Lecture notes
Control of Stochastic Processes

Adam Shwartz, Electrical Engineering, Technion, Israel

Version of April 20, 1998

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).

2 Markov chains and Controlled Markov chains

2.1 Markov chains

Definition 2.1 *A discrete time stochastic process* $\{x_0, x_1, \ldots\}$ *is called a Markov chain if, for all* n*, all sets* A *and points* $b_0, b_1, \ldots,$

$$\mathbb{P}\{x_{n+1} \in A \mid x_i = b_i,\ i \le n\} = \mathbb{P}\{x_{n+1} \in A \mid x_n = b_n\} \ . \tag{2.1}$$

The state *of the Markov chain at time* n *is the value of* x_n*. The* state space $\mathcal{S}$ *of the chain is the space of all possible states.*

Fine points: for non countable state spaces, need to define the conditional probability in the right way.

We shall deal with *discrete state spaces*, e.g. integers. In this case we shall denote

$$p_{ij} \stackrel{def}{=} \mathbb{P}\{x_{n+1} = j \mid x_n = i\} \ . \tag{2.2}$$

Definition 2.2 *If the rhs of* (2.2) *does not depend on* n*, we call the Markov chain* homogeneous*, and the notation of* (2.2) *is valid. We denote* $P \stackrel{def}{=} \{p_{ij}\}$ *and call it the transition matrix. We define the row-vector*

$$\mu_n(j) \stackrel{def}{=} \mathbb{P}\{x_n = j\} \ . \tag{2.3}$$

Obviosly, μ_n depends on the distribution of x_0.

Homogeneous MC are sometimes called stationary: a poor terminology.

Example 2.3 *The sequence of coin tosses (get $+1$ for heads, -1 for tails) is a Markov chain. Write the distribution of x_1 as a function of x_0, and as a function of μ_0. What is the structure of the transition matrix?*

Theorem 2.4 *Given an initial state x and a transition matrix P, there exists a probability meansre $\mathbb{P}_x$ and a Markov proceses $x_0, x_1, \ldots$ so that*

$$\mathbb{P}_x\{x_0 = x\} = 1, \tag{2.4}$$

$$\mathbb{P}_x\{x_n = y\} = \mu_n(y), \tag{2.5}$$

$$\mu_{n+1} = \mu_n P. \tag{2.6}$$

Exercise 2.5 *Let $\xi_0, \xi_1, \ldots$ be a sequence of i.i.d. random variables. Consider the general recursion*

$$x_{n+1} = f\left(x_n, \xi_n\right) \ , \quad x_0 = x. \tag{2.7}$$

Show that $x_0, x_1, \ldots$ is a Markov process. Under what conditions is the state space equal $\mathbb{N}$? In this case write an expression for p_{ij}.

Exercise 2.6 *This complements the previous exercise. Let $y_0, y_1, \ldots$ be a homogeneous Markov process with state space $\mathbb{N}$ and transition probabilities p_{ij}. Let $\xi_0, \xi_1, \ldots$ be a sequence of Uniform [0,1], i.i.d. random variables. Show that it is possible to construct a function f so that the Markov chain* *given by*

$$x_{n+1} = f\left(x_n, \xi_n\right) \ , \quad x_0 = y_0, \tag{2.8}$$

has the same distribution as the process y.

2.2 Controlled Markov chains

The latest text on the subject is [9].

A controlled Markov chain differs from a Markov chain by the inclusion of a "control" or "action" variable a. This variable may change the transition probabilities. The process is now defined (recursively as before) through the following.

Definition 2.7 *A* Controlled Markov chain *is a process where, given $a_n = a$, we have*

$$p_{ij}(a) \stackrel{def}{=} \mathbb{P}\{x_{n+1} = j \mid x_n = j,\ a_n = a\} \tag{2.9}$$

$$= \mathbb{P}\{x_{n+1} \in A \mid x_n = j,\ a_n = a,\ x_i = b_i,\ a_i = \alpha_i,\ i < n\} \ . \tag{2.10}$$

Example 2.8 *Suppose I have two coins in my pocket: a normal (fair) coin, and a coin that gives probability $3/4$ to heads. I throw the fair coin. If I get tails, I throw the same coin 2 more times. If I get heads, I throw the biased coin two more times. Is the sequence of coin tosees (get $+1$ for heads, -1 for tails) a Markov chain? Is it a controlled chain?*

Definition 2.9 *Histories, actions, decision rules and policies. The existence of a Probability $\mathbb{P}_x^\pi$ and of a controlled stochastic process. Denote by Π the collection of all admissible policies.*

Definition 2.10 *The notation* p_{ij}, $p_{ij}(a)$ *and* $p_{ij}(g)$ *when* g *is a deterministic or randomized decision rule.*

Exercise 2.11 *Suppose* $p_{ij}(g;n)$ *depends explicitly on time. Embed this chain in a new homogeneous Controlled chain. Give the new state and action spaces as well as transitions explicitely.*

Exercise 2.12 *Extend exercises 2.5–2.6 to the controlled case, where*

$$x_{n+1} = f\left(x_n, a_n, \xi_n\right) \ . \tag{2.11}$$

Suppose we fix an initial state x and a policy π. We then have a probability $\mathbb{P}_x^\pi$ and expectation (operator) $\mathbb{E}_x^\pi$.

Example 2.13 *Let* μ *be a row vector and think of a function* f *as a column vector (its easy of* f *is defined for a finite number of arguments). Give an interpretation of the following objects and write them in terms of* $\mathbb{P}_x^a$, $\mathbb{E}_x^a$ *and* x_1*:* $P(a)$, $\mu P(a)$, μf, $P(a) f$, $\mu P(a) f$. *Suppose we choose a feedback control, so that when in state* x*, we choose action* $g(x)$*. What is the form of the transition matrix? Suppose we use* g_0 *at time* 0 *and* g_1 *at time* 1*, so that* $\pi = \{g_0, g_1\}$*. Express* $\mathbb{E}_x^\pi f(x_2)$ *in terms of transition matrices.*

The theorem below is stated and proved for the case where $\mathcal{S}$ and $\mathcal{A}$ are both finite (just a bit more care would extend to the countable case). It is, however, far more general. See [7, Theorem 13.2].

Theorem 2.14 *Let π be any admissible policy and fix an initial state x. Define a Markov (non stationary!) policy g through its decision rules*

$$g_n(i;a) \stackrel{def}{=} \frac{\mathbb{P}_x^\pi\left(x_n = i, a_n = a\right)}{\mathbb{P}_x^\pi\left(x_n = i\right)} \ . \tag{2.12}$$

Then

$$\mathbb{P}_x^\pi\left(x_n = i, a_n = a\right) = \mathbb{P}_x^g\left(x_n = i, a_n = a\right) \qquad \textit{for all } n, i, a. \tag{2.13}$$

Proof. By induction. For $n = 0$, we only need to check for $i = x$, and then

$$\begin{aligned}
\mathbb{P}_x^\pi\left(x_0 = x, a_0 = a\right) &= \frac{\mathbb{P}_x^\pi\left(x_0 = x, a_0 = a\right)}{\mathbb{P}_x^\pi\left(x_0 = x\right)} && (2.14)\\
&\stackrel{def}{=} g_0(x;a) && (2.15)\\
&= \mathbb{P}_x^g\left(a_0 = a\right) && (2.16)\\
&= \frac{\mathbb{P}_x^g\left(x_0 = x, a_0 = a\right)}{\mathbb{P}_x^g\left(x_0 = x\right)} && (2.17)\\
&= \mathbb{P}_x^g\left(x_0 = x, a_0 = a\right) \ . && (2.18)
\end{aligned}$$

Now suppose the theorem is true up to n. Then

$$\begin{aligned}
\mathbb{P}_x^g\left(x_{n+1} = j\right) &= \sum_i \mathbb{P}_x^g\left(x_{n+1} = j, x_n = i\right) && (2.19)\\
&= \sum_i \mathbb{P}_x^g\left(x_{n+1} = j \mid x_n = i\right) \mathbb{P}_x^g\left(x_n = i\right) && (2.20)\\
&= \sum_i \sum_a p_{ij}(a) g_n(i;a)\, \mathbb{P}_x^g\left(x_n = i\right) && (2.21)\\
&= \sum_i \sum_a p_{ij}(a) \frac{\mathbb{P}_x^\pi\left(x_n = i, a_n = a\right)}{\mathbb{P}_x^\pi\left(x_n = i\right)} \mathbb{P}_x^g\left(x_n = i\right) && (2.22)\\
&= \sum_i \sum_a p_{ij}(a)\, \mathbb{P}_x^\pi\left(x_n = i, a_n = a\right) && (2.23)\\
&= \mathbb{P}_x^\pi\left(x_{n+1} = j\right) \ . && (2.24)
\end{aligned}$$

The rest is by definition of g and of a controlled Markov process. $\blacksquare$

3 Markov decision processes

Definition 3.1 *General cost function, immediate cost function. Finite cost, discounted cost and its interpretation. Average cost: definition.*

Definition 3.2 *The formal standard definition of MDP. A Markov decision process is a 5-tuple* $\{\mathcal{S}, \mathcal{A}, P, c, V\}$ *where*

- $\mathcal{S}$ *is the state space,*
- $\mathcal{A}$ *is the action space,*
- $P(a) = \{p_{ij}(a)\}$ *is the transition matrix, when action* a *is taken,*
- $c = c(x, a)$ *is the immediate cost when at state* x *and using action* a*, and*
- $V = V(x; \pi)$ *is the value of the criterion when starting at* x *and using policy* π*.*

More complicated models allow for the avaliable actions to depend on the state as $\mathcal{A}(x)$. Allowing explicit time dependence is conceptually trivial although could be practically annoying. Slight complications arise when we wish to include events in our model, since often there are events (or pairs of events) that lead to no change in the state.

The standard Markov decision problem is to maximize (or minimize) some V: different problems may differ in the structure of V. The standard structures of V are defined and investigated in subsequent chapters. They are all the expectation of a weighed sum of immediate cost functions, that is, for some sequence $\alpha_n, \ n \geq 0$ they have the form

$$V(x;\pi) = \mathbb{E}_x^\pi \sum_{n=0}^{\infty} \alpha_n c(x_n, a_n) \tag{3.1}$$

or a similar expression (with the usual stipulation that the expression is well defined!). The less standard Markov decision problems are, for example, constrained (through several cost functionals), multicriteria etc.

Corollary 3.3 *If V is a weighted sum of immediate cost functions, then Markov policies suffice.*

Note that, in the most abstract form, we can view our variable as a policy (the argument of V) or, more abstractly, as a probability measure in the space of measures which can be generated by policies.

We now turn to modelling.

Exercise 3.4 *Consider a buffer where, at each unit of time, one of the following happens: a new request for service arrives w.p. p_r, and if the buffer is not empty, a service is completed w.p. p_s where $p_r + p_s < 1$. Is there a state such that the processes is a Markov chain? Now suppose the available action is to accept an arrival or not. Give the parameters of the CMC. Write an immediate cost function that gives a fixed reward for accepting a request, and a fixed penalty for each request in the buffer for each unit of time.*

Exercise 3.5 *Repeat Exercise 3.4 in the case that, if an arrival occurs, it carries requests according to an integer valued random variable (with finite* *number of values): this is called batch arrivals. Now allow the control action to accept any subset of the arrival batch. Now allow batch departures.*

Exercise 3.6 *Let $d_n; n \geq 0$ be a sequence of iid, positive, integer valued random variables representing a demand process. Suppose the available action is to order any (positive) number into stock; but the order arrives at the next time unit. Write the CMC in the case that unfulfilled demand is lost, and in the case that it is "backlogged." Write an immediate cost which is negative and affine in the order size, gives a fixed reward for each fulfilled unit of demand, and a penalty for a parially completed request.*

Exercise 3.7 *Let f be an arbitrary function and let consider a CMC x_n. Fix $\lambda > 0$ and suppose we want to maximize*

$$\mathbb{P}_x^\pi \left(\max_{0 \leq n \leq N} f(x_n \geq \lambda \right). \tag{3.2}$$

Write this in a standard form. Hint: define a variable y_n which keeps the last largest value of $f(x_n)$.

Let $V^i(x; \pi),\ i = 1, 2, \ldots, K$ be some cost functions. The two main problems we shall consider below are

$$\begin{aligned} &\text{Optimization:} \\ &\text{Maximize} && V^1(x;\pi) && (3.3) \\ &\text{Constrained Optimization:} \\ &\text{Maximize} && V^1(x;\pi) && (3.4) \\ &\text{Subject to} && V^i(x;\pi) \geq v_i,\ i = 2, \ldots, K. && (3.5) \end{aligned}$$

and we shall denote the value of the problem by $V^1(x)$: that is, this is the maximal value obtained in (3.3), or the maximal value obtained in (3.4) under the constraints (3.5). A policy π^* satisfying the constraints in (3.5) is called feasible at x. A (feasible) policy π^* satisfying $V^1(x;\pi) = V^1(x)$ will be called optimal at x. If it is optimal for all x, it will be called optimal.

Note that, in general, this value depends on the initial state x:

Example 3.8 *Let $\mathcal{S} = \{0,1\}$ with a single action a. Let $p_{10}(a) = p_{00}(a) = 1$. Let $c(1,a) = 1$ and $c(0,a) = 0$. Consider the ocnstrained optimization problem where the immediate cost associated with the constraint satisfies $c_2(1,a) = 1$, $c_2(0,a) = 0$. Then for $v_2 = \alpha_0/2$ the optimization problem is not even feasible starting at state 1.*

4 The finite horizon MDP.

The finite-horizon problem is defined by setting

$$V_{fh}(x;\pi) = \mathbb{E}_x^\pi \left(\left[\sum_{n=0}^{N-1} c(x_n, a_n) \right] + c_0(x_N) \right) . \tag{4.1}$$

Definition 4.1 *The finite horizon MDP is to Maximize $V_{fh}(x;\pi)$ over all policies π.*

Example 4.2 *Cheapest path on a graph.*

Example 4.3 *See [10, I.2]: note that n counts steps to go.*

Corollary 4.4 *For the finite horizon MDP (optimization and constrained optimization), Markov policies suffice.*

Once we have restricted to Markov policies, we can establish the principle of optimality.

Definition 4.5 *For a policy π we use the notation*

$$\pi = \{\pi_0^k, \pi_{k+1}^N\} \quad \textit{where} \quad \pi_0^k = \{\pi_0, \dots, \pi_k\} \quad \textit{and} \quad \pi_{k+1}^n = \{\pi_{k+1}, \dots, \pi_N\} . \tag{4.2}$$

We also recall that

$$\mathbb{P}_x(\cdot) = \mathbb{P}(\cdot \mid x_0 = x) \tag{4.3}$$

so that, due to assumed homogeneity we can write

$$\mathbb{P}\left(x_N = y \mid x_k = x\right) = \mathbb{P}_x\left(x_{N-k} = y\right) \ . \tag{4.4}$$

Note that the two sides have a different meaning, but the values agree.

Theorem 4.6 *Let π be an optimal Markov policy for a finite horizon MDP. Then, for each $k < N$, the policy π_{k+1}^N is optimal for the finite horizon MDP with objective*

$$V_{fh;k}(x;\pi) = \mathbb{E}_x^{\pi}\left(\left[\sum_{n=0}^{N-k-2} c(x_n, a_n)\right] + c_0(x_{N-k-1})\right) \tag{4.5}$$

for each x such that $\mathbb{P}_x^{\pi_0^k}\left(x_{k+1} = x\right) > 0$. Conversly, suppose π_{k+1}^N is optimal for the $N-k-1$-step finite horizon MDP with the above objective, for all initial states x. Then given any optimal policy σ, the policy $\{\sigma_0^k, \pi_{k+1}^N$ is optimal.

Note: we use the convention that $\sum_{i=0}^{-1} a_i = 0$. Note also that in order to calculate the distribution of x_{k+1}, we only need to know x and π_0^k.

Proof. By contradiction. Let σ be an optimal Markov policy for $V_{fh;k}$. Consider the policy

$$\tilde{\pi} = \{\pi_0^k, \sigma\} \tag{4.6}$$

and note that this is a Markov policy. Write

$$V_{fh}(x;\pi) = \mathbb{E}_x^{\pi}\left[\sum_{n=0}^{k} c(x_n, a_n)\right] + \mathbb{E}_x^{\pi}\left(\left[\sum_{n=k+1}^{N-1} c(x_n, a_n)\right] + c_0(x_N)\,\middle|\, x_{k+1}\right) \ . \tag{4.7}$$

Since by assumption σ is optimal and π_{k+1}^N is not, we have for all x, (in obvious notation)

$$\left(\mathbb{E}_x^{\sigma} - \mathbb{E}_x^{\pi_{k+1}^N}\right)\left(\left[\sum_{n=0}^{N-k-2} c(x_n, a_n)\right] + c_0(x_{N-k-1})\right) > 0 \tag{4.8}$$

where we have shifted the indices as in definition 4.5. But this is exactly the second term in our evaluation of V_{fh} above. Therefor, $\tilde{\pi}$ performs strictly better than π, a contradiction. The proof of the converse is left as an exercise.

■

Definition 4.7 *The optimality equation for the finite horizon problem is*

$$V_n(x) = \max_a \left(c(x, a) + \mathbb{E}_x^a V_{n-1}(x_1)\right) \;\; ; \quad V_0(x) = c_0(x) \; . \tag{4.9}$$

Theorem 4.8 *Consider a finite MDP with finite horizon. Then there exists an optimal policy which is Markov and deterministic. The value function can be computed by backward induction using the optimality equation, and any Markov policy which satisfies the optimality equation is optimal.*

Proof. Note that the index n in the optimality equation counts steps to go, and not running time index.

By the principle of optimality, we should do the last step in an optimal way; this is the case $n = 1$. This shows that we can choose a Markov decision rule at time $N - 1$, so that the action depends only on the state x at that time, and is not randomized. moreover, this decision rule is optimal for all x. Now apply the principle of optimality twice: the optimal policy must do the last two steps in an optimal way, and our just-computed last step can be used as the last part of the two-step problem. Thus, again from the optimality equation, the last two steps use a non-randomized Markov policy. Continue $N - 1$ times. ∎

end of Lecture 3, March 23, 1998

5 Linear Programming, Occupation measures

A linear program is an optimization problem of the following type.

Definition 5.1 *A Linear Program in the variables z is specified through a vector of weights c, a matrix A specifying inequality constraints, and a matrix C specifying equality constraints:*

$$\textit{Maximize} \qquad c \cdot z \tag{5.1}$$

$$\textit{Subject to} \qquad A \cdot z + b \geq 0 \tag{5.2}$$

$$\textit{and} \qquad C \cdot z + d = 0. \tag{5.3}$$

This is a very general structure, but there are efficient methods to solve such problems (polynomial-although it seems that the non-polynomial method works better ...). Many MDP's can be solved through LP methods. Note in particular that the finite-horizon cost is linear in the space of measures, since the cost can be written as

$$
\begin{aligned}
&V_{fh}(x;\pi) &(5.4)\\
&\quad = \mathbb{E}_x^\pi \left(\left[\sum_{n=0}^{N-1} c(x_n, a_n) \right] + c_0(x_N) \right) &(5.5)\\
&\quad = \sum_{n=0}^{N-1} \mathbb{E}_x^\pi \, c(x_n, a_n) + \mathbb{E}_x^\pi \, c_0(x_N) &(5.6)\\
&\quad = \left(\sum_{n=0}^{N-1} \sum_{y,a} \mathbb{P}_x^\pi \left(x_n = y, a_n = a\right) c(y,a) \right) + \sum_{y,a} \mathbb{P}_x^\pi \left(x_N = y, a_N = a\right) c_0(y,a) &(5.7)\\
&\quad = \sum_{y,a} \left(\sum_{n=0}^{N-1} \mathbb{P}_x^\pi \left(x_n = y, a_n = a\right) c(y,a) + \mathbb{P}_x^\pi \left(x_N = y, a_N = a\right) c_0(y,a) \right) &(5.8)
\end{aligned}
$$

which is linear in $\mathbb{P}_x^\pi$. However, we need to choose our variables so that the constraints are linear, and so that we can recover π from the variables.

Example 5.2 *Write the optimality equation as a linear program. Hint: the variable is the distribution of the actions. This give the optimal cost (and policy) through a sequence of linear programs. Now write all of them using a single problem (hint: the programs are related through constraints). What is the size of this program?*

There is a more natural way to write a linear program for the N-step problem, which is related to LP's for other criteria. To define the LP, it is convenient to introduce "occupation measures."

Definition 5.3 *The* occupation measure Q_{fh} *for the finite horizon process is defined by*

$$Q_{fh}(x_0, \pi; n, x, a) = \mathbb{P}^{\pi}_{x_0}(x_n = x, a_n = a). \tag{5.9}$$

Note that, for each n, Q_{fh} is a probability measure. We can write a linear program, extending 5.2, as follows.

Definition 5.4 *Linear program for a finite MDP (in the variables $z = z(n, x, a)$).*

$$\textit{Maximize} \quad \sum_{y, a \in \mathcal{A}(y)} \sum_{n=0}^{N-1} c(y, a) z(n, y, a) + c_0(y) z(N, y, a) \tag{5.10}$$

$$\textit{Subject to} \quad \sum_{a \in \mathcal{A}(y)} z(0, y, a) = \mathbf{1}_{\{y=x\}} \tag{5.11}$$

$$\textit{and} \sum_{a \in \mathcal{A}(y)} z(n, y, a) - \sum_{u \in \mathcal{S}} \sum_{a \in \mathcal{A}(u)} p_{uy}(a) z(n-1, u, a) = 0 \tag{5.12}$$

$$\textit{and} \quad z(n, y, a) \geq 0. \tag{5.13}$$

The interpretation of the variables is that $z(n, y, a) = Q_{fh}(\pi, x; n, y, a)$. Therefore, if π is a Markov policy,

$$\frac{z(n, y, a)}{\sum_a z(n, y, a)} \tag{5.14}$$

is exactly the decision rule at time n, since by definition,

$$\mathbb{P}^{\pi}_{x}(x_n = y, a_n = a) = \mathbb{P}^{\pi}_{x}(x_n = y) \cdot \pi_n(a \mid y). \tag{5.15}$$

Note that the decision rule is a non-linear function of the variables of the LP! The maximization is exactly of the cost. The first constraint is that the initial condition is x. The second is that the occupation measure is consistent with the transition probabilities and with the chosen actions, and the last is obvious.

Exercise 5.5 *Discuss the computational complexity of solving the finite horizon MDP via the two algorithms: the optimality equation (backward induction), and linear programming.*

6 Super Modularity and optimization

It is often impossible or impractical to solve an MDP via any of the computational methods we have seen. This is certainly the case if the state (or action) spaces are infinite. However, it may still be possible to obtain structural results. These are important for two reasons. First, they provide some information about the control, and in this way help design "good" controls. But more importantly, often the structural results show that we need only search within a fairly small class of candidates for optimality, thus making the computation feasible.

One type of structural results relies on "supermodularity." We illustrate via an example. The two key points to keep in mind are: first, the use of supermodularity. Second, the way the result is obtained, namely by identifying a set of properties that are satisfied by the final cost, and propagate through the backward induction.

Example 6.1 *[10, I.4]*

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 $\alpha < 1$,

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

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

Corollary 7.3 *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 7.4 *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. ∎

Definition 7.5 *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:*

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

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

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

Lemma 7.6 *If the cost is bounded then T and, for each d, L_d are contraction operators under the sup norm. Therefor 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

$$\sup_x |(L_d f)(x) - (L_d g)(x)| = \beta \sup_x \left| \mathbb{E}_x^{d(x)} [f(x_1) - g(x_1)] \right| \tag{7.5}$$
$$\leq \beta \sup_x \mathbb{E}_x^{d(x)} |f(x_1) - g(x_1)| \tag{7.6}$$
$$\leq \beta \sup_x |f(x) - g(x)| \ . \tag{7.7}$$

Therefor 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 7.7 *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 7.6 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 7.8 *Consider a Discounted MDP with bounded costs. There 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. 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 $c_0(x) = \beta^N \cdot V(x)$. We then see that $V(x)$ solves the finite horizon optimality equation, so that $\pi^N = \{d, d, \dots, d\}$ is an optimal policy. Note however that, for any policy σ,

$$V(x;\sigma) \le \mathbb{E}_x^\sigma \sum_{n=0}^{N} \beta^n c(x_n, a_n) + \frac{C\beta^N}{1-\beta} \tag{7.17}$$

$$\le \mathbb{E}_x^{\pi^N} \sum_{n=0}^{N} \beta^n c(x_n, a_n) + \mathbb{E}_x^{\pi^N} \beta^N V(x_N) + \left(\mathbb{E}_x^{\pi^N} \left|\beta^N V(x_N)\right| + \frac{C\beta^N}{1-\beta} \right) \tag{7.18}$$

$$\le \mathbb{E}_x^{\pi^\infty} \sum_{n=0}^{\infty} \beta^n c(x_n, a_n) + \beta^N C_1 \tag{7.19}$$

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 7.9 *Consider a problem with finite state space. Under the conditions of Lemma 7.7, 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 as a matrix, 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 7.10 *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),

$$\begin{aligned}
V(x; d_1^\infty) &\leq c(x, d_2(x)) + \beta\, \mathbb{E}_x^{d_2} V(x_1; d_1^\infty) && (7.22)\\
&\leq c(x, d_2(x)) + \beta\, \mathbb{E}_x^{d_2} \left[c(x_1, d_2(x_1) + \beta\, \mathbb{E}_{x_1}^{d_2} V(x_1; d_1^\infty)\right] && (7.23)\\
& && (7.24)\\
&= \mathbb{E}_x^{d_2^\infty} \left[c(x, d_2(x)) + \beta c(x_1, d_2(x_1)) + \beta^2 V(x_2, d_1^\infty)\right] && (7.25)\\
&\leq \mathbb{E}_x^{d_2^\infty} \left[\sum_{n=0}^{N-1} \beta^n c(x_n, d_2(x_n)) + \beta^N V(x_N, d_1^\infty)\right] && (7.26)\\
&= V(x, d_2^\infty) + \frac{2C\beta^N}{1-\beta} && (7.27)
\end{aligned}$$

and the result is established. ■

Theorem 7.11 (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 the lemma 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. Therefor, it must converge, say to some $W(x)$. Hence

$$W(x) = \lim_{n\to\infty} V(x, d_n^\infty) \tag{7.28}$$

$$= \lim_{n\to\infty} \max_a \left[c(x,a) + \beta\, \mathbb{E}_x^a\, V\left(x_1, d_{n-1}^\infty\right)\right] \tag{7.29}$$

$$= \max_a \left[c(x,a) + \lim_{n\to\infty} \beta\, \mathbb{E}_x^a\, V\left(x_1, d_{n-1}^\infty\right)\right] \tag{7.30}$$

$$= \max_a \left[c(x,a) + \beta\, \mathbb{E}_x^a \lim_{n\to\infty} V\left(x_1, d_{n-1}^\infty\right)\right] \tag{7.31}$$

$$= \max_a \left[c(x,a) + \beta\, \mathbb{E}_x^a\, W(x_1)\right] \tag{7.32}$$

so that W solves the optimality equation and is therefore the optimal solution. Since there 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

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

Exercise 7.12 *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 7.13 *Let $u(x)$ be a bounded function satisfying*

$$u(x) \geq \max_a \left(c(x,a) + \beta \, \mathbb{E}^a_x \, V(x_1) \right). \tag{7.33}$$

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

Exercise 7.14 *Prove Lemma 7.13. 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 7.15 *Let* $\alpha(x)$ *be positive constants so that* $\sum_{x\in\mathcal{S}} \alpha(x) = 1$*. Suppose* $u(x)$ *solves*

$$\text{Minimize} \qquad \sum_{x\in\mathcal{S}} \alpha(x)u(x) \tag{7.34}$$

$$\text{Subject to} \qquad u(x) \geq (c(x,a) + \beta\,\mathbb{E}_x^a\, V(x_1) \quad \text{for all } x, a. \tag{7.35}$$

Then $u \equiv V_\beta$*.*

Proof. Follows immediately from Lemma 7.13. ∎

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 7.15. It is, however, more intuitive and more useful. Define

$$f_\beta(x, \pi; y, a) \stackrel{def}{=} (1-\beta) \sum_{t=0}^{\infty} \mathbb{P}_x^\pi\left(x(t) = y, a(t) = a\right) \ . \tag{7.36}$$

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

$$f_\beta(x, \pi; y, a) \geq 0 \quad \text{for all } y, a, \tag{7.37}$$

$$\sum_{y,a} f_\beta(x, \pi; y, a) = 1 \ . \tag{7.38}$$

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

$$\sum_a f_\beta(x,\pi;y,a) = \begin{cases} \sum_{s\in\mathcal{S}} \sum_a \beta f_\beta(x,\pi;y,a) p_{yz}(a) & \text{if } y \neq x, \\ \sum_{s\in\mathcal{S}} \sum_a \beta f_\beta(x,\pi;y,a) p_{yz}(a) + (1-\beta) & \text{if } y = x. \end{cases} \tag{7.39}$$

To establish (7.39), we use the fact that, for $t \geq 1$,

$$\sum_a \mathbb{P}_x^\pi\left(x(t+1) = y, a(t+1) = a\right) = \mathbb{P}_x^\pi\left(x(t+1) = y\right) \tag{7.40}$$

$$= \sum_{s\in\mathcal{S}} \sum_a \mathbb{P}_x^\pi\left(x(t) = s, a(t) = a\right) p_{sy}(a). \tag{7.41}$$

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

$$\sum_a f_\beta(x,\pi;y,a) = \sum_a (1-\beta) \sum_{t=0}^{\infty} \mathbb{P}_x^\pi\left(x(t) = y, a(t) = a\right) \tag{7.42}$$

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

$$= (1-\beta)\beta \sum_{t=0}^{\infty} \sum_a \mathbb{P}_x^\pi\left(x(t+1) = y, a(t+1) = a\right) \tag{7.44}$$

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

$$= \beta \sum_{s\in\mathcal{S}} \sum_a (1-\beta) \sum_{t=0}^{\infty} \mathbb{P}_x^\pi\left(x(t) = s, a(t) = a\right) p_{sy}(a) \tag{7.46}$$

$$= \beta \sum_{s\in\mathcal{S}} \sum_a f_\beta(x,\pi;s,a)\ . \tag{7.47}$$

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

$$\sum_a f_\beta(x,\pi;y,a) = \sum_{s\in\mathcal{S}}\sum_a \beta f_\beta(x,\pi;y,a)p_{yz}(a) + \mathbf{1}_{\{x=y\}}(1-\beta)\ . \tag{7.48}$$

On the other hand, suppose we a set of numbers $f(y,a)$ that satisfy (7.37), (7.38) and (7.48). Define

$$d(y;a) \stackrel{def}{=} \frac{f(y,a)}{\sum_a f(y,a)}\ . \tag{7.49}$$

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

Exercise 7.16 *Prove the last assertion. Generalize the argument to the case of initial distribution μ.*

So, we have established the following.

Theorem 7.17 *Consider a finite Discounted MDP. A stationary decision rule d is optimal if and only if it satisfies* (7.49) *for some $f(y,a)$, where $f(y,a)$ solves*

$$\text{Maximize} \qquad \sum_a\sum_y f(y,a) \tag{7.50}$$

$$\text{Subject to} \qquad f(y,a)\geq 0 \quad \text{for all } y,a, \tag{7.51}$$

$$\sum_{y,a} f_\beta(x,\pi;y,a) = 1 \tag{7.52}$$

$$\sum_a f_\beta(x,\pi;y,a) - \sum_{s\in\mathcal{S}}\sum_a \beta f_\beta(x,\pi;y,a)p_{yz}(a) = \mathbf{1}_{\{x=y\}}(1-\beta) \tag{7.53}$$

Exercise 7.18 *State and prove te analogue for the Constrained Optimization problem. Assume the existence of an optimal stationary policy.*

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 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 $\mathbb{P}^\pi_x\{\tau < \infty\} = 1$ then*

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

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

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

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

But

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

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

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*

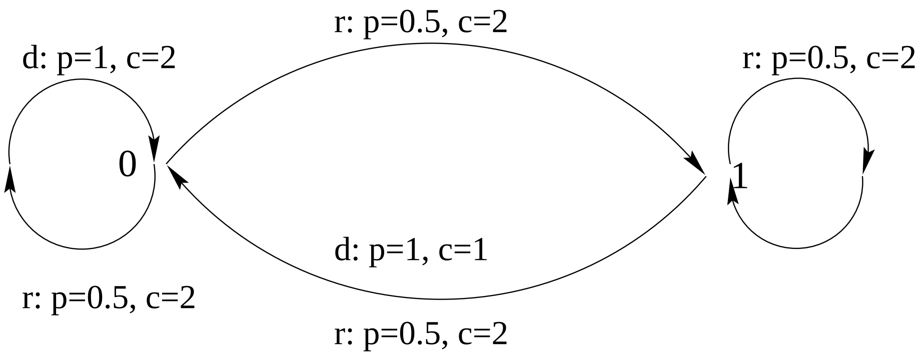

Figure 1: Optimal policies may not solve the Optimality Equation

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!*

References

[1] E. Altman *Constrained Markov decision processes* To appear, 1998.

New book, advanced.

[2] D. P. Bertsekas *Dynamic programming and stochastic control* Academic Press 1976

Engineering approach, basics.

[3] D. P. Bertsekas and S. E. Shreve *Stochastic Optimal Control* Academic Press New York 1978

Very mathematical.

[4] Cyrus Derman *Finite state Markovian decision processes* Academic Press New York 1970

Standard reference for the finite case.

[5] D. P. Heyman and M. J. Sobel *Stochastic Methods in Operations Research I: Stochastic Processes and Operating Characteristics* McGraw-Hill New York 1982

[6] D. P. Heyman and M. J. Sobel *Stochastic Methods in Operations Research II: Stochastic Optimization* McGraw-Hill New York 1984

A standard reference; covers much more.

[7] A. Hordijk Dynamic programming and Markov potential theory, Mathematical Centre Tracts 51, Mathematisch Centrum, Amsterdam, 1977.

Deep, theoretical classic, hard to read and somewhat old style.

[8] H.J. Kushner Introduction to stochastic control, Hold Reinhart Winston 1971.

Sample path approach, quite different from the rest and worthwhile. Watch out for the many typoes!

[9] M. L. Puterman *Markov decision processes* Elsevier Science Publishers 1990

Very detailed, very strong on computational techniques.

[10] S. M. Ross *Introduction to Stochastic Dynamic Programming* Academic Press 1984

Elementary, very good intuition, but be careful with proofs!

[11] H. C. Tijms *Stochastic Modelling and Analysis: a computational approach* John Wiley New York 1986

Good introduction to the modelling aspect.

[12] P. Whittle *Optimization over time; dynamic programming and stochastic control* Wiley 1983

Comprehensive, good intuition, sometimes sloppy proofs.