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**

Solution.

1. We claim that

$$A_I = \cup_{i\in[n]}(\{I = i\} \cap A_i)\,. \tag{1}$$

Call the set on the LHS B. Now, if $\omega \in A_I$, then $\omega \in A_{I(\omega)}$. Let $i = I(\omega)$. Then $\omega \in \{I = i\}$ and $\omega \in A_i$. Hence, $\omega \in \{I = i\} \cap A_i$, and also $\omega \in B$. Now, for every $i \in [n]$, $\{I = i\} \cap A_i$, the intersection of two events is an event (here, $\{I = i\}$ is an event, because I is a random variable). The countable union of events is also an event, hence B is an event.

2. Let $I(\omega) = n$ if $\omega \notin \cup_i A_i$. Otherwise, let $I(\omega) = \min\{i \in [n] : \omega \in A_i\}$. We claim that I is a random variable. For $A \subset \Omega$, let χ_A be the characteristic (indicator) function of A: $\chi_A(\omega) = 1$ if $\omega \in A$ and $\chi_A(\omega) = 0$ otherwise. As it is immediate from the definition of random variables, χ_A is a random variable. Now note that

$$I = \min(J_1, \ldots, J_n, n)$$

where for $i \in [n]$, $J_i = i\chi_{A_i} + (n + 1)\chi_{A_i^c}$. Since the complement of an event is an event and linear combination of random variables gives random variables, J_i for $i \in [n]$ are random variables. Constant functions are also random variables. The minimum of random variables is a random variable, too. Hence, I is a random variable.

Now, clearly, $A_I \subset \cup_{i\in[n]} A_i$ (this follows directly (1)). To show the reverse, assume that $\omega \in \cup_{i\in[n]} A_i$. Then, $\omega \in A_i$ for some $i \in [n]$. Let i be the smallest index for which $\omega \in A_i$. Then, by its definition, $I(\omega) = i$ and thus $\omega \in \{I = i\} \cap A_i \subset A_I$ by (1).

Since $A_I = \cup_{i\in[n]} A_i$, they have the same probability.

3. The first solution does not need to be changed, except for the obvious replacement of $[n]$ with $\mathbb{N}$.

The second solution will need to be updated. Specifically, I can no longer be shown to be a random variable by using the alternate definition of

$$I = \min(J_1, \ldots, J_n, n),$$

since n can be arbitrarily large now.

To fix this problem we first let $I(\omega) = 1$ if $\omega \notin \cup_i A_i$ and now we will show that I must be a random variable by showing that it satisfies the definiton of a measurable map. We have that $I : \Omega \to \mathbb{N}$ and we claim that for any $b \in P(\mathbb{N})$ that $I^{-1}(b) \in \mathcal{F}$. Note that $P(\mathbb{N})$ is a valid σ-algebra on the set $\mathbb{N}$ and is the largest possible σ-algebra on the set $\mathbb{N}$. Thus, if we can show that I is a $\mathcal{F}/P(\mathbb{N})$ measurable map, then we have shown that I is also a $\mathcal{F}/\mathcal{G}$ measurable map for any $\mathcal{G}$ that is also a σ-algebra on $\mathbb{N}$. Since, for any $c \in \mathcal{G}$ we know that $c \in P(\mathbb{N})$. To show that $I^{-1}(b) \in \mathcal{F}$ for any $b \in P(\mathbb{N})$ it will be sufficient to show that $I^{-1}(d) \in \mathcal{F}$ for any $d \in \mathcal{D}$, where $\mathcal{D}$ is a generator of $P(\mathbb{N})$. This is true due to the well known fact that the pre-image of set functions preserves set operations. More precisely: for a set function $F : \mathbb{X} \to \mathbb{Y}$ and $B, B_i \subset \mathbb{Y},\ i \in \mathbb{N}$ we have that

$$F^{-1}(\cup_i B_i) = \cup_i F^{-1}(B_i),$$

$$F^{-1}(\mathbb{Y}\backslash B) = \mathbb{X}\backslash F^{-1}(B).$$

Thus, the pre-images of elements of a generator can be used to construct the pre-image of any element in the σ-algebra generated by the generator. The generator of $P(\mathbb{N})$ we choose is $\mathbb{N}$. It is clear that $I^{-1}(i) = A_i \backslash \cup_{k=1}^{i-1} A_i$ for any $i \in \mathbb{N}\backslash 1$, and $I^{-1}(1) = A_1 \cup (\cup_i A_i)^c$. Thus, since a σ-algebra is closed under countable unions, intersections and compliments (note that set subtraction can be written using intersection and complimentation), and $A_i \in \mathcal{F},\ \forall i \in \mathbb{N}$, we have that $I^{-1}(i) \in \mathcal{F}$. Showing that I is a random variable (measurable map). No further changes need to be made.

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, $C_k(s,a) = [S'_1(k,s,a), \ldots, S'_m(k,s,a)]$ 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 $Q_k(s,\cdot) = \hat{T}_k \ldots \hat{T}_1 \mathbf{0}(s,\cdot)$.

   5 points

2. Fix $H > 0$. Define a sequence of sets $\mathcal{S}_0, \ldots, \mathcal{S}_H$ 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, $\ldots$, C_{h+1} only (and is not a function of $C_h, \ldots, C_1$).

   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**

Solution.

1. We will show the result $Q_k(s,\cdot) = \hat{T}_k \dots \hat{T}_1 \mathbf{0}(s,\cdot)$ by induction on k. First, the base case with $k = 0$

$$Q_0(s,\cdot) = \texttt{q(0, s)} = \mathbf{0}. \qquad \text{Where } \mathbf{0} \text{ is a } |\mathcal{A}| \text{ length vector of zeros}$$

   Now, assume that $Q_{k-1}(s,\cdot) = \hat{T}_{k-1} \dots \hat{T}_1 \mathbf{0}(s,\cdot)$ holds. We show that $Q_k(s,\cdot) = \hat{T}_k \dots \hat{T}_1 \mathbf{0}(s,\cdot)$ holds

$$\begin{aligned}
&Q_k(s,\cdot) = q(k,s) \\
&= \texttt{[r(s,a) + gamma/m * sum([max(q(k-1,s')) for s' in C(k,s,a)]) for a in A]} \\
&= r_a(s) + \frac{\gamma}{m} \sum_{s' \in C_k(s,a)} \max_{a'} Q_{k-1}(s',a'),\ \forall a \in \mathcal{A} \\
&= r_a(s) + \frac{\gamma}{m} \sum_{s' \in C_k(s,a)} \max_{a'} \left(\hat{T}_{k-1} \dots \hat{T}_1 \mathbf{0}(s',a') \right),\ \forall a \in \mathcal{A} \quad \text{By inductive assumption} \\
&= \hat{T}_k \hat{T}_{k-1} \dots \hat{T}_1 \mathbf{0}(s,a),\ \forall a \in \mathcal{A} \\
&= \hat{T}_k \dots \hat{T}_1 \mathbf{0}(s,\cdot)
\end{aligned}$$

2. As usual, let $C_k(s) = \cup_a C_k(s,a)$. We let $\mathcal{S}_1 = \cup_{s \in \mathcal{S}_0} C_H(s) (= C_H(s_0))$ and, more generally, for $i > 0$, $\mathcal{S}_i = \cup_{s \in \mathcal{S}_{i-1}} C_{H-i+1}(s)$.

   (a) Let $h > 0$. To show $\delta_h \le \gamma \delta_{h-1} + \|\hat{T}_h q^* - q^*\|_{\mathcal{S}_{H-h}}$, note that, by the triangle inequality and the definition of $\hat{T}_h$,

$$\begin{aligned}
\delta_h &= \|Q_h - q^*\|_{\mathcal{S}_{H-h}} \\
&= \|\hat{T}_h Q_{h-1} - \hat{T}_h q^*\|_{\mathcal{S}_{H-h}} + \|\hat{T}_h q^* - q^*\|_{\mathcal{S}_{H-h}} .
\end{aligned}$$

   For the first term,

$$\begin{aligned}
\|\hat{T}_h Q_{h-1} - \hat{T}_h q^*\|_{\mathcal{S}_{H-h}} &\le \frac{\gamma}{m} \max_{s \in \mathcal{S}_{H-h}, a \in \mathcal{A}} \sum_{s' \in C_h(s,a)} |M Q_{h-1}(s') - v^*(s')| \\
&\le \gamma \max_{s \in \mathcal{S}_{H-h}, a \in \mathcal{A}} \max_{s' \in C_h(s,a)} |M Q_{h-1}(s') - v^*(s')| \\
&= \gamma \max_{s \in \mathcal{S}_{H-h+1}} |M Q_{h-1}(s) - v^*(s)| \\
&\le \gamma \|Q_{h-1}(s) - q^*(s)\|_{\mathcal{S}_{H-h+1}} .
\end{aligned}$$

   (b) This follows by induction starting with $h = H - 1$. Clearly, $\mathcal{S}_1$ is a function of C_H only. Assume that we already established that for $0 < h < H$, $\mathcal{S}_{H-h}$ is a function of $C_H, \dots, C_{h+1}$ only. Then, by its definition, $\mathcal{S}_{H-(h-1)} = \mathcal{S}_{H-h+1} = \cup_{s \in \mathcal{S}_{H-h}} C_{H-(H-h)}(s) = \cup_{s \in \mathcal{S}_{H-h}} C_h(s)$. And now, by the induction hypothesis, the claim follows: $\mathcal{S}_{H-(h-1)}$ is a function of $C_H, \dots, C_{h+1}$ and C_h only.

3. Fix h. For a fixed state $s \in \mathcal{S}$ and a fixed action $a \in \mathcal{A}$, w.p. at least $1-\zeta$,

$$\begin{aligned}
\overbrace{\left|\hat{T}_h q^*(s,a) - q^*(s,a)\right|}^{\Delta_h(s,a):=} &= \gamma \left| \frac{1}{m} \sum_{s' \in C_h(s,a)} v^*(s') - \gamma \langle P_a(s), v^* \rangle \right| \\
&< \gamma \|v^*\|_\infty \sqrt{\frac{\log(2/\zeta)}{2m}} && \text{(using Hoeffding's inequality)} \\
&\le \underbrace{\frac{\gamma}{1-\gamma} \sqrt{\frac{\log(2/\zeta)}{2m}}}_{=:f(\zeta)}. && \text{(since rewards lie in } [0,1])
\end{aligned}$$

We claim that w.p. at least $1-\zeta$,

$$\left|\hat{T}_h q^*(s,a) - q^*(s,a)\right| < \frac{\gamma}{1-\gamma} \sqrt{\frac{\log(2\mathrm{A}|\mathcal{S}_{H-h}|/\zeta)}{2m}}, \quad \text{for all } s \in \mathcal{S}_{H-h}, a \in \mathcal{A}. \tag{2}$$

To prove this, for $\mathcal{S}' \subset \mathcal{S}$ nonempty and $s \in \mathcal{S}, a \in \mathcal{A}$ we let

$$F_{s,a}(\mathcal{S}') = \{\Delta_h(s,a) \ge f(\zeta/(\mathrm{A}|\mathcal{S}'|))\}.$$

Display (2) is equivalent to

$$\mathbb{P}(\cup_{s \in \mathcal{S}_{H-h}, a \in \mathcal{A}} F_{s,a}(\mathcal{S}_{H-h})) \le \zeta.$$

To verify this inequality we use the law of total probability:

$$\begin{aligned}
\mathbb{P}(\cup_{s \in \mathcal{S}_{H-h}, a \in \mathcal{A}} F_{s,a}(\mathcal{S}_{H-h})) &= \sum_{\mathcal{S}" \subset \mathcal{S}, \mathcal{S}' \neq \emptyset} \mathbb{P}(\cup_{s \in \mathcal{S}_{H-h}, a \in \mathcal{A}} F_{s,a}(\mathcal{S}_{H-h}), \mathcal{S}_{H-h} = \mathcal{S}') \\
&= \sum_{\mathcal{S}" \subset \mathcal{S}, \mathcal{S}' \neq \emptyset} \mathbb{P}(\cup_{s \in \mathcal{S}', a \in \mathcal{A}} F_{s,a}(\mathcal{S}'), \mathcal{S}_{H-h} = \mathcal{S}') \\
&\overset{(*)}{=} \sum_{\mathcal{S}" \subset \mathcal{S}, \mathcal{S}' \neq \emptyset} \mathbb{P}(\cup_{s \in \mathcal{S}', a \in \mathcal{A}} F_{s,a}(\mathcal{S}')) \mathbb{P}(\mathcal{S}_{H-h} = \mathcal{S}') \\
&\le \zeta \sum_{\mathcal{S}" \subset \mathcal{S}, \mathcal{S}' \neq \emptyset} \mathbb{P}(\mathcal{S}_{H-h} = \mathcal{S}') = \zeta,
\end{aligned}$$

where the equality marked with $(*)$ used that by part (b) of Q2 and the definitions, $\mathcal{S}_{H-h} = \mathcal{S}'$ and $\cup_{s \in \mathcal{S}', a \in \mathcal{A}} F_{s,a}(\mathcal{S}')$ are independent (the latter only depends on C_h, the former only depends on $C_H, \cdots, C_{h+1}$, which are independent), and the last inequality follows from a union bound. Indeed, for $\mathcal{S}' \subset \mathcal{S}$ nonemtpy,

$$\mathbb{P}(\cup_{s \in \mathcal{S}', a \in \mathcal{A}} F_{s,a}(\mathcal{S}')) \le \sum_{s \in \mathcal{S}', a \in \mathcal{A}} \mathbb{P}(F_{s,a}(\mathcal{S}')) \le |\mathcal{S}'| \mathrm{A} \frac{\zeta}{\mathrm{A}|\mathcal{S}'|} = \zeta.$$

Now, from (2) we get that w.p. at least $1-\zeta$,

$$\begin{aligned}
\|\hat{T}_h q^* - q^*\|_{\mathcal{S}_{H-h}} &= \max_{s \in \mathcal{S}_{H-h}} \max_{a \in \mathcal{A}} \left|\hat{T}_h q^*(s,a) - q^*(s,a)\right| \\
&\le \frac{\gamma}{1-\gamma} \sqrt{\frac{\log(2\mathrm{A}|\mathcal{S}_{H-h}|/\zeta)}{2m}} \\
&\le \frac{1}{1-\gamma} \sqrt{\frac{\log(2\mathrm{A}|\mathcal{S}_{H-h}|/\zeta)}{2m}}.
\end{aligned}$$

4. We first use the recurrence relation $\delta_h \le \gamma\delta_{h-1} + \|\hat{T}_h q^* - q^*\|_{\mathcal{S}_{H-h}}$ from Q2, to obtain an expression for $\delta_H := \|Q_H - q^*\|_{\mathcal{S}_0} := \max_a |Q_H(s_0,a) - q^*(s_0,a)|$ as follows

$$\begin{aligned}
\delta_H &\le \gamma\delta_{H-1} + \|\hat{T}_H q^* - q^*\|_{\mathcal{S}_0} \\
&\le \gamma^2\delta_{H-2} + \gamma\|\hat{T}_{H-1}q^* - q^*\|_{\mathcal{S}_1} + \|\hat{T}_H q^* - q^*\|_{\mathcal{S}_0} \\
&\vdots \\
&\le \gamma^H\delta_0 + \sum_{k=0}^{H-1}\gamma^k\|\hat{T}_{H-k}q^* - q^*\|_{\mathcal{S}_k} \\
&\le \frac{\gamma^H}{1-\gamma} + \sum_{k=0}^{H-1}\gamma^k\|\hat{T}_{H-k}q^* - q^*\|_{\mathcal{S}_k}. \qquad \text{(since } \delta_0 := \|Q_0 - q^*\|_{\mathcal{S}_H} = \|q^*\|_{\mathcal{S}_H} \le \tfrac{1}{1-\gamma}\text{)}
\end{aligned}$$

Using the result from Q3 with the fact that $|\mathcal{S}_k| = (m\mathrm{A})^k$, we get that for a fixed k, w.p. $1-\zeta$, $\|\hat{T}_{H-k}q^* - q^*\|_{\mathcal{S}_k} \le \frac{1}{1-\gamma}\sqrt{\frac{\log(2\mathrm{A}(m\mathrm{A})^k/\zeta)}{2m}}$. Now we can employ union bound over the index $0 \le k \le H-1$, in the equation for δ_H given above, to get w.p. $1-\zeta$

$$\delta_H = \max_a |Q_H(s_0,a) - q^*(s_0,a)| \le \frac{\gamma^H}{1-\gamma} + \frac{1}{1-\gamma}\sum_{k=0}^{H-1}\gamma^k\sqrt{\frac{\log(2H\mathrm{A}(m\mathrm{A})^k/\zeta)}{2m}} =: \Delta(m,H,\zeta),$$

where the extra H comes in the numerator comes from union bound over the index $0 \le k \le H-1$.[1]

From above equation we get that the policy induced by the local planner $\pi(s_0) = \arg\max_a Q_H(s_0,a)$ is $2\Delta(m,H,\zeta)$-optimizing. Then using policy error bound II from Lecture 6, we get that π is $\varepsilon(m,H,\zeta)$-optimal with $\varepsilon(m,H,\zeta)$ defined as

$$\frac{2\Delta(m,H,\zeta) + 2\zeta\|q^*\|_\infty}{1-\gamma} \le \frac{2}{(1-\gamma)^2}\left[\gamma^H + \sum_{k=0}^{H-1}\gamma^k\sqrt{\frac{\log(2H\mathrm{A}(m\mathrm{A})^k/\zeta)}{2m}} + \zeta\right] =: \varepsilon(m,H,\zeta).$$

5. The bound given in the lecture 6 was $\varepsilon_{\text{lec}}(m,H,\zeta) := \frac{2}{(1-\gamma)^2}\left[\gamma^H + \frac{1}{1-\gamma}\sqrt{\frac{\log(2n\mathrm{A}/\zeta)}{2m}} + \zeta\right]$ with $n = (m\mathrm{A})^H$. Therefore, we need to compare the two terms $T_1 = \sum_{k=0}^{H-1}\gamma^k\sqrt{\frac{\log(2H\mathrm{A}(m\mathrm{A})^k/\zeta)}{2m}}$ and $T_2 = \frac{1}{1-\gamma}\sqrt{\frac{\log(2\mathrm{A}(m\mathrm{A})^H/\zeta)}{2m}}$. It is likely that $T_1 \le T_2$ in general, since for all but large k, $H(m\mathrm{A})^k \ll (m\mathrm{A})^H$. Thus, this bound is tighter than that given in the lecture. In fact, a quick calculation gives that one saves a factor of H on setting m this way.

We could obtain a bound, similar to the one obtained here, in the lecture if we don't use the relaxation $\|\hat{T}q^* - q^*\|_{\mathcal{S}_{H-h}} \le \|\hat{T}q^* - q^*\|_{\mathcal{S}_{H-1}}$ while deriving the recurrence for δ_h (see lecture 6 notes) and in fact the bound then would save (an insignificant) H in the logarithm.

6. It is straightforward to see that the computational complexity of the algorithm is $O((m\mathrm{A})^H)$. If we use the result $\varepsilon(m,H,\zeta) \le \varepsilon_{\text{lec}}(m,H,\zeta)$, hypothesized in Q5, then the computation complexity is the same as given in lecture 6 notes with $m \ge m^*$ (Eq. 8 in the lecture 6 notes). A tighter analysis should also be possible.

Paraphrasing from Kearns, Mansour, & Ng (2002), this algorithm is based on the idea of sparse sampling. As we showed above, a randomly sampled look-ahead tree that covers only a fraction (given

[1] It might be possible to obtain a tighter bound by using separate $\zeta_k = x_k\zeta$, with x_ks found by solving the optimization problem $\min \sum_{k=0}^{H-1}\sqrt{\log(e_k/x_k)}$ subject to $\sum_{k=0}^{H-1} x_k = 1$. But maybe $x_k = 1/H$ is a good enough solution; at least it's super simple! Another choice could be $x_k = e_k/\sum_{i=1}^{H} e_i$.

by the value m^*) of the full look-ahead tree suffices to compute near-optimal actions from any state s_0 of an MDP. Therefore, this approach is called a sparse-lookahead tree approach. Here, we can think of the computation as building out a lookahead tree of depth H from s_0 and then using this tree to back-propagate action-values using value iteration.

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).[2]

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

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

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

10 points

[2] A terrible name.

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

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

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 *e*xpected 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**

Solution.

1. Note that

$$\theta_k = G_\rho^{-1} \underbrace{\sum_{z \in C} \rho(z)\varphi(z) Y_k(z)}_{=:B_k} , \tag{3}$$

where $Y_k(z) = \hat{T}_k \Pi f_{\theta_{k-1}}(z)$. For z fixed, $Y_k(z)$ can be computed in $O(mAd)$ steps. All $Y_k(\cdot)$ is computed in $O(|C|mAd) = O(d^3 mA)$ steps. Given these $O(d^2)$ values, $B_k \in \mathbb{R}^d$ can be calculated in time $O(|C|d) = O(d^3)$ and thus the total cost of calculating B_k is $O(d^3 mA)$. The matrix inverse G_ρ^{-1} needs only to be computed once, at the cost of, say $O(d^3)$. The cost of matrix vector multiplication is $O(d^2)$. The total cost of calculating θ_k is dominated by $O(d^3 mA)$. Multiply this by K to get the total cost of the procedure.

For storage, one can invert a matrix in place. Besides the matrix G_ρ^{-1}, one needs to store only d-dimensional vectors. Hence, the storage cost is $O(d^2)$.

The query complexity of calculating comes from the need to access $C_k(z)$ for $z \in C$. Hence, the query cost is $O(d^2 m)$. Multiply this by K to get the total number of queries.

2. Choose θ_k^* as the minimizer of $\theta \mapsto g(\theta) := \|Tq_{k-1} - f_\theta\|_\infty$. We argue that this exists. Indeed, g is continuous and nonnegative. Hence, there exists a sequence $(\theta_i)_i$ such that $g(\theta_i) \to \inf_\theta g(\theta)$. Note that G_ρ is full rank because φ spans $\mathbb{R}^d$. There are two cases: Either $\sup_i \|\theta_i\|_{G_\rho}$ is finite, or it is infinite. If it is finite, by the completeness of $\mathbb{R}^d$, a subsequence of θ_i converges to a minimizer of g by the continuity of g. In the opposite case, from $\|\theta_i\|_{G_\rho}^2 = \sum_{z\in C} \rho(z) f_{\theta_i}^2(z)$ we see that, $(f_{\theta_i}^2(z))_i$ must be unbounded for at least one $z \in C$. Hence, for this z,

$$g(\theta_i) = \|f_{\theta_i} - Tq_{k-1}\|_\infty \geq |f_{\theta_i}(z)| - |Tq_{k-1}(z)| \,.$$

Hence,

$$\limsup_{i\to\infty} g(\theta_i) \geq \limsup_{i\to\infty} |f_{\theta_i}(z)| - |Tq_{k-1}(z)| = \infty \,,$$

which contradict to that $\limsup_{i\to\infty} g(\theta_i) = \inf_\theta g(\theta) \leq g(0) < \infty$.

3. We have

$$\begin{aligned}
\|q_k - Tq_{k-1}\|_\infty &= \|\Pi f_{\theta_k} - Tq_{k-1}\|_\infty \\
&\leq \|f_{\theta_k} - (f_{\theta_k^*} + \varepsilon_k)\|_\infty \\
&\leq \|f_{\theta_k} - f_{\theta_k^*}\|_\infty + \varepsilon_{\mathrm{apx}} \\
&\leq \|f_{\theta_k} - f_{\theta_k^*}\|_\infty + \varepsilon_{\mathrm{apx}} \\
&\leq \sqrt{d} \max_{z\in C} |\varepsilon(z)| + \varepsilon_{\mathrm{apx}} \,,
\end{aligned}$$

where the first equality uses the definition of q_k, the next inequality uses that $Tq_{k-1} \in [0, 1/(1-\gamma)]$, hence dropping the truncation can only increase the values (at the same place we also used the definition of θ_k^* and ε_k). The next inequality uses the triangle inequality and that by definition $\|\varepsilon_k\|_\infty \leq \varepsilon_{\mathrm{apx}}$, and for the last inequality we use an appropriately defined function $\varepsilon : C \to \mathbb{R}$. For the definition recall the corollary of Lecture 8 that states that $\|f_{\hat\theta} - f_\theta\|_\infty \leq \sqrt{d} \max_{z\in C} |\varepsilon(z)|$ holds for $\hat\theta = G_\rho^{-1} \sum_{z\in Z} \rho(z)\varphi(z)(f_\theta(z) + \varepsilon(z))$. Now, recall the definition of θ_k from (3). Writing

$$Y_k(z) = (Tq_{k-1})(z) + \hat\varepsilon(z) = f_{\theta_k^*}(z) + \hat\varepsilon(z) + \varepsilon_k(z) \,,$$

where the first equality defines $\hat\varepsilon(z)$, we see that above we can use $\varepsilon(z) = \hat\varepsilon(z) + \varepsilon_k(z)$. Now, note from Hoeffding's inequality that with probability $1-\zeta$,

$$\max_{z\in C} |\hat\varepsilon(z)| \leq H_\gamma \sqrt{\frac{\log\left(\frac{2|C|}{\zeta}\right)}{2m}} \,,$$

where we use that $\mathbb{E}[Y_k(z)|q_{k-1}] = (Tq_{k-1})(z)$ and that $S_1'(k,z), \ldots, S_m'(k,z)$ are independent given q_{k-1}, hence, $(Mq_{k-1}(S_j'(k,z))_j$ is an i.i.d. sequence, and it takes values in the interval $[0, H_\gamma]$. We also have

Cs: this independence is not quite well explained.

$$|\varepsilon(z)| \leq \varepsilon_{\mathrm{apx}} + H_\gamma \sqrt{\frac{\log\left(\frac{2|C|}{\zeta}\right)}{2m}} \,,$$

Putting everything together gives the desired claim.

4. Let $\delta_k = \|q_k - q^*\|_\infty$. For $k > 0$ we have

$$\delta_k \leq \|q_k - Tq_{k-1}\|_\infty + \|Tq_{k-1} - Tq^*\|_\infty \leq \|q_k - Tq_{k-1}\|_\infty + \gamma\|q_{k-1} - q^*\|_\infty \leq \|q_k - Tq_{k-1}\|_\infty + \gamma\delta_{k-1} \,.$$

Unfolding this and using $\delta_0 \le H_\gamma$,

$$\delta_K \le \gamma^K H_\gamma + H_\gamma \max_{1 \le k \le K} \|q_k - T q_{k-1}\|_\infty .$$

Taking a union bound over $k \in [K]$ and plugging in the bound from the previous item, we get

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

Finally, by our policy error bound,

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

5. Let $\hat{\pi}$ be the random policy computed by the algorithm. Let $\mathcal{E}$ be the event of the previous part. By the previous part, on $\mathcal{E}$,

$$v^* - v^{\hat{\pi}} \le \delta \mathbf{1} .$$

Now, for any fixed $s \in \mathcal{S}$,

$$\begin{aligned}
\mathbb{E}[v^*(s) - v^{\hat{\pi}}(s)] &= \mathbb{E}[(v^*(s) - v^{\hat{\pi}}(s))\mathbb{I}_{\mathcal{E}}] + \mathbb{E}[(v^*(s) - v^{\hat{\pi}}(s))\mathbb{I}_{\mathcal{E}^c}] \\
&\le \mathbb{E}[\delta \mathbb{I}_{\mathcal{E}}] + \mathbb{E}[H_\gamma \mathbb{I}_{\mathcal{E}^c}] \\
&= \delta \mathbb{P}(\mathcal{E}) + H_\gamma \mathbb{P}(\mathcal{E}^c) \\
&\le \delta + H_\gamma \zeta ,
\end{aligned}$$

where the last inequality used Q3.

Now, from the result of Q4,

$$\delta + H_\gamma \zeta \le 2H_\gamma^2 (1+\sqrt{d})\varepsilon_{\text{apx}} + \underbrace{2\, H_\gamma^2 \left[\gamma^K + \sqrt{d} H_\gamma \sqrt{\frac{\log\left(\frac{2|C|K}{\zeta}\right)}{2m}} + \frac{\zeta}{2H_\gamma} \right]}_{\le \varepsilon} \le 2H_\gamma^2 (1+\sqrt{d})\varepsilon_{\text{apx}} + 2\varepsilon.$$

Letting each of the last three terms above to be less than $\varepsilon/3$, we get the following conditions:

$$\begin{aligned}
K &\ge \frac{\log\left(3H_\gamma^2/\varepsilon\right)}{\log(1/\gamma)}, \\
\zeta &\le 2\varepsilon/(3H_\gamma), \quad \text{and} \\
m &\ge \frac{9H_\gamma^6 d}{2\varepsilon^2} \left[\log(2|C|) + \log K + \log\left(3H_\gamma/(2\varepsilon)\right) \right].
\end{aligned}$$

Recalling that $|C| \le d(d+1)/2$ gives us the desired result.

6. From Q1, we know that the query cost $O(Kd^2m)$, the runtime complexity $O(Kd^3mA)$, and the space complexity $O(d^2)$ are all polynomial in A, d, K, and m. Therefore, the result follows from Q5, which shoes that both K and m themselves have polynomial dependence on H_γ, d, and $1/\varepsilon$, and that policy is δ-suboptimal with $\delta = 2H_\gamma^2(1+\sqrt{d})_{\varepsilon_{\text{apx}}} + 2\varepsilon$.

7. It suffices to see that for any $q \in \mathbb{R}^{\mathcal{S}\times\mathcal{A}}$, $Tq \in \mathcal{F}_\varphi$. Indeed, then for any θ, $T\Pi f_\theta \in \mathcal{F}_\varphi$, which means that $\inf_{\theta'} \|f_{\theta'} - T\Pi f_\theta\|_\infty = 0$.

   Fix now $q \in \mathbb{R}^{\mathcal{S}\times\mathcal{A}}$. Let $v = Mq$. Letting $Z \in \mathbb{R}^{d\times \mathrm{S}}$ be defined using $Z(i, s') = \mu_i(s')$, notice that for $P \in \mathbb{R}^{\mathrm{SA}\times\mathrm{S}}$ it holds that $P = \Phi Z$ while $r = \Phi\theta_r$. Hence,

$$Tq = r + \gamma Pv = \Phi\theta_r + \gamma\Phi Zv = \Phi(\theta_r + \gamma Zv)\,,$$

   which shows that $Tq \in \mathcal{F}_\varphi$, finishing the proof.

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.