CMPUT 605: Theoretical Foundations of Reinforcement Learning, Winter 2023
Homework #3

Instructions

Submissions You need to submit a single PDF file, named `p03_<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: March 12 at 11:55 pm

Average vs. mixed policies

Fix policies $\pi^{(1)},\dots,\pi^{(k)}$ of some finite discounted MDP $M=(\mathcal{S},\mathcal{A},P,r,\gamma)$. There are two ways of combining these policies with some weights $\alpha\in\mathcal{M}_1([k])$. The first way is to choose one of the policies at random from the multinomial parameterized by α and then follow the resulting policy for all the time steps. Formally, one would choose an index $I\in[k]$ at random such that $\mathbb{P}(I=i)=\alpha_i$ and then follow the policy $\pi^{(I)}$ for whichever state one encounters. The second way is to choose the policy to follow at random in each time step. Call the policy that is obtained following the first method the (α-weighted) **mixture of** $\pi^{(1)},\dots,\pi^{(k)}$. Call the policy that is obtained following the second method the (α-weighted) **average of** $\pi^{(1)},\dots,\pi^{(k)}$.

Intuitively, a distribution $\mu\in\mathcal{M}_1(\mathcal{S})$ over the states and the interconnection of a mixture policy and M gives rise to a probability space $(\Omega,\mathcal{F},\mathbb{P})$ that carries the random elements $I,S_0,A_0,S_1,A_1,\dots$ with $I\in[k]$, $S_t\in\mathcal{S}$ and $A_t\in\mathcal{A}$ for $t\ge 0$ and such that for $H_t=(S_0,A_0,S_1,\dots,A_{t-1},S_t)$,

1. $\mathbb{P}(S_0=s|I)=\mu(s)$ for all $s\in\mathcal{S}$,
2. $\mathbb{P}(A_t=a|I,H_t)=\pi_t^{(I)}(a|H_t)$ for all $a\in\mathcal{A},t\ge 0$,
3. $\mathbb{P}(S_{t+1}=s'|I,H_t,A_t)=P_{A_t}(S_t,s')$ for all $s'\in\mathcal{S}$, and
4. $\mathbb{P}(I=i)=\alpha_i$ for all $i\in[k]$.

Note that all first three criteria are modified to express that the laws that govern S_0, the action distribution and the next state distribution are as before even when conditioning on I. A new, fourth criterion is added that expresses that the distribution of I follows the multinomial distribution with parameter α. That the probability distribution $\mathbb{P}$ with the above properties exists is guaranteed again by the Ianescu-Tulcea theorem. As usual, when needed, we use $\mathbb{P}_\mu$ to indicate the dependence of $\mathbb{P}$ on μ.

Finally some notation: For a probability measure $\mathbb{P}$ on a measurable space $(\Omega,\mathcal{F})$ and a sub-sigma algebra $\mathcal{G}$ of $\mathcal{F}$, let $\mathbb{P}|_{\mathcal{G}}$ be the probability measure on $(\Omega,\mathcal{G})$ obtained from $\mathbb{P}$ by restricting it to $\mathcal{G}$: $\mathbb{P}|_{\mathcal{G}}(U)=\mathbb{P}(U)$ for any $U\in\mathcal{G}$.

Question 1. Unless otherwise specified let $\pi^{(1)}, \dots, \pi^{(k)}$ be arbitrary policies of M and let $\alpha \in \mathcal{M}_1([k])$, $\mu \in \mathcal{M}_1(\mathcal{S})$ be also arbitrary. Also, let $(\Omega, \mathcal{F}, \mathbb{P})$ as above (we shall also use $\mathbb{P}_\mu$ when the dependence on μ is important). Let $Z = (S_0, A_0, S_1, A_1, \dots)$. Show that the following hold:

1. Z is random element between $(\Omega, \mathcal{F})$ and $((\mathcal{S} \times \mathcal{A})^{\mathbb{N}}, \mathcal{G}')$ where $\mathcal{G}'$ is the product σ-algebra on $(\mathcal{S} \times \mathcal{A})^{\mathbb{N}}$ induced by the discrete topology on $\mathcal{S} \times \mathcal{A}$.

   5 points

2. Show that there is a policy $\bar{\pi}$ of the MDP M such that for any $\mu \in \mathcal{M}_1(\mathcal{S})$, the pushforward of $\mathbb{P}_\mu$ under Z, $(\mathbb{P}_\mu)_Z$ satisfies
$$(\mathbb{P}_\mu)_Z = \mathbb{P}_\mu^{\bar{\pi}}$$
where $\mathbb{P}_\mu^{\bar{\pi}}$ is the unique probability measure on the canonical space $((\mathcal{S} \times \mathcal{A})^{\mathbb{N}}, \mathcal{G}')$ induced by the interconnection of $\bar{\pi}$ and the MDP, given the initial state distribution μ. That is, a mixture policy induces a policy $\bar{\pi}$ of the MDP M.

   20 points

3. Let $R = \sum_{t=0}^{\infty} \gamma^t r_{A_t}(S_t)$ and let $\mathbb{P}$ be as above with the choice $\mu = \delta_s$. Let $\mathbb{E}$ be the expectation operator corresponding to $\mathbb{P}$. Show that $v(s) = \mathbb{E}[R]$ is well-defined: That is, for any $(\Omega, \mathcal{F}, \mathbb{P})$ and $(\Omega, \mathcal{F}, \mathbb{P}')$ as long as $\mathbb{P}$ and $\mathbb{P}'$ satisfy the above four properties, $\mathbb{E}[R] = \mathbb{E}'[R]$ where $\mathbb{E}'$ is the expectation operator underlying $\mathbb{P}'$.

   10 points

4. Show that $v(s) = v^{\bar{\pi}}(s)$.

   5 points

5. Let $\mathbb{P}_\mu^{\pi^{(i)}}$ ($\mathbb{P}_\mu^{\bar{\pi}}$) be the probability measures induced on the canonical space $((\mathcal{S} \times \mathcal{A})^{\mathbb{N}}, \mathcal{G}')$ by the initial state distribution μ and the interconnection of $\pi^{(i)}$ (respectively, $\bar{\pi}$) with the MDP M. Show that $\mathbb{P}_\mu^{\bar{\pi}} = \sum_{i=1}^{k} \alpha_i \mathbb{P}_\mu^{\pi^{(i)}}$.

   10 points

6. Mixing is guaranteed to keep performance bounds: if for some $v : \mathcal{S} \to \mathbb{R}$ and for all $i \in [k]$, $v^{\pi^{(i)}} \geq v$ then $v^{\bar{\pi}} \geq v$.

   5 points

7. Averaging is not guaranteed to keep performance bounds: For any $\gamma > 1/2$ there exists an MDP with state space $\mathcal{S}$, $k \geq 2$, policies $\pi_1, \dots, \pi_k$, a function $v : \mathcal{S} \to \mathbb{R}$ and $\alpha \in \mathcal{M}_1([k])$ such that $v^{\pi_i} \geq v$ holds for all $i \in [k]$, yet if π is the α-average of $\pi_1, \dots, \pi_k$ then $v^\pi < v$.

   10 points

Hint: Recall the change-of-variables formula: For a random element X taking values in some measurable set $\mathcal{X}$, the pushforward $\mathbb{P}_X$ of X satisfies

$$\mathbb{E}[f(X)] = \int f(x)\mathbb{P}_X(dx)\,.$$

Recall also that integration is linear in measures. In particular, for any measures $\mathbb{P}_i$ and nonnegative coefficients α_i, $i \in [k]$ and f which is $(\sum_{i=1}^k \alpha\mathbb{P}_i)$-integrable, $\int f d(\sum_{i=1}^k \alpha\mathbb{P}_i) = \sum_{i=1}^k \alpha_i \int f d\mathbb{P}_i$ (this also extends to signed measures, but we won't need this extension).

Total: **65 points**

Finding needles with high probability

The high-probability needle lemma is as follows:

Lemma 1 (High-probability needle lemma)**.** *Any algorithm that correctly identifies the single nonzero entry in any binary array of length k with probability at least* 0.91 *has the property that on some input the expected number of queries that the algorithm uses is at least* $\Omega(k)$*.*

Question 2. Prove Lemma 1. Note that the algorithms are allowed to randomize.

Total: **30 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 $\phi : \mathcal{S} \times \mathcal{A} \to \mathbb{R}^d$ which spans $\mathbb{R}^d$. Let $\mathcal{F} = \{f_\theta \,:\, f_\theta(s,a) = \phi(s,a)^\top\theta, \theta \in \mathbb{R}^d\}$ 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 $G_\rho = \sum_{z\in C} \rho(z)\phi(z)\phi(z)^\top$, $\max_{z\in\mathcal{Z}} \|\phi(z)\|_{G_\rho^{-1}} \le \sqrt{d}$.

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

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

Further, let $\Pi : \mathbb{R}^{\mathcal{Z}} \to \mathbb{R}^{\mathcal{Z}}$ 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]

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

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

[1] A terrible name.

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

   10 points

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

   5 points

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

   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), $P_a(s, s') = \langle \phi(s, a), \mu(s') \rangle$. Show that if $\mathcal{M}$ is linear in ϕ then $\varepsilon_{\text{apx}} = 0$.

   10 points

Total: **55 points**

Total for all questions: 150. Of this, 30 are bonus marks (i.e., 120 marks worth 100% on this problem set).