requirement that another discounted loss must not exceed a specified value, almost surely. We show that the problem can be reformulated as a standard MDP and solved using the Dynamic Programming approach. An example on a controlled queue is presented. In the last section, we briefly reinforce the connection of the Dynamic Programming approach to another close problem statement and present the corresponding example. Several other types of constraints are discussed, as well.

Keywords: Markov decision process (MDP), constraints, optimization, dynamic programming, myopic control strategy, queuing system.

1. Introduction

Constrained problems appear naturally when one considers more than one objective. Examples can be found in Altman (1999), Chen (2004), Chen and Blankenship (2004), Piunovskiy (1997), Piunovskiy and Mao (2000), Sniedovich (1980), Yakowitz (1982), and in other monographs and articles. One can formulate many versions of constrained optimal control problems. Let h be a trajectory of the system and $R(h)$ and $S(h)$ be the cost functions, and let π denote a control strategy. (The rigorous mathematical constructions are given in Section 2.)

Version 1: $E^\pi[R(h)] \to \inf_\pi, \quad E^\pi[S(h)] \le d.$

Version 2: $E^\pi[R(h)] \to \inf_\pi, \quad S(h) \le d \; P^\pi$-a.s.

Version 3: $E^\pi[R(h)] \to \inf_\pi, \quad P^\pi[S(h) \le d] \ge \alpha \; (\alpha \in [0,1]).$

Version 4: $E^\pi[R(h)] \to \inf_\pi, \quad Var^\pi[S(h)] \le d.$

The main objective can also be of a different form (e.g. $P^\pi[R(h) \le c] \to \sup_\pi$); there can be several constraints (e.g. $E^\pi[S^1(h)] \le d^1$; $P^\pi[S^2(h) \le d^2] \ge \alpha$), and so on. Of course, the cost function S itself can be vector-valued.

Version 1 was studied in Altman (1999), Chen and Blankenship (2004), Piunovskiy (1997), Piunovskiy and Mao (2000), Feinberg and Shwartz (1999), Tanaka (1991), and in many other papers (see the survey in Piunovskiy, 1998). The main instrument here was the Convex Analytic Approach leading to Linear Programs. In Chen and Blankenship (2004), Piunovskiy and Mao (2000), and partially in Feinberg and Shwartz (1999), Dynamic Programming is used to tackle the Version 1 type problems. At the same time, this method is also numerically convenient; moreover, it allows to build ALL (Markov) optimal control strategies.

Versions 2-4 received less attention. Type 2 is only mentioned in Chen and Blankenship (2004), Piunovskiy and Mao (2000). This is possibly because Version 2 frequently has no feasible strategies at all. In contrast to Version 1, the Convex Analytic Approach is problematic here, but the Dynamic Programming Approach is natural and straightforward. Clearly, many resource allocation problems can be reformulated as Version 2 problems. (For instance, the example solved in Section 4 belongs to this class.) In such a situation it is possible to satisfy constraints almost surely, because the value of resource allocated at each step is non-random under the known current state of the process.

Version 3 (entirely probabilistic criteria: $P^\pi[R(h) \le c] \to \sup_\pi$) was studied in Chen (2004), see also Yakowitz (1982). Version 4 was investigated in Sniedovich (1980). These papers are mainly based on the dynamic programming method.

In the present article (Sections 2 and 3), we focus on Version 2. The main idea is similar to the penalty functions method; as a result, we obtain an equivalent unconstrained optimization problem and apply the Dynamic Programming method to it (Statement 3.1). The model under consideration is Markovian, with the total cost optimality criterion. Section 4 is devoted to an example of a controlled queue. In Section 5, we briefly discuss other versions of constrained problems. In contrast to Chen and Blankenship (2004), we study the general Borel model. Several statements deal with the so called 'myopic' strategies which are of interest on their own.

Dynamic Programming is widely used for solving real life problems. As for stochastic constrained versions, one can find examples from Queuing Theory, Reservoir Management, Radar Systems, Resource Allocation and Reliability Theory in Chen (2004), Chen and Blankenship (2004), Piunovskiy and Mao (2000), Sniedovich (1980), Yakowitz (1982).

2. Model description and auxiliary results

Consider the controlled model $Z = \{X, A, p\}$ where X is the Borel state space; A is the action space (metric compact); $p_t(dy|x,a)$ is the continuous transition probability, that is $\int_X c(y)p_t(dy|x,a)$ is a continuous function for each continuous bounded function $c(\cdot)$. As usual a control strategy π is a sequence of measurable stochastic kernels $\pi_t(da|h_{t-1})$ on A where $h_{t-1} = (x_0, a_1, x_1, \dots, a_{t-1}, x_{t-1})$. A strategy is called Markov if it is of the form $\pi_t(da|h_{t-1}) = \pi_t^m(da|x_{t-1})$ and is called stationary if $\pi_t(da|h_{t-1}) = \pi^s(da|x_{t-1})$. A strategy is called pure if each stochastic kernel π_t is concentrated at the point $\varphi(h_{t-1})$. Similarly, a measurable function $\varphi(x_{t-1})$ $(\varphi_t(x_{t-1}))$ defines the pure stationary (Markov) strategy.

It is well known (Piunovskiy, 1997; Bertsekas and Shreve, 1978) that for a fixed initial probability distribution $P_0(dx) \in P(X)$, each strategy defines the unique probability measure P^π on the trajectories space $H_\infty = X \times (A \times X)^\infty$, whose generic element will be denoted as h. Here and further, $P(Y)$ is the space of all probability measures on a Borel space Y, equipped with the weak topology. The integral with respect to the measure P^π is denoted by E^π.

The traditional optimal control problem consists of the minimization of the following functional

$$\mathbf{R}(\pi) = E^\pi[R(h)] = E^\pi\left[\sum_{t=1}^{\infty} \beta_0^{t-1} r_t(x_{t-1}, a_t)\right] \longrightarrow \inf_{\pi\in\Pi}, \tag{1}$$

where $r_t(\cdot)$ is a one-step cost function and $\beta_0 > 0$ is a discount factor; Π is the set of all strategies. If there are no costs in (1) beyond time T then one puts $\beta_0 = 1$ (the case of a finite horizon). If the cost function r and the transition probability p do not depend on time, then we deal with the homogeneous model.

Let us assume that $r_t(x,a)$ is a lower-semicontinuous lower-bounded function and the transition probability $p_t(dy|x,a)$ is continuous. Suppose that a lower-semicontinuous lower-bounded function $s_t(x,a)$ is given as well as the discount factor $\beta_1 > 0$ and a real number d. A strategy π is called feasible if the following inequality is satisfied

$$S(h) = \sum_{t=1}^{\infty} \beta_1^{t-1} s_t(x_{t-1}, a_t) \le d \tag{2}$$

P^π-almost surely. In what follows, the expressions (1) and (2) are assumed to be well defined. To be more specific, we study either the model with a finite horizon, or the case of a discounted model $\beta_1 \in (0,1)$. One must build an optimal feasible strategy; in other words, one must solve problem (1) under constraints (2).

Remark 2.1 *(a) We intend to investigate Version 2 for the scalar function S, but the case $S \in \mathbb{R}^N$ can be treated basically in the same way.*

(b) One step loss $s_t(x,a)$ can be interpreted as the value of some resource; the total discounted resource should not exceed d.

If the stationary strategy $\pi^*(da|x_{t-1})$ is optimal in the homogeneous one-step model ($T = 1$) then it is called myopic. Sometimes a myopic strategy is optimal in the homogeneous model, independently on the length of the planning horizon.

Lemma 2.1 *Consider a homogeneous unconstrained model. If a myopic strategy π^* is optimal for any horizon T then it is optimal in the discounted version of the model, under any $\beta_0 \in (0, 1)$.*

Proof. The proof is really simple, but the author could not find an appropriate reference. Clearly, under any fixed $\beta_0 \in (0, 1)$, $\mathbf{R}(\pi) = E^T[R(\pi, T)]$, where

$$R(\pi, T) = E^\pi \left[\sum_{t=1}^{T} r(x_{t-1}, a_t) \right],$$

E^T is the expectation with respect to T, and T is the random variable with distribution

$$P\{T = i\} = (1 - \beta_0)\beta_0^{i-1}, \qquad i = 1, 2, \ldots$$

Now $\forall \pi$, $\forall T$, $R(\pi, T) \geq R(\pi^*, T)$; hence $\mathbf{R}(\pi) \geq \mathbf{R}(\pi^*)$. ■

One is tempted to think that the converse is also true, but that is not the case.

Counter example. (See Fig. 1.) Let $X = \{1, 2, 3\}$, $A = \{1, 2\}$, $p(1|1, 1) = 1$, $p(2|1, 2) = 1$, $p(3|2, a) = 1$, $p(3|3, a) = 1$, $r(1, 1) = -2$, $r(1, 2) = -3$, $r(2, a) = 0$, $r(3, a) = -3$.

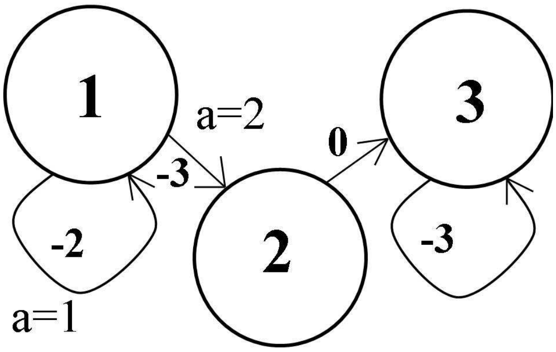

Figure 1. Counter-example.

Then the strategy $\varphi(x) \equiv 2$ is myopic and it is optimal in the discounted model, under any value of $\beta_0 \in (0, 1)$: the Bellman function

$$v(x) \stackrel{\triangle}{=} \inf_{\pi \in \Pi} E^\pi[R(h)|x_0 = x]$$

equals

$$v(1) = -3 - \frac{3\beta_0^2}{1-\beta_0}; \quad v(2) = -\frac{3\beta_0}{1-\beta_0}; \quad v(3) = -\frac{3}{1-\beta_0}.$$

At the same time, obviously, in the two-step model, with $T = 2$ and $\beta_0 = 1$, the Markov pure strategy $\varphi_1(x) = 1$, $\varphi_2(x) = 2$ is optimal, and the myopic strategy is not optimal.

3. Dynamic programming approach

It will be convenient to assume that $s_t(\cdot) \geq 0$, $d \geq 0$, $r_t(\cdot) \geq 0$. Obviously, in the cases of a finite horizon and of a discounted model with $\beta_n \in (0,1), n = 0, 1$, this assumption is just a technicality.

The main concept we deploy here is similar to the penalty function method. A new model is built in which the losses generated by non-feasible strategies are equal to $+\infty$. If a strategy is feasible then the value of the main functional (1) does not change.

The state in the new model is the pair (x_t, W_t), where W_t is the accumulated loss generated by the function s:

$$W_0 = 0; \quad W_t = W_{t-1} + \beta_1^{t-1} \cdot s_t(x_{t-1}, a_t) = W_t(W_{t-1}, x_{t-1}, a_t). \qquad (3)$$

It remains only to adjust the loss function r:

$$\tilde{r}_t(x, W, a) = \begin{cases} \beta_0^{t-1} r_t(x, a), & \text{if } W \leq d, \\ +\infty & \text{otherwise.} \end{cases}$$

There is a one-to-one correspondence between the strategies in the initial model and in the new one; if $\pi \leftrightarrow \tilde{\pi}$ then

$$\tilde{\mathbf{R}}(\tilde{\pi}) = \begin{cases} \mathbf{R}(\pi) & \text{if (2) holds,} \\ +\infty & \text{otherwise.} \end{cases}$$

Therefore, it is sufficient to solve problem

$$\tilde{\mathbf{R}}(\tilde{\pi}) = E^{\tilde{\pi}} \left[\sum_{t=1}^{\infty} \tilde{r}_t(x_{t-1}, W_{t-1}, a_t) \right] \longrightarrow \inf_{\tilde{\pi}}. \qquad (4)$$

Assume that the function $s_t(\cdot)$ is finite and continuous. Then the mapping W_t in (3) is continuous. Under all previously made assumptions, the new (tilde) model is 'semicontinuous' and the following statement holds (Bertsekas and Shreve, 1978).

Statement 3.1 *(a) Problem (4) is solvable: there exists an optimal pure Markov strategy $\varphi_t^*(x_{t-1}, W_{t-1})$. (Note that in terms of the initial problem (1),(2), this strategy is not Markov.)*

(b) Bellman equation for (4) is as follows

$$v_t(x,W) = \inf_{a\in A}\left\{\tilde r_{t+1}(x,W,a) + \int_X p_{t+1}(dy|x,a)v_{t+1}(y, W+\beta_1^t s_{t+1}(x,a)\right\}, \quad t \ge 0. \tag{5}$$

(c) The Bellman function coincides with the minimal non-negative solution to (5) that can be obtained by the successive approximations

$$v_t^0 \equiv 0,$$

$$v_t^{k+1}(x,W) = \inf_{a\in A}\{\tilde r_{t+1}(x,W,a) + \int_X p_{t+1}(dy|x,a)v_{t+1}^k(y,W + \beta_1^t s_{t+1}(x,a)\}, \quad t \ge 0, \quad k = 0,1,2,\ldots$$

(d) Function $v_t(x,W)$ is non-negative and lower-semicontinuous.

Note that the Bellman function can be degenerate, that is $v_t(x,0)=\infty$, in which case there would be no feasible strategies for $x_0 = x$ in the original model.

Recall that a Markov (pure) strategy $a_{t+1} = \varphi_{t+1}(x_t, W_t)$ is optimal in problem (4) iff for each $t \ge 0$ it provides the infimum in (5) for P^φ-almost all values (x_t, W_t). Therefore, the Dynamic Programming Approach makes it possible to build ALL optimal Markov strategies.

If we deal with the model with the finite horizon T then the sequence v^k converges in a finite number of steps: $v^{T+1} = v^\infty$.

Let us consider the homogeneous case when all the cost functions and the transitional probabilities do not depend on t. We introduce the new variable

$$\hat d_t \triangleq \frac{d - W_t}{\beta_1^t}.$$

It equals the accumulated loss generated by s, which is feasible on the remaining interval $\{t+1, t+2, \ldots\}$. The stage loss function, r, can be rewritten in the form

$$\hat r(x, \hat d, a) = \begin{cases} r(x,a), & \text{if } \hat d \ge 0; \\ +\infty & \text{otherwise.} \end{cases}$$

We deal with the standard discounted model

$$\hat{\mathbf{R}}(\hat\pi) = E^{\hat\pi}\left[\sum_{t=1}^\infty \beta_0^{t-1}\hat r(x_{t-1}, \hat d_{t-1}, a_t)\right] \longrightarrow \inf_{\hat\pi}, \tag{6}$$

assuming that $\beta_0 \in (0,1)$, where the dynamics of the component $\hat d$ is defined by the following equation

$$\hat d_t = \frac{1}{\beta_1}\left[\frac{d - W_{t-1} - \beta_1^{t-1}s(x_{t-1},a_t)}{\beta_1^{t-1}}\right]$$
$$= \frac{\hat d_{t-1} - s(x_{t-1},a_t)}{\beta_1} \triangleq D(\hat d_{t-1}, x_{t-1}, a_t).$$

The initial values come from the original problem: $\hat{d}_0 = d$. The 'hat' strategies $\hat{\pi}$ are defined in the standard way; in fact, one can replace them with $\tilde{\pi}$. Note that $\beta_1 > 0$ can be arbitrary.

The Bellman equation for problem (6) is of the form

$$\hat{v}(x,\hat{d}) = \inf_{a\in A}\left\{\hat{r}(x,\hat{d},a) + \beta_0 \int_X p(dy|x,a)\hat{v}(y, D(\hat{d},x,a))\right\}, \tag{7}$$

and we are interested in its minimal non-negative solution. Note that equation (7) is in agreement with the Bellman equation obtained in Chen and Blankenship (2004), where the model with finite state and action spaces was considered.

Equation (7) can also be solved by the successive approximations method. It is simpler than equation (5) since the time-dependence is absent.

In practice, it is sometimes possible to build the domain $\overline{G}$ where $\hat{v}(P,\hat{d}) = \infty$ (there are no feasible strategies). One can show that $\overline{G}$ is an open set. Let

$$G = \{(x,\hat{d}) \in X \times \mathbb{R} : \quad \hat{v}(x,\hat{d}) < +\infty\} = \{X \times \mathbb{R}\} \setminus \overline{G}.$$

Then, for every pair $(x,\hat{d}) \in G$, there exists $a \in A$ such that $(y, D(\hat{d},x,a)) \in G$ almost surely wrt $p(\cdot|x,a)$. If function $r(\cdot)$ is bounded, then equation (7) has a unique lower-semicontinuous uniformly bounded solution on G. (The Bellman operator on the right-hand side of the equation is a contraction mapping in the space of lower-semicontinuous bounded functions on G.) It should be emphasized that this solution, extended by infinity on $\overline{G}$, provides the minimal nonnegative solution of equation (5), using

$$v_t(x,W) = \beta_0^t \hat{v}\left(x, \frac{d-W}{\beta_1^t}\right).$$

Equation (7) cannot have any other bounded solutions on G. In contrast, if v is a solution of (5) then $v+c$ is also the solution of equation (5) for any constant c.

Remark 3.1 *Exactly the same reasoning holds in the case of finite horizon problem ($\beta_i \equiv 1$):*

$$E^{\pi}\left[\sum_{t=1}^{T} r_t(x_{t-1},a_t)\right] \longrightarrow \inf_{\pi}, \qquad \sum_{t=1}^{T} s_t(x_{t-1},a_t) \le d.$$

We deal with Bellman equation

$$\hat{v}_t(x,\hat{d}) = \inf_{a\in A}\left\{\hat{r}_{t+1}(x,\hat{d},a) + \int_X p_{t+1}(dy|x,a)\hat{v}_{t+1}(y, \hat{d} - s_{t+1}(x,a))\right\};$$
$$\hat{v}_T(x,\hat{d}) = 0$$

corresponding to problem

$$E^{\hat{\pi}}\left[\sum_{t=1}^{T} \hat{r}(x_{t-1},\hat{d}_{t-1},a_t)\right] \longrightarrow \inf_{\hat{\pi}}. \tag{8}$$

4. Examples

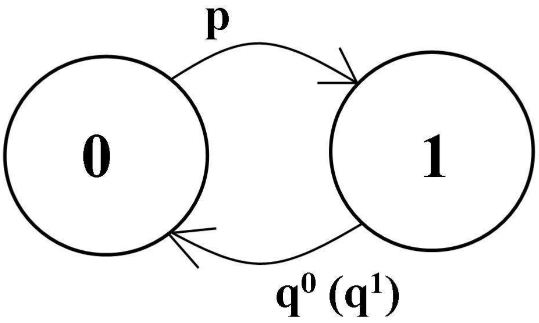

Figure 2. Transition probabilities.

Let us consider the one-channel Markov queueing system with losses. (See Fig. 2.) Set $X = \{0, 1\}$ where $x_t = 0$ ($x_t = 1$) means that the system is free (busy) at time t; $A = \{0, 1\}$ where $a_t = 0$ ($a_t = 1$) means that the system effects less intensive (more intensive) servicing at the interval $(t - 1, t]$. The initial probability $P_0(1)$ that the system is busy, is known. The transition probability at stage t is expressed by the formula

$$p_t(y|x, a) = \begin{cases} p, & \text{if } x = 0,\ y = 1; \\ 1 - p, & \text{if } x = 0,\ y = 0; \\ q^a, & \text{if } x = 1,\ y = 0; \\ 1 - q^a, & \text{if } x = 1,\ y = 1. \end{cases}$$

Here, p is the probability of a customer arriving in the interval $(t - 1, t]$; q^0 (q^1) is the probability that the service will be completed in the interval $(t - 1, t]$ for the less (more) intensive regime; $0 < q^0 < q^1 < 1$. The more intensive regime is connected with additional cost e. Lastly, $c > 0$ is the penalty caused by the loss of an order which is paid off only if a customer came into the busy system and was rejected. We have to minimise these discounted expected penalties under the a.s.-constraint on the service consumption. Therefore, we put

$$r(x, a) = xpc; \qquad s(x, a) = e \cdot I\{a = 1\},$$

where $I\{\cdot\}$ is the characteristic function.

First, let us consider the case where $\beta_1 = 1$ and $\beta_0 \in (0, 1)$:

$$\left.\begin{array}{l} E^\pi \left[\sum_{t=1}^{\infty} \beta_0^{t-1} r(x_{t-1}, a_t) \right] \longrightarrow \inf_\pi \\ \sum_{t=1}^{\infty} s(x_{t-1}, a_t) \le d \qquad P^\pi - \text{a.s.} \end{array}\right\} \tag{9}$$

STATEMENT 4.1 *[proof in the appendix] Solution to problem (9) is given by the myopic control strategy of the form*

$$a_t = \varphi^*(x_{t-1}, \hat{d}_{t-1}) = x_{t-1} \cdot I\{\hat{d}_{t-1} \geq e\}. \tag{10}$$

Here, as usual,

$$\hat{d}_t = d - W_t; \qquad W_t = \sum_{i=0}^{t} s(x_{i-1}, a_i).$$

Note that φ^* is the myopic strategy in the 'hat' model (see (6) and (8)).

REMARK 4.1 *Analysing the discounted Dynamic Programming equation, it can be shown that no other control strategy provides a solution to (9).*

Now, consider the more general situation:

$$\left.\begin{array}{l} \mathbf{R}(\pi) = E^\pi \left[\sum_{t=1}^{\infty} \beta_0^{t-1} r(x_{t-1}, a_t) \right] \longrightarrow \inf_\pi; \\ \sum_{t=1}^{\infty} \beta_1^{t-1} s(x_{t-1}, a_t) \leq d \quad P^\pi - \text{a.s.}, \end{array}\right\} \tag{11}$$

where $\beta_1 > 0$ can be arbitrary.

STATEMENT 4.2 *[proof in the appendix] If $\beta_0 \leq \frac{1}{2}$ then (for any $\beta_1 > 0$) myopic control strategy (10) solves problem (11).*

Note that here $\hat{d}_t = \frac{d - W_t}{\beta_1^t}$, $W_t = \sum_{i=1}^{t} \beta_1^{i-1} s(x_{i-1}, a_i)$.

REMARK 4.2 *In case $d > 0$, but not large enough to make the constraint redundant, inequalities (16) and (17) are strict, meaning that no other control strategy is optimal in problem (11).*

5. Other constrained problems and example

If one considers Version 1:

$$\mathbf{R}(\pi) = E^\pi \left[\sum_{t=1}^{\infty} \beta_0^{t-1} r_t(x_{t-1}, a_t) \right] \longrightarrow \inf_\pi;$$

$$\mathbf{S}(\pi) = E^\pi \left[\sum_{t=1}^{\infty} \beta_1^{t-1} s_t(x_{t-1}, a_t) \right] \leq d,$$

the theory becomes more complicated (Altman, 1999; Piunovskiy, 1997; Feinberg and Shwartz, 1999; Feinberg and Shwartz, 1995). The main theoretical tool in this case is 'Convex Analytic Approach' and/or Linear Programming.

The Dynamic Programming Approach was used only in Chen and Blankenship (2004), Piunovskiy and Mao (2000). Similarly to Section 3, where we passed to pairs (x, W) or $(x, \hat{d})$, one should extend the notion of the state. In Chen and Blankenship (2004) the authors suggest a similar extension and introduce new actions of the form $(a, \gamma(\cdot))$, where $\gamma(y)$ is the new allowed value of the threshold at the next step, if y is the next state of the process. In Piunovskiy and Mao (2000) the new state is the pair $(P, \hat{d})$, where P is the probability distribution on X; and the action is the transition probability from X to A. From the computational point of view, both approaches are almost equivalent; perhaps the method suggested in Chen and Blankenship (2004) is a little more economical. An example similar to those presented in Section 4 was studied in Piunovskiy and Mao (2000). In the current section, we give the solution to a slightly different modification, consistent with the examples from Section 4. Namely, we consider the same queuing system with the same loss functions and with $\beta_0 = \beta_1 = \beta$.

As was shown in Piunovskiy and Mao (2000), the sufficient statistics (i.e. the arguments of the Bellman function) are the current probability of the system being busy, $P(1)$, and the vector of expected feasible losses $\hat{d}$. The initial values $P_0(1)$ and $\hat{d}_0 = d$ are given. The action is the (conditional) probability of applying $a = 1$ in state 1, denoted as $\tilde{a}(1|1) \in [0, 1]$. The Bellman equation for the newly constructed model, similar to (7) or, more specifically, to (15), is as follows:

$$\begin{aligned}\hat{v}(P(1), \hat{d}) = &\inf_{0 \le \tilde{a}(1|1) \le \min\{\frac{\hat{d}}{P(1)e};\ 1\}} \Bigg\{ P(1)pc \\ &+\beta\hat{v}\left(p - P(1)[p - 1 + q^0 + \tilde{a}(1|1)(q^1 - q^0)],\ \ \frac{\hat{d} - P(1)\tilde{a}(1|1)e}{\beta}\right)\Bigg\}.\end{aligned} \tag{12}$$

All the details can be found in Piunovskiy and Mao (2000).

If $d < 0$ there are no feasible strategies and $\hat{v}(P(1), \hat{d}) = +\infty$, if $\hat{d} < 0$.

If $d \ge \frac{e\beta p + (1-\beta)eP_0(1)}{(1-\beta)(1-\beta+\beta q^1+\beta p)}$ then the constraint is inessential and $\varphi^*(1) \equiv 1$. In the region where this inequality holds for $(P(1), \hat{d})$, we have $\hat{v}(P(1), \hat{d}) = \frac{pc(P(1)(1-\beta)+\beta p)}{(1-\beta)(1-\beta+\beta q^1+\beta p)}$.

Suppose

$$0 \le d \le \frac{e\beta p + (1-\beta)eP_0(1)}{(1-\beta)(1-\beta+\beta q^1+\beta p)}.$$

In the region where this inequality holds for $(P(1), \hat{d})$, the solution to the Bellman equation (12) has the form

$$\hat{v}(P(1), \hat{d}) = \frac{pc[\beta p + P(1)(1-\beta) - \beta(1-\beta)(q^1 - q^0)\hat{d}/e]}{(1-\beta)(1-\beta+\beta q^0+\beta p)}.$$

In this region, one can choose

$$\tilde{a}(1|1) \in \left[\max\left\{0;\ \frac{\hat{d}(1-\beta+\beta q^1+\beta p)}{P(1)e(1-\beta+\beta q^0+\beta p)} + \frac{\beta(p-1+q^0)}{1-\beta+\beta q^0+\beta p} - \frac{\beta p}{P(1)(1-\beta)(1-\beta+\beta q^0+\beta p)}\right\},\quad \min\left\{\frac{\hat{d}}{P(1)e};\ 1\right\}\right]$$

arbitrarily. All these actions provide the minimum in the Bellman equation (12). This means that there exist many optimal control strategies in the constrained optimal control problem for the queuing system under consideration. (Compare with Remarks 4.1 and 4.2.)

Dynamic Programming approach to Versions 3 and 4 is more complicated (Chen, 2004; Chen and Blankenship, 2004). For example in Chen and Blankenship (2004), Version 3 is handled with the aid of an extended state of the form $(x, \hat{d}, \hat{\alpha})$, where $\hat{d}$ is the value of the threshold, $\hat{\alpha}$ is the value of probability; the new action is the pair $(a, \gamma(\cdot))$, where $\gamma(y)$, like previously, is the new allowed value of the threshold.

Note that all constraints in versions 1-3 can be expressed in terms of expectations: $P^\pi[S(h) \leq d] = E^\pi[I\{S(h) - d\}]$; Version 2 coincides with Version 3 at $\alpha = 1$. In contrast, the variance that is present in Version 4, contains the square of expectation. Such problems were studied in Sniedovich (1980), Krawczyk (1990). For example, in Sniedovich (1980) the following method is suggested for calculation of the Lagrange function

$$L(\lambda) = \inf_{\pi\in\Pi}\{E^\pi[R(h)] + \lambda\ Var^\pi[S(h)]\} :$$

$$L(\lambda) = \min_{u\in\mathbf{R}^1}\inf_{\pi\in\Pi}\{E^\pi[R(h) + \lambda(S(h)-u)^2]\}.$$

If functions $R(\cdot)$ and $S(\cdot)$ are additive, like (1) and (2), the infimum with respect to π can be again found using Dynamic Programming.

6. Conclusion

The Dynamic Programming Approach remains effective in nonstandard (constrained) optimization problems. Moreover, in contrast to the Convex Analytic Approach, it is possible to construct ALL optimal Markov strategies. However, this may require modifications in the standard model such as extension of the state variable, introduction of suitable penalties and so on.

In the last sections, we investigated a simple controlled queue and found that the myopic control strategy is the only optimal one for the Version 2. In contrast, typically Version 1 has many optimal solutions. It seems, that this is the case in many other problem instances.

Acknowledgement

The author is thankful to the referee for helpful remarks.

Appendix

Proof of Statement 4.1. First, consider the undiscounted finite-horizon model:

$$E^{\pi}\left[\sum_{t=1}^{T} r(x_{t-1},a_t)\right] \longrightarrow \inf_{\pi}, \tag{13}$$

$$\sum_{t=1}^{T} s(x_{t-1},a_t) \le d, \quad P^{\pi} - \text{a.s.} \tag{14}$$

and prove that the myopic strategy is optimal. (We also denote it with φ^*.) To put it differently, we should apply action $a_t = 1$ always, if $x_{t-1} = 1$ and if we do not violate constraint (14).

Strategy φ^* is obviously optimal, if $T = 1$, so that it is indeed myopic. Suppose it is optimal for some $T \ge 1$ and consider the case $T + 1$. We must analyse only the situation $x_0 = 1$; starting from $t = 2$, the myopic strategy is optimal by induction. The value of $d \ge e$ is fixed.

(i) Suppose $\tau \triangleq \min\{t : \sum_{i=1}^{t} s(x_{t-1},a_t) + e > d\} < T + 1$ and $\tau_1 \triangleq \min\{t > \tau : x_t = 1\} < T + 1$. This condition is denoted by C. Under C, it does not matter whether we apply $a_1 = 0$ or $a_1 = 1$. Indeed, let $a_1 = 0$, and the control strategy is myopic thereafter. Then there are four different types of trajectories:

$$\text{Type 1:}\quad x_0 = 1 \xrightarrow{q^0} x_1 = 0 \longrightarrow \ldots \longrightarrow x_{\tau_1} = 1 \xrightarrow{q^1} x_{\tau_1+1} = 0 \longrightarrow \ldots$$

$$\text{Type 2:}\quad x_0 = 1 \xrightarrow{q^0} x_1 = 0 \longrightarrow \ldots \longrightarrow x_{\tau_1} = 1 \xrightarrow{1-q^1} x_{\tau_1+1} = 1 \longrightarrow \ldots$$

$$\text{Type 3:}\quad x_0 = 1 \xrightarrow{1-q^0} x_1 = 1 \longrightarrow \ldots \longrightarrow x_{\tau_1} = 1 \xrightarrow{q^1} x_{\tau_1+1} = 0 \longrightarrow \ldots$$

$$\text{Type 4:}\quad x_0 = 1 \xrightarrow{1-q^0} x_1 = 1 \longrightarrow \ldots \longrightarrow x_{\tau_1} = 1 \xrightarrow{1-q^1} x_{\tau_1+1} = 0 \longrightarrow \ldots$$

In the case $a_1 = 1$, the marks $q^0(q^1)$ should be changed to $q^1(q^0)$. It is essential that all unmarked arrows have the same probabilities, independently of a_1. Hence, all trajectories of type 1 and 4 have the same probabilities in both cases $a_1 = 0$ and $a_1 = 1$. As for type 2 and 3, there exists the trivial 1-1 correspondence between trajectories-2 and trajectories-3 which obviously preserves the total amount of 'ones'. The probabilities of the image (under $a_1 = 1$) and the preimage (under $a_1 = 0$) coincide. Therefore,

$$E^{0,\varphi^*}\left[\sum_{t=1}^{T+1} r(x_{t-1},a_t)|C\right] = E^{1,\varphi^*}\left[\sum_{t=1}^{T+1} r(x_{t-1},a_t)|C\right].$$

Here and below, $[0, \varphi^*]$ is the natural composition of $a_1 = 0$ and the myopic strategy φ^* at $t = 2, 3, \dots$.

(ii) Suppose Condition C is violated. Now we have only two types of trajectories:

$$x_0 = 1 \longrightarrow x_1 = 0 \longrightarrow \dots$$
$$x_0 = 1 \longrightarrow x_1 = 1 \longrightarrow \dots$$

The first arrow in the first trajectory has probability $q^0 (q^1)$ in case $a_1 = 0$ $(a_1 = 1)$. The first arrow in the second trajectory has probability $1 - q^0$ $(1 - q^1)$ in case $a_1 = 0$ $(a_1 = 1)$. All the remaining arrows have the same probabilities, independently of a_1. Since the main loss (13) is larger for the second type, and $q^0 < q^1$, we conclude that

$$E^{0,\varphi^*}\left[\sum_{t=1}^{T+1} r(x_{t-1}, a_t)|\bar{C}\right] > E^{1,\varphi^*}\left[\sum_{t=1}^{T+1} r(x_{t-1}, a_t)|\bar{C}\right],$$

and the myopic strategy is optimal in case $T + 1$ for problem (13),(14). Note that no other strategy is optimal, if $T > 1$; in case $T = 1$, all strategies are optimal (if $d \geq e$).

In accordance with Remark 3.1 in Section 3, the myopic control strategy (10) is optimal in problem (8). According to Lemma 2.1, it is also optimal in problem

$$E^{\hat{\pi}}\left[\sum_{t=1}^{\infty} \beta_0^{t-1} \hat{r}(x_{t-1}, \hat{d}_{t-1}, a_t)\right] \longrightarrow \min_{\hat{\pi}}$$

which is equivalent to problem (9). ■

Proof of Statement 4.2. Clearly, if $d = 0$, then the only feasible strategy is given by $a_t \equiv 0$, for which $\mathbf{R} = P_0(0)W(0) + P_0(1)W(1)$, where $W(\cdot)$ satisfies equations

$$\left.\begin{array}{l} W(0) = \beta_0[pW(1) + (1-p)W(0)]; \\ W(1) = pc + \beta_0[q^0 W(0) + (1 - q^0)W(1)]. \end{array}\right\}$$

The actual Bellman function, under arbitrary fixed $d \geq 0$, satisfies the Bellman equation (see (7))

$$
\begin{aligned}
\hat{v}(0,\hat{d}) = \quad & \min\left\{\beta_0\left[p\hat{v}\left(1,\frac{\hat{d}}{\beta_1}\right)+(1-p)\hat{v}\left(0,\frac{\hat{d}}{\beta_1}\right)\right];\right.\\
& \left.\beta_0\left[p\hat{v}\left(1,\frac{\hat{d}-e}{\beta_1}\right)+(1-p)\hat{v}\left(0,\frac{\hat{d}-e}{\beta_1}\right)\right]\right\}\\
\hat{v}(1,\hat{d}) = \quad & \min\left\{pc+\beta_0\left[q^0\hat{v}\left(0,\frac{\hat{d}}{\beta_1}\right)+(1-q^0)\hat{v}\left(1,\frac{\hat{d}}{\beta_1}\right)\right];\right.\\
& \left.pc+\beta_0\left[q^1\hat{v}\left(0,\frac{\hat{d}-e}{\beta_1}\right)+(1-q^1)\hat{v}\left(1,\frac{\hat{d}-e}{\beta_1}\right)\right]\right\}.
\end{aligned}
\qquad (15)
$$

It is bounded above by $W(\cdot)$:

$$
\begin{aligned}
\hat{v}(0,\hat{d}) \leq W(0) = \frac{p^2\beta_0 c}{(1-\beta_0)(1-\beta_0+\beta_0 q^0+\beta_0 p)};\\
\hat{v}(1,\hat{d}) \leq W(1) = \frac{pc(1-\beta_0+\beta_0 p)}{(1-\beta_0)(1-\beta_0+\beta_0 q^0+\beta_0 p)}.
\end{aligned}
\qquad (16)
$$

Here $D(\hat{d},x,a) = D(\hat{d},a) = \begin{cases} \hat{d}/\beta_1, & \text{if } a=0;\\ (\hat{d}-e)/\beta_1, & \text{if } a=1.\end{cases}$

Note that the first minimum in (15) is provided by the first expression corresponding to $a=0$ because functions $\hat{v}(0,\hat{d})$ and $\hat{v}(1,\hat{d})$ are non-increasing in $\hat{d}$. (This can be easily proved using the successive approximations discussed in Section 3.) Of course, we consider only the case $\hat{d} \geq 0$, since otherwise $\hat{v}(x,\hat{d}) = +\infty$.

A sufficiently large value of d makes the constraint superfluous and leads to the lower bound

$$
\begin{aligned}
\hat{v}(0,\hat{d}) \geq \frac{p^2\beta_0 c}{(1-\beta_0)(1-\beta_0+\beta_0 q^1+\beta_0 p)};\\
\hat{v}(1,\hat{d}) \geq \frac{pc(1-\beta_0+\beta_0 p)}{(1-\beta_0)(1-\beta_0+\beta_0 q^1+\beta_0 p)}.
\end{aligned}
\qquad (17)
$$

Now it follows from (16) and (17) that, for any value $\hat{d} \geq e$,

$$
\begin{aligned}
& q^0 \hat{v}\left(0, \frac{\hat{d}}{\beta_1}\right) + (1-q^0)\hat{v}\left(1, \frac{\hat{d}}{\beta_1}\right) - q^1 \hat{v}\left(0, \frac{\hat{d}-e}{\beta_1}\right) + (1-q^1)\hat{v}\left(1, \frac{\hat{d}-e}{\beta_1}\right) \\
& \quad \geq \frac{q^0 p^2 \beta_0 c}{(1-\beta_0)(1-\beta_0+\beta_0 q^1+\beta_0 p)} + \frac{(1-q^0)pc(1-\beta_0+\beta p)}{(1-\beta_0)(1-\beta_0+\beta_0 q^1+\beta_0 p)} \\
& \quad -\frac{q^1 p^2 \beta_0 c}{(1-\beta_0)(1-\beta_0+\beta_0 q^0+\beta_0 p)} - \frac{(1-q^1)pc(1-\beta_0+\beta p)}{(1-\beta_0)(1-\beta_0+\beta_0 q^0+\beta_0 p)} \\
& \quad = \frac{pc(q^1-q^0)[(1-2\beta_0)(1-\beta_0+\beta_0 p)+\beta_0(1-\beta_0)(q^0+q^1)]}{(1-\beta_0)(1-\beta_0+\beta_0 q^1+\beta_0 p)(1-\beta_0+\beta_0 q^0+\beta_0 p)} \geq 0,
\end{aligned}
$$

and the myopic strategy satisfies Bellman equation. ■

References

Altman, E. (1999) *Constrained Markov Decision Processes.* Chapman and Hall, London.

Bertsekas, D.P. and Shreve, S.E. (1978) *Stochastic Optimal Control.* Academic Press, New York.

Chen, R. (2004) Constrained stochastic control with probabilistic criteria and search optimization. *Preprints of the 43-th IEEE Conf. on Decision and Control*, Bahamas.

Chen, R.C. and Blankenship, G.L. (2004) Dynamic programming equations for discounted constrained stochastic control. *IEEE Trans. on Aut. Control* **49**, 699–709.

Feinberg, E.A. and Shwartz, A. (1995) Constrained Markov decision models with discounted rewards. *Math. of Oper. Res.* **20**, 302–320.

Feinberg, E.A. and Shwartz, A. (1999) Constrained dynamic programming with two discount factors: applications and an algorithm. *IEEE Trans. on Aut. Control* **44**, 628–631.

Krawczyk, J.B. (1990) On variance constrained programming. *Asia-Pacific J. of Oper. Res.* **7**, 190–206.

Piunovskiy, A.B. (1997) *Optimal Control of Random Sequences in Problems with Constraints.* Kluwer Academic Publishers, Dordrecht-Boston-London.

Piunovskiy, A.B. (1998) Controlled random sequences: the convex analytic approach and constrained problems. *Russ. Math. Surveys* **53**, 1233–1293.

Piunovskiy, A.B. and Mao, X. (2000) Constrained Markov decision processes: the dynamic programming approach. *Oper. Res. Letters* **27**, 119-126.

Sniedovich, M. (1980) A variance-constrained reservoir control problem. *Water Resources Research* **16**, 271–274.

Tanaka, K. (1991) On discounted dynamic programming with constraints. *J. of Mathem. Anal. and Appl.* **155**, 264–277.

Yakowitz, S. (1982) Dynamic programming applications in water resources. *Water Resources Research* **18**, 673-696.