Master MVA: Reinforcement Learning **Lecture: 2**

Markov Decision Processes and Dynamic Programming

Lecturer: *Alessandro Lazaric* *http://researchers.lille.inria.fr/∼lazaric/Webpage/Teaching.html*

Objectives of the lecture

1. **Understand:** Markov decision processes, Bellman equations and Bellman operators.
2. **Use:** dynamic programming algorithms.

1 The Markov Decision Process

1.1 Definitions

Definition 1 (Markov chain)**.** *Let the state space X be a bounded compact subset of the Euclidean space, the discrete-time dynamic system $(x_t)_{t\in\mathbb{N}} \in X$ is a Markov chain if*

$$\mathbb{P}(x_{t+1} = x \,|\, x_t, x_{t-1}, \dots, x_0) = \mathbb{P}(x_{t+1} = x \,|\, x_t), \tag{1}$$

*so that all the information needed to predict (in probability) the future is contained in the current state (**Markov property**). Given an initial state $x_0 \in X$, a Markov chain is defined by the **transition probability** p such that*

$$p(y|x) = \mathbb{P}(x_{t+1} = y | x_t = x). \tag{2}$$

Remark: notice that in some cases we can turn a higher-order Markov process into a Markov process by including the past as a new state variable. For instance, in the control of an inverted pendulum, the state that can be observed is only the angular position θ_t. In this case the system is non-Markov since the next position depends on the previous position but also on the angular speed, which is defined as $d\theta_t = \theta_t - \theta_{t-1}$ (as a first approximation). Thus, if we define the state space as $x_t = (\theta_t, d\theta_t)$ we obtain a Markov chain.

Definition 2 (Markov decision process [Bellman, 1957, Howard, 1960, Fleming and Rishel, 1975, Puterman, 1994, Bertsekas and Tsitsiklis, 1996])**.** *A **Markov decision process** is defined as a tuple $M = (X, A, p, r)$ where*

- *X is the **state** space (finite, countable, continuous),*[1]
- *A is the **action** space (finite, countable, continuous),*

[1] In most of our lectures it can be consider as finite such that $|X| = N$.

- $p(y|x,a)$ *is the* ***transition probability*** *(i.e., environment dynamics) such that for any* $x \in X$, $y \in X$, *and* $a \in A$
$$p(y|x,a) = \mathbb{P}(x_{t+1} = y | x_t = x, a_t = a),$$
is the probability of observing a next state y *when action* a *is taking in* x,
- $r(x,a,y)$ *is the* ***reinforcement*** *obtained when taking action* a*, a transition from a state* x *to a state* y *is observed.*[2]

Definition 3 (Policy)**.** *At time* $t \in \mathbb{N}$ *a* ***decision rule*** π_t *is a mapping from states to actions, in particular it can be*

- *Deterministic:* $\pi_t : X \to A$, $\pi_t(x)$ *denotes the action chosen at state* x *at time* t,
- *Stochastic:* $\pi_t : X \to \Delta(A)$, $\pi_t(a|x)$ *denotes the probability of taking action* a *at state* x *at time* t.

A ***policy*** *(strategy, plan) is a sequence of decision rules. In particular, we have*

- *Non-stationary:* $\pi = (\pi_0, \pi_1, \pi_2, \dots)$,
- *Stationary (Markovian):* $\pi = (\pi, \pi, \pi, \dots)$.

Remark: an MDP M together with a deterministic (stochastic) stationary policy π forms a dynamic process $(x_t)_{t\in\mathbb{N}}$ (obtained by taking the actions $a_t = \pi(x_t)$) which corresponds to a Markov chain of state X and transition probability $p(y|x) = p(y|x,\pi(x))$.

Time horizons

- *Finite time horizon* T: the problem is characterized by a deadline at time T (e.g., the end of the course) and the agent only focuses on the sum of the rewards up to that time.
- *Infinite time horizon with discount*: the problem never terminates but rewards which are closer in time receive a higher importance.
- *Infinite time horizon with absorbing (terminal) state*: the problem never terminates but the agent will eventually reach a termination state.
- *Infinite time horizon with average reward*: the problem never terminates but the agent only focuses on the average of the rewards.

Definition 4 (The state value function)**.** *Depending on the time horizon we consider we have*

- *Finite time horizon* T
$$V^\pi(t,x) = \mathbb{E}\big[\sum_{s=t}^{T-1} r(x_s, \pi_s(x_s)) + R(x_T) |\, x_t = x; \pi\big], \tag{3}$$
where R *is a reward function for the final state.*

[2] Most of the time we will use either $r(x,a)$ (which can be considered as the expected value of $r(x,a,y)$) or $r(x)$ as a state-only function.

- *Infinite time horizon with discount*

$$V^\pi(x) = \mathbb{E}\big[\sum_{t=0}^{\infty} \gamma^t r(x_t, \pi(x_t)) \,|\, x_0 = x; \pi\big], \tag{4}$$

  where $0 \leq \gamma < 1$ is a discount factor (i.e., small *= focus on short-term rewards,* big *= focus on long term rewards).*[3]

- *Infinite time horizon with absorbing (terminal) states*

$$V^\pi(x) = \mathbb{E}\big[\sum_{t=0}^{T} r(x_t, \pi(x_t)) | x_0 = x; \pi\big], \tag{5}$$

  where T is the first (random) time when the agent achieves a absorbing state.

- *Infinite time horizon with average reward*

$$V^\pi(x) = \lim_{T\to\infty} \frac{1}{T}\mathbb{E}\big[\sum_{t=0}^{T-1} r(x_t, \pi(x_t)) \,|\, x_0 = x; \pi\big]. \tag{6}$$

Remark: the previous expectations refer to all the possible stochastic trajectories. More precisely, let us consider an MDP $M = (X, A, p, r)$, if an agent follows a non-stationary policy π from the state x_0, then it observes the random sequence of states $(x_0, x_1, x_2, \ldots)$ and the corresponding sequence of rewards $(r_0, r_1, r_2, \ldots)$ where $r_t = r(x_t, \pi_t(x_t))$. The corresponding value function (in the infinite horizon with discount setting) is then defined as

$$V^\pi(x) = \mathbb{E}_{(x_1, x_2, \ldots)}\Big[\sum_{t=0}^{\infty} \gamma^t r(x_t, \pi(x_t)) \,|\, x_0 = x; \pi\Big],$$

where each $x_t \sim p(\cdot|x_{t-1}, a_t = \pi(x_t))$ is a random realization from the transition probability of the MDP.

Definition 5 (Optimal policy and optimal value function)**.** *The solution to an MDP is an optimal policy π^* satisfying*

$$\pi^* \in \arg\max_{\pi\in\Pi} V^\pi$$

in all the states $x \in X$, where Π is some policy set of interest. The corresponding value function is the optimal value function $V^* = V^{\pi^*}$.

Remark: $\pi^* \in \arg\max(\cdot)$ and not $\pi^* = \arg\max(\cdot)$ because an MDP may admit more than one optimal policy.

Beside the state value function, we can also introduce an alternative formulation, the state-action value function. For infinite horizon discounted problems we have the following definition (similar for other settings).

Definition 6 (The state-action value function)**.** *In discounted infinite horizon problems, for any policy π, the state-action value function (or Q-function) $Q^\pi : X \times A \mapsto \mathbb{R}$ is defined as*

$$Q^\pi(x, a) = \mathbb{E}\Big[\sum_{t\geq 0} \gamma^t r(x_t, a_t) | x_0 = x, a_0 = a, a_t = \pi(x_t), \forall t \geq 1\Big],$$

and the corresponding optimal Q-function is

$$Q^*(x, a) = \max_\pi Q^\pi(x, a).$$

[3] Mathematical interpretation: for any $\gamma \in [0, 1)$ the series always converge (for bounded rewards).

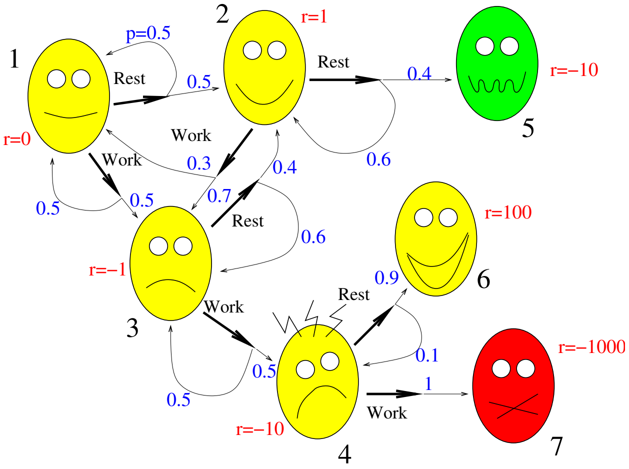
The relationships between the V-function and the Q-function are:

$$\begin{aligned}
Q^\pi(x,a) &= r(x,a) + \gamma \sum_{y \in X} p(y|x,a) V^\pi(y) \\
V^\pi(x) &= Q^\pi(x, \pi(x)) \\
Q^*(x,a) &= r(x,a) + \gamma \sum_{y \in X} p(y|x,a) V^*(y) \\
V^*(x) &= Q^*(x, \pi^*(x)) = \max_{a \in A} Q^*(x,a).
\end{aligned}$$

Finally, from an optimal Q-function we can deduce the optimal policy as $\pi^*(x) \in \arg\max_{a \in A} Q^*(x,a)$.

1.2 Examples

Example 1 *(the MVA student dilemma).*



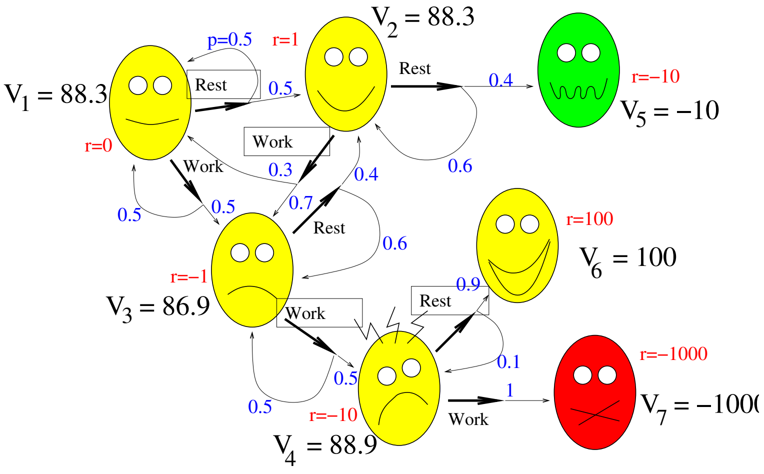
- *Model*: all the transitions are Markov since the probability of transition only depend on the previous state. The states x_5, x_6, x_7 are terminal (absorbing) states.

- *Setting*: infinite horizon with terminal states.

- *Objective*: find the policy that maximizes the expected sum of rewards before achieving a terminal state.

- *Problem 1*: if a student knows the transition probabilities and the rewards, how does he/she compute the optimal strategy?

Solution:



$$
\begin{aligned}
&V_5 = -10, V_6 = 100, V_7 = -1000\\
&V_4 = -10 + 0.9V_6 + 0.1V_4 \simeq 88.9\\
&V_3 = -1 + 0.5V_4 + 0.5V_3 \simeq 86.9\\
&V_2 = 1 + 0.7V_3 + 0.3V_1\\
&V_1 = \max\{0.5V_2 + 0.5V_1, 0.5V_3 + 0.5V_1\}\\
&V_1 = V_2 = 88.3
\end{aligned}
$$

- *Problem 2*: what if the student doesn't know the MDP?

Example 2 *(The retail store management problem).*

At each month t, a store contains x_t items of a specific goods and the demand for that goods is D_t. At the end of each month the manager of the store can order a_t more items from his supplier. Furthermore we know that

- The cost of maintaining an inventory of x is $h(x)$.
- The cost to order a items is $C(a)$.
- The income for selling q items is $f(q)$.
- If the demand D is bigger than the available inventory x, the customers that cannot be served will leave and move to another store.
- The income of the remaining inventory at the end of the year is $g(x)$.
- Constraint: the store has a maximum capacity M.
- **Objective**: the performance of the manager is evaluated as the profit obtained over an horizon of a year ($T = 12$ months).

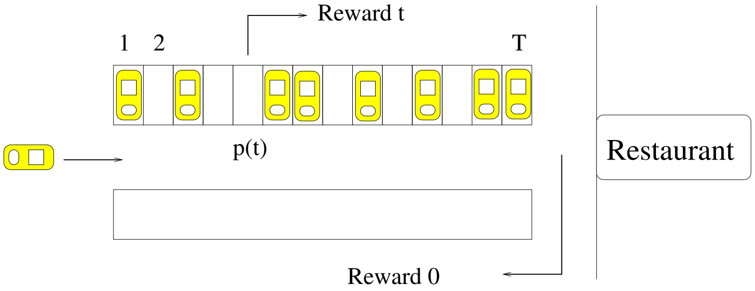
Problem: How do we formalize the problem?

Solution: we define the following simplified model using the MDP formalism.

- *State space*: $x \in X = \{0, 1, \dots, M\}$.
- *Action space*: it is not possible to order more items that the capacity of the store, then the action space should depend on the current state. Formally, at state$x$, $a \in A(x) = \{0, 1, \dots, M - x\}$.
- *Dynamics*: $x_{t+1} = [x_t + a_t - D_t]^+$.
  Problem: the dynamics should be Markov and stationary. The demand D_t is stochastic and time-independent. Formally, $D_t \overset{i.i.d.}{\sim} \mathcal{D}$.
- *Reward*: $r_t = -C(a_t) - h(x_t + a_t) + f([x_t + a_t - x_{t+1}]^+)$.
- *Objective function*: $\mathbb{E}\big[\sum_{t=1}^{T-1} r_t + g(x_T)\big]$

Example 3 *(The parking problem).*

A driver wants to park his car as close as possible to the restaurant.



We know that

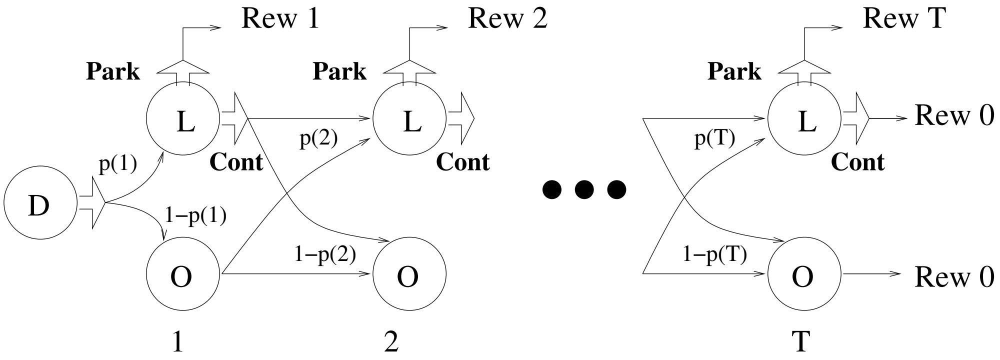
- The driver cannot see whether a place is available unless he/she is in front of it.
- There are P places.
- At each place i the driver can either move to the next place or park (if the place is available).
- The closer to the restaurant the parking, the higher the satisfaction.
- If the driver doesn't park anywhere, then he/she leaves the restaurant and has to find another one.
- **Objective**: maximize the satisfaction.

Problem: How do we formalize the problem?

Solution: we define the following simplified model using the MDP formalism.

- *State space*: $x \in X = \{(1, T), (1, A), (2, T), (2, A), \dots, (P, T), (P, A), \text{parked}, \text{left}\}$. Each of the P places can be (A)vailable or (T)aken. Whenever the driver parks, he reaches the state *parked*, while if he never parks then the state becomes *left*. Both these states are *terminal* states.
- *Action space*: $A(x) = \{\text{park}, \text{continue}\}$ if $x = (\cdot, A)$ or $A(x) = \{\text{continue}\}$ if $x = (\cdot, T)$.
- *Dynamics*: (see graph). We assume that the probability of a place i to be available is $\rho(i)$.
- *Reward*: for any $i \in [1, P]$ we have $r(x, \text{park}, \text{parked}) = i$ and 0 otherwise.

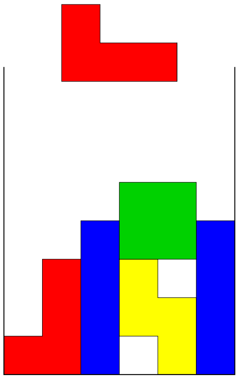
- *Objective function*: $\mathbb{E}\big[\sum_{t=1}^{T} r_t\big]$ with T the random time when a terminal state is reached.



Example 4 *(The tetris game).*

Model:

- *State space*: configuration of the wall and next piece and terminal state when the well reach the maximum height.
- *Action space*: position and orientation of the current space in the wall.
- *Dynamics*: new configuration of the well and new *random* piece.
- *Reward*: number of deleted rows.
- *Objective function*: $\mathbb{E}\big[\sum_{t=1}^{T} r_t\big]$ with T the random time when a terminal state is reached. (*remark:* it has been proved that the game eventually terminates with probability 1 for any playing strategy).



Problem: Compute the optimal strategy.

Solution: unknown! The state space is huge! $|X| = 10^{61}$ for a problem with maximum height 20, width 10 and 7 different pieces.

1.3 Finite Horizon Problems

Proposition 1. For any horizon T and a (non-stationary) policy $\pi = (\pi_0, \ldots, \pi_{T-1})$, the state value function at a state $x \in X$ at time $t \in \{0, \ldots, T\}$ satisfies the **Bellman equation**:

$$V^{\pi}(t,x) = \mathbb{E}\big[\sum_{s=t}^{T-1} r(x_s, \pi_s(x_s)) + R(x_T) \,|\, x_t = x; \pi\big]. \qquad (7)$$

Proof. Recalling equation 3 and Def. 5, we have that

$$
\begin{aligned}
V^\pi(t,x) &= \mathbb{E}\big[\sum_{s=t}^{T-1} r(x_s,\pi_s(x_s)) + R(x_T)\,|\,x_t=x;\pi\big] \\
&\overset{(a)}{=} r(x,\pi_t(x)) + \mathbb{E}_{x_{t+1},x_{t+2},\dots,x_{T-1}}\Bigg[\sum_{s=t+1}^{T-1} r(x_s,\pi_s(x_s)) + R(x_T)\,|\,x_t=x;\pi\Bigg] \\
&= r(x,\pi_t(x)) + \sum_{y\in X}\mathbb{P}[x_{t+1}=y|x_t=x;a_t=\pi(x_t)]\mathbb{E}_{x_{t+2},\dots,x_{T-1}}\Bigg[\sum_{s=t+1}^{T-1} r(x_s,\pi_s(x_s)) + R(x_T)\,|\,x_{t+1}=y;\pi\Bigg] \\
&\overset{(b)}{=} r(x,\pi_t(x)) + \sum_{y\in X} p(y|x,\pi(x))\mathbb{E}\big[\sum_{s=t+1}^{T-1} r(x_s,\pi_s(x_s)) + R(x_T)\,|\,x_{t+1}=x;\pi\big] \\
&\overset{(c)}{=} r(x,\pi_t(x)) + \sum_{y\in X} p(y|x,\pi(x))V^\pi(t+1,y),
\end{aligned}
$$

where

(a) The expectation is conditioned on $x_t = x$.

(b) Application of the law of total expectation.

(c) Definition of the MDP dynamics p and of the value function.

□

Bellman's Principle of Optimality [Bellman, 1957]:

> "An optimal policy has the property that, whatever the initial state and the initial decision are, the remaining decisions must constitute an optimal policy with regard to the state resulting from the first decision."

Proposition 2. The optimal value function $V^*(t,x)$ (i.e., $V^* = \max_\pi V^\pi$) is the solution to the **optimal Bellman equation**:

$$
\begin{aligned}
V^*(t,x) &= \max_{a\in A}\big[r(x,a) + \sum_{y\in X} p(y|x,a)V^*(t+1,y)\big], \text{ with } 0\le t<T \qquad (8)\\
V^*(T,x) &= R(x),
\end{aligned}
$$

and the policy

$$\pi_t^*(x) \in \arg\max_{a\in A}\big[r(x,a) + \sum_{y\in X} p(y|x,a)V^*(t+1,y)\big], \text{ with } 0\le t<T.$$

is an optimal policy.

Proof. By definition we have $V^*(T,x) = R(x)$, then V^* is defined by backward propagation for any $t < T$. Any policy π applied from time $t < T$ on at state x can be written as $\pi = (a, \pi')$ with $a \in A$ is the action taken at time t in x and $\pi' = (\pi_{t+1}, \dots, \pi_{T-1})$. Then we can show that

$$\begin{aligned}
V^*(t,x) &\overset{(a)}{=} \max_\pi \mathbb{E}\big[\sum_{s=t}^{T-1} r(x_s, \pi_s(x_s)) + R(x_T) | x_t = x; \pi\big] \\
&\overset{(b)}{=} \max_{(a,\pi')} \Big[r(x,a) + \sum_{y \in X} p(y|x,a) V^{\pi'}(t+1,y)\Big] \\
&\overset{(c)}{=} \max_{a \in A} \Big[r(x,a) + \sum_{y \in X} p(y|x,a) \max_{\pi'} V^{\pi'}(t+1,y)\Big] \\
&\overset{(d)}{=} \max_{a \in A} \big[r(x,a) + \sum_{y \in X} p(y|x,a) V^*(t+1,y)\big].
\end{aligned}$$

where

- (a) : definition of value function from equation 3,
- (b) : decomposition of policy $\pi = (a, \pi')$ and recursive definition of the value function,
- (c) : follows from
  - Trivial inequality

$$\max_{\pi'} \sum_y p(y|x,a) V^{\pi'}(t+1,y) \leq \sum_y p(y|x,a) \max_{\pi'} V^{\pi'}(t+1,y)$$

  - Let $\bar{\pi} = (\bar{\pi}_{t+1}, \dots)$ a policy such that $\bar{\pi}_{t+1}(y) = \arg\max_{b \in A} \max_{(\pi_{t+2},\dots)} V^{(b,\pi_{t+2},\dots)}(t+1,y)$. Then

$$\sum_y p(y|x,a) \max_{\pi'} V^{\pi'}(t+1,y) = \sum_y p(y|x,a) V^{\bar{\pi}}(t+1,y) \leq \max_{\pi'} \sum_y p(y|x,a) V^{\pi'}(t+1,y).$$

- (d) : definition of optimal value function.

Finally the optimal policy simply takes the action for which the maximum is attained at each iteration. □

Remark. The previous Bellman equations can be easily extended to the case of state-action value functions.

Parking example. Let $V^*(p,A)$ (resp. $V^*(p,T)$) the optimal value function when at position p and the place is *available* (resp. *taken*). The optimal value function can be constructed using the optimal Bellman equations as:

$$\begin{aligned}
V^*(P,A) &= \max\{P, 0\} = P, \quad V^*(P,T) = 0, \\
V^*(P-1,A) &= \max\{P-1, \rho(P)V^*(P,A) + (1-\rho(P))V^*(P,T))\} \\
V^*(p,A) &= \max\{t, \rho(p+1)V^*(p+1,A) + (1-\rho(p+1))V^*(p+1,T)\},
\end{aligned}$$

where the max corresponds to the two possible choices (park or continue) and the corresponding optimal policy can be derived as the arg max of the previous maxima.

1.4 Discounted Infinite Horizon Problems

1.4.1 Bellman Equations

Proposition 3. For any stationary policy $\pi = (\pi, \pi, \dots)$, the state value function at a state $x \in X$ satisfies the **Bellman equation**:

$$V^\pi(x) = r(x, \pi(x)) + \gamma \sum_y p(y|x, \pi(x)) V^\pi(y). \tag{9}$$

Proof. For any policy π,

$$\begin{aligned} V^\pi(x) &= \mathbb{E}\big[\sum_{t\geq 0} \gamma^t r(x_t, \pi(x_t)) \,|\, x_0 = x; \pi\big] \\ &= r(x, \pi(x)) + \mathbb{E}\big[\sum_{t\geq 1} \gamma^t r(x_t, \pi(x_t)) \,|\, x_0 = x; \pi\big] \\ &= r(x, \pi(x)) + \gamma \sum_y \mathbb{P}(x_1 = y \,|\, x_0 = x; \pi(x_0)) \mathbb{E}\big[\sum_{t\geq 1} \gamma^{t-1} r(x_t, \pi(x_t)) \,|\, x_1 = y; \pi\big] \\ &= r(x, \pi(x)) + \gamma \sum_y p(y|x, \pi(x)) V^\pi(y). \end{aligned}$$

□

Proposition 4. The optimal value function V^* (i.e., $V^* = \max_\pi V^\pi$) is the solution to the **optimal Bellman equation**:

$$V^*(x) = \max_{a \in A} \big[r(x, a) + \gamma \sum_y p(y|x, a) V^*(y) \big]. \tag{10}$$

Proof. For any policy $\pi = (a, \pi')$ (possibly non-stationary),

$$\begin{aligned} V^*(x) &\overset{(a)}{=} \max_\pi \mathbb{E}\big[\sum_{t\geq 0} \gamma^t r(x_t, \pi(x_t)) \,|\, x_0 = x; \pi\big] \\ &\overset{(b)}{=} \max_{(a,\pi')} \Big[r(x, a) + \gamma \sum_y p(y|x, a) V^{\pi'}(y) \Big] \\ &\overset{(c)}{=} \max_a \Big[r(x, a) + \gamma \sum_y p(y|x, a) \max_{\pi'} V^{\pi'}(y) \Big] \\ &\overset{(d)}{=} \max_a \Big[r(x, a) + \gamma \sum_y p(y|x, a) V^*(y) \Big]. \end{aligned} \tag{11}$$

where

(a) : definition of value function from equation 4,

(b) : decomposition of policy $\pi = (a, \pi')$ and recursive definition of the value function,

(c) : follows from

- Trivial inequality

$$\max_{\pi'} \sum_y p(y|x,a) V^{\pi'}(y) \leq \sum_y p(y|x,a) \max_{\pi'} V^{\pi'}(y)$$

- Let $\bar{\pi}(y) = \arg\max_{\pi'} V^{\pi'}(y)$. Then

$$\sum_y p(y|x,a) \max_{\pi'} V^{\pi'}(y) = \sum_y p(y|x,a) V^{\bar{\pi}}(y) \leq \max_{\pi'} \sum_y p(y|x,a) V^{\pi'}(y).$$

(d) : definition of optimal value function.

□

The equivalent Bellman equations for state-action value functions are:

$$\begin{aligned} Q^\pi(x,a) &= r(x,a) + \gamma \sum_{y \in X} p(y|x,a) Q^\pi(y, \pi(y)) \\ Q^*(x,a) &= r(x,a) + \gamma \sum_{y \in X} p(y|x,a) \max_{b \in A} Q^*(y,b). \end{aligned}$$

1.4.2 Bellman Operators

Notation. W.l.o.g. from this moment on we consider a discrete state space $|X| = N$, so that for any policy π, V^π is a vector of size N (i.e., $V^\pi \in \mathbb{R}^N$).

Definition 7. *For any $W \in \mathbb{R}^N$, the **Bellman operator** $\mathcal{T}^\pi : \mathbb{R}^N \to \mathbb{R}^N$ is defined as*

$$\mathcal{T}^\pi W(x) = r(x, \pi(x)) + \gamma \sum_y p(y|x, \pi(x)) W(y), \tag{12}$$

*and the **optimal Bellman operator** (or dynamic programming operator) is defined as*

$$\mathcal{T} W(x) = \max_{a \in A} \big[r(x,a) + \gamma \sum_y p(y|x,a) W(y) \big]. \tag{13}$$

Proposition 5. The Bellman operators enjoy a number of properties:

1. *Monotonicity*: for any $W_1, W_2 \in \mathbb{R}^N$, if $W_1 \leq W_2$ component-wise, then

$$\begin{aligned} \mathcal{T}^\pi W_1 &\leq \mathcal{T}^\pi W_2, \\ \mathcal{T} W_1 &\leq \mathcal{T} W_2. \end{aligned}$$

2. *Offset*: for any scalar $c \in \mathbb{R}$,

$$\begin{aligned} \mathcal{T}^\pi (W + c I_N) &= \mathcal{T}^\pi W + \gamma c I_N, \\ \mathcal{T} (W + c I_N) &= \mathcal{T} W + \gamma c I_N, \end{aligned}$$

3. *Contraction in L_∞-norm*: for any $W_1, W_2 \in \mathbb{R}^N$

$$\begin{aligned} ||\mathcal{T}^\pi W_1 - \mathcal{T}^\pi W_2||_\infty &\leq \gamma ||W_1 - W_2||_\infty, \\ ||\mathcal{T} W_1 - \mathcal{T} W_2||_\infty &\leq \gamma ||W_1 - W_2||_\infty. \end{aligned}$$

4. *Fixed point*: For any policy π

$$\begin{aligned} &V^\pi \text{ is the unique fixed point of } \mathcal{T}^\pi, \\ &V^* \text{ is the unique fixed point of } \mathcal{T}. \end{aligned}$$

Furthermore for any $W \in \mathbb{R}^N$ and any stationary policy π

$$\begin{aligned} \lim_{k\to\infty} (\mathcal{T}^\pi)^k W &= V^\pi, \\ \lim_{k\to\infty} (\mathcal{T})^k W &= V^*. \end{aligned}$$

Proof. Monotonicity (1) and offset (2) directly follow from the definitions.
The contraction property (3) holds since for any $x \in X$ we have

$$\begin{aligned} |\mathcal{T}W_1(x) - \mathcal{T}W_2(x)| &= \Big| \max_a \big[r(x,a) + \gamma \sum_y p(y|x,a) W_1(y) \big] - \max_{a'} \big[r(x,a') + \gamma \sum_y p(y|x,a') W_2(y) \big] \Big| \\ &\overset{(a)}{\leq} \max_a \Big| \big[r(x,a) + \gamma \sum_y p(y|x,a) W_1(y) \big] - \big[r(x,a) + \gamma \sum_y p(y|x,a) W_2(y) \big] \Big| \\ &= \gamma \max_a \sum_y p(y|x,a) |W_1(y) - W_2(y)| \\ &\leq \gamma ||W_1 - W_2||_\infty \max_a \sum_y p(y|x,a) = \gamma ||W_1 - W_2||_\infty, \end{aligned}$$

where in (a) we used $\max_a f(a) - \max_{a'} g(a') \leq \max_a (f(a) - g(a))$.

The fixed point property follows from the Bellman equations (eq. 9 and 10) and Banach fixed point theorem (see Theorem 11). The property regarding the convergence of the limit of the Bellman operators is actually used in the proof of the Banach theorem. □

Remark 1 (optimal policy): Any stationary policy $\pi^*(x) \in \arg\max_{a\in A} \big[r(x,a) + \gamma \sum_y p(y|x,a) V^*(y) \big]$ is optimal. In fact, from the definition of π^* we have that $\mathcal{T}^{\pi^*} V^* = \mathcal{T} V^* = V^*$ where the first equality follows from the fact that the optimal Bellman operator coincides with the action taken by π^* and the second from the fixed point property of $\mathcal{T}$. Furthermore, by the fixed point property of $\mathcal{T}^\pi$ we have that V^{π^*} is the fixed point of $\mathcal{T}^{\pi^*}$ which is also unique. Then $V^{\pi^*} = V^*$ thus implying that π^* is an optimal policy.

Remark 2 (value/policy iteration): most of the dynamic programming algorithms studied in Section 2 will heavily rely on property (4).

Bellman operators for Q-functions.

Both the Bellman $\mathcal{T}^\pi$ and the optimal Bellman operators $\mathcal{T}$ can be extended to Q-functions. Thus for any function[4] $W : X \times A \to \mathbb{R}$, we have

$$\begin{aligned} \mathcal{T}^\pi W(x,a) &= r(x,a) + \gamma \sum_{y \in X} p(y|x,a) W(y, \pi(y)), \\ \mathcal{T} W(x,a) &= r(x,a) + \gamma \sum_{y \in X} p(y|x,a) \max_{b \in A} W(y,b). \end{aligned}$$

This allows to extend also all the properties in Proposition 5, notably that Q^π (resp. Q^*) is the fixed point of $\mathcal{T}^\pi$ (resp. $\mathcal{T}$).

1.5 Undiscounted Infinite Horizon Problems

1.5.1 Bellman equations

Recall the definition of value function for infinite horizon problems with absorbing states in eq. 5:

$$V^\pi(x) = \mathbb{E}\big[\sum_{t=0}^{\infty} r(x_t, \pi(x_t)) | x_0 = x; \pi\big] = \mathbb{E}\big[\sum_{t=0}^{T} r(x_t, \pi(x_t)) | x_0 = x; \pi\big],$$

where T is the first (random) time when the agent achieves a absorbing state. The equivalence follows from the fact that we assume that once the absorbing state is achieved the system stays in that state indefinitely long with a constant reward of 0.

Definition 8. *A stationary policy π is **proper** if there exists an integer $n \in \mathbb{N}$ such that from any initial state $x \in X$ the probability of achieving the terminal state (denoted by $\bar{x}$) after n steps is strictly positive. That is*

$$\rho_\pi = \max_x \mathbb{P}(x_n \neq \bar{x} \,|\, x_0 = x, \pi) < 1.$$

Proposition 6. For any proper policy π with parameter ρ_π after n steps, the value function is bounded as

$$||V^\pi||_\infty \leq r_{\max} \sum_{t \geq 0} \rho_\pi^{\lfloor t/n \rfloor}.$$

Proof. We have that by def. 8

$$\mathbb{P}(x_{2n} \neq \bar{x} \,|\, x_0 = x, \pi) = \mathbb{P}(x_{2n} \neq \bar{x} \,|\, x_n \neq \bar{x}, \pi) \times \mathbb{P}(x_n \neq \bar{x} \,|\, x_0 = x, \pi) \leq \rho_\pi^2.$$

Then for any $t \in \mathbb{N}$

$$\mathbb{P}(x_t \neq \bar{x} \,|\, x_0 = x, \pi) \leq \rho_\pi^{\lfloor t/n \rfloor},$$

which implies that *eventually* the terminal state $\bar{x}$ is achieved with probability 1. Then

$$||V^\pi||_\infty = \max_{x \in X} \mathbb{E}\big[\sum_{t=0}^{\infty} r(x_t, \pi(x_t)) | x_0 = x; \pi\big] \leq r_{\max} \sum_{t > 0} \mathbb{P}(x_t \neq \bar{x} \,|\, x_0 = x, \pi) \leq n r_{\max} + r_{\max} \sum_{t \geq n} \rho_\pi^{\lfloor t/n \rfloor}.$$

□

[4] It is simpler to state W as function rather than a vector as done for value functions.

1.5.2 Bellman operators

Assumption. There exists at least one *proper* policy and for any non-proper policy π there exists at least one state x where the corresponding value function is negatively unbounded, i.e., $V^\pi(x) = -\infty$, which corresponds to the existence of a cycle with only negative rewards.

Proposition 7. [Bertsekas and Tsitsiklis, 1996] Under the previous assumption, the optimal value function is bounded, i.e., $||V^*||_\infty < \infty$ and it is the unique fixed point of the **optimal** Bellman operator $\mathcal{T}$ such that for any vector $W \in \mathbb{R}^n$

$$\mathcal{T}W(x) = \max_{a\in A}\big[r(x,a) + \sum_y p(y|x,a)W(y)\big].$$

Furthermore, we have that $V^* = \lim_{k\to\infty}(\mathcal{T})^k W$.

Proposition 8. Let all the policies π be **proper**, then there exist a vector $\mu \in \mathbb{R}^N$ with $\mu > \mathbf{0}$ and a scalar $\beta < 1$ such that, $\forall x, y \in X_N$, $\forall a \in A$,

$$\sum_y p(y|x,a)\mu(y) \le \beta\mu(x).$$

Then it follows that both operators $\mathcal{T}$ and $\mathcal{T}^\pi$ are contraction in the weighted norm $L_{\infty,\mu}$, that is

$$||\mathcal{T}W_1 - \mathcal{T}W_2||_{\infty,\mu} \le \beta||W_1 - W_2||_{\infty,\mu}.$$

Proof. Let μ be the maximum (over all policies) of the average time to the terminal goal. This can be easily casted to a MDP where for any action and any state the rewards are 1 (i.e., for any $x \in X$ and $a \in A$, $r(x,a) = 1$). Under the assumption that all the policies are proper, then μ is finite and it is the solution to the dynamic programming equation

$$\mu(x) = 1 + \max_a \sum_y p(y|x,a)\mu(y).$$

Then $\mu(x) \ge 1$ and for any $a \in A$, $\mu(x) \ge 1 + \sum_y p(y|x,a)\mu(y)$. Furthermore,

$$\sum_y p(y|x,a)\mu(y) \le \mu(x) - 1 \le \beta\mu(x),$$

for

$$\beta = \max_x \frac{\mu(x)-1}{\mu(x)} < 1.$$

From this definition of μ and β we obtain the contraction property of $\mathcal{T}$ (similar for $\mathcal{T}^\pi$) in norm $L_{\infty,\mu}$:

$$\begin{aligned}
||\mathcal{T}W_1 - \mathcal{T}W_2||_{\infty,\mu} &= \max_x \frac{|\mathcal{T}W_1(x) - \mathcal{T}W_2(x)|}{\mu(x)} \\
&\le \max_{x,a} \frac{\sum_y p(y|x,a)}{\mu(x)}|W_1(y) - W_2(y)| \\
&\le \max_{x,a} \frac{\sum_y p(y|x,a)\mu(y)}{\mu(x)}\|W_1 - W_2\|_\mu \\
&\le \beta\|W_1 - W_2\|_\mu
\end{aligned}$$

□

2 Dynamic Programming

2.1 Value Iteration (VI)

The idea. We build a sequence of value functions. Let V_0 be any vector in R^N, then iterate the application of the optimal Bellman operator so that given V_k at iteration k we compute

$$V_{k+1} = \mathcal{T} V_k.$$

From Proposition 5 we have that $\lim_{k \to \infty} V_k = V^*$, thus a repeated application of the Bellman operator make the initial guess V_0 closer and closer to the actual optimal value function. At any iteration K, the algorithm returns the *greedy* policy

$$\pi_K(x) \in \arg\max_{a \in A} \big[r(x,a) + \gamma \sum_y p(y|x,a) V_K(y) \big].$$

Guarantees. Using the contraction property of the Bellman operator we obtain the convergence of V_0 to V^*. In fact,

$$||V_{k+1} - V^*||_\infty = ||\mathcal{T} V_k - \mathcal{T} V^*||_\infty \leq \gamma ||V_k - V^*||_\infty \leq \gamma^{k+1} ||V_0 - V^*||_\infty \to 0.$$

This also provides the convergence rate of VI. Let $\epsilon > 0$ be a desired level of accuracy in approximating V^* and $||r||_\infty \leq r_{\max}$, then after *at most*

$$K = \frac{\log(r_{\max}/\epsilon)}{\log(1/\gamma)}$$

iterations VI returns a value function V_K such that $||V_K - V^*||_\infty < \epsilon$. In fact, after K iterations we have

$$||V_K - V^*||_\infty < \gamma^{K+1} ||V_0 - V^*||_\infty \leq \gamma^{K+1} r_{\max} \leq \epsilon.$$

Computational complexity. One application of the optimal Bellman operator takes $O(N^2 A)$ operations.

Remark: Notice that the previous guarantee is for the *value function* returned by VI and not the corresponding *policy* π_K.

Q-iteration. Exactly the same algorithm can be applied to Q-functions using the corresponding optimal Bellman operator.

Implementation. There exist several implementations depending on the order used to update the different components of V_k (i.e., the states). For instance, in asynchronous VI, at each iteration k, one single state x_k is chosen and only the corresponding component of V_k is updated, i.e., $V_{k+1}(x_k) = \mathcal{T} V_k(x_k)$, while all the other states remain unchanged. If all the states are selected **infinitely often**, then $V_k \to V^*$.

Pros: each iteration is very computationally efficient.

Cons: convergence is only asymptotic.

2.2 Policy Iteration (PI)

Notation. Further extending the vector notation, for any policy π we introduce the reward vector $r^\pi \in \mathbb{R}^N$ as $r^\pi(x) = r(x, \pi(x))$ and the transition matrix $P^\pi \in \mathbb{R}^{N\times N}$ as $P^\pi(x,y) = p(y|x,\pi(x))$

The idea. We build a sequence of policies. Let π_0 be any stationary policy. At each iteration k we perform the two following steps:

1. **Policy evaluation** given π_k, compute V^{π_k}.
2. **Policy improvement**: we compute the *greedy* policy π_{k+1} from V^{π_k} as:

$$\pi_{k+1}(x) \in \arg\max_{a\in A} \big[r(x,a) + \gamma \sum_y p(y|x,a) V^{\pi_k}(y)\big].$$

The iterations continue until $V^{\pi_k} = V^{\pi_{k+1}}$.

Remark: Notice π_{k+1} is called *greedy* since it is maximizing the current estimate of the future rewards, which corresponds to the application of the optimal Bellman operator, since $\mathcal{T}^{\pi_{k+1}} V^{\pi_k} = \mathcal{T} V^{\pi_k}$.

Remark: The PI algorithm is often seen as an *actor-critic* algorithm.

Guarantees.

Proposition 9. The policy iteration algorithm generates a sequences of policies with non-decreasing performance (i.e., $V^{\pi_{k+1}} \geq V^{\pi_k}$)) and it converges to π^* in a finite number of iterations.

Proof. From the definition of the operators $\mathcal{T}$, $\mathcal{T}^{\pi_k}$, $\mathcal{T}^{\pi_{k+1}}$ and of the greedy policy π_{k+1}, we have that

$$V^{\pi_k} = \mathcal{T}^{\pi_k} V^{\pi_k} \leq \mathcal{T} V^{\pi_k} = \mathcal{T}^{\pi_{k+1}} V^{\pi_k}, \tag{14}$$

and from the monotonicity property of $\mathcal{T}^{\pi_{k+1}}$, it follows that

$$\begin{aligned} V^{\pi_k} &\leq \mathcal{T}^{\pi_{k+1}} V^{\pi_k}, \\ \mathcal{T}^{\pi_{k+1}} V^{\pi_k} &\leq (\mathcal{T}^{\pi_{k+1}})^2 V^{\pi_k}, \\ &\dots \\ (\mathcal{T}^{\pi_{k+1}})^{n-1} V^{\pi_k} &\leq (\mathcal{T}^{\pi_{k+1}})^n V^{\pi_k}, \\ &\dots \end{aligned}$$

Joining all the inequalities in the chain we obtain

$$V^{\pi_k} \leq \lim_{n\to\infty} (\mathcal{T}^{\pi_{k+1}})^n V^{\pi_k} = V^{\pi_{k+1}}.$$

Then $(V^{\pi_k})_k$ is a non-decreasing sequence. Furthermore, in a finite MDP we have a finite number of policies, then the termination condition is always met for a specific k. Thus eq. 14 holds with an equality and we obtain

$$V^{\pi_k} = \mathcal{T} V^{\pi_k}$$

and $V^{\pi_k} = V^*$ which implies that π_k is an optimal policy. □

Remark: More recent and refined proofs of the convergence rate of PI are available.

Policy evaluation. At each iteration k the value function V^{π_k} corresponding to the current policy π_k is computed. There are a number of different approaches to compute V^{π_k}.

- **Direct computation.** From the vector notation introduced before and the Bellman equation in Proposition 4 we have that for any policy π:

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

  By rearranging the terms we write the previous equation as

  $$(I - \gamma P^\pi)V^\pi = r^\pi.$$

  Since P^π is a stochastic matrix, then all its eigenvalues are ≤ 1. Thus the eigenvalues of the matrix $(I - \gamma P^\pi)$ are bounded by $\geq 1 - \gamma$, which guarantees that it is invertible. Then we can compute V^π as

  $$V^\pi = (I - \gamma P^\pi)^{-1} r^\pi.$$

  The inversion can be done with different methods. For instance, the Gauss-Jordan elimination algorithm has a complexity $O(N^3)$, which can be lowered to $O(N^{2.807})$ using Strassen's algorithm.

- **Iterative policy evaluation.** For any policy π from the properties of the policy Bellman operator $\mathcal{T}^\pi$, for any V_0 we have that $\lim_{n\to\infty} \mathcal{T}^\pi V_0 = V^\pi$. Thus an approximation of V^π could be computed by re-iterating the application $\mathcal{T}^\pi$ for n steps. In particular, in order to achieve an ϵ-approximation of V^π we need $O(N^2 \frac{\log 1/\epsilon}{\log 1/\gamma})$ steps.

- **Monte-Carlo simulation.** In each state x, we simulate n trajectories $((x_t^i)_{t\geq 0,})_{1\leq i\leq n}$ following policy π, where for any $i = 1, \ldots, n$, $x_0^i = x$ and $x_{t+1}^i \sim p(\cdot|x_t^i, \pi(x_t^i))$. We compute

  $$\hat{V}^\pi(x) \simeq \frac{1}{n} \sum_{i=1}^{n} \sum_{t\geq 0} \gamma^t r(x_t^i, \pi(x_t^i)).$$

  The approximation error is of order $O(1/\sqrt{n})$.

- **Temporal-difference (TD)** see next lecture.

Policy improvement. The computation of π_{k+1} requires the computation of an expectation w.r.t. the next state, which might be as expensive as $O(N|A|)$. If we move to policy iteration of Q-functions, then the policy improvement step simplifies to

$$\pi_{k+1}(x) \in \arg\max_{a\in A} Q(x, a),$$

which has a computational cost of $O(|A|)$. Furthermore, this could allow to compute the greedy policy even when the dynamics p is not known.

Pros: converge in a finite number of iterations (often small in practice).

Cons: each iteration requires a full policy evaluation and it might be expensive.

2.3 Linear Programming (LP)

Geometric interpretation of PI. We first provide a different interpretation of PI as a Newton method trying to find the zero of the Bellman residual operator. Let $B = \mathcal{T} - I$ the Bellman residual operator. From the definition of V^* as the fixed point of $\mathcal{T}$, it follows that $\mathcal{B}V^* = 0$. Thus, computing V^* as the fixed point of $\mathcal{T}$ is equivalent to compute the zero of $\mathcal{B}$. We first prove the following proposition.

Proposition 10. Let $\mathcal{B} = \mathcal{T} - I$ the Bellman residual operator and $\mathcal{B}'$ its derivative. Then sequence of policies π_k generated by PI satisfies

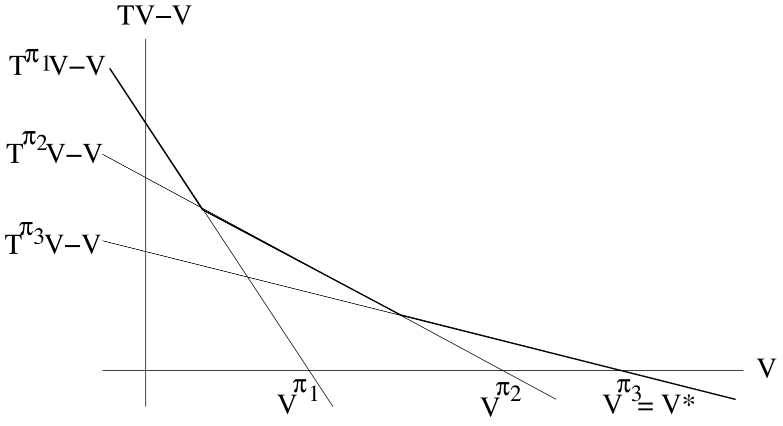
$$\begin{aligned} V^{\pi_{k+1}} &= V^{\pi_k} - (\gamma P^{\pi_{k+1}} - I)^{-1}[\mathcal{T}^{\pi_{k+1}} V^{\pi_k} - V^{\pi_k}] \\ &= V^{\pi_k} - [\mathcal{B}']^{-1} \mathcal{B} V^{\pi_k}, \end{aligned}$$

which coincides with the standard formulation of the Newton's method.

Proof. By the definition of V^{π_k} and the Bellman operators we have

$$\begin{aligned} V^{\pi_{k+1}} &= (I - \gamma P^{\pi_{k+1}})^{-1} r^{\pi_{k+1}} - V^{\pi_k} + V^{\pi_k} \\ &= V^{\pi_k} + (I - \gamma P^{\pi_{k+1}})^{-1} [r^{\pi_{k+1}} + (\gamma P^{\pi_{k+1}} - I) V^{\pi_k}] \end{aligned}$$

□



Following the geometric interpretation, we have that V^{π_k} is the zero of the linear operator $\mathcal{T}^{\pi_k} - I$. Since the application $V \to \mathcal{T}V - V = \max_\pi T^\pi V - V$ is convex, then we have that the Newton's method is guaranteed to converge for any V_0 such that $\mathcal{T}V_0 - V_0 \geq 0$.

Linear Programming The previous intuition is at the basis of the observation that V^* is the smallest vector V such that $V \geq \mathcal{T}V$. In fact,

$$V \geq \mathcal{T}V \quad \Longrightarrow \quad V \geq \lim_{k \to \infty} (\mathcal{T})^k V = V^*.$$

Thus, we can compute V^* as the solution to the **linear program**:

- Min $\sum_x V(x)$,

- Subject to (finite number of linear inequalities which defines a polyhidron in $\mathbb{R}^N$):

$$V(x) \geq r(x,a) + \gamma \sum_y p(y|x,a)V(y), \quad \forall x \in X, \forall a \in A$$

which contains N variables and $N \times |A|$ constraints.

A Probability Theory

Definition 9 (Conditional probability)**.** *Given two events A and B with $\mathbb{P}(B) > 0$, the conditional probability of A given B is*

$$\mathbb{P}(A|B) = \frac{\mathbb{P}(A \cup B)}{\mathbb{P}(B)}.$$

Similarly, if X and Y are non-degenerate and jointly continuous random variables with density $f_{X,Y}(x,y)$ then if B has positive measure then the conditional probability is

$$\mathbb{P}(X \in A | Y \in B) = \frac{\int_{y \in B} \int_{x \in A} f_{X,Y}(x,y) dx dy}{\int_{y \in B} \int_{x} f_{X,Y}(x,y) dx dy}.$$

Definition 10 (Law of total expectation)**.** *Given a function f and two random variables X, Y we have that*

$$\mathbb{E}_{X,Y}\big[f(X,Y)\big] = \mathbb{E}_X\Big[\mathbb{E}_Y\big[f(x,Y)|X=x\big]\Big].$$

B Norms, Contractions, and Banach's Fixed-Point Theorem

Definition 11. *Given a vector space $\mathcal{V} \subseteq \mathbb{R}^d$ a function $f : \mathcal{V} \to \mathbb{R}_0^+$ is a norm if an only if*

- *If $f(v) = 0$ for some $v \in \mathcal{V}$, then $v = 0$.*
- *For any $\lambda \in \mathbb{R}, v \in \mathcal{V}$, $f(\lambda v) = |\lambda| f(v)$.*
- ***Triangle inequality:** For any $v, u \in \mathcal{V}$, $f(v + u) \leq f(v) + f(u)$.*

A list of useful norms.

- L_p-norm

$$||v||_p = \left(\sum_{i=1}^{d} |v_i|^p\right)^{1/p}.$$

- L_∞-norm

$$||v||_\infty = \max_{1 \leq i \leq d} |v_i|.$$

- $L_{\mu,p}$-norm

$$||v||_{\mu,p} = \left(\sum_{i=1}^{d} \frac{|v_i|^p}{\mu_i}\right)^{1/p}.$$

- $L_{\mu,p}$-norm

$$||v||_{\mu,\infty} = \max_{1 \leq i \leq d} \frac{|v_i|}{\mu_i}.$$

- $L_{2,P}$-matrix norm (P is a positive definite matrix)

$$||v||_P^2 = v^\top P v.$$

Definition 12. *A sequence of vectors $v_n \in \mathcal{V}$ (with $n \in \mathbb{N}$) is said to converge in norm $||\cdot||$ to $v \in \mathcal{V}$ if*

$$\lim_{n\to\infty} ||v_n - v|| = 0.$$

Definition 13. *A sequence of vectors $v_n \in \mathcal{V}$ (with $n \in \mathbb{N}$) is a Cauchy sequence if*

$$\lim_{n\to\infty} \sup_{m\geq n} ||v_n - v_m|| = 0.$$

Definition 14. *A vector space $\mathcal{V}$ equipped with a norm $||\cdot||$ is complete if every Cauchy sequence in $\mathcal{V}$ is convergent in the norm of the space.*

Definition 15. *An operator $\mathcal{T} : \mathcal{V} \to \mathcal{V}$ is L-Lipschitz if for any $v, u \in \mathcal{V}$*

$$||\mathcal{T}v - \mathcal{T}u|| \leq L||u - v||.$$

*If $L \leq 1$ then $\mathcal{T}$ is a **non-expansion**, while if $L < 1$ then $\mathcal{T}$ is a L-**contraction**.*

*If $\mathcal{T}$ is Lipschitz then it is also **continuous**, that is*

$$\textit{if } v_n \to_{||\cdot||} v \textit{ then } \mathcal{T}v_n \to_{||\cdot||} \mathcal{T}v.$$

Definition 16. *A vector $v \in \mathcal{V}$ is a fixed point of the operator $\mathcal{T} : \mathcal{V} \to \mathcal{V}$ if $\mathcal{T}v = v$.*

Proposition 11. Let $\mathcal{V}$ be a complete vector space equipped with the norm $||\cdot||$ and $\mathcal{T} : \mathcal{V} \to \mathcal{V}$ be a γ-contraction mapping. Then

1. $\mathcal{T}$ admits a **unique fixed point** v.
2. For any $v_0 \in \mathcal{V}$, if $v_{n+1} = \mathcal{T}v_n$ then $v_n \to_{||\cdot||} v$ with a geometric convergence rate:

$$||v_n - v|| \leq \gamma^n ||v_0 - v||.$$

Proof. We first derive the *fundamental contraction property.* For any $v, u \in \mathcal{V}$:

$$\begin{aligned} ||v - u|| &= ||v - \mathcal{T}v + \mathcal{T}v - \mathcal{T}u + \mathcal{T}u - u|| \\ &\overset{(a)}{\leq} ||v - \mathcal{T}v|| + ||\mathcal{T}v - \mathcal{T}u|| + ||\mathcal{T}u - u|| \\ &\overset{(b)}{\leq} ||v - \mathcal{T}v|| + \gamma||v - u|| + ||\mathcal{T}u - u||, \end{aligned}$$

where (a) follows from the triangle inequality and (b) from the contraction property of $\mathcal{T}$. Rearranging the terms we obtain:

$$||v - u|| \leq \frac{||v - \mathcal{T}v|| + ||\mathcal{T}u - u||}{1 - \gamma}. \tag{15}$$

If v and u are both fixed points then $||\mathcal{T}v - v|| = 0$ and $||\mathcal{T}u - u|| = 0$, thus from the previous inequality $||v - u|| \leq 0$ which implies $v = u$. Then $\mathcal{T}$ admits *at most* one fixed point.

Let $v_0 \in \mathcal{V}$, for any $n, m \in \mathbb{N}$ an iterative application of eq. 15 gives

$$\begin{aligned}
||\mathcal{T}^n v_0 - \mathcal{T}^m u_0|| &\leq \frac{||\mathcal{T}^n v_0 - \mathcal{T}\mathcal{T}^n v_0|| + ||\mathcal{T}\mathcal{T}^m v_0 - \mathcal{T}^m v_0||}{1-\gamma} \\
&= \frac{||\mathcal{T}^n v_0 - \mathcal{T}^n \mathcal{T} v_0|| + ||\mathcal{T}^m \mathcal{T} v_0 - \mathcal{T}^m v_0||}{1-\gamma} \\
&\overset{(a)}{\leq} \frac{\gamma^n ||v_0 - \mathcal{T} v_0|| + \gamma^m ||\mathcal{T} v_0 - v_0||}{1-\gamma} \\
&= \frac{\gamma^n + \gamma^m}{1-\gamma} ||v_0 - \mathcal{T} v_0||,
\end{aligned}$$

where in (a) we used the fact that $\mathcal{T}^n$ is a γ^n contraction. Since $\gamma < 1$ we have that

$$\lim_{n\to\infty} \sup_{m \geq n} ||\mathcal{T}^n v_0 - \mathcal{T}^m u_0|| \leq \lim_{n\to\infty} \sup_{m\geq n} \frac{\gamma^n + \gamma^m}{1-\gamma} ||v_0 - \mathcal{T} v_0|| = 0,$$

which implies that $\{\mathcal{T}^n v_0\}_n$ is a Cauchy sequence. Since $\mathcal{V}$ is complete by assumption then $\{\mathcal{T}^n v_0\}_n$ is convergent to a vector v. Recalling that $v_{n+1} = \mathcal{T} v_n$ and by taking the limit on both sides we obtain

$$\begin{aligned}
\lim_{n\to\infty} v_{n+1} &\overset{(a)}{=} \lim_{n\to\infty} \mathcal{T}^{n+1} v_0 = v, \\
\lim_{n\to\infty} \mathcal{T} v_n &\overset{(b)}{=} \mathcal{T} \lim_{n\to\infty} v_n = \mathcal{T} v,
\end{aligned}$$

where (a) follows from the definition of v_{n+1} and the fact that $\{\mathcal{T}^n v_0\}_n$ is a convergence Cauchy sequence, while (b) follows from the fact that a contraction operator $\mathcal{T}$ is also continuous. Joining the two equalities we obtain $v = \mathcal{T} v$ which is the definition of fixed point. □

C Linear Algebra

Eigenvalues of a matrix (1). Given a square matrix $A \in \mathbb{R}^{N\times N}$, a vector $v \in \mathbb{R}^N$ and a scalar λ are *eigenvector* and *eigenvalue* of the matrix if

$$Av = \lambda v.$$

Eigenvalues of a matrix (2). Given a square matrix $A \in \mathbb{R}^{N\times N}$ with eigenvalues $\{\lambda_i\}_{i=1}^N$, then the matrix $(I - \alpha A)$ has eigenvalues $\{\mu_i = 1 - \alpha\lambda_i\}_{i=1}^N$.

Stochastic matrix. A square matrix $P \in \mathbb{R}^{N\times N}$ is a stochastic matrix if

1. all non-zero entries, $\forall i, j$, $[P]_{i,j} \geq 0$
2. all the rows sum to one, $\forall i$, $\sum_{j=1}^N [P]_{i,j} = 1$.

All the eigenvalues of a stochastic matrix are bounded by 1, i.e., $\forall i$, $\lambda_i \leq 1$.

Matrix inversion. A square matrix $A \in \mathbb{R}^{N\times N}$ can be inverted if and only if $\forall i$, $\lambda_i \neq 0$.

References

[Bellman, 1957] Bellman, R. E. (1957). *Dynamic Programming.* Princeton University Press, Princeton, N.J.

[Bertsekas and Tsitsiklis, 1996] Bertsekas, D. and Tsitsiklis, J. (1996). *Neuro-Dynamic Programming*. Athena Scientific, Belmont, MA.

[Fleming and Rishel, 1975] Fleming, W. and Rishel, R. (1975). *Deterministic and stochastic optimal control*. Applications of Mathematics, 1, Springer-Verlag, Berlin New York.

[Howard, 1960] Howard, R. A. (1960). *Dynamic Programming and Markov Processes*. MIT Press, Cambridge, MA.

[Puterman, 1994] Puterman, M. (1994). *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. John Wiley & Sons, Inc., New York, Etats-Unis.