Homework Assignment # 1

Due: Friday, Feb. 5, 2021, 11:59 p.m. Mountain time

Total marks: 100

Question 1. [15 MARKS]

Let X be a random variable with outcome space $\Omega = \{a, b, c\}$ and $p(a) = 0.1, p(b) = 0.2$, and $p(c) = 0.7$. Let

$$f(x) = \begin{cases} 10 & \text{if } x = a \\ 5 & \text{if } x = b \\ 10/7 & \text{if } x = c \end{cases}$$

(a) [4 MARKS] What is $E[f(X)]$?

(b) [3 MARKS] What is $E[1/p(X)]$?

(c) [4 MARKS] For an arbitrary pmf p, what is $E[1/p(X)]$?

(d) [4 MARKS] What is $E[f(X)^2]$ and $E[f(X)]^2$?

Question 2. [15 MARKS]

Suppose you have three coins. Coin A has a probability of heads of 0.75, Coin B has a probability of heads of 0.5, and Coin C has a probability of heads of 0.25.

(a) [5 MARKS] Suppose you flip all three coins at once, and let X be the number of heads you see (which will be between 0 and 3). What is the expected value of X, $E[X]$?

(b) [10 MARKS] Suppose instead you put all three coins in your pocket, select one at random, and then flip that coin 5 times. You notice that 3 of the 5 flips result in heads while the other 2 are tails. What is the probability that you chose Coin C? Hint: Define random variable D as the observed data, and notice that you can compute $p(D = 3 \text{ heads and } 2 \text{ tails} | \text{Coin} = C)$.

Question 3. [10 MARKS]

Alberta Hospital occasionally has electrical problems. It can take some time to find the problem, though it is always found in no more than 10 hours. The amount of time is variable; for example, one time it might take 0.3 hours, and another time it might take 5.7 hours. The time (in hours) necessary to find and fix an electrical problem at Alberta Hospital is a random variable, say X, whose density is given by the following uniform distribution

$$p(x) = \begin{cases} \frac{1}{10} & \text{if } 0 < x < 10 \\ 0 & \text{otherwise} \end{cases}$$

Such electrical problems can be costly for the Hospital, more so the longer it takes to fix it. The cost of an electrical breakdown of duration x is x^3. What is the expected cost of an electrical breakdown? Show your work.

Question 4. [35 MARKS]

To better visualize random variables and get some intuition for sampling, this question involves some simple simulations, which is a central theme in machine learning. You will also get some

experience using `numpy`, a scientific computating library commonly used in machine learning, which you will also need to use in later assignments. Use the attached code called `simulate.py`. This code is a simple script for sampling and plotting with python; play with some of the parameters to see what it is doing. Calling `simulate.py` runs with default parameters; `simulate.py 1 100` simulates 100 samples from a 1d Gaussian. Note that if you do not have matplotlib installed, you will have to install it.

For the first two questions, the goal is to understand how much estimators themselves can vary: how different our estimate would have been under a different randomly sampled dataset. In the real world, we do not get to obtain different estimators, we will only have one; in this controlled setting, though, we can actually simulate how different the estimators could be.

For the second two questions, the goal is to understand how we to obtain confidence intervals for our single sample average estimator.

(a) [5 MARKS] Change the code such that it prints the mean and variance of your samples, using Numpy functions. You are only required to add these two lines in simulate.py.

(b) [7 MARKS] Run the code for 10 samples with dim=1 and $\sigma^2 = 1.0$. Write down the sample average that you obtain. Now do this another 4 times, giving you 5 estimates of the sample average M_1, M_2, M_3, M_4 and M_5. What is the sample variance of these 5 estimates?

(c) [7 MARKS] Now run the same experiment, but use 100 samples for each sample average estimate. What is the sample variance of these 5 estimates? How is it different from the variance when you used 10 samples to compute the estimates?

(d) [8 MARKS] Now let us consider a higher variance situation, where $\sigma^2 = 10.0$. Imagine you know this variance, and that the data comes from a Gaussian, but that you do not know the true mean. Run the code to get 30 samples, and compute one sample average M. What is the 95% confidence interval around this M? Give actual numbers.

(e) [8 MARKS] Now assume you know less: you **do not know** the data is Gaussian, though you still know the variance is $\sigma^2 = 10.0$. Use the same 30 samples from (d) and resulting sample average M. Give a 95% confidence interval around M, now without assuming the samples are Gaussian.

Question 5. [25 MARKS]

We have talked about the fact that the sample mean estimator $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ is an unbiased estimator of the mean μ for identically distributed $X_1, X_2, \ldots, X_n$: $\mathbb{E}[\bar{X}] = \mu$. The sample variance, on the other hand, is not an unbiased estimate of the true variance σ^2: for $\bar{V} = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2$, we get that $\mathbb{E}[\bar{V}] = (1 - \frac{1}{n})\sigma^2$. Instead, the following bias-corrected sample variance estimator is unbiased: $\bar{V}_b = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$.

(a) [15 MARKS] Use the fact that $\mathbb{E}[\bar{V}_b] = \sigma^2$ to show that $\mathbb{E}[\bar{V}] = (1 - \frac{1}{n})\sigma^2$. **Hint:** The proof is short, it can be done in a few lines.

(b) [10 MARKS] We also discussed the variance of the sample mean estimator, and concluded that $\mathrm{Var}[\bar{X}] = \frac{1}{n}\sigma^2$, for iid variables with variance σ^2. We can similarly ask what the variance is of the sample variance estimator. Deriving the formula is a bit more complex for general random variables, so let's assume the X_i are zero-mean Gaussian. Then we know that the following is true (though we omit the derivation): $\mathrm{Var}[\bar{V}] = \frac{2(n-1)}{n^2}\sigma^4$.

This variance enables us to use Chebyshev's inequality, to get a confidence estimate. Recall that Chebyshev's inequality states that for a random variable Y with known variance v, we know that

$\Pr(|Y - \mathbb{E}[Y]| < \epsilon) > 1 - v/\epsilon^2$. Note that for zero-mean Gaussian X_i, we can use $\bar{V} = \frac{1}{n}\sum_{i=1}^{n} X_i^2$, which is unbiased (i.e., $\mathbb{E}[\bar{V}] = \sigma^2$). After seeing 10 samples from a distribution, do you think you will have a tighter confidence estimate around the sample mean $\bar{X}$ or the sample variance $\bar{V}$? Explain why.

Homework policies:

Your assignment should be submitted as a single pdf document and a zip file with code, on eClass. The answers must be written legibly and scanned or must be typed (e.g., Latex). All code should be turned in when you submit your assignment.

Because assignments are more for learning, and less for evaluation, grading will be based on coarse bins. **The grading is atypical**. For grades between (1) 80-100, we round-up to 100; (2) 60-80, we round-up to 80; (3) 40-60, we round-up to 60; and (4) **0-40, we round down to 0**. The last bin is to discourage quickly throwing together some answers to get some marks. The goal for the assignments is to help you learn the material, and completing less than 50% of the assignment is ineffective for learning.

We will not accept late assignments. Plan for this and aim to submit at least a day early. If you know you will have a problem submitting by the deadline, due to a personal issue that arises, please contact the instructor as early as possible to make a plan. If you have an emergency that prevents submission near the deadline, please contact the instructor right away. Retroactive reasons for delays are much harder to deal with in a fair way.

All assignments are individual. All the sources used for the problem solution must be acknowledged, e.g. web sites, books, research papers, personal communication with people, etc. Academic honesty is taken seriously; for detailed information see the University of Alberta Code of Student Behaviour.

Good luck!