

# Reinforcement Learning — Exercises Lectures 1-5

## Instructions

This is the first exercise booklet for Reinforcement Learning. It covers both ungraded exercises to practice at home or during the tutorial sessions as well as graded homework exercises. The graded homework assignments are clearly marked. Please note the following:

- Make sure you deliver answers in a clear and structured format.  $\text{\LaTeX}$  has our preference. Messy handwritten answers will not be graded.
- Pre-pend the name of your TA to the file name you hand in and remember to put your name and student ID on the submission;
- The deadline for this first assignment is *September 20th 2019* and will cover the material of lectures 1-5 (all questions marked ‘homework’ in this booklet).

## Contents

<b>Prior knowledge self-test</b>	<b>2</b>
0.1 Linear algebra and multivariable derivatives. . . . .	2
0.2 Probability theory . . . . .	2
0.3 OLS, linear projection, and gradient descent. . . . .	3
<b>1 Introduction</b>	<b>5</b>
1.1 Introduction . . . . .	5
1.2 Exploration . . . . .	5
<b>2 MDPs and dynamic programming</b>	<b>7</b>
2.1 Markov Decision Processes . . . . .	7
2.2 Homework: Dynamic Programming . . . . .	7
<b>3 Monte Carlo methods</b>	<b>8</b>
3.1 Homework: Monte Carlo . . . . .	8
<b>4 Temporal Difference methods</b>	<b>9</b>
4.1 Temporal Difference Learning (Application) . . . . .	9
4.2 Contraction Mapping . . . . .	9
4.3 Temporal Difference Learning (Theory) . . . . .	10
4.4 Homework: Maximization Bias . . . . .	11

# Lecture 0: Prior knowledge self-test

## 0.1 Linear algebra and multivariable derivatives.

Consider the following matrices and vectors:

$$A = \begin{pmatrix} a_{11} & 0 \\ 0 & a_{22} \end{pmatrix} \quad B = \begin{pmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \end{pmatrix} \quad c = \begin{pmatrix} y - x^2 \\ \frac{\ln x}{y} \end{pmatrix} \quad d = \begin{pmatrix} d_1 \\ d_2 \end{pmatrix} \quad e = \begin{pmatrix} x \\ y \end{pmatrix}$$

1. Compute  $AB$ ,  $AB^T$ , and  $d^T B d$ .
2. Find the inverses of  $A$  and  $B$ .
3. Compute  $\frac{\partial c}{\partial x}$  and  $\frac{\partial c}{\partial e}$ .
4. Consider the function  $f(\mathbf{x}) = \sum_i^N i x_i$ , which maps an  $N$ -dimensional vector of real numbers  $\mathbf{x}$  to a real number. Find an expression for  $\frac{\partial f}{\partial \mathbf{x}}$  in terms of integers 1 to  $N$ .

## 0.2 Probability theory

Assume  $X$  and  $Y$  are two independent random variables, respectively with mean  $\mu$ ,  $\nu$ , and variance  $\sigma^2$ ,  $\tau^2$ .

1. What is the expected value of  $X + \alpha Y$ , where  $\alpha$  is some constant?
2. What is the variance of  $X + \alpha Y$ ?

Suppose now that we have a training set consisting of points  $(x_1, x_2, \dots, x_n)$  and real values  $y_i$  associated to each point  $x_i$ . Assume that there is a function  $f$  such that  $y = f(x) + \epsilon$ , where  $\epsilon$  is a noise vector with zero mean and variance  $\sigma^2$ .

Assume we are looking for a function  $\hat{f}$  that minimises the squared error  $(y - \hat{f}(x))^2$ . Recall the bias-variance decomposition of this squared error on an unseen sample  $x$ :

$$\mathbb{E} \left[ (y - \hat{f}(x))^2 \right] = \left( \text{Bias}[\hat{f}(x)] \right)^2 + \text{Var}[\hat{f}(x)] + \sigma^2 \quad (1)$$

where

$$\text{Bias}[\hat{f}(x)] = \mathbb{E}[\hat{f}(x)] - f(x) \quad \text{Var}[\hat{f}(x)] = \mathbb{E}[\hat{f}(x)^2] - \mathbb{E}[\hat{f}(x)]^2$$

and the expectation  $\mathbb{E}[\cdot]$  ranges over different choices of the training set  $(x_1, x_2, \dots, x_n; y_1, y_2, \dots, y_n)$ , all sampled from the same joint distribution  $P(X, Y)$ .

3. Explain what errors the various terms in equation (1) represent (for instance, under what circumstances is a particular term large or small?).
4. Explain why this decomposition is also known as the ‘bias-variance trade-off’.

### 0.3 OLS, linear projection, and gradient descent.

Suppose now that we have a training set  $\mathbf{X}$  consisting of  $n$  vectors (datapoints) of size  $m$  (features). Associated to this we have an  $n$ -dimensional vector of real values  $\mathbf{y}$ . Suppose we want to regress a linear model to this data using ordinary least-squares (OLS). That is, we fit linear model  $f_\beta(\mathbf{X}) = \mathbf{X}\beta$ , such that we minimise  $(\mathbf{y} - f_\beta(\mathbf{X}))^2$  over the parameters  $\beta$ .

1. What is the dimensionality of the parameter vector  $\beta$ ?
2. Show by differentiation that the OLS estimator  $\hat{\beta}$  equals  $(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y}$ .

Another method for deriving the OLS estimator is by thinking geometrically. Imagine  $\mathbf{y}$  and  $\mathbf{X}\beta$  as vectors in an  $n$ -dimensional vector space. Further note that the regressors  $\mathbf{X}\beta$  actually span an  $m$ -dimensional subspace of this larger vector space (the column space of  $\mathbf{X}$ ), see Figure 1.

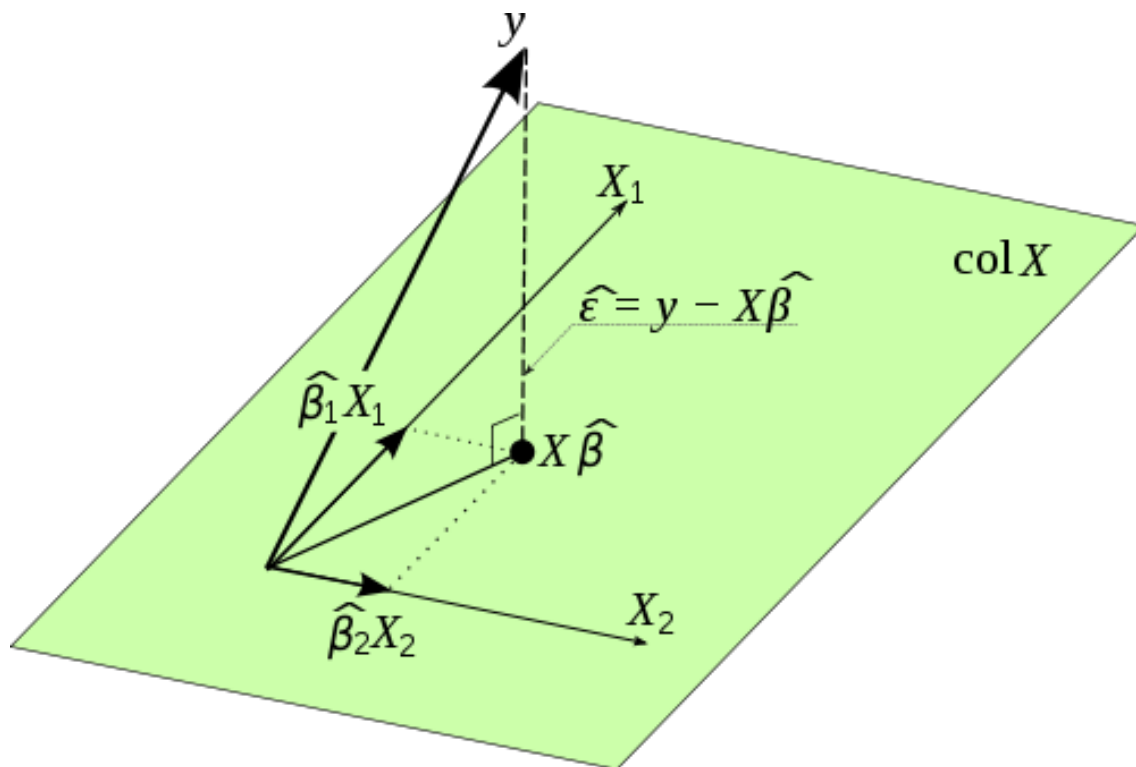


Figure 1: Geometric representation of OLS, from Wikipedia. The  $X_i$  are columns of  $\mathbf{X}$ .

3. Define as  $\epsilon_\beta$  the residual vector  $(\mathbf{y} - \mathbf{X}\beta)$ . Argue that minimising the  $L_2$  norm of this vector over  $\beta$  (as we do in OLS) is equivalent to choosing  $\beta$  such that  $\mathbf{X}\beta = P(\mathbf{y})$ , where  $P(\cdot)$  orthogonally projects  $\mathbf{y}$  onto the linear subspace spanned by the regressors.
4. Argue that this is equivalent to setting  $\mathbf{X}^T \epsilon_\beta = 0$ .
5. Show that from this also follows that  $\hat{\beta} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y}$ .

When our models get more complicated, we quickly lose the ability to optimise them analytically. Instead, we use numerical optimisation methods such as gradient descent. Although we do not need gradient descent to fit the model described above, we can still do so.

6. In gradient descent we minimise a loss function  $L_\beta(\mathbf{y}, \mathbf{X})$  over the parameter values  $\beta$ . What is the loss function corresponding to the OLS description from before?
7. Assume we use a learning rate  $\alpha$ . Write the update rule for a single gradient descent step on our parameters  $\beta$ .

# Lecture 1: Introduction

## 1.1 Introduction

1. Explain what is meant by the ‘curse of dimensionality’.
2. Suppose you are trying to design a predator agent that can learn to catch a randomly moving prey on a  $5 \times 5$  **toroidal** grid. You have been given the  $(x, y)$ -coordinates of the predator and the  $(x, y)$ -coordinates of the prey to use as the state.
  - (a) How many possible states are there in this naive approach?
  - (b) There is a way of reducing the state space considerably a priori. Write down how you would adapt the given state representation to reduce the size of the state space. Note that you only care about a representation that you can use to solve the problem.
  - (c) How many possible states are there now?
  - (d) What is the advantage of doing this?
  - (e) Consider the Tic-Tac-Toe example in Chapter 1.5 of the book. Here, too, we can exploit certain properties of the problem to reduce the size of the state space. Give an example of a property you can exploit.
3. Suppose you want to implement a Reinforcement Learning agent that learns to play Tic-Tac-Toe, as outlined in Chapter 1 of the book.
  - (a) Which agent do you think would learn a better policy in the end: a greedy agent that always chooses the action it currently believes is best, or a non-greedy agent that sometimes tries new actions? Why?
4. Assume we start with an exploration rate of  $\epsilon$ , meaning that whenever the agent chooses an action, it has a probability of  $\epsilon$  to pick an action at random, and a probability of  $1 - \epsilon$  to pick the greedy action. If we assume the environment has been sufficiently explored, we may want to reduce the amount of exploration after some time.
  - (a) Write down how you would do this.
  - (b) Does your method work if the opponent changes strategies? Why/why not? If not, provide suggestions on a heuristic that can adapt to changes in the opponent’s strategy.

## 1.2 Exploration

1. In  $\epsilon$ -greedy action-selection for the case of  $n$  actions, what is the probability of selecting the greedy action?
2. Consider a 3-armed bandit problem with actions 1, 2, 3. If we use  $\epsilon$ -greedy action-selection, initialization at 0, and **sample-average** action-value estimates, which of the following sequence of actions are certain to be the result of exploration?  $A_0 = 1, R_1 = -1, A_1 = 2, R_2 = 1, A_2 = 2, R_3 = -2, A_3 = 2, R_4 = 2, A_4 = 3, R_5 = 1$ .
3. You are trying to find the optimal policy for a two-armed bandit. You try two approaches: in the pessimistic approach, you initialize all action-values at -5, and in the optimistic approach you initialize all action-values at +5. One arm gives a reward of +1, one arm

gives a reward of -1. Using a greedy policy to choose actions, compute the resulting Q-values for both actions after three interactions with the environment. In case of a tie between two Q-values, break the tie at random. *Note*: the initialization is *not* a sample.

4. Which initialization leads to a higher (undiscounted) return? What if you had broken the tie differently?
5. Which initialization leads to a better estimation of the Q-values?
6. Explain why one of the two initialization methods is better for exploration.

# Lecture 2: MDPs and dynamic programming

## 2.1 Markov Decision Processes

1.
  - (a) For the first four examples outlined in Section 1.2 of the book, describe the state space, action space and reward signal.
  - (b) Come up with an example of your own that you might model with an MDP. State the action space, state space and reward signal.
  - (c) Come up with an example for a problem that you might have trouble solving with an MDP. Why doesn't this fit the framework?
  - (d) In mazes, the agent's position is often seen as the state. However, the agent's position alone is not always a sufficient description. Come up with an example where the state consists of the agent's location and one or more other variables.
  - (e) Consider the example in exercise 3.3 of the book. Why might you choose to view the actions as handling the accelerator, brake and steering wheel? What is the disadvantage of doing this?
  - (f) Why might you choose to view the actions as choosing where to drive? What is the disadvantage of doing this?
  - (g) Can you think of some way to combine both approaches?
2.
  - (a) Eq. 3.8 in the book gives the discounted return for the continuous case. Write down the formula for the discounted return in the episodic case.
  - (b) Show that  $\sum_{k=0}^{\infty} \gamma^k = \frac{1}{1-\gamma}$  if  $0 \leq \gamma < 1$ . (Hint: if you're stuck, have a look at the Wikipedia page on *geometric series*)
  - (c) Consider exercise 3.7 in the book. Why is there no improvement in the agent?
  - (d) How would adding a discount factor of  $\gamma < 1$  help solve this problem?
  - (e) How might changing the reward function help solve this problem?

## 2.2 Homework: Dynamic Programming

1. Write the value,  $v^{\pi}(s)$ , of a state  $s$  under policy  $\pi$ , in terms of  $\pi$  and  $q^{\pi}(s, a)$ . Write down both the stochastic and the deterministic policy case.
2. In Policy Iteration, we first evaluate the policy by computing its value function, and then update it using a Policy Improvement step. You will now change Policy Iteration as given on page 80 of the book to compute action-values. First give the new policy evaluation update in terms of  $Q^{\pi}(s, a)$  instead of  $V^{\pi}(s)$ . Note that Policy Improvement uses deterministic policies.
3. Now change the Policy Improvement update in terms of  $Q^{\pi}(s, a)$  instead of  $V^{\pi}(s)$ .
4. The Value Iteration update, given in the book in Eq. 4.10 can also be rewritten in terms of Q-values. Give the Q-value Iteration update.

# Lecture 3: Monte Carlo methods

## 3.1 Homework: Monte Carlo

1. Consider an MDP with a single state  $s_0$  that has a certain probability of transitioning back onto itself with a reward of 0, and will otherwise terminate with a reward of 5. Your agent has interacted with the environment and has gotten the following three trajectories:  $[0, 0, 5]$ ,  $[0, 0, 0, 0, 5]$ ,  $[0, 0, 0, 5]$ . Use  $\gamma = 0.9$ .
  - (a) Estimate the value of  $s_0$  using first-visit MC.
  - (b) Estimate the value of  $s_0$  using every-visit MC.
2. What is a disadvantage of using *ordinary importance sampling* in off-policy Monte Carlo?
3. What is a disadvantage of using *weighted importance sampling* in off-policy Monte Carlo?



# Lecture 4: Temporal Difference methods

## 4.1 Temporal Difference Learning (Application)

Consider an undiscounted Markov Decision Process (MDP) with two states A and B, each with two possible actions 1 and 2, and a terminal state T with  $V(T) = 0$ . The transition and reward functions are unknown, but you have observed the following episode using a random policy:

$$\bullet A \xrightarrow[r_3=-3]{a_3=1} B \xrightarrow[r_4=4]{a_4=1} A \xrightarrow[r_5=-4]{a_5=2} A \xrightarrow[r_6=-3]{a_6=1} B \xrightarrow[r_7=1]{a_7=2} T$$

where the arrow ( $\rightarrow$ ) indicates a transition and  $a_t$  and  $r_t$  take the values of the observed actions and rewards respectively.

1. What are the state(-action) value estimates  $V(s)$  (or  $Q(s, a)$ ) after observing the sample episode when applying:
  - (a) TD(0)
  - (b) 3-step TD
  - (c) SARSA
  - (d) Q-learning

where we initialize state values to 0 and use a learning rate  $\alpha = 0.1$

2. Choose a deterministic policy that you think is better than the random policy given the data. Refer to any of the state(-action) value estimates to explain your reasoning.
3. Let  $\pi_{random}$  denote the random policy used so far and  $\pi_{student}$  denote the new policy you proposed. Suppose you can draw new sample episodes indefinitely until convergence of the value estimates.
  - (a) Discuss how do you expect the final value estimates to differ if you ran Q-Learning with  $\pi_{random}$  as compared to  $\pi_{student}$ .
  - (b) What problems may arise with  $\pi_{random}$  or  $\pi_{student}$  respectively?
  - (c) Do you think using an  $\epsilon$ -greedy policy as behavior policy would be beneficial? Explain why/why not?

## 4.2 Contraction Mapping

Consider the following Theorem, from Csaba Szepesvari's *Algorithms for Reinforcement Learning*:

**Theorem 1.** (Banach's Fixed Point Theorem). Let  $V$  be a Banach space and  $T : V \rightarrow V$  be a contraction mapping. Then  $T$  has a unique fixed point. Further, for any  $v_0 \in V$ , if  $v_{n+1} = Tv_n$ , then  $v_n \rightarrow_{\|\cdot\|} v$ , where  $v$  is the unique fixed point of  $T$  and the convergence is geometric:

$$\|v_n - v\| \leq \gamma^n \|v_0 - v\| \tag{2}$$

1.
  - (a) Consider the contraction mapping  $T(x) := 1 + \frac{1}{3}x$ . Find the fixed point of this mapping.
  - (b) Consider the contraction mapping  $(Tf)(s) := \frac{-1}{2}f(s) + g(s)$ . Find the fixed point of this mapping in terms of  $g(s)$ .

2. Consider an MDP with finite state space  $\mathcal{S}$ , finite action space  $\mathcal{A}$ , and discount factor  $\gamma$ . Let  $V$  be the value function which gives the value per state, and  $R(s, a)$  the expected immediate reward for action  $a$  in state  $s$  and  $Pr(s'|s, a)$  the probability of transitioning to state  $s'$  after taking action  $a$  in state  $s$ . Then the *Bellman optimality operator*  $B^*$  is given by

$$(B^*V)(s) = \max_a \left[ R(s, a) + \gamma \sum_{s' \in \mathcal{S}} Pr(s'|s, a)V(s') \right] \quad (3)$$

(a) Show that

$$\left| \max_z f(z) - \max_z h(z) \right| \leq \max_z |f(z) - h(z)| \quad (4)$$

for arbitrary  $f(z)$  and  $h(z)$ .

Let  $(X, \|\cdot\|)$  be a metric space.  $T : X \rightarrow X$  is said to be a contraction if there exists a constant  $0 \leq c < 1$  such that

$$\|T(x) - T(y)\| \leq c\|x - y\| \quad \forall x \in X, \forall y \in X \quad (5)$$

In our case, the metric space is  $(\mathcal{B}(\mathcal{S}), \|\cdot\|_\infty)$ , where  $\mathcal{B}(\mathcal{S})$  is the space of bounded functions on  $\mathcal{S}$ .  $\mathcal{B}(\mathcal{S}) = \{V : \mathcal{S} \rightarrow \mathbb{R} : \|V\|_\infty < +\infty\}$ .  $(\mathcal{B}(\mathcal{S}), \|\cdot\|_\infty)$  is a normed vector space. Thus, the distance between value functions is given by the supremum norm  $\|V_1 - V_2\|_\infty = \max_{s \in \mathcal{S}} |V_1(s) - V_2(s)|$ , or the largest difference between state values.

(b) Show that  $B^*$  is a contraction mapping in supremum norm, i.e.

$$\|(B^*V_1) - (B^*V_2)\|_\infty \leq c\|V_1 - V_2\|_\infty \quad (6)$$

(c) Why is this an important result?

### 4.3 Temporal Difference Learning (Theory)

1. We can use Monte Carlo to get value estimates of a state with  $V_M(S) = \frac{1}{M} \sum_{n=1}^M G_n(S)$  where  $V_M(S)$  is the value estimate of state  $S$  after  $M$  visits of the state and  $G_n(S)$  the return of an episode starting from  $S$ . Show that  $V_M(S)$  can be written as the update rule  $V_M(S) = V_{M-1}(S) + \alpha_M[G_M(S) - V_{M-1}(S)]$  and identify the learning rate  $\alpha_M$ .
2. Consider the TD-error

$$\delta_t = R_{t+1} + \gamma V(S_{t+1}) - V(S_t). \quad (7)$$

- (a) What is  $\mathbb{E}[\delta_t | S_t = s]$  if  $\delta_t$  uses the true state-value function  $V^\pi$
  - (b) What is  $\mathbb{E}[\delta_t | S_t = s, A_t = a]$  if  $\delta_t$  uses the true state-value function  $V^\pi$
3. The Monte-Carlo error can be written as the sum of TD-errors if the value estimates don't change (cf. equation 6.6 in the book) as

$$G_t - V(S_t) = \sum_{k=t}^{T-1} \gamma^{k-t} \delta_k. \quad (8)$$

(a) Show that the n-step error as used in

$$V_{t+n}(S_t) = V_{t+n-1}(S_t) + \alpha[G_{t:t+n} - V_{t+n-1}(S_t)], \quad 0 \leq t < T \quad (9)$$

can also be written as a sum TD errors (assuming the value estimates don't change).

#### 4.4 Homework: Maximization Bias

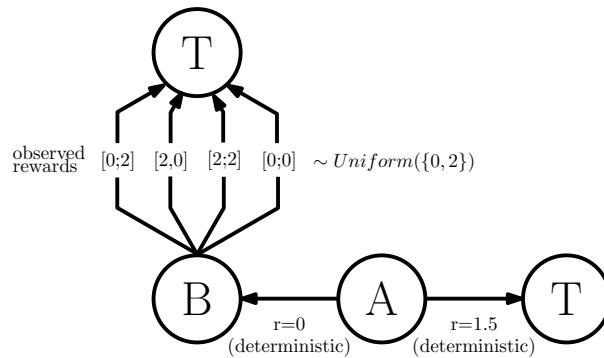


Figure 2: MDP: Maximization Bias

Consider the undiscounted MDP in Figure 2 where we have a state A with two actions, one ending in the terminal state T (with  $V(T) = 0$ ) and always giving a reward of 1.5 and another action that transitions to state B with zero reward. From state B we can take a total of four different actions each transitioning to the terminal state and giving a reward of either 0 or 2 with equal probability. Suppose we sample all possible episodes two times and record the rewards. The observed rewards for each of the four actions from B are indicated in the Figure, e.g. the leftmost action received one time a reward of 0 and one time a reward of 2.

1. We repeatedly apply Q-learning and SARSA on the observed data until convergence. Give all final state-action values for Q-learning and SARSA respectively.
2. This problem suffers from maximization bias. Explain where this can be observed. Do both Q-learning and SARSA suffer from this bias? Why/why not?
3. To circumvent the issue of maximization bias, we can apply Double Q-learning. Use the given example to explain how Double Q-learning alleviates the problem of maximization bias.
4. What are the true state-action values that we would expect to get (after convergence) if we continued sampling episodes.