Homework 6

Due Tuesday, November 3

Please either give the assignment to Loraine at the CDS or send it via email to the graders **before noon**.

1. *Uniform distribution.* You have data that you have reasons to think can be well modeled as iid samples from a uniform distribution from 0 to u (where u is an unknown parameter).
   a. Given the realization $x_1, x_2, \ldots, x_n$ of an iid vector $X_1, X_2, \ldots, X_n$ what are the possible values that the parameter u could have?
   b. Compute the maximum-likelihood estimator $\widehat{U}_{\text{ML}}$ of u given n iid samples.
   c. Compute the pdf of $\widehat{U}_{\text{ML}}$.
   d. Is the estimator unbiased?
   e. Show that $\widehat{U}_{\text{ML}}$ converges to u in probability.
2. *Half life* The half life of a radioactive isotope is the time it takes for half the amount of the isotope to decay (assuming that the original amount is not extremely small). In probabilistic terms, if T represents the waiting time until a particular atom of the isotope decays,

$$P\left(T \leq \text{half time}\right) = \frac{1}{2}. \tag{1}$$

   Assume that we have a sample from an unknown isotope that could be carbon-15 (half life: 2.45 s), carbon-10 (half life: 19.29 s) or seaborgium-266 (no, we are not making the name up, half life: 30 s). We aim to identify what isotope we are dealing with.
   a. If the isotopes decay according to an exponential distribution, what is the parameter of the exponential distribution of each element?
   b. We monitor n atoms from the unknown isotope and determine their decay times. What is the maximum-likelihood estimate of the parameter of the exponential distribution given these decay times? Does it make sense to restrict the set of possible values of the parameter or would you choose it to be $[0, \infty]$?
   c. If we have a prior for the probability of the sample being made of each of the isotopes, what is the form of the posterior distribution of the parameter given the data?
   d. Does it make sense to use the posterior mean as a point estimate for the parameter of the exponential? If you don't think so make an alternative suggestion.
   e. A chemist working with you determines that the probabilities that the isotope is carbon-15, carbon-10 and seaborgium-266 are 50%, 49% and 1% (seaborgium is a synthetic element) respectively. Complete the script hw6pb2.py, which plots the posterior distribution for 10 and 100 samples taken from an exponential distribution with the parameter corresponding to seaborgium-266. Explain the results but don't submit the plots.
3. *Empirical probability mass function.* You are hired to estimate the distribution of ages (in years) in New York. You sample n people uniformly at random in the city and ask them their age in years.

a. What is a reasonable nonparametric estimator for the pmf of a random variable representing the age of people in New York?

b. Is the estimator unbiased?

c. Is the estimator consistent? Use convergence in mean square as a criterion for consistency.

d. What problem will we run into if we record the age at a higher granularity (for example in minutes) and apply this estimator?

4. *Call center* A company that runs a call center hires you as a consultant. They are interested in the probability of receiving a certain number of calls in any 10 minute interval between 8 pm and 12 am during October in order to decide whether to hire more operators. Your task is to estimate these probabilities from data provided by the company[1]. In particular, you have available two datasets:

   - *2-day data:* Number of calls in 10-minute intervals between 8 pm and 12 am from two days at the beginning of October.
   - *September data:* Number of calls in 10-minute intervals between 8 pm and 12 am from the whole month of September.

   You decide to estimate the distribution of the calls in October from these two datasets by using a parametric and a nonparametric approach.

   a. Telephone calls are often modeled using Poisson distributions. What is the maximum-likelihood estimator of the parameter of a Poisson random variable given iid samples?

   b. Complete the code in hw6pb4.py. This code estimates the Poisson parameter corresponding to the 2-day data and the September data. It compares the corresponding pmfs to the pmf of October and computes the estimation error (using the ℓ_1 norm of the difference between the distributions). Then it computes the empirical pmf of the 2-day data and the September data and again compares them to the pmf of October. What errors do you obtain? (You don't need to submit the plots, just look at them.)

   c. What is the method that performs best for the 2-day data? What method is better for the September data? What does this suggest about parametric against non-parametric estimation depending on the amount of data available?

5. *Method of moments* The method of moments is an alternative way of fitting parameters from iid data within a frequentist point of view. It is of interest theoretically or in scenarios where computing the maximum likelihood estimator is computationally intractable. The idea is very simple. Assume that we want to fit the parameter θ of a distribution. Let μ be the mean of the distribution of interest. We define the function g as a function such that

$$\mu = g\left(\theta\right). \tag{2}$$

   Now, let $\bar{X}_n$ be the sample mean of the samples. For any realization, we can set

$$\bar{x}_n = g\left(\hat{\theta}\right) \tag{3}$$

[1]The data in this problem, which was compiled from a real call center in Israel, is available at `http://iew3.technion.ac.il/serveng/callcenterdata`.

and solve the equation to obtain an estimator for θ. If we want to estimate more parameters we can use higher moments to set up analogous equations. Compute the method-of-moments estimator for the parameter of the following distributions and compare it to the ML estimator (if the ML estimator has been derived in the notes or in another problem *don't* recompute it).

a. Bernoulli.

b. Geometric.

c. Poisson.

d. Exponential.