Homework 0: Review Problems
CS 6789: Foundations in Reinforcement Learning

Cornell University

Due Feb 1st 11:59pm ET

1 Policies *[0 points]*

Please read these policies. **Please answer the questions below and include your answers marked in a "problem 1" in your solution set.** Homeworks which do not include these answers will not be graded.

Homework Scoring: All bonus questions on homeworks are worth a maximum of 5 points. Note bonus points can only be used to get you a maximum of a 100% on the assignment.

Gradescope submission: When submitting your HW, please tag your pages correctly as is requested in gradescope. Untagged homeworks will not be graded, until the tagging is fixed.

Readings: Read the notes and required material.

Submission format: Submit your report as a *single* pdf file. Please typeset your writing using LaTeX.

Collaboration: It is acceptable for you to discuss problems with other students; it is not acceptable for students to look at another students written answers. Each student must understand, write, and hand in their own answers.

Acknowledgments: If students find out solutions in published material, on the web, or from other textbooks, or chatGPT / GPT-X (this unlikely gonna happen for this class, but who knows), this must be acknowledged. If students find proofs in existing papers, it is ok to use these for guidance; students must acknowledge this, and students should first make an attempt at the answer on their own. All students must understand all the written steps that they write.

1.1 List of Collaborators

List the names of all people you have collaborated with and for which question(s).

1.2 List of Acknowledgements

If you find an assignment's answer or use a another source for help, acknowledge for which question and provide an appropriate citation (there is no penalty, provided you include the acknowledgement). If not, then write "none".

1.3 Certify that you have read the instructions

Write "I have read these policies" to certify this.

2 Certify that you have read the website *[0 points]*

Please read the course policies on the website (up until the Lecture Notes section) and write"I have read the course policies on the website". It is your responsibility to understand and follow these policies. If further clarification is needed, please post to the discussion board.

3 Martingales *[20 points]*

An agent participates in a betting game, where in each round, it can pay any p_t such that $0 < p_t < 1$, and receives a reward of 1 with probability p_t (independently). That is, after the agent chooses p_t, the reward y_t is drawn from a Bernoulli distribution with parameter p_t and the net gain of the agent is $y_t - p_t$.

Note that the agent's choice p_t may depend on the full history of choices and outcomes but the reward at time t is solely based on an independent Bernoulli sample.

1. ***[10 points]*** Show that S_t, the agent's total net gain after the t-th round, is a martingale.
2. ***[10 points]*** Using Azuma-Hoeffding's inequality, show that the probability that the agent's net gain exceeds h at any of the first n rounds is at most $\exp(-\frac{h^2}{2n})$ (i.e. $P(\max_{1\le t\le n} S_t > h) \le \exp(-\frac{h^2}{2n})$)

4 Concentration inequalities *[25 points]*

Consider the problem of binary classification with n labeled examples of the form $(x_i, y_i)_{i=1}^n$ in your dataset with $i \in \mathcal{X}$ and $y_i \in \{0, 1\}$. Suppose you have a finite set $\mathcal{H}$ of binary classifiers where each $h \in \mathcal{H}$ is a mapping of the form $h : \mathcal{X} \to \{0, 1\}$. Let $\mathbf{1}(h(x) \neq y)$ be an indicator which takes the value 0 if $h(x) = y$ and 1 otherwise. If the samples are drawn i.i.d. according to a joint distribution D over (x, y), show that the following statements are true:

1. ***[8 points]*** By Hoeffding's inequality, for a fixed $h \in \mathcal{H}$, with probability at least $1 - \delta$:
$$\left| \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}(h(x_i) \neq y_i) - \mathbb{E}_{(X,Y)\sim D}[\mathbf{1}(h(X) \neq Y)] \right| \le \sqrt{\frac{1}{2n} \log \frac{2}{\delta}}.$$
2. ***[5 points]*** Let $\widehat{h} = \arg\min_{h\in\mathcal{H}} \sum_{i=1}^{n} \mathbf{1}(h(x_i) \neq y_i)$ and $h^\star = \arg\min_{h\in\mathcal{H}} \text{err}(h)$, where we define $\text{err}(h) = \mathbb{E}_{(X,Y)\sim D}\, \mathbf{1}(h(X) \neq Y)$. The previous result, combined with a union bound yields with probability at least $1 - \delta$:
$$\text{err}(\widehat{h}) - \text{err}(h^\star) \le \sqrt{\frac{2}{n} \log \frac{2|\mathcal{H}|}{\delta}}.$$

3. ***[12 points]*** Using Bernstein's inequality instead of Hoeffding's in the first bound, show the following more refined guarantee. With probability at least $1-\delta$:

$$\left|\frac{1}{n}\sum_{i=1}^{n}\mathbf{1}(h(x_i)\neq y_i)-\text{err}(h)\right|\leq\sqrt{\frac{2\text{err}(h)(1-\text{err}(h))}{n}\log\frac{2}{\delta}}+\frac{2}{3n}\log\frac{2}{\delta}.$$

Hint: Use the identity $\sqrt{a+b}\leq\sqrt{a}+\sqrt{b}$.

4. **[Bonus]** Using the inequality above, along with a union bound, show that with probability at least $1-\delta$:

$$\text{err}(\widehat{h})-\text{err}(h^\star)\leq 8\sqrt{\frac{2\text{err}(h^\star)}{n}\log\frac{2|\mathcal{H}|}{\delta}}+\frac{14}{n}\log\frac{2|\mathcal{H}|}{\delta}\,.$$

Hint: Use the identities $\text{err}(\widehat{h})=(\text{err}(\widehat{h})-\text{err}(h^\star))+\text{err}(h^\star)$ and $\sqrt{a+b}\leq\sqrt{a}+\sqrt{b}$.

5 Non-convex optimization and convergence rates to stationary points *[25 points]*

Let us say a function $F:\mathbb{R}^d\to\mathbb{R}$ is L *-smooth* if

$$\|\nabla F(w)-\nabla F(w')\|\leq L\|w-w'\|\,,$$

where the norm $\|\cdot\|$ is the Euclidean norm. In other words, the derivatives of F do not change too quickly.

Gradient descent, with a constant learning rate, is the algorithm:

$$w^{(k+1)}=w^{(k)}-\eta\cdot\nabla F(w^{(k)})$$

In this question, we do *not* assume that F is convex. If you find it helpful, you can assume that F is twice differentiable.

1. ***[15 points]*** Now let us bound the function value decrease at every step. In particular, show that the following holds (for all η):

$$F(w^{(k+1)})\leq F(w^{(k)})-\eta\|\nabla F(w^{(k)})\|^2+\frac{1}{2}\eta^2L\|\nabla F(w^{(k)})\|^2$$

It is fine to prove a looser version of this bound, where the factor of $1/2$ is replaced by 1. Brownie points if you get the factor of $1/2$. (Hint: Taylor's theorem is the natural starting point. You may also want to consider what smoothness implies about the second derivative. If you think about the intermediate value theorem, you can actually get the factor of $1/2$).

2. ***[3 points]*** Let us now show that if the gradient is large, then it is possible to substantially decrease the function value. Precisely, show that with an appropriate setting of η, we have that:

$$F(w^{(k+1)})\leq F(w^{(k)})-\frac{1}{2L}\|\nabla F(w^{(k)})\|^2$$

3. ***[7 points]*** Now let $F(w_*)$ be the minimal function value (i.e. the value at the global minima). Argue that gradient descent will find a $w^{(k)}$ that is "almost" a stationary point in a bounded (and polynomial) number of steps. Precisely, show that there exist some k where:

$$k \leq \frac{2L(F(w^{(0)}) - F(w_*))}{\epsilon}$$

such that

$$\|\nabla F(w^{(k)})\|^2 \leq \epsilon\,.$$

6 Markov chains and convergence to stationary distributions *[30 points]*

Let $G = (V, E)$ be an undirected graph. Let P be the transition matrix which specifies a Markov chain on G. Let X_t denote the node of the graph which the Markov chain is in, after t steps, and let P be the transition probability matrix underlying the Markov chain. That is, $P(i, j) = \mathbb{P}(X_{t+1} = i \mid X_t = j)$. Note that the probability does not depend on the path taken by the chain before time t because of the Markov assumption in a Markov chain. Let d_i denote the degree of node i (number of neighbors of i) in G and let $N = |V|$ be the total number of nodes in G. A vector v is called the stationary distribution of the Markov chain induced by P if v satisfies $v = Pv$. Intuitively, v_i represents $\mathbb{P}(X_t = i)$ for t sufficiently large. Note that this probability is not always independent of the starting state X_1, meaning that a stationary distribution need not always exist. A Markov chain is called ergodic if it has a unique stationary distribution v.

1. ***[5 points]*** Show that P satisfies the regularity properties: $P(i, j) \geq 0$ and $\sum_{i=1}^{N} P(i, j) = 1$. If P is started in an initial distribution u_0, show that the distribution u_t after t steps of the chain is given by $P^t u_0$.
2. ***[5 points]*** Suppose that P is uniform, that is $P(i, j) = 1/d_j$ whenever (i, j) is an edge in E. Show that the vector $v_i = d_i / \sum_{j \in V} d_j$ satisfies $v = Pv$. That is, the stationary distribution of the uniform Markov chain on an undirected graph is proportional to the degree distribution.
3. ***[10 points]*** It is the case that all the eigenvalues of P are smaller than or equal (in magnitude) to 1. Show this for the case of symmetric P so that all the eigenvalues are real.
4. **[*Bonus*]** Show the last question for general P with complex eigenvalues.
5. ***[10 points]*** Assume that G is a d-regular graph, that is $d_i = d$ for all nodes i. Suppose P is a uniform Markov chain on G, that is, $P(i, j) = 1/d$ for all $(i, j) \in E$. Show that starting from any distribution u_0, the Markov chain converges to the stationary distribution v as $t \to \infty$. Let λ be the second largest eigenvalue of P in absolute value. Given a bound on $\|u_t - v\|_1$ in terms of λ using the first three parts.

7 L_p norms *[20 points]*

For a vector $x \in \mathbb{R}^d$, the L_p norm is defined as:

$$\|x\|_p = (|x_1|^p + |x_2|^p + \ldots |x_d|^p)^{1/p}$$

We will be frequently making use of $p = 1, 2, \infty$ norms.

1. ***[5 points]*** Give a simple formula for the L_∞ norm, which is defined as $\|x\|_\infty = \lim_{p \to \infty} \|x\|_p$.
2. ***[5 points]*** Verify that $\|x\|_\infty$ is a norm on vector space, i.e. for vectors a and b and scalar c, it satisfies the following three properties: $\|a + b\|_\infty \leq \|a\|_\infty + \|b\|_\infty$; $\|ca\|_\infty = |c|\|a\|_\infty$; if $\|a\|_\infty = 0$ then a is the all 0's vector.
3. ***[5 points]*** Show that:

$$\|x\|_\infty \leq \|x\|_2 \leq \|x\|_1$$

4. ***[5 points]*** Verify the following special case of Holder's inequality for vectors a and b:

$$a \cdot b \leq \|a\|_1 \|b\|_\infty$$

8 Importance Sampling *[25 points]*

Suppose we want to evaluate the expected value of a function f under a distribution P, that is $\mu(f) := \mathbb{E}_{X \sim p} f(X)$. If we have access to samples from p, an estimate of this quantity can be obtained using a sample average, that is, $\widehat{\mu}(f) = \frac{1}{n} \sum_{i=1}^{n} f(x_i)$. In many cases, samples from p are either not available, or can be expensive to obtain (e.g. running a particle accelerator for validating theoretical models in physics). Often we have access instead to samples from another distribution q which is easy to sample from (such as a from a high-fidelity simulator). An important technique to use the samples from q to evaluate $\mu(f)$ is called importance sampling, where we estimate:

$$\widehat{\mu}(f) = \frac{1}{n} \sum_{i=1}^{n} \frac{p(x_i)}{q(x_i)} f(x_i). \tag{1}$$

We are assuming for now that p and q are the probability mass functions of a distribution, with X taking at most M distinct values. That is, we compute a similar sample average as when we have access to samples from p, but additionally reweight each sample by the ratio of probabilities under p and q. In this question, we will study some basic properties of this importance sampling technique.

1. ***[5 points]*** Show that $\widehat{\mu}(f)$ is not an unbiased estimator of $\mu(f)$ in general with an example. Based on the example, suggest a regularity condition relating p and q under which the estimator is unbiased.
2. ***[10 points]*** What is the variance of $\widehat{\mu}(f)$? Which distribution q minimizes this variance while still yielding an unbiased estimator $\widehat{\mu}(f)$?
3. ***[10 points]*** Suppose that $f(X) \in [0, 1]$ for all x. Using Hoeffding's inequality, show that with probability at least $1 - \delta$

$$|\widehat{\mu}(f) - \mu(f)| \leq \max_x \frac{p(x)}{q(x)} \sqrt{\frac{1}{2n} \log \frac{2}{\delta}}.$$

4. **[Bonus]** Can you obtain an improved bound using Bernstein's inequality?