CMPUT 653: Theoretical Foundations of Reinforcement Learning, Winter 2022
Homework #2

Instructions

Submissions You need to submit a zip file, named `p02_<name>.zip` where `<name>` is your name. The zip file should include a report in PDF, typed up (we strongly encourage to use pdfLaTeX) and the code that we asked for. Write your name on your solution. I provide a template that you are encouraged to use. You have to submit the zip file on the eclass website of the course.

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 14 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**

Local planning revisited

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

1. `define q(k,s):`
2. `if k = 0 return [0 for a in A] # base case`
3. `return [ r(s,a) + gamma/m * sum( [max(q(k-1,s')) for s' in C(k,s,a)] ) for a in A ]`
4. `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 $(S'_i(k,s,a)) \stackrel{\text{iid}}{\sim} P_a(s)$. 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 $\hat{T}_k : \mathbb{R}^{\mathcal{S}\times\mathcal{A}} \to \mathbb{R}$ 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 $k \geq 0$, 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 $0 \leq h \leq H$:

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

   5 points

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

   5 points

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

   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**

Fitted Value Iteration

Assume that the rewards belong to the $[0,1]$ interval and fix the discount factor γ. Let $H_\gamma = 1/(1-\gamma)$. Assume we are given a feature map $\varphi : \mathcal{S} \times \mathcal{A} \to \mathbb{R}^d$ which spans $\mathbb{R}^d$. Let $\mathcal{F} = \{f_\theta : f_\theta(s,a) = \varphi(s,a)^\top \theta, \theta \in \mathbb{R}^d\}$ be the span of the features. Let $C \subset \mathcal{Z} := \mathcal{S} \times \mathcal{A}$ be the set whose existence is guaranteed by the Kiefer-Wolfowitz theorem for the feature map φ and let $\rho : C \to [0,1]$ be the corresponding weighting function. In particular, $|C| \le d(d+1)/2$, $\sum_{z \in C} \rho(z) = 1$ and with $G_\rho = \sum_{z \in C} \rho(z)\varphi(z)\varphi(z)^\top$, $\max_{z \in \mathcal{Z}} \|\varphi(z)\|_{G_\rho^{-1}} \le \sqrt{d}$.

For $k \ge 1$, $(s,a) \in \mathcal{S} \times \mathcal{A}$, let $C_k(s,a) = [S'_1(k,s,a), \dots, S'_m(k,s,a)]$ be so that all the $(C_k(s,a))_{k,s,a}$ are independent of each other, and for any k, s, a, $S'_1(k,s,a), \dots, S'_m(k,s,a) \overset{\text{iid}}{\sim} P_a(s)$. For $k \ge 1$ let $\hat{T}_k : \mathbb{R}^{\mathcal{S} \times \mathcal{A}} \to \mathbb{R}$ be defined by

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

Further, let $\Pi : \mathbb{R}^{\mathcal{Z}} \to \mathbb{R}$ be defined by $(\Pi f)(z) = \max(\min(f(z), H_\gamma), 0)$: In words, Π truncates the values of its argument to the $[0, H_\gamma]$ interval.

Consider the following procedure, which we call fitted q iteration (FQI).[1]

1. $\theta_0 = \mathbf{0}$
2. `for` $k = 1, 2, \dots, K$ `do`
3. $\quad \theta_k = \operatorname{argmin}_{\theta \in \mathbb{R}^d} \sum_{z \in C} \rho(z)(f_\theta(z) - (\hat{T}_k \Pi f_{\theta_{k-1}})(z))^2$
4. `return` θ_K

Let $\varepsilon_{\text{apx}} = \sup_\theta \inf_{\theta'} \|f_{\theta'} - T\Pi f_\theta\|_\infty$.

Question 3. Prove that the following hold:

1. The computation cost of FQI is $O(Kd^3mA)$ and it needs $O(d^2)$ space (all in the RAM model of computation). The query cost is $O(Kd^2m)$. Explain how you get the bounds.

5 points

2. Fix $k \ge 0$. Let $q_k = \Pi f_{\theta_k}$. For $k > 0$, let $\varepsilon_k : \mathcal{Z} \to R$ and $\theta_k^* \in \mathbb{R}^d$ be such that $Tq_{k-1} = f_{\theta_k^*} + \varepsilon_k$ and $\|\varepsilon_k\|_\infty \le \varepsilon_{\text{apx}}$. Show that ε_k and θ_k^* are well-defined (i.e., they exist).

10 points

[1] A terrible name.

3. Show that for any $k \geq 1$, $0 \leq \zeta \leq 1$, with probability at least $1 - \zeta$,

$$\|q_k - Tq_{k-1}\|_\infty \leq (1 + \sqrt{d})\varepsilon_{\text{apx}} + \sqrt{d} H_\gamma \sqrt{\frac{\log\left(\frac{2|C|}{\zeta}\right)}{2m}} .$$

10 points

4. Show that, on the same event as in the previous part, the policy π that is greedy with respect to q_K is δ-optimal with

$$\delta \leq 2H_\gamma^2 \left\{ (1 + \sqrt{d})\varepsilon_{\text{apx}} + \gamma^K + \sqrt{d} H_\gamma \sqrt{\frac{\log\left(\frac{2|C|K}{\zeta}\right)}{2m}} \right\} .$$

10 points

5. Fix $\varepsilon > 0$. Argue that K, m and ζ can be chosen as a polynomial function of $H_\gamma, d, 1/\varepsilon$ so that the expected suboptimality of the policy π is bounded by $2H_\gamma^2(1 + \sqrt{d})\varepsilon_{\text{apx}} + 2\varepsilon$. Show the choices you made.

5 points

6. Argue that with a query, runtime and space cost that is polynomial in $H_\gamma, d, 1/\varepsilon, A$, the procedure obtains a policy π that is at most δ-optimal with $\delta = 2H_\gamma^2(1 + \sqrt{d})\varepsilon_{\text{apx}} + 2\varepsilon$.

5 points

7. The MDP $\mathcal{M} = (\mathcal{S}, \mathcal{A}, P, r, \gamma)$ is called linear in φ if it holds that with some $\theta_r \in \mathbb{R}^d$, $r_a(s) = f_{\theta_r}(s, a)$ holds for all (s, a) and if for some $\mu : \mathcal{S} \to \mathbb{R}^d$, for any (s, a), $P_a(s, s') = \langle \varphi(s, a), \mu(s') \rangle$. Show that if $\mathcal{M}$ is linear in φ then $\varepsilon_{\text{apx}} = 0$.

10 points

Total: **55 points**

Total for all questions: 120. Of this, up to 20 can be bonus marks You can receive bonus marks by asking/upvoting questions, for a total of 20 bonus marks! You must ask at least one question in one of the Lecture Discussion Threads by the Assignment 2 deadline to receive 10 bonus marks. You can also receive 2 bonus marks for upvoting at least one question before 8am on the day of each lecture, for a maximum of 2 marks x 5 lectures = 10 marks for upvoting. Your assignment will be marked out of 120 minus the bonus marks you received.