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).

7 Discounted Cost

Definition 7.1 *The discounted cost with discount factor β, under policy π with initial state x is*

$$V_\beta(x;\pi) = \mathbb{E}_x^\pi \left[\sum_{n=0}^{\infty} \beta^n c(x_n, a_n) \right] . \tag{7.1}$$

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

- c is bounded below,
- c is bounded above,
- for some $\alpha < 1$ and C,

$$\mathbb{E}_x^\pi \left| c(x_n, a_n) \right| \le C \cdot \left(\frac{\alpha}{\beta} \right)^n .$$

Under the first two conditions, the cost may be inifinite, but it is well defined. Under the last condition, the cost is finite, and the sum converges exponentially fast. However, this condition depends on both initial state and policy! The last condition may seem impossible to verify: this is not so.

Example 7.2 *Suppose the state space is the positive integers, and that the transitions up are bounded, that is, for some K*

$$\mathbb{P}(x_{n+1} \ge i + k | x_n = i, a_n) = 0 \quad \textit{for all } k \ge K.$$

In this case $x_n \le x + n \cdot K$, and so if c grows at most polynomially fast with x, uniformly in a, that is if

$$\sup_a |c(x,a)| \le C \cdot x^d \quad \textit{for some } C \textit{ and } d,$$

then the bound holds, with any $\beta < \alpha < 1$.
This can be extended in an obvious way to more general state spaces—for example the N dimensional lattice. In fact, this is exactly the situation in most models of queues and communications networks, where the state space is the length of various queues, and where typically the number of new jobs/arrivals at any one slot is bounded.

The discounted cost is used when the future is less important than the past—a rather common condition. This particular form is chosen, in this case, for ease of computation. In addition, the discounted cost has an appealing time-homogeneity: shifting time by one unit amounts to multiplying all costs by a single factor—β. This property also gives the discounted cost its economic interpretation. For suppose I am promised an income stream of the form c_n. To calculate the present value of this future stream, we perform the following mental procedure. We take a loan so that the payments we need to make are exactly the stream c_n. But if we can return c_1 after one period of time and the interest rate is α, then we can get a loan (now) of l_1 so that

$$l_1(1+\alpha) = c_1$$

and if we are to return c_n after n periods of time, we can obtain for this a loan (now) of l_n so that

$$l_n(1+\alpha)^n = c_n$$

and the total loan is then given by

$$\sum_{n=0}^{\infty} l_n = \sum_{n=0}^{\infty} \frac{c_n}{(1+\alpha)^n} = \sum_{n=0}^{\infty} \beta^n c_n$$

where $\beta = 1/(1+\alpha)$.

The discounted cost also appears naturally as a model of learning in the context of manufacturing. It is an observed fact, for example, that prices of mass-storage devices decrease exponentially (hard-discs, for example), in terms of the cost per unit of storage. The same can be said for the price of a unit of computation speed. Although (for physical reasons) it is obvious that such changes cannot persist forever, if the horizon is long enough it is convenient to use the discounted cost model.

Finally, as we shall see later, the discounted cost is relatively easy to handle from a mathematical and a computational point of view. It is therefore used as an approximation to other, more challanging cost structures (such as the average cost).

Definition 7.3 *The Discounted MDP is to Maximize $V_\beta(x;\pi)$ over all policies π.*

Corollary 7.4 *For the Discounted MDP (both Optimization and Constrained Optimization), Markov policies suffice.*

Once we have restricted to Markov policies, we can establish the principle of optimality. We use the notation of Section 4 and Definition 4.5.

Theorem 7.5 *Let π be an optimal Markov policy for a Discounted MDP. Then, for each $k > 0$, the policy π_{k+1}^{∞} is optimal for the discounted MDP with objective $V_\beta(x;\pi)$ for each x such that $\mathbb{P}_x^{\pi_0^k}(x_{k+1} = x) > 0$.*

Proof. Left as an exercise. ∎

We now turn to a more abstract approach to the discounted MDP.

Definition 7.6 *The one-step operator $L_d : \mathcal{S} \to \mathcal{S}$ corresponding to a decision rule d, and the optimality operator $T : \mathcal{S} \to \mathcal{S}$ are defined as follows:*

$$(L_d f)(x) = c(x, d(x)) + \beta\, \mathbb{E}_x^d\, f(x_1) \tag{7.2}$$

$$(T f)(x) = \max_a \left(c(x, a) + \beta\, \mathbb{E}_x^a\, f(x_1) \right) \ . \tag{7.3}$$

The optimality equation for the Discounted problem is $TV = V$, that is

$$V(x) = \max_a \left(c(x, a) + \beta\, \mathbb{E}_x^a\, V(x_1) \right). \tag{7.4}$$

Note that under the boundedness assumption $|c(x, a)| \leq C$ we have $|V(x; \pi)| \leq C/(1-\beta)$, so that the objective function is bounded, uniformly in x and in π.

Definition 7.7 *A metric $d(x, y)$ is a real valued function that satisfies*

1. *Positivity: $d(x, y) \geq 0$*
2. *Symmetry: $d(x, y) = d(y, x)$*
3. *$d(x, y) = 0$ iff $x = y$*
4. *Triangle inequality: $d(x, y) \leq d(x, z) + d(z, y)$.*

A Metric space is a space (collection of points) with a metric. A Metric space is complete if every Cauchy sequence converges. That is, if

$$\lim_{n,m\to\infty} d(x_n, x_m) \to 0 \quad \textit{implies that } \lim_{n\to\infty} d(x_n, x_0) = 0 \textit{ for some } x_0.$$

The real line is a metric space, with $d(x, y) = |x - y|$. But it is also a metric space with the metric $d(x, y) = \infty$ for all $x \neq y$. The set of integers is a metric space with either of the metrics above.

Definition 7.8 *A function f on a metric space is a contraction if there exists a constant $K_f < 1$ so that $d(f(x), f(y)) \leq K_f d(x, y)$.*

Thus, a contraction moves points closer.

Example 7.9 *The following real functions are easily seen to be contractions: $f(x) = x/2 + b$ (any b), $f(x) = (x + \sin x)/3$.*
In N dimensions, fix some rotation matrix Θ and any matrix A. Then the function (with vectors as arguments and as values!) $f(x) = \frac{1}{2}\Theta \cdot x + A$ is a contraction. Moreover, we could add any nonlinear term, provided it reduces the size by at least 3.

Definition 7.10 *A point x is a fixed point of a function f if $f(x) = x$.*

Lemma 7.11 *If f is a contraction mapping on a complete metric space then f has a unique fixed point x_0. Moreover, for any x, the iteration $f^{(n)}(x)$ converges to x_0 geometrically fast. That is, if we define*

$$f^{(1)} = f, \quad f^{(n+1)}(x) = f^{(n)}(f(x))$$

then for each x there is a constant C_x so that

$$d(f^{(n)}(x), x_0) \leq C_x K_f^n \,.$$

Lemma 7.12 *If the cost is bounded then T and, for each d, L_d are contraction operators under the sup norm. Therefore there exists a unique bounded solution to each of the equations* (7.2) *and* (7.3)*.*

Proof. Fix d, and let f and g be bounded functions. Then

$$\sup_x |(L_d f)(x) - (L_d g)(x)| = \beta \sup_x \left| \mathbb{E}_x^{d(x)} [f(x_1) - g(x_1)] \right| \tag{7.5}$$

$$\leq \beta \sup_x \mathbb{E}_x^{d(x)} |f(x_1) - g(x_1)| \tag{7.6}$$

$$\leq \beta \sup_x |f(x) - g(x)| \ . \tag{7.7}$$

Therefore L_d is a contraction with constant $\beta < 1$. As for T,

$$\sup_x |(Tf)(x) - (Tg)(x)| = \sup_x \left| \sup_a (L_a f)(x) - \sup_a (L_a g)(x) \right| \tag{7.8}$$

$$\leq \sup_x \sup_a |(L_a f)(x) - (L_a g)(x)| \tag{7.9}$$

$$= \sup_a \left[\sup_x |(L_a f)(x) - (L_a g)(x)| \right] \tag{7.10}$$

$$\leq \sup_a \left[\beta \sup_x |f(x) - g(x)| \right] \tag{7.11}$$

$$= \beta \sup_x |f(x) - g(x)| \tag{7.12}$$

where the last inequality follows from the proof for L_d. Thus T is also a contraction with constant β. The rest follows from properties of contraction operators. ∎

Lemma 7.13 *Consider a Discounted MDP with bounded costs. Let $\pi = \{d, d, \dots\}$ be a stationary policy. Then the discounted cost $V(x; \pi)$ is the* *unique bounded solution of $L_d W = W$, that is, of*

$$W(x) = c(x, d(x)) + \beta\, \mathbb{E}_x^{d(x)}\, W(x_1). \tag{7.13}$$

Proof. By Lemma 7.12 there exists a unique bounded solution. Iterating (7.13), we obtain

$$\begin{aligned} W(x) &= c(x, d(x)) + \beta\, \mathbb{E}_x^{d(x)}\, W(x_1) && (7.14)\\ &= \mathbb{E}_x^{\pi}\left(c(x_0, a_0) + \beta c(x_1, a_1) + \beta^2 W(x_2)\right) && (7.15)\\ &= \mathbb{E}_x^{\pi}\left(\sum_{n=0}^{N} \beta^n c(x_n, a_n)\right) + \mathbb{E}_x^{\pi}\, \beta^{N+1} W(x_{N+1}) \ . && (7.16) \end{aligned}$$

But the first term in the last line converges to the cost under π, and the second is bounded by $\frac{C\beta^N}{1-\beta}$, which converges to zero as $N \to \infty$. ∎

Theorem 7.14 *Consider a Discounted MDP with bounded costs. There exists an optimal policy which is Markov stationary 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. Note that since by assumption $|c(x, a)| \le C$ we have $|V(x; \pi)| \le C/(1 - \beta)$, so that the objective function is bounded. We first show that any bounded solution of the optimality equation provides the optimal cost and an optimal decision rule. Let $V(x)$ be a solution of the optimality equation and let $d(x)$ be a (deterministic) decision rule so that

$$V(x) = \left(c(x, d(x)) + \mathbb{E}_x^d(x) V(x_1)\right) \ .$$

(The maximizer exists if $\mathcal{A}$ is finite, and otherwise under appropriate continuity conditions which we ignore here). Now fix N and consider the finite horizon problem with cost $\beta^n \cdot c$ and terminal cost $c_0(x) = \beta^N \cdot V(x)$. Using the optimality equation that V satisfies we see that $V(x)$ solves the N-step finite horizon optimality equation, so that $\pi^N = \{d, d, \dots, d\}$ is an optimal policy. Note however that, for any policy σ,

$$
\begin{aligned}
V(x;\sigma) &\leq \mathbb{E}_x^\sigma \sum_{n=0}^N \beta^n c(x_n, a_n) + \frac{C\beta^N}{1-\beta} && (7.17)\\
&\leq \mathbb{E}_x^{\pi^N} \sum_{n=0}^N \beta^n c(x_n, a_n) + \mathbb{E}_x^{\pi^N} \beta^N V(x_N) + \left(\mathbb{E}_x^{\pi^N} \left|\beta^N V(x_N)\right| + \frac{C\beta^N}{1-\beta} \right) && (7.18)\\
&\leq \mathbb{E}_x^{\pi^\infty} \sum_{n=0}^\infty \beta^n c(x_n, a_n) + \beta^N C_1 && (7.19)
\end{aligned}
$$

where, in the first inequality, we used the fact that the costs are bounded, in the second we used the optimality of π^N for the finite horizon problem, and we added a positive term. In the last inequality we collected all terms using a new constant C_1. The first term in the last line is, by the previous lemma, exactly the cost of the policy $\pi^\infty = \{d, d, \dots\}$. This shows that there exists an optimal, Markov, deterministic policy. The rest is left as an exercise. ■

We now turn to methods for comparing and computing costs, values and optimal policies. Most of the methods are developed for the optimization problem: however, the Linear Programming methods work for the Constrained Optimization problem as well. We start with the computation of the cost. Note that if the state space is finite, we can consider the function $V(\cdot;\pi)$ as a column vector, which we denote by $V(\pi)$. Similarly we denote by $c(d)$

the column vector with entries $c(x, d(x))$, and by $\mathbb{E}^{\pi} f(x_1)$ the column vector with entries $\mathbb{E}_x^{\pi} f(x_1)$.

Corollary 7.15 *Consider a problem with finite state space. Under the conditions of Lemma 7.13, the discounted cost of a stationary policy $\pi = \{d, d, \dots\}$ is the solution of the system of linear equations*

$$V(\pi) = [I - L_d]^{-1} c(d) \ . \tag{7.20}$$

Proof. Note first that

$$[I - L_d]^{-1} = \sum_{k=0}^{\infty} L_d^k \tag{7.21}$$

as is easy to verify (at least formally) by multiplying the right hand side by $I - L_d$ (both from the left and from the right). The right hand side is well defined, in the sense that we can apply it to any bounded function, and obtain a bounded function: indeed, we obtain exactly the discounted cost. If the state space is finite, this is a set of linear equations, with the dimension of the state space. ∎

In the finite case, the operator L_d can be written using a matrix and a vector, so that the equation can be stated in explicit terms: it depends on the transitions and the costs, and is linear in both.

The two most common algorithms for computing the optimal cost and policy are Policy Iteration and Value Iteration. We begin with a comparison result. Denote by d^{∞} the policy which uses the decision rule d at each stage, that is, $d^{\infty} = \{d, d, \dots\}$.

Lemma 7.16 *Assume the costs are bounded and let d_i be deterministic decision rules. If*

$$L_{d_2} V(x; d_1^\infty) \geq V(x; d_1^\infty)$$

for all x, then $V(x; d_2^\infty) \geq V(x; d_1^\infty)$ for all x.

Proof. By assumption (and watch out for the notation, especially in line 2),

$$\begin{aligned}
V(x; d_1^\infty) &\leq c(x, d_2(x)) + \beta\, \mathbb{E}_x^{d_2}\, V(x_1; d_1^\infty) && (7.22)\\
&\leq c(x, d_2(x)) + \beta\, \mathbb{E}_x^{d_2} \left[c(x_1, d_2(x_1) + \beta\, \mathbb{E}_{x_1}^{d_2}\, V(x_1; d_1^\infty)\right] && (7.23)\\
&= \mathbb{E}_x^{d_2^\infty} \left[c(x, d_2(x)) + \beta c(x_1, d_2(x_1)) + \beta^2 V(x_2, d_1^\infty)\right] && (7.24)\\
&\leq \mathbb{E}_x^{d_2^\infty} \left[\sum_{n=0}^{N-1} \beta^n c(x_n, d_2(x_n)) + \beta^N V(x_N, d_1^\infty)\right] && (7.25)\\
&= V(x, d_2^\infty) + \frac{2C\beta^N}{1-\beta} && (7.26)
\end{aligned}$$

and the result is established. ∎

Theorem 7.17 (Policy iteration) *Consider a discounted optimization problem with bounded costs and finite action and state spaces. Fix an arbitrary stationary, deterministic policy $\pi_0 = d_0^\infty$. Define*

$$d_{n+1}(x) = \arg\max_a \left\{c(x, a) + \beta\, \mathbb{E}_x^a\, V(x_1, d_n^\infty\right\}.$$

Then $V(x, d_n^\infty) \uparrow V(x)$. In addition, we have convergence in a finite number of steps, and moreover, we can choose d_n so that $d_n(x) \to d(x)$. Any such d defines an optimal policy $\pi^* = d^\infty$.

Note that there may be more than one "argmax:" although it does not matter which one we choose, we need to fix our choice if we wish for converngence.

Proof. By Lemma 7.16 and the definition of d_n, the sequence $V(x, d_n)$ is increasing for each x. Since the costs are bounded, this sequence is bounded above. Therefore, it must converge, say to some $W(x)$. Hence

$$
\begin{aligned}
W(x) &= \lim_{n\to\infty} V(x, d_n^\infty) && (7.27)\\
&= \lim_{n\to\infty} \max_a \left[c(x,a) + \beta\, \mathbb{E}_x^a\, V\left(x_1, d_{n-1}^\infty\right)\right] && (7.28)\\
&= \max_a \left[c(x,a) + \lim_{n\to\infty} \beta\, \mathbb{E}_x^a\, V\left(x_1, d_{n-1}^\infty\right)\right] && (7.29)\\
&= \max_a \left[c(x,a) + \beta\, \mathbb{E}_x^a \lim_{n\to\infty} V\left(x_1, d_{n-1}^\infty\right)\right] && (7.30)\\
&= \max_a \left[c(x,a) + \beta\, \mathbb{E}_x^a\, W(x_1)\right] && (7.31)
\end{aligned}
$$

so that W solves the optimality equation and is therefore the optimal solution. Since there is strict imrovement at every step (otherwise we have the optimal solution), and the number of policies is finite, we must have convergence in a finite number of steps. Convergence of the policies follows by the same argument, provided we have a fixed rule to choose when there is more than one "argmax." ∎

As can be seen, some of the steps work without difficulty without the finiteness assumption, but some require more care.

To compute the optimal cost using policy iteration we need to solve for $V(x, d_n^\infty)$ at each step—that is, solve a system of $|\mathcal{S}|$ linear equations. The expectation requires $|\mathcal{S}|$ sums, for each pair (x, a), for a total of roughly

$|\mathcal{S}|^2|\mathcal{A}|$ operations. In addition we need to compare $|\mathcal{A}|$ terms $|\mathcal{S}|$ times at each iteration.

The *Value iteration* algorithm is defined as follows:

- Set $V_0(x) \equiv 0$,
- Define

$$V_{n+1}(x) = \max_a \left[c(x,a) + \beta\, \mathbb{E}_x^a\, V_n(x_1) \right],$$

Exercise 7.18 *Prove (and give conditions so) that the values computed by the* value iteration *algorithm converge to the optimal value. Prove that this remains true if $V_0(x)$ is chosen as any bounded function. Show that if the state and action spaces are finite, then the optimal policy can be computed as the maximizing argument. Does convergence occur in a finite number of steps? Estimate the error in the policy after n steps, and number of steps until the correct policy is found. Hint: the algorithm computes the optimal policy for the finite horizon problem.*

7.1 Discounted cost and Linear Programs: I

There is a direct way to derive a linear program, whose solution gives the optimal cost. We restrict our attention to bounded costs.

Lemma 7.19 *Let $u(x)$ be a bounded function satisfying*

$$u(x) \geq \max_a \left(c(x,a) + \beta\, \mathbb{E}_x^a\, V(x_1) \right). \tag{7.32}$$

Then $u(x) \geq V_\beta(x)$.

Exercise 7.20 *Prove Lemma 7.19. Hint: add a new state Δ and action a_Δ so that $p_{x\Delta}(a_\Delta) = 1$ and $p_{\Delta\Delta}(a) = 1$ for all a. Set $c(x, \Delta) = u(x)$ for all $x \neq \Delta$, while $c(\Delta, \Delta) = 0$. Now interpret u as the optimal cost.*

Theorem 7.21 *Let $\alpha(x)$ be positive constants so that $\sum_{x\in\mathcal{S}} \alpha(x) = 1$. Suppose $u(x)$ solves*

$$\begin{aligned} &\textit{Minimize} \qquad && \sum_{x\in\mathcal{S}} \alpha(x)u(x) && (7.33)\\ &\textit{Subject to} \qquad && u(x) \geq c(x, a) + \beta\, \mathbb{E}^a_x V(x_1) \quad \textit{for all } x, a. && (7.34) \end{aligned}$$

Then $u \equiv V_\beta$.

Proof. Follows immediately from Lemma 7.19. ∎

Note that we could require only $\sum_{x\in\mathcal{S}} \alpha(x) < \infty$, but this is equivalent.

There is a different linear program we could use which, as it turns out, is the dual of the one in Theorem 7.21. It is, however, more intuitive and more useful. Define

$$f_\beta(x, \pi; y, a) \stackrel{def}{=} (1 - \beta) \sum_{t=0}^{\infty} \beta^t\, \mathbb{P}^\pi_x \left(x(t) = y, a(t) = a\right) \ . \tag{7.35}$$

Note that f_β satisfies the following (here we hold β, x and π fixed):

$$f_\beta(x, \pi; y, a) \geq 0 \quad \text{for all } y, a, \tag{7.36}$$

$$\sum_{y,a} f_\beta(x, \pi; y, a) = 1 \ . \tag{7.37}$$

Moreover, let π be a stationary policy $\{d, d, \dots\}$. Then

$$\sum_a f_\beta(x,\pi;y,a) = \begin{cases} \sum_{s\in\mathcal{S}} \sum_a \beta f_\beta(x,\pi;y,a) p_{yz}(a) & \text{if } y \neq x, \\ \sum_{s\in\mathcal{S}} \sum_a \beta f_\beta(x,\pi;y,a) p_{yz}(a) + (1-\beta) & \text{if } y = x. \end{cases} \tag{7.38}$$

To establish (7.38), we use the fact that, for $t \geq 1$,

$$\sum_a \mathbb{P}_x^\pi \left(x(t+1) = y, a(t+1) = a\right) = \mathbb{P}_x^\pi \left(x(t+1) = y\right) \tag{7.39}$$

$$= \sum_{s\in\mathcal{S}} \sum_a \mathbb{P}_x^\pi \left(x(t) = s, a(t) = a\right) p_{sy}(a). \tag{7.40}$$

If $y \neq x$ then $\mathbb{P}_x^\pi \left(x(0) = y, a(0) = a\right) = 0$ so that

$$\sum_a f_\beta(x,\pi;y,a) = \sum_a (1-\beta) \sum_{t=0}^{\infty} \beta^t \, \mathbb{P}_x^\pi \left(x(t) = y, a(t) = a\right) \tag{7.41}$$

$$= (1-\beta) \sum_{t=1}^{\infty} \sum_a \beta^t \, \mathbb{P}_x^\pi \left(x(t) = y, a(t) = a\right) \tag{7.42}$$

$$= (1-\beta)\beta \sum_{t=0}^{\infty} \sum_a \beta^t \, \mathbb{P}_x^\pi \left(x(t+1) = y, a(t+1) = a\right) \tag{7.43}$$

$$= (1-\beta)\beta \sum_{t=0}^{\infty} \sum_{s\in\mathcal{S}} \sum_a \beta^t \, \mathbb{P}_x^\pi \left(x(t) = s, a(t) = a\right) p_{sy}(a) \tag{7.44}$$

$$= \beta \sum_{s\in\mathcal{S}} \sum_a (1-\beta) \sum_{t=0}^{\infty} \beta^t \, \mathbb{P}_x^\pi \left(x(t) = s, a(t) = a\right) p_{sy}(a) \tag{7.45}$$

$$= \beta \sum_{s\in\mathcal{S}} \sum_a f_\beta(x,\pi;s,a) p_{sy}(a) \ . \tag{7.46}$$

If $y = x$ then, in the first equality above we have an additional term

$$(1-\beta)\sum_a \mathbb{P}_x^\pi\left(x(0) = y, a(0) = a\right) = 1 - \beta\ ,$$

and (7.38) is established. We can re-write it in the form

$$\sum_a f_\beta(x,\pi;y,a) = \sum_{s\in\mathcal{S}}\sum_a \beta f_\beta(x,\pi;y,a)p_{yz}(a) + \mathbf{1}_{\{x=y\}}(1-\beta)\ . \quad (7.47)$$

On the other hand, suppose we have a set of numbers $f(y,a)$ that satisfy (7.36), (7.37) and (7.47). Define

$$d(y;a) \stackrel{def}{=} \frac{f(y,a)}{\sum_a f(y,a)}\ . \quad (7.48)$$

Then (for $\pi = \{d, d, \ldots\}$) we have $f(y,a) = f_\beta(x,\pi;y,a)$.

Exercise 7.22 *Prove the last assertion. Generalize the argument to the case of initial distribution* μ.

So, we have established the following.

Theorem 7.23 *Consider a finite Discounted MDP. A stationary decision rule* d *is optimal if and only if it satisfies* (7.48) *for some* $f(y,a)$*, where* $f(y,a)$ *solves*

Maximize
$$\sum_a\sum_y f(y,a) \quad (7.49)$$

Subject to
$$f(y,a) \geq 0 \quad \textit{for all } y, a, \quad (7.50)$$

$$\sum_{y,a} f_\beta(x,\pi;y,a) = 1 \quad (7.51)$$

$$\sum_a f_\beta(x,\pi;y,a) - \sum_{s\in\mathcal{S}}\sum_a \beta f_\beta(x,\pi;y,a)p_{yz}(a) = \mathbf{1}_{\{x=y\}}(1-\beta) \quad (7.52)$$

Exercise 7.24 *State and prove te analogue for the Constrained Optimization problem. Assume the existence of an optimal stationary policy.*

Example 7.25 ([11, Exampe 3.1]) *Consider the following model of windowing. Information is transmitted in a window of size i. The window size changes from one time slot to the next according to some transition probabilities $\{p_{ij}\}$. The only action we can take is to decide to reset—that is, set the window size to 1. We pay $c(i)$ for using a window of size i, and we pay a penalty of R for resetting the window size. We make the following assumptions:*

$$c(i) \quad \textit{is increasing in } i \tag{7.53}$$

$$\sum_{j=k}^{\infty} p_{ij} \quad \textit{is increasing in } i\textit{, for each } k. \tag{7.54}$$

The first condition is certainly reasonable. The second is consistent with windowing protocols: these usually increase window size if transission was successful, and decrease if not, and the increase/decrease are either additive of multiplicative. Therefore, our chances of having a large window in the next step indeed increases if the present window size is increased.

Let us show that the optimal policy (minimizing the discounted cost) is a threshold policy. That is, there exists some I so that we should reset if window size is larger than I, and do nothing otherwise.

Our first step is to show that the value function $V(i)$ is increasing in i. Note that our second assumption can be written in the following way:

$$\sum_{j=1}^{\infty} p_{ij}\mathbf{1}_{\{j\geq k\}} \quad \textit{is increasing in } i\textit{, for each } k.$$

The function $\mathbf{1}_{\{j\geq k\}}$ is, for each k, an increasing function of j. It follows *that this condition is equivalent to the condition*

$$\sum_{j=1}^{\infty} p_{ij} f(j) \quad \textit{is increasing in } i\textit{, for each increasing function } f.$$

Now apply our contraction mapping result: we start with a function $V_1 = c$ and iterate using the operator T. In other words, we apply the optimality operator:

$$V_{n+1}(i) = c(i) + \min\left\{ R + \beta V_n(1), \beta \sum_j p_{ij} V_n(j) \right\}.$$

By our first assumption V_1 is increasing. Make the induction hypothesis that so is V_n. As argued above, our second assumption implies that if V_n is increasing then

$$\beta \sum_j p_{ij} V_n(j)$$

is also increasing. Therefore V_{n+1} is increasing, and hence also the limit, as $n \to \infty$, which is the value V.

Now looking at the optimality equation

$$V(i) = c(i) + \min\left\{ R + \beta V(1), \beta \sum_j p_{ij} V(j) \right\}$$

we note that our optimal action is to reset if and only if

$$R + \beta V(1) \leq \beta \sum_j p_{ij} V(j) \tag{7.55}$$

(actually, if there is equality, it does not matter which action we take). However, the term on the left does not depend on i, while the term on the right is increasing. So, if there is a finite I so that (7.55) holds, then we should reset at that I and also at every state $i > I$. If there is no such I then we set $I = \infty$.

Note that in order to obtain the structure of the optimal policy we do not need to perform any computation. Moreover, the result is generic in that it does not depend on specific parameters of the model: only on our two monotonicity assumptions. In addition, this result makes the computation of the optimal policy a much simpler matter, since we only need to find one number I. This number can be obtained analytically, by approximating V and using error bounds until we can decide whether 7.55 holds, by numerical computations or by simulation.

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,

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

Under any of these conditions, the expected immediate cost may be inifinite, but it is well defined. The 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* $x \geq 0$ *then under action* u *we go up one state, while under action* s *we 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. However, 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*

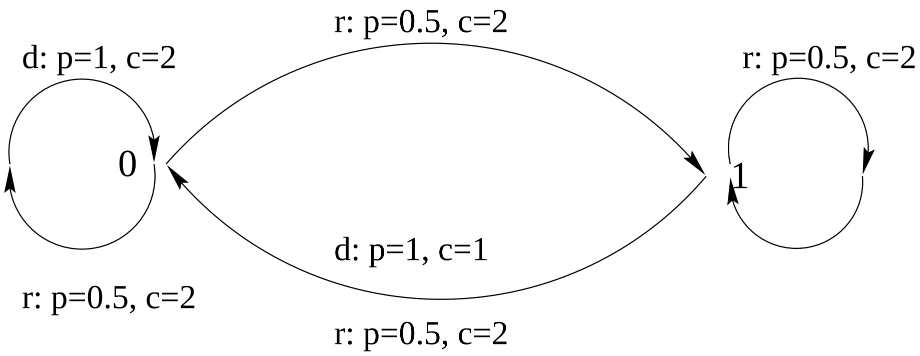

Figure 1: Optimal policies may not solve the Optimality Equation

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!