DSCC/CSC 435 & ECE 412 Optimization for Machine Learning Lecture 16

Reinforcement Learning

Lecturer: Jiaming Liang *November 30, 2023*

1 Markov decision process model

Reinforcement learning is the study of planing and learning in a scenario where a learner/agent actively interacts with the environment to maximize the reward she receives from the environment.

We first introduce the model of Markov decision processes (MDPs), a model of the environment and interactions with the environment widely adopted in reinforcement learning. We then introduce several algorithms for the planning problem, which corresponds to the case where the environment model is known to the agent, and a series of learning algorithms for the more general case of an unknown model.

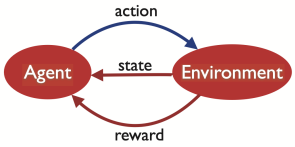

Figure 1: The general scenario of reinforcement learning

An MDP is a Markovian process defined as follows.

Definition 1. *A Markov decision process is defined by:*

- *a set of states S, possibly infinite;*
- *a start state or initial state $s_0 \in S$;*
- *a set of actions A, possibly infinite;*
- *a transition probability $\mathbb{P}(s' \mid s, a)$: distribution over destination states $s' = \delta(s, a)$;*
- *a reward probability $\mathbb{P}(r' \mid s, a)$: distribution over rewards returned $s' = \delta(s, a)$.*

The model is Markovian because the transition and reward probabilities depend only on the current state s and not the entire history of states and actions taken.

Example: whether to fish salmons this year.

- state space (salmon population): $S = \{\text{empty, low, medium, high}\}$;

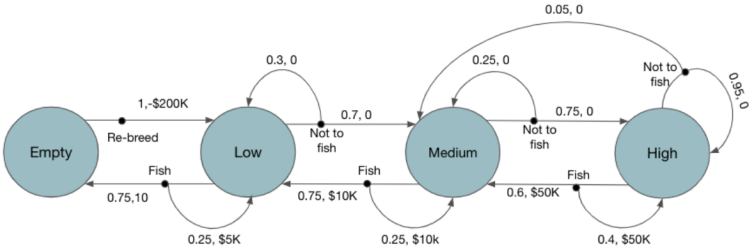

Figure 2: An example of MDP

- action space: $A = \{\text{fish, not to fish, re-breed}\}$;
- transition probabilities and rewards labeled in graph.

Definition 2. ***(Policy)*** *A policy is a mapping $\pi : S \to A$.*

More precisely, this is the definition of a stationary policy since the choice of the action does not depend on the time. More generally, we could define a non-stationary policy as a sequence of mappings $\pi_t : S \to A$ indexed by t.

The agent's objective is to find a policy that maximizes her expected (reward) return. The return she receives following a policy π along a specific sequence of states $s_t, \ldots, s_T$ is defined as follows:

- finite horizon $(T < \infty) : \sum_{\tau=0}^{T-t} r\left(s_{t+\tau}, \pi\left(s_{t+\tau}\right)\right)$;
- infinite horizon $(T = \infty) : \sum_{\tau=0}^{T-t} \gamma^{\tau} r\left(s_{t+\tau}, \pi\left(s_{t+\tau}\right)\right)$, where $\gamma \in [0, 1)$ is a constant factor less than one used to discount future rewards.

Definition 3. ***(Policy value)*** *The value $V_\pi(s)$ of a policy π at state $s \in S$ is defined as the expected reward returned when starting at s and following policy π:*

- *finite horizon:* $V_\pi(s) = \mathbb{E}\left[\sum_{\tau=0}^{T-t} r\left(s_{t+\tau}, \pi\left(s_{t+\tau}\right)\right) \mid s_t = s\right]$;
- *infinite discounted horizon:* $V_\pi(s) = \mathbb{E}\left[\sum_{\tau=0}^{T-t} \gamma^{\tau} r\left(s_{t+\tau}, \pi\left(s_{t+\tau}\right)\right) \mid s_t = s\right]$,

where the expectations are over the random selection of the states s_t and the reward values r_{t+1}. An infinite undiscounted horizon is also often considered based on the limit of the average reward, when it exists.

Proposition 1. ***(Bellman equation)*** *The values $V_\pi(s)$ of policy π at states $s \in S$ for an infinite horizon MDP obey the following system of linear equations:*

$$V_\pi(s) = \mathbb{E}[r(s, \pi(s)] + \gamma \sum_{s'} \mathbb{P}\left(s' \mid s, \pi(s)\right) V_\pi\left(s'\right), \quad \forall s \in S. \tag{1}$$

Definition 4. ***(Optimal policy)*** *A policy π^* is optimal if it has maximal value for all states $s \in S$.*

Definition 5. ***(State-action value function)*** *The optimal state-action value function Q^* is defined for all $(s, a) \in S \times A$ as the expected return for taking action $a \in A$ at state $s \in S$ and then following the optimal policy:*

$$Q^*(s,a) = \mathbb{E}[r(s,a)] + \gamma \sum_{s' \in S} \mathbb{P}\left(s' \mid s, a\right) V^*\left(s'\right). \tag{2}$$

It is not hard to see then that the optimal policy values are related to Q^* via

$$V^*(s) = \max_{a \in A} Q^*(s,a), \quad \forall s \in S. \tag{3}$$

Observe also that, by definition of the optimal policy, we have

$$\pi^*(s) = \operatorname*{argmax}_{a \in A} Q^*(s,a), \quad \forall s \in S.$$

Replacing Q^* by its definition in (3) gives the following system of equations for the optimal policy values $V^*(s)$:

$$V^*(s) = \max_{a \in A} \left\{ \mathbb{E}[r(s,a)] + \gamma \sum_{s' \in S} \mathbb{P}\left(s' \mid s, a\right) V^*\left(s'\right) \right\},$$

also known as Bellman equations. Note that this new system of equations is not linear due to the presence of the max operator. It is distinct from the previous linear system we defined under the same name in (1).

2 Planning algorithms

In this section, we assume that the environment model is known. That is, the transition probability $\mathbb{P}\left(s' \mid s, a\right)$ and the expected reward $\mathbb{E}[r(s,a)]$ for all $s, s' \in S$ and $a \in A$ are assumed to be given. The problem of finding the optimal policy then does not require learning the parameters of the environment model or estimating other quantities helpful in determining the best course of actions, it is purely a planning problem.

Three standard algorithms for this planning problem are the value iteration algorithm, the policy iteration algorithm, and a linear programming formulation of the problem. We will skip the details of these algorithms since the learning algorithms are more related to optimization for machine learning.

3 Stochastic approximation

The estimation and algorithmic methods adopted for learning in reinforcement learning are closely related to the concepts and techniques in stochastic approximation. Thus, we start by introducing several useful results of this field that will be needed for the proofs of convergence of the reinforcement learning algorithms presented.

Theorem 1. ***(Mean estimation)*** *Let X be a random variable taking values in $[0,1]$ and let $x_0, \ldots, x_m$ be i.i.d. values of X. Define the sequence $(\mu_m)_{m\in\mathbb{N}}$ by*

$$\mu_{m+1} = (1-\alpha_m)\,\mu_m + \alpha_m x_m,$$

with $\mu_0 = x_0, \alpha_m \in [0,1], \sum_{m\geq 0}\alpha_m = +\infty$ and $\sum_{m\geq 0}\alpha_m^2 < +\infty$. Then,

$$\mu_m \xrightarrow{a.s.} \mathbb{E}[X].$$

Stochastic optimization is the general problem of finding the solution to the equation

$$x = H(x),$$

where $x \in \mathbb{R}^n$, when

- $H(x)$ cannot be computed, for example, because H is not accessible or because the cost of its computation is prohibitive;
- but an i.i.d. sample of m noisy observations $H(x_i) + w_i$ are available, $i \in [1,m]$, where the noise random variable w has expectation zero: $\mathbb{E}[w] = 0$.

This problem arises in a variety of different contexts and applications. As we shall see, it is directly related to the learning problem for MDPs.

One general idea for solving this problem is to use an iterative method and define a sequence $\{x_t\}_{t\in\mathbb{N}}$ in a way similar to what is suggested by Theorem 1:

$$\begin{aligned} x_{t+1} &= (1-\alpha_t)\,x_t + \alpha_t\left[H\left(x_t\right) + w_t\right] \\ &= x_t + \alpha_t\left[H\left(x_t\right) + w_t - x_t\right], \end{aligned}$$

where $\{\alpha_t\}_{t\in\mathbb{N}}$ follow conditions similar to those assumed in Theorem 1. More generally, we consider sequences defined via

$$x_{t+1} = x_t + \alpha_t D(x_t, w_t),$$

where D is a function mapping $\mathbb{R}^n \times \mathbb{R}^n$ to $\mathbb{R}^n$. There are many different theorems guaranteeing the convergence of this sequence under various assumptions. We will present one of the most general forms of such theorems, which relies on the following general result.

Theorem 2. ***(Supermartingale convergence)*** *Let $\{X_t\}_{t\in\mathbb{N}}$, $\{Y_t\}_{t\in\mathbb{N}}$, and $\{Z_t\}_{t\in\mathbb{N}}$ be sequences of non-negative random variables such that $\sum_{t=0}^{\infty} Y_t < \infty$. Let $\mathcal{F}_t$ denote all the information for $t' \leq t : \mathcal{F}_t = \left\{(X_{t'})_{t'\leq t}, (Y_{t'})_{t'\leq t}, (Z_{t'})_{t'\leq t}\right\}$. Then, if $\mathbb{E}\left[X_{t+1} \mid \mathcal{F}_t\right] \leq X_t + Y_t - Z_t$, the following holds:*

- *X_t converges to a limit (with probability one);*
- *$\sum_{t=0}^{\infty} Z_t < \infty$.*

The following is one of the most general forms of such theorems.

Theorem 3. *Let D be a function mapping $\mathbb{R}^n \times \mathbb{R}^n$ to $\mathbb{R}^n$, $\{x_t\}_{t\in\mathbb{N}}$ and $\{w_t\}_{t\in\mathbb{N}}$ be two sequences in $\mathbb{R}^n$, and $\{\alpha_t\}_{t\in\mathbb{N}}$ be a sequence of real numbers with $x_{t+1} = x_t + \alpha_t D(x_t, w_t)$. Let $\mathcal{F}_t$ denote the entire history for $t' \leq t$, that is: $\mathcal{F}_t = \{\{x_{t'}\}_{t'\leq t}, \{w_{t'}\}_{t'\leq t}, \{\alpha_{t'}\}_{t'\leq t}\}$.*

Let Ψ denote $x \mapsto \frac{1}{2}\left\|x - x^*\right\|_2^2$ for some $x^* \in \mathbb{R}^n$ and assume that D and $\{\alpha_t\}_{t\in\mathbb{N}}$ verify the following conditions:

- *$\exists K_1, K_2 \in \mathbb{R} : \mathbb{E}\left[\left\|D\left(x_t, w_t\right)\right\|_2^2 \middle| \mathcal{F}_t\right] \leq K_1 + K_2\Psi\left(x_t\right)$;*
- *$\exists c \geq 0 : \nabla\Psi\left(x_t\right)^\top \mathbb{E}\left[D\left(x_t, w_t\right) \mid \mathcal{F}_t\right] \leq -c\Psi\left(x_t\right)$;*
- *$\alpha_t > 0$, $\sum_{t=0}^{\infty} \alpha_t = \infty$, $\sum_{t=0}^{\infty} \alpha_t^2 < \infty$.*

Then, the sequence $\{x_t\}$ converges almost surely to x^*:

$$x_t \xrightarrow{a.s.} x^*.$$

Theorem 4. *Let H be a function mapping $\mathbb{R}^n$ to $\mathbb{R}^n$, $\{x_t\}_{t\in\mathbb{N}}$, $\{w_t\}_{t\in\mathbb{N}}$, and $\{\alpha_t\}_{t\in\mathbb{N}}$ be three sequences in $\mathbb{R}^n$ with*

$$x_{t+1}(s) = x_t(s) + \alpha_t(s)[H(x_t)(s) - x_t(s) + w_t(s)], \quad \forall s \in [1, n].$$

Let $\mathcal{F}_t$ denote the entire history for $t' \leq t$, that is: $\mathcal{F}_t = \{\{x_{t'}\}_{t'\leq t}, \{w_{t'}\}_{t'\leq t}, \{\alpha_{t'}\}_{t'\leq t}\}$ and assume the following conditions hold:

- *$\exists K_1, K_2 \in \mathbb{R} : \mathbb{E}\left[w_t^2(s) \mid \mathcal{F}_t\right] \leq K_1 + K_2\left\|x_t\right\|^2$ for some norm $\|\cdot\|$;*
- *$\mathbb{E}[w_t \mid \mathcal{F}_t] = 0$;*
- *$\sum_{t=0}^{\infty} \alpha_t = \infty$ and $\sum_{t=0}^{\infty} \alpha_t^2 < \infty$ for every $s \in [1, n]$;*
- *H is a $\|\cdot\|_\infty$-contraction with fixed point x^*.*

Then, the sequence $\{x_t\}$ converges almost surely to x^*:

$$x_t \xrightarrow{a.s.} x^*.$$

4 Learning algorithms

This section considers the more general scenario where the environment model of an MDP, that is the transition and reward probabilities , is unknown.

There are two main learning approaches that can be adopted. One known as the model-free approach consists of learning an action policy directly. Another one, a model-based approach, consists of first learning the environment model, and then use that to learn a policy. The temporal difference (TD) and Q-learning algorithms we present for this problem are widely adopted in reinforcement learning and belong to the family of model-free approaches.

Both TD and Q-learning algorithms resemble stochastic formulations of the value iteration algorithm for planning. We will skip the policy gradient methods that resembles the policy iteration algorithm for planning.

4.1 Temporal difference algorithm

Our goal is to estimate the polivy value function under a stationary policy π. The TD algorithm is based on Bellman's linear equations giving the value of a policy π (see (1)):

$$\begin{aligned} V_\pi(s) &= \mathbb{E}[r(s,\pi(s))] + \gamma \sum_{s' \in S} \mathbb{P}\left(s' \mid s, \pi(s)\right) V_\pi\left(s'\right) \\ &= \mathbb{E}\left[r(s,\pi(s)) + \gamma V_\pi(s') \mid s\right]. \end{aligned}$$

The probability distribution according to which this last expectation is defined is not known. The TD algorithm consists of:

- sampling a new state s';
- updating the policy values according to the following, which justifies the name of the algorithm:

$$\begin{aligned} V(s) &\leftarrow (1-\alpha)V(s) + \alpha\left[r(s,\pi(s)) + \gamma V\left(s'\right)\right] \\ &= V(s) + \alpha[\underbrace{r(s,\pi(s)) + \gamma V\left(s'\right) - V(s)}_{\text{temporal difference of } V \text{ values}}]. \end{aligned}$$

  Here, the parameter α is a function of the number of visits to the state s.

Algorithm 1 TD

Input: Initial policy value vector $V_0 \in \mathbb{R}^{|S|}$, sample an initial state s
for $t = 0$ to T **do**
  Observe reward $r' = r(s, \pi(s))$ and next state s'
  Compute $V(s) \leftarrow (1 - \alpha)V(s) + \alpha[r' + \gamma V(s')]$
  Update $s \leftarrow s'$
end for
Return V

4.2 Q-learning algorithm

Our goal is to find an optimal policy and the core problem is to solve the Bellman equation. The Q-learning algorithm is based on the equations giving the optimal state-action value function Q^* (see (2) and (3)):

$$\begin{aligned} Q^*(s,a) &= \mathbb{E}[r(s,a)] + \gamma \sum_{s' \in S} \mathbb{P}\left(s' \mid s, a\right) V^*\left(s'\right) \\ &= \mathbb{E}\left[r(s,a) + \gamma \max_{a \in A} Q^*(s,a)\right]. \end{aligned}$$

The distribution model is not known. Thus, the Q-learning algorithm consists of the following main steps:

- sampling a new state s';
- updating the policy values according to the following:

$$Q(s,a) \leftarrow (1-\alpha)Q(s,a) + \alpha[r(s,a) + \gamma \max_{a' \in A} Q(s',a')],$$

  where the parameter α is a function of the number of visits to the state s.

Algorithm 2 Q-learning

Input: Initial state-action value function $Q_0 \in \mathbb{R}^{|S| \times |A|}$, sample an initial state s
for $t = 0$ to T **do**
  Select an action a from the current state s using a policy π derived from Q
  Observe reward $r' = r(s, \pi(s))$ and next state s'
  Compute $Q(s,a) \leftarrow (1-\alpha)Q(s,a) + \alpha[r' + \gamma \max_{a' \in A} Q(s',a')]$
  Update $s \leftarrow s'$
end for
Return Q

Theorem 5. *Consider a finite MDP. Assume that for all $s \in S$ and $a \in A$, $\sum_{t=0}^{\infty} \alpha_t(s,a) = \infty$ and $\sum_{t=0}^{\infty} \alpha_t^2(s,a) < \infty$ with $\alpha_t(s,a) \in [0,1]$. Then, the Q-learning algorithm converges to the optimal value Q^* (with probability one).*

Note that the conditions on $\alpha_t(s,a)$ impose that each state-action pair is visited infinitely many times.

Proof. Let $\{Q_t(s,a)\}_{t\geq 0}$ denote the sequence of state-action value functions at $(s,a) \in S \times A$ generated by the algorithm. By definition of the Q-learning updates,

$$Q_{t+1}(s_t,a_t) = Q_t(s_t,a_t) + \alpha \left[r(s_t,a_t) + \gamma \max_{a'} Q_t\left(s_{t+1},a'\right) - Q_t(s_t,a_t) \right].$$

This can be rewritten as the following for all $s \in S$ and $a \in A$:

$$\begin{aligned} Q_{t+1}(s,a) =& Q_t(s,a) + \alpha_t(s,a) \left[r(s,a) + \gamma \mathop{\mathbb{E}}_{s'\sim\mathbb{P}(\cdot|s,a)} \left[\max_{a'} Q_t\left(s',a'\right) \right] - Q_t(s,a) \right] \\ &+ \gamma \alpha_t(s,a) \left[\max_{a'} Q_t\left(s',a'\right) - \mathop{\mathbb{E}}_{s'\sim\mathbb{P}(\cdot|s,a)} \left[\max_{a'} Q_t\left(s',a'\right) \right] \right], \end{aligned} \tag{4}$$

if we define $\alpha_t(s,a)$ as 0 if $(s,a) \neq (s_t,a_t)$ and $\alpha_t(s_t,a_t)$ otherwise. Now,

$$w_t(s') = \max_{a'} Q_t\left(s',a'\right) - \mathop{\mathbb{E}}_{s'\sim\mathbb{P}(\cdot|s,a)} \left[\max_{a'} Q_t\left(s',a'\right) \right],$$

and

$$H(Q_t)(s,a) = r(s,a) + \gamma \mathop{\mathbb{E}}_{s'\sim\mathbb{P}(\cdot|s,a)} \left[\max_{a'} Q_t\left(s',a'\right) \right].$$

In view of (4),

$$Q_{t+1}(s,a) = Q_t(s,a) + \alpha_t(s,a) \left[H\left(Q_t\right)(s,a) - Q_t(s,a) + \gamma w_t(s) \right], \quad \forall (s,a) \in S \times A.$$

We now show that the hypotheses of Theorem 4 hold for Q_t and w_t, which will imply the convergence of Q_t to Q^*. The conditions on α_t hold by assumption. By definition of w_t, $\mathbb{E}\left[w_t \mid \mathcal{F}_t\right] = 0$. Also, for any $s' \in S$,

$$\begin{aligned} \left|w_t\left(s'\right)\right| &\leq \max_{a'} \left|Q_t\left(s',a'\right)\right| + \left| \mathop{\mathbb{E}}_{s'\sim\mathbb{P}[\cdot|s,a]} \left[\max_{a'} Q_t\left(s',a'\right) \right] \right| \\ &\leq 2 \max_{s'} \left| \max_{a'} Q_t\left(s',a'\right) \right| = 2 \left\|Q_t\right\|_\infty. \end{aligned}$$

Thus, $\mathbb{E}\left[w_t^2(s) \mid \mathcal{F}_t\right] \leq 4\left\|Q_t\right\|_\infty^2$. Finally, H is a γ-contraction for $\|\cdot\|_\infty$ since for any $Q_1', Q_2'' \in \mathbb{R}^{|S|\times|A|}$, and $(s,a) \in S \times A$, we can write

$$\begin{aligned}
\left|H\left(Q_2\right)(x,a) - H\left(Q_1'\right)(x,a)\right| &= \left|\gamma \mathop{\mathbb{E}}_{s'\sim\mathbb{P}[\cdot|s,a]}\left[\max_{a'} Q_2\left(s',a'\right) - \max_{a'} Q_1\left(s',a'\right)\right]\right| \\
&\leq \gamma \mathop{\mathbb{E}}_{s'\sim\mathbb{P}[\cdot|s,a]}\left[\left|\max_{a'} Q_2\left(s',a'\right) - \max_{a'} Q_1\left(s',a'\right)\right|\right] \\
&\leq \gamma \mathop{\mathbb{E}}_{s'\sim\mathbb{P}[\cdot|s,a]} \max_{a'}\left[\left|Q_2\left(s',a'\right) - Q_1\left(s',a'\right)\right|\right] \\
&\leq \gamma \max_{s'} \max_{a'}\left[\left|Q_2\left(s',a'\right) - Q_1\left(s',a'\right)\right|\right] \\
&= \gamma\left\|Q_2'' - Q_1'\right\|_\infty .
\end{aligned}$$

Since H is a contraction, it admits a fixed point $Q^* : H\left(Q^*\right) = Q^*$. □