

UNIVERSITY OF MANNHEIM

Bachelor Thesis

# Markov-Decision Processes

by

**Felix Benning**

born on the 27.11.1996 in Nürtingen

matriculation number 1501817

in the

Fakulty for Mathematics in Business and Economics

Supervisor: Prof. Dr. Leif Döring

Due Date: ???



# Declaration of Authorship

I hereby declare that the thesis submitted is my own unaided work. All direct or indirect sources used are acknowledged as references.

This thesis was not previously presented to another examination board and has not been published.

City, Date

Signature



# Contents

<b>1</b>	<b>Markov Decision Processes</b>	<b>1</b>
1.1	Introduction . . . . .	1
1.2	Model Formulation . . . . .	2
1.3	Value functions . . . . .	5
<b>2</b>	<b>Title Chapter 2</b>	<b>13</b>
<b>A</b>	<b>Appendix</b>	<b>15</b>
A.1	Basic Probability Theory . . . . .	15
A.2	Analysis . . . . .	15



# Chapter 1

## Markov Decision Processes

### 1.1 Introduction

A Markov Process is a random process in a state space with no memory of where it was, that is, only the current state influences where it goes next. While Markov Processes allow to model random phenomena evolving over time and make predictions about certain events (e.g. terminal states), they are unable to model the interaction of an actor with such a processes. *Markov Decision Processes* (MDPs) introduce *actions* and *rewards* to the state space and transition probabilities of Markov Processes, and shift the focus from *describing* terminal distributions, absorption times, etc. towards *finding* the optimal action(s) to take in each state (if such an action exists).

The MDP model inherits the restriction of Markov Chains to have no memory of past states. We will also not consider changing transition probabilities over time. Rather the transition probabilities will only be influenced by the state and the action.

Both of these limitations could in principle be circumvented by including the time in the state space at the expense of a larger state space. Although it is questionable whether such a construct would yield any interesting results, as then no state is visited twice. So it is of no use to an actor to learn the value of an action in a certain state without further assumptions.

To illustrate the uses of such a framework, I have selected a few examples from White (1985):

1. Resource Management: The state is the resource level
  - Inventory Management: The resource is the inventory, the possible action is to order resupply, influencing the inventory (state) together with the stochastic demand, and the reward is the profit. The essential trade-off is the cost of storage versus lost sales from a stock-out.
  - Fishing: The resource is the amount of fish, the action is the amount fished, the reward is directly proportional to the amount fished, and

the repopulation is the random element.

- Pumped storage Hydro-power: The state is the amount of water in the higher reservoir and the electricity price, the action is to use water to generate electricity or wait for higher prices.
  - Beds in a hospital: How many empty beds are needed for emergencies?
2. Stock trading: The state is the price level and stock and liquidity owned.
  3. Maintenance: When does a car/road become too expensive to repair?
  4. Evacuation in response to flood forecasts

## 1.2 Model Formulation

Most of the definitions in this chapter are adaptations from Szepesvári (2010). But to properly define the transition probabilities given an action in a certain state, let us define a probability kernel first.

**Definition 1.2.1.** (Kernel) Let  $(Y, \sigma_Y), (X, \sigma_X)$  be measure spaces.

$$\begin{aligned} \lambda: X \times \sigma_Y &\rightarrow \mathbb{R} \text{ is a (probability) kernel} \\ : \iff \lambda(\cdot, A): x &\mapsto \lambda(x, A) \text{ measurable} \\ \lambda(x, \cdot): A &\mapsto \lambda(x, A) \text{ a (probability) measure} \end{aligned}$$

Since we will interpret probability kernels as distributions over  $Y$  given a certain condition  $x \in X$ , the notation  $\lambda(\cdot | x) := \lambda(x, \cdot)$  helps this intuition.

**Definition 1.2.2.**

$\mathcal{M} = (\mathcal{X}, \mathcal{A}, \mathcal{P}_0)$  is called a (*finite*) *Markov Decision Process* (MDP). Where:

$\mathcal{X}$  is a countable (finite) set of states.

$\mathcal{A}$  is a countable (finite) set of actions.

And  $\mathcal{P}_0: (\mathcal{X} \times \mathcal{A}) \times \sigma_{\mathcal{X} \times \mathbb{R}} \rightarrow \mathbb{R}$  is a probability kernel.

$\mathcal{X} \times \mathbb{R}$  represents the next state and the reward. So  $\mathcal{P}_0(\cdot | x, a)$  represents the probability distribution over the next states and rewards given an action  $a$  in the state  $x$ .

$\mathcal{P}_0$  is called the *transition probability kernel* or in short, transition kernel.

*Remark 1.2.3.* Instead of a transition probability kernel  $\mathcal{P}_0$ , sometimes a *transition function*  $f$  with a and an exogenous random element  $D_t$  (e.g. Demand) is used to define the next state and reward:  $(X_{t+1}, R_{t+1}) = f(X_t, A_t, D_t)$

Some authors include a Time set  $T$  in the tuple (e.g. Puterman 2014) this allows for finite horizons but not for continuous time, since the transition kernel



is defined for discrete steps. Most authors split the transition kernel into a state transition kernel and a reward kernel (e.g. Puterman 2014). But since it is easier to define a marginal distribution from a joint distribution than vice versa, and since this notation is more compact I will stick to the definition from Szepesvári (2010).

According to Puterman (2014) some authors call this tuple a Markov Decision Problem instead of Markov Decision Process, presumably to reserve the term Markov Decision Process for the resulting sequence of states, actions and rewards  $(X_t, A_t, R_{t+1}, t \in \mathbb{N}_0)$ , aligning the Definition with the definition of a Markov process. Although this does not appear to be common practice.

I can find no explanation for this deviation from the notation of Markov processes. So I offer my own interpretation:

The objective of the theory of MDPs is to find an optimal action selection rule (behavior). And without a fixed behavior the sequence  $(X_t, A_t, R_{t+1}, t \in \mathbb{N}_0)$  is undefined, since the  $(A_t, t \in \mathbb{N}_0)$  are not defined. But fixing the behavior defeats the purpose of modeling decisions. As it would not make sense to talk about optimal behaviors in an MDP if every behavior creates its own MDP.

Nevertheless we still need to construct a stochastic process from the MDP when we have an action selection rule.

First we need to select the random variable  $X_0$  of the initial state. The initial state is not included in the definition of an MDP because later objects will be defined conditional on the current state. They are thus invariant to different starting distributions, as long as  $\mathbb{P}(X_0 = x) > 0$  holds for all  $x \in \mathcal{X}$  ensuring that conditioning on every state is possible.

To inductively define a stochastic process we need an action selection rule, more formally:

**Definition 1.2.4.** An  $A_t$  selection-rule  $\pi = (\pi_t, t \in \mathbb{N}_0)$  is called *behavior*, where

$$\pi_t: \begin{cases} [(\mathcal{X} \times \mathcal{A} \times \mathbb{R})^t \times \mathcal{X}] \times \sigma_{\mathcal{A}} \rightarrow \mathbb{R} \\ (y, A) \mapsto \pi_t(A \mid y) \end{cases} \quad \text{is a probability kernel,}$$

and  $A_t \sim \pi_t(\cdot \mid (X_0, A_0, R_1), \dots, (X_{t-1}, A_{t-1}, R_t), X_t)$ .

Special cases:

1. *Deterministic stationary policies* specified with some abuse of notation:

$$\pi: \mathcal{X} \rightarrow \mathcal{A} \text{ with } A_t = \pi(X_t)$$

2. *(Stochastic) stationary policies* specified by:

$$\pi: \begin{cases} \mathcal{X} \times \sigma_{\mathcal{A}} \rightarrow \mathbb{R} \\ (x, A) \mapsto \pi(A \mid x) \end{cases} \quad \text{with } A_t \sim \pi(\cdot \mid X_t)$$

$A_t$  is here selected such that it has the markov property (well defined c.f. 1.2.5), i.e.:

$$\mathbb{P}[A_t = a \mid X_t] = \mathbb{P}[A_t = a \mid X_t, (X_{t-1}, A_{t-1}, R_t), \dots (X_0, A_0, R_1)]$$

$\Pi$  is the set of behaviors,

$\Pi_{\text{stat}}$  is the set of (stochastic) stationary policies,

$\Pi_{\text{stat}}^{\text{det}}$  is the set of deterministic stationary policies (note  $\Pi_{\text{stat}}^{\text{det}} \subseteq \Pi_{\text{stat}} \subseteq \Pi$ )

Now we define inductively:  $(X_{t+1}, R_{t+1}) \sim \mathcal{P}_0(\cdot \mid X_t, A_t)$  with the Markov property (well defined c.f. 1.2.5), i.e.:

$$\begin{aligned} \mathbb{P}[(X_{t+1}, R_{t+1}) = (x, r) \mid (X_t, A_t)] \\ = \mathbb{P}[(X_{t+1}, R_{t+1}) = (x, r) \mid (X_t, A_t), (X_{t-1}, A_{t-1}, R_t), \dots (X_0, A_0, R_1)] \end{aligned} \quad (1.1)$$

resulting in the stochastic process  $((X_t, A_t, R_{t+1}), t \in \mathbb{N}_0)$

*Remark 1.2.5.*  $(X_{t+1}, R_{t+1}) \sim \mathcal{P}_0(\cdot \mid X_t, A_t)$  with the Markov property, is well defined i.e.:

$\exists (X_{t+1}, R_{t+1}) \mathcal{X} \times \mathbb{R}$ -valued random variable :

$(X_{t+1}, R_{t+1}) \sim \mathcal{P}_0(\cdot \mid X_t, A_t)$  and satisfies the Markov property

(analogous  $A_t$  well defined)

*Proof.* **TODO** □

*Remark 1.2.6.* A stationary policy  $\pi$  induces a *time-homogenous* Markov chain  $(X_t, A_t, R_{t+1}, t \in \mathbb{N}_0)$ .

*Proof.*  $\mathcal{H}_s^t := \{(X_t, A_t, R_{t+1}) \in H_t, \dots, (X_s, A_s, R_{s+1}) \in H_s\}$ ,  $H_i \in \sigma_{\mathcal{X} \times \mathcal{A} \times \mathbb{R}}$   
Because of

$$\mathbb{P}(A \cap B \mid C) = \frac{\mathbb{P}(A \cap B \cap C)}{\mathbb{P}(B \cap C)} \frac{\mathbb{P}(B \cap C)}{\mathbb{P}(C)} = \mathbb{P}(A \mid B \cap C) \mathbb{P}(B \mid C)$$

we can show:

$$\begin{aligned} & \mathbb{P}[(X_t, A_t, R_{t+1}) \in \{(x, a)\} \times U \mid \mathcal{H}_0^{t-1}] \\ &= \mathbb{P}[R_{t+1} \in U \mid (X_t, A_t) = (x, a), \mathcal{H}_0^{t-1}] \mathbb{P}[(X_t, A_t) = (x, a) \mid \mathcal{H}_0^{t-1}] \\ &\stackrel{(1.1)}{=} \mathbb{P}[R_{t+1} \in U \mid (X_t, A_t) = (x, a)] \underbrace{\mathbb{P}[A_t = a \mid X_t = x]}_{=\pi(a|x)} \mathbb{P}[X_t = x \mid \mathcal{H}_{t-1}^{t-1}] \\ &\stackrel{(*)}{=} \mathbb{P}[R_{t+1} \in U \mid (X_t, A_t) = (x, a), \mathcal{H}_{t-1}^{t-1}] \mathbb{P}[(X_t, A_t) = (x, a) \mid \mathcal{H}_{t-1}^{t-1}] \\ &= \mathbb{P}[(X_t, A_t, R_{t+1}) \in \{(x, a)\} \times U \mid \mathcal{H}_{t-1}^{t-1}] \end{aligned}$$

(\*) *Some* of the History is irrelevant if all of the History is irrelevant. (**ap-  
pendix?**) □

*Remark 1.2.7.* Sometimes not all actions make sense in all states. A simple fix would be to set the immediate reward functions for those actions very low, or (if possible) redirect them to the closest possible action.

A more formal approach would be to introduce an additional mapping, which assigns the set of admissible actions to each state  $\mathcal{X} \rightarrow \mathcal{P}(\mathcal{A})$  and define  $\mathcal{X} \times \mathcal{A} := \{(x, a) : x \in \mathcal{X}, a \in f(x)\}$ .

If there is just one admissible action in every state, the MDP is equivalent to a normal Markov Process. Since then there is a mapping  $f: \mathcal{X} \rightarrow \mathcal{A}$  mapping the state to the only admissible action. Which implies that  $A_t = f(X_t)$  which forces every behavior to be equal to  $f$ . And since  $f$  is a deterministic stationary behavior, 1.2.6 applies.

**Definition 1.2.8.** An MDP together with a discount factor  $\gamma \in [0, 1]$  is a

*discounted reward MDP* for  $\gamma < 1$

*undiscounted reward MDP* for  $\gamma = 1$

This allows us to define the *return*:

$$\mathcal{R} := \sum_{t=0}^{\infty} \gamma^t R_{t+1}$$

**Definition 1.2.9.** Let  $(Y_{(x,a)}, R_{(x,a)}) \sim \mathcal{P}_0(\cdot | x, a)$  be a random variable.

$r(x, a) := \mathbb{E}[R_{(x,a)}]$  is the *immediate reward function*.

**Definition 1.2.10.**  $\mathcal{M} = (\mathcal{X}, \mathcal{A}, \mathcal{P}_0)$  a MDP

$x \in \mathcal{X}$  is a *terminal (absorbing) state* :  $\iff \forall s \in \mathbb{N} : \mathbb{P}(X_{t+s} = x | X_t = x) = 1$

An MDP with such states is called *episodic*.

An *episode* is the random time period  $(1, \dots, T)$  until a terminal state is reached.

*Remark 1.2.11.*

- The reward in a terminal state is by convention zero, i.e.  $x$  terminal state implies  $\forall a \in \mathcal{A} : R_{(x,a)} = 0$ .
- Episodic MDPs are often undiscounted

To avoid clutter we will from now on we assume an underlying MDP with the accompanying definitions and notation.

## 1.3 Value functions

The goal in this section is to

- define Value functions which assign states (and actions) a value, which allow the agent to make a more nuanced decisions than comparing immediate rewards of different actions

- explore the relation of different value functions
- show uniqueness of optimal value functions with the Banach fixpoint theorem, yielding a simple approximation methode along the way
- demonstrate that in MDPs deterministic stationary policies are generally a large enough set of policies to choose from

**Definition 1.3.1.** Let  $\pi$  be a behavior. Select  $X_0$  such that  $\forall x \in \mathcal{X} : \mathbb{P}(X_0 = x) > 0$ , be  $((X_t, A_t, R_{t+1}), t \in \mathbb{N}_0)$  the resulting stoch. process.

$$\begin{aligned}
 V^\pi &: \begin{cases} \mathcal{X} \rightarrow \mathbb{R} \\ x \mapsto \mathbb{E}[\mathcal{R} \mid X_0 = x] \end{cases} && \text{is the \textit{value function} for } \pi^1 \\
 Q^\pi &: \begin{cases} \mathcal{X} \times \mathcal{A} \rightarrow \mathbb{R} \\ (x, a) \mapsto \mathbb{E}[\mathcal{R} \mid X_0 = x, A_0 = a] \end{cases} && \text{is the \textit{action value} \\ && \text{function for } \pi^2 \\
 V^* &: \begin{cases} \mathcal{X} \rightarrow \mathbb{R} \\ x \mapsto \sup_{\pi \in \Pi} V^\pi(x) \end{cases} && \text{is the \textit{optimal value function} \\
 Q^* &: \begin{cases} \mathcal{X} \times \mathcal{A} \rightarrow \mathbb{R} \\ (x, a) \mapsto \sup_{\pi \in \Pi} Q^\pi(x, a) \end{cases} && \text{is the \textit{optimal action} \\ && \text{value function}
 \end{aligned}$$

$\pi$  is *optimal* :  $\iff V^* = V^\pi$

*Remark 1.3.2.* With the distribution of  $X_0$  set (or  $X_0$  being realized with a fixed value  $x$ ), the distribution of  $X_t, A_t, R_{t+1}$  is determined for all  $t \in \mathbb{N}_0$ . The conditional expectation is thus unique for a given  $X_0 = x$ , for all possible realizations of the MDP with a given behavior.

This means  $V^\pi, Q^\pi$  are well defined.

From now on we assume:

**Assumption 1.**  $\forall (x, a) \in \mathcal{X} \times \mathcal{A} : |R_{(x,a)}| \leq R \in \mathbb{R}$  almost surely.

This implies that the immediate reward is bounded:

$$\|r\|_\infty = \sup_{(x,a) \in \mathcal{X} \times \mathcal{A}} |\mathbb{E}[R_{(x,a)}]| \leq R$$

But more importantly:

$$|\mathcal{R}| \leq \sum_{t=0}^{\infty} \gamma^t |R_{t+1}| \leq \frac{R}{1-\gamma} \quad \text{a.s.} \implies V^\pi, V^* \in B(\mathcal{X}), Q^\pi, Q^* \in B(\mathcal{X} \times \mathcal{A})$$

with  $B(M) := \{f : M \rightarrow \mathbb{R} : \|f\|_\infty < \infty\}$  the set of bounded functions on  $\mathcal{X}$ .

As mentioned previously, the marginal probability distribution of the state instead of the joint distribution with the reward, will now makes some notation shorter.

<sup>1</sup>Well defined because  $\mathbb{P}(X_0 = x) > 0$

<sup>2</sup>Well defined because  $A_1 \sim \pi_1(\cdot \mid (x, a, r_0), x_1)$  is defined for all  $a$  regardless of  $\pi_0$

**Definition 1.3.3.**

$$p: \begin{cases} (\mathcal{X} \times \mathcal{A}) \times \sigma_{\mathcal{X}} \rightarrow \mathbb{R} \\ (x, a, Y) \mapsto \mathcal{P}_0(Y \times \mathbb{R} \mid x, a) \end{cases} \quad \text{is the state transition kernel.}$$

*Notation:*  $p(y \mid x, a) := p(\{y\} \mid x, a)$  with  $(x, a, y) \in \mathcal{X} \times \mathcal{A} \times \mathcal{X}$

We start by setting  $V^\pi$  and  $Q^\pi$  in relation to each other.

**Proposition 1.3.4.** *Let  $\pi \in \Pi_{\text{stat}}^{\text{det}}$  be a deterministic stationary behavior, then:*

$$Q^\pi(x, a) = r(x, a) + \gamma \sum_{y \in \mathcal{X}} p(y \mid x, a) V^\pi(y)$$

*Proof.* split equation or accept eq sticking out a bit?

$$\begin{aligned} Q^\pi(x, a) &= \mathbb{E}[\mathcal{R}(\pi) \mid X_0 = x, A_0 = a] \\ &= \mathbb{E}[R_1(\pi) \mid X_0 = x, A_0 = a] + \gamma \mathbb{E} \left[ \sum_{t=0}^{\infty} \gamma^t R_{t+2}(\pi) \mid X_0 = x, A_0 = a \right] \\ &= \mathbb{E}[R_{(x,a)}] + \gamma \sum_{y \in \mathcal{X}} \mathbb{E} \left[ \sum_{t=0}^{\infty} \gamma^t R_{t+2}(\pi) \mid X_0 = x, A_0 = a, X_1 = y \right] p(y \mid x, a) \\ &\stackrel{\text{Markov}}{=} r(x, a) + \gamma \sum_{y \in \mathcal{X}} \underbrace{\mathbb{E} \left[ \sum_{t=0}^{\infty} \gamma^t R_{t+2}(\pi) \mid X_1 = y, A_1 = \pi(y) \right]}_{= \mathbb{E} \left[ \sum_{t=0}^{\infty} \gamma^t R_{t+2}(\pi) \mid X_1 = y \right]} p(y \mid x, a) \\ &\stackrel{(*)}{=} \mathbb{E} \left[ \sum_{t=0}^{\infty} \gamma^t \tilde{R}_{t+1}(\pi) \mid \tilde{X}_0 = y \right] = V^\pi(y) \end{aligned}$$

(\*) Rename:  $\tilde{X}_t := X_{t+1}, \tilde{A}_t := A_{t+1}, \tilde{R}_t := R_{t+1}$ , then  $(\tilde{X}_t, \tilde{A}_t, \tilde{R}_{t+1}, t \in \mathbb{N}_0)$  is an "evaluation" / "Markov Action Process" of the MDP with the (stationary!) policy  $\pi$ .  $\square$

**Corollary 1.3.5.** *For  $\pi \in \Pi_{\text{stat}}^{\text{det}}$  this holds:*

$$\begin{aligned} V^\pi(x) &= Q^\pi(x, \pi(x)) \\ &= r(x, \pi(x)) + \gamma \sum_{y \in \mathcal{X}} p(y \mid x, \pi(x)) V^\pi(y) \end{aligned}$$

*Proof.* Since  $\pi$  is a deterministic stationary policy:

$$V^\pi(x) = \mathbb{E}[\mathcal{R}(\pi) \mid X_0 = x] = \mathbb{E}[\mathcal{R}(\pi) \mid X_0 = x, A_0 = \pi(x)] = Q^\pi(x, \pi(x))$$

The rest follows from 1.3.4  $\square$

With this relation we can use the Banach fix-point theorem (BFT) for the first time.

**Definition 1.3.6.** For policy  $\pi \in \Pi_{\text{stat}}^{\text{det}}$  is the mapping  $T^\pi: B(\mathcal{X}) \rightarrow B(\mathcal{X})$  with:

$$T^\pi V(x) := r(x, \pi(x)) + \gamma \sum_{y \in \mathcal{X}} p(y | x, \pi(x)) V(y) \quad V \in B(\mathcal{X}), x \in \mathcal{X}$$

called the *Bellman Operator*.

*Remark 1.3.7.*  $\forall \pi \in \Pi_{\text{stat}}^{\text{det}} : T^\pi V^\pi = V^\pi$  (c.f. 1.3.5)

$T^\pi$  meets the requirements of the Banach fixed-point theorem [appendix?](#) for  $\gamma < 1$ , this implies that  $V^\pi$  for  $\pi \in \Pi_{\text{stat}}^{\text{det}}$  is a *unique* fixpoint and can be approximated with the canonical iteration

*Proof.*  $(B(\mathcal{X}), \|\cdot\|_\infty)$  is a non-empty, complete metric space [appendix?](#) and the mapping maps onto itself. It is left to show, that  $T^\pi$  is a contraction. Be  $V, W \in B(\mathcal{X})$ :

$$\begin{aligned} \|T^\pi V - T^\pi W\|_\infty &= \left\| \gamma \sum_{y \in \mathcal{X}} p(y | \cdot, \pi(\cdot)) (V(y) - W(y)) \right\|_\infty \\ &\leq \gamma \sup_{x \in \mathcal{X}} \left\{ \sum_{y \in \mathcal{X}} p(y | x, \pi(x)) \|V - W\|_\infty \right\} \\ &= \gamma \|V - W\|_\infty \sup_{x \in \mathcal{X}} \underbrace{\left\{ \sum_{y \in \mathcal{X}} p(y | x, \pi(x)) \right\}}_{=1} \\ &= \gamma \|V - W\|_\infty \end{aligned}$$

□

*Remark 1.3.8.* Some observations which will come in useful later:

1.  $T^\pi$  is an affine operator
2.  $W_1, W_2 \in B(\mathcal{X})$ , write  $W_1 \leq W_2$  for  $\forall x \in \mathcal{X} : W_1(x) \leq W_2(x)$ , then:

$$W_1 \leq W_2 \implies T^\pi W_1 \leq T^\pi W_2$$

*Proof.* Be  $W_1, W_2 \in B(\mathcal{X})$ ,  $W_1 \leq W_2$  and  $x \in \mathcal{X}$ :

$$T^\pi W_2(x) - T^\pi W_1(x) = \gamma \sum_{y \in \mathcal{X}} p(y | x, \pi(x)) \underbrace{(W_2(y) - W_1(y))}_{\geq 0} \geq 0$$

□

Now we get to the more interesting but also harder optimal value functions. We will later see that taking the supremum over all behaviors is the same as taking it just over the deterministic stationary behaviors. But for now we need to make the distinction:

**Definition 1.3.9.**

$$\tilde{V}(x) := \sup_{\pi \in \Pi_{\text{stat}}^{\text{det}}} V^{\pi}(x)$$

The goal is to show that these two different optimal value functions are actually the same using the uniqueness of the fix-point of the BFT. To achieve that we need to show that they are fix points of the following mapping. And that this mapping satisfies the requirements of the BFT.

**Definition 1.3.10.** The mapping  $T^*: B(\mathcal{X}) \rightarrow B(\mathcal{X})$  with:

$$T^*V(x) := \sup_{a \in \mathcal{A}} \left\{ r(x, a) + \sum_{y \in \mathcal{X}} p(y \mid x, a) V(y) \right\} \quad V \in B(\mathcal{X}), x \in \mathcal{X}$$

is called the *Bellman Optimality Operator*.

**Lemma 1.3.11.**  $\mathcal{M} = (\mathcal{X}, \mathcal{A}, \mathcal{P}_0)$  MDP

- (i)  $\tilde{V}(x) = \sup_{a \in \mathcal{A}} r(x, a) + \gamma \sum_{y \in \mathcal{X}} p(y \mid x, a) \tilde{V}(y)$
- (ii)  $V^*(x) = \sup_{a \in \mathcal{A}} Q^*(x, a)$
- (iii)  $Q^*(x, a) = r(x, a) + \gamma \sum_{y \in \mathcal{X}} p(y \mid x, a) V^*(y)$

*Proof.* (i) By 1.3.5 we know  $V^{\pi}(x) = Q^{\pi}(x, \pi(x))$  thus:

$$\begin{aligned} \tilde{V}(x) &= \sup_{\pi \in \Pi_{\text{stat}}^{\text{det}}} V^{\pi}(x) \\ &= \sup_{\pi \in \Pi_{\text{stat}}^{\text{det}}} \left\{ r(x, \pi(x)) + \gamma \sum_{y \in \mathcal{X}} p(y \mid x, \pi(x)) V^{\pi}(y) \right\} \\ &\stackrel{(*)}{\leq} \sup_{a \in \mathcal{A}} \left\{ r(x, a) + \gamma \sum_{y \in \mathcal{X}} p(y \mid x, a) \sup_{\pi \in \Pi_{\text{stat}}^{\text{det}}} V^{\pi}(y) \right\} \\ &= \sup_{a \in \mathcal{A}} \left\{ r(x, a) + \gamma \sum_{y \in \mathcal{X}} p(y \mid x, a) \tilde{V}(y) \right\} \end{aligned}$$

Assume  $(*)$  is a true inequality for some  $x \in \mathcal{X}$ , since the supremum can be arbitrarily closely approximated:

$$\exists \pi, \exists a : \tilde{V}(x) < r(x, a) + \gamma \sum_{y \in \mathcal{X}} p(y \mid x, a) V^\pi(y)$$

Define a slightly changed deterministic policy with this  $\pi, a$ :

$$\hat{\pi} : \begin{cases} \mathcal{X} \rightarrow \mathcal{A} \\ y \mapsto \begin{cases} \pi(y) & y \neq x \\ a & y = x \end{cases} \end{cases}$$

Define  $W_n := (T^{\hat{\pi}})^n V^\pi$ , then:

$$\begin{aligned} W_1(y) &= T^{\hat{\pi}} V^\pi(y) \stackrel{y \neq x}{=} T^\pi V^\pi(y) = V^\pi(y) \\ &\stackrel{y=x}{=} r(x, \hat{\pi}(x)) + \gamma \sum_{z \in \mathcal{X}} p(z \mid x, \hat{\pi}(x)) V^\pi(z) \\ &= r(x, a) + \gamma \sum_{z \in \mathcal{X}} p(z \mid x, a) V^\pi(z) \\ &> \tilde{V}(x) \geq V^\pi(x) \end{aligned}$$

In either case we get  $W_1(y) \geq V^\pi(y) = W_0(y)$ . By induction with 1.3.8 we get:  $W_{n+1} = T^{\hat{\pi}} W_n \geq T^{\hat{\pi}} W_{n-1} = W_n$ , thus:

$$\begin{aligned} V^{\hat{\pi}}(x) &= \lim_{n \rightarrow \infty} (T^{\hat{\pi}})^n V^\pi(x) = \lim_{n \rightarrow \infty} W_n(x) \geq W_1(x) \\ &= r(x, a) + \gamma \sum_{z \in \mathcal{X}} p(z \mid x, a) V^\pi(z) \\ &> \tilde{V}(x) \quad \text{⚡} \quad \hat{\pi} \in \Pi_{\text{stat}}^{\text{det}} \end{aligned}$$

□

**Corollary 1.3.12.**

$$\begin{aligned} T^* \tilde{V} &= \tilde{V} \\ T^* V^* &= V^* \end{aligned}$$

*Proof.*

$$V^*(x) \stackrel{\text{(ii)}}{=} \sup_{a \in \mathcal{A}} Q^*(x, a) \stackrel{\text{(iii)}}{=} \sup_{a \in \mathcal{A}} \left\{ r(x, a) + \sum_{y \in \mathcal{X}} p(y \mid x, a) V^*(y) \right\} = T^* V^*(x)$$

$\tilde{V}$  analogous

□



**Theorem 1.3.13.**  $\mathcal{M} = (\mathcal{X}, \mathcal{A}, \mathcal{P}_0)$  MDP

$T^*$  satisfies the requirements of the Banach fixpoint theorem, in particular:

$$V^*(x) = \sup_{\pi \in \Pi_{\text{stat}}} V^\pi(x) = \tilde{V}(x)$$

is the unique fixpoint of  $T^*$

**Lemma 1.3.14.** (Blackwell's condition for contraction)

*Proof.* <https://math.stackexchange.com/questions/1087885/blackwells-condition-for-a-contraction-why-is-boundedness-neccessary?rq=1>  $\square$

*Proof (Theorem).*  $\square$

**Proposition 1.3.15.**  $\mathcal{M} = (\mathcal{X}, \mathcal{A}, \mathcal{P}_0)$  MDP

The following statements are equivalent:

- (i)  $\pi \in \Pi_{\text{stat}}$  is optimal ( $V^* = V^\pi$ )
- (ii)  $\forall x \in \mathcal{X} : V^*(x) = \sum_{a \in \mathcal{A}} \pi(a | x) Q^*(x, a)$
- (iii)  $\forall x \in \mathcal{X} : \pi = \arg \max_{\pi \in \Pi_{\text{stat}}} \sum_{a \in \mathcal{A}} \pi(a | x) Q^*(x, a)$
- (iv)  $\pi(a | x) > 0 \iff Q^*(x, a) = V^*(x) = \sup_{b \in \mathcal{A}} Q^*(x, b)$   
 “actions are concentrated on the set of actions that maximize  $Q^*(x, \cdot)$ ”  
 (this also implies:  $Q^*(x, a) < V^*(x) \implies \pi(a | x) = 0$ )

*Proof.*  $\square$

**Definition 1.3.16.**  $Q: \mathcal{X} \times \mathcal{A} \rightarrow \mathbb{R}$  an action value function,  $\tilde{\pi}: \mathcal{X} \rightarrow \mathcal{A}$  with:

$$\tilde{\pi}(x) := \arg \max_{\pi \in \Pi_{\text{stat}}} \sum_{a \in \mathcal{A}} \pi(a | x) Q(x, a) \quad x \in \mathcal{X}$$

$\tilde{\pi}(x)$  is called *greedy* with respect to  $Q$  in  $x \in \mathcal{X}$

$\tilde{\pi}$  is called *greedy* w.r.t.  $Q$

*Remark 1.3.17.*

- 1.3.15(iii) implies that greedy w.r.t.  $Q^*$  is optimal. This means that knowledge of  $Q^*$  is sufficient to select the best action.
- 1.3.11 implies that knowledge of  $V^*, r, p$  is sufficient as well.



## Chapter 2

### Title Chapter 2



## Appendix A

# Appendix

### A.1 Basic Probability Theory

**Lemma A.1.1.**  $\mathbb{P}(A \cap B \mid C) = \mathbb{P}(A \mid B \cap C)\mathbb{P}(B \mid C)$

*Proof.*

$$\mathbb{P}(A \cap B \mid C) = \frac{\mathbb{P}(A \cap B \cap C)}{\mathbb{P}(B \cap C)} \frac{\mathbb{P}(B \cap C)}{\mathbb{P}(C)}$$

□

### A.2 Analysis



# Bibliography

- Puterman, Martin L. (Aug. 28, 2014). *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. John Wiley & Sons. 615 pp. ISBN: 978-1-118-62587-3.
- Szepesvári, Csaba (Jan. 1, 2010). “Algorithms for Reinforcement Learning”. In: *Synthesis Lectures on Artificial Intelligence and Machine Learning* 4.1, pp. 1–103. ISSN: 1939-4608. DOI: 10.2200/S00268ED1V01Y201005AIM009. URL: <https://www.morganclaypool.com/doi/abs/10.2200/S00268ED1V01Y201005AIM009> (visited on 02/06/2019).
- White, Douglas J. (Dec. 1985). “Real Applications of Markov Decision Processes”. In: *Interfaces* 15.6, pp. 73–83. ISSN: 0092-2102, 1526-551X. DOI: 10.1287/inte.15.6.73. URL: <http://pubsonline.informs.org/doi/abs/10.1287/inte.15.6.73> (visited on 02/05/2019).