1. For a given real vector with elements $z = [z_1, \ldots, z_n]^T$, let the function $s : \mathbb{R}^n \to \mathbb{R}^n$ be,

$$s(z) = \frac{1}{\sum_{i=1}^n e^{z_i}} [e^{z_1} \quad \ldots \quad e^{z_n}]^T.$$

This is sometimes called the *softmax* function. Establish the following properties of $s(\cdot)$:

(a) $s(z) = s(z + c\mathbf{1})$ where $\mathbf{1}$ is a vector of 1's and c is a constant.

(b) For any vector z, $s(z)$ is a probability vector. That is,

$$\sum_{i=1}^n s(z)_i = 1, \qquad \text{and} \qquad s(z)_i \in [0, 1].$$

(c) Take now a sequence of vectors $z^{(1)}, z^{(2)}, \ldots$ that converges to a limit,

$$\lim_{t\to\infty} z^{(t)} = z^*,$$

with $z^*_{i^*} > z^*_j$ for all $j \neq i^*$. Further consider the sequence $s(z^{(1)}), s(z^{(2)}), \ldots$. Establish that,

$$\lim_{t\to\infty} s(z^{(t)}) = s(z^*),$$

and that $z^*_{i^*}$ is the *mode* of this limiting probability distribution.

Solution: (a) For any vector $z = [z_1, \ldots, z_n]^T \in \mathbb{R}^n$,

$$z + c\mathbf{1} = [z_1, \ldots, z_n]^T + [c, \ldots, c]^T = [z_1 + c, \ldots, z_n + c]^T.$$

Hence, from the definition of s,

$$\begin{aligned} s(z + c\mathbf{1}) &= \frac{1}{\sum_{i=1}^n e^{z_i+c}} [e^{z_1+c} \quad \ldots \quad e^{z_n+c}]^T \\ &= \frac{e^c}{\sum_{i=1}^n e^{z_i+c}} [e^{z_1} \quad \ldots \quad e^{z_n}]^T \\ &= \frac{1}{\sum_{i=1}^n e^{z_i+c-c}} [e^{z_1} \quad \ldots \quad e^{z_n}]^T \\ &= s(z). \end{aligned}$$

(b) Note that $e^x > 0$ for any real value x. Hence, for any real valued vector $z = [z_1, \ldots, z_n]^T$, for all $i = 1, \ldots, n$,

$$0 < e^{z_i} < \sum_{i=1}^n e^{z_i},$$

and thus,

$$0 < s(z)_i < 1.$$

Furthermore,

$$\sum_{i=1}^n s(z)_i = \sum_{i=1}^n \frac{e^{z_i}}{\sum_{j=1}^n e^{z_j}} = 1.$$

(c) Since $s(z)_i = e^z_i / \sum_{j=1}^n e^{z_j}$ is differentiable at every $z_i \in \mathbb{R}$, from Theorem 5.2 of [Rud76], $s(z_i)$ is continuous for all $i = 1, \ldots, n$. This continuity implies that

$$\lim_{t\to\infty} s(z^{(t)})_i = s(\lim_{t\to\infty} z^{(t)})_i = s(z^*)_i$$

for all $i = 1, \ldots, n$. As a consequence of Theorem 4.10 (a) of [Rud76],

$$\lim_{t \to \infty} s(z^{(t)}) = s(z^*).$$

From the definition, it is easy to observe that $s(z)_i > s(z)_j$ if and only if $e^{z_i} > e^{z_j}$ for any $i, j \in \{1, \ldots, n\}$ and $z \in \mathbb{R}^n$. Furthermore, it is easy to show that the exponential function e^x is monotone in x, that is, $e^x > e^{x'}$ if and only if $x > x'$ for any $x, x' \in \mathbb{R}$. Together, we conclude that $s(z)_i > s(z)_j$ if and only if $z_i > z_j$.

From (b), $s(z^*)$ is a probability vector. Hence, the mode of $s(z^*)$ is $z^*_{i^*}$ since i^* is such that $z^*_{i^*} > z^*_j$ for all $j \neq i^*$, which implies $s(z^*_j) > s(z^*_{i^*})$.

2. Let the function $\sigma(u) = max(u, 0)$. This is sometimes called the *ReLU activation function.* Consider the matrices $A_1 \in \mathbb{R}^{2\times 3}$ and $A_2 \in \mathbb{R}^{1\times 2}$, and the vectors $b_1 \in \mathbb{R}^2$ and $b_2 \in \mathbb{R}$. Consider now the function $\rho : \mathbb{R}^3 \to \mathbb{R}$ constructed via,

$$\rho(x) = \sigma(A_2\sigma(A_1 x + b_1) + b_2),$$

where $\sigma(\cdot)$ applied to a vector element-wise.

(a) Assume b_1 and b_2 are zero vectors and that A_1 and A_2 have all entries 1. Determine $\rho([1 \;\; 1 \;\; 1]^T)$.

(b) Assume A_1 has first two columns as the 2×2 identity matrix and third column with zero entries. Further assume $b_1 = [0 \;\; 0]^T$, $A_2 = [1 \;\; 0]$, and $b_2 = 10$. For what values of x do we have $\rho(x) \neq 0$?

(c) It can be shown that for any values of A_1, A_2, b_1, and b_2 we can represent $\rho(\cdot)$ as,

$$\rho(x) = C_{i(x)}\; x + d_{i(x)},$$

where C_i and d_i are fixed for each $i = 1, \ldots, K$ and $i(x) : \mathbb{R}^3 \to \{1, \ldots, K\}$ determines a *piece-wise affine* partition of $\mathbb{R}^3$. That is, it partitions $\mathbb{R}^3$ into K polytopes and returns the index of the associated polytope. Determine an upper bound on K.

Solution: (a)

$$\begin{aligned}
\rho([1 \;\; 1 \;\; 1]^T) &= \sigma(A_2\sigma(A_1[1 \;\; 1 \;\; 1]^T + 0) + 0) \\
&= \sigma(A_2\sigma([3 \;\; 3]^T)) \\
&= \sigma(A_2[3 \;\; 3]^T)) \\
&= \sigma(6) \\
&= 6.
\end{aligned}$$

(b)

$$\begin{aligned}
\rho([x_1 \;\; x_2 \;\; x_3]^T) &= \sigma(A_2\sigma(A_1[x_1 \;\; x_2 \;\; x_3]^T + [0 \;\; 0]^T) + 10) \\
&= \sigma(A_2\sigma([x_1 \;\; x_2]^T) + 10) \\
&= \sigma(A_2[\sigma(x_1) \;\; \sigma(x_2)]^T) + 10) \\
&= \sigma(\sigma(x_1) + 10) \\
&= \sigma(x_1) + 10 \\
&\geq 10.
\end{aligned}$$

Hence for any input vector x, $\rho(x) \neq 0$.

(c) The vector $z^{(1)} := A_1 x + b_1$ is 2 dimensional and the vector $\sigma(z^{(1)})$ can be one of the four values, $[0 \;\; 0]^T$, $[z_1^{(1)} \;\; 0]^T$, $[0 \;\; z_2^{(1)}]^T$, or $[z_1^{(1)} \;\; z_2^{(1)}]^T$. This depends on the signs of the two entries of $z^{(1)}$ and hence partitions the input space $\mathbb{R}^3$ into 4 polytopes. For example, the polytope associated with $\sigma(z^{(1)}) = [0 \;\; z_2^{(1)}]^T$ is the set,

$$\mathcal{P} = \{x \in \mathbb{R}^3 \mid \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix} (A_1 x + b_1) \leq \begin{bmatrix} 0 \\ 0 \end{bmatrix}\}.$$

That is, whenever $x \in \mathcal{P}$ we have that $\sigma(z^{(1)}) = [0 \;\; z_2^{(1)}]^T$. Similar polytopes exist for the other three possibilities.

Now $z^{(2)} := A_2 \sigma(z^{(1)}) + b_2)$ is a scalar and $\rho(x) = \sigma(z^{(2)})$ can be either 0 or $z^{(2)}$ depending on the sign of $z^{(2)}$. Here, for every value of $\sigma(z^{(1)}) \in \mathbb{R}^2$ we have that the scalar $z^{(2)}$ is positive or not, so there are again two polytopes, now partitioning $\mathbb{R}^2$.

Hence in total there are $K = 4 \times 2 = 8$ possible polytopes that partition the input space $\mathbb{R}^3$ and

$$\rho(x) = C_{i(x)} \; x + d_{i(x)},$$

where $i(x)$ determines which polytope we are in.

As an example, assume that the inner $\sigma(\cdot)$ zeros out the second coordinate but not the first, and further assume that the outer $\sigma(\cdot)$ does not zero out its (scalar) input. Then for x that cause such behavior we have,

$$\rho(x) = A_2 \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix} (A_1 x + b_1) + b_2.$$

Denote this case by i^*, then

$$C_{i^*} = A_2 \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix} A_1, \qquad \text{and} \qquad d_{i^*} = A_2 \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix} b_1 + b_2.$$

Now this particular i^* occurs for the polytope which is the intersection of the two polytopes

$$\mathcal{P}_{i^*}^{(1)} = \{x \in \mathbb{R}^3 \mid \begin{bmatrix} -1 & 0 \\ 0 & 1 \end{bmatrix} (A_1 x + b_1) \leq \begin{bmatrix} 0 \\ 0 \end{bmatrix}\},$$

and,

$$\mathcal{P}_{i^*}^{(2)} = \{x \in \mathbb{R}^3 \mid -A_2 \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix} (A_1 x + b_1) - b_2 \leq 0\}.$$

Note that an intersection of two such polytopes can be represented as a third polytope, say $\mathcal{P}_{i^*}$ by combining the inequalities the define the polytopes. Hence $i(x) = i^*$ whenever $x \in \mathcal{P}_{i^*}$. Similar polytopes and C, d elements may be defined for each of the 8 possibilities.

3. We consider now the least squares problem, where we attempt to solve the linear system of equations $Aw = b$, with A an $m \times n$ matrix and b an m-vector. Consider the loss function,

$$L(w) = ||Aw - b||^2,$$

where the norm is the L_2 norm and we are seeking w that minimizes $L(w)$.

(a) Explain why the minimizer of $L(w)$ is the same as the minimizer of $||Aw - b||$.

(b) Show that the gradient of $L(\cdot)$ is $\nabla L(w) = 2A^T(Aw - b)$.

(c) Establish that $L(w)$ has a unique minimizer if and only if the columns of A are linearly independent.

Solution: (a) Say that w^* minimizes $L(\cdot)$. This means,

$$L(w^*) \leq L(w), \qquad \forall w \in \mathbb{R}^n.$$

Now we can consider the minimized function as $L(w) = g\big(f(w)\big)$ where $f(w) = ||Aw - b||$ and $g(u) = u^2$. Note that $f(w) \geq 0$ and that for $u \geq 0$ the function $g(\cdot)$ is monotonic increasing and continuous. This means that $g^{-1}(z)$ exists for $z \geq 0$. In this case $g^{-1}(z) = \sqrt{z}$. That is,

$$g\big(f(w^*)\big) \leq g\big(f(w)\big), \qquad \forall w \in \mathbb{R}^n.$$

Now apply $g^{-1}(\cdot)$ to both sides to obtain,

$$f(w^*) \leq f(w), \qquad \forall w \in \mathbb{R}^n.$$

or,

$$||Aw^* - b|| \leq ||Aw - b||, \qquad \forall w \in \mathbb{R}^n.$$

This shows that the minimizer of $L(w)$ is the same as minimizer of $||Aw - b||$. We can also go the other way and show that the minimizer of $||Aw - b||$ is the minimizer of $L(w)$.

(b) One can obtain this formula directly by considering,

$$L(w) = (Aw - b)^T(Aw - b) = \sum_{i=1}^{m}(Aw - b)_i^2 = \sum_{i=1}^{m}\Big(\big(\sum_{j=1}^{n} A_{ij}w_j\big) - b_i\Big)^2$$

and then taking the k'th element of $\nabla L(w)$ as the derivative of $L(w)$ with respect to w_k. However we will do it by considering $L(w)$ as composition of the function $h(x) = ||x||^2$ and the function $g(x) = Ax - b$.

In general when considering a function $f(x) = h(g(x))$ where say $g : \mathbb{R}^n \to \mathbb{R}^p$ and $h : \mathbb{R}^p \to \mathbb{R}^q$, the Jacobian of f at the point x, denoted $Df(x)$ can be represented via the Jacobians of h and g, namely $Dh(\cdot)$ and $Dg(\cdot)$. This is the multivariate chain rule:

$$Df(x) = Dh(g(x))Dg(x).$$

Here the Jacobian $Df(x)$ is an $q \times n$ matrix with (i, j) coordinate being $\partial f(x)_i/\partial x_j$. Similarly, the Jacobian $Dh(\cdot)$ is a $q \times p$ matrix and $Dg(\cdot)$ is a $p \times n$ matrix. In our case $q = 1$ and $Df(x) = \nabla f(x)^T$. Further in our case

$$h(u) = ||u||^2 = \sum_{i=1}^{n} u_i^2$$

and thus $\nabla h(u) = 2u$ or $Dh(u) = 2u^T$. Now $g(u) = Au - b$, an affine function, and thus $Dg(u) = A$. Hence using the chain rule,

$$\nabla L(w) = Df(w)^T = Dg(w)^T Dh(g(w))^T = A^T 2g(w)^{T^T} = A^T 2(Aw-b) = 2A^T(Aw-b).$$

(c) First observe that the objective, $L(w)$ is a convex function. We can do this in different ways, one of which is considering the Hessian matrix. The Hessian matrix $H(w)$ is the Jacobian of the gradient evaluated at any point $w \in \mathbb{R}^n$. In this case it is independent of w and is,

$$H(w) = 2A^T A.$$

In this case of the columns of A linearly independent we have that for any $y \in \mathbb{R}^n$ with $y \neq 0$,

$$y^T H(w) y = y^T 2A^T Ay = 2||Ay||^2 > 0.$$

Hence the Hessian is positive definite (everywhere - for every w). Proceeding similarly, even if the columns of A are linearly dependent then for any y,

$$y^T H(w) y = y^T 2A^T A y = 2||Ay||^2 \geq 0.$$

Hence the Hessian matrix is positive semi-definite. Thus in any case, we see $L(w)$ is convex via the Hessian. Now a convex differentiable function possesses a global minimum at the point w with $\nabla L(w) = 0$. That is, we equate the gradient to the zero vector to obtain the normal equations,

$$A^T A w = A^T b.$$

Now if A has linearly independent columns then $A^T A$ is non-singular and there is a unique solution to these normal equations via,

$$w^* = (A^T A)^{-1} A^T b := A^\dagger b.$$

To see this assume a vector x such that,

$$A^T A x = 0.$$

Showing that $A^T A$ is non-singular will be done if can show $x = 0$. To do so, left multiply by x^T to get,

$$x A^T A x = (Ax)^T Ax = ||Ax||^2 = 0.$$

This then means that $Ax = 0$. But since the columns of A are linearly independent it means that $x = 0$.

Proceeding in the other direction, if A has linearly independent columns then there is an $x \neq 0$ such that, $Ax = 0$. Left multiplying by A^T implies that,

$$A^T A x = 0, \qquad \text{with} \qquad x \neq 0.$$

That is, the nullspace of $A^T A$ is not just the zero vector and hence there multiple solutions to the normal equations.

4. Consider a sequence of i.i.d. (independent and identically distributed) uniform$(0, 1)$ random variables denoted $U_1, U_2, \ldots$.

   (a) Define the sequence of random variables $Y_1, Y_2, \ldots$ via,

   $$Y_i = \frac{U_1 - a}{b},$$

   for constants a and $b > 0$. Find a and b such that the expectation of Y_1 is 0 and the variance is 1.

   (b) Consider now sequence, $\overline{X}_1, \overline{X}_2, \ldots$, where

   $$\overline{X}_n = \frac{1}{n} \sum_{i=1}^{n} U_i.$$

   You are interested in the probability

   $$\mathbb{P}\Big(2\sqrt{3n}\left(|\overline{X}_n| - 0.5\right) > 2\Big).$$

   Use the central limit theorem to approximate this probability for non-small n.

(c) Estimate the above probability for $n = 3$ using Monte-Carlo by generating a million instances of $\overline{X}_3$ and compare to the normal approximation above. Repeat for $n = 30$ and compare again. Use the programming language of your choice (R, Python, Matlab, Julia, etc...).

Solution: (a) Since U_1 is a uniform$(0, 1)$ random variable, first note that $\mathbb{E}[U_1] = 1/2$ and $\mathbb{V}\mathsf{ar}(U_1) = 1/12$. From the relation between Y_1 and U_1, for any constants a, b, the expectation

$$\mathbb{E}[Y_1] = \frac{1}{b}\mathbb{E}[U_1 - a] = \frac{1}{b}(\mathbb{E}[U_1] - a) = \frac{1}{b}(1/2 - a),$$

and the variance

$$\begin{aligned}\mathbb{V}\mathsf{ar}(Y_1) &= \frac{1}{b^2}\mathbb{V}\mathsf{ar}(U_1 - a) &(1)\\ &= \frac{1}{b^2}\mathbb{V}\mathsf{ar}(U_1) &(2)\\ &= \frac{1}{12\,b^2},\end{aligned}$$

where (1) holds from the fact that for any random variable X and a constant c, the variance of $c\,X$ is equal to c^2 time the variance of X. The equality in (2) holds because we can not change the variance of a random variable by adding a constant to it, that is, $\mathbb{V}\mathsf{ar}(X) = \mathbb{V}\mathsf{ar}(X + c)$.

Given that $\mathbb{E}[Y_1] = 0$ and $\mathbb{V}\mathsf{ar}[Y_1] = 1$. By solving

$$\frac{(1/2 - a)}{b} = 0, \quad \text{and} \quad \frac{1}{12\,b^2} = 1,$$

we obtain $a = 1/2$ and $b = 2\sqrt{3}$.

(b) Since the mean and the variance of a uniform$(0, 1)$ are $1/2$ and $1/12$, respectively, from the central limit theorem for i.i.d. random variables,

$$\sqrt{n}\left(\frac{\overline{X}_n - 0.5}{1/12}\right) \xrightarrow{d} \mathcal{N}(0, 1), \quad \text{as } n \to \infty,$$

where $\mathcal{N}(0, 1)$ is a standard normal random variable (i.e., normal random variable with mean 0 and variance 1), and $\xrightarrow{d}$ denotes the convergence in distribution. Therefore, for non-small n, the probability distribution of $\sqrt{n}\left(\frac{\overline{X}_n - 0.5}{1/12}\right)$ can be approximated by the standard normal, that is, for any constant c,

$$\mathbb{P}\left(\sqrt{n}\left(\frac{\overline{X}_n - 0.5}{1/12}\right) > c\right) = \mathbb{P}\left(2\sqrt{3n}(\overline{X}_n - 0.5) > c\right) \approx \frac{1}{\sqrt{2\pi}}\int_c^\infty e^{-t^2/2}\mathsf{d}t.$$

In particular, with the choice of $c = 2$ and the observation that $|\overline{X}_n| = \overline{X}_n$,

$$\mathbb{P}\left(2\sqrt{3n}\big(|\overline{X}_n| - 0.5\big) > 2\right) \approx 0.02275.$$

(c) Sample code in Python:

```
"""
Quiz 0 Question 4 (c)
"""
import numpy as np
np.random.seed(12345)

def MC_estimate(n, n_iter=10**6):
    n_succ = 0
    for i in range(n_iter):
        x_bar = 0.0
        for _ in range(n):
            x_bar += np.random.random_sample()
        x_bar = x_bar/n
        y = 2*np.sqrt(3*n)*(x_bar - 0.5)
        if y > 2:
            n_succ += 1
        est_prob = n_succ/n_iter
    return est_prob

print('MC Estimation with n=3 is ', MC_estimate(3))
print('MC Estimation with n=30 is ', MC_estimate(30))
```

Output:
MC Estimation with $n = 3$ is 0.020952
MC Estimation with $n = 30$ is 0.022546

5. Write programs or script using the programming language of your choice that perform the following tasks. Use programming primitives as opposed to dedicated library functions:
   (a) Given an input matrix, check if it is: Symmetric, upper-triangular, lower-triangular.
   (b) Given a list of numbers, return a sorted list of numbers.
   (c) Given a sequence of n input vectors, each of length p, compute the $p \times p$ sample covariance matrix of these vectors.
   (d) Use brute-force Riemann sums to illustrate numerically that

$$\int_{x_1=-5}^{5} \int_{x_2=-5}^{5} \frac{1}{2\pi} e^{-\frac{x_1^2+x_2^2}{2}} \, dx_1 \, dx_2 \approx 1.$$

Solution:

See Python based solution:

https://colab.research.google.com/drive/1Afh08Y7_jWV9rx1cBImTXxjZ7UbWfApy?usp=sharing.

See Julia based solution:

https://nbviewer.jupyter.org/github/yoninazarathy/MathematicalEngineeringDeepLearning/blob/master/Quiz0/OpeningQuiz1JuliaSol.ipynb

See R based solution:

https://github.com/yoninazarathy/MathematicalEngineeringDeepLearning/blob/master/Quiz0/QuizOSol_5_Rlanguage.Rmd

References

[Rud76] Walter Rudin. *Principles of mathematical analysis*. International series in pure and applied mathematics. McGraw-Hill, New York, 3rd ed. edition, 1976.