Homework #1

ECE 461/661: Introduction to Machine Learning for Engineers
Prof. Gauri Joshi and Prof. Carlee Joe-Wong
Due: Monday, January 27, 2020 at 8:59 pm PT/11:59 pm ET

Please remember to show your work for all problems and to write down the names of any students that you collaborate with. The full collaboration and grading policies are available on the course website: https://www.andrew.cmu.edu/course/18-661/.

Your solutions should be uploaded to Gradescope (https://www.gradescope.com/) in PDF format by the deadline. We will not accept hardcopies. If you choose to hand-write your solutions, please make sure the uploaded copies are legible. Gradescope will ask you to identify which page(s) contain your solutions to which problems, so make sure you leave enough time to finish this before the deadline. We will give you a 30-minute grace period to upload your solutions in case of technical problems.

1 Warm-up *[20 points]*

a. (5 points) **Multivariable Calculus**: Consider a real function $f(x, z) = x\cos(z)e^{-3x+z}$, where $x, z \in \mathbb{R}$. What is the partial derivative of $f(x, z)$ with respect to x?

b. (5 points) **Mean and Variance**: If the variance of a zero-mean random variable X is σ^2, what is the variance of $2X$? What about the variance of $X + 2$?

c. (5 points) **Probability**: Consider the following joint distribution between X and Y.

$P(X,Y)$		Y: a	b	c
X	T	0.2	0.1	0.2
	F	0.05	0.15	0.3

What is $P(X = T|Y = b)$?

d. (5 points) Show that the function $f(x) = |x| + \exp(x)$ over the domain $x \in \mathbb{R}$ is convex.

2 Linear algebra *[15 points]*

a. (3 points) The covariance matrix $\boldsymbol{\Sigma}$ of a random column vector $\boldsymbol{X}$ is defined as $\boldsymbol{\Sigma} = \mathbb{E}[(\boldsymbol{X} - \mathbb{E}\boldsymbol{X})(\boldsymbol{X} - \mathbb{E}\boldsymbol{X})^\top]$, where $\mathbb{E}\boldsymbol{X}$ is the expectation of $\boldsymbol{X}$. Is $\boldsymbol{\Sigma}$ positive-semidefinite? Why? Recall a matrix is positive semi-definite if for any vector x, $x'Ax \geq 0$.

b. (6 points) Let $\boldsymbol{A}, \boldsymbol{B} \in \mathbb{R}^{n\times n}$ be two symmetric matrices. Suppose $\boldsymbol{A}$ and $\boldsymbol{B}$ have the exact same set of eigenvectors $\boldsymbol{u}_1, \boldsymbol{u}_2, \cdots, \boldsymbol{u}_n$ with the corresponding eigenvalues $\alpha_1, \alpha_2, \cdots, \alpha_n$ for $\boldsymbol{A}$, and $\beta_1, \beta_2, \cdots, \beta_n$ for $\boldsymbol{B}$. Write down the eigenvectors and their corresponding eigenvalues for the following matrices:

(i) $\boldsymbol{C} = \boldsymbol{A} + \boldsymbol{B}$

(ii) $\boldsymbol{D} = \boldsymbol{A} - \boldsymbol{B}$

(iii) $\boldsymbol{E} = \boldsymbol{AB}$

(iv) $\boldsymbol{F} = \boldsymbol{A}^{-1}\boldsymbol{B}$ (assume $\boldsymbol{A}$ is invertible)

c. (6 points) Let $\boldsymbol{A} \in \mathbb{R}^{m \times n}$, $\boldsymbol{b} \in \mathbb{R}^m$ be given. For a given value of m, under what conditions on $\boldsymbol{A}$ and $\boldsymbol{b}$ will the equation $\boldsymbol{Ax} = \boldsymbol{b}$ have

(i) no solution?

(ii) one solution?

(iii) infinitely many solutions?

3 Vector-valued functions *[15 points]*

Compute the first and second derivatives of the following functions:

a. $f(\mathbf{x}) = \mathbf{c}^\top \mathbf{x}$, where $\mathbf{x}, \mathbf{c} \in \mathbb{R}^m$.

b. $f(\mathbf{x}) = \mathbf{M}^\top \mathbf{M}\mathbf{x}$, where $\mathbf{x} \in \mathbb{R}^m$ and $\mathbf{M} \in \mathbb{R}^{m \times m}$. What happens if $\mathbf{M} = \mathbf{M}^\top$?

4 Entropy *[10 points]*

Let X be a random variable with the Bernoulli distribution $B(1, p)$ where $0 < p < 1$. The entropy of X is defined as $H(p) = -p \log p - (1-p)\log(1-p)$.

a. Derive the second derivative $H''(p)$ of $H(p)$. If $H''(p) < 0$, $H(p)$ is called concave. Is $H(p)$ a concave function of p?

b. Find the value of $p \in (0, 1)$ that maximizes $H(p)$.

5 MLE *[25 points]*

Suppose that $\mathbf{x} \in \mathbb{R}^d$ is a random variable with probability distribution D, *i.e.*, $\mathbf{x} \sim D$. Moreover, assume that $\mathbf{w} \in \mathbb{R}^d$ is a parameter vector and let

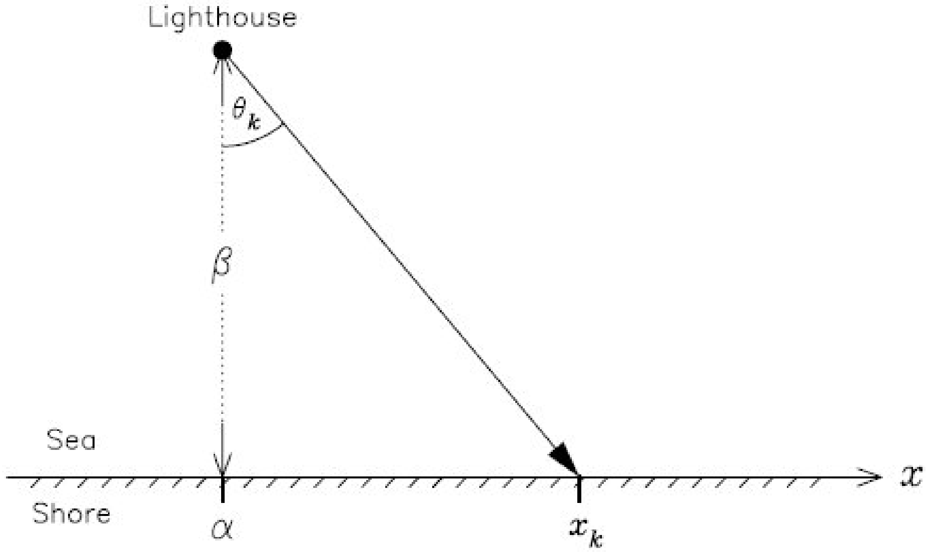
$$y = \langle \mathbf{x}, \mathbf{w} \rangle + \epsilon, \text{ where } \epsilon \sim \mathcal{N}(0, \sigma^2)\,. \tag{1}$$

Therefore, y is a linear function of $\mathbf{x}$ with Gaussian noise added.

a. (5 points) Suppose we take $\mathbf{x}$ and $\mathbf{w}$ to be fixed and given. We then see that $y|\mathbf{x}, \mathbf{w}$ is a random variable: it is the sum of the fixed, given constant $\langle \mathbf{x}, \mathbf{w} \rangle$ and the Gaussian random variable ϵ. What is the conditional distribution of $y|\mathbf{x}, \mathbf{w}$?

b. (6 points) Assume that we (independently) draw n pairs $(\mathbf{x}_i, y_i) \in \mathbb{R}^d \times \mathbb{R}$ from the above model, *i.e.* we draw $\mathbf{x}_i$ from D and then, given $\mathbf{x}_i$, we draw a value for y_i according to (1). What is the distribution of $(y_1, \ldots, y_n)|(\mathbf{x}_1, \ldots, \mathbf{x}_n), \mathbf{w}$?

c. (3 points) Write down the log-likelihood function of $(y_1, \ldots, y_n)|(\mathbf{x}_1, \ldots, \mathbf{x}_n), \mathbf{w}$, using your results from part (b) above.



d. (6 points) We now wish to solve the MLE problem, i.e., to find the parameter $\mathbf{w}$ that maximizes the log-likelihood function you derived in part (c). Suppose that $d = 1$, i.e., that $\mathbf{x} \in \mathbb{R}$ and $\mathbf{w} \in \mathbb{R}$; we can then write $\langle \mathbf{x}, \mathbf{w} \rangle$ as the product $\mathbf{xw}$. Derive the optimal parameter $\mathbf{w}$ that maximizes the log-likelihood.

   Hint: The log-likelihood is a concave function of $\mathbf{w}$.

e. (5 points) Now suppose that $d > 1$. Generalize your result from part (d) to derive the parameter $\mathbf{w}$ that maximizes the log-likelihood of the conditional model from part (c).

6 Python *[20 points]*

Follow the instructions below and use the included Python code along with your own code to solve the Lighthouse problem.

When uploading to Gradescope, you will need to produce a PDF version of your solutions and code. One way to do this is to use a notebook (`https://jupyter.org`); if you wish to use this, we have provided a Jupyter version of the problem where you can fill in your solutions in hw1.ipynb, which can be downloaded from `https://www.andrew.cmu.edu/course/18-661/`.

6.1 Problem: the lighthouse

(from D. Sivia's book, "Data Analysis - A Bayesian Tutorial"):

A lighthouse is somewhere off a piece of straight coastline at a position α along the shore and a distance β out at sea. It emits a series of short highly collimated flashes. The flashes are sent out at random azimuths, where the azimuth are chosen uniformly at random between $[0, \pi]$. These pulses are intercepted on the coast by photo-detectors that record only the fact that a flash has occurred, but not the angle from which it came. N flashes have been recorded so far at positions $\{x_k\}$.

Suppose β is given. Where is the lighthouse?

6.2 Guided solution

We need to estimate the parameter α. Let us start by writing the likelihood for this problem; since the flashes are thrown at random azimuths, we know that:

$$P(\theta_k|\alpha, \beta) = \frac{1}{\pi}.$$

Moreover,

$$\beta \tan(\theta_k) = x_k - \alpha,$$

and by changing variables we get

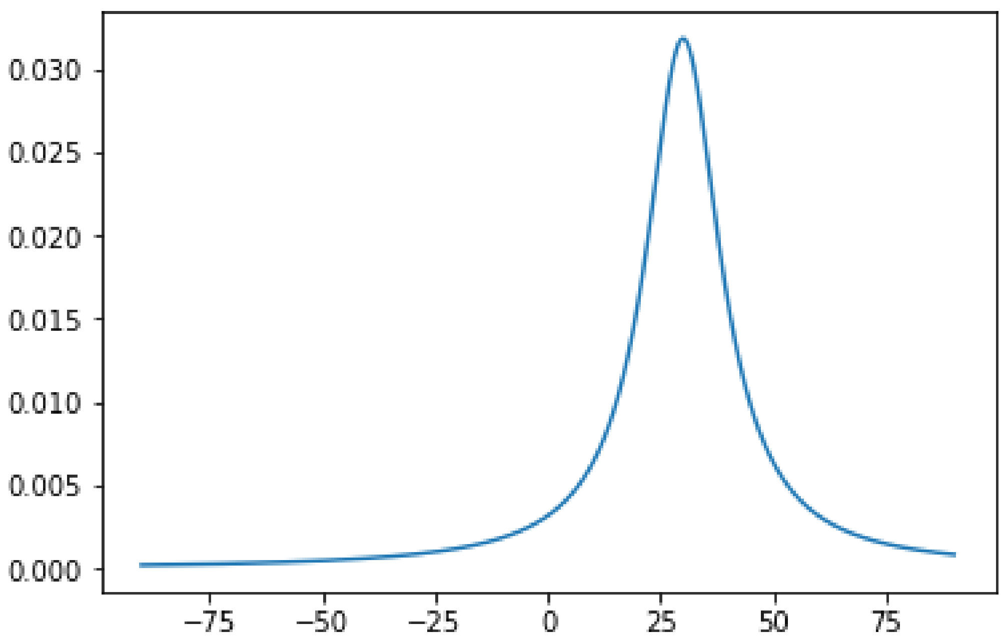
$$P(x_k|\alpha, \beta) = \frac{\beta}{\pi\left[\beta^2 + (x_k - \alpha)^2\right]}.$$

```
In [2]: # Scientific computing and plotting packages
        import numpy as np
        import matplotlib.pyplot as plt

        # Likelihood definition
        def likelihood(x, alpha, beta):
            return beta / (np.pi * (beta ** 2 + (x - alpha) ** 2))

        # Parameters
        alpha = 30.0 # alpha appears here, only for simmulations purposes, we want to find the value of
        beta  = 10.0 # beta is given

        #Compute and plot the likelihood
        x = np.linspace(-90, 90, 1001)
        plt.plot(x, likelihood(x, alpha, beta))
        plt.show()
```



The above likelihood is the a Cauchy or Lorentz distribution. We will sample from it so that we can have some synthetic data to work with.

6.3 Generate synthetic data

In [3]:
```
from scipy.stats import cauchy
samples = cauchy.rvs(loc = alpha, scale = beta, size = 1000)
```

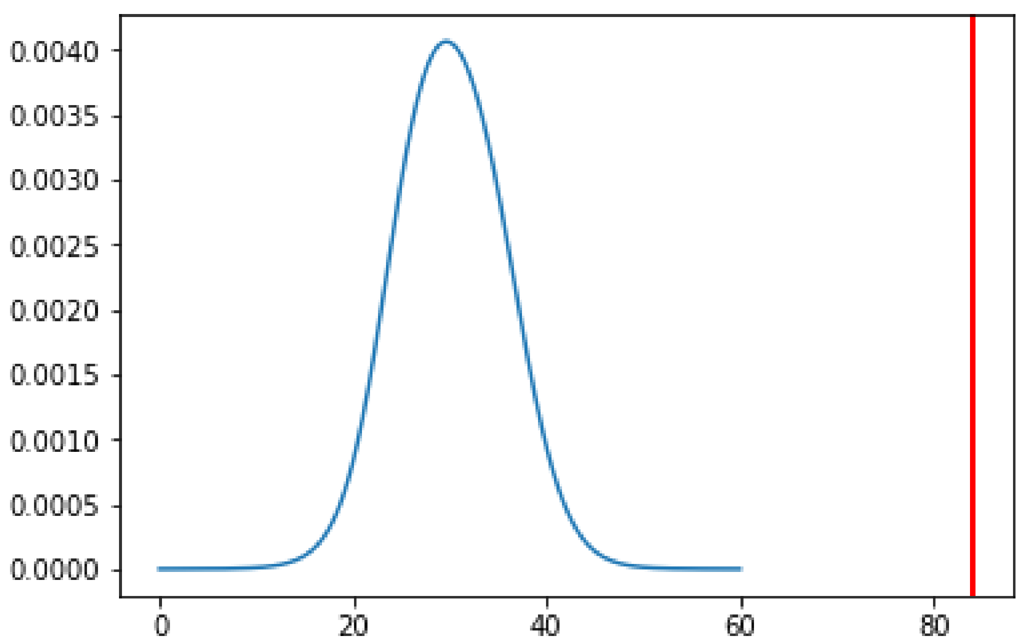
Assuming our prior $P(\alpha)$ is a uniform distribution, the posterior probability is

$$P(\alpha|\{x_k\},\beta) \propto \prod_{k=1}^{N} P(x_k|\alpha,\beta)P(\alpha|\beta) \propto \prod_{k=1}^{N} P(x_k|\alpha,\beta)$$

In [4]:
```
# Computes the (unnormalized) posterior for a given set of samples
def posterior(x, alpha, beta):
    post = np.ones(len(alpha))
    for x_k in x:
        post *= likelihood(x_k, alpha, beta)
        post /= np.sum(post)
    return post

def plot_posterior(n_samples):
    alphas = np.linspace(0, 60, 1001)
    plt.plot(alphas, posterior(samples[:n_samples], alphas, beta))
    plt.axvline(np.mean(samples[:n_samples]), c = "r", lw = 2)

plot_posterior(10)
plt.show()
```



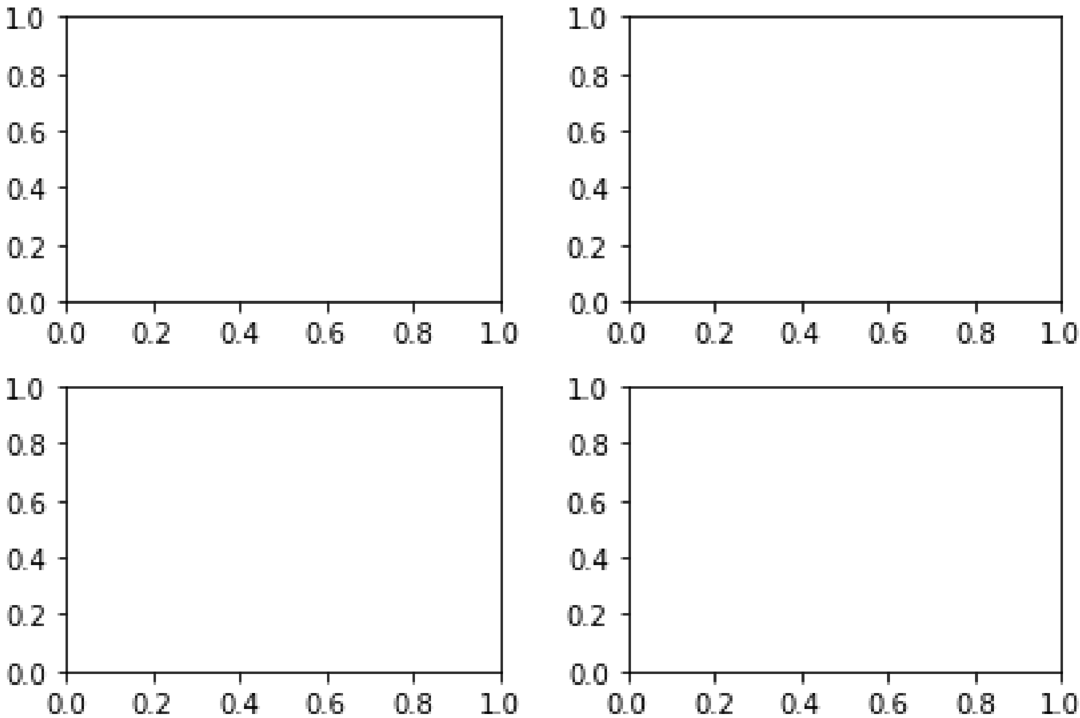
Exercise 1: Create 4 subplots for different values of $N = 2, 5, 20, 100$.

```
In [5]: fig, axs = plt.subplots(2, 2)
        fig.tight_layout()

        alphas = np.linspace(0, 60, 1001)

        # Your solution goes here

        plt.show()
```



Note the mean does not coincide with the mode of the posterior!

Why is that? Will they coincide in the $N \to \infty$ limit?

Now compute the value of α that maximizes the posterior (and the likelihood, since our prior here is uniform). The log-likelihood reads:

$$\mathcal{L}(\alpha) = \sum_k \log P(x_k|\alpha, \beta) = -\sum_k \log[\beta^2 + (x_k - \alpha)^2)] + \mathrm{c},$$

where c is a constant.

Hence the maximum is obtained at

$$2\sum_k \frac{x_k - \alpha^*}{\beta^2 + (x_k - \alpha^*)^2} = 0.$$

Now let's solve this numerically for different values of N.

Exercise 2: Plot the ML estimate of α for N between 10 and 1000.

In [6]:
```
# Use a off the shelf method to find a root of a function on an interval - ex: bisect, brentq,
from scipy.optimize import bisect # Bisection method is probably the simpler to understand

# Your solution goes here
```