List of Corrections

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

Chapter 1

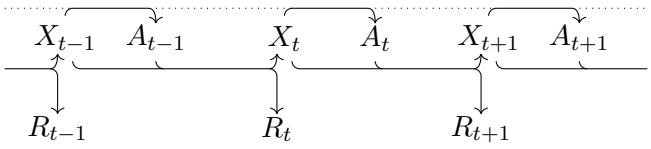
Markov Decision Processes

1.1 Introduction

In this introduction I try to convey an intuition for Markov Decision Processes. Readers familiar with the subject can skip directly to the more 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). That 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 case of 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 behavior, a decision. So defining MDPs as a stochastic process, would result in every behavior resulting in a different MDP. But since MDPs want to model decisions, this does not make much sense. Talking about optimal behaviors in a framework which can only be defined for a set behavior, is nonsensical. What is needed is a model framework which is invariant to different

behaviors. Therefore MDPs are not defined as a stochastic process, but rather as a rulebook on how to create a stochastic process from a given behavior.

The questions we ask the model are also different. MDPs are more interested into evaluating different behaviors and finding optimal ones, while Markov Processes try to describe an existing phenomenon. Though we will find, that – given a behavior 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.

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

FiXme: more diverse examples

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.

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,

$\mathcal{A} \coloneqq \mathbb{N}$ be the set of possible orders for resupply.

We will now introduce the mechanics of actions and rewards without stochastic behavior 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

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

We express this with the *state transition function* f defined as

$$\begin{aligned} &f(x, a, d) \coloneqq (x - d \wedge x + a,\ p(d \wedge x) - a - q(x - d \wedge x)) \\ &\implies f(x_t, a_t, d_t) = (x_{t+1}, r_{t+1}) \end{aligned}$$

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) neccessarily 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

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

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 behavior in this example would be a probability distribution over the possible orders, given the current state

$$\pi(\cdot \mid x) \text{ Probability distribution over } \mathcal{A}$$

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

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

Non-stationary behaviors 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.

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

Since we will interpret probability kernels as distributions over Y given a certain condition $x \in X$, the notation $\lambda(\cdot \mid x) \coloneqq \lambda(x, \cdot)$ 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 A_x the countable (finite) set of possible actions in state x,
$\mathcal{P}\colon (\mathcal{X} \times \mathcal{A}) \times \sigma_{\mathcal{X} \times \mathbb{R}} \to \mathbb{R}$ is a probability kernel, with the notation

$$\mathcal{X} \times \mathcal{A} \coloneqq \{(x, a) : x \in \mathcal{X}, a \in \mathcal{A}_x\} \tag{1.1}$$

$\mathcal{X} \times \mathbb{R}$ 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.

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.

FiXme: Make this a definition of some sort?

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) *behavior (policy)*, where

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

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

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

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

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

2. *Stationary policies* are policies which do not change over time i.e. $\pi_{t+1} = \pi_t$. Since π_0 can only depend on the first state, they are naturally markov and can be defined as

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

3. *Determinisitic stationary policies* are specified by[2]

$$\pi \colon \mathcal{X} \to \mathcal{A}_x \quad \text{with } A_t = \pi(X_t)$$

[1] 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))
[2] $\pi \colon \mathcal{X} \to \mathcal{A}_x$ is notation for $\pi \colon \mathcal{X} \to \bigcup_{x \in \mathcal{X}} \mathcal{A}_x : \pi(x) \in \mathcal{A}_x$

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

$$\begin{array}{lccccc} & \text{Stationary} & & \text{Markov} & & \text{History dependent} \\ \text{Deterministic} & \Pi_S^D & \subseteq & \Pi_M^D & \subseteq & \Pi^D \\ & \cap\!\mid & & \cap\!\mid & & \cap\!\mid \\ \text{Stochastic} & \Pi_S & \subseteq & \Pi_M & \subseteq & \Pi \end{array}$$

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.

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

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 $cX \times \mathbb{R}$ in $\mathbb{R}$. Since $\mathcal{X}$ is countable there exists an injective measurable mapping $g\colon \mathcal{X} \to \mathbb{N}$ and because there exist an injective sigmoid function $\sigma\colon \mathbb{R} \to (0,1)$, there exists an injective measurable function

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

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

$$\begin{aligned} \tilde{\mathcal{P}}(\cdot \mid x, a) &\coloneqq \mathbb{P}_{f \circ Y} \text{ where } Y \sim \mathcal{P}(\cdot \mid x, a) \\ F_{x,a}(y) &\coloneqq \tilde{\mathcal{P}}((-\infty, y] \mid x, a) \end{aligned}$$

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

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

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

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

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

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

where H_i is defined as

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

With this notation we can now finish the proof

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

Proposition 1.2.6. *A (stationary) markov policy π 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

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

$(*)$ *Some* of the History is irrelevant if all of the History is irrelevant.

FiXme: some history irrelevant in appendix?

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 $f\colon \mathcal{X} \to \mathcal{A}_x$ 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. Therefore f is a deterministic stationary behavior 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 behaviors. 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*

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

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

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

With the return we can define value functions.

Definition 1.2.10. Let $\mathcal{M} = (\mathcal{X}, \mathcal{A}, \mathcal{P})$ be a MDP. Assume an exploring start, i.e. select X_0 such that $\mathbb{P}(X_0 = x) > 0$ for all states x. Then every behavior π 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 behavior generated the process in the following definitions.

$$V^\pi\colon \begin{cases} \mathcal{X} \to \mathbb{R} \\ x \mapsto \mathbb{E}^\pi[\mathcal{R} \mid X_0 = x] \end{cases} \quad \text{is called } \textit{value function} \text{ for } \pi,\ ^1$$

$$Q^\pi\colon \begin{cases} \mathcal{X} \times \mathcal{A} \to \mathbb{R} \\ (x,a) \mapsto \mathbb{E}^\pi[\mathcal{R} \mid X_0 = x, A_0 = a] \end{cases} \quad \text{is called } \textit{action value function} \text{ for } \pi,\ ^2$$

$$V^*\colon \begin{cases} \mathcal{X} \to \mathbb{R} \\ x \mapsto \sup\limits_{\pi \in \Pi} V^\pi(x) \end{cases} \quad \text{is called } \textit{optimal value function},$$

$$Q^*\colon \begin{cases} \mathcal{X} \times \mathcal{A} \to \mathbb{R} \\ (x,a) \mapsto \sup\limits_{\pi \in \Pi} Q^\pi(x,a) \end{cases} \quad \text{is called } \textit{optimal action value function},$$

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

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

[2] Well defined because $A_1 \sim \pi_1(\cdot \mid (x, a, r_0), x_1)$ is defined for all a regardless of π_0

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 behavior. This means V^π, Q^π 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

FiXme: he quotes Blackwell 1965 - quote blackwell directly?

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

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 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. So we will define the terms around terminal states here although we will not make use of them.

FiXme: We will not use terminal states - true?

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

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

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

1.2.1 Outlook

To avoid clutter we will from now on we 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 fix-point equation. And that this fix-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^π and Q^π which fulfill 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 fix-point equations for the optimal value functions. One intuitive example for such a use, is that you can try to show that V^* fulfills the fixpoint equation for V^π and use uniqueness to argue that π has to be optimal. Other cases are a bit more technical and will use a type of monotonicity of the fix point equation.

To find these fixpoint equations we will try to set V^π and Q^π 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.

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

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

Now we can start to explore the relation of V^π and Q^π.

Proposition 1.3.2. *Let $\pi \in \Pi_S$ be a 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. We will use A.1.1

$$\mathbb{E}[X \mid A] = \sum_{n\in\mathbb{N}} \mathbb{E}[X \mid A\cap B_n]\mathbb{P}(B_n \mid A) \text{ for } \mathbb{P}\left(\biguplus_{n\in\mathbb{N}} B_n\right) = 1$$

quite a bit in this proof.

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

This is almost what we want since $r(x,a) = \mathbb{E}[R_{(x,a)}]$, so we just need to look at the conditional expectation

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

$(*)$ We rename

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

then $(\tilde{X}_t, \tilde{A}_t, \tilde{R}_{t+1}, t \in \mathbb{N}_0)$ is an evaluation of the MDP with the (stationary!) policy π. □

FiXme: or realization or "Markov Action Process"? or something else?

Corollary 1.3.3. *For $\pi \in \Pi_S$, this fix 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 $\pi \in \Pi_S^D$ this fix-point equation holds

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

and this fix-point equation

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

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

$$\begin{aligned}
V^\pi(x) &= \mathbb{E}^\pi[\mathcal{R} \mid X_0 = x] \\
&= \sum_{a \in \mathcal{A}_x} \underbrace{\mathbb{E}^\pi[\mathcal{R} \mid X_0 = x, A_0 = a]}_{=Q^\pi(x,a)} \pi(a \mid x) \qquad (1.4) \\
&= \sum_{a \in \mathcal{A}_x} \left(r(x,a) + \gamma \sum_{y \in \mathcal{X}} p(y \mid x,a) V^\pi(y) \right) \pi(a \mid x) \\
&= \sum_{a \in \mathcal{A}_x} r(x,a)\pi(a \mid x) + \gamma \sum_{(y,a) \in \mathcal{X} \times \mathcal{A}_x} V^\pi(y) p(y \mid x,a) \pi(a \mid x) \qquad (1.5) \\
&= \mathbb{E}^\pi[r(x, A_0) \mid X_0 = x] + \gamma \sum_{(y,a) \in \mathcal{X} \times \mathcal{A}_x} V^\pi(y) \mathbb{P}^\pi(X_1 = y, A_0 = a \mid X_0 = x) \\
&= \mathbb{E}^\pi[r(x, A_0) \mid X_0 = x] + \gamma \sum_{y \in \mathcal{X}} V^\pi(y) \mathbb{P}^\pi(X_1 = y \mid X_0 = x)
\end{aligned}$$

And for the case, where π 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^π fix 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

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

□

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

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

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

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

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

$$B(M) \coloneqq \{f \colon M \to \mathbb{R} : \|f\|_\infty < \infty\}$$

which happens to be a complete metric space which we will need for the Banach fix-point Theorem to work. Bounded Value functions will also be neccessary 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 $T^\pi \colon B(\mathcal{X}) \to B(\mathcal{X})$ with

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

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

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

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

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

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

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^π for a proof about optimal policies later on. In contrast we only need the the special case for the operator on action value functions Q^π.

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).

T^π 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: proof BFT? - 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^π is a contraction. Be $V, W \in B(\mathcal{X})$:

FiXme: show: is com
metric space - apper

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

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

Lemma 1.3.6. *Observe that*

1. *T^π 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^π is monotonous,

FiXme: comma, fullstop, nothing?

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

Proof. Be $W_1, W_2 \in B(\mathcal{X})$, $W_1 \leq W_2$ and $x \in \mathcal{X}$, 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 a disctinction 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 uniquness of the fix-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 theses 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 adress 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 apporximate 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 show 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 achived by the seqences of behaviors

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

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. 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:*

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

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

$$\hat{\pi} : \begin{cases} \mathcal{X} \to \mathcal{A} \\ x \mapsto \pi_{\underset{i=1,\dots,n}{\operatorname{argmax}} V^{\pi_i}(x)}(x) \end{cases}$$

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 achive 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

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

And be

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

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

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

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

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

thus

$$V^{\hat{\pi}} = \lim_{n \to \infty} (T^{\hat{\pi}})^n V \geq V \geq V^{\pi_i} \qquad \forall i = 1, \dots, n$$ □

1.4.2 Fix-Point Equations for Optimal Value functions

With this statement we can now prove the building blocks for the fix-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 γ, 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 $M_\delta \coloneqq \{y \in \mathcal{X} : p(y \mid x,a) > \delta\}$ is finite for all $\delta > 0$, and

$$\begin{aligned}1 = p(\mathcal{X} \mid x,a) &= p\left(\bigcup_{\delta \to 0} M_\delta \,\middle|\, x,a \right) = \lim_{\delta \to 0} p(M_\delta \mid x,a) \\ &= \lim_{\delta \to 0} \sum_{y \in M_\delta} p(y \mid x,a)\end{aligned}$$

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

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

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

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

And also because M_δ is finite and 1.4.3 we know that

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

FiXme: slightly ugly vs. focus on important bits

This finally implies

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

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 $V^\pi(x) = Q^\pi(x, \pi(x))$ thus

$$\begin{aligned}
\tilde{V}(x) &= \sup_{\pi \in \Pi_S^D} V^\pi(x) = \sup_{\pi \in \Pi_S^D} Q^\pi(x, \pi(x)) \\
&\leq \sup_{a \in \mathcal{A}_x} \sup_{\pi \in \Pi_S^D} Q^\pi(x, a) = \sup_{a \in \mathcal{A}_x} \tilde{Q}(x, a) \qquad (1.9) \\
&\overset{\text{(i)}}{=} \sup_{a \in \mathcal{A}_x} \left\{ r(x, a) + \gamma \sum_{y \in \mathcal{X}} p(y \mid x, a) \sup_{\pi \in \Pi_S^D} V^\pi(y) \right\} \qquad (1.10)
\end{aligned}$$

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

FiXme: exist in set?
ters up the equation
ng it harder to read,
math. more correct

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

Define a slightly changed deterministic policy with this π and a

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

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

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

In either case we get

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

By induction using the monotonicity of T^{π} (1.3.6) we get

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

thus

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

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

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

($*$) "$\leq$" is trivial, since a deterministic π_0 can achieve the same as fixing an action $A_0 = a$.

"$\geq$" follows from this inequality which holds for all π (especially for all π_0)

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

(iv) For an arbitrary policy π we can expand the action value function like this

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

This implies

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

The last equality follows from the fact, that the behaviors 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

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

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

$$\tilde{X}_t \coloneqq X_{t+1}, \quad \tilde{A}_t \coloneqq A_{t+1}, \quad \tilde{R}_t \coloneqq R_{t+1}, \quad \tilde{\pi}_t \coloneqq \pi_{t+1}$$ □

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

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

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

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

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

Corollary 1.4.6. *All the optimal value functions are fix-points of T^**

$$T^*\tilde{V} = \tilde{V} \qquad T^*\tilde{Q} = \tilde{Q}$$
$$T^*V^* = V^* \qquad T^*Q^* = Q^*$$

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

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

The others are analogous □

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. This proof was taken from 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

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

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 $V, W \in B(\mathcal{X})$

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

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

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

□

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) $\forall x \in \mathcal{X} : V^*(x) = \sum_{a \in \mathcal{A}_x} \pi^*(a \mid x) Q^*(x,a)$

(iii) $\forall x \in \mathcal{X} : \pi^* = \arg\max_{\pi^* \in \Pi_S} \sum_{a \in \mathcal{A}_x} \pi^*(a \mid x) Q^*(x,a)$

(iv) $\pi^*(a \mid x) > 0 \implies Q^*(x,a) = V^*(x) = \sup_{b \in \mathcal{A}_x} Q^*(x,b)$
"actions are concentrated on the set of actions that maximize $Q^*(x,\cdot)$"
(this also implies: $Q^*(x,a) < V^*(x) \implies \pi^*(a \mid x) = 0$)

Proof. "(i) $\Rightarrow$ (ii)" Let $x \in \mathcal{X}$ be arbitrary, then

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

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

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

Therefore $V^*(x)$ is an upper bound for every $\pi \in \Pi_S$, and since π^* attains this upper bound it is a maximum.
"(iii) $\Rightarrow$ (iv)" Be $\pi^*(a \mid x) > 0$ 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 behavior

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

This results in

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

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

$$M_x \coloneqq \{a \in \mathcal{A}_x : \pi^*(a \mid x) > 0\}$$

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

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

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

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

Therefore $V^{\pi^*} = V^*$ since the fix-point of T^{π^*} is unique (1.3.5). □

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. $Q\colon \mathcal{X} \times \mathcal{A} \to \mathbb{R}$ an action value function, $\tilde{\pi}\colon \mathcal{X} \to \mathcal{A}_x$ with

$$\tilde{\pi}(x) \coloneqq \arg\max_{a \in \mathcal{A}_x} Q(x,a) \qquad x \in \mathcal{X}$$

$\tilde{\pi}(x)$ 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 π^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 π.

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

$$\pi^x_t(\cdot \mid y) \coloneqq \mathbb{P}^\pi(A_t \mid X_t = y, X_0 = x) \tag{1.14}$$

We will now show the statement for this π^x by induction. The base case is

$$\begin{aligned}\mathbb{P}^\pi(X_0 = y, A_0 = a \mid X_0 = x) &= \mathbb{P}^\pi(A_0 = a \mid X_0 = x)\delta_{xy} \\ &= \pi^x_0(a \mid x)\delta_{xy} = \mathbb{P}^{\pi^x}(A_0 = a \mid X_0 = x)\delta_{xy} \\ &= \mathbb{P}^{\pi^x}(X_0 = y, A_0 = a \mid X_0 = x)\end{aligned}$$

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

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

together with A.1.1, we get

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

Using A.1.1 again we get

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

□

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

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

Proof.

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

□

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. For the optimal policy π exist $\pi^x, x \in \mathcal{X}$ markovian policies with

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

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

$$\hat{\pi}(\cdot \mid x) \coloneqq \pi_0^x(\cdot \mid x) \qquad x \in \mathcal{X} \tag{1.15}$$

For which we now need to show that it is optimal.

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

Where the conditional probability is

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

And the conditional expectation is

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

with

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

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 $\tilde{\pi}^x$. We can conclude that

FiXme: evaluated MDP?

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

otherwise we can use the fact that $V^{\pi^z}(z) = V^*(z)$ and improve the optimal policy π^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 again).

FiXme: ugly - better solution?

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

If you apply the same steps to $\hat{\pi}^x$ as to π^x in (1.16), everything but the summand z in the large sum will stay the same. And this summand increases because

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

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

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

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

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 (ε-Optimal Policies)**.** *For every $\varepsilon > 0$ exists a deterministic stationary policy π^ε 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

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

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

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

By induction we get

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

using the monotonicity of T^{π^ε} (1.3.6) we can make the induction step

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

This concludes our proof with

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

□

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 $(Q_k, k \in \mathbb{N}_0)$ is called a *value iteration* if

$$V_{k+1} \coloneqq T^* V_k \qquad\qquad Q_{k+1} \coloneqq T^* Q_k$$

This converges geometrically with regard to the $\|\cdot\|_\infty$-norm due to T^* fulfilling the requirements of the BFT (1.4.7). In general this iteration can of course only work on a finite MDP, since taking the supremum over an infinite amount of actions T^* for an infinite amount of states is impossible without some assumptions.

To find a close to optimal action this iteration can be stopped after a finite amount of steps, since from $\|V_k - V^*\|_\infty \leq \delta/2$ follows

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

So picking a π^ε as 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 a *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*

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

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

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

which means

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

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

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

This implies by induction $V^{\pi_k} \geq (T^*)^k V^{\pi_0}$

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

The statement follows from the fact, that Q^{π_k} can be expressed as a formula of V^{π_k} (c.f. 1.3.3) □

Converges faster than value iteration with every step (c.f. (1.23)) but requires calculation of Q^{π_k} in every step.

But in the finite case policy iteration has some interesting properties

Corollary 1.6.4.

(i) *Policy iteration in finite MDPs converges in finitely many steps*

(ii) V^π *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 fix 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

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

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

(ii) Let $\mathcal{X} = \{x_1, \ldots, x_n\}$ be the finite state set. Then we can interpret $V\colon \mathcal{X} \to \mathbb{R}$ as a vector of $\mathbb{R}^n$ and define

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

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

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 $(k \to k+1)$

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

we get

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

FiXme: Write down Algorithm?

Chapter 2

Reinforcement Learning Algorithms

2.1 Introduction

Dynamic Programming usually breaks down in the real world for two reasons:

1. The transition probabilities and immediate rewards are not known or hard to calculate.

2. The state and action space is too large to even compute one iteration of Dynamic Programming for every state-action tuple (e.g. possible positions and possible moves in every position in chess).

This is where *Reinforcement Learning* algorithms come in, which try to find solutions without having to sweep the entire state space. In this chapter based on Sutton and Barto (2018) we will examine advantages and disadvantages of various algorithms and discuss possible variations and extensions. In the next chapter we will show almost sure convergence of the basic algorithms introduced in this chapter and illustrate their connection to sotchastic approximation.

FiXme: check outline of the plan

We separate their introduction and the convergence proofs, because – while the guarantee of almost sure convergence is reassuring – it is of little use for comparing various algorithms. Since one of the reasons for moving away from Dynamic Programming in the first place was, that we could not calculate the value function for the entire state space within a reasonable time frame. As the size of the state space is often too large for that. Therefore almost sure convergence should not be viewed as more than an entry requirement. For this reason most papers compare algorithms empirically on various example problems. And for some of the more complex algorithms convergence proofs simply do not exist yet.

So since the theoretical convergence properties are usually only ever an afterthought, it is more natural to introduce the various algorithms heuristically, explaining what specific problems they try to address with examples.

But let us ignore the second problem for a moment and consider the case where we only have the first problem (p and r unknown). Then we can learn from a sample $((X_t, A_t, Y_t, R_t), t \in \{0, \dots, T\})$ with

$$(Y_t, R_t) \sim \mathcal{P}(\cdot \mid X_t, A_t)$$

and with $Y_t = X_{t+1}$ if the sample is generated sequentially by a behavior. But there is no reason not to allow this more general sample which might be useful in cases where you can jump around in the statespace and try different transitions at will.

We can then use this batch of transitions to calculate estimators $\hat{p}, \hat{r}$ for the state transitions and immediate rewards $\hat{r}$. And use Dynamic Programming on these estimators.

Algorithm 1 Naive Batch Learning Algorithm

1. Generate the history $(X_t, A_t, Y_t, R_t, t \in \{0, \dots, T\})$

```
for (y, x, a) ∈ X × X × A do                    ▷ initialize variables
    rewards[x, a]←list()
    stateTransitions[y | x, a]← 0
    totalTransitions[x, a]← 0
end for
for t = 0, ..., T do
    rewards[X_t, A_t].append(R_{t+1})
    stateTransitions[Y_t | X_t, A_t]++
    totalTransitions[X_t, A_t]++
end for
for (x, a) ∈ X × A do
    r̂(x, a) ← average(rewards[x,a])
    for y ∈ X do
        p̂(y | x, a) ← stateTransitions[y | x, a]/totalTransitions[x, a]
    end for
end for
```

2. Use Dynamic Programming on $\hat{r}$, $\hat{p}$

FiXme: add code/ ref code Dynamic Programming

FiXme: numerical stability analysis?

If the batch was generated by an exploration policy we separated the *exploration* from the later *exploitation*. The methods which do that are often grouped under the term *Batch Reinforcement Learning* or *off-line* learning.

This method works fine, if the state space is small and one can sample enough observations for every state and action in a reasonable timeframe. But if our state space is too large for that, it is impractical to wait for this.

2.2 Monte Carlo

One idea to tackle larger state spaces is, that it is often unneccessary to know the value function in every state.

To visualize this idea it is useful to imagine the state space to be a room the agent has to navigate.

	γ^2	γ	G	
	γ^3			
	γ^4			
	γ^5			
	S			

The state space $\mathcal{X}$ are the tiles on the floor, the actions $\mathcal{A}$ are the adjacent tiles, where the next state is with probability one equal to the action, and the reward is 0 for every transition but the transition to the goal (G) where the reward is 1 and the game ends. The start state is S. The value of choosing a tile is indicated in grey on the tile.

It is often enough for the agent to know the action value function for the states on his path, without knowing the value of states in the corners. Since he can then follow these "breadcrumbs" to find the goal reliably again. And it will quickly stop walking in circles if it goes into the direction of the highest value next time.

This idea is the basis for Monte Carlo algorithms. To make notation more brief we will introduce the algorithm to calculate the value function, the action value function will be analogous. Consider an episodic MDP and a behavior π then

$$\sum_{t=0}^{\infty} \gamma^t R_{t+1} = \sum_{t=0}^{T} \gamma^t R_{t+1}$$

is a bias free estimator for $V^\pi(X_0)$ for episode lenght T. As the definition was

$$V^\pi(x) = \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t R_{t+1} \,\middle|\, X_0 = x \right]$$

And because of the markov property

$$\sum_{t=k}^{T} \gamma^t R_{t+1}$$

is a bias free estimator for $V^\pi(X_k)$. *First-visit Monte Carlo* uses the return after the first visit of a state $x \in \{X_1, \ldots, X_t\}$ as an estimator for $V^\pi(x)$.

Because of the strong law of large numbers, first-visit Monte Carlo algorithm causes the value function estimation $\hat{V}^\pi(x)$ to converge with probability 1 to

Algorithm 2 First-visit Monte Carlo

for $x \in \mathcal{X}$ **do** ▷ initializing
    Returns$(x) \leftarrow$ list()
    $V^\pi(x) \leftarrow 0$
end for
while true **do** (forever) ▷ learning
    Generate an episode $((X_t, R_{t+1}), t \in \{0, \dots, T\})$ with behavior π
    for $x \in \{X_0, \dots, X_T\}$ **do**
        $k \leftarrow \min\{t \mid X_t = x\}$
        Returns(x).append$(\sum_{t=k}^{T} \gamma^t R_{t+1})$
        $V^\pi(x) \leftarrow$ average(Returns(x))
    end for
end while

$V^\pi(x)$ for every state $x \in \mathcal{X}$ which is visited with positive probability given behavior π and starting distribution X_0.

Every-visit Monte Carlo uses the Return after every visit of a state x to estimate $V^\pi(x)$. Since this means that the tail of the rewards can be included in multiple returns, the returns are not independent from each other anymore which make proving convergence a little bit more difficult than simply applying the strong law of large numbers. But every visit Monte Carlo will turn out to be a special case of $TD(\lambda)$.

FiXme: will TD proof work?

The same method can be applied to learn the action value function Q^π. In this case we take the returns following a state *and* action as estimators for Q^π. But if the model is known and our problem is just a large state space calculating V^π is preferrable, since $|\mathcal{X}| \leq |\mathcal{X} \times \mathcal{A}|$ means that the first algorithm needs fewer observations to converge and Q^π can be calculated with V^π given r and p.

2.2.1 From π to π^*

Let us assume exploring starts

$$\mathbb{P}(X_0 = x) > 0 \qquad \forall x \in \mathcal{X}$$

for a moment. Then Monte Carlo converges whatever policy we select. Similar to policy iteration (1.6.2) we can then alternate between calculating Q^{π_n} and selecting π_{n+1} as a greedy policy with regard to Q^{π_n}. This is referred to as generalized policy iteration (GPI). If we would wait for our Monte Carlo approximation of Q^{π_n} to converge, π_n would converge to π^* for the same reason as policy iteration converges. But remember we are doing Monte Carlo in the first place, because the state space is too large to wait for an evaluation of every state. Which means in practice algorithms alternate beteen choosing a greedy policy with regard to V^{π_n} and generating an episode with policy π_n, averaging

the estimates from this episode with the estimates of $V^{\pi_{n-1}}$. But since we have to assume

$$V^{\pi_{n-1}} \neq V^{\pi_n}$$

in general, using these old estimates is not bias free. It is easy to see that this procedure can only converge to π^* if it converges at all. Since there are only a finite number of deterministic stationary policies it would have to stay constant at some point, but for a constant policy Monte Carlo will converge, and then the policy can only stay constant if it is greedy with regards to its own value function, which forces it to be optimal (1.22). But it is still an open problem whether or not this alternating procedure converges at all (Sutton and Barto 2018, p. 99).

If we remove the assumption of exporing starts, we still need to ensure that every state is visited with positive probability for MC to converge. This requires the policy to do the exploring. There are multiple approaches to exploration which we will discuss later. We will discuss the most straightforward example here, under the assumption that we can calculate Q^{π_n} somehow.

FiXme: check claim

Definition 2.2.1. A stationary policy π is called

soft if it fulfills

$$\pi(a \mid x) > 0 \qquad \forall (x, a) \in \mathcal{X} \times \mathcal{A}$$

ε-soft if it fulfills

$$\pi(a \mid x) > \frac{\varepsilon}{|\mathcal{A}_x|} \quad \forall (x, a) \in \mathcal{X} \times \mathcal{A}$$

ε-greedy with regard to Q, if it selects the greedy action w.r.t. Q with probability $(1 - \varepsilon)$ and a (uniform) random action with probability ε, i.e.

$$\pi(a \mid x) = \begin{cases} (1 - \varepsilon) + \frac{\varepsilon}{|\mathcal{A}_x|} & a \text{ is greedy}^1 \\ \frac{\varepsilon}{|\mathcal{A}_x|} & a \text{ is not greedy} \end{cases}$$

Note that an ε-greedy policy is ε-soft.

An ε-soft policy π^* is called *ε-soft optimal* if

$$V^{\pi^*}(x) = \sup_{\pi \ \varepsilon\text{-soft}} (x) V^{\pi} =: \tilde{V}^*(x)$$

Proposition 2.2.2. *Generalized policy iteration with ε-greedy policies converges to a ε-soft optimal policy*

Proof. To use the same argument as with the standard policy iteration, we would need

$$T^*(V^{\pi_n}) = T^{\pi_{n+1}}(V^{\pi_n})$$

[1] w.l.o.g. only one greedy action

But this is not true. To circumvent this problem we "move the requirement of ε-softness into the environment". I.e. we define an MDP $\tilde{M}$ where

$$\tilde{\mathcal{P}}(\cdot \mid x, a) \coloneqq (1-\varepsilon)\mathcal{P}(\cdot \mid x, a) + \frac{\varepsilon}{|\mathcal{A}_x|} \sum_{b \in \mathcal{A}_x} \mathcal{P}(\cdot \mid x, b)$$

This means, that with probability ε the transition kernel will ignore the selected action and behave as if an uniformly random action was chosen.

We can transform ε-soft policies π from the old MDP to stationary policies $\tilde{\pi}$ of $\tilde{M}$ with a mapping f, where

$$f(\pi)(a \mid x) = \tilde{\pi}(a \mid x) \coloneqq \frac{\pi(a \mid x) - \frac{\varepsilon}{|\mathcal{A}_x|}}{1-\varepsilon} \geq 0$$

And for every stationary policy $\tilde{\pi}$ of $\tilde{M}$ we can define the ε-soft policy π by

$$\pi(a \mid x) \coloneqq (1-\varepsilon)\tilde{\pi}(a \mid x) + \frac{\varepsilon}{|\mathcal{A}_x|}$$

which is the inverse mapping. Therefore f is a bijection between the ε-soft policies in the old MDP and all stationary policies in the new MDP. We now show that the Value functions V^π stay invariant with regard to this mapping. First note that

$$\begin{aligned}
\tilde{p}(y \mid x, a) &= \tilde{\mathcal{P}}(\{y\} \times \mathbb{R} \mid x, a) \\
&= (1-\varepsilon)p(y \mid x, a) + \frac{\varepsilon}{|\mathcal{A}_x|} \sum_{b \in \mathcal{A}_x} p(y \mid x, b)
\end{aligned}$$

and together with $q(B \mid x, a) \coloneqq \mathcal{P}(\mathcal{X} \times B \mid x, a)$ the complementary marginal distribution we can show

$$\begin{aligned}
\tilde{r}(x, a) &= \int t \; d\tilde{q}(t \mid x, a) = \int t \; d \left((1-\varepsilon) q(\cdot \mid x, a) + \frac{\varepsilon}{|\mathcal{A}_x|} \sum_{b \in \mathcal{A}_x} q(\cdot \mid x, b) \right)(t) \\
&= (1-\varepsilon) \int t \; dq(t \mid x, a) + \frac{\varepsilon}{|\mathcal{A}_x|} \sum_{b \in \mathcal{A}_x} \int t \; dq(t \mid x, a) \\
&= (1-\varepsilon) r(x, a) + \frac{\varepsilon}{|\mathcal{A}_x|} \sum_{b \in \mathcal{A}_x} r(x, b)
\end{aligned}$$

Therefore we know

$$\begin{aligned}
&\sum_{a\in\mathcal{A}_x} \tilde{\pi}(a \mid x)\tilde{r}(x,a) \\
&= \sum_{a\in\mathcal{A}_x} \left(\frac{\pi(a \mid x) - \frac{\varepsilon}{|\mathcal{A}_x|}}{1-\varepsilon}\right)\left((1-\varepsilon)r(x,a) + \frac{\varepsilon}{|\mathcal{A}_x|}\sum_{b\in\mathcal{A}_x} r(x,b)\right) \\
&= \sum_{a\in\mathcal{A}_x} \left(\pi(a \mid x) - \frac{\varepsilon}{|\mathcal{A}_x|}\right) r(x,a) + \frac{\varepsilon}{|\mathcal{A}_x|}\sum_{b\in\mathcal{A}_x} r(x,b) \underbrace{\sum_{a\in\mathcal{A}_x} \frac{\pi(a \mid x) - \frac{\varepsilon}{|\mathcal{A}_x|}}{1-\varepsilon}}_{=1} \\
&= \sum_{a\in\mathcal{A}_x} \pi(a \mid x) r(x,a)
\end{aligned}$$

and similarly

$$\begin{aligned}
&\mathbb{P}^{\tilde{\pi}}(\tilde{X}_1 = y \mid \tilde{X}_0 = x) = \sum_{a\in\mathcal{A}_x} \tilde{\pi}(a \mid x)\tilde{p}(y \mid x,a) \\
&= \sum_{a\in\mathcal{A}_x} \left(\frac{\pi(a \mid x) - \frac{\varepsilon}{|\mathcal{A}_x|}}{1-\varepsilon}\right)\left((1-\varepsilon)p(y \mid x,a) + \frac{\varepsilon}{|\mathcal{A}_x|}\sum_{b\in\mathcal{A}_x} p(y \mid x,b)\right) \\
&= \ldots = \sum_{a\in\mathcal{A}_x} \pi(a \mid x) p(y \mid x,a) = \mathbb{P}^{\pi}(X_1 = y \mid X_0 = x)
\end{aligned}$$

The last two equations together imply

$$\begin{aligned}
\tilde{T}^{\tilde{\pi}} V^{\pi}(x) &= \sum_{a\in\mathcal{A}_x} \tilde{\pi}(a \mid x)\tilde{r}(x,a) + \gamma \sum_{y\in\mathcal{X}} \mathbb{P}^{\tilde{\pi}}(\tilde{X}_1 = y \mid X_0 = x) V^{\pi}(y) \\
&= \sum_{a\in\mathcal{A}_x} \pi(a \mid x) r(x,a) + \gamma \sum_{y\in\mathcal{X}} \mathbb{P}^{\pi}(X_1 = y \mid X_0 = x) V^{\pi}(y) \\
&= T^{\pi} V^{\pi}(x) = V^{\pi}(x)
\end{aligned}$$

Since the fixpoint is unique it follows that

$$V^{\tilde{\pi}} = V^{\pi}$$

Since f is bijective this implies

$$\sup_{\tilde{\pi}\in\tilde{\Pi}_S} V^{\tilde{\pi}}(x) = \sup_{\pi \ \varepsilon\text{-soft}} V^{\pi}(x)$$

as well as

$$\begin{aligned} Q^{\tilde{\pi}}(x,a) &= \tilde{r}(x,a) + \gamma \sum_{y \in \mathcal{X}} \tilde{p}(y \mid x,a) V^{\pi}(y) \\ &= (1-\varepsilon) \left(r(x,a) + \gamma \sum_{y \in \mathcal{X}} p(y \mid x,a) V^{\pi}(y) \right) \\ &\qquad + \frac{\varepsilon}{|\mathcal{A}_x|} \sum_{b \in \mathcal{A}_x} \left(r(x,b) + \gamma \sum_{y \in \mathcal{X}} p(y \mid x,b) \right) \\ &= (1-\varepsilon) Q^{\pi}(x,a) + \frac{\varepsilon}{|\mathcal{A}_x|} \sum_{b \in \mathcal{A}_x} \left(r(x,b) + \gamma \sum_{y \in \mathcal{X}} p(y \mid x,b) \right) \end{aligned}$$

Which implies that

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

Threfore greedy with regard to Q^{π} and $Q^{\tilde{\pi}}$ is the same. Let π_n be an ε-soft policy, and let π_{n+1} be ε-greedy with regard to Q^{π_n}. Then $\tilde{\pi}_{n+1} \coloneqq f(\pi_{n+1})$ is greedy w.r.t. $Q^{\tilde{\pi}_n}$

$$\begin{aligned} \tilde{\pi}_{n+1}(a \mid x) = f(\pi_{n+1})(a \mid x) &= \frac{\pi(a \mid x) - \frac{\varepsilon}{|\mathcal{A}_x|}}{1-\varepsilon} \\ &= \begin{cases} \frac{\left((1-\varepsilon) + \frac{\varepsilon}{|\mathcal{A}_x|}\right) - \frac{\varepsilon}{|\mathcal{A}_x|}}{1-\varepsilon} = 1 & a \text{ is greedy} \\ \frac{\frac{\varepsilon}{|\mathcal{A}_x|} - \frac{\varepsilon}{|\mathcal{A}_x|}}{1-\varepsilon} = 0 & a \text{ is not greedy} \end{cases} \end{aligned}$$

This finishes our proof

$$\sup_{\pi \ \varepsilon\text{-soft}} V^{\pi}(x) = \sup_{\tilde{\pi} \in \tilde{\Pi}_S} V^{\tilde{\pi}}(x) = \lim_{n \to \infty} V^{\tilde{\pi}_n}(x) = \lim_{n \to \infty} V^{\pi_n}(x) \qquad \square$$

2.2.2 The Weakness of Monte Carlo

Sutton and Barto provides a very instructive example for the weakness of Monte Carlo. Recall that Monte Carlo tries to estimate V^{π}. So the example assumes a given behavior and tries to evaluate the value of a situation.

Example 2.2.3 (Driving Home)**.** Let us assume that John Doe works in city A and drives to his home in city B after work every day. Since he wants to get home as quickly as possible, the value of the state is antiproportional to the time it will take him from a certain position home. This means that estimating the value of a certain state is equivalent to estimating the time left to drive. The longer John drives this route the better he will get at estimating the time

it will take him to drive certain sections of the road. If there is a delay in an earlier section he has a good estimate for the remaining time once he cleared the obstruction. Now imagine he is driving home from a doctors appointment from city A. As soon as he enters the highway to city B, he is able to use his experience driving this route to estime the remaining time quite precisely.

The Monte Carlo algorithm on the other hand *never* uses existing value estimations to estimate the value of a different state. If John Doe uses Monte Carlo estimates to guess the remaining time from the doctor, the accuracy of his estimates will only ever increase with the times he actually starts driving from the doctor's office.

Since Monte Carlo has more possible "starting points" or generally earlier points in the chain in case of estimating Q^π, it is worse in this case. An extreme example would be two actions in the same state which have the same effect. Even though Monte Carlo might have a good estimate of the value of the first action it will start completely anew for the second action.

FiXme: illustrations?

2.3 Temporal Difference Learning TD

2.4 Mixing Both – The Generalization TD(λ)

2.5 Q-learning

2.6 Exploration

Chapter 3

Stochastic Approximation – Convergence Proofs

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* $\mathbb{P}\left(\biguplus_{n\in\mathbb{N}} B_n\right) = 1$

(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* $\mathbb{P}\left(\biguplus_{n\in\mathbb{N}} B_n\right) = 1$

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)

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

(iii)

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

□

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

$$f\colon \Omega \to \mathbb{R} \text{ injective and measureable,}$$
$$f^{-1}\colon f(\Omega) \to \Omega \text{ measureable.}$$

Then for X Ω-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

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

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

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

"$\Rightarrow$" Let $A \in \mathcal{A}$, then

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

□

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

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

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) "$\Rightarrow$"

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

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)

$$\mathbb{P}(F^{\leftarrow}(U) \leq x) \overset{\text{(i)}}{=} \mathbb{P}(U \leq F(x)) = F(x)$$

□

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.

Sutton, Richard S. and Andrew G. Barto (2018). *Reinforcement Learning: An Introduction*. Second edition. Adaptive Computation and Machine Learning Series. Cambridge, Massachusetts: The MIT Press. 526 pp. ISBN: 978-0-262-03924-6.

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.