Week 6 Exercise Sheet Solutions

The following exercises have different levels of difficulty indicated by (*), (**), (***). An exercise with (*) is a simple exercise requiring less time or effort to solve compared to an exercise with (***), which is a more complex exercise.

1 Logistic Regression

1. (**) Derive π from $\log\left(\frac{\pi}{1-\pi}\right) = \mathbf{w}^T\mathbf{x}$, i.e. derive the logistic sigmoid function from the logit function.

Solution:
To find the probability π as a function of the weighted input we first start by applying the exponential to both sides of the above function (which will cancel the log function on the left handside):

$$\log\left(\frac{\pi}{1-\pi}\right) = \mathbf{w}^T\mathbf{x} \tag{1}$$

$$\left(\frac{\pi}{1-\pi}\right) = \exp(\mathbf{w}^T\mathbf{x}). \tag{2}$$

From here we can multiply both sides by $1-\pi$ then start re-arranging the terms as follows:

$$\pi = (1-\pi)\exp(\mathbf{w}^T\mathbf{x}) \tag{3}$$

$$\pi + \pi\exp(\mathbf{w}^T\mathbf{x}) = \exp(\mathbf{w}^T\mathbf{x}) \tag{4}$$

$$(1+\exp(\mathbf{w}^T\mathbf{x}))\pi = \exp(\mathbf{w}^T\mathbf{x}) \tag{5}$$

$$\pi = \frac{\exp(\mathbf{w}^T\mathbf{x})}{1+\exp(\mathbf{w}^T\mathbf{x})}. \tag{6}$$

Now if we factorise the $\exp(\mathbf{w}^T\mathbf{x})$ term in the denominator it will cancel with the numerator as follows:

$$\pi = \frac{\exp(\mathbf{w}^T\mathbf{x})}{\exp(\mathbf{w}^T\mathbf{x})\left(-\exp(\mathbf{w}^T\mathbf{x})+1\right)} \tag{7}$$

$$\pi = \frac{1}{1+\exp(-\mathbf{w}^T\mathbf{x})}, \tag{8}$$

which is what we defined as the logsitic sigmoid in the lecture. Note you could alternatively find 1/ equation 2 to give

$$\frac{1-\pi}{\pi} = \exp(-\mathbf{w}^T\mathbf{x}) \tag{9}$$

$$\frac{1}{\pi} - 1 = \exp(-\mathbf{w}^T\mathbf{x}) \tag{10}$$

$$\frac{1}{\pi} = 1 + \exp(-\mathbf{w}^T\mathbf{x}) \tag{11}$$

$$\pi = \frac{1}{1+\exp(-\mathbf{w}^T\mathbf{x})}. \tag{12}$$

---

2. (*) In a binary (two-class) logistic regression model, the weight vector $\mathbf{w} = [4, -2, 5, -3, 11, 9]$. We apply it to some object that we'd like to classify; the vectorised feature representation of this object is $\mathbf{x} = [6, 8, 2, 7, -3, 5]$. What is the probability, according to the model, that this instance belongs to the positive class (i.e y=1)?

Solution:
The probability that this instance belongs to the positive class (y=1) is

$$P(y=1|\mathbf{x}) = \frac{1}{1+\exp(-\mathbf{w}^T\mathbf{x})}$$

To calculate this we first need to calculate the weighted sum of the inputs, which is

$$\begin{aligned}
\mathbf{w}^T\mathbf{x} &= (4\times 6) + (-2\times 8) + (5\times 2) + (-3\times 7) + (11\times -3) + (9\times 5) && (13)\\
&= 24 - 16 + 10 - 21 - 33 + 45 && (14)\\
&= 9. && (15)
\end{aligned}$$

We can then substitute this value into our probability function

$$\begin{aligned}
P(y=1|\mathbf{x}) &= \frac{1}{1+\exp(-9)} && (16)\\
&= 0.99988 && (17)
\end{aligned}$$

to 5 significant figures.

---

3. (***) Consider flipping a coin 20 times and recording each result $y_i = \{0, 1\}$. Using the log likelihood $l(\pi; \mathbf{y})$ derive the maximum likelihood estimation (MLE) for π by finding the derivative w.r.t π and setting this equal to zero ($\partial l(\pi; \mathbf{y})/\partial \pi = 0$). Is the result what you would have expected?

Solution:
While not strictly asked in the question, this first part is given to help build understanding. Note, usually capital letters are used to represent to outcomes of random events. The probability mass function (i.e probability of each possible output) of the coin flips is given by the Bernoulli distribution, which is defined as

$$P(y_i; \pi) = \pi^{y_i}(1-\pi)^{(1-y_i)}, \tag{18}$$

with y_i the possible outcome and π the probability that the outcome is a 1.

The likelihood is the probability of observing each of the these random variables (RVs), which since they are all independent is the product of each of the independent probabilities:

$$\begin{aligned} L(\pi; \mathbf{y}) &= \prod_{i=1}^{N} P(y_i; \pi) &(19) \\ &= \prod_{i=1}^{N} \pi^{y_i}(1-\pi)^{(1-y_i)} &(20) \end{aligned}$$

where $N = 20$ is the number of trials. $\mathbf{y}$ has been used to indicate that L is a function of all the outcomes measure, i.e $y_1, \ldots, y_N$. The log likelihood is then the logarithm of this equation:

$$\begin{aligned} l(\pi; \mathbf{y}) &= \log\left(\prod_{i=1}^{N} \pi^{y_i}(1-\pi)^{(1-y_i)}\right) &(21) \\ &= \sum_{i=1}^{N} \log\left(\pi^{y_i}(1-\pi)^{(1-y_i)}\right) &(22) \\ &= \sum_{i=1}^{N} \left(\log(\pi^{y_i}) + \log((1-\pi)^{(1-y_i)})\right) &(23) \\ &= \sum_{i=1}^{N} \left(y_i \log(\pi) + (1-y_i)\log(1-\pi)\right). &(24) \end{aligned}$$

Note: since we are trying to find where the derivative is equal to zero, then either the log likelihood or the *negative* log likelihood will work as the -1 factor will multiply out. The derivative of the log-likelihood is

$$\frac{\partial l}{\partial \pi} = \sum_{i=1}^{N} \frac{\partial}{\partial \pi} \left[y_i \log(\pi) + (1 - y_i) \log(1 - \pi) \right] \tag{25}$$

$$= \sum_{i=1}^{N} y_i \frac{\partial \log(\pi)}{\partial \pi} + (1 - y_i) \frac{\partial \log(1 - \pi)}{\partial \pi} \tag{26}$$

$$= \sum_{i=1}^{N} \frac{y_i}{\pi} - \frac{1 - y_i}{1 - \pi} \tag{27}$$

$$= \sum_{i=1}^{N} \frac{y_i(1 - \pi) - (1 - y_i)\pi}{\pi(1 - \pi)} \tag{28}$$

$$= \sum_{i=1}^{N} \frac{y_i - \pi}{\pi(1 - \pi)} \tag{29}$$

Now, to try and find the MLE for this distribution we need to find an expression for π when this derivative is zero. So

$$\frac{\partial l}{\partial \pi} = 0 = \sum_{i=1}^{N} \frac{y_i - \pi}{\pi(1 - \pi)}. \tag{30}$$

Since π is a constant and independent of the index i, we can multiply through by the denominator and since we have 0 on the LHS it disappears. This leaves

$$0 = \sum_{i=1}^{N} (y_i - \pi) \tag{31}$$

$$= (\sum_{i=1}^{N} y_i) - (\pi \sum_{i=1}^{N} 1) \tag{32}$$

$$= (\sum_{i=1}^{N} y_i) - \pi N \tag{33}$$

$$\pi N = \sum_{i=1}^{N} y_i \tag{34}$$

$$\pi = \frac{1}{N} \sum_{i=1}^{N} y_i. \tag{35}$$

This is telling us that when using the Bernoulli distribution, the probability can be estimated by using the average of all the outcomes. Hopefully, this is what we expected but note we cannot apply this to cases where the probability is a function of some input features as in logistic regression.

---

2 Automatic differentiation

Let $\mathbf{f}$ be a vector-valued function that maps from $\mathbb{R}^3$ to $\mathbb{R}^2$

$$y_1 = f_1(x_1, x_2, x_3) = x_1 x_3 + \log(x_2 + x_1) \times \exp(-x_3), \tag{36}$$
$$y_2 = f_2(x_1, x_2, x_3) = \exp(-x_2) + \cos(x_1 x_3). \tag{37}$$

1. (*) Compute the Jacobian using manual differentiation and evaluate the Jacobian at the point $(x_1 = 3, x_2 = 5, x_3 = 1)$

Solution:
The Jacobian $\mathbf{J}$ has dimensions 2 by 3 and is given by

$$\mathbf{J} = \begin{bmatrix} \frac{\partial f_1}{\partial x_1}, \frac{\partial f_1}{\partial x_2}, \frac{\partial f_1}{\partial x_3} \\ \frac{\partial f_2}{\partial x_1}, \frac{\partial f_2}{\partial x_2}, \frac{\partial f_2}{\partial x_3} \end{bmatrix} \tag{38}$$

where each of these derivatives in the first row are

$$\frac{\partial f_1}{\partial x_1} = x_3 + \frac{\exp(-x_3)}{x_2 + x_1} \tag{39}$$
$$\frac{\partial f_1}{\partial x_2} = \frac{\exp(-x_3)}{x_2 + x_1} \tag{40}$$
$$\frac{\partial f_1}{\partial x_3} = x_1 - \log(x_2 + x_1)\exp(-x_3) \tag{41}$$

and in the second row are

$$\frac{\partial f_2}{\partial x_1} = -x_3 \sin(x_1 x_3) \tag{42}$$
$$\frac{\partial f_2}{\partial x_2} = -\exp(-x_2) \tag{43}$$
$$\frac{\partial f_2}{\partial x_3} = -x_1 \sin(x_1 x_3). \tag{44}$$

Thus, the Jacobian at the point $(x_1 = 3, x_2 = 5, x_3 = 1)$ is

$$\mathbf{J}(x_1 = 3, x_2 = 5, x_3 = 1) = \begin{bmatrix} 1.0459849301 & 0.0459849301 & 2.2350162077 \\ -0.1411200081 & -0.006737947 & -0.4233600242 \end{bmatrix}. \tag{45}$$

```python
import numpy as np

def f1(x1, x2, x3):
  return x1*x3 + np.log(x2+x1)*np.exp(-x3)

def f2(x1, x2, x3):
  return np.exp(-x2) + np.cos(x1*x3)

x1_0 = 3
x2_0 = 5
x3_0 = 1
epsilon = 1e-6

df1dx1_numerical = (f1(x1_0+epsilon, x2_0, x3_0) - f1(x1_0, x2_0, x3_0))/epsilon
df1dx2_numerical = (f1(x1_0, x2_0+epsilon, x3_0) - f1(x1_0, x2_0, x3_0))/epsilon
df1dx3_numerical = (f1(x1_0, x2_0, x3_0+epsilon) - f1(x1_0, x2_0, x3_0))/epsilon

df2dx1_numerical = (f2(x1_0+epsilon, x2_0, x3_0) - f2(x1_0, x2_0, x3_0))/epsilon
df2dx2_numerical = (f2(x1_0, x2_0+epsilon, x3_0) - f2(x1_0, x2_0, x3_0))/epsilon
df2dx3_numerical = (f2(x1_0, x2_0, x3_0+epsilon) - f2(x1_0, x2_0, x3_0))/epsilon
```

2. (*) Compute the Jacobian at the same point that in the previous point, but using finite difference approximation.

Solution:
We can compute the finite difference approximations for the partial derivatives using the Python code shown in following snippet: The Jacobian computed using finite differences is approximated as

$$\mathbf{J}(x_1 = 3, x_2 = 5, x_3 = 1) \approx \begin{bmatrix} 1.0459849276 & 0.0459849274 & 2.2350165896 \\ -0.1411195131 & -0.0067379436 & -0.4233555692 \end{bmatrix}. \tag{46}$$

---

3. (*) Draw the computational graph.

Solution:
The computational graph is shown in figure 1 below:

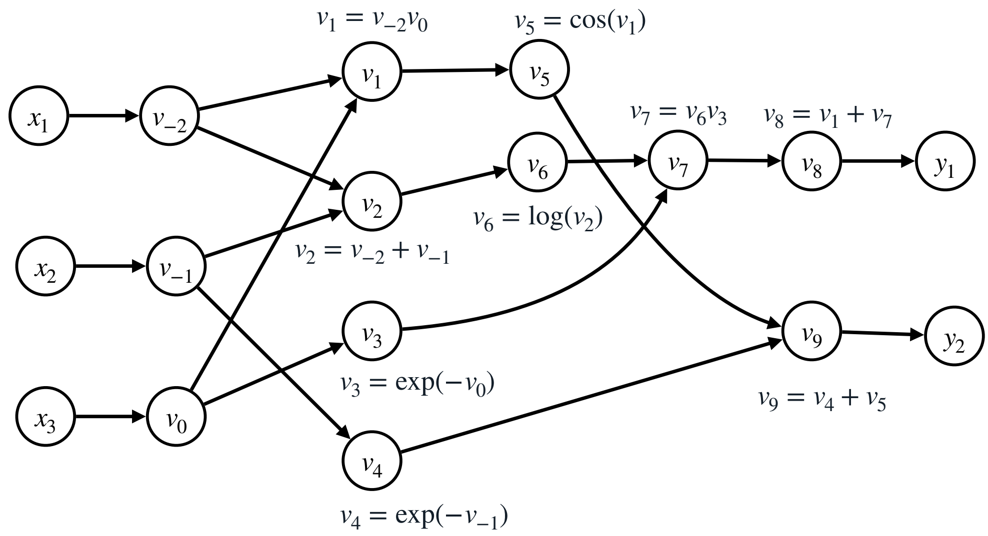

Figure 1: Computational graph for the vector-valued function.

---

4. (**) Compute the Jacobian using AD in forward mode. Write the expressions for all the intermediate variables $\dot{v}_i$ in the forward tangent trace.

Solution:
In forward mode AD we need to compute the derivative of each node with respect to the inputs $\dot{v}_i = \partial v_i / \partial x_j$ (where $j = 0, 1, 2$ to represent any of the inputs). The forward tangent trace is summarise in table 1 but here I will explain the derivation for a couple of the intermediate variables. For v_1 we note that it depends on both v_{-2} and v_0 so when we apply the chain rule we need to consider both these dependencies:

$$\dot{v}_1 = \frac{\partial v_1}{\partial x_j} = \frac{\partial v_1}{\partial v_{-2}} \frac{\partial v_{-2}}{\partial x_j} + \frac{\partial v_1}{\partial v_0} \frac{\partial v_0}{\partial x_j} \tag{47}$$

Since $\dot{v}_{-2} = \partial v_{-2} / \partial x_j$ and $\dot{v}_0 = \partial v_0 / \partial x_j$ we can substitute these into the expression and take the derivative of $v_1 = v_{-2} v_0$ with respect to v_{-2} and v_0 respectively.

This gives

$$\dot{v}_1 = v_0\dot{v}_{-2} + v_{-2}\dot{v}_0. \tag{48}$$

For v_2, the process is the same. It depends on v_{-2} and v_{-1} so we must include both terms when applying the chain rule

$$\dot{v}_2 = \frac{\partial v_2}{\partial x_j} = \frac{\partial v_2}{\partial v_{-2}}\frac{\partial v_{-2}}{\partial x_j} + \frac{\partial v_2}{\partial v_{-1}}\frac{\partial v_{-1}}{\partial x_j} \tag{49}$$

$$= \dot{v}_{-2} + \dot{v}_{-1}. \tag{50}$$

Forward primal trace	Forward tangent trace
$v_{-2} = x_1$	$\dot{v}_{-2} = \dot{x}_1$
$v_{-1} = x_1$	$\dot{v}_{-1} = \dot{x}_2$
$v_0 = x_1$	$\dot{v}_0 = \dot{x}_3$
$v_1 = v_{-2}v_0$	$\dot{v}_1 = v_0\dot{v}_{-2} + v_{-2}\dot{v}_0$
$v_2 = v_{-2} + v_{-1}$	$\dot{v}_2 = \dot{v}_{-2} + \dot{v}_{-1}$
$v_3 = \exp(-v_0)$	$\dot{v}_3 = -\exp(-v_0)\dot{v}_0 = -v_3\dot{v}_0$
$v_4 = \exp(-v_{-1})$	$\dot{v}_4 = -\exp(-v_{-1})\dot{v}_{-1} = -v_4\dot{v}_{-1}$
$v_5 = \cos(v_1)$	$\dot{v}_5 = -\sin(v_1)\dot{v}_1$
$v_6 = \log(v_2)$	$\dot{v}_6 = \dot{v}_2/v_2$
$v_7 = v_3v_6$	$\dot{v}_7 = v_6\dot{v}_3 + \dot{v}_6v_3$
$v_8 = v_1 + v_7$	$\dot{v}_8 = \dot{v}_1 + \dot{v}_7$
$v_9 = v_4 + v_5$	$\dot{v}_8 = \dot{v}_4 + \dot{v}_5$
$y_1 = v_8$	$\dot{y}_1 = \dot{v}_8$
$y_2 = v_9$	$\dot{y}_2 = \dot{v}_9$

Table 1: Forward primal and tangent trace.

Using these expressions for the tangents, we can now compute the Jacobian using forward mode. Table 2 shows the values when computing the first column of the Jacobian, dy_1/dx_1 and dy_2/dx_1. Recall that forward mode calculates each column with a single forward pass, to select different columns the initial values for $\dot{x}_1$,$\dot{x}_2$ and $\dot{x}_3$ are set 1 or 0 depending which derivative is being sought. Values are only given here for x_1 but the process is the same for the other columns.

Forward primal trace	Value	Forward tangent trace	Value
$v_{-2} = x_1$	$= 3$	$\dot{v}_{-2} = \dot{x}_1$	$= 1$
$v_{-1} = x_1$	$= 5$	$\dot{v}_{-1} = \dot{x}_2$	$= 0$
$v_0 = x_1$	$= 1$	$\dot{v}_0 = \dot{x}_3$	$= 0$
$v_1 = v_{-2}v_0$	$= 3$	$\dot{v}_1 = v_0\dot{v}_{-2} + v_{-2}\dot{v}_0$	$= 1$
$v_2 = v_{-2} + v_{-1}$	$= 8$	$\dot{v}_2 = \dot{v}_{-2} + \dot{v}_{-1}$	=1
$v_3 = \exp(-v_0)$	= 0.367	$\dot{v}_3 = -\exp(-v_0)\dot{v}_0 = -v_3\dot{v}_0$	$= 0$
$v_4 = \exp(-v_{-1})$	= 0.006	$\dot{v}_4 = -\exp(-v_{-1})\dot{v}_{-1}$	= -0.1411
$v_5 = \cos(v_1)$	= -0.989	$\dot{v}_5 = -\sin(v_1)\dot{v}_1$	$= 0$
$v_6 = \log(v_2)$	= 2.079	$\dot{v}_6 = \dot{v}_2/v_2$	= 0.125
$v_7 = v_3v_6$	= 0.765	$\dot{v}_7 = v_6\dot{v}_3 + \dot{v}_6v_3$	= 0.0459
$v_8 = v_1 + v_7$	= 3.765	$\dot{v}_8 = \dot{v}_1 + \dot{v}_7$	= 1.0459
$v_9 = v_4 + v_5$	= -0.983	$\dot{v}_8 = \dot{v}_4 + \dot{v}_5$	= -0.1411
$y_1 = v_7$	= 3.765	$\dot{y}_1 = \dot{v}_7$	= 1.0459
$y_2 = v_8$	= -0.983	$\dot{y}_2 = \dot{v}_8$	= -0.1411

Table 2: Value and derivatives for $j = 1$ giving dy_1/dx_1 and dy_2/dx_1 when $x_1 = 3, x_2 = 5, x_3 = 1$.

5. (**) Compute the Jacobian using AD in reverse mode. Write the expressions for all the adjoints $\bar{v}_i$ in the reverse derivative trace.

Solution:

To compute the Jacobian using reverse mode, we need to compute the adjoints which are defined as

$$\bar{v}_i = \frac{\partial y_j}{\partial v_i}. \tag{51}$$

To calculate these adjoints we need to keep track of how the output depends on the variable of interest. For the output variables, the adjoints are equal to the derivative of the outputs, which like in the forward mode, will be $= 1$ for the output for which we are finding the derivative for and $= 0$ otherwise.

$$\bar{v}_9 = \frac{\partial y_j}{\partial v_9} = \bar{y}_2 \tag{52}$$

$$\bar{v}_8 = \frac{\partial y_j}{\partial v_8} = \bar{y}_1 \tag{53}$$

$$\tag{54}$$

Now we can move backward through the network, apply the chain rule to make the derivative dependent on the other adjoints. For the v_7 to v_2 each variable has only 1 other variable that depends on it so there is only 1 term when we apply the chain rule:

$$\bar{v}_7 = \frac{\partial y_j}{\partial v_7} = \frac{\partial y_j}{\partial v_8}\frac{\partial v_8}{\partial v_7} = \bar{v}_8 \frac{\partial}{\partial v_7}(v_1 + v_7) = \bar{v}_8 \tag{55}$$

$$\bar{v}_6 = \frac{\partial y_j}{\partial v_6} = \frac{\partial y_j}{\partial v_7}\frac{\partial v_7}{\partial v_6} = \bar{v}_7 \frac{\partial}{\partial v_6}(v_6 v_3) = \bar{v}_7 v_3 \tag{56}$$

$$\bar{v}_5 = \frac{\partial y_j}{\partial v_5} = \frac{\partial y_j}{\partial v_9}\frac{\partial v_9}{\partial v_5} = \bar{v}_9 \tag{57}$$

$$\bar{v}_4 = \frac{\partial y_j}{\partial v_4} = \frac{\partial y_j}{\partial v_9}\frac{\partial v_9}{\partial v_4} = \bar{v}_9 \tag{58}$$

$$\bar{v}_3 = \frac{\partial y_j}{\partial v_3} = \frac{\partial y_j}{\partial v_7}\frac{\partial v_7}{\partial v_3} = \bar{v}_7 v_6 \tag{59}$$

$$\bar{v}_2 = \frac{\partial y_j}{\partial v_2} = \frac{\partial y_j}{\partial v_6}\frac{\partial v_6}{\partial v_2} = \frac{\bar{v}_6}{v_2} \tag{60}$$

$$\tag{61}$$

For v_1, v_0, v_{-1} and v_{-2}, we can see from the computational graph that 2 variables in the middle of the network depend on these variables so when we apply the chain rule we need to consider each of these dependencies.

$$\bar{v}_1 = \frac{\partial y_j}{\partial v_1} = \frac{\partial y_j}{\partial v_5}\frac{\partial v_5}{\partial v_1} + \frac{\partial y_j}{\partial v_8}\frac{\partial v_8}{\partial v_1} \tag{62}$$

$$= -\sin(v_1)\bar{v}_5 + \bar{v}_8 \tag{63}$$

$$\bar{v}_0 = \frac{\partial y_j}{\partial v_0} = \frac{\partial y_j}{\partial v_1}\frac{\partial v_1}{\partial v_0} + \frac{\partial y_j}{\partial v_3}\frac{\partial v_3}{\partial v_0} \tag{64}$$

$$= \bar{v}_1 v_{-2} - \exp(-v_0)\bar{v}_3 \tag{65}$$

$$\bar{v}_{-1} = \frac{\partial y_j}{\partial v_{-1}} = \frac{\partial y_j}{\partial v_2}\frac{\partial v_2}{\partial v_{-1}} + \frac{\partial y_j}{\partial v_4}\frac{\partial v_4}{\partial v_{-1}} \tag{66}$$

$$= \bar{v}_2 - \exp(-v_{-1})\bar{v}_4 \tag{67}$$

$$\bar{v}_{-2} = \frac{\partial y_j}{\partial v_{-2}} = \frac{\partial y_j}{\partial v_1}\frac{\partial v_1}{\partial v_{-2}} + \frac{\partial y_j}{\partial v_2}\frac{\partial v_2}{\partial v_{-2}} \tag{68}$$

$$= \bar{v}_1 v_0 + \bar{v}_2 \tag{69}$$

Now if we wish to calculate the value of the derivatives, in reverse mode we get a row in the Jacobian, i.e $(\partial y_1/\partial x_1, \partial y_1/\partial x_2, \partial y_1/\partial x_3)$. Let's calculate these values now, we can use the same primal trace values from the previous question. When calculating the adjoints, we set $\bar{y}_1 = 1$ and $\bar{y}_2 = 0$ (so $j = 1$). The adjoints are therefore:

$$\bar{v}_9 = \bar{y}_2 = 0 \tag{70}$$
$$\bar{v}_8 = \bar{y}_1 = 1 \tag{71}$$
$$\bar{v}_7 = \bar{v}_8 = 1 \tag{72}$$
$$\bar{v}_6 = \bar{v}_7 v_3 = 0.367 \tag{73}$$
$$\bar{v}_5 = \bar{v}_9 = 0 \tag{74}$$
$$\bar{v}_4 = \bar{v}_9 = 0 \tag{75}$$
$$\bar{v}_3 = \bar{v}_7 v_6 = 2.079 \tag{76}$$
$$\bar{v}_2 = \bar{v}_6 \frac{1}{v_2} = 0.0459 \tag{77}$$
$$\bar{v}_1 = -\sin(v_1)\bar{v}_5 + \bar{v}_8 = 1 \tag{78}$$
$$\bar{v}_0 = \bar{v}_1 v_{-2} - \exp(-v_0)\bar{v}_3 = 2.235 \tag{79}$$
$$\bar{v}_{-1} = \bar{v}_2 - \exp(-v_{-1})\bar{v}_4 = 0.0459 \tag{80}$$
$$\bar{v}_{-2} = \bar{v}_1 v_0 + \bar{v}_2 = 1.0459 \tag{81}$$

These final 3 values are the top row of the Jacobian (you can compare these values to your manual and numerical solutions. Some precision might have been lost in the process. If you results are different it may be down to which log function you are using, here we use the natural logarithm.