

# List of Corrections

Make this a definition of some sort? . . . . .	5
remark? lemma? prop? . . . . .	6
some history irrelevant in appendix? . . . . .	8
does this series exist? argument neccessary? c.f. assumption 1 - enough? . . . . .	8
he quotes Blackwell 1965 - quote blackwell directly? . . . . .	9
proof BFT? - appendix? . . . . .	14
show: is complete metric space - appendix? . . . . .	14
comma, fullstop, nothing? . . . . .	14
exist in set? clutters up the equation making it harder to read, but is math. more correct . . . . .	19
ugly - better solution? . . . . .	28
is this Title fix? . . . . .	29
Write down Algorithm? . . . . .	33



UNIVERSITY OF MANNHEIM

Bachelor Thesis

# Reinforcement Learning

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: 11.05.2019



# 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>Introduction</b>	<b>vii</b>
<b>1 Markov Decision Processes</b>	<b>1</b>
1.1 Introduction . . . . .	1
1.2 Model Formulation . . . . .	4
1.3 Fix-Point Equations for Value Functions . . . . .	10
1.4 Optimal Value Functions . . . . .	15
1.5 Optimal policies . . . . .	23
1.6 Dynamic Programming . . . . .	30
<b>2 Stochastic Approximation – Convergence Proofs</b>	<b>35</b>
2.1 Introduction . . . . .	35
<b>A Appendix</b>	<b>37</b>
A.1 Basic Probability Theory . . . . .	37
A.2 Analysis . . . . .	39





# Introduction

Let us consider a system to be a intelligent, if it collects information from its sensors and transforms this information into an “appropriate” response. Then the question where this input comes from (eyes, ears, cameras, etc.), or what kind of input it is (visual, audio, etc.), is as irrelevant for this definition, as the question what a response is (motor activity, sound, etc.). An intelligent system is simply a function mapping inputs to “appropriate” outputs. The question what “appropriate” means, is in contrast a lot more important and difficult to answer. But if we assume that there exists a “correct” response, then this correct response is itself a function from the input space to the output space. Artificial intelligence is therefore simply an artificial implementation of this correct response function.

If we have complete knowledge of this function, we can encode this function as an executable program manually. But more often than not, we do not really understand the rules according to which the inputs are supposed to be transformed into outputs.

For this reason we might want a machine to “learn” these rules by itself, i.e. we want it to approximate the correct response function. There are two larger subcategories in this category of *machine learning*.

In *supervised learning* we do not know the correct response function, but we know the correct response for certain inputs intuitively. An example for this case is image classification of animals. There we do not know how the input (pixels of the picture) map to the output (name of the animal), but we can tell for a given picture what animal is depicted intuitively; i.e. we can provide examples of the correct response function to the learning algorithm. Supervised Learning is therefore concerned with generalization from examples. And while approximating a function with a finite sample of (error prone) function evaluations is a relatively old problem in numerical analysis and statistics, supervised learning is often associated with more recent approaches like artificial neuronal networks.

While we are still able to provide examples in supervised learning, *unsupervised learning* has to work with even less. This category includes clustering and principle component analysis, as well as *reinforcement learning*. Reinforcement learning is applied to problems, where we “know the correct response if we see it”, but can not provide examples ourselves. This is for example the case with

walking. We know how walking looks like, but we have a hard time giving an example for the correct output signals sent to the motors in the leg. It is also used in cases where we do not know the correct response and have never seen it, but are able to compare different response functions. Examples for this range from games like chess to profit maximization of a company.

In general reinforcement learning is used in cases, where humans or the environment in general can rate the solution, and provide rewards for good outcomes. Similar to the conditioning of animals, desirable actions are *reinforced* with rewards.

But in order to write algorithms which maximize their rewards, we first need to formulate the relationship of possibly delayed rewards with actions (outputs) of the learning algorithm in certain states (given certain inputs). The model for this relationship – which virtually all modern reinforcement learning algorithms are based on – is the Markov Decision Process.

We will therefore introduce this model and its properties in the first chapter, continue with a review of common reinforcement algorithms in the second chapter, and explore the relation between the theory of stochastic approximation and reinforcement learning in the third chapter, which yields proofs of convergence to the optimal policy for a number of reinforcement algorithms.

# Chapter 1

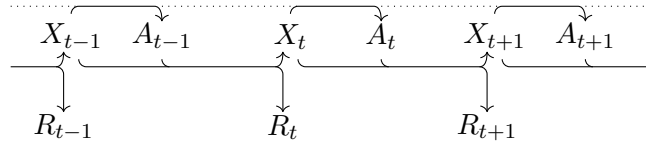
## Markov Decision Processes

### 1.1 Introduction

I will try to convey intuition for Markov Decision Processes in this Introduction. Readers familiar with the subject can skip directly to the formal model formulation.

A Markov Decision Process appears to be quite similar to a Markov Process at first glance. Where a Markov Process is a stochastic process, i.e. a sequence of random variables  $(X_t, t \in \mathbb{N}_0)$ , which is memoryless (Markov). This means knowledge of the current state, makes knowledge about past states useless for predicting future states.

Markov Decision Processes (MDPs) introduce actions  $(A_t, t \in \mathbb{N}_0)$  and rewards  $(R_t, t \in \mathbb{N})$  to this model. Where the transition to the next state  $X_{t+1}$  given a state  $X_t$  and action  $A_t$  is Markov (memoryless), just like in Markov Processes.



But this apparent similarity can be misleading. While Markov Processes are defined as a stochastic process (i.e. a sequence of states) with properties as described above, MDPs cannot be defined like that. The reason for this is, that the next state  $X_{t+1}$  depends on the current state and action  $A_t$ , which means that the sequence of states cannot be defined without the sequence of actions. But defining the sequence of actions requires an action selection rule, a behaviour, a decision. So defining MDPs as a stochastic process, would result in every behaviour resulting in a different MDP. But since MDPs want to model decisions, this does not make much sense. Talking about optimal behaviours in a framework which can only be defined for a set behaviour, is nonsensical. What is needed is a model framework which is invariant to different

behaviours. Therefore MDPs are not defined as a stochastic process, but rather as a rulebook on how to create a stochastic process given a behaviour.

The questions we ask the model are also different. MDPs are more interested into evaluating different behaviours and finding optimal ones, while Markov Processes try to describe an existing phenomenon. Though we will find, that – given a behaviour which is memoryless as well (without the dotted lines in the diagram) – the resulting stochastic process  $(X_t, A_t, R_{t+1}, t \in \mathbb{N}_0)$  is a Markov Process, so the theory of Markov Processes could in principle be applied. But this will not be our focus.

Just like with Markov chains, the memorylessness could be circumvented by including the history in the current state, massively increasing the state space in the process. Which means it is questionable whether this 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.

While the theory of MDPs found an extremely broad application as the backbone for reinforcement algorithms, and by proxy robotics, more classical applications include (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

Lastly, to ease ourselves into the abstract definition of MDPs, let us do one example in more depth. Since examples in reinforcement learning, which make use of stochasticity, quickly become overly complex, we will stay with a classical example. Examples for reinforcement learning will naturally come up in the second chapter.

**Example 1.1.1** (Inventory Management). We will look at a retail store with just one good for simplicity. Let

$\mathcal{X} := \mathbb{N}$  be the set of possible quantities of goods in stock,  
 $\mathcal{A} := \mathbb{N}$  be the set of possible orders for resupply.

We will now introduce the mechanics of actions and rewards without stochastic behaviour and then change that later.

Let us assume that the ordered goods  $a_t \in \mathcal{A}$  arrive in the morning the next day. The goods sold are the demand  $d_t$ , if the current stock  $x_t \in \mathcal{X}$  can meet the demand. So the amount sold is actually  $d_t \wedge x_t = \min\{d, x_t\}$ . Assume orders are paid for in advance and bought at price 1, while the goods are sold at price  $p$ . The cost of storage is  $q$  per item and day. Then the profit at the end of day  $t$  is

$$r_{t+1} = p(d \wedge x_t) - a_t - q(x_t - d \wedge x_t).$$

And the stock level on the next day will be

$$x_{t+1} = x_t - d_t \wedge x_t + a_t.$$

Defining the *state transition function*

$$f(x, a, d) := (x - d \wedge x + a, p(d \wedge x) - a - q(x - d \wedge x)),$$

allows us to express the day to day transition as

$$f(x_t, a_t, d_t) = (x_{t+1}, r_{t+1}).$$

Now assume that the demand is a random variable  $D_t$  which are independent and identically distributed (iid) for all  $t$ . This makes all the other objects (except for the actions) necessarily random variables too. Then distribution of  $(X_{t+1}, R_{t+1})$  conditional on  $X_t = x, A_t = a$  is defined to be the *transition probability kernel*  $\mathcal{P}$  with

$$\mathcal{P}(\cdot \mid x, a) := \mathbb{P}_{f(x, a, D_t)} \stackrel{\text{iid}}{=} \mathbb{P}_{f(x, a, D_0)}.$$

Note that knowledge of the transition kernel  $\mathcal{P}$  is equivalent to knowledge of the transition function  $f$  and the distribution of the demand. Transition kernels reduces the clutter caused by exogenous random variables which are different from application to application and are unknown until the next state is realized at which point their knowledge is useless. We will therefore define this MDP to be  $\mathcal{M} = (\mathcal{X}, \mathcal{A}, \mathcal{P})$ .

A stationary behaviour in this example would be a probability distribution over the possible orders, given the current state, i.e.

$$\pi(\cdot \mid x) \text{ is a probability distribution over } \mathcal{A}.$$

Actions are then selected by picking a random variable from this distribution,

$$A_t \sim \pi(\cdot \mid X_t).$$

Non-stationary behaviours are probability distributions over the action space which can depend on the entire history of states, actions and rewards.

We will just make one more generalization in the actual definition of MDPs. We will allow the action space  $\mathcal{A}$  to depend on the state  $x \in \mathcal{X}$ .

## 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.

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

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

**Definition 1.2.2.**  $\mathcal{M} = (\mathcal{X}, \mathcal{A}, \mathcal{P})$  is called a (*finite*) *Markov Decision Process* (MDP), where

- $\mathcal{X}$  is a countable (finite) set of states,
- $\mathcal{A} = (\mathcal{A}_x)_{x \in \mathcal{X}}$  with  $\mathcal{A}_x$  the countable (finite) set of possible actions in state  $x$ ,
- $\mathcal{P}: (\mathcal{X} \times \mathcal{A}) \times \sigma_{\mathcal{X} \times \mathbb{R}} \rightarrow \mathbb{R}$  is a probability kernel, with the notation

$$\mathcal{X} \times \mathcal{A} := \{(x, a) : x \in \mathcal{X}, a \in \mathcal{A}_x\}. \quad (1.1)$$

- $\mathcal{X} \times \mathbb{R}$  represents the next state and the reward. So  $\mathcal{P}(\cdot \mid x, a)$  represents the probability distribution over the next states and rewards given an action  $a$  in the state  $x$ .
- $\mathcal{P}$  is called the *transition probability kernel*, or in short transition kernel.

*Remark 1.2.3.* Most authors split the transition kernel into a state transition kernel and a reward kernel (e.g. Puterman 2005). 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).

We will spare ourselves the complications of changing transition kernels over time. Just like the memorylessness this could be circumvented by including

time in the state. But if an algorithm is to learn which actions are optimal in which state, then both changing transition kernels over time and state spaces where you never visit any state twice (since the time is included) make it impossible to learn “the rules of the game” without assumptions on how the rules change over time. If such assumptions can be made it is usually easier to bake them into the state space. For example if you would like to introduce changing demand distributions over time in our inventory management example (1.1.1), you could add the current expected demand to the state and the current market penetration. Then changes in the demand distribution can be modelled with a stationary transition kernel.

Some authors prefer to use “cost” over “rewards”. This simply flips the notation and maximization problems become minimization problems. I prefer rewards, as negative rewards have an easier interpretation. According to Puterman (2005) 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.

Nevertheless we still need to construct that 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.

Second, we need an action selection rule

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

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

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

The following special cases can be viewed as policy subsets with some abuse of notation:

1. *Markov policies*  $\pi = (\pi_t, t \in \mathbb{N}_0)$  are memoryless policies, i.e.

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

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

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

---

<sup>1</sup>note that  $\mathcal{X} \times \sigma_{\mathcal{A}} := \{(x, A) : x \in \mathcal{X}, A \in \sigma_{\mathcal{A}_x}\}$  is defined just like  $\mathcal{X} \times \mathcal{A}$  (c.f. (1.1))

FiXme: Make this a definition of some sort?

2. *Stationary policies* are policies which do not change over time i.e.  $\pi_{t+1} = \pi_t$ . Since  $\pi_0$  can only depend on the first state, they are naturally Markov and can be defined as

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

3. *Deterministic stationary policies* are specified by<sup>2</sup>

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

Similarly, deterministic Markov policies and deterministic history dependent policies can be defined. The sets of these policies will be denoted like this

	Stationary		Markov		History dependent
Deterministic	$\Pi_S^D$	$\subseteq$	$\Pi_M^D$	$\subseteq$	$\Pi^D$
	$\cap$		$\cap$		$\cap$
Stochastic	$\Pi_S$	$\subseteq$	$\Pi_M$	$\subseteq$	$\Pi$

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

$$\begin{aligned} &\mathbb{P}[(X_{t+1}, R_{t+1}) = (x, r) \mid (X_t, A_t)] \\ &= \mathbb{P}[(X_{t+1}, R_{t+1}) = (x, r) \mid (X_t, A_t), (X_{t-1}, A_{t-1}, R_t), \dots (X_0, A_0, R_1)] \end{aligned} \quad (1.2)$$

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

FiXme: remark? lemma? prop?

*Remark 1.2.5.*  $(X_{t+1}, R_{t+1}) \sim \mathcal{P}(\cdot \mid X_t, A_t)$  with the Markov property, is well defined i.e.: There exists a  $\mathcal{X} \times \mathbb{R}$ -valued random variable  $(X_{t+1}, R_{t+1})$  such that

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

(analogous  $A_t$  well defined).

*Proof.* As we want to use cumulative distribution functions (cdf) for this proof, we first embed the  $\mathcal{X} \times \mathbb{R}$  in  $\mathbb{R}$ . Since  $\mathcal{X}$  is countable there exists an injective measurable mapping  $g: \mathcal{X} \rightarrow \mathbb{N}$ , and because there exist a homeomorphism<sup>3</sup>  $\sigma: \mathbb{R} \rightarrow (0, 1)$ , there exists an injective measurable function

$$f: \begin{cases} \mathcal{X} \times \mathbb{R} \rightarrow \mathbb{R} \\ (x, r) \rightarrow g(x) + \sigma(r). \end{cases}$$

<sup>2</sup> $\pi: \mathcal{X} \rightarrow \mathcal{A}_x$  is notation for  $\pi: \mathcal{X} \rightarrow \bigcup_{x \in \mathcal{X}} \mathcal{A}_x : \pi(x) \in \mathcal{A}_x$ .

<sup>3</sup>continuous, bijective, inverse continuous – but we only really need measurable.



This allows us to define a probability kernel on the real numbers

$$\tilde{\mathcal{P}}(\cdot \mid x, a) := \mathbb{P}_{f \circ Y} \quad \text{where} \quad Y \sim \mathcal{P}(\cdot \mid x, a),$$

and the corresponding cdf

$$F_{x,a}(y) := \tilde{\mathcal{P}}((-\infty, y] \mid x, a).$$

Now we select  $U \sim \mathcal{U}(0, 1)$  independent from the entire history, then

$$F_{x,a}^{\leftarrow}(U) \sim \tilde{\mathcal{P}}(\cdot \mid x, a) \xrightarrow{\text{A.1.2}} f^{-1} \circ F_{x,a}^{\leftarrow}(U) \sim \mathcal{P}(\cdot \mid x, a),$$

where  $F^{\leftarrow}$  is the pseudo-inverse (A.1.3). Thus

$$(X_{t+1}, R_{t+1}) := f^{-1} \circ F_{X_t, A_t}^{\leftarrow}(U) \sim \mathcal{P}(\cdot \mid X_t, A_t)$$

fulfils the first requirement and is also Markov. To show the Markov property we first define a shorthand for the history

$$\mathcal{H}_s^t := \{(X_t, A_t, R_{t+1}) \in H_t, \dots, (X_s, A_s, R_{s+1}) \in H_s\},$$

where  $H_i$  is defined as

$$H_i := \{(x_i, a_i)\} \times U_i \in \sigma_{\mathcal{X} \times \mathcal{A} \times \mathbb{R}}.$$

With this notation we can now finish the proof:

$$\begin{aligned} \mathbb{P}(f^{-1} \circ F_{X_t, A_t}^{\leftarrow}(U) \in A \mid \mathcal{H}_0^t) &= \frac{\mathbb{P}(f^{-1} \circ F_{x_t, a_t}^{\leftarrow}(U) \in A, \mathcal{H}_0^t)}{\mathbb{P}(\mathcal{H}_0^t)} \\ &\stackrel{U \text{ indep.}}{=} \mathbb{P}(f^{-1} \circ F_{x_t, a_t}^{\leftarrow}(U) \in A) \\ &= \dots = \mathbb{P}(f^{-1} \circ F_{X_t, A_t}^{\leftarrow}(U) \in A \mid \mathcal{H}_t^t) \end{aligned}$$

This proof relies heavily on our countable state set, more general cases would warrant the use of Kolmogorov's extension theorem.  $\square$

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

*Proof.* Using the shorthand for the history defined in the last proof together with

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

we can show the Markov property of the resulting chain:

$$\begin{aligned}
& \mathbb{P}[(X_t, A_t, R_{t+1}) \in \{(x, a)\} \times U \mid \mathcal{H}_0^{t-1}] \\
&= \mathbb{P}[R_{t+1} \in U \mid (X_t, A_t) = (x, a), \mathcal{H}_0^{t-1}] \mathbb{P}[(X_t, A_t) = (x, a) \mid \mathcal{H}_0^{t-1}] \\
&\stackrel{(1.2)}{=} \mathbb{P}[R_{t+1} \in U \mid (X_t, A_t) = (x, a)] \underbrace{\mathbb{P}[A_t = a \mid X_t = x]}_{=\pi_t(a|x)} \mathbb{P}[X_t = x \mid \mathcal{H}_{t-1}^{t-1}] \\
&\stackrel{(*)}{=} \mathbb{P}[R_{t+1} \in U \mid (X_t, A_t) = (x, a), \mathcal{H}_{t-1}^{t-1}] \mathbb{P}[(X_t, A_t) = (x, a) \mid \mathcal{H}_{t-1}^{t-1}] \\
&= \mathbb{P}[(X_t, A_t, R_{t+1}) \in \{(x, a)\} \times U \mid \mathcal{H}_{t-1}^{t-1}]
\end{aligned} \tag{1.3}$$

FiXme: some history  
irrelevant in appendix?

(\*) *Some* of the history is irrelevant if all of the history is irrelevant.

Left to show is, that for stationary policies the Markov chain is time homogeneous. When the policy is stationary, equation (1.3) simplifies to

$$\mathcal{P}(\mathcal{X} \times U \mid x, a) \pi(a \mid x) \mathcal{P}(\{x\} \times \mathbb{R} \mid x_{t-1}, a_{t-1}).$$

So the distribution of  $(X_t, A_t, R_{t+1})$  given  $\mathcal{H}_{t-1}^{t-1}$  is independent of  $t$ . The Markov chain is thus time homogeneous.  $\square$

*Remark 1.2.7.* If there is just one possible action in every state, the MDP is equivalent to a normal Markov Process. Since then there is a mapping  $f: \mathcal{X} \rightarrow \mathcal{A}_x$  mapping the state to the only admissible action. Which implies that  $A_t = f(X_t)$  which forces every behaviour to be equal to  $f$ . Therefore  $f$  is a deterministic stationary behaviour and 1.2.6 applies.

We can mostly ignore Markov policies because our transition kernel is stationary, which makes stationary policies the appropriate set of behaviours. Allowing changing transition probabilities over time, breaks virtually all of the following proofs. In such an environment of changing transition probabilities, Markov policies are the appropriate set to consider (c.f. Puterman 2005).

**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.** Define random variables  $(Y_{(x,a)}, R_{(x,a)}) \sim \mathcal{P}(\cdot \mid x, a)$ , then

$$r(x, a) := \mathbb{E}[R_{(x,a)}] \quad \text{is called } \textit{immediate reward function}.$$

With the return we can define value functions.

FiXme: does this series  
exist? argument  
neccessary? c.f.  
assumption 1 - enough?

**Definition 1.2.10.** Let  $\mathcal{M} = (\mathcal{X}, \mathcal{A}, \mathcal{P})$  be an MDP. Assume an exploring start, i.e. select  $X_0$  such that  $\mathbb{P}(X_0 = x) > 0$  for all states  $x$ . Then every behaviour  $\pi$  results in a stochastic process  $((X_t, A_t, R_{t+1}), t \in \mathbb{N}_0)$ . We indicate with the superscript (e.g.  $\mathbb{E}^\pi$ ) which behaviour generated the process in the following definitions.

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

and  $\pi$  is called *optimal* :  $\Longleftrightarrow V^* = V^\pi$ .

*Remark 1.2.11.* 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 inductively 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 behaviour. This means  $V^\pi, Q^\pi$  are well defined.

*Remark 1.2.12.* While the following proofs do not require discrete state and action spaces per se (as all the sums could just be replaced by integrals), the optimal value functions need not be measurable (Puterman 2005, p. 157) . To remedy this Puterman suggests to either

- impose regularity conditions which would ensure measurability,
- use outer integrals to avoid measurability issues altogether,
- or extend notions of measurability.

FiXme: he quotes Blackwell 1965 - quote blackwell directly?

But since the real world is finite in basically every case anyway we will spare ourselves the trouble.

Some authors include a Time set  $T$  in the tuple  $\mathcal{M} = (\mathcal{X}, \mathcal{A}, \mathcal{P})$  (e.g. Puterman 2005) this allows for finite horizons but not for continuous time, since the transition kernel is defined for discrete steps.

As finite processes are not our focus, we will not do that. But it is possible to achieve finite processes in infinite times with the use of terminal states. The

<sup>1</sup>Well defined because  $\mathbb{P}(X_0 = x) > 0$ . Note that the definition is independent of the distribution of  $X_0$  and  $V^\pi$  can thus be defined as an inherent property of the MDP even when the actual start of the MDP is not exploring.

<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  $\pi_0$

catch is that – unless you include the time in the state – you can not guarantee the end of a process at a fixed time. Although including a finite time set in the state space does not blow it up as much as an infinite time set does. So this solution could possibly be acceptable.

**Definition 1.2.13.** Let  $\mathcal{M} = (\mathcal{X}, \mathcal{A}, \mathcal{P})$  be an MDP

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

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

*Remark 1.2.14.* The reward in a terminal state is zero by convention, i.e.  $x$  terminal state implies for all actions  $a \in \mathcal{A}_x$ , that  $R_{(x,a)} = 0$ .

### 1.2.1 Outlook

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

In the following sections we will try to show that deterministic stationary policies are generally a large enough set of policies to choose from when searching for optimal or close to optimal policies. To be exact: We will find, that the supremum over all policies of value functions is the same as the supremum over just the deterministic stationary policies. The other types of policies will get sandwiched in between.

We will show this fact by proving that both solve the same fixed point equation. And that this fixed point equation satisfies the requirements of the Banach Fix-point Theorem (BFT). Which means we will get a free approximation method of the (optimal) value functions along the way.

## 1.3 Fix-Point Equations for Value Functions

We will start by finding fixpoint equations for  $V^\pi$  and  $Q^\pi$  which fulfil the requirements of the BFT. These can be used to simply calculate the value of one policy. But we will primarily use them as tools to show the fixed point equations for the optimal value functions. One intuitive example for such a use, is that you can try to show that  $V^*$  fulfils the fixpoint equation for  $V^\pi$  and use uniqueness to argue that  $\pi$  has to be optimal. Other cases are a bit more technical and will use a type of monotonicity of the fixed point equation.

To find these fixpoint equations we will try to set  $V^\pi$  and  $Q^\pi$  in relation to each other. For the most part we can focus on deterministic stationary policies, as the other policies will get sandwiched between these policies and the set of all policies.

To make some notation shorter, we will now define the state transition kernel as the marginal probability distribution of the entire transition kernel.

**Definition 1.3.1.**

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

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

Now we can start to explore the relation of  $V^\pi$  and  $Q^\pi$ .

**Proposition 1.3.2.** *Let  $\pi \in \Pi_S$  be a stationary behaviour, then*

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

*Proof.*

$$\begin{aligned} & Q^\pi(x, a) \\ &= \mathbb{E}^\pi[\mathcal{R} \mid X_0 = x, A_0 = a] \\ &= \mathbb{E}^\pi[R_1 \mid X_0 = x, A_0 = a] + \gamma \mathbb{E}^\pi \left[ \sum_{t=0}^{\infty} \gamma^t R_{t+2} \mid X_0 = x, A_0 = a \right] \\ &\stackrel{\text{A.1.1}}{=} \underbrace{\mathbb{E}[R_{(x,a)}]}_{r(x,a)} + \gamma \sum_{y \in \mathcal{X}} \mathbb{E}^\pi \left[ \sum_{t=0}^{\infty} \gamma^t R_{t+2} \mid X_0 = x, A_0 = a, X_1 = y \right] p(y \mid x, a) \end{aligned}$$

This has almost the desired form. Just looking at the conditional expectation we get

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

(\*) by renaming

$$\tilde{X}_t := X_{t+1}, \quad \tilde{A}_t := A_{t+1}, \quad \tilde{R}_t := R_{t+1}.$$

Since then  $(\tilde{X}_t, \tilde{A}_t, \tilde{R}_{t+1}, t \in \mathbb{N}_0)$  is another sequence generated by the MDP with the (stationary!) policy  $\pi$ .  $\square$

**Corollary 1.3.3.** *For  $\pi \in \Pi_S$ , this fixed point equation holds:*

$$V^\pi(x) = \mathbb{E}^\pi[r(x, A_0) \mid X_0 = x] + \gamma \sum_{y \in \mathcal{X}} \mathbb{P}^\pi(X_1 = y \mid X_0 = x) V^\pi(y)$$

For  $\pi \in \Pi_S^D$  the fixed point equation

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

and this fixed point equation hold:

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

*Proof.* Be  $\pi \in \Pi_S$ , then:

$$\begin{aligned} V^\pi(x) &= \mathbb{E}^\pi[\mathcal{R} \mid X_0 = x] \\ &= \sum_{a \in \mathcal{A}_x} \underbrace{\mathbb{E}^\pi[\mathcal{R} \mid X_0 = x, A_0 = a]}_{=Q^\pi(x, a)} \pi(a \mid x) \end{aligned} \tag{1.4}$$

$$\begin{aligned} &= \sum_{a \in \mathcal{A}_x} \left( r(x, a) + \gamma \sum_{y \in \mathcal{X}} p(y \mid x, a) V^\pi(y) \right) \pi(a \mid x) \\ &= \sum_{a \in \mathcal{A}_x} r(x, a) \pi(a \mid x) + \gamma \sum_{(y, a) \in \mathcal{X} \times \mathcal{A}_x} V^\pi(y) p(y \mid x, a) \pi(a \mid x) \\ &= \mathbb{E}^\pi[r(x, A_0) \mid X_0 = x] + \gamma \sum_{(y, a) \in \mathcal{X} \times \mathcal{A}_x} V^\pi(y) \mathbb{P}^\pi(X_1 = y, A_0 = a \mid X_0 = x) \\ &= \mathbb{E}^\pi[r(x, A_0) \mid X_0 = x] + \gamma \sum_{y \in \mathcal{X}} V^\pi(y) \mathbb{P}^\pi(X_1 = y \mid X_0 = x) \end{aligned} \tag{1.5}$$

And for the case, where  $\pi$  is a deterministic stationary policy:

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

The rest follows from 1.3.2.

Alternatively: the  $V^\pi$  fixed point equation is a special case of the equation above. One just needs to realize that  $r(x, \pi(x)) = \mathbb{E}^\pi[r(x, A_0) \mid X_0 = x]$  and

$$\begin{aligned} &\mathbb{P}^\pi(X_1 = y \mid X_0 = x) \\ &= \sum_{a \in \mathcal{A}_x} \mathbb{P}^\pi(X_1 = y \mid X_0 = x, A_0 = a) \underbrace{\mathbb{P}^\pi(A_0 = a \mid X_0 = x)}_{=\delta_{a\pi(x)}} \\ &= \mathbb{P}^\pi(X_1 = y \mid X_0 = x, A_0 = \pi(x)) \\ &= p(y \mid x, \pi(x)) \end{aligned} \tag{1.6}$$

$\square$

With this relation we can use the Banach fixed point theorem (BFT) for the first time. But to do that we first need to make an assumption.

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

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 it implies

$$\mathbb{E}^\pi[|\mathcal{R}| \mid X_0 = x] \leq \mathbb{E}^\pi \left[ \sum_{t=0}^{\infty} \gamma^t |R_{t+1}| \mid X_0 = x \right] \leq \frac{R}{1-\gamma},$$

which implies

$$\|V^\pi\|_\infty, \|V^*\|_\infty, \|Q^\pi\|_\infty, \|Q^*\|_\infty \leq R/(1-\gamma).$$

In particular they are bounded functions. We will denote the set of bounded function on  $M$  with

$$B(M) := \{f : M \rightarrow \mathbb{R} : \|f\|_\infty < \infty\},$$

which happens to be a complete metric space which we will need for the Banach fixed point Theorem to work. Bounded Value functions will also be necessary for a lot of other arguments later on, since the discount factor allows us to discard rewards after a finite time period as negligible. In finite MDPs this assumption is of course always fulfilled.

**Definition 1.3.4.** For a policy  $\pi \in \Pi_S$ , the mapping  $T^\pi : B(\mathcal{X}) \rightarrow B(\mathcal{X})$  with

$$T^\pi V(x) := \mathbb{E}^\pi[r(x, A_0) \mid X_0 = x] + \gamma \sum_{y \in \mathcal{X}} \mathbb{P}^\pi(X_1 = y \mid X_0 = x) V(y)$$

for  $V \in B(\mathcal{X})$  and  $x \in \mathcal{X}$  is called the *Bellman Operator*.

For the special case of  $\pi \in \Pi_S^D$  this mapping can be written as

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

With some abuse of notation we can define the Bellman operator on action value functions  $T^\pi : B(\mathcal{X} \times \mathcal{A}) \rightarrow B(\mathcal{X} \times \mathcal{A})$  for  $\pi \in \Pi_S^D$  with

$$T^\pi Q(x, a) := r(x, a) + \gamma \sum_{y \in \mathcal{X}} p(y \mid x, a) Q(y, \pi(y)) \quad Q \in B(\mathcal{X} \times \mathcal{A}).$$

As mentioned previously, we can mostly ignore stochastic stationary policies as their optimal value functions will get sandwiched between deterministic and general policies. But we will need the slightly more general version of the operator for the value functions  $V^\pi$  for a proof about optimal policies later on. In contrast we only need the special case for the operator on action value functions  $Q^\pi$ .

*Remark 1.3.5.* For all stationary policies  $T^\pi V^\pi = V^\pi$  holds and for deterministic stationary policies  $T^\pi Q^\pi = Q^\pi$  holds (c.f. 1.3.3).

FiXme: proof BFT? -  
appendix?

$T^\pi$  meets the requirements of the Banach fixed-point theorem for  $\gamma < 1$ , this implies that the fixpoints above are *unique* fixpoints and can be approximated with the canonical iteration.

FiXme: show: is complete  
metric space - appendix?

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

$$\begin{aligned} \|T^\pi V - T^\pi W\|_\infty &= \left\| \gamma \sum_{y \in \mathcal{X}} \mathbb{P}^\pi(X_1 = y \mid X_0 = \cdot) (V(y) - W(y)) \right\|_\infty \\ &\leq \gamma \sup_{x \in \mathcal{X}} \left\{ \sum_{y \in \mathcal{X}} \mathbb{P}^\pi(X_1 = y \mid X_0 = x) \|V - W\|_\infty \right\} \\ &= \gamma \|V - W\|_\infty \sup_{x \in \mathcal{X}} \underbrace{\left\{ \sum_{y \in \mathcal{X}} \mathbb{P}^\pi(X_1 = y \mid X_0 = x) \right\}}_{=1} \\ &= \gamma \|V - W\|_\infty \end{aligned}$$

The proof for  $T^\pi: B(\mathcal{X} \times \mathcal{A}) \rightarrow B(\mathcal{X} \times \mathcal{A})$  is analogous. □

**Lemma 1.3.6.** *Observe that*

1.  $T^\pi$  is an affine operator.
2. For  $W_1, W_2 \in B(\mathcal{X})$  write

$$W_1 \leq W_2 : \iff W_1(x) \leq W_2(x) \quad \forall x \in \mathcal{X}$$

*Then the operator  $T^\pi$  is monotonous,*

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

FiXme: comma, fullstop,  
nothing?

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

$$T^\pi W_2(x) - T^\pi W_1(x) = \gamma \sum_{y \in \mathcal{X}} \mathbb{P}^\pi(X_1 = y \mid X_0 = x) \underbrace{(W_2(y) - W_1(y))}_{\geq 0} \geq 0 \quad \square$$



## 1.4 Optimal Value Functions

Until we have shown that the supremum over the small subset of deterministic stationary behaviors is the same as over all behaviors, we need to make the notational distinction between different optimal value functions.

**Definition 1.4.1.**

$$\begin{aligned}\tilde{V}(x) &:= \sup_{\pi \in \Pi_S^D} V^\pi(x) \\ \tilde{Q}(x, a) &:= \sup_{\pi \in \Pi_S^D} Q^\pi(x, a)\end{aligned}$$

As previously stated: The goal is to show that these two pairs of optimal value functions are actually the same using the uniqueness of the fixed point of the BFT.

The intuition for why this should be the case comes from the fact that the MDP is memory-less. So winding up in the same state results in the agent having the exact same decision problem. And if the agent decided for one optimal action in the past, then this action will again be optimal. For this reason the supremum over all behaviors should be equal to the supremum over stationary behaviors.

And if an optimal policy randomizes over different actions, then the values of these actions must all be equally high which means that the set of deterministic policies is large enough. Since just picking one and sticking with it should be just as good.

### 1.4.1 Finite Outmatching

Before we get to tackle this problem, we first need to address another one. Notice how we take the supremum for every state  $x$  individually?

$$\tilde{V}(x) = \sup_{\pi \in \Pi_S^D} V^\pi(x)$$

Since the stationary policy still allows us to condition on the state, it is intuitive to assume that we do not need different sequences of policies to approximate  $V^*$  for each state  $x \in \mathcal{X}$ . This will eventually turn out to be true using statements about the optimal value functions. We would not be successful to show this directly from the definition. As sequences approximating the different suprema are indexed by the (possibly countable) state space, what we would need to find is a single sequence of policies, which matches all the policies in this countable set of sequences in every step of the sequence. This turns out to be an impossible task. Here is an example for that.

**Example 1.4.2** (The genie cubicles). Imagine an infinite (countable) amount of cubicles  $\mathbb{N} \subset \mathcal{X}$ , with a genie in every cubicle. You can wish for an arbitrary

amount of money  $\mathcal{A} = \mathbb{N}$ , after that you have to leave (end up in the terminal state 0, i.e.:  $\mathcal{X} = \mathbb{N}_0$ ). Then of course  $V^*(x) = \infty$  for  $x \in \mathbb{N}$ , is achieved by the sequences of behaviors

$$(\pi_x^{(n)}, n \in \mathbb{N}) \quad \text{with } \pi_x^{(n)}(y) = x + n \quad \forall y \in \mathcal{X}$$

Then there is no policy  $\pi^{(n)}$  which can match every  $\pi_x^{(n)}$  for all  $x \in \mathbb{N}$ .

Even if  $\tilde{V}$  was finite, we could modify the example by cutting gifts off above a certain threshold with behaviors approaching that threshold.

So we can not match an infinite set of behaviors. Which means we will have to settle for a finite version right now. This version will help us to handle the optimal value functions. But with the later facts about optimal value functions we can then define a sequence of policies which approach the optimal value functions uniformly confirming our earlier suspicions, that the suprema can be attained with a single sequence of policies.

**Proposition 1.4.3.** *Be  $n \in \mathbb{N}$  and  $\{\pi_1, \dots, \pi_n\} \subseteq \Pi_S^D$ , then:*

$$\exists \hat{\pi} \in \Pi_S^D : \quad \max_{i=1, \dots, n} V^{\pi_i}(x) \leq V^{\hat{\pi}}(x) \quad \forall x \in \mathcal{X}$$

*Proof.* The idea is to pick the same action in state  $x$ , as the policy which generates the most value out of this state, i.e.  $\max_{i=1, \dots, n} V^{\pi_i}(x)$ . So define

$$\hat{\pi} : \begin{cases} \mathcal{X} \rightarrow \mathcal{A} \\ x \mapsto \pi_{\arg\max_{i=1, \dots, n} V^{\pi_i}(x)}(x). \end{cases}$$

One might be surprised that such an appearingly short sighted policy should surpass every policy in the finite set. At first it might seem that different policies achieve their values of their state through different, maybe incompatible strategies. Would a strategy which takes a policy because it changes transition probabilities in a way that leads to a high-payoff state not be sabotaged by switching policies erratically? It is important to realize that if a policy A attaches a high value to a state because it provides easy access to states which can yield a high payoff, then the other policies will either exploit the high payoff later on as well, or the policy A will once again be the maximum. Either way it would result in  $\hat{\pi}$  exploiting the high payoff later on.

For the maximum value we write

$$V(x) := \max_{i=1, \dots, n} V^{\pi_i}(x).$$

And be

$$m(x) := \arg \max_{i=1, \dots, n} V^{\pi_i}(x),$$

then the modified policy  $\hat{\pi}$  improves  $V$  in all states:

$$\begin{aligned} T^{\hat{\pi}}V(x) &= r(x, \pi_{m(x)}(x)) + \gamma \sum_{y \in \mathcal{X}} p(y \mid x, \pi_{m(x)}(x))V(y) \\ V &\geq V^{\pi_{m(x)}} \\ &\geq r(x, \pi_{m(x)}(x)) + \gamma \sum_{y \in \mathcal{X}} p(y \mid x, \pi_{m(x)}(x))V^{\pi_{m(x)}}(y) \\ &\stackrel{1.3.3}{=} V^{\pi_{m(x)}}(x) = V(x) \end{aligned}$$

By using the monotonicity of  $T^{\hat{\pi}}$  (1.3.6) inductively with  $(T^{\hat{\pi}})^1V \geq (T^{\hat{\pi}})^0V$ , we get

$$(T^{\hat{\pi}})^nV \geq (T^{\hat{\pi}})^{n-1}V,$$

and thus finish the proof with:

$$V^{\hat{\pi}} = \lim_{n \rightarrow \infty} (T^{\hat{\pi}})^nV \geq V \geq V^{\pi_i} \quad \forall i = 1, \dots, n \quad \square$$

### 1.4.2 Fix-Point Equations for Optimal Value functions

With this statement we can now prove the building blocks for the fixed point equations.

**Lemma 1.4.4.**

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

*Proof.* (i) The smaller or equal part is easy:

$$\begin{aligned} \tilde{Q}(x, a) &= \sup_{\pi \in \Pi_S^D} \left\{ r(x, a) + \gamma \sum_{y \in \mathcal{X}} p(y \mid x, a)V^{\pi}(y) \right\} \\ &\leq r(x, a) + \gamma \sum_{y \in \mathcal{X}} p(y \mid x, a) \underbrace{\sup_{\pi \in \Pi_S^D} V^{\pi}(y)}_{=\tilde{V}(y)} \end{aligned}$$

For the other direction we need to do a bit more work. Since the  $r(x, a)$  and  $\gamma$ , are unaffected by the supremum what is left to show is:

$$\sup_{\pi \in \Pi_S^D} \sum_{y \in \mathcal{X}} p(y \mid x, a)V^{\pi}(y) \geq \sum_{y \in \mathcal{X}} p(y \mid x, a)\tilde{V}(y)$$

This problem should look familiar. It is the question whether there is a single sequence of policies that can match multiple sequences of policies indexed by the state space  $y \in \mathcal{X}$ . As we can find a policy which can outmatch a finite set of policies (1.4.3), we will consider only the  $y \in \mathcal{X}$  with the largest probability to occur and match these sequences with a single sequence. Since we then have just a single policy in the sum, we can estimate up by taking the outer supremum over deterministic policies. That is the idea, which can be executed as follows:

The set  $M_\delta := \{y \in \mathcal{X} : p(y \mid x, a) > \delta\}$  is finite for all  $\delta > 0$ . Together with the continuity of measures this yields:

$$1 = p\left(\bigcup_{\delta \rightarrow 0} M_\delta \mid x, a\right) = \lim_{\delta \rightarrow 0} p(M_\delta \mid x, a) = \lim_{\delta \rightarrow 0} \sum_{y \in M_\delta} p(y \mid x, a)$$

Therefore for all  $\varepsilon > 0$  exists a  $\delta > 0$  such that

$$\sum_{y \in M_\delta^c} p(y \mid x, a) < \frac{\varepsilon/4}{R/(1-\gamma)}. \quad (1.6)$$

Be  $(\pi_y^{(n)}, n \in \mathbb{N})$  with  $V^{\pi_y^{(n)}}(y) \nearrow \tilde{V}(y) \quad (n \rightarrow \infty)$ . Since  $M_\delta$  is finite, there exists  $N \in \mathbb{N}$  such that

$$|\tilde{V}(y) - V^{\pi_y^{(n)}}(y)| < \varepsilon/2 \quad \forall n \geq N, \forall y \in M_\delta^c. \quad (1.7)$$

And also because  $M_\delta$  is finite, and 1.4.3 we know that

$$\exists \hat{\pi}^{(n)} \in \Pi_S^D : \quad V^{\hat{\pi}^{(n)}} \geq V^{\pi_y^{(n)}} \quad \forall y \in M_\delta. \quad (1.8)$$

This finally implies

$$\begin{aligned} & \sum_{y \in \mathcal{X}} p(y \mid x, a) \tilde{V}(y) - \sum_{y \in \mathcal{X}} p(y \mid x, a) V^{\hat{\pi}^{(n)}}(y) \\ & \leq \sum_{y \in M_\delta} p(y \mid x, a) (\tilde{V}(y) - V^{\hat{\pi}^{(n)}}(y)) + \sum_{y \in M_\delta^c} p(y \mid x, a) \underbrace{|\tilde{V}(y) - V^{\hat{\pi}^{(n)}}(y)|}_{\leq \|\tilde{V}\|_\infty + \|V^{\hat{\pi}^{(n)}}\|_\infty} \\ & \hspace{15em} \leq 2R/(1-\gamma) \\ & \stackrel{(1.8)}{\leq} \sum_{y \in M_\delta} p(y \mid x, a) \underbrace{(\tilde{V}(y) - V^{\pi_y^{(n)}}(y))}_{< \varepsilon/2 \quad (1.7)} + 2R/(1-\gamma) \sum_{y \in M_\delta^c} p(y \mid x, a) \\ & \hspace{15em} \underbrace{< \frac{\varepsilon/4}{R/(1-\gamma)}}_{(1.6)} \\ & \leq \varepsilon \quad \forall n \geq N. \end{aligned}$$

This results in

$$\begin{aligned} \sum_{y \in \mathcal{X}} p(y \mid x, a) \tilde{V}(y) &\leq \varepsilon + \sum_{y \in \mathcal{X}} p(y \mid x, a) V^{\hat{\pi}^{(n)}}(y) \\ &\leq \varepsilon + \sup_{\pi \in \Pi_S^D} \sum_{y \in \mathcal{X}} p(y \mid x, a) V^\pi(y). \end{aligned}$$

Since this equation holds for all  $\varepsilon > 0$  we are finished.

(ii) By 1.3.3 we know  $V^\pi(x) = Q^\pi(x, \pi(x))$  and thus:

$$\begin{aligned} \tilde{V}(x) &= \sup_{\pi \in \Pi_S^D} V^\pi(x) = \sup_{\pi \in \Pi_S^D} Q^\pi(x, \pi(x)) \\ &\leq \sup_{a \in \mathcal{A}_x} \sup_{\pi \in \Pi_S^D} Q^\pi(x, a) = \sup_{a \in \mathcal{A}_x} \tilde{Q}(x, a) \end{aligned} \quad (1.9)$$

$$\stackrel{(i)}{=} \sup_{a \in \mathcal{A}_x} \left\{ r(x, a) + \gamma \sum_{y \in \mathcal{X}} p(y \mid x, a) \sup_{\pi \in \Pi_S^D} V^\pi(y) \right\} \quad (1.10)$$

Assume (1.9) is a true inequality for some  $x \in \mathcal{X}$ . Since the suprema in (1.10) can be arbitrarily closely approximated we get

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

FiXme: exist in set?  
clutters up the equation  
making it harder to read,  
but is math. more correct

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

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

Define  $W_n := (T^{\hat{\pi}})^n V^\pi$ , then a case by case analysis results in:

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

So in either case we get

$$W_1(y) \geq V^\pi(y) = W_0(y).$$

By induction using the monotonicity of  $T^\pi$  (1.3.6) we get

$$W_{n+1} = T^{\hat{\pi}} W_n \geq T^{\hat{\pi}} W_{n-1} = W_n,$$

which leads to a contraction:

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

(iii) When taking the supremum over two variables the order does not matter, therefore:

$$\begin{aligned}
\sup_{a \in \mathcal{A}_x} Q^*(x, a) &= \sup_{a \in \mathcal{A}_x} \sup_{\pi \in \Pi} \mathbb{E}^{\pi}[\mathcal{R} \mid X_0 = x, A_0 = a] \\
&= \sup_{(\pi_t, t \in \mathbb{N}_0)} \sup_{a \in \mathcal{A}_x} \mathbb{E}^{\pi}[\mathcal{R} \mid X_0 = x, A_0 = a] \\
&\stackrel{A_0=a}{=} \sup_{(\pi_t, t \in \mathbb{N})} \sup_{a \in \mathcal{A}_x} \mathbb{E}^{\pi}[\mathcal{R} \mid X_0 = x, A_0 = a] \\
&\stackrel{(*)}{=} \sup_{(\pi_t, t \in \mathbb{N})} \sup_{\pi_0} \mathbb{E}^{\pi}[\mathcal{R} \mid X_0 = x] \\
&= V^*(x)
\end{aligned} \tag{1.11}$$

(\*) “ $\leq$ ” is trivial, since a deterministic  $\pi_0$  can achieve the same as fixing an action  $A_0 = a$ .

“ $\geq$ ” follows from this inequality which holds for all  $\pi$  (especially for all  $\pi_0$ ):

$$\begin{aligned}
\mathbb{E}^{\pi}[\mathcal{R} \mid X_0 = x] &\stackrel{A.1.1}{=} \sum_{a \in \mathcal{A}_x} \mathbb{E}^{\pi}[\mathcal{R} \mid X_0 = x, A_0 = a] \pi_0(a \mid x) \\
&\leq \sup_{a \in \mathcal{A}_x} \mathbb{E}^{\pi}[\mathcal{R} \mid X_0 = x, A_0 = a] \sum_{a \in \mathcal{A}_x} \pi_0(a \mid x) \\
&= \sup_{a \in \mathcal{A}_x} \mathbb{E}^{\pi}[\mathcal{R} \mid X_0 = x, A_0 = a]
\end{aligned}$$

(iv) For an arbitrary policy  $\pi$  we can expand the action value function

$$\begin{aligned}
Q^{\pi}(x, a) &= \mathbb{E}^{\pi}[R_1 \mid X_0 = x, A_0 = a] + \mathbb{E} \left[ \sum_{t=1}^{\infty} \gamma^t R_{t+1} \mid X_0 = x, A_0 = a \right] \\
&= r(x, a) + \gamma \mathbb{E} \left[ \sum_{t=0}^{\infty} \gamma^t R_{t+2} \mid X_0 = x, A_0 = a \right].
\end{aligned} \tag{1.12}$$

This implies

$$\begin{aligned}
Q^*(x, a) &= \sup_{\pi \in \Pi} Q^\pi(x, a) \\
&\stackrel{(1.12)}{=} r(x, a) + \gamma \sup_{(\pi_t, t \in \mathbb{N})} \mathbb{E}^\pi \left[ \sum_{t=0}^{\infty} \gamma^t R_{t+2} \mid X_0 = x, A_0 = a \right] \\
&\stackrel{A.1.1}{=} r(x, a) + \gamma \sup_{(\pi_t, t \in \mathbb{N})} \sum_{y \in \mathcal{X}} p(y \mid x, a) \mathbb{E}^\pi \left[ \sum_{t=0}^{\infty} \gamma^t R_{t+2} \mid X_0 = x, A_0 = a, X_1 = y \right] \\
&= r(x, a) + \gamma \sum_{y \in \mathcal{X}} p(y \mid x, a) \underbrace{\sup_{(\pi_t, t \in \mathbb{N})} \mathbb{E}^\pi \left[ \sum_{t=0}^{\infty} \gamma^t R_{t+2} \mid X_0 = x, A_0 = a, X_1 = y \right]}_{(*)}.
\end{aligned}$$

The last equality follows from the fact, that the behaviours can condition on the entire history. In particular it can condition on  $X_1 = y$  which means that moving the supremum into the sum does not actually open any room for improvement.

What is left to show, is that  $(*)$  equals  $V^*(y)$ . To show this we use the same trick as in (1.11) twice

$$\begin{aligned}
(*) &\stackrel{(1.11)}{=} \sup_{(\pi_t, t \geq 2)} \sup_{a_1 \in \mathcal{A}_y} \mathbb{E}^\pi \left[ \sum_{t=0}^{\infty} \gamma^t R_{t+2} \mid X_0 = x, A_0 = a, X_1 = y, A_1 = a_1 \right] \\
&\stackrel{\text{Markov}}{=} \sup_{(\pi_t, t \geq 2)} \sup_{a_1 \in \mathcal{A}_y} \mathbb{E}^\pi \left[ \sum_{t=0}^{\infty} \gamma^t R_{t+2} \mid X_1 = y, A_1 = a_1 \right] \\
&\stackrel{(1.11)}{=} \sup_{(\pi_t, t \geq 2)} \sup_{\pi_1} \mathbb{E}^\pi \left[ \sum_{t=0}^{\infty} \gamma^t R_{t+2} \mid X_1 = y \right] \\
&= \sup_{(\tilde{\pi}_t, t \in \mathbb{N}_0)} \mathbb{E}^{\tilde{\pi}} \left[ \sum_{t=0}^{\infty} \gamma^t \tilde{R}_{t+1} \mid \tilde{X}_0 = y \right] = V^*(y)
\end{aligned}$$

Where  $(\tilde{X}_t, \tilde{A}_t, \tilde{R}_{t+1}, t \in \mathbb{N}_0)$  is also a sequence generated by  $\mathcal{M}$  with

$$\tilde{X}_t := X_{t+1}, \quad \tilde{A}_t := A_{t+1}, \quad \tilde{R}_t := R_{t+1}, \quad \tilde{\pi}_t := \pi_{t+1}. \quad \square$$

With this Lemma it will be easy to show that the optimal value functions are fixed points of the following mappings.

**Definition 1.4.5.** The mapping  $T^*: B(\mathcal{X}) \rightarrow B(\mathcal{X})$  with

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

is called the *Bellman Optimality Operator*. With some more abuse of notation, define the mapping  $T^*: B(\mathcal{X} \times \mathcal{A}) \rightarrow B(\mathcal{X} \times \mathcal{A})$  with

$$T^*Q(x, a) := r(x, a) + \sum_{y \in \mathcal{X}} p(y \mid x, a) \sup_{a' \in \mathcal{A}_y} Q(y, a') \quad V \in B(\mathcal{X}), (x, a) \in \mathcal{X} \times \mathcal{A}.$$

**Corollary 1.4.6.** *All the optimal value functions are fixed points of  $T^*$ , i.e.*

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

*Proof.* We simply assemble the building blocks from the previous lemma 1.4.4:

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

The others are analogous.  $\square$

Now we just need to show that the Bellman Optimality Operator satisfies the Banach Fix-point Theorem.

**Theorem 1.4.7.**  *$T^*$  satisfies the requirements of the Banach fixpoint theorem for  $\gamma < 1$ , in particular*

$$V^*(x) = \tilde{V}(x)$$

*is the unique fixpoint of  $T^*$ . The suprema over all the other policy sets get sandwiched in between.*

*Proof.* (Szepesvári 2010, p. 79) Since  $B(\mathcal{X})$  and  $B(\mathcal{X} \times \mathcal{A})$  are complete metric spaces and the mapping  $T^*$  maps onto itself we only need to show that it is a contraction. For that we first need to show:

$$\left| \sup_{a \in \mathcal{A}_x} f(a) - \sup_{b \in \mathcal{A}_x} g(b) \right| \leq \sup_{a \in \mathcal{A}_x} |f(a) - g(a)| \quad (1.13)$$

To do that, we assume without loss of generality that the absolute value on the left side is itself (otherwise switch  $f$  and  $g$ ):

$$\sup_{a \in \mathcal{A}_x} f(a) - \sup_{b \in \mathcal{A}_x} g(b) \leq \sup_{a \in \mathcal{A}_x} f(a) - g(a) \leq \sup_{a \in \mathcal{A}_x} |f(a) - g(a)|$$

Now let  $V, W \in B(\mathcal{X})$ , then we get a contraction:

$$\begin{aligned} \|T^*V - T^*W\|_\infty &= \sup_{x \in \mathcal{X}} |T^*V(x) - T^*W(x)| \\ &\stackrel{(1.13)}{\leq} \sup_{x \in \mathcal{X}} \sup_{a \in \mathcal{A}_x} \left| \gamma \sum_{y \in \mathcal{X}} p(y \mid x, a) (V(y) - W(y)) \right| \\ &\leq \sup_{x \in \mathcal{X}} \sup_{a \in \mathcal{A}_x} \gamma \underbrace{\sum_{y \in \mathcal{X}} p(y \mid x, a)}_{=1} \|V - W\|_\infty \\ &= \gamma \|V - W\|_\infty \end{aligned}$$



Similarly for  $Q, P \in B(\mathcal{X} \times \mathcal{A})$ :

$$\begin{aligned}
\|T^*Q - T^*P\|_\infty &= \sup_{x \in \mathcal{X}} \left| \gamma \sum_{y \in \mathcal{Y}} p(y | x, a) \left( \sup_{a' \in \mathcal{A}_y} Q(y, a') - \sup_{b' \in \mathcal{A}_y} P(y, b') \right) \right| \\
&\stackrel{(1.13)}{\leq} \sup_{x \in \mathcal{X}} \gamma \sum_{y \in \mathcal{Y}} p(y | x, a) \sup_{a' \in \mathcal{A}_y} |Q(y, a') - P(y, a')| \\
&\leq \gamma \|Q - P\|_\infty
\end{aligned}
\quad \square$$

Alternatively one can show that  $T^*$  fulfills Blackwell's condition for a contraction.

## 1.5 Optimal policies

Now that we proved the uniqueness of the optimal value functions we need to ask the question whether this supremum can be attained with a single policy. And if not, if it can be approximated over all states by the same sequence of policies.

Let us first consider the case where the supremum can be attained, since this includes the important special case of finite MDPs.

**Proposition 1.5.1.** *Be  $\pi^* \in \Pi_S$ , then the following statements are equivalent:*

- (i)  $\pi^* \in \Pi_S$  is optimal ( $V^* = V^{\pi^*}$ )
- (ii)  $\forall x \in \mathcal{X} : V^*(x) = \sum_{a \in \mathcal{A}_x} \pi^*(a | x) Q^*(x, a)$
- (iii)  $\forall x \in \mathcal{X} : \pi^* = \arg \max_{\pi \in \Pi_S} \sum_{a \in \mathcal{A}_x} \pi(a | x) Q^*(x, a)$
- (iv)  $\pi^*(a | x) > 0 \implies Q^*(x, a) = V^*(x) = \sup_{b \in \mathcal{A}_x} Q^*(x, b)$   
*“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.* “(i)  $\Rightarrow$  (ii)” Let  $x \in \mathcal{X}$  be arbitrary, then:

$$\begin{aligned}
V^*(x) &= V^{\pi^*}(x) \stackrel{(1.4)}{=} \sum_{a \in \mathcal{A}_x} \pi^*(a | x) Q^{\pi^*}(x, a) \\
&\leq \sum_{a \in \mathcal{A}_x} \pi^*(a | x) Q^*(x, a) \\
&\leq \underbrace{\sum_{a \in \mathcal{A}_x} \pi^*(a | x)}_{=1} \sup_{b \in \mathcal{A}_x} Q^*(x, b) \\
&\stackrel{1.4.4}{=} V^*(x)
\end{aligned}$$

“(ii)  $\Rightarrow$  (iii)” Let  $\pi \in \Pi_S$ ,  $(x, a) \in \mathcal{X} \times \mathcal{A}$  be arbitrary. Then with (1.4) we have

$$\sum_{a \in \mathcal{A}_x} \pi(a | x) Q^*(x, a) \leq \underbrace{\sum_{a \in \mathcal{A}_x} \pi(a | x)}_{=1} \sup_{b \in \mathcal{A}_x} Q^*(x, b) = V^*(x).$$

Therefore  $V^*(x)$  is an upper bound for every  $\pi \in \Pi_S$ , and since  $\pi^*$  attains this upper bound it is a maximum.

“(iii)  $\Rightarrow$  (iv)” Be  $\pi^*(a | x) > 0$  for some  $a \in \mathcal{A}_x, x \in \mathcal{X}$ . Then there exists no  $b \in \mathcal{A}_x$  with  $Q^*(x, b) > Q^*(x, a)$ . Otherwise we can define the behaviour

$$\hat{\pi}(\cdot | x): \begin{cases} \mathcal{A}_x \rightarrow [0, 1] \\ c \mapsto \begin{cases} 0 & c = a \\ \pi^*(b | x) + \pi^*(a | x) & c = b \\ \pi^*(c | x) & \text{else.} \end{cases} \end{cases}$$

This results in a contradiction:

$$\begin{aligned} & \sum_{c \in \mathcal{A}_x} \hat{\pi}(c | x) Q^*(x, c) \\ &= [\pi^*(b | x) + \underbrace{\pi^*(a | x)}_{>0}] \underbrace{Q^*(x, b)}_{>Q^*(x, a)} + \sum_{c \in \mathcal{A}_x \setminus \{a, b\}} \pi^*(c | x) Q^*(x, c) \\ &> \sum_{c \in \mathcal{A}_x} \pi^*(c | x) Q^*(x, c) \quad \nrightarrow \pi^* \text{ attains maximum} \end{aligned}$$

“(iv)  $\Rightarrow$  (ii)” Be  $x \in \mathcal{X}$ , since  $\pi^*(\cdot | x)$  is a probability distribution on  $\mathcal{A}_x$  there exists  $a \in \mathcal{A}_x$  such that  $\pi^*(a | x) > 0$ . Define

$$M_x := \{a \in \mathcal{A}_x : \pi^*(a | x) > 0\}.$$

Then  $Q^*(x, a) = V^*(x)$  for all  $a \in M_x$  by prerequisite, therefore

$$\begin{aligned} V^*(x) &= \sum_{a \in M_x} \pi^*(a | x) V^*(x) = \sum_{a \in M_x} \pi^*(a | x) Q^*(x, a) \\ &= \sum_{a \in \mathcal{A}_x} \pi^*(a | x) Q^*(x, a). \end{aligned}$$

“(ii)  $\Rightarrow$  (i)” Using (iv) from 1.4.4 for an  $x \in \mathcal{X}$ :

$$\begin{aligned}
 V^*(x) &= \sum_{a \in \mathcal{A}_x} \pi^*(a | x) Q^*(x, a) \\
 &= \sum_{a \in \mathcal{A}_x} \pi^*(a | x) \left( r(x, a) + \gamma \sum_{y \in \mathcal{X}} p(y | x, a) V^*(y) \right) \\
 &= \sum_{a \in \mathcal{A}_x} \pi^*(a | x) r(x, a) + \gamma \sum_{y \in \mathcal{X}} \sum_{a \in \mathcal{A}_x} \pi^*(a | x) p(y | x, a) V^*(y) \\
 &\stackrel{(1.5)}{=} \mathbb{E}[r(x, A_0) | X_0 = x] + \gamma \sum_{y \in \mathcal{X}} \mathbb{P}(X_1 = y | X_0 = x) V^*(y) \\
 &= T^{\pi^*} V^*(x)
 \end{aligned}$$

Therefore  $V^{\pi^*} = V^*$  since the fixed point of  $T^{\pi^*}$  is unique (1.3.5).  $\square$

*Remark 1.5.2.* From (iv) follows: if an optimal stochastic stationary policy exists, then there exists an optimal deterministic stationary policy as well. Since such a policy can be constructed by choosing one of the actions  $a$  for every  $x \in \mathcal{X}$  which result in  $Q^*(x, a) = V^*(x)$ . These actions have to exist if there is a stochastic stationary policy which is optimal.

Before we show, that if an arbitrary (history dependent) optimal policy exists, there also exists a stationary policy which is optimal, I want to note that from (iii) follows a first heuristic to find an optimal value function. This heuristic only works of course, if the maximum exists. This is never a problem in finite MDPs.

**Definition 1.5.3.**  $Q: \mathcal{X} \times \mathcal{A} \rightarrow \mathbb{R}$  an action value function,  $\tilde{\pi}: \mathcal{X} \rightarrow \mathcal{A}_x$  with

$$\tilde{\pi}(x) := \arg \max_{a \in \mathcal{A}_x} Q(x, a) \quad x \in \mathcal{X}$$

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

*Remark 1.5.4.*

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

Now we show how we can construct an optimal stationary policy from an arbitrary optimal policy. We will break this problem into two parts. First we construct an optimal Markovian policy, and then we construct an optimal stationary policy from the optimal Markovian policy. The first step is taken from Puterman (2005, pp. 134–137).

**Proposition 1.5.5** (Puterman). *Let  $\pi \in \Pi$ . Then, for each  $x \in \mathcal{X}$ , there exists a Markovian policy  $\pi^x$  satisfying*

$$\mathbb{P}^{\pi^x}(X_t = y, A_t = a \mid X_0 = x) = \mathbb{P}^\pi(X_t = y, A_t = a \mid X_0 = x) \quad \forall t \in \mathbb{N}_0,$$

where  $\mathbb{P}^\pi$  indicates that the sequence  $((X_t, A_t, R_{t+1}), t \in \mathbb{N}_0)$  is constructed with policy  $\pi$ .

*Proof.* Fix  $x \in \mathcal{X}$ . For each  $(y, a) \in \mathcal{X} \times \mathcal{A}$  define  $\pi^x := (\pi_t^x, t \in \mathbb{N}_0)$  with

$$\pi_t^x(\cdot \mid y) := \mathbb{P}^\pi(A_t \mid X_t = y, X_0 = x). \quad (1.14)$$

We will now show the statement for this  $\pi^x$  by induction. The base case is:

$$\begin{aligned} \mathbb{P}^\pi(X_0 = y, A_0 = a \mid X_0 = x) &= \mathbb{P}^\pi(A_0 = a \mid X_0 = x) \delta_{xy} \\ &= \pi_0^x(a \mid x) \delta_{xy} = \mathbb{P}^{\pi^x}(A_0 = a \mid X_0 = x) \delta_{xy} \\ &= \mathbb{P}^{\pi^x}(X_0 = y, A_0 = a \mid X_0 = x) \end{aligned}$$

Assume the claim holds for  $\{0, \dots, t-1\}$ . Then using Markovian transition kernel

$$\mathbb{P}^\pi[X_t = y \mid (X_{t-1}, A_{t-1}) = (z, a), X_0 = x] = p(y \mid z, a),$$

together with A.1.1, we get:

$$\begin{aligned} \mathbb{P}^\pi(X_t = y \mid X_0 = x) &= \sum_{(z,a) \in \mathcal{X} \times \mathcal{A}} p(y \mid z, a) \mathbb{P}^\pi[(X_{t-1}, A_{t-1}) = (z, a) \mid X_0 = x] \\ &\stackrel{\text{ind.}}{=} \sum_{(z,a) \in \mathcal{X} \times \mathcal{A}} p(y \mid z, a) \mathbb{P}^{\pi^x}[(X_{t-1}, A_{t-1}) = (z, a) \mid X_0 = x] \\ &= \mathbb{P}^{\pi^x}(X_t = y \mid X_0 = x) \end{aligned}$$

Using A.1.1 again we get:

$$\begin{aligned} &\mathbb{P}^\pi[X_t = y, A_t = a \mid X_0 = x] \\ &= \underbrace{\mathbb{P}^\pi[A_t = a \mid X_t = y, X_0 = x]}_{=\pi_t^x(a \mid y)} \mathbb{P}^\pi[X_t = y \mid X_0 = x] \\ &= \mathbb{P}^{\pi^x}[A_t = a \mid X_t = y, X_0 = x] \mathbb{P}^{\pi^x}[X_t = y \mid X_0 = x] \\ &= \mathbb{P}^{\pi^x}[X_t = y, A_t = a \mid X_0 = x] \quad \square \end{aligned}$$

**Corollary 1.5.6** (Puterman). *For (an optimal) policy  $\pi \in \Pi$  exist Markovian policies  $\pi^x$  for all  $x \in \mathcal{X}$  such that*

$$V^\pi(x) = V^{\pi^x}(x).$$

*Proof.*

$$\begin{aligned}
V^\pi(x) &= \mathbb{E}^\pi \left[ \sum_{t=0}^{\infty} \gamma^t R_{t+1} \mid X_0 = x \right] \\
&\stackrel{\text{Fubini}}{=} \sum_{t=0}^{\infty} \gamma^t \mathbb{E}^\pi [R_{t+1} \mid X_0 = x] \\
&= \sum_{t=0}^{\infty} \gamma^t \sum_{(y,a) \in \mathcal{X} \times \mathcal{A}} \mathbb{E}^\pi [R_{t+1} \mid X_t = y, A_t = a, X_0 = x] \cdot \mathbb{P}^\pi(X_t = y, A_t = a \mid X_0 = x) \\
&= \sum_{t=0}^{\infty} \gamma^t \sum_{(y,a) \in \mathcal{X} \times \mathcal{A}} \mathbb{E}[R_{(x,a)}] \mathbb{P}^{\pi^x}(X_t = y, A_t = a \mid X_0 = x) \\
&= \dots \stackrel{\text{analogous}}{=} V^{\pi^x}(x) \quad \square
\end{aligned}$$

Using the results from Puterman we can now show:

**Theorem 1.5.7.** *Let  $\pi \in \Pi$  be an optimal policy (i.e.  $V^* = V^\pi$ ) then there exists an optimal stationary policy  $\hat{\pi}$ .*

*Proof.* Because of 1.5.5, for the optimal policy  $\pi$  exist Markovian policies  $\pi^x$  for  $x \in \mathcal{X}$  with

$$V^*(x) = V^\pi(x) = V^{\pi^x}(x).$$

From these Markovian policies we can define the stationary policy  $\hat{\pi}$  with

$$\hat{\pi}(\cdot \mid x) := \pi_0^x(\cdot \mid x) \quad x \in \mathcal{X}. \quad (1.15)$$

For which we need to show that it is optimal. To start we have:

$$\begin{aligned}
V^*(x) &= V^{\pi^x}(x) = \mathbb{E}^{\pi^x}[R_1 \mid X_0 = x] + \gamma \mathbb{E}^{\pi^x} \left[ \sum_{t=0}^{\infty} \gamma^t R_{t+2} \mid X_0 = x \right] \\
&= \sum_{a \in \mathcal{A}_x} \mathbb{E}[R_{(x,a)}] \pi_0^x(a \mid x) \\
&\quad + \gamma \sum_{y \in \mathcal{X}} \mathbb{P}^{\pi^x}(X_1 = y \mid X_0 = x) \mathbb{E}^{\pi^x} \left[ \sum_{t=0}^{\infty} \gamma^t R_{t+2} \mid X_1 = y, X_0 = x \right] \quad (1.16)
\end{aligned}$$

Where the conditional probability equals:

$$\begin{aligned}
& \mathbb{P}^{\pi^x}(X_1 = y \mid X_0 = x) \\
&= \sum_{a \in \mathcal{A}_x} \mathbb{P}^{\pi^x}(X_1 = y, A_0 = a \mid X_0 = x) \\
&= \sum_{a \in \mathcal{A}_x} \mathbb{P}^{\pi^x}(X_1 = y \mid X_0 = x, A_0 = a) \mathbb{P}^{\pi^x}(A_0 = a \mid X_0 = x) \\
&= \sum_{a \in \mathcal{A}_x} p(y \mid x, a) \pi_0^x(a \mid x) = \sum_{a \in \mathcal{A}_x} p(y \mid x, a) \hat{\pi}(a \mid x) \\
&= \mathbb{P}^{\hat{\pi}}(X_1 = y \mid X_0 = x)
\end{aligned} \tag{1.17}$$

And the conditional expectation is:

$$\begin{aligned}
& \mathbb{E}^{\pi^x} \left[ \sum_{t=0}^{\infty} \gamma^t R_{t+2} \mid X_1 = y, X_0 = x \right] \\
&= \sum_{a \in \mathcal{A}_y} \mathbb{E}^{\pi^x} \left[ \sum_{t=0}^{\infty} \gamma^t R_{t+2} \mid X_1 = y, A_1 = a, X_0 = x \right] \pi_1^x(a \mid y) \\
&\stackrel{\text{Markov}}{=} \sum_{a \in \mathcal{A}_y} \mathbb{E}^{\pi^x} \left[ \sum_{t=0}^{\infty} \gamma^t R_{t+2} \mid X_1 = y, A_1 = a \right] \pi_1^x(a \mid y) \\
&= \mathbb{E}^{\pi^x} \left[ \sum_{t=0}^{\infty} \gamma^t R_{t+2} \mid X_1 = y \right] = \mathbb{E}^{\tilde{\pi}^x} \left[ \sum_{t=0}^{\infty} \gamma^t \tilde{R}_{t+1} \mid \tilde{X}_0 = y \right] \\
&= V^{\tilde{\pi}^x}(y)
\end{aligned} \tag{1.18}$$

Where we renamed the variables

$$\tilde{X}_t := X_{t+1}, \quad \tilde{A}_t := A_{t+1}, \quad \tilde{R}_t := R_{t+1}, \quad \tilde{\pi}_t^x = \pi_{t+1}^x,$$

creating the sequence  $((\tilde{X}_t, \tilde{A}_t, \tilde{R}_{t+1}), t \in \mathbb{N}_0)$  consistent with the MDP together with policy  $\tilde{\pi}^x$ . And if it were not the case, that

$$V^{\tilde{\pi}^x}(z) = V^*(z) \quad \forall z \in \mathcal{X} : \mathbb{P}^{\pi^x}(X_1 = z \mid X_0 = x) > 0, \tag{1.19}$$

we can use the property of the policy  $\pi^z$

$$V^{\pi^z}(z) = V^*(z),$$

and improve the optimal policy  $\pi^x$  by swapping out the policy  $(\pi_t^x, t \in \mathbb{N})$  with  $(\pi_{t-1}^z, t \in \mathbb{N})$  conditional on  $X_0 = x$  and  $X_1 = z$  (losing the Markov property in the process):

Fixme: ugly - better solution?

$$\begin{aligned}
\hat{\pi}_0^x(a \mid x) &:= \pi_0^x(a \mid x) \\
\hat{\pi}_t^x(a \mid X_0 = x, X_1 = z, X_t = x_t) &:= \pi_{t-1}^z(a \mid x_t) && \text{for } t \geq 1 \\
\hat{\pi}_t^x(a \mid X_0 = x, X_1 \neq z, X_t = x_t) &:= \pi_t^x(a \mid x_t) && \text{for } t \geq 1 \\
\hat{\pi}_t^x(a \mid y) &:= \pi_t^x(a \mid y) && \text{for } y \neq x
\end{aligned}$$

And if the same steps are applied to  $\hat{\pi}^x$  as to  $\pi^x$  in (1.16), everything but the summand  $z$  in the large sum will stay the same. And since we assumed 1.19 were false, this summand increases because

$$V^*(z)\mathbb{P}^{\pi^x}(X_1 = z \mid X_0 = x) > V^{\hat{\pi}^x}(z)\mathbb{P}^{\pi^x}(X_1 = z \mid X_0 = x).$$

But that implies that  $V^*(x) < V^{\hat{\pi}^x}(x)$ , which is a contradiction. Therefore (1.19) holds. Using this fact, we can conclude:

$$\begin{aligned} V^*(x) &= V^{\pi^x}(x) \\ &= \sum_{a \in \mathcal{A}_x} \mathbb{E}[R_{(x,a)}] \pi_0^x(a \mid x) + \gamma \sum_{y \in \mathcal{X}} \mathbb{P}^{\pi^x}(X_1 = y \mid X_0 = x) V^{\hat{\pi}^x}(y) \\ &= \sum_{a \in \mathcal{A}_x} \mathbb{E}[R_{(x,a)}] \hat{\pi}(a \mid x) + \gamma \sum_{y \in \mathcal{X}} \mathbb{P}^{\hat{\pi}}(X_1 = y \mid X_0 = x) V^*(y) \\ &= T^{\hat{\pi}} V^*(x) \end{aligned}$$

Since  $V^*$  is a fixed point of  $T^{\hat{\pi}}$  and the fixed point is unique we have completed our proof, since then  $V^{\hat{\pi}} = V^*$  which means  $\hat{\pi}$  is optimal.  $\square$

In Summary: With an arbitrary optimal policy we can construct an optimal stationary policy with 1.5.7, and if there is an optimal stationary policy, there is also an optimal deterministic stationary policy (1.5.2). Therefore the set of deterministic stationary policies is large enough to choose from if you are looking for an optimal policy.

But what happens if there are no optimal policies? If there are no optimal policies, can one sequence of deterministic policies get arbitrarily close to the optimal value function for *every* starting state? We are now able to answer the question we were not quite ready to answer in Finite Outmatching (c.f 1.4.3).

FiXme: is this Title fix?

**Proposition 1.5.8** ( $\varepsilon$ -Optimal Policies). *For every  $\varepsilon > 0$  exists a deterministic stationary policy  $\pi^\varepsilon$  such that  $V^* \leq V^{\pi^\varepsilon} + \varepsilon$ .*

*Proof.* Be  $\varepsilon > 0$  arbitrary, define  $\delta := \varepsilon(1 - \gamma)$ . Because of this

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

there exists an  $\pi^\varepsilon(x)$  for all  $x \in \mathcal{X}$  such that

$$r(x, \pi^\varepsilon(x)) + \gamma \sum_{y \in \mathcal{X}} p(y \mid x, \pi^\varepsilon(x)) V^*(y) + \delta \geq V^*(x). \quad (1.20)$$

This defines a mapping  $\pi^\varepsilon: \mathcal{X} \rightarrow \mathcal{A}_x$  with this property

$$\begin{aligned} T^{\pi^\varepsilon} V^*(x) &= r(x, \pi^\varepsilon(x)) + \gamma \sum_{y \in \mathcal{X}} p(y \mid x, \pi^\varepsilon(x)) V^*(y) \\ &\geq V^*(x) - \delta = (V^* - \delta \mathbf{1})(x). \end{aligned}$$

By induction we get

$$(T^{\pi^\varepsilon})^n V^*(x) \geq (V^* - (\delta \sum_{k=0}^n \gamma^k) \mathbf{1})(x).$$

Using the monotonicity of  $T^{\pi^\varepsilon}$  (1.3.6) we can make the induction step:

$$\begin{aligned} (T^{\pi^\varepsilon})^{n+1} V^*(x) &= T^{\pi^\varepsilon} (T^{\pi^\varepsilon})^n V^*(x) \\ &\stackrel{\text{ind.}}{\geq} T^{\pi^\varepsilon} (V^* - (\delta \sum_{k=0}^n \gamma^k) \mathbf{1})(x) \\ &\stackrel{\text{affine}}{=} T^{\pi^\varepsilon} V^*(x) - \gamma \sum_{y \in \mathcal{X}} p(y \mid x, a) (\delta \sum_{k=0}^n \gamma^k) \mathbf{1}(y) \\ &\geq V^*(x) - \delta - \delta \sum_{k=0}^n \gamma^{k+1} \sum_{y \in \mathcal{X}} p(y \mid x, a) \\ &= V^*(x) - \delta \sum_{k=0}^{n+1} \gamma^k \end{aligned}$$

This concludes our proof with:

$$\begin{aligned} V^{\pi^\varepsilon}(x) &= \lim_{n \rightarrow \infty} (T^{\pi^\varepsilon})^n V^* \geq V^*(x) - \delta \lim_{n \rightarrow \infty} \sum_{k=0}^n \gamma \\ &= V^*(x) - \delta/(1 - \gamma) = V^*(x) - \varepsilon \end{aligned}$$

□

## 1.6 Dynamic Programming

If we have full knowledge of all transition probabilities and immediate rewards, we can use *value iteration* and *policy iteration* to approach the optimal value functions. These methods developed by Richard Bellman are known as “Dynamic Programming” as they recursively break down the problem using the Bellman equations. Since they use estimates of  $V^*$  to create better estimates, they are also a case of *bootstrapping*.

**Definition 1.6.1.** Let  $V_0 \in B(\mathcal{X})$  or  $Q_0 \in B(\mathcal{X} \times \mathcal{A})$  be arbitrary, then  $(V_k, k \in \mathbb{N}_0)$  or  $(Q_k, k \in \mathbb{N}_0)$  are called *value iterations* if

$$V_{k+1} := T^* V_k \quad \text{or} \quad Q_{k+1} := T^* Q_k$$

This converges geometrically with regard to the  $\|\cdot\|_\infty$ -norm, due to  $T^*$  fulfilling the requirements of the BFT (1.4.7).

To find an  $\varepsilon$ -optimal policy, define  $\delta := \varepsilon(1 - \gamma)$  as in 1.5.8. This iteration can then be stopped when  $\|V_k - V^*\|_\infty \leq \delta/2$ , which also implies

$$\|T^* V_k - V^*\|_\infty = \|T^* V_k - T^* V^*\|_\infty \leq \gamma \delta/2 \leq \delta/2.$$



So picking  $\pi^\varepsilon$  with regard to  $V_k$  instead of  $V^*$  like in 1.5.8 results in

$$\begin{aligned} r(x, \pi^\varepsilon(x)) + \gamma \sum_{y \in \mathcal{X}} p(y \mid x, \pi^\varepsilon(x)) V_k(y) + \delta/2 &\geq T^* V_k(x) + \delta/2 \\ &\geq V^*(x) + \delta \end{aligned}$$

From there the proof is the same as from (1.20).

The other iteration, policy iteration, works if a greedy policy can always be selected. Since this requires a supremum over actions to be attained, this can not be guaranteed in general. Though for finite MDPs this is always the case.

**Definition 1.6.2.** Let  $\pi_0 \in \Pi_S^D$  be an arbitrary deterministic stationary policy. Then  $(\pi_k, k \in \mathbb{N}_0)$  is called *policy iteration* if

$$\pi_{k+1} \in \Pi_S^D \text{ is greedy with regard to } Q^{\pi_k}.$$

**Proposition 1.6.3.** Let  $(\pi_k, k \in \mathbb{N}_0)$  be a policy iteration, then

$$\|Q^{\pi_k} - Q^*\|_\infty \rightarrow 0 \quad (k \rightarrow \infty).$$

*Proof.* Let  $(\pi_k, k \in \mathbb{N}_0)$  be a policy iteration. Then by definition

$$\pi_{k+1}(x) = \arg \max_{a \in \mathcal{A}_x} Q^{\pi_k}(x, a), \quad (1.21)$$

which means:

$$\begin{aligned} V^{\pi_k} &\leq T^* V^{\pi_k}(x) \\ &= \sup_{a \in \mathcal{A}_x} \left\{ r(x, a) + \gamma \sum_{y \in \mathcal{X}} p(y \mid x, a) V^{\pi_k}(y) \right\} \\ &\stackrel{1.3.2}{=} \sup_{a \in \mathcal{A}_x} Q^{\pi_k}(x, a) \\ &\stackrel{(1.21)}{=} Q^{\pi_k}(x, \pi_{k+1}(x)) \\ &= T^{\pi_{k+1}} V^{\pi_k}(x) \end{aligned} \quad (1.22)$$

Because of the monotonicity of  $T^{\pi_{k+1}}$  (1.3.6) this implies

$$V^{\pi_{k+1}} = \lim_{n \rightarrow \infty} (T^{\pi_{k+1}})^n V^{\pi_k} \geq T^{\pi_{k+1}} V^{\pi_k} = T^* V^{\pi_k}. \quad (1.23)$$

Which in turn implies by induction  $V^{\pi_k} \geq (T^*)^k V^{\pi_0}$ , and therefore:

$$\lim_{k \rightarrow \infty} \|V^* - V^{\pi_k}\|_\infty \leq \lim_{k \rightarrow \infty} \|V^* - (T^*)^k V^{\pi_0}\|_\infty \stackrel{1.4.7}{=} 0$$

This is easily translated into the action value version:

$$\begin{aligned} \|Q^* - Q^{\pi_k}\|_\infty &\stackrel{1.3.2}{\leq} \sup_{(x,a) \in \mathcal{X} \times \mathcal{A}} \left| \gamma \sum_{y \in \mathcal{X}} p(y \mid x, a) (V^*(y) - V^{\pi_k}(y)) \right| \\ &\leq \gamma \|V^* - V^{\pi_k}\|_\infty \end{aligned} \quad \square$$

Policy iteration converges faster than value iteration with every step (c.f. (1.23)), in the case of finite MDPs it actually converges in finitely many steps. And while it requires the calculation of  $Q^{\pi_k}$  in every step, this turns out to be less of a problem than one might think:

**Corollary 1.6.4.**

- (i) *Policy iteration in finite MDPs converges in finitely many steps*
- (ii)  *$V^\pi$  (and thus  $Q^\pi$ ) can be calculated in one step*

*Proof.* (i) (Szepesvári 2010) Consider equation (1.22).

If it is an equality, then the current step is optimal since the value function is a fixed point of  $T^*$ .

If it is a true inequality, the value function truly increases in this iteration. But since there are only a finite set of deterministic stationary policies on finite MDPs

$$|\Pi_S^D| = |\{\pi : \mathcal{X} \rightarrow \mathcal{A}_x\}| = \prod_{x \in \mathcal{X}} |\mathcal{A}_x| < \infty$$

there is only a finite set of value functions  $V^\pi$  which means that it has to stop increasing after a finite number of steps.

(ii) Since the finite case is analogous to the (countably) infinite case, but easier to write down and understand, we will prove the finite case only. Let  $\mathcal{X} = \{x_1, \dots, x_n\}$  be the finite state set. Then we can interpret  $V : \mathcal{X} \rightarrow \mathbb{R}$  as a vector of  $\mathbb{R}^n$  and define

$$r^\pi := (r(x_1, \pi(x_1)), \dots, r(x_n, \pi(x_n)))^T \in \mathbb{R}^n$$

$$P^\pi := \begin{pmatrix} p(x_1 | x_1, \pi(x_1)) & \cdots & p(x_n | x_1, \pi(x_1)) \\ \vdots & & \vdots \\ p(x_1 | x_n, \pi(x_n)) & \cdots & p(x_n | x_n, \pi(x_n)) \end{pmatrix}$$

This allows us to write

$$T^\pi V = r^\pi + \gamma P^\pi V.$$

And since the Banach fixpoint iteration can be started with any bounded value function, it can in particular be started with the zero function or rather vector. Therefore by induction with induction basis ( $k = 1$ )

$$(T^\pi)^k 0 = r^\pi = \sum_{i=0}^{k-1} (\gamma P^\pi)^i r^\pi,$$

and induction step ( $k \rightarrow k + 1$ )

$$(T^\pi)^{k+1} 0 = T^\pi \left( \sum_{i=0}^{k-1} (\gamma P^\pi)^i r^\pi \right) = r^\pi + \sum_{i=0}^{k-1} (\gamma P^\pi)^{i+1} r^\pi = \sum_{i=0}^k (\gamma P^\pi)^i r^\pi,$$

we get

$$V^\pi = \sum_{k=0}^{\infty} (\gamma P^\pi)^k r^\pi = (1 - \gamma P^\pi)^{-1} r^\pi.$$

□

Fixme: Write down  
Algorithm?



## Chapter 2

# Stochastic Approximation – Convergence Proofs

### 2.1 Introduction



# Appendix A

## Appendix

### A.1 Basic Probability Theory

**Lemma A.1.1.**

- (i)  $\mathbb{P}(A \cap B \mid C) = \mathbb{P}(A \mid B \cap C)\mathbb{P}(B \mid C)$
- (ii)  $\mathbb{P}(A \mid C) = \sum_{n \in \mathbb{N}} \mathbb{P}(A \mid B_n \cap C)\mathbb{P}(B_n \mid C)$  for  $\mathbb{P}(\biguplus_{n \in \mathbb{N}} B_n) = 1$
- (iii)  $\mathbb{E}[X \mid C] = \sum_{n \in \mathbb{N}} \mathbb{E}[X \mid C \cap B_n]\mathbb{P}(B_n \mid C)$  for  $\mathbb{P}(\biguplus_{n \in \mathbb{N}} B_n) = 1$

*Proof.* (i)

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

(ii)

$$\begin{aligned} \mathbb{P}(A \mid C) &= \mathbb{P}\left(A \cap \biguplus_{n \in \mathbb{N}} B_n \mid C\right) = \sum_{n \in \mathbb{N}} \mathbb{P}(A \cap B_n \mid C) \\ &\stackrel{(i)}{=} \sum_{n \in \mathbb{N}} \mathbb{P}(A \mid B_n \cap C)\mathbb{P}(B_n \mid C) \end{aligned}$$

(iii)

$$\begin{aligned} \mathbb{E}[X \mid C] &= \frac{1}{\mathbb{P}(C)} \int_C X d\mathbb{P} = \sum_{n \in \mathbb{N}} \frac{1}{\mathbb{P}(C)} \int_{C \cap B_n} X d\mathbb{P} \\ &= \sum_{n \in \mathbb{N}} \frac{\mathbb{P}(C \cap B_n)}{\mathbb{P}(C)} \frac{1}{\mathbb{P}(C \cap B_n)} \int_{C \cap B_n} X d\mathbb{P} \\ &= \sum_{n \in \mathbb{N}} \mathbb{E}[X \mid C \cap B_n]\mathbb{P}(B_n \mid C) \end{aligned} \quad \square$$

**Lemma A.1.2.** Let  $(\Omega, \mathcal{A}, \mu)$  be a measure space and a function  $f$  exists with

$$f: \Omega \rightarrow \mathbb{R} \text{ injective and measurable,}$$

$$f^{-1}: f(\Omega) \rightarrow \Omega \text{ measurable.}$$

Then for  $X$   $\Omega$ -valued random variable and  $Y$   $f(\Omega)$ -valued random variable

$$\mathbb{P}_{f \circ X} = \mathbb{P}_Y \iff \mathbb{P}_X = \mathbb{P}_{f^{-1} \circ Y}$$

*Proof.* “ $\Leftarrow$ ” Let  $A \in \mathcal{B}(\mathbb{R})$ , then w.l.o.g.  $A \subseteq f(\Omega)$  otherwise

$$\mathbb{P}_{f \circ X}(A) = \mathbb{P}(A \cap f(\Omega)) + \underbrace{\mathbb{P}_{f \circ X}(A \cap f(\Omega)^c)}_{=0} = \dots = \mathbb{P}_Y(A)$$

Thus  $f \circ f^{-1}(A) = A$  holds, which finishes this direction with

$$\begin{aligned} \mathbb{P}_{f \circ X}(A) &= \mathbb{P}(X^{-1} \circ f^{-1}(A)) = \mathbb{P}_X(f^{-1}(A)) \\ &= \mathbb{P}_{f^{-1} \circ Y}(f^{-1}(A)) = \mathbb{P}(Y^{-1} \circ f \circ f^{-1}(A)) \\ &= \mathbb{P}_Y(A) \end{aligned}$$

“ $\Rightarrow$ ” Let  $A \in \mathcal{A}$ , then

$$\begin{aligned} \mathbb{P}_X(A) &= \mathbb{P}(X^{-1} \circ f^{-1} \circ f(A)) = \mathbb{P}_{f \circ X}(f(A)) \\ &= \mathbb{P}_Y(f(A)) = \mathbb{P}(Y^{-1} \circ f(A)) \\ &= \mathbb{P}_{f^{-1} \circ Y}(A) \end{aligned} \quad \square$$

**Definition A.1.3** (Pseudo-inverse). Let  $F$  be a cumulative distribution function, then

$$F^{\leftarrow}(y) := \inf\{x \in \mathbb{R} : F(x) \geq y\}$$

is called the *Pseudo-inverse* of  $F$ .

**Lemma A.1.4.** Let  $F$  be a cdf, then

- (i)  $F^{\leftarrow}(y) \leq x \iff y \leq F(x)$
- (ii)  $U \sim \mathcal{U}(0, 1) \implies F^{\leftarrow}(U) \sim F$

*Proof.* (i) “ $\Rightarrow$ ”

$$\begin{aligned} y &\leq \inf_{x \in \{z \in \mathbb{R} : F(z) \geq y\}} F(x) \stackrel{\text{F right-continuous}}{=} F(\inf\{z \in \mathbb{R} : F(z) \geq y\}) \stackrel{\text{def.}}{=} F(F^{\leftarrow}(y)) \\ &\leq F(x) \end{aligned}$$

Where the last inequality follows from the assumption  $F^{\leftarrow}(y) \leq x$  and  $F$  being non decreasing.

“ $\Leftarrow$ ” Follows simply from the fact that  $x$  is included in the set of the infimum.

$$y \leq F(x) \implies F^{\leftarrow}(y) = \inf\{z \in \mathbb{R} : F(z) \geq y\} \leq x$$

(ii) is a simple corollary from (i)

$$\mathbb{P}(F^{\leftarrow}(U) \leq x) \stackrel{(i)}{=} \mathbb{P}(U \leq F(x)) = F(x) \quad \square$$



## A.2 Analysis



# Bibliography

- Puterman, Martin L. (2005). *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. Wiley Series in Probability and Statistics. OCLC: 254152847. Hoboken, NJ: Wiley-Interscience. 649 pp. ISBN: 978-0-471-72782-8.
- 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.
- 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.