University of Mannheim

Bachelor Thesis

Markov-Decision Processes

by

Felix Benning

born on the 27.11.1996 in Nürtingen
matriculation number 1501817

in the
Fakulty for Mathematics in Business and Economics
Supervisor: Prof. Dr. Leif Döring

Due Date: ???

Declaration of Authorship

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

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

City, Date

Signature

Contents

Chapter 1

Markov Decision Processes

1.1 Introduction

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

The MDP model inherits the restriction of Markov Chains to have no memory of past states. We will also not consider changing transition probabilities over time. Rather the transition probabilities will only be influenced by the state and the action.
Both of these limitations could in principle be circumvented by including the time in the state space at the expense of a larger state space. Although it is questionable whether such a construct would yield any interesting results, as then no state is visited twice. So it is of no use to an actor to learn the value of an action in a certain state without further assumptions.

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

1. Resource Management: The state is the resource level

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

the repopulation is the random element.

- Pumped storage Hydro-power: The state is the amount of water in the higher reservoir and the electricity price, the action is to use water to generate electricity or wait for higher prices.
- Beds in a hospital: How many empty beds are needed for emergencies?

2. Stock trading: The state is the price level and stock and liquidity owned.
3. Maintenance: When does a car/road become too expensive to repair?
4. Evacuation in response to flood forecasts

1.2 Model Formulation

Most of the definitions in this chapter are adaptions 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.

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

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

Now we can define a Markov Decision Process

Definition 1.2.2.
$\mathcal{M} = (\mathcal{X}, \mathcal{A}, \mathcal{P}_0)$ is called a *(finite) Markov Decision Process* (MDP). Where:
$\mathcal{X}$ is a countable (finite) set of states.
$\mathcal{A}$ is a countable (finite) set of actions.
And $\mathcal{P}_0\colon (\mathcal{X} \times \mathcal{A}) \times \sigma_{\mathcal{X}\times\mathbb{R}} \to \mathbb{R}$ is a probability kernel.

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

$\mathcal{P}_0$ is called the *transition probability kernel.*

Remark 1.2.3. Some authors include a Time set T in the tuple (e.g. Puterman 2014). Most authors split the transition kernel into a state transition kernel and a reward kernel (e.g. Puterman 2014). But since it is easier to define a marginal distribution from a joint distribution than vice versa, and since this notation is more compact I will stick to the definition from Szepesvári (2010).

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

I can find no explanation for this deviation from the notation of Markov processes. So I offer my own interpretation:
The objective of the theory of MDPs is to find an optimal action selection rule (behavior). And without a fixed behavior the sequence $(X_t, A_t, R_{t+1}, t \in \mathbb{N}_0)$ is undefined, since the $(A_t, t \in \mathbb{N}_0)$ are not defined. But fixing the behavior defeats the purpose of modeling decisions. As it would not make sense to talk about optimal behaviors in an MDP if every behavior creates its own MDP.

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

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

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

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

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

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

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

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

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

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

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

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

$$\begin{aligned}&\mathbb{P}[(X_{t+1}, R_{t+1}) = (x, r) \mid (X_t, A_t) = (x_t, a_t), \dots, (X_0, A_0) = (x_0, a_0)]\\ &= \mathbb{P}[(X_{t+1}, R_{t+1}) = (x, r) \mid (X_t, A_t) = (x_t, a_t)]\end{aligned}$$

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

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

$$\begin{aligned}&\exists (X_{t+1}, R_{t+1})\ \mathcal{X} \times \mathbb{R}\text{-valued random variable :}\\ &(X_{t+1}, R_{t+1}) \sim \mathcal{P}_0(\cdot \mid X_t, A_t) \text{ and satisfies the Markov property}\end{aligned}$$

Proof. □

Remark 1.2.6. A stationary policy induces a *time-homogenous* Markov chain.

Definition 1.2.7. An MDP together with a discount factor $\gamma \in [0, 1]$ is a

discounted reward MDP for $\gamma < 1$

undiscounted reward MDP for $\gamma = 1$

This allows us to define the *return*:

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

Definition 1.2.8. Let $(Y_{(x,a)}, R_{(x,a)}) \sim \mathcal{P}_0(\cdot \mid x, a)$ be a random variable.

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

Remark 1.2.9.

1. From now on we assume that $\forall (x, a) \in \mathcal{X} \times \mathcal{A} : |R_{(x,a)}| \leq R \in \mathbb{R}$ almost surely. This also implies: $\|r\|_\infty = \sup_{(x,a)\in\mathcal{X}\times\mathcal{A}} |\mathbb{E}[R_{(x,a)}]| \leq R$
$$|\mathcal{R}| \leq \sum_{t=0}^{\infty} \gamma^t |R_{t+1}| \leq \frac{R}{1-\gamma} \text{ a.s.}$$

2. Sometimes not all actions make sense in all states. A simple fix would be to set the immediate reward functions for those actions very low, or (if possible) redirect them to the closest possible action.
A more formal approach would be to introduce an additional mapping, which assigns the set of admissible actions to each state $\mathcal{X} \to \mathcal{P}(\mathcal{A})$, or alternatively define a (binary) relation on $\mathcal{X} \times \mathcal{A}$.

3. If there is just one admissible action in every state, the MDP is equivalent to a normal Markov Process.

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

Definition 1.2.10. $\mathcal{M} = (\mathcal{X}, \mathcal{A}, \mathcal{P}_0)$ a MDP
$x \in \mathcal{X}$ is a *terminal (absorbing)* state $:\iff \forall s \in \mathbb{N} : \mathbb{P}(X_{t+s} = x \mid X_t = x) = 1$
An MDP with such states is called *episodic.*
An *episode* is the random time period $(1, \dots, T)$ until a terminal state is reached.

Remark 1.2.11.

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

- Episodic MDPs are often undiscounted

Definition 1.2.12. (Markov Reward Process - MRP)

1.3 Value functions

The goal in this section is to

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

- explore the relation of different value functions

- show uniqueness of optimal value functions with the Banach fixpoint theorem, yielding a simple approximation methode along the way

- demonstrate that in MDPs deterministic stationary policies are generally a large enough set of policies to choose from

Definition 1.3.1. $\mathcal{M} = (\mathcal{X}, \mathcal{A}, \mathcal{P}_0)$ MDP, π Behavior
Select X_0 such that $\forall x \in \mathcal{X} : \mathbb{P}(X_0 = x) > 0$ and evaluate the MDP with $((X_t, A_t, R_{t+1}), t \in \mathbb{N}_0)$ the resulting stoch. process.

$$V^\pi\colon \begin{cases} \mathcal{X} \to \mathbb{R} \\ x \mapsto \mathbb{E}[\mathcal{R} \mid X_0 = x] \end{cases}$$ is the *value function* for π[1]

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

$$V^*\colon \begin{cases} \mathcal{X} \to \mathbb{R} \\ x \mapsto \sup\limits_{\pi \text{ Behav.}} V^\pi(x) \end{cases}$$ is the *optimal value function*

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

π is *optimal* $:\Longleftrightarrow V^* = V^\pi$

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

Definition 1.3.3. $\mathcal{M} = (\mathcal{X}, \mathcal{A}, \mathcal{P}_0)$ MDP
Sometimes we don't care about the probability distribution of the reward, so we define:

$$p\colon \begin{cases} \mathcal{X} \times \mathcal{A} \times \mathcal{P}(\mathcal{X}) \to \mathbb{R} \\ (x,a,Y) \mapsto \mathcal{P}_0(Y \times \mathbb{R} \mid x,a) \end{cases}$$ the <u>state</u> *transition kernel.*

And use the notation $p(y \mid x,a) \coloneqq p(\{y\} \mid x,a)$ with $(x,a,y) \in \mathcal{X} \times \mathcal{A} \times \mathcal{X}$

Proposition 1.3.4. $\mathcal{M} = (\mathcal{X}, \mathcal{A}, \mathcal{P}_0)$ *MDP,* $\pi \in \Pi^{\text{det}}_{\text{stat}}$

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

[1] Well defined because $\mathbb{P}(X_0 = x) > 0$

[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

Proof.

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

$(*)$ Rename: $\tilde{X}_t \coloneqq X_{t+1}, \tilde{A}_t \coloneqq A_{t+1}, \tilde{R}_t \coloneqq R_{t+1}$, 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 □

Corollary 1.3.5. *$\mathcal{M} = (\mathcal{X}, \mathcal{A}, \mathcal{P}_0)$ MDP, $\pi \in \Pi^{\text{det}}_{\text{stat}}$*

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

Proof. Since π is a deterministic stationary policy:

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

The rest follows from 1.3.4 □

Definition 1.3.6. $\mathcal{M} = (\mathcal{X}, \mathcal{A}, \mathcal{P}_0)$ MDP, $\pi \in \Pi^{\text{det}}_{\text{stat}}$
The mapping $T^\pi \colon \mathbb{R}^{\mathcal{X}} \to \mathbb{R}^{\mathcal{X}}$ with:

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

is called the *Bellman Operator*

Remark 1.3.7.

1. $\forall \pi \in \Pi^{\text{det}}_{\text{stat}} : T^\pi V^\pi = V^\pi$ (c.f. 1.3.5)
2. T^π meets the requirements of the Banach fixed-point theorem for $\gamma < 1$, this implies that V^π for $\pi \in \Pi^{\text{det}}_{\text{stat}}$ is a *unique* fixpoint and can be approximated with the canonical iteration

3. T^π is an affine operator

4. $W_1, W_2 \in \mathbb{R}^{\mathcal{X}}$, write $W_1 \leq W_2$ for $\forall x \in \mathcal{X} : W_1(x) \leq W_2(x)$, then:

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

Proof. 2. $(\mathbb{R}^{\mathcal{X}}, \|\cdot\|_\infty)$ 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 $V, W \in \mathbb{R}^{\mathcal{X}}$:

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

4. Be $W_1, W_2 \in \mathbb{R}^{\mathcal{X}}$, $W_1 \leq W_2$ and $x \in \mathcal{X}$:

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

□

Definition 1.3.8. $\mathcal{M} = (\mathcal{X}, \mathcal{A}, \mathcal{P}_0)$ MDP

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

Definition 1.3.9. $\mathcal{M} = (\mathcal{X}, \mathcal{A}, \mathcal{P}_0)$ MDP
The mapping $T^* \colon \mathbb{R}^{\mathcal{X}} \to \mathbb{R}^{\mathcal{X}}$ with:

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

is called the *Bellman Optimality Operator*

Lemma 1.3.10. *$\mathcal{M} = (\mathcal{X}, \mathcal{A}, \mathcal{P}_0)$ MDP*

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

(ii) $V^*(x) = \sup_{a \in \mathcal{A}} Q^*(x, a)$

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

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

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

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

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

Define a slightly changed deterministic policy with this π, a:

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

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

$$\begin{aligned}
W_1(y) &= r(y,\hat{\pi}(y)) + \gamma \sum_{z\in\mathcal{X}} p(z \mid y,\hat{\pi}(y)) V^\pi(z) \\
&= \begin{cases} r(y,\pi(y)) + \gamma \sum_{z\in\mathcal{X}} p(z \mid y,\pi(y)) V^\pi(z) = V^\pi(x) & y \neq x \\ r(x,a) + \gamma \sum_{z\in\mathcal{X}} p(z \mid x,a) V^\pi(z) > \tilde{V}(x) & y = x \end{cases} \\
&\geq V^\pi(y) = W_0(y)
\end{aligned}$$

By induction with 1.3.7 (4.): $W_{n+1} = T^{\hat{\pi}} W_n \geq T^{\hat{\pi}} W_{n-1} = W_n$, thus:

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

□

Corollary 1.3.11.

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

Proof.

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

$\tilde{V}$ analogous □

Theorem 1.3.12. *$\mathcal{M} = (\mathcal{X}, \mathcal{A}, \mathcal{P}_0)$ MDP*
T^* satisfies the requirements of the Banach fixpoint theorem, in particular:

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

is the unique fixpoint of T^*

Lemma 1.3.13. *(Blackwell's condition for contraction)*

Proof. https://math.stackexchange.com/questions/1087885/blackwells-condition-for-a-contraction-why-is-boundedness-neccessary?rq=1 □

Proof (Theorem). □

Proposition 1.3.14. *$\mathcal{M} = (\mathcal{X}, \mathcal{A}, \mathcal{P}_0)$ MDP*
The following statements are equivalent:

(i) $\pi \in \Pi_{\text{stat}}$ *is optimal ($V^* = V^\pi$)*

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

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

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

Proof. □

Definition 1.3.15. $Q\colon \mathcal{X} \times \mathcal{A} \to \mathbb{R}$ an action value function, $\tilde{\pi}\colon \mathcal{X} \to \mathcal{A}$ with:

$$\tilde{\pi}(x) \coloneqq \arg\max_{\pi\in\Pi_{\text{stat}}} \sum_{a\in\mathcal{A}} \pi(a \mid x) Q(x,a) \qquad x \in \mathcal{X}$$

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

Remark 1.3.16.

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

Chapter 2

Title Chapter 2

Bibliography

Puterman, Martin L. (Aug. 28, 2014). *Markov Decision Processes: Discrete Stochastic Dynamic Programming.* John Wiley & Sons. 615 pp. ISBN: 978-1-118-62587-3.

Szepesvári, Csaba (Jan. 1, 2010). “Algorithms for Reinforcement Learning”. In: *Synthesis Lectures on Artificial Intelligence and Machine Learning* 4.1, pp. 1–103. ISSN: 1939-4608. DOI: `10.2200/S00268ED1V01Y201005AIM009`. URL: `https://www.morganclaypool.com/doi/abs/10.2200/S00268ED1V01Y201005AIM009` (visited on 02/06/2019).

White, Douglas J. (Dec. 1985). “Real Applications of Markov Decision Processes”. In: *Interfaces* 15.6, pp. 73–83. ISSN: 0092-2102, 1526-551X. DOI: `10.1287/inte.15.6.73`. URL: `http://pubsonline.informs.org/doi/abs/10.1287/inte.15.6.73` (visited on 02/05/2019).