Foundations of Reinforcement Learning
Assignment 1

Issue date: October 11, 2021
Due date: October 29, 2021

Coverage: Basics of MDP, Bellman equation, value iteration, policy iteration

Instructions

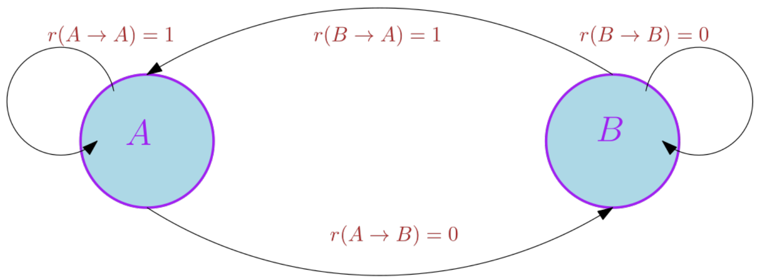
- Where to submit: Please submit your solution as a PDF on Moodle. File name should follow the format `Assignment1-Lastname-Firstname.pdf`.
- How to write solutions: You should type your solution using LaTeX and following the template. Handwritten solutions will not be graded. Keep in mind the following premise:
  - When writing in English, write short, simple sentences.
  - When writing a proof, write clear, precise statements.

  You can use previous points of the same problem without proving them. You can use results from the lectures if you reference them properly.
- Discussion: You may discuss only at a high level with classmates. You should not dig around for homework solutions; if you do rely upon external resources, cite them, and write solutions in your own words. We ask you to please follow the ETH Disciplinary Code.
- Grading: Grading will be based on the completeness and correctness of your solution according to points assigned for each exercise. `Final grade = min(regular points + bonus points, 100)`.

  We reserve the right to deduct points on sloppy LaTeX, minor errors in calculations, and unclear writing in general.
- Re-grading: You may request for regrading within one week after the grade is released, with a written justification of why your solution deserves more points.
- Encountering problems?
  - If you think some exercise is unclear or wrong use the Forum *Assignments* in Moodle or reach out to the TA in charge of the exercise.
  - If you have trouble submitting your solution to Moodle within six hours before the deadline due to technical problems, you can send your PDF solution to our head TA.

1 Basic Concepts (30 points)

Consider a Markov decision process with deterministic dynamics on two states A and B. In both states, there are two actions: stay or switch. The reward for entering A is 1, and the reward for entering B is 0.



In other words, we have $\mathcal{S} = \{A, B\}$, $\mathcal{A} = \{\text{switch}, \text{stay}\}$,

$$r(A, \text{switch}) = 0, r(A, \text{stay}) = 1, r(B, \text{switch}) = 1, r(B, \text{stay}) = 0,$$

$$P(A|A, \text{switch}) = 0, P(A|A, \text{stay}) = 1, P(B|B, \text{switch}) = 0, P(B|B, \text{stay}) = 1.$$

Let $\gamma \in (0, 1)$ be the discount factor. Let π_0 be the policy which always switches states.

a) **[5 points]** Compute the value function $\mathbf{V}^{\pi_0}$ of π_0.

b) **[5 points]** Compute the optimal value function $\mathbf{V}^*$ and optimal policy π^*.

c) **[5 points]** Compute $\mathcal{T}\mathbf{V}^{\pi_0}$ where $\mathcal{T}$ is the Bellman optimality operator.

d) **[5 points]** Compute the greedy policy based on $\mathbf{V}^{\pi_0}$.

e) **[5 points]** Compute the first 5 updates of Value Iteration algorithm initialized with $\mathbf{V}^{\pi_0}$.

f) **[5 points]** Compute the first 5 updates of Policy Iteration algorithm initialized with π_0.

Remark 1 *What happens if we evaluate the first* $10, 20, 30\dots$ *steps of Value Iteration? This example also implies that iterating* n *times the Bellman optimality operator is not guaranteed to lead exactly to* $\mathbf{V}^*$ *for any finite* $n \in \mathbb{N}$.

Contact: `nuria.armengolurpi@inf.ethz.ch`

Solution. Put your solution here.

2 Convergence of Policy Iteration (15 points)

Consider the policy iteration algorithm for infinite horizon MDPs with a discount factor $\gamma < 1$. Let π_t and π_{t+1} be the respective policies at time steps t and $t+1$, where π_{t+1} is the greedy policy based on the value function $\mathbf{V}^{\pi_t}$. From the above example (see Exercise 1), one can observe that the Bellman optimality operator applied to V^{π_t} does not in general yield $\mathbf{V}^{\pi_{t+1}}$, i.e., $\mathcal{T}\mathbf{V}^{\pi_t} \neq \mathbf{V}^{\pi_{t+1}}$.

In slide 39/53 of Lecture 2, there is a typo:

$$\|\mathbf{V}^{\pi_{t+1}} - \mathbf{V}^*\|_\infty = \|\mathcal{T}\mathbf{V}^{\pi_t} - \mathcal{T}\mathbf{V}^*\|_\infty \tag{2.1}$$

This should instead be

$$\|\mathbf{V}^{\pi_{t+1}} - \mathbf{V}^*\|_\infty \leq \|\mathcal{T}\mathbf{V}^{\pi_t} - \mathcal{T}\mathbf{V}^*\|_\infty \tag{2.2}$$

Here we aim to prove the linear convergence of policy iteration in a precise manner.

a) **[5 points]** Prove

$$\mathcal{T}\mathbf{V}^{\pi_t}(s) \geq \mathbf{V}^{\pi_t}(s) \qquad \forall s \in \mathcal{S}.$$

b) **[5 points]** Prove

$$\mathbf{V}^{\pi_{t+1}}(s) \geq \mathcal{T}\mathbf{V}^{\pi_t}(s) \qquad \forall s \in \mathcal{S}.$$

c) **[5 points]** Using the above, prove the linear convergence of the policy iteration algorithm.

Contact: daniel.paleka@math.ethz.ch

Solution. Put your solution here.

3 Bounding Suboptimality via Bellman Error (35 points)

Consider a tabular MDP with finite state space $\mathcal{S}$ and action space $\mathcal{A}$.

a) **[10 points]** Recall the Bellman optimality operator $\mathcal{T}$ and the Bellman expectation operator $\mathcal{T}^\pi$ from the lecture, defined on the space of value functions. Formally define analogous operators $\mathcal{T}_Q$ and $\mathcal{T}_Q^\pi$ on the space of state-action value Q-functions.

b) **[10 points]** For any state-action value functions $\mathbf{Q}, \mathbf{Q}'$, prove:

$$\left| \max_{a \in \mathcal{A}} \mathbf{Q}(s,a) - \max_{a \in \mathcal{A}} \mathbf{Q}'(s,a) \right| \le \|\mathbf{Q} - \mathbf{Q}'\|_\infty \forall s \in \mathcal{S}.$$

Show both defined operators are γ-contractions under the ℓ_∞-norm.

c) **[5 points]** For any state-action value function $\mathbf{Q}$, prove:

$$\|\mathbf{Q} - \mathbf{Q}^*\|_\infty \le \frac{\|\mathbf{Q} - \mathcal{T}_Q \mathbf{Q}\|_\infty}{1-\gamma}.$$

d) **[10 points]** Now we are ready for the main point. Let Q be a state-action value function, and let π be the greedy policy with respect to Q, that is, $\pi = \arg\max_a Q(\cdot, a)$. Show that the value function V^π for this policy satisfies

$$\|\mathbf{V}^\pi - \mathbf{V}^*\|_\infty \le \frac{2\|\mathbf{Q} - \mathcal{T}_Q \mathbf{Q}\|_\infty}{1-\gamma}.$$

where $\mathcal{T}_Q$ is the Bellman optimality operator on the space of state-action value Q-functions.

Contact: `nuria.armengolurpi@inf.ethz.ch`

Solution. Put your solution here.

4 Reading: The Value Function Polytope (20 points)

In the reading exercise, you are expected to read some material and write a paragraph for each question. The questions may not be well-posed, and are intended to encourage reading and thinking; the grading here will be more lenient than in the previous problems.

In [**?**] the authors characterize the geometry of the space of all possible value functions given a Markov Decision Process. Then, they illustrate the dynamics of RL algorithms in the value function space, including value iteration, policy iteration, policy gradient methods, etc. Read the paper [**?**] and think about the following questions. [1]

a) **[5 points]** On the first page, Figure 1 shows a convex space of policies mapped into a non-convex space of value functions. Why does this happen, philosophically? (You can think of your explanation, as we do not have one single correct answer.)

b) **[5 points]** In Theorem 1, the authors consider the set of policies that differ only in one state s. Why is it necessary that policies differ only in a single state? What happens if we interpolate between two policies that differ in more states instead of only in a single one? (Answer in simple heuristic terms, no proofs needed.)

   Hint: look at Lemma 3.

c) **[10 points]** What is your take-away from this paper? What is the main limitation of the paper?

Contact: `daniel.paleka@math.ethz.ch`

Solution. Put your solution here.

[1] Extending [**?**] by investigating the interplay between non-convexity and dynamics may be a nice idea for the course project.

5 Bonus: Convergence of Inexact Policy Iteration (10 points)

Consider a tabular MDP with finite state space $\mathcal{S}$ and action space $\mathcal{A}$ and discount factor γ. In Exercise 2, we proved the linear convergence of the policy iteration algorithm.

Now consider the "inexact" version of the policy iteration algorithm, i.e., in each step we compute a function $\mathbf{V}_t$ such that

$$\max_{s \in \mathcal{S}} |\mathbf{V}_t(s) - \mathbf{V}^{\pi_t}(s)| \le \varepsilon$$

for all steps $t \ge 0$ and some $\varepsilon > 0$. Then the algorithm sets π_{t+1} to the greedy policy with respect to $\mathbf{V}_t$. We can interpret ε as an error incurred during the policy evaluation step, as for example errors due to simulation.
Show that the sequence of policies π_t generated by this algorithm satisfies:

$$\limsup_{t \to \infty} \max_{s \in \mathcal{S}} |\mathbf{V}^{\pi_t}(s) - \mathbf{V}^*(s)| \le \frac{2\gamma\varepsilon}{(1-\gamma)^2}.$$

Contact: daniel.paleka@math.ethz.ch

Solution. Put your solution here.

References

[1] Robert Dadashi, Adrien Ali Taiga, Nicolas Le Roux, Dale Schuurmans, and Marc G Bellemare. The value function polytope in reinforcement learning. In International Conference on Machine Learning, pages 1486–1495. PMLR, 2019.