

2 Finite Horizon Markov Decision Problems

In this chapter we solve finite horizon Markov decision problems. We are describing a policy evaluation algorithm and the Bellman equations, which are necessary and sufficient optimality conditions for Markov decision problems. Then we are constructing optimal policies out of the solution of the Bellman equations. We will see that the class of Markov deterministic policies—that are easier to handle—contain, under assumptions which are often satisfied in practise, optimal policies. Finally, we describe how optimal policies can be calculated, based on a backward induction algorithm. This chapter is based on [Put94], [Whi93], and [Der70].

2.1 Optimal Policies and the Bellman Equations

In order to be able to speak about optimal policies we need a method for comparing two policies. In the finite horizon case we can simply choose the so called expected reward criterion. Let π be a history dependent randomized policy and define

$$v_N^\pi(s) := E_s^\pi \left(\sum_{t=1}^{N-1} r_t(X_t, Y_t) + r_N(X_N) \right),$$

where the right hand side is the conditional expectation of the sum of the expected rewards $r_t(X_t, Y_t)$ and the final reward $r_N(X_N)$ conditioned on $X_1 = s$. Since the last reward $r_N(X_N)$ only depends on the last state and not on an action anymore, we have to write it down separately. Recall that we defined $X_t := s_t$ and $Y_t := a_t$. To

be able to speak about expectations we need a suitable sample space $(\Omega, \mathcal{A}, P_N^\pi)$, which we defined for discrete A and S in Section 1.2. If we have $\pi \in \mathcal{D}^{DM}$ the action chosen at decision epoch t is determined by $d_t(s_t)$, so in this case we have

$$v_N^\pi(s) = E_s^\pi \left(\sum_{t=1}^{N-1} r_t(X_t, d_t(X_t)) + r_N(X_N) \right).$$

Definition 2.1 (Optimal policies and ε -optimal policies). *A policy π^* is called optimal if for every starting state $s \in S$ and all $\pi \in \mathcal{D}^{RH}$*

$$v_N^{\pi^*}(s) \geq v_N^\pi(s).$$

Fix $\varepsilon > 0$. We say that the policy π_ε^ is ε -optimal if for every starting state $s \in S$ and all $\pi \in \mathcal{D}^{RH}$*

$$v_N^{\pi_\varepsilon^*}(s) + \varepsilon > v_N^\pi(s).$$

Definition 2.2 (Value of a Markov decision problem). *We define the value of a Markov decision problem as*

$$v_N^*(s) := \sup_{\pi \in \mathcal{D}^{RH}} v_N^\pi(s).$$

If S and A are finite and the rewards are bounded then the supremum exists and can be replaced by the maximum of the right hand side of the equation above. Clearly, in this case we have $v_N^*(s) = v_N^{\pi^*}(s)$, so the expected total reward equals the value of a Markov decision problem if an optimal policy π^* is used. If we use a ε -optimal policy we have $v_N^{\pi_\varepsilon^*}(s) + \varepsilon > v_N^*(s)$.

2.1.1 Policy Evaluation

Now we want to find a method which allows us to calculate the expected total reward $v_N^\pi(s)$ for a given policy π . We want to do this in a backward inductive way. We define functions $u_t^\pi : H_t \rightarrow \mathbb{R}$,

$$u_t^\pi(h_t) := E_{h_t}^\pi \left(r_N(X_N) + \sum_{n=t}^{N-1} r_n(X_n, Y_n) \right), \quad (2.1)$$

the expected total reward from decision epoch t on given the history h_t up to time t . Furthermore we define $u_N^\pi(h_N) := r_N(s_N)$ where $h_N = (h_{N-1}, a_{N-1}, s_N)$. Note that if $h_1 = s$ and $t = 1$ we have

$$u_1^\pi(s) = E_s^\pi \left(r_N(X_N) + \sum_{n=1}^{N-1} r_n(X_n, Y_n) \right) = v_N^\pi(s). \quad (2.2)$$

So if we are able to calculate $u_1^\pi(s)$ we know the value of the Markov decision problem if an optimal policy is used. Algorithm 1 calculates the functions u_t^π in a backward inductive way.

Algorithm 1 Finite Horizon Policy Evaluation Algorithm for $\pi \in \mathcal{D}^{RH}$

- 1: Set $u_N^\pi(h_N) = r_N(s_N)$ for each $h_N = (h_{N-1}, a_{N-1}, s_N) \in H_N$.
- 2: $t \leftarrow N$
- 3: **for** $t \neq 1$ **do**
- 4: $t \leftarrow t - 1$
- 5: Compute

$$u_t^\pi(h_t) = \sum_{a \in A_{s_t}} q_{d_t(h_t)}(a) \left(r_t(s_t, a) + \sum_{\sigma \in S} p_t(\sigma | s_t, a) u_{t+1}^\pi((h_t, a, \sigma)) \right)$$

for each $h_t \in H_t$.

- 6: **end for**
-

Of course we need to show that the $u_t^\pi(h_t)$ constructed by the algorithm are the same as these defined in (2.1).

Theorem 2.3. *Let π be a randomized history dependent policy and $u_t^\pi(h_t)$ be constructed by Algorithm 1. Then these functions are equal to the right hand side of (2.1) for all $t \leq N$, particularly $u_1^\pi(s)$ is the value of the underlying Markov decision problem if an optimal policy is used.*

Proof. We prove the claim with backward induction. Within this proof u_t^π always denotes the functions generated by Algorithm 1. We see that $u_N^\pi(h_N) = r_N(s_N)$ coincides with the definition in (2.1). Assume that $u_t^\pi(h_t) = E_{h_t}^\pi \left(r_N(X_N) + \sum_{k=t}^{N-1} r_k(X_k, Y_k) \right)$ is true for all $t = n+1, \dots, N$. Now we calculate for $t = n$

$$\begin{aligned}
u_n^\pi(h_n) &= \sum_{a \in A_{s_n}} q_{d_n(h_n)}(a) r_n(s_n, a) \\
&\quad + \sum_{\sigma \in S} \sum_{a \in A_{s_n}} q_{d_n(h_n)} p_n(\sigma | s_n, a) u_{n+1}^\pi(h_n, a, \sigma) \\
&\stackrel{(1.1)}{=} r_n(s_n, d_n(h_n)) \\
&\quad + \sum_{\sigma \in S} p_n(\sigma | s_n, d_n(h_n)) u_{n+1}^\pi(h_n, d_n(h_n), \sigma) \\
&= r_n(s_n, d_n(h_n)) + E_{h_n}^\pi (u_{n+1}^\pi(h_n, d_n(h_n), X_{n+1})) \\
&= r_n(s_n, d_n(h_n)) \\
&\quad + E_{h_n}^\pi \left(E_{h_{n+1}}^\pi \left(\sum_{k=n+1}^{N-1} r_k(X_k, Y_k) + r_N(X_N) \right) \right) \\
&= r_n(s_n, d_n(h_n)) + E_{h_n}^\pi \left(\sum_{k=n+1}^{N-1} r_k(X_k, Y_k) + r_N(X_N) \right) \\
&= E_{h_n}^\pi \left(r_n(X_n, Y_n) + \sum_{k=n+1}^{N-1} r_k(X_k, Y_k) + r_N(X_N) \right) \\
&= E_{h_n}^\pi \left(\sum_{k=n}^{N-1} r_k(X_k, Y_k) + r_N(X_N) \right),
\end{aligned}$$

which finishes the main part of the proof. As we have seen in (2.2) we also have $u_1^\pi(s) = v_N^\pi(s)$, so if we have an optimal policy we know the value of the underlying Markov decision problem. \square

We want to have a look at the complexity of the policy evaluation algorithm. The crucial point is that in every decision epoch

we have to evaluate u_t^π for every possible history h_t . Let there be α states and β actions and N decision epochs. Then there are altogether $\alpha^N \beta^{N-1}$ possible histories since histories are of the form $h_N = (s_1, a_1, s_2, a_2, \dots, a_{N-1}, s_N)$. Up to time t there are $\alpha^t \beta^{t-1}$ possible histories. Now fix a policy π , a decision epoch t and a particular history h_t up to time t . To calculate $u_t^\pi(h_t)$ Algorithm 1 needs $\alpha\beta$ multiplications, see line five. To construct u_t^π we need to evaluate $u_t^\pi(h_t)$ for every $h_t \in H_t$, so $\alpha\beta\alpha^t\beta^{t-1} = \alpha^{t+1}\beta^t$ multiplications are needed. Now we iterate over all decision epochs, so all together $\sum_{t=1}^{N-1} \alpha^{t+1}\beta^t$ multiplications are needed to evaluate a single policy. Additionally we have to store $\alpha^N \beta^{N-1}$ numbers in the beginning of the algorithm in order to construct u_N^π .

So this is quite a lot of work to do. Luckily we will see that we only have to give attention to deterministic Markovian decision rules. In the Markovian case u_t^π are actually functions from S to \mathbb{R} since the actions chosen only depend on the current state s_t and not on the entire past h_t . Under this assumption we have

$$\begin{aligned} u_t^\pi(h_t) &= E_{h_t}^\pi \left(\sum_{n=t}^{N-1} r_n(X_n, Y_n) + r_N(X_N) \right) \\ &= E_{s_t}^\pi \left(\sum_{n=t}^{N-1} r_n(X_n, Y_n) + r_N(X_N) \right). \end{aligned}$$

Consequently we do not have to look at each history h_t like we did in Algorithm 1. We rewrite the policy evaluation algorithm for the case $\pi \in \mathcal{D}^{DM}$. Now we have a look at line five of Algorithm 2. We need α multiplications for a single fixed s_t , so constructing u_t^π needs only α^2 multiplications. We again iterate over the entire time horizon which leads to $\sum_{t=1}^{N-1} \alpha^2 = (N-1)\alpha^2$ multiplications altogether. Additionally we only have to store α numbers to construct u_N^π . Theorem 2.3 also includes the correctness of Algorithm 2 if deterministic Markovian policies are interpreted as degenerated probability measures.

Algorithm 2 Finite Horizon Policy Evaluation Algorithm for $\pi \in \mathcal{D}^{DM}$

- 1: for $t = N$ set $u_N^\pi(s_N) = r_N(s_N) \quad \forall s_N \in S$.
 - 2: $t \leftarrow N$
 - 3: **for** $t \neq 1$ **do**
 - 4: $t \leftarrow t - 1$
 - 5: $u_t^\pi(s_t) = r_t(s_t, d(s_t)) + \sum_{\sigma \in S} p_t(\sigma | s_t, d_t(s_t)) u_{t+1}^\pi(\sigma)$ for all $s_t \in S$
 - 6: **end for**
-

2.1.2 The Bellman Equations

Now we can take a look at the Bellman equations. The solution of the equations will help us to find optimal policies.

We define the functions

$$u_t^*(h_t) = \sup_{\pi \in \mathcal{D}^{RH}} u_t^\pi(h_t) \quad (2.3)$$

which are the best possible expected total rewards from decision epoch t onwards if the history until t equals h_t and the supremum is attained. This is, e.g., the case if we deal with finite S and A .

The Bellman equations, often also called optimality equations, are given by

$$u_t(h_t) = \sup_{a \in A_{s_t}} \left(r_t(s_t, a) + \sum_{\sigma \in S} p_t(\sigma | s_t, a) u_{t+1}(h_t, a, \sigma) \right) \quad (2.4a)$$

for $1 \leq t \leq N - 1$, and

$$u_N(h_N) = r_N(s_N). \quad (2.4b)$$

The solution of the Bellman equations is a sequence of functions $u_t : H_t \rightarrow \mathbb{R}$ fulfilling (2.4a) and (2.4b). We will prove that the solution of the Bellman equations fulfil (2.3), so if we solve the Bellman equations we obtain a finite sequence $u_t : H_t \rightarrow \mathbb{R}$ of functions that tell us what the best possible expected reward from decision epoch t

onwards is if the history up to time t equals h_t . Before we do this we need a small lemma.

Lemma 2.4. *Let f be a real-valued function on a discrete set Ω and let $p(\cdot)$ be a probability distribution on Ω . Then we have*

$$\sup_{\omega \in \Omega} f(\omega) \geq \sum_{\omega \in \Omega} p(\omega) f(\omega).$$

Proof. Set $\omega^* := \sup_{\omega \in \Omega} f(\omega)$. We easily calculate

$$\omega^* = \omega^* \sum_{\omega \in \Omega} p(\omega) = \sum_{\omega \in \Omega} p(\omega) \omega^* \geq \sum_{\omega \in \Omega} p(\omega) f(\omega).$$

□

Now we can state the main property of the solution of the Bellman equations.

Theorem 2.5. *Suppose the family $u_t, t = 1, \dots, N$ is a solution of the Bellman equations. Then we have*

$$u_t(h_t) = u_t^*(h_t)$$

for all $h_t \in H_t$ and $t = 1, \dots, N$. Moreover we have $u_1(s) = v_N^*(s)$ for all $s \in S$, i.e., u_1 equals the value of the underlying Markov decision problem.

Proof. Within this proof we denote the solution¹ of the Bellman equations (assuming it exists) by $u_t, t = 1, \dots, N$. We are starting with proving by backward induction that

$$u_t(h_t) \geq u_t^*(h_t) \quad \forall h_t \in H_t, \quad t = 1, \dots, N. \quad (2.5)$$

Note that we have for an arbitrary policy π by definition $u_N^\pi(h_N) = E_{h_N}(r_N(X_N)) = r_N(s_N)$ since by conditioning on an arbitrary history

¹If S and A are finite we only have to assume that $r_t(\cdot, \cdot)$ and $r_N(\cdot)$ are bounded, then a unique solution always exists.

$h_N \in H_N$ the random variable X_N is known. By (2.4b) we have $u_N(h_N) = r_N(s_N) = u_N^\pi(h_N)$ for all $h_N \in H_N$ and an arbitrary $\pi \in \mathcal{D}^{RH}$. So consequently we have $u_N(h_N) = u_N^*(h_N)$ for all $h_N \in H_N$ and of course therefore $u_N(h_N) \geq u_N^*(h_N)$ for all $h_N \in H_N$. Now assume that

$$u_t(h_t) \geq u_t^*(h_t) \quad \forall h_t \in H_t, \quad t = n+1, \dots, N,$$

and let $\tilde{\pi} := (\tilde{d}_1, \dots, \tilde{d}_N)$ be an arbitrary randomized history dependent policy. For $t = n$ we have

$$\begin{aligned} u_n(h_n) &\stackrel{(2.4)}{=} \sup_{a \in A_{s_n}} \left(r_n(s_n, a) + \sum_{\sigma \in S} p_n(\sigma | s_n, a) u_{n+1}(h_n, a, \sigma) \right) \\ &\stackrel{\text{i.h.}}{\geq} \sup_{a \in A_{s_n}} \left(r_n(s_n, a) + \sum_{\sigma \in S} p_n(\sigma | s_n, a) u_{n+1}^*(h_n, a, \sigma) \right) \\ &\stackrel{(2.3)}{\geq} \sup_{a \in A_{s_n}} \left(r_n(s_n, a) + \sum_{\sigma \in S} p_n(\sigma | s_n, a) u_{n+1}^{\tilde{\pi}}(h_n, a, \sigma) \right) \\ &\stackrel{2.4}{\geq} \sum_{a \in A} q_{\tilde{d}_n(h_n)}(a) \\ &\quad \left(r_n(s_n, a) + \sum_{\sigma \in S} p_n(\sigma | s_n, a) u_{n+1}^{\tilde{\pi}}(h_n, a, \sigma) \right) \\ &\stackrel{(2.3)}{=} u_n^{\tilde{\pi}}(h_n). \end{aligned}$$

Because $\tilde{\pi}$ was arbitrary we showed $u_t(h_t) \geq u_t^*(h_t)$ for all $h_t \in H_t$, $t = 1, \dots, N$. So we have proved the claim (2.5). Now we want to show that, for an arbitrary $\varepsilon > 0$, there exists a policy π for which we have

$$u_t^\pi(h_t) + (N - t)\varepsilon \geq u_t(h_t) \quad \forall h_t \in H_t, \quad t = 1, \dots, N. \quad (2.6)$$

To do so we choose any policy² $\pi = (d_1, \dots, d_{N-1})$ which fulfills for all $t = 1, \dots, N$

$$r_t(s_t, d_t(h_t)) + \sum_{\sigma \in S} p_t(\sigma | s_t, d_t(h_t)) u_{t+1}(s_t, d_t(h_t), \sigma) + \varepsilon \geq u_t(h_t).$$

We again proof the claim (2.6) by backward induction. We have $u_N^\pi(h_N) = u_N(h_N)$ for an arbitrary policy π , so (2.6) clearly holds for $t = N$. Now assume that (2.6) is valid for all $t = n + 1, \dots, N$. Then we have

$$\begin{aligned} u_n^\pi(h_n) &= r_n(s_n, d_n(h_n)) + \sum_{\sigma \in S} p_n(\sigma | s_n, d_n(h_n)) u_{n+1}^\pi(s_n, d_n(h_n), \sigma) \\ &\geq r_n(s_n, d_n(h_n)) \\ &\quad + \sum_{\sigma \in S} p_n(\sigma | s_n, d_n(h_n)) (u_{n+1}(h_n, d_n(h_n), \sigma) - (N - n - 1)\varepsilon) \\ &= -(N - n)\varepsilon + r_n(s_n, d_n(h_n)) \\ &\quad + \sum_{\sigma \in S} p_n(\sigma | s_n, d_n(h_n)) u_{n+1}(h_n, d_n(h_n), \sigma) + \varepsilon \\ &\geq u_n(h_n) - (N - n)\varepsilon. \end{aligned}$$

This proves the claim (2.6). By definition we have $u_t^*(\cdot) \geq u_t^\pi(\cdot)$ for all possible policies. Therefore we have

$$u_t^*(h_t) + (N - t)\varepsilon \geq u_t^\pi(h_t) + (N - t)\varepsilon \stackrel{(2.6)}{\geq} u_t(h_t) \stackrel{(2.5)}{\geq} u_t^*(h_t).$$

Now let us set $\tilde{\varepsilon} := (N - t)\varepsilon$. Then we have

$$u_t^*(h_t) + \tilde{\varepsilon} \geq u_t(h_t) \geq u_t^*(h_t),$$

which means $u_t^*(h_t) = u_t(h_t)$ since ε was arbitrary. Moreover because of $v_N^*(s) = u_1^*(s)$ we also have $u_1(s) = v_N^*(s)$. \square

²Such a policy clearly exists. Note that without adding ε we would have equality by definition if the image of h_t under d_t equals the best possible action.

Now we show how to use the solution of the Bellman equations to construct optimal policies. At first we take a look at the case when the Bellman equations attain the suprema.

Theorem 2.6. *Let $u_t^*, t = 1, \dots, N$ be a finite sequence of functions which solve the Bellman equations (2.4a) and (2.4b) and assume that the policy $\mathcal{D}^{DH} \ni \pi^* := (d_1^*, \dots, d_{N-1}^*)$ satisfies*

$$d_t^*(h_t) \in \operatorname{argmax}_{a \in A_{s_t}} r_t(s_t, a) + \sum_{\sigma \in S} p_t(\sigma | s_t, a) u_{t+1}^*(h_t, a, \sigma). \quad (2.7)$$

Then we have

$$u_t^{\pi^*}(h_t) = u_t^*(h_t), \quad h_t \in H_t.$$

Moreover π^* is an optimal policy since we have $v_N^{\pi^*}(s) = v_N^*(s)$.

Remark 2.7 (Randomized vs. deterministic policies). Note that Theorem 2.6 is already stated for deterministic history dependent policies. This is not a restriction if seen in the following way: Let us rewrite (2.7) for randomized history dependent policies,

$$q_{d_t(h_t)}^*(\cdot) \in \operatorname{argmax}_{q_{d_t(h_t)}(\cdot) \in \mathcal{P}(A_{s_t})} \mathcal{U}(q_{d_t(h_t)})$$

with

$$\mathcal{U}(q_{d_t}) := \sum_{a \in A_{s_t}} q_{d_t}(a) \left(r_t(s_t, a) + \sum_{\sigma \in S} p_t(\sigma | s_t, a) u_{t+1}^*(h_t, a, \sigma) \right).$$

We maximize over all possible probability distributions, which are by definition randomized policies, on A_{s_t} . Now let us assume that we have found an optimal probability distribution $q_{d_t(h_t)}^*$. Then we can guarantee the existence of a deterministic history dependent policy. To construct it set $f_{h_t}(a) := r_t(s_t, a) + \sum_{\sigma \in S} p_t(\sigma | s_t, a) u_{t+1}^*(h_t, a, \sigma)$ and $p(a) := q_{d_t(h_t)}^*(a)$ and use Lemma 2.4, which basically tells us that we are also doing well with a degenerated probability distribution $\tilde{q}_{d_t(h_t)}^*$. To construct $\tilde{q}_{d_t(h_t)}^*$ fix any $a^* \in \operatorname{argmax}_{a \in A_{s_t}} f_{h_t}(a)$ and define

$$\tilde{q}_{d_t(h_t)}^*(a) = \begin{cases} 1 & \text{if } a = a^* \\ 0 & \text{otherwise.} \end{cases}$$

<http://www.springer.com/978-3-658-08343-4>

Optimized Response-Adaptive Clinical Trials
Sequential Treatment Allocation Based on Markov
Decision Problems

Ondra, Th.

2015, XV, 102 p. 14 illus., Softcover

ISBN: 978-3-658-08343-4