CMPUT 653: Theoretical Foundations of Reinforcement Learning, Winter 2022
Homework #4

Instructions

Submissions You need to submit a zip file, named `p4_<name>.zip` or `p4_<name>.pdf` 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: March 30 at 11:55 pm

Large action set query lower bound

We recall a few definitions and results from Lecture 9. For a featurized MDP (M, ϕ), let

$$\varepsilon^*(M, \Phi) := \sup_{\pi \text{ memoryless}} \inf_{\theta \in \mathbb{R}^d} \|\Phi\theta - q^\pi\|_\infty . \tag{1}$$

Definition 1. An online planner is (δ, ε)-sound if for any finite discounted MDP $M = (\mathcal{S}, \mathcal{A}, P, r, \gamma)$ and feature-map $\varphi : \mathcal{S} \times \mathcal{A} \to \mathbb{R}^d$ such that $\varepsilon^*(M, \Phi) \le \varepsilon$, when interacting with (M, φ), the planner induces a δ-suboptimal policy of M.

The following was proven in the said lecture:

Theorem 1 (Query lower bound: large action sets). *For any $\varepsilon > 0$, $0 < \delta \le 1/2$, positive integer d and for any (δ, ε)-sound online planner $\mathcal{P}$ there exists a featurized-MDP (M, φ) with rewards in $[0, 1]$ with $\varepsilon^*(M, \Phi) \le \varepsilon$ such that when interacting with a simulator of (M, φ), the expected number of queries used by $\mathcal{P}$ is at least $\Omega(f(d, \varepsilon, \delta))$ where*

$$f(d, \varepsilon, \delta) = \exp\left(\frac{1}{32}\left(\frac{\sqrt{d}\varepsilon}{\delta}\right)^2\right) .$$

Question 1. The lecture notes provide a proof sketch for this theorem. Formally prove this theorem, explicitly explain each step of your proof.

Total: **20 points**

Fixed-horizon fundamental theorem

The same lecture stated the fundamental theorem for fixed-horizon problems, which we copy here for convenience. For the definitions of the quantities used in the theorem, see the lecture notes.

Theorem 2 (Fixed-horizon fundamental theorem)**.** *We have $v_0^* \equiv \mathbf{0}$ and for any $h \ge 0$, $v_{h+1}^* = T v_h^*$. Furthermore, for any $\pi_0^*, \dots, \pi_h^*, \dots$ such that for $i \ge 0$, π_i^* is greedy with respect to v_i^*, for any $h > 0$ it holds that $\pi = (\pi_{h-1}^*, \dots, \pi_0^*, \dots)$ (i.e., the policy which in step 0 uses π_{h-1}^*, in step 1 uses π_{h-2}^*, $\dots$, in step $(h-1)$ uses π_0^*, after which it continues arbitrarily) is h-step optimal:*

$$v_h^\pi = v_h^* \,.$$

In the lecture notes we did not give a proof.

Question 2. Prove Theorem 2. **Hint**: Use induction and mimic the previous proofs.

Total: **50 points**

Statisticians also have limits

Let $\mathcal{X}$ be a subset of a Euclidean space equipped with the usual Borel σ-algebra, $\mathcal{P} \subset \mathcal{M}_1(\mathcal{X})$ a set of probability distributions over $\mathcal{X}$. Let $f : \mathcal{P} \to \mathbb{R}$ be a fixed function. We consider statistical estimation problems where a random "data" $X \in \mathcal{X}$ is observed from an unknown $P \in \mathcal{P}$ and the job of the statistician is to produce an estimate of $f(P)$.

That is, the statistician needs to design an estimator; for simplicity we assume that the estimators are not randomizing (an extension to randomizing estimators is trivial). A non-randomizing estimator maps the data to a real; thus, any such estimator is a map $g : \mathcal{X} \to \mathbb{R}$. We assume that g is measurable so that we can talk about the probability of errors.

In particular, for $\delta \in [0, 1]$ and $\varepsilon > 0$, we say that g is (δ, ε)**-sound** for the problem specified by $(\mathcal{P}, f)$ if for any $P \in \mathcal{P}$,

$$P(|g(X) - f(P)| > \varepsilon) \le \delta \,. \tag{2}$$

Here, $X : \mathcal{X} \to \mathcal{X}$ is treated as the identity map, as usual: $X(x) = x$, $x \in \mathcal{X}$. Thus, the above probability is the probability assigned by P to the set

$$\{x \in \mathcal{X} \,:\, |g(x) - f(P)| > \varepsilon\}$$

and condition (2) has the equivalent form that for any $P \in \mathcal{P}$,

$$P(\{x \in \mathcal{X} \,:\, |g(x) - f(P)| > \varepsilon\}) \le \delta \,.$$

It is just shorter and more elegant to write Eq. (2), hence, we will stick to this usual form.

For two probability measures, P, Q, over the same measurable space $(\Omega, \mathcal{F})$, we define their **relative entropy** by

$$D(P, Q) = \begin{cases} \int \log \frac{dP}{dQ}(\omega)\, dP(\omega)\,, & \text{if } P \ll Q \\ +\infty\,, & \text{otherwise}\,. \end{cases}$$

The relative entropy is also known as the Kullback-Leibler divergence between P and Q (see Chapter 14 in the bandit book for an explanation of its origin and some examples).

The following result, which is Theorem 14.12 in that book, will be useful:

Theorem 3 (Bretagnolle–Huber inequality)**.** *Let P and Q be probability measures on the same measurable space $(\Omega, \mathcal{F})$, and let $A \in \mathcal{F}$ be an arbitrary event. Then,*

$$P(A) + Q(A^c) \geq \frac{1}{2} \exp\left(-D(P,Q)\right) , \tag{3}$$

where $A^c = \Omega \setminus A$ is the complement of A.

Question 3. Show that if there is an (δ, ε)-sound estimator for $(\mathcal{P}, f)$ then

$$\log\left(\frac{1}{4\delta}\right) \leq \inf\{D(P_0, P_1) \,:\, P_0, P_1 \in \mathcal{P} \text{ s.t. } |f(P_0) - f(P_1)| > 2\varepsilon\} .$$

In words, distributions whose f-values are separated by 2ε cannot be too close to each other if a (δ, ε)-sound estimator exist. This should be quite intuitive.

Total: **20 points**

In what follows, we will deal with Bernoulli random variables. The relative entropy between Bernoulli distributions has special properties which we will find useful. The next problem asks you to prove some of these properties.

Let $\mathrm{Ber}(p)$ denote the Bernoulli distribution with parameter $p \in [0,1]$. As it is well known (and not hard to see from the definition),

$$D(\mathrm{Ber}(p), \mathrm{Ber}(q)) = d(p,q)$$

where $d(p,q)$ is the so-called **binary relative entropy function**, which is defined as

$$d(p,q) = p\log(p/q) + (1-p)\log((1-p)/(1-q)) .$$

Question 4. Show that for $p, q \in (0,1)$, defining p^* to be p or q depending on which is further away from $1/2$,

$$d(p,q) \leq \frac{(p-q)^2}{2p^*(1-p^*)} . \tag{4}$$

Hint: Notice that $d(p,q) = D_R((p, 1-p), (q, 1-q))$, where D_R is Bregman divergence with respect to our old friend, the unnormalized negentropy R over $[0,\infty)^2$. Then use Theorem 26.12 from the bandit book.

Total: **20 points**

Now, for $n > 0$ let $\mathrm{Ber}^{\otimes n}(p)$ denote the n-fold product of $\mathrm{Ber}(p)$ with itself, so that if $X \sim \mathrm{Ber}^{\otimes n}(p)$ then $X = (X_1, \ldots, X_n)$ where $X_i \sim \mathrm{Ber}(p)$ and $(X_1, \ldots, X_n)$ is an independent sequence.

Take $\mathcal{X} = \{0,1\}^n$ and $\mathcal{P}_n = \{\mathrm{Ber}^{\otimes n}(p) \,:\, p \in [0,1]\}$. Let $f : \mathcal{P}_n \to [0,1]$ be defined by $f(\mathrm{Ber}^{\otimes n}(p)) = p$. The problem specified by $(\mathcal{P}_n, f)$ is the problem of estimating the parameter of a Bernoulli distribution given n independent observations from the said, unknown distribution.

Question 5. Show that for the Bernoulli estimation problem described above, for $\delta \in [0,1]$ and $0 \leq \varepsilon^2 < 1/32$ fixed, there is no (δ, ε)-sound estimator of the common mean, unless $n \geq \frac{\log(1/(4\delta))}{16\varepsilon^2}$.
Hint: Use that $D(P^{\otimes n}, Q^{\otimes n}) = nD(P,Q)$ and the statements from the previous two problems.

Total: **20 points**

Now consider the problem when the definition of f is changed to

$$f_\gamma(\mathrm{Ber}^{\otimes n}(p)) = \frac{1}{1-\gamma p}\,, \tag{5}$$

where $0 < \gamma < 1$.

Question 6. Show that for the Bernoulli estimation problem described above with $f = f_\gamma$ as in Eq. (5), with some constants $\gamma_0 > 0$ and $c_0, c_1 > 0$, for $\delta \in [0,1]$, $\varepsilon \le c_0/(1-\gamma)$, $\gamma \ge \gamma_0$, the necessary condition for the existence of (δ, ε)-sound estimator for $(\mathcal{P}_n, f_\gamma)$ is that $n \ge c_1 \frac{\log(1/(4\delta))}{(1-\gamma)^3\varepsilon^2}$.
Hint: Use the same strategy as in the solution of the previous exercise.

Total: **40 points**

Total for all questions: 170. Of this, up to 70 can be bonus marks. You can receive bonus marks by asking/upvoting questions, for a total of 70 bonus marks! You must ask at least one question in one of the Lecture Discussion Threads by the Assignment 4 deadline to receive 50 bonus marks. You can also receive 5 bonus marks for upvoting at least one question before 8am on the day of each lecture, for a maximum of 5 marks x 4 lectures = 20 marks for upvoting. Your assignment will be marked out of 170 minus the bonus marks you received.