6.231 Dynamic Programming and Optimal Control
Midterm Exam, Fall 2015
Prof. Dimitri Bertsekas

Problem 1 (50 points)

Consider the scalar linear system $x_{k+1} = ax_k + bu_k$, where the nonzero scalars a and b are known. At each period k we have the option of using a control u_k and incurring a cost $qx_k^2 + ru_k^2$, or else stopping and incurring a stopping cost tx_k^2 (we may have $t \neq q$). At the final period N, if we have not already stopped, the terminal cost is the same as the stopping cost, i.e., $g_N(x_N) = tx_N^2$. We assume that the scalars q, r, and t are positive.

(a) Consider first the restricted optimization over policies that never stop, except at time N, so we obtain a standard linear quadratic problem, whose optimal cost function has the form

$$J_0(x_0) = K_0 x_0^2,$$

where K_0 is a positive scalar that depends on N. Write the Riccati equation that yields K_0 as well as the steady-state equation that has $\bar{K}$, the limit of K_0 as $N \to \infty$, as its solution. Does the steady-state equation have any other solutions?

(b) Consider the unrestricted optimization where stopping is also allowed. Write the DP algorithm for this problem.

(c) Show that for the problem of part (b) there is a threshold value $\bar{t}$ such that if $t > \bar{t}$ immediate stopping is optimal at every state, and if $t < \bar{t}$ continuing at every state x_k and period k is optimal. How are the scalars $\bar{K}$ and $\bar{t}$ of parts (a) and (c) related?

(d) State an extension of the result of part (c) for the case of a multidimensional system.

Solution: (a) Consider the restricted version of the problem where no stopping is allowed. The cost-to-go functions for this problem solve the system:

$$J_k(x_k) = \min_{u_k} \big\{ qx_k^2 + ru_k^2 + J_{k+1}(ax_k + bu_k) \big\}$$

and $J_N(x_N) = tx_N^2$. These cost-to-go functions are obtained from the Riccati equation. We have

$$V_k(x) = K_k x^2$$

where:

$$K_k = a^2[K_{k+1} - K_{k+1}^2 b^2 (r + b^2 K_{k+1})^{-1}] + q = \frac{a^2 r K_{k+1}}{r + b^2 K_{k+1}} + q,$$

and $K_N = t$. The steady-state equation is

$$K = \frac{a^2 K r}{r + b^2 K} + q, \tag{1}$$

and has two solutions, one positive and one negative (cf. Fig. 4.1.2). The positive solution is $\bar{K}$, the limit of the equation as $k \to -\infty$.

(b) For the original problem, where a stopping control is allowed, the DP algorithm is given by

$$J_N(x_N) = t x_N^2,$$

$$J_k(x_k) = \min\Big[t x_k^2, V_k(x_k)\big\}\Big],$$

where

$$V_k(x_k) = \min_{u_k} \big\{ q x_k^2 + r u_k^2 + J_{k+1}(a x_k + b u_k) \big\}.$$

Assume that

$$\frac{a^2 t r}{r + b^2 t} + q \geq t. \tag{2}$$

Then

$$J_{N-1}(x_{N-1}) = \min\big[t x_{N-1}^2, V_{N-1}(x_{N-1})\big] = t x_{N-1}^2$$

and stopping is optimal for all x_{N-1}. Since

$$J_{N-1}(x) = J_N(x) = t x^2,$$

the argument can be repeated for stage $N-2$ all the way to stage 0.

Now assume that

$$\frac{a^2 t r}{t b^2 + r} + q \leq t. \tag{3}$$

Then

$$V_{N-1}(x_{N-1}) \leq t x_{N-1}^2, \qquad \forall\ x,$$

so not stopping is optimal for all x_{N-1}. Since the system and cost per stage are stationary, the monotonicity property of DP yields

$$J_{N-1}(x_{N-1}) \leq J_N(x_{N-1}), \quad \forall\ x \qquad \Rightarrow \qquad J_k(x) \leq J_{k+1}(x), \quad \forall\ x,\ k$$

Assume that stopping is not optimal for all x_k. Then

$$\begin{aligned} t x^2 &\geq \min_u \big\{ q x^2 + r u^2 + J_{k+1}(a x + b u) \big\} \\ &\geq \min_u \big\{ q x^2 + r u^2 + J_k(a x + b u) \big\} \\ &= V_{k-1}(x), \end{aligned}$$

so stopping is not optimal for all x_{k-1}.

(c) The threshold value $\bar{t}$ is the positive solution of the equation

$$\frac{a^2tr}{r + b^2t} + q = t.$$

Thus $\bar{t}$ is equal to $\bar{K}$ [cf. Eqs. (1)-(3)].

(d) Consider a quadratic stopping cost of the form $x'Tx$ where T is positive definite symmetric. Then if $T - \bar{K}$ is negative semidefinite, stopping at every state is optimal, while if $T - \bar{K}$ is positive semidefinite, never stopping is optimal.

Problem 2 (50 points)

Consider a situation involving a blackmailer and his victim. At each stage the blackmailer has a choice of:

(1) Retiring with his accumulated blackmail earnings.

(2) Demanding a payment of \$ 1, in which case the victim will comply with the demand (this happens with probability p, independently of the past history), or will refuse to pay and denounce the blackmailer to the police (this happens with probability $1 - p$).

Once denounced to the police, the blackmailer loses all of his accumulated earnings and cannot blackmail again. Also, the blackmailer will retire once he reaches accumulated earnings of \$ n, where n is a given integer that may be assumed very large for the purposes of this problem. The blackmailer wants to maximize the expected amount of money he ends up with.

(a) Formulate the problem as a stochastic shortest path problem with states $i = 0, 1, \ldots, n$, plus a termination state,

(b) Write Bellman's equation and justify that its unique solution is the optimal value function $J^*(i)$.

(c) Use value iteration to show that $J^*(i)$ is monotonically increasing with i, and that $J^*(i) = i$ for all i larger than a suitable scalar.

(d) Start policy iteration with the policy where the blackmailer retires at every i. Derive the sequence of generated policies and the optimal policy. How many iterations are needed for convergence?

Solution: (a) The state of the SSP problem is the accumulated earnings of the blackmailer, and takes values $i = 0, 1, \ldots, n$. The termination state, denoted t, is reached in two ways: 1) Upon retirement at state i, in which case the reward for the transition from i to t is \$ i, or 2) Upon denouncement to the police, which happens upon blackmail from state i with probability $1 - p$, with transition reward equal to 0. The other transitions are from i

to $i+1$, which happen upon blackmail from state i with probability p, and have reward 0. This is a full description of the SSP problem.

(b) Bellman's equation is

$$J^*(i) = \max\big[i,\, pJ^*(i+1)\big], \qquad i = 0, 1, \ldots, n-1,$$

$$J^*(n) = n.$$

Its unique solution is the optimal value function J^*, since termination occurs within n steps from every initial state and under every policy.

(c) We consider VI starting from the identically zero function. It has the form

$$J_0(i) = 0, \qquad i = 0, 1, \ldots, n,$$

$$J_{k+1}(i) = \begin{cases} \max\big[i,\, pJ_k(i+1)\big], & \text{if } i = 0, 1, \ldots, n-1, \\ n, & \text{if } i = n. \end{cases}$$

We have $J_1(i) = i$ for all i, so $J_1 \geq J_0$, and J_1 is monotonically increasing as a function of i. By monotonicity of the DP mapping, it follows that $J_{k+1} \geq J_k$ for all k, and by induction it is seen that $J_k(i)$ is monotonically increasing as a function of i for all k. Since $J_k(i) \uparrow J^*(i)$, it follows that $J^*(i)$ is also monotonically increasing as a function of i.

Since $J_1(i) = i$ for all i, we have

$$J_2(i) = \begin{cases} \max\big[i,\, p(i+1)\big], & \text{if } i = 0, 1, \ldots, n-1, \\ n, & \text{if } i = n, \end{cases}$$

so $J_2(i) = i$ for $i \geq p(i+1)$ or $i \geq \frac{p}{1-p}$. We use this as the first step in an induction that will show that for all k, we have $J_k(i) = i$ for $i \geq \frac{p}{1-p}$.

Indeed by using the induction hypothesis, we have

$$J_{k+1}(i) = \max\big[i,\, pJ_k(i+1)\big] = \max\big[i,\, p(i+1)\big], \qquad \forall\, i \geq \frac{p}{1-p}.$$

Since we have

$$i \geq \frac{p}{1-p} \qquad \text{if and only if} \qquad \max\big[i,\, p(i+1)\big] = i,$$

it follows that $J_{k+1}(i) = i$ for all $i \geq \frac{p}{1-p}$, thus completing the induction. Since $J_k(i) \uparrow J^*(i)$, it follows that $J^*(i) = i$ for all $i \geq \frac{p}{1-p}$.

(d) With μ^0 being the policy that retires for all i, we have the policy evaluation

$$J_{\mu^0}(i) = i, \qquad i = 0, 1, \ldots, n.$$

In the corresponding policy improvement, we compare i with $pJ_{\mu^0}(i+1) = p(i+1)$, and we obtain that the improved policy μ^1 retires if and only if $i \geq \frac{p}{1-p}$. The policy evaluation of μ^1 yields

$$J_{\mu^1}(i) = i, \qquad \text{if } i \geq \frac{p}{1-p}.$$

In the corresponding policy improvement, we compare i with $pJ_{\mu^1}(i+1)$ which is equal to $p(i+1)$ for $i \geq \frac{p}{1-p}$. It follows if $i \geq \frac{p}{1-p}$, then μ^2 retires. Also $i \leq \frac{p}{1-p}$, we have

$$i \leq p(i+1) \leq pJ_{\mu^1}(i+1),$$

since $J_{\mu^1}(i+1) \geq i+1$ by the fact that μ^1 is an improved policy over μ^0. This shows that if $i \leq \frac{p}{1-p}$, then μ^2 does not retire. Thus μ^2 retires if and only if $i \geq \frac{p}{1-p}$, so it is identical to μ^1. It follows that policy iteration terminates with μ^2, which is an optimal policy.