You can use Mathematica as an aid for many of the computations, however make sure to do hand calculations where suitable as well.
Note: The Mathematica based solution are in `https://github.com/yoninazarathy/MATH7501-2021/blob/master/Assignment3Sol/Sol3Mathematica.pdf`

1. Consider the exponential distribution with probability density function $f(x) = \lambda e^{-\lambda x}$ defined on $x \geq 0$ and with parameter $\lambda > 0$.

   (a) Show that $f(x)$ is a valid probability density function by showing that the integral over $[0, \infty)$ is unity.

   (b) Use integration to show that the mean of the distribution is $\frac{1}{\lambda}$.

   (c) Use integration to show that the variance of the distribution is $\frac{1}{\lambda^2}$.

   (d) Determine the median of the distribution. The median is the number M such that,

   $$\int_0^M f(x)\,dx = \frac{1}{2}.$$

   (e) The quantile function of the distribution, $q(u)$ for $u \in [0, 1)$, is defined as follows: For each u, we should have,

   $$\int_0^{q(u)} f(x)\,dx = u.$$

   Determine an expression for $q(u)$.

   (f) Say that U is a uniformly distributed random variable on $[0, 1]$. If you set a new random variable X, via $X = q(U)$, then the distribution of X is exponential (for $q(\cdot)$ evaluated for an exponential distribution as in the item above). Show this empirically for $\lambda = 3$ by generating 10^6 uniform random variables, and comparing the empirical quantile of this data with $q(\cdot)$.

Solution:
(a)

$$\int_0^\infty \lambda e^{-\lambda x}\,dx = \lim_{L\to\infty} \int_0^L \lambda e^{-\lambda x}\,dx = \lim_{L\to\infty} \left[-e^{-\lambda x}\right|_{x=0}^{x=L} = \lim_{L\to\infty} (-e^{-\lambda L} - (-1)) = 1.$$

(b) First compute $\int x\lambda e^{-\lambda x}\,dx$. Using integration by parts with $u = x$ and $v' = \lambda e^{-\lambda x}$, and thus $u' = 1$ and $v = -e^{-\lambda x}$ we get,

$$\int x\lambda e^{-\lambda x}\,dx = uv - \int u'v\,dx = -xe^{-\lambda x} + \int e^{-\lambda x}dx = -xe^{-\lambda x} - \frac{1}{\lambda}e^{-\lambda x}.$$

Compute now,

$$\lim_{L\to\infty} \left[-xe^{-\lambda x} - \frac{1}{\lambda}e^{-\lambda x}\right|_0^L = 0 - 0 - (-0 - \frac{1}{\lambda}) = \frac{1}{\lambda}.$$

(c) We'll remember that one way to compute the variance is,

$$\text{variance} = \text{second moment} - \text{mean}^2.$$

Hence we'll focus on computation of the second moment,

$$\int_0^\infty x^2\lambda e^{-\lambda x}\,dx.$$

To compute $\int x^2\lambda e^{-\lambda x}\,dx$ we'll need integration by parts twice. Set $u = x^2$ and $v' = \lambda e^{-\lambda x}$, and thus $u' = 2x$ and $v = -e^{-\lambda x}$. We then get,

$$\int x^2\lambda e^{-\lambda x}\,dx = uv - \int u'v\,dx = -x^2e^{-\lambda x} + \int 2xe^{-\lambda x}dx = -x^2e^{-\lambda x} + \frac{2}{\lambda}\int x\lambda e^{-\lambda x}dx$$

Now observe that the last integral is the same we computed above and thus we obtain,

$$\int x^2\lambda e^{-\lambda x}\,dx = -x^2e^{-\lambda x} + \frac{2}{\lambda}\Big(-xe^{-\lambda x} - \frac{1}{\lambda}e^{-\lambda x}\Big).$$

Now considering the improper integral we again take $xto\infty$ and subtract at the value at $x = 0$ to get that the second moment is $2/\lambda$. Hence the variance is,

$$\frac{2}{\lambda^2} - \frac{1}{\lambda^2} = \frac{1}{\lambda^2}.$$

(d) For both this item and the next we'll observe that compute the value of the so called CDF (Cumulative Distribution Function):

$$F(x) = \int_0^x \lambda e^{-\lambda u}\,du = 1 - e^{-\lambda x}.$$

Observe that $F(0) = 0$. Observe that $F(x)$ is a non-decreasing (actually strictly increasing) function. Finally observe that $\lim_{x\to\infty} F(x) = 1$.

Now to compute the median we want $F(M) = \frac{1}{2}$ or ,

$$1 - e^{-\lambda M} = \frac{1}{2},$$

or after solving for M,

$$M = \frac{1}{\lambda}\log 2.$$

(e) In a similar vain to the previous question we need $F\big(q(u)\big) = u$, or,

$$1 - e^{-\lambda q(u)} = u.$$

Or,

$$q(u) = -\frac{1}{\lambda}\log(1-u).$$

(f) See Mathematica file for solution.

2. Consider the normal probability distribution with parameters $\mu \in \mathbb{R}$ and $\sigma > 0$. The probability density is,

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{(x-\mu)^2}{2\sigma^2}}.$$

   (a) Showing that $f(x)$ is a valid probability density function is not immediate. Do this first numerically for $\mu = 2$ and $\sigma = 3$ by approximating the integral via a discretization sum over 100, 1,000, 10,000, and 10^5 terms. You should observe that as the number of terms grows, the value of the sum approaches 1.

   (b) Use integration to show that the mean of the distribution is μ.

   (c) Use integration to show the variance of the distribution is σ^2.

(d) The k'th moment of the distribution, denoted m_k for $k = 1, 2, 3, \ldots$, is

$$m_k = \int_{-\infty}^{\infty} x^k f(x)\, dx.$$

Based on the previous items, $m_0 = 1$, $m_1 = \mu$, and $m_2 = \mu^2 + \sigma^2$. Show that for higher valued $k > 2$, we have,

$$m_k = \mu\, m_{k-1} + (k-1)\sigma^2 m_{k-2}.$$

(e) Use this recurrence relation to compute m_4 for $\mu = 2$ and $\sigma = 3$. Compare this value to a numerical computation of the integral both using a discretization, and Mathematica's in-built, `NIntegrate[]` function.

(f) Show analytically that,

$$\int_{-\infty}^{\infty} f(x)\, dx = 1.$$

You may use material from the lecture or online material, but must justify and explain your calculations.

Solution:

(a) See Mathematica file for solution.

(b) See solution in last lecture of class.

(c) See solution in last lecture of class.

(d) See for example `https://people.smp.uq.edu.au/YoniNazarathy/teaching_projects/studentWork/EricOrjebin_TruncatedNormalMoments.pdf`.

(e) We get $m_4 = \mu^4 + 6\mu^2\sigma^2 + 3\sigma^4$. See Mathematica file for numerical part.

(f) See solution in last lecture of class. Also see pages 154-155 in the course reader.

3. Assume you are presented with univariate data of a random sample, $x_1, \ldots, x_n$ and wish to find a single number, x^* that summarizes $x_1, \ldots, x_n$ as best as possible. One way to specify this in terms of a loss function is to seek a value x^* that minimizes,

$$L(u) = \sum_{i=1}^{n} (x_i - u)^2.$$

Analytically, it is very easy to show that $x^* = \sum_{i=1}^{n} x_i/n$, the sample mean. Nevertheless, it is good to see how this number can be reached via a gradient descent algorithm. Set $\eta > 0$, start with some arbitrary initial $x(0)$. Then you get a sequence of points $x(t)$, for $t = 1, 2, 3, \ldots$ via,

$$x(t+1) = x(t) - \eta \nabla L\big(x(t)\big).$$

(a) Show that,

$$x(t) = \alpha^t x(0) + \beta \frac{1-\alpha^t}{1-\alpha}.$$

for some α and β (specify these values in terms of of the problem parameters and data).

(b) Determine the range of η values for which $x(t)$ will converge to x^*.

Solution:

See this taken from Chapter 9 of "Statistics with Julia":

For this we momentarily return to the basic statistical inference setting of Chapter **??** and assume we are presented with univariate data of a random sample, $x_1, \ldots, x_n$. In the context of machine learning, this can be viewed as unsupervised learning because there are only features and no labels. Assume now that we wish to find a single number, x^* that summarizes $x_1, \ldots, x_n$ as best as possible. One way to specify this in terms of a loss function is to seek a value x^* that minimizes,

$$L(u) = \sum_{i=1}^{n} (x_i - u)^2. \tag{9.7}$$

Analytically, it is very easy to show that $x^* = \overline{x}$, the sample mean. This can be done by taking the derivative (gradient in one dimension) of the loss function. The gradient is,

$$\nabla L(u) = -2 \sum_{i=1}^{n} (x_i - u) = -2\Big(\sum_{i=1}^{n} x_i\Big) + 2\,n\,u. \tag{9.8}$$

If we equate it to 0 and solve for u, we obtain $u = \overline{x}$. Further, the second derivative is $2n > 0$ (positive) indicating that $\overline{x}$ is a minimum. Alternatively, without using calculus, we may represent $L(u)$ as an upward facing parabola in u via,

$$\underbrace{\text{UQ username}}_{a} + \underbrace{\Big(-2\sum_{i=1}^{n} x_i\Big)}_{b} u + \underbrace{\sum_{i=1}^{n} x_i^2}_{c}.$$

This then allows us to read off the minimum value of the parabola at $u = -\frac{b}{2a} = \overline{x}$. In any case we see that the sample mean has the interpretation of minimizing the sum of squared deviations in the data.

With such a simple solution for the minimization of (9.7) there is no practical reason to execute gradient descent for this problem. All one needs to do is compute the sample mean. Nevertheless, to get a better feel for gradient descent it is useful to consider how the algorithm performs on this problem. For this consider (9.6) and use the gradient expression (9.8). This means, that,

$$\begin{aligned}\theta(t+1) &= \theta(t) - \eta\Big(-2\Big(\sum_{i=1}^{n} x_i\Big) + 2\,n\,\theta(t)\Big) \\ &= \underbrace{(1 - 2n\eta)}_{\alpha}\,\theta(t) + \underbrace{2\eta \sum_{i=1}^{n} x_i}_{\beta}.\end{aligned}$$

Now a recursion of the form $\theta(t+1) = \alpha\theta(t) + \beta$, starting at some value $\theta(0)$ yields,

$$\theta(t) = \alpha^t \theta(0) + \beta \frac{1 - \alpha^t}{1 - \alpha}.$$

Such a form can be obtained by iterating the recursion and summing up a geometric sum. This means that if $|\alpha| < 1$ then,

$$\lim_{t \to \infty} \theta(t) = \frac{\beta}{1 - \alpha} = \frac{2\eta \sum_{i=1}^{n} x_i}{1 - (1 - 2n\eta)} = \overline{x}.$$

The condition $|\alpha| < 1$ is equivalent to $\eta < n^{-1}$. That is we see that if η is not too large, the sequence converges to the minimum fr UQ username ng point $\theta(0)$. Otherwise it does not. Further it can be seen that the fastest convergence occurs when η is arbitrarily close to n^{-1} from below. This type of behavior of the learning rate is typical: On the one hand, too large of a learning rate implies that gradient descent does not converge. On the other hand, too low of a learning rate implies that convergence is very slow. In this simple toy example we can analytically analyze the learning rate. However in more realistic examples that follow, experimentation is needed. This falls under the umbrella of *hyper-parameter tuning*.

4. Consider the simple linear regression problem where you are presented with data points $(x_1, y_1), \ldots, (x_n, y_n)$. You seek β_0 and β_1 to fit the line,

$$y = \beta_0 + \beta_1 x,$$

by minimizing the loss function

$$L(\beta_0, \beta_1) = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2.$$

This minimization is often carried out by methods others than gradient descent, but for the purposes of this exercise you will use gradient descent.

(a) Compute an expression for the gradient $\nabla L(\beta_0, \beta_1)$.

(b) Return to question 5 from Assignment 1. In that question you dealt with data points,

$$(2.4, 3.1), (4.7, 2.7), (4.9, 4.8), (2.9, 7.6), (8.1, 5.4),$$

and fit a line parameterized by β_0 and β_1. Do this now numerically using a gradient descent algorithm using the expression for the gradient you developed above. Make sure that the learning rate (step size rate), η is small enough for convergence. Illustrate your numerical experiments.

Solution:
This problem was essentially solved in the last lecture.