CMPUT 605: Theoretical Foundations of Reinforcement Learning, Winter 2023
Homework #2

Instructions

Submissions You need to submit a single PDF file, named `p02_<name>.pdf` where `<name>` is your name. The PDF file should include your typed up solutions (we strongly encourage to use pdfLaTeX). Write your name in the title of your PDF file. We provide a LaTeXtemplate that you are encouraged to use. To submit your PDF file you should send the PDF file via private message to Vlad Tkachuk on Slack before the deadline.

Collaboration and sources Work on your own. You can consult the problems with your classmates, use books or web, papers, etc. Also, the write-up must be your own and you must acknowledge all the sources (names of people you worked with, books, webpages etc., including class notes.) Failure to do so will be considered cheating. Identical or similar write-ups will be considered cheating as well. Students are expected to understand and explain all the steps of their proofs.

Scheduling Start early: It takes time to solve the problems, as well as to write down the solutions. Most problems should have a short solution (and you can refer to results we have learned about to shorten your solution). Don't repeat calculations that we did in the class unnecessarily.

Deadline: February 12 at 11:55 pm

Problems

Union bounds

Question 1. Let $A_1, \dots, A_n$ be events of a probability space $(\Omega, \mathcal{F}, \mathbb{P})$. Note that finite (and actually discrete) sets are always equipped with the discrete σ-algebra (power set) unless otherwise specified. Show that the following hold:

1. Show that for any random variable I taking values in $[n]$, A_I, which is naturally defined as

$$A_I = \{\omega \in \Omega \,:\, \omega \in A_{I(\omega)}\},$$

is an event.

5 points

2. Show that there exist a random variable I taking values in $[n]$, such that $\mathbb{P}(A_I) = \mathbb{P}(\cup_{i=1}^n A_i)$.

10 points

3. Show that the first two claims hold even if I takes values in $\{1, 2, \dots\}$ and $(A_i)_{i=1,2,\dots}$ is a countably infinite sequence of events. (It suffices to explain which parts of the solution to the first two questions need to be changed.)

5 points

Total: **20 points**

Online planning revisited

In the next problem we consider the variant of online planner that uses a fresh sample in each call of function q. In particular, consider the following algorithm:

```
define q(k,s):
if k = 0 return [0 for a in A] # base case
return [ r(s,a) + gamma/m * sum( [max(q(k-1,s')) for s' in C(k,s,a)] ) for a in A ]
end
```

Here, the lists C(k,s,a), which in what follows will be denoted by $C_k(s,a)$ are as usual: They are created independently of each other for each (s,a) and k and they have m mutually independent elements, sampled from $P_a(s)$. In particular, $C_k(s,a) = [S'_1(k,s,a), \dots, S'_m(k,s,a)]$ where $(S'_i(k,s,a)) \overset{\text{iid}}{\sim} P_a(s)$. The planner is used the same way as before: when asked for an action at state s_0, it returns $\arg\max_{a\in\mathcal{A}} q(k,s_0)$ with an appropriate choice of k (and m).

Let $\hat{T}_k : \mathbb{R}^{\mathcal{S}\times\mathcal{A}} \to \mathbb{R}^{\mathcal{S}\times\mathcal{A}}$ be defined by

$$\hat{T}_k q(s,a) = r_a(s) + \frac{\gamma}{m} \sum_{s' \in C_k(s,a)} \max_{a'} q(s',a')\,.$$

Question 2. Assume that the rewards belong to the $[0,1]$ interval. Show that the following hold:

1. For $k \ge 0$, let $Q_k(s,\cdot)$ be the values returned by the call q(k,s) with a particular value of s and k. Show that $Q_k(s,\cdot) = \hat{T}_k \dots \hat{T}_1 \mathbf{0}(s,\cdot)$.

   5 points

2. Fix $H > 0$. Define a sequence of sets $\mathcal{S}_0, \dots, \mathcal{S}_H$ with $|\mathcal{S}_h| = O((mA)^h)$ and $\mathcal{S}_0 = \{s_0\}$ such that with $\delta_h = \|Q_h - q^*\|_{\mathcal{S}_{H-h}}$, the following hold for any $0 \le h \le H$:

   (a) If also $h > 0$, $\delta_h \le \gamma\delta_{h-1} + \|\hat{T}_h q^* - q^*\|_{\mathcal{S}_{H-h}}$;

   5 points

   (b) If also $h < H$, $\mathcal{S}_{H-h}$ is a function of $C_H, \dots, C_{h+1}$ only (and is not a function of $C_h, \dots, C_1$).

   5 points

3. Show that with probability $1-\zeta$, $\|\hat{T}_h q^* - q^*\|_{\mathcal{S}_{H-h}} \le \frac{1}{1-\gamma}\sqrt{\frac{\log(2|\mathcal{S}_{H-h}||A|/\zeta)}{2m}}$.

   10 points

4. Let π be the policy induced by the modified planner. Give a bound on the suboptimality of π (make it as tight as you can using the usual tools).

   10 points

5. Compare the bound to the one we obtained for the case when the same sets are used in the algorithm throughout.

   5 points

6. Bound the computational complexity of the algorithm; argue why one would call this the "sparse lookahead tree approach".

5 points

Total: **45 points**

Tightness of performance bounds of greedy policies

Error bounds for greedy policies are at the heart of many of the upper bounds we obtained. Here you will be asked to show that these bounds are unimprovable. For example, in Lecture 6, the following is stated in Part II of the "Policy error bound - I." lemma:

Lemma 1. Let π be a memoryless policy and choose a function $q : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ and $\varepsilon \geq 0$. Then, if π is greedy with respect to q then

$$v^{\pi} \geq v^{*} - \frac{2\|q - q^{*}\|_{\infty}}{1-\gamma}\mathbf{1}\,.$$

Question 3. Show that for any $\gamma \in [0, 1)$ and $\varepsilon > 0$ there is a finite discounted MDP $M = (\mathcal{S}, \mathcal{A}, P, r, \gamma)$ and $q : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ such that the following hold:

1. $\|q - q^{*}\|_{\infty} = \varepsilon$;
2. There is policy π that is greedy with respect to q such that $\|v^{\pi} - v^{*}\|_{\infty} = \frac{2\varepsilon}{1-\gamma}$.

Total: **10 points**

Total for all questions: 75. Of this, 10 are bonus marks. Your assignment will be marked out of 65.