EE-411, HomeWork 1 : Maximum Likelihood & Probability

This homework involves some coding, with the language of your choice. Present your results with graphs, plots, and data. Jupyter notebooks are a good option, and we recommend you to send your work as a notebook on Google colab.

1 First passage time of a 1D Brownian particle

A bit of context : Particles suspended in fluids can be described using the theory of random processes, and in particular *Brownian motion.* In 1D, the motion is described by the diffusion equation $\partial_t p(x,t|x_0) = D\partial_x^2 p(x,t|x_0)$, where x_0 is the particle starting position and D the *diffusion constant.* It is possible to show, see e.g. Wikipedia, that the (density of) probability that a particle first reaches a point x_c (at distance $d = |x_c - x_0|$ from the starting point) at time t is given by

$$p(t|d,D) = \frac{d}{\sqrt{4\pi D t^3}} \exp\left(-\frac{d^2}{4Dt}\right) , \tag{1}$$

which is the so-called *Lévy distribution.* In the following, we will fix the distance d, and starting from the observation of set of times $\{t_i\}$ we aim to find an estimate of the the diffusion constant D.

1. Write explicitly the probability $p(\{t_i\}_{i=1}^n|d,D)$ to observe a set of n *independent* events at times $\{t_i\}_{i=1}^n$ and the normalized log-likelihood $\mathcal{L}(\{t_i\}_{i=1}^n|d,D)) = \frac{\log(p(\{t_i\}_{i=1}^n|d,D))}{n}$.
2. Write a program that simulates n such observations sampled from the probability distribution (1) for fixed $d = 2$ and D generic. This can be done, for instance, with the scipy.stats.levy method in python, choosing as parameters $loc = 0$ and $scale = d^2/(2D)$. Start with $n = 10$ observations and plot the likelihood as a function of D. Repeat for $n = 20, 100$, discuss and comment on what you see.
3. We now assume that we are given a set of n observations $\{t_i\}_{i=1}^n$, without being told the true value D^*. We consider the maximum likelihood estimator

   $$\hat{D}_{\text{ML}}(\{t_i\}_{i=1}^n) = \arg\max_D \left(\log\left(p(\{x_i\}_{i=1}^n|d,D)\right)\right) \tag{2}$$

   and we shall define the squared error as $\text{SE} = (\hat{D}_{ML}(\{t_i\}_{i=1}^n) - D^*)^2$.
   Create some data sets with $n = 10, 100, 1000$ for different values of $D^* \in (0,5]$ and see how the ML estimator performs. Note that finding the maximizer $\hat{D}_{ML}$ can be done numerically in python, for instance using the scipy.minimize method.
4. One can show (*Bonus 1 : prove it!*) that the Fisher information in our problem is given by

   $$I_n(D) = nI(D) = n\mathbb{E}_t\left[\left(\frac{\partial}{\partial D}\mathcal{L}(t|d,D))\right)^2\right] = \frac{n}{2D^2} . \tag{3}$$

   Knowing this, another interesting estimator is given by the maximum as posteriori (MAP) with the Jeffreys prior :

   $$\hat{D}_{\text{J}}(\{x_i\}_{i=1}^n) = \arg\max_D \left(\log\left(p\left(\{x_i\}_{i=1}^n|d,D\right)\sqrt{I(D)}\right)\right) . \tag{4}$$

   Implement this estimator using scipy and repeat the analysis done in point 3.
5. If we average on many realizations (say about a hundred) we can obtain numerically the averaged mean squared error $MSE(D^*, \hat{D}, n)$ which is thus a function of n, D^* and of the estimator $\hat{D}$. Compute and plot, for $n = 10, 100, 1000$, the curves $MSE(D^*, \hat{D}_{\text{ML}}, n)$ and $MSE(D^*, \hat{D}_{\text{J}}, n)$ as a function of D^*.
6. How do the MSE curves at various n compare with the Cramér-Rao bound for unbiased estimator $\text{MSE}(\hat{D}) \geq \frac{1}{nI(D^*)}$ (where $I(D)$ is the Fisher information)? How do the Jeffrey and ML estimator behave? Which one would you choose?
7. *Bonus 2* : Look at the median of the Lévy distribution on Wikipedia. Can you design an estimator for D from it? Repeat the analysis at point 3 and at points 5 and 6 for this third estimator and compare its performances to the other two.

2 Probability bounds and a pooling problem

We are going to follow the steps we took in lecture 1 and prove an interesting inequality : Let $Z_1, \dots, Z_m$ be independent random variables *Bernoulli*-distributed, i.e. such that $Z_i = 1$ with probability p, and 0 with probability $1-p$. Then, for any $\epsilon \geq 0$ we have

$$\mathbb{P}\left(\frac{1}{m}\sum_i Z_i \geq p + \epsilon\right) \leq e^{-2m\epsilon^2} \tag{5}$$

1. Our starting point is to realize that $\mathbb{P}(a \geq b) = \mathbb{P}(e^{\lambda a} \geq e^{\lambda b})$ for any $\lambda \geq 0$. Using Markov inequality and the proof strategy discussed in the lectures, show that :

$$\mathbb{P}\left(\frac{1}{m}\sum_i Z_i \geq p + \epsilon\right) \leq \left(\frac{pe^{\lambda} + (1-p)}{e^{\lambda(p+\epsilon)}}\right)^m \tag{6}$$

Hint : Since the Z_i are independent, we can write $\mathbb{E}_{\mathbf{Z}}[\exp \sum_i Z_i] = \prod_i \mathbb{E}_{Z_i}[\exp Z_i]$.

2. Using the value of λ that minimizes the right-hand-side of the former equation, show that

$$\mathbb{P}\left(\frac{1}{m}\sum_i Z_i \geq p + \epsilon\right) \leq e^{-mf(p,\epsilon)}$$

with

$$f(p,\epsilon) = (p+\epsilon)\log\left(\frac{p+\epsilon}{p}\right) + (1-(p+\epsilon))\log\left(\frac{1-(p+\epsilon)}{1-p}\right)$$

Hint : Instead of differentiating the right-hand-side of (6), one can differentiate its log*, since the logarithm is monotone and thus its extrema are the same of its argument.*

3. Show that

$$f(p,\epsilon=0) = 0, \ \left.\frac{\partial f(p,\epsilon)}{\partial \epsilon}\right|_{\epsilon=0} = 0, \ \text{and that } \frac{\partial^2 f(p,\epsilon)}{\partial \epsilon^2} \geq 4 \ \text{ for any } \epsilon.$$

4. Use Taylor's theorem (that states that $f(p,\epsilon) = f(p,0) + \epsilon f'(p,0) + \epsilon^2 f''(p,\tilde{\epsilon})/2$ for some unknown $\tilde{\epsilon}$, and where the prime stands for derivative with respect to ϵ) to show that $f(p,\epsilon) \geq 2\epsilon^2$, and prove the inequality (5).

Similarly, it is possible to show (*Bonus 3 : prove it*) that

$$\mathbb{P}\left(\frac{1}{m}\sum_i Z_i \leq p - \epsilon\right) \leq e^{-2m\epsilon^2} \qquad \text{so that} \qquad \mathbb{P}\left(\left|\frac{1}{m}\sum_i Z_i - p\right| \geq \epsilon\right) \leq 2e^{-2m\epsilon^2} \tag{7}$$

5. The most important use of such a bound is in terms of pooling problems. Suppose you want to know what fraction of the population in a country approves its current president : how many people should you ask to be confident, with probability at least 95 percent, that the error in estimating the fraction of people who approves the president is correct within one percent (so that $\hat{p}$ is in $[p-0.01, p+0.01]$ with 95% probability) ?

6. Compare the number m^* you find this way with what you observe when performing numerical experiments in python :
   - Define a function that takes the number of people m and the probability p as arguments and returns a random array of m votes. *Hint : You can generate Bernoulli-distributed samples in python using the* scipy.stats.bernoulli *method.*
   - Starting with fixed values of $m = m^*$ and $p \in \{0.2, 0.5, 0.8\}$
     - Use this function evaluated in m^* and p to simulate polls.
     - Just by averaging the generated votes, estimate p.
     - Quantify the probability that $\hat{p}$ is correct within one percent.
   - Which values of p seem to be harder to estimate ? Do you find that the bound is accurate, or does it grossly overestimate the needed number ?
   - For each p, repeat for different values of m to find the value that (more or less) gives an estimate which is correct within one percent with 95% probability.
   - *Bonus 4* : Plot the behaviour of the probability of error $\mathbb{P}\big(\hat{p} \notin [p-0.01, p+0.01]\big)$ as a function of p for values of $m \in [10, 10^4]$, and compare it with the theoretical m^*.