Lecture notes
Control of Stochastic Processes

Adam Shwartz, Electrical Engineering, Technion, Israel

Version of November 20, 2003
For the latest see
http://www.ee.technion.ac.il/~adam/GRADUATES/048913

Abstract

These are lecture notes for the course: written in LaTeX2$_\varepsilon$ since I need much mathematics, and in English since there is no standard for Hebrew (yet).

1 Introduction

This took place in class no. 1, using notes I posted on the WEB (word-7).

8 Average Cost

Definition 8.1 *The Average cost under policy π with initial state x is*

$$V_a(x;\pi) = \liminf_{T\to\infty} \frac{1}{T}\,\mathbb{E}_x^\pi\left[\sum_{n=0}^{T} c(x_n,a_n)\right] \ . \tag{8.1}$$

Note that this is well defined under any of the following conditions:

- c is bounded below,
- c is bounded above,
- for each n,

$$\mathbb{E}_x^\pi\,|c(x_n,a_n)| < \infty \ .$$

Under the first two conditions, the expected immediate cost may be inifinite, but it is well defined. Under the last condition, the expected immediate cost is finite. However, this condition depends on both initial state and policy! Note that neither condition guarantees the existence of a limit: this is why we resort to limit inferior.

The average cost is used when we do not care about "trasient behavior." Shifting time by one unit amounts to starting at a different initial state. As we shall see, under some conditions, the cost is unchanged.

The average cost also measures "steady state" performance (if there is any). It is more amenable to analysis then the steady state behavior: the latter is known to produce many difficulties, since it is very sensitive. The average cost, on the other hand, involves a Cesaro sum, and is relatively insensitive to short-term phenomena.

Finally, as we shall see later, the average cost is reasonably easy to handle from a mathematical and a computational point of view—at least if costs are bounded and the chain is "indecomposeable" in the right sense. It is therefore used as an approximation to other, even more challanging cost structures (such as finite horizon cost with very long horizon).

Suppose $\pi = \{g, g, \dots\}$ is a stationary deterministic policy. Then it is easy to see that there is a Markov chain and action sequence $\{x_n^\pi,\ a_n^\pi,\ n \geq 0\}$ where $a_n^\pi = g(x_n^\pi)$. More generally, each policy π and initial state x induce a process (not necessarily Markov!) $\{x_n^\pi,\ a_n^\pi,\ n \geq 0\}$ (where we omit the initial state from the notation).

Definition 8.2 *The Ergodic cost under policy π with initial state x is*

$$V_e(x;\pi) = \liminf_{T\to\infty} \frac{1}{T}\left[\sum_{n=0}^{T} c(x_n^\pi, a_n^\pi)\right] \ . \tag{8.2}$$

This is a random variable, and we would expect that its mean is equal to the average cost. In fact, under some structural conditions, even more is true. The ergodic cost is a relatively new concept, and we postpone its investigation.

Definition 8.3 *The average MDP is to Maximize $V_a(x;\pi)$ over all policies π.*

Corollary 8.4 *For the average MDP (both Optimization and Constrained Optimization), Markov policies suffice.*

Before we embark on structural results, let us develop some feeling for the difficulties and pitfalls of the average criterion. First note that this criterion is insensitive to finite-time phenomena, in a rather strong sense.

Lemma 8.5 *Assume the costs are bounded. Fix x and π and let τ be a positive random variable taking integer values. If $\mathbb{P}_x^\pi\{\tau < \infty\} = 1$ then*

$$V_e(x;\pi) = \liminf_{T\to\infty} \frac{1}{T}\left[\sum_{n=\tau}^{T} c(x_n^\pi, a_n^\pi)\right] , \tag{8.3}$$

$$V_a(x;\pi) = \liminf_{T\to\infty} \mathbb{E}_x^\pi \frac{1}{T}\left[\sum_{n=\tau}^{T} c(x_n, a_n)\right] . \tag{8.4}$$

Proof. The ergodic case follows from the definition. For the average cost, note that the sum is interpreted as 0 if $\tau > T$. Fix an increasing function $f(T)$ so that

$$\lim_{T\to\infty} f(T) = \infty \tag{8.5}$$

$$\lim_{T\to\infty} \frac{f(T)}{T} = 0 . \tag{8.6}$$

Denote $C = \sup_{x,a} |c(x,a)|$. Now write

$$V_a(x;\pi) = \liminf_{T\to\infty} \mathbb{E}_x^\pi \frac{1}{T}\left[\sum_{n=1}^{\tau\wedge T} c(x_n, a_n) + \sum_{n=\tau\wedge T}^{T} c(x_n, a_n)\right] . \tag{8.7}$$

But

$$\mathbb{E}_x^\pi \frac{1}{T}\sum_{n=1}^{\tau\wedge T} |c(x_n, a_n)| \le \mathbb{E}_x^\pi \frac{C}{T}\left[f(T)\mathbf{1}_{\{\tau\le f(T)\}} + T\mathbf{1}_{\{\tau > f(T)\}}\right] \tag{8.8}$$

$$\le C\frac{f(T)}{T} + C\,\mathbb{P}_x^\pi\{\tau > f(T)\} . \tag{8.9}$$

Both terms go to zero as $T \to \infty$ by the assumptions on τ and on $f(T)$. This proves the result. ∎

Example 8.6 (Ross, V.1.1) *This example shows that optimal policies need not exist. The state space consists of all integers except* 0. *If* $x \geq$ *then under action* u *we go up one state, while under action* s *which switch signs to* $(-x)$, *both with probability* 1. *Once* $x < 0$ *you stay there. The cost is* 0 *at* $x > 0$ *and, for* $x < 0$, $c(x, a) = 1 + 1/x$. *Obviously,* $V_a(1; \pi) < 1$, *and* $V_a(1) = 1$ *but there is no optimal policy.*

Example 8.7 (Ross, V.1.2) *This example shows that stationary deterministic policies need not be optimal. The state space consists of all positive integers and under action* u *we go up, while under* s *we stay put. We recieve* $c(i, u) = 0$ *for going up and* $c(i, s) = 1 - 1/i$ *for staying put. Obviously, the value of the maximization problem starting at* $x(0) = 1$ *is* 1. *Now let* g *be a stationary policy which, in state* j, *chooses action* s. *Then* $V_a(1; g) \leq a - 1/j$. *However, as the following exercise shows, we can come up with an optimal policy.*

Exercise 8.8 *Let* π *be the non-stationary policy which upon entering state* i *chooses action* s i *times and then chooses action* u. *Compute* $V_a(1, \pi)$. *Let* g *be a stationary randomized policy defined by*

$$p(u \mid i) = q_i \ .$$

Can you achieve $V_a(1)$? *Hint: try* q_i *decreasing.*

In the previous example, although there is no optimal stationary policy, you can get as close as you wish to the optimal cost, using stationary policies. Hoever, even this is not guaranteed.

Example 8.9 (Ross, V.1.3) *Let $\mathcal{S}$ consist of all integers. Under action u we have $p_{ii+1}(u) = 1$ for all i not equal to 0 or (-1). Also $p_{-11}(u) = 1$. Action s is available only for $i > 0$ and*

$$p_{i-i}(s) = \alpha_i = 1 - p_{i0}(s) \ .$$

Finally, $p_{00}(u) = 1$. The immediate cost satisfies $c(i,u) = 2$ for $i < 0$ and is 0 otherwise. The constants α_n are chosen so that

$$\alpha_n < 1 \ , \quad \prod_{n=1}^{\infty} \alpha_n = \frac{3}{4} \ .$$

Suppose we start at state 1. Under a stationary policy, if we never use action s then the cost is 0. But if we ever do, then each time state 1 is visited, there is a fixed positive chance of never returning to state 1. But this implies that

$$\sum_{n=0}^{T} c(x_n, a_n) < \infty$$

almost surely, so that the Cesaro limit is zero and so is its expectation. However, suppose we choose the following non stationary policy. Initially use s. On our nth visit to state 1, choose action u n times and then choose s. In this case, the probability that the process never enters state 1 is

$$\prod_{n=1}^{\infty} \alpha_n = \frac{3}{4}$$

and, since we get an immediate reward of 2 for exactly half of the time (and 0 the rest of the time) we have $V_a(1,\pi) = 3/4$.

The fact that Markov policies suffice follows from our usual argument, since this cost functional depends only on the one-dimensional marginals. The next step should be to establish the principle of optimality.

Exercise 8.10 *What would the principle of optimality say in this case?*

Definition 8.11 *The Poisson equation associated with a stationary policy* f *is*

$$g + h(x) = c(x, f(x)) + \mathbb{E}_x^f\, h(x_1) \tag{8.10}$$

and (g, h) *solving* (8.10) *are called a solution of the Poisson equation. The optimality equation for the average cost problem is*

$$g + h(x) = \max_a \left(c(x, a) + \mathbb{E}_x^a\, h(x_1) \right) \ . \tag{8.11}$$

A pair (g, h) *solving* (8.11) *is called a solution of the optimality equation.*

Lemma 8.12 *Supopse the cost is bounded and let* (g, h) *be a solution of the Poisson equation for the policy* f*, with* h *bounded. Then* $g = V_a(x; f)$ *for all* x*.*

Proof. Just iterate, sum and then divide by the number of steps. ∎

Theorem 8.13 *Consider an average MDP with bounded costs. Let* (g, h) *be a solution of the optimality equation, with* h *bounded. Then* $g = V_a(x)$ *for all* x*.*

Proof. see Ross. ∎

Theorem 8.14 *Consider an average MDP with bounded costs. If there is a bounded solution to the optimality equation, then here exists an optimal policy which is Markov and deterministic. The value function is the unique bounded solution of the optimality equation, and any Markov policy which satisfies the optimality equation is optimal.*

Proof. See Ross. ∎

In contrast to the discounted case, the optimality equation need not find all the optimal policies. This is not very well known (but, nonetheless, true).

Example 8.15 (K.W. Ross) *Let* $\mathcal{S} = \{0,1\}$ *and* $\mathcal{A} = \{d, r\}$. *Under* d

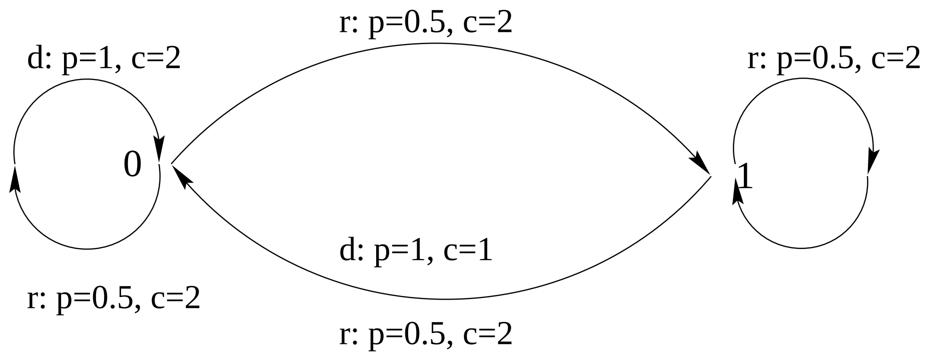

Figure 1: Optimal policies may not solve the Optimality Equation

we go down from 1 *to* 0 *with probability* 1 *(or stay in* 0*). Under action* r *we randomize: with probability* 0.5 *we switch states, and with probability* 0.5 *we stay put. The cost is* $c(1,u) = 1$ *and* $c(x,a) = 2$ *otherwise. Then the* *stationary (deterministic) policies*

$$g_1(0) = g_1(1) = d \tag{8.12}$$

$$g_2(0) = g_2(1) = r \tag{8.13}$$

are both optimal, since under g_1 *we move to state* 0 *and then always get* 2*, while under* g_2 *we always get* 2*. However, it is clear that under the discounted cost (for any discount factor), using* d *at state* 1 *is not optimal, so we expect that* g_1 *may not solve the optimality equation! Indeed, the optimality equations for this example are*

$$h(0) + 2 = \max\left[2 + h(0); 2 + 0.5h(0) + 0.5h(1)\right] \tag{8.14}$$

$$h(1) + 2 = \max\left[1 + h(0); 2 + 0.5h(0) + 0.5h(1)\right] \tag{8.15}$$

and indeed, in state 1 *the action* d *is not maximizing!*