CSE 5522 HOMEWORK 3

BRENDAN WHITAKER

1. *Regularized linear regression.*
   (a) *Regularized linear regression is the solution to:*

$$\min_{\mathbf{w}} Err(\mathbf{w}) = \frac{1}{N}||Y - X\mathbf{w}||^2 + \frac{\lambda}{2}||\mathbf{w}||^2. \tag{1}$$

What is the gradient $\nabla_w Err(w)$?
Behold:

$$\begin{aligned}
\nabla_w Err(\mathbf{w}) &= \nabla_{\mathbf{w}} \left[\frac{1}{N}||Y - X\mathbf{w}||^2 + \frac{\lambda}{2}||\mathbf{w}||^2 \right] \\
&= \nabla_{\mathbf{w}} \left[\frac{1}{N}(Y - X\mathbf{w})^T(Y - X\mathbf{w}) + \frac{\lambda}{2}\mathbf{w}^T\mathbf{w} \right] \\
&= \nabla_{\mathbf{w}} \left[\frac{1}{N}\left(Y^TY - 2X^TY\mathbf{w}^T + X^TX\mathbf{w}^T\mathbf{w}\right) + \frac{\lambda}{2}\mathbf{w}^T\mathbf{w} \right] \\
&= \frac{1}{N}\left(-2X^TY + 2X^TX\mathbf{w}\right) + \lambda\mathbf{w}.
\end{aligned} \tag{2}$$

   (b) *What is the closed-form solution of* $\min_{\mathbf{w}} Err(\mathbf{w})$*?*
In order to minimize, we set the gradient equal to zero and solve for $\mathbf{w}$. Observe:

$$\begin{aligned}
\frac{1}{N}\left(-2X^TY + 2X^TX\mathbf{w}\right) + \lambda\mathbf{w} &= 0 \\
-2X^TY + 2X^TX\mathbf{w} + N\lambda\mathbf{w} &= 0 \\
2X^TX\mathbf{w} + N\lambda\mathbf{w} &= 2X^TY \\
(2X^TX + N\lambda I)\mathbf{w} &= 2X^TY \\
\mathbf{w} &= \frac{2X^TY}{2X^TX + N\lambda I}.
\end{aligned} \tag{3}$$

2. *Logistic regression.*
   (a) *Let L be the logistic regression loss function* $L(y, f(x)) = \log\left(1 + e^{-y\mathbf{w}^Tx}\right)$. *Compute the gradient $\nabla_{\mathbf{w}} L$ step-by-step to show that:*

$$\nabla_{\mathbf{w}} L = \frac{-yxe^{-y\mathbf{w}^Tx}}{1 + e^{-y\mathbf{w}^Tx}}. \tag{4}$$

Date: SP18.

Proof. We compute the gradient:

$$\begin{aligned}\nabla_{\mathbf{w}} L(y, f(x)) &= \nabla_{\mathbf{w}} \log\left(1 + e^{-y\mathbf{w}^T x}\right) \\ &= \frac{\nabla_{\mathbf{w}}\left(1 + e^{-y\mathbf{w}^T x}\right)}{1 + e^{-y\mathbf{w}^T x}} \\ &= \frac{e^{-y\mathbf{w}^T x}\nabla_{\mathbf{w}}\left(-y\mathbf{w}^T x\right)}{1 + e^{-y\mathbf{w}^T x}} \\ &= \frac{-yxe^{-y\mathbf{w}^T x}}{1 + e^{-y\mathbf{w}^T x}}.\end{aligned} \tag{5}$$

□

(b) *Let L be the softmax loss function $L(y, f(x)) = -\mathbf{w}_y^T x + \log\left(\sum_k e^{\mathbf{w}_k^T x}\right)$. Compute the gradient $\nabla_{\mathbf{w}_k} L$ step-by-step to show that:*

$$\nabla_{\mathbf{w}_k} L = x \cdot \left(-\mathbb{1}_{y=k} + \frac{e^{\mathbf{w}_k^T x}}{\sum_j e^{\mathbf{w}_j^T x}}\right). \tag{6}$$

Proof. Observe:

$$\nabla_{\mathbf{w}_k} L(y, f(x)) = \nabla_{\mathbf{w}_k}\left[-\mathbf{w}_y^T x + \log\left(\sum_j e^{\mathbf{w}_j^T x}\right)\right]. \tag{7}$$

Now note that $\nabla_{\mathbf{w}_k}\mathbf{w}_y^T = 0$ if $k \neq y$ since in this case we are not taking the gradient with respect to $\mathbf{w}_y$, so $\mathbf{w}_y$ is treated like a constant. Thus we have:

$$\begin{aligned}\nabla_{\mathbf{w}_k} L(y, f(x)) &= -x\mathbb{1}_{y=k} + \nabla_{\mathbf{w}_k} \log\left(\sum_j e^{\mathbf{w}_j^T x}\right) \\ &= -x\mathbb{1}_{y=k} + \frac{\nabla_{\mathbf{w}_k}\sum_j e^{\mathbf{w}_j^T x}}{\sum_j e^{\mathbf{w}_j^T x}}.\end{aligned} \tag{8}$$

Now here we note that when taking the gradient of the sum $\sum_j e^{\mathbf{w}_j^T x}$, since we are taking the gradient with respect to $\mathbf{w}_k$, when $j \neq k$, we treat the term $e^{\mathbf{w}_j^T x}$ as a constant, hence its gradient is zero. And thus all the terms go to zero when we take the gradient except for the term where $j = k$, hence we have:

$$\begin{aligned}\nabla_{\mathbf{w}_k} L(y, f(x)) &= -x\mathbb{1}_{y=k} + \frac{\nabla_{\mathbf{w}_k} e^{\mathbf{w}_k^T x}}{\sum_j e^{\mathbf{w}_j^T x}} \\ &= -x\mathbb{1}_{y=k} + \frac{xe^{\mathbf{w}_k^T x}}{\sum_j e^{\mathbf{w}_j^T x}} \\ &= x \cdot \left(-\mathbb{1}_{y=k} + \frac{e^{\mathbf{w}_k^T x}}{\sum_j e^{\mathbf{w}_j^T x}}\right).\end{aligned} \tag{9}$$

□

3. *Hard-margin Support Vector Machine: to make sure you understand the derivation of SVM in dual form, you will repeat the derivation from the slides yourself.*
   (a) *Define the Lagrangian and the dual variables.*
   The Lagrangian is given by:

$$\begin{aligned} L(\mathbf{w}, b, \mathbf{a}) &= \frac{1}{2}||\mathbf{w}||^2 - \sum_p a_p(y_p(\mathbf{w}^T\mathbf{x}_p + b) - 1) \\ &= \frac{1}{2}\mathbf{w}^T\mathbf{w} - \sum_p a_p(y_p\mathbf{w}^T\mathbf{x}_p + y_p b - 1) \\ &= \frac{1}{2}\mathbf{w}^T\mathbf{w} - \sum_p \left(a_p y_p \mathbf{w}^T\mathbf{x}_p + a_p y_p b - a_p\right). \end{aligned} \tag{10}$$

   And the dual variables are $\mathbf{a}$.
   (b) *Find the dual function.* So we find the dual function ($= \inf_{\mathbf{w},b} L(\mathbf{w}, b, \mathbf{a})$) by taking derivatives of the Lagrangian and setting them equal to zero:

$$\begin{aligned} \frac{\partial}{\partial \mathbf{w}} L &= \mathbf{w} - \sum_p a_p y_p \mathbf{x}_p = 0 \\ \mathbf{w} &= \sum_p a_p y_p \mathbf{x}_p. \\ \frac{\partial}{\partial b} L &= \sum_p a_p y_p = 0. \end{aligned} \tag{11}$$

   Now we plug these expressions back into the Lagrangian:

$$\begin{aligned} L(\mathbf{w}, b, \mathbf{a}) &= \frac{1}{2}\mathbf{w}^T\mathbf{w} - \sum_p \left(a_p y_p \mathbf{w}^T\mathbf{x}_p + a_p y_p b - a_p\right) \\ &= \frac{1}{2}\mathbf{w}^T\mathbf{w} - \sum_p a_p y_p \mathbf{w}^T\mathbf{x}_p - \sum_p a_p y_p b + \sum_p a_p. \end{aligned} \tag{12}$$

   Plugging in the expression we derived for $\mathbf{w}$, we have:

$$\mathbf{w}^T\mathbf{w} = \mathbf{w}^T \sum_p a_p y_p \mathbf{x}_p = \sum_p a_p y_p \mathbf{w}^T\mathbf{x}_p = \sum_p \sum_q a_p a_q y_p y_q \mathbf{x}_p^T \mathbf{x}_q. \tag{13}$$

   And then substituting this expression in the first two terms of the Lagrangian, we have:

$$L(\mathbf{w}, b, \mathbf{a}) = \left(\frac{1}{2} - 1\right) \sum_p \sum_q a_p a_q y_p y_q \mathbf{x}_p^T \mathbf{x}_q - b \sum_p a_p y_p + \sum_p a_p. \tag{14}$$

   Then using the other equality we derived after taking the derivative with respect to b, the third term is annihilated:

$$\begin{aligned} L(\mathbf{w}, b, \mathbf{a}) &= \left(\frac{1}{2} - 1\right) \sum_p \sum_q a_p a_q y_p y_q \mathbf{x}_p^T \mathbf{x}_q + \sum_p a_p \\ &= \frac{1}{2} \sum_p \sum_q a_p a_q y_p y_q \mathbf{x}_p^T \mathbf{x}_q + \sum_p a_p. \end{aligned} \tag{15}$$

And this is exactly our dual function $\tilde{L}(\mathbf{a}) = \frac{1}{2}\sum_p \sum_q a_p a_q y_p y_q \mathbf{x}_p^T \mathbf{x}_q + \sum_p a_p$.

(c) *Write the dual problem.*
Thus the dual form of the problem is to find $\max_{\mathbf{a}} \tilde{L}(\mathbf{a})$ such that $a_p \geq 0\ \forall p$ and $\sum_p a_p y_p = 0$.

4. *Soft-margin Support Vector Machine: derive the soft-margin SVM in dual from similarly to the hard-margin SVM.*
   (a) *Define the Lagrangian and the dual variables.* Define slack variables $\xi_1, ..., \xi_N$ which are all nonegative. Note the primal problem is to solve:

$$\min_{\mathbf{w}, b, \{\, \xi_p \geq 0 \,\}} \frac{1}{N}\sum_p \xi_p + \frac{\lambda}{2}||\mathbf{w}||^2 \tag{16}$$

subject to $y_p(\mathbf{w}^T\mathbf{x}_p + b) \geq 1 - \xi_p$, $p = 1, ..., N$. We first define the Lagrangian multipliers $\mathbf{a} = (a_1, ..., a_N)$ where $a_p \geq 0\ \forall p$. We also define a second set of dual variables $\mathbf{c} = (c_1, ..., c_N)$ which are nonnegative. Then our Lagrangian is:

$$\begin{aligned}
L(\mathbf{w}, b, \{\, \xi_p \,\}, \mathbf{a}, \mathbf{c}) &= \frac{1}{N}\sum_p \xi_p + \frac{\lambda}{2}||\mathbf{w}||^2 - \sum_p a_p \left(y_p(\mathbf{w}^T\mathbf{x}_p + b) - 1 + \xi_p\right) - \sum_p c_p \xi_p \\
&= \frac{1}{N}\sum_p \xi_p + \frac{\lambda}{2}\mathbf{w}^T\mathbf{w} - \sum_p a_p(y_p\mathbf{w}^T\mathbf{x}_p + y_p b - 1 + \xi_p) - \sum_p c_p \xi_p \\
&= \frac{1}{N}\sum_p \xi_p + \frac{\lambda}{2}\mathbf{w}^T\mathbf{w} - \sum_p \left(a_p y_p \mathbf{w}^T\mathbf{x}_p + a_p y_p b - a_p + a_p \xi_p\right) - \sum_p c_p \xi_p.
\end{aligned} \tag{17}$$

And the dual variables will be $\mathbf{a}, \mathbf{c}$.

   (b) *Derive the dual function.* We take derivatives with respect to $\mathbf{w}, b, \xi$ and set equal to zero to derive relations:

$$\begin{aligned}
\frac{\partial}{\partial \mathbf{w}} L &= \lambda\mathbf{w} - \sum_p a_p y_p \mathbf{x}_p = 0 \\
\mathbf{w} &= \frac{1}{\lambda}\sum_p a_p y_p \mathbf{x}_p. \\
\frac{\partial}{\partial b} L &= -\sum_p a_p y_p = 0. \\
\frac{\partial}{\partial \xi_p} L &= \frac{1}{N} - a_p - c_p = 0 \\
\frac{1}{N} &= a_p + c_p \\
c_p &= \frac{1}{N} - a_p.
\end{aligned} \tag{18}$$

Note since we added the constraint that $c_p \geq 0 \; \forall p$, we know $\frac{1}{N} - a_p \geq 0 \; \forall p$. Note again we have:

$$\begin{aligned} \mathbf{w}^T\mathbf{w} &= \frac{1}{\lambda} \sum_p a_p y_p \mathbf{w}^T \mathbf{x}_p \\ &= \frac{1}{\lambda^2} \sum_p \sum_q a_p a_q y_p y_q \mathbf{x}_p^T \mathbf{x}_q. \end{aligned} \tag{19}$$

Plugging these relations into our Lagrangian, we have:

$$\begin{aligned} L(\mathbf{w}, b, \{ \xi_p \}, \mathbf{a}, \mathbf{c}) &= \frac{1}{2\lambda} \sum_p \sum_q a_p a_q y_p y_q \mathbf{x}_p^T \mathbf{x}_q - \frac{1}{\lambda} \sum_p \sum_q a_p a_q y_p y_q \mathbf{x}_p^T \mathbf{x}_q \\ &- b \sum_p a_p y_p + \sum_p a_p + \sum_p \frac{1}{N} \xi_p - \sum_p a_p \xi_p - \sum_p c_p \xi_p \\ &= -\frac{1}{2\lambda} \sum_p \sum_q a_p a_q y_p y_q \mathbf{x}_p^T \mathbf{x}_q - b \cdot 0 + \sum_p a_p \\ &+ \sum_p \xi_p \left(\frac{1}{N} - a_p - c_p \right) \\ &= -\frac{1}{2\lambda} \sum_p \sum_q a_p a_q y_p y_q \mathbf{x}_p^T \mathbf{x}_q + \sum_p a_p. \end{aligned} \tag{20}$$

Thus the dual function is $\tilde{L}(\mathbf{a}) = -\frac{1}{2\lambda} \sum_p \sum_q a_p a_q y_p y_q \mathbf{x}_p^T \mathbf{x}_q + \sum_p a_p$.

(c) *Write the dual problem.*
The dual problem is to find $\max_{\mathbf{a}} \tilde{L}(\mathbf{a})$ subject to the constraints $a_p \geq 0$ for all p, $\sum_p a_p y_p = 0$, and $\frac{1}{N} - a_p \geq 0 \; \forall p$.

5. *Soft-margin linear SVM in unconstrained form.*
(a) *Compute the gradient $\nabla_{\mathbf{w}} f(\mathbf{w})$ of the (regularized) empirical risk:*

$$f(\mathbf{w}) = \frac{1}{N} \sum_p \max \left\{ 0, 1 - y_p \left(\mathbf{w}^T \mathbf{x}_p \right) \right\} + \frac{\lambda}{2} ||\mathbf{w}||^2. \tag{21}$$

Case 1: $1 - y_p \left(\mathbf{w}^T \mathbf{x}_p \right) > 0$. Then we have:

$$\begin{aligned} \nabla_{\mathbf{w}} f(\mathbf{w}) &= \nabla_{\mathbf{w}} \left[\frac{1}{N} \sum_p \max \left\{ 0, 1 - y_p \left(\mathbf{w}^T \mathbf{x}_p \right) \right\} + \frac{\lambda}{2} ||\mathbf{w}||^2 \right] \\ &= \frac{1}{N} \sum_p \nabla_{\mathbf{w}} \max \left\{ 0, 1 - y_p \left(\mathbf{w}^T \mathbf{x}_p \right) \right\} + \nabla_{\mathbf{w}} \frac{\lambda}{2} ||\mathbf{w}||^2 \\ &= \frac{1}{N} \sum_p \nabla_{\mathbf{w}} \left(1 - y_p \left(\mathbf{w}^T \mathbf{x}_p \right) \right) + \lambda \mathbf{w} \\ &= -\frac{1}{N} \sum_p y_p \mathbf{x}_p + \lambda \mathbf{w}. \end{aligned} \tag{22}$$

Case 2: $1 - y_p\left(\mathbf{w}^T\mathbf{x}_p\right) < 0$. Then we have:

$$\begin{aligned}\nabla_{\mathbf{w}} f(\mathbf{w}) &= \nabla_{\mathbf{w}}\left[\frac{1}{N}\sum_p \max\left\{\,0, 1 - y_p\left(\mathbf{w}^T\mathbf{x}_p\right)\,\right\} + \frac{\lambda}{2}||\mathbf{w}||^2\right] \\ &= \frac{1}{N}\sum_p \nabla_{\mathbf{w}} \max\left\{\,0, 1 - y_p\left(\mathbf{w}^T\mathbf{x}_p\right)\,\right\} + \nabla_{\mathbf{w}}\frac{\lambda}{2}||\mathbf{w}||^2 \\ &= \frac{1}{N}\sum_p \nabla_{\mathbf{w}}(0) + \lambda\mathbf{w} \\ &= \lambda\mathbf{w}.\end{aligned} \tag{23}$$

Case 3: $1 - y_p\left(\mathbf{w}^T\mathbf{x}_p\right) = 0$. Then the gradient of the max is undefined, so the gradient of the entire function is undefined.
Let $\alpha = 1 - y_p\left(\mathbf{w}^T\mathbf{x}_p\right)$. So we have:

$$\nabla_{\mathbf{w}} f(\mathbf{w}) = \begin{cases} -\frac{1}{N}\sum_p y_p\mathbf{x}_p + \lambda\mathbf{w} & \text{if } \alpha > 0 \\ \lambda\mathbf{w} & \text{if } \alpha < 0 \\ \text{undefined} & \text{otherwise} \end{cases}. \tag{24}$$

6. *Neural Network.*
   (a) *Implement the logical OR with a single neuron, similar to the logical AND example from Lecture 10. In other words, find w_0, w_1, w_2.*
   Set $w_0 = 0.5, w_1 = 1, w_2 = 1$. We show this is a correct implementation of OR. Let $x_1 = 0, x_2 = 0$. Then $a = g(0.5(-1) + 1(0) + 1(0)) = g(-0.5) = 0$. Let $x_1 = 1, x_2 = 0$. Then $a = g(0.5(-1) + 1(1) + 1(0)) = g(0.5) = 1$. Let $x_1 = 0, x_2 = 1$. Then $a = g(0.5(-1)+1(0)+1(1)) = g(0.5) = 1$. Let $x_1 = x_2 = 1$. Then $a = g(0.5(-1) + 1(1) + 1(1)) = g(1.5) = 1$. So it is correct.
   (b) *Implement the logical XOR with three neurons from Lecture 10, using the weights for AND and OR from previous problems. What are the missing values in the three empty boxes.*
   The first box should be 0.5, the second should be -1, and the third 1. We let the weights for the first two $\wedge$ and $\vee$ neurons use weights $1.5, 1, 1$ and $0.5, 1, 1$ respectively, so they behave like OR and AND as proved in class and in the previous exercise. Then let $x_1 = 1, x_2 = 0$. Then the AND output is 0, and the OR output is 0. So $a = g(0.5(-1) + -1(0) + 1(0)) = g(-0.5) = 0$. Let $x_1 = 1, x_2 = 0$. Then the AND output is 0, the OR output is 1. So we have $a = g(0.5(-1) + -1(0) + 1(1)) = g(0.5) = 1$, and we have the same case when $x_1 = 0, x_2 = 1$. Now let $x_1 = x_2 = 1$. Then the AND output is 1 and the OR output is 1. So we have $a = g(0.5(-1) + -1(1) + 1(1)) = g(-0.5) = 0$. So the XOR implementation is correct.
7. *Multilayer Neural Network.*
   (a) *Forward propagation. What are the values of $a_1, a_2, ..., a_9$?*
   The first three are just inputs:

$$\begin{aligned} a_1 &= 1 \\ a_2 &= 0 \\ a_3 &= -1. \end{aligned} \tag{25}$$

We compute the hidden layer, where $g(t) = \max\{0, t\}$:

$$\begin{aligned} a_4 &= g\left(w_{14}a_1 + w_{24}a_2 + w_{34}a_3\right) \\ &= g\left(1(1) + 0 + -1(-1)\right) \\ &= g(2) = 2. \\ a_5 &= g\left(w_{15}a_1 + w_{25}a_2 + w_{35}a_3\right) \\ &= g\left(-1(1) + 0 + -1(-1)\right) \\ &= g(0) = 0. \\ a_6 &= g\left(w_{16}a_1 + w_{26}a_2 + w_{36}a_3\right) \\ &= g\left(0 + 0 + 1(-1)\right) \\ &= g(-1) = 0. \end{aligned} \tag{26}$$

We compute the output layer, where $g(t)$ is softmax:

$$\begin{aligned} n_7 &= w_{47}a_4 + w_{57}a_5 + w_{67}a_6 \\ &= 0 \\ n_8 &= w_{48}a_4 + w_{58}a_5 + w_{68}a_6 \\ &= -2 \\ n_9 &= w_{49}a_4 + w_{59}a_5 + w_{69}a_6 \\ &= 2. \\ a_7 &= \frac{e^{n_7}}{e^{n_7} + e^{n_8} + e^{n_9}} \\ &= \frac{1}{1 + e^{-2} + e^{2}} \\ &= 0.117. \\ a_8 &= \frac{e^{n_8}}{e^{n_7} + e^{n_8} + e^{n_9}} \\ &= \frac{e^{-2}}{1 + e^{-2} + e^{2}} \\ &= 0.016. \\ a_9 &= \frac{e^{n_9}}{e^{n_7} + e^{n_8} + e^{n_9}} \\ &= \frac{e^{2}}{1 + e^{-2} + e^{2}}. \\ &= 0.867. \end{aligned} \tag{27}$$

(b) *Backward propagation.*

We compute using $L = L_2 = L(y, f(x)) = ||y - f(x)||^2$:

$$\begin{aligned}
L &= (y_1 - a_7)^2 + (y_2 - a_8)^2 + (y_3 - a_9)^2 \\
\frac{\partial L}{\partial a_7} &= -2(y_1 - a_7) \\
&= -2(1 - 0.117) \\
&= -1.766 \\
\frac{\partial L}{\partial a_8} &= -2(y_2 - a_8) \\
&= -2(-1 - 0.016) \\
&= 2.032 \\
\frac{\partial L}{\partial a_9} &= -2(y_3 - a_9) \\
&= -2(1 - 0.867) \\
&= -0.266.
\end{aligned} \tag{28}$$

(c) *What is the value of* $\frac{\partial L}{\partial w_{47}}$*?*
Observe we have:

$$\begin{aligned}
\frac{\partial L}{\partial w_{47}} &= \frac{\partial L}{\partial a_7}\frac{\partial a_7}{\partial n_7}\frac{\partial n_7}{\partial w_{47}} \\
&= (-1.766)\frac{\frac{\partial}{\partial n_7}\left(e^{n_7}\right)\left(e^{n_7} + e^{n_8} + e^{n_9}\right) - \frac{\partial}{\partial n_7}\left(e^{n_7} + e^{n_8} + e^{n_9}\right)e^{n_7}}{\left(e^{n_7} + e^{n_8} + e^{n_9}\right)^2}a_4 \\
&= (-1.766\frac{e^{n_7}\left(e^{n_8} + e^{n_9}\right)}{\left(e^{n_7} + e^{n_8} + e^{n_9}\right)^2}a_4 \\
&= (-1.766\frac{e^0\left(e^{-2} + e^2\right)}{\left(e^0 + e^{-2} + e^2\right)^2}a_4 \\
&= (-1.766\frac{\left(e^{-2} + e^2\right)}{\left(1 + e^{-2} + e^2\right)^2}a_4 \\
&= (-1.766)(0.104) \cdot 2 \\
&= -0.368.
\end{aligned} \tag{29}$$

(d) *What is the value of* $\frac{\partial L}{\partial a_4}$*?*
Observe by the chain rule we have:

$$\begin{aligned}
\frac{\partial L}{\partial a_4} &= \frac{\partial L}{\partial a_7}\frac{\partial a_7}{\partial n_7}w_{47} + \frac{\partial L}{\partial a_8}\frac{\partial a_8}{\partial n_8}w_{48} + \frac{\partial L}{\partial a_9}\frac{\partial a_9}{\partial n_7}w_{49} \\
&= (-1.766)(0.104)w_{47} + (2.032)\frac{e^{n_8}\left(e^{n_7} + e^{n_9}\right)}{\left(e^{n_7} + e^{n_8} + e^{n_9}\right)^2}w_{48} + (-0.266)\frac{e^{n_9}\left(e^{n_8} + e^{n_7}\right)}{\left(e^{n_7} + e^{n_8} + e^{n_9}\right)^2}w_{49} \\
&= 0 + (2.032)(0.016)(-1) + (-0.266)(0.115) \\
&= -0.062.
\end{aligned} \tag{30}$$

(e) *Finish formula for* $\frac{\partial L}{\partial a_1}, \frac{\partial L}{\partial a_2}, \frac{\partial L}{\partial a_3}$.

Behold:

$$
\begin{aligned}
\frac{\partial L}{\partial a_1} &= \frac{\partial L}{\partial a_4}\frac{\partial a_4}{\partial n_4}w_{14} + \frac{\partial L}{\partial a_5}\frac{\partial a_5}{\partial n_5}w_{15} + \frac{\partial L}{\partial a_6}\frac{\partial a_6}{\partial n_6}w_{16} \\
&= \left(\frac{\partial L}{\partial a_7}\frac{\partial a_7}{\partial n_7}w_{47} + \frac{\partial L}{\partial a_8}\frac{\partial a_8}{\partial n_8}w_{48} + \frac{\partial L}{\partial a_9}\frac{\partial a_9}{\partial n_7}w_{49}\right)\frac{\partial a_4}{\partial n_4}w_{14} \\
&+ \left(\frac{\partial L}{\partial a_7}\frac{\partial a_7}{\partial n_7}w_{57} + \frac{\partial L}{\partial a_8}\frac{\partial a_8}{\partial n_8}w_{58} + \frac{\partial L}{\partial a_9}\frac{\partial a_9}{\partial n_7}w_{59}\right)\frac{\partial a_5}{\partial n_5}w_{15} \\
&+ \left(\frac{\partial L}{\partial a_7}\frac{\partial a_7}{\partial n_7}w_{67} + \frac{\partial L}{\partial a_8}\frac{\partial a_8}{\partial n_8}w_{68} + \frac{\partial L}{\partial a_9}\frac{\partial a_9}{\partial n_7}w_{69}\right)\frac{\partial a_6}{\partial n_6}w_{16}. \\
\frac{\partial L}{\partial a_2} &= \frac{\partial L}{\partial a_4}\frac{\partial a_4}{\partial n_4}w_{24} + \frac{\partial L}{\partial a_5}\frac{\partial a_5}{\partial n_5}w_{25} + \frac{\partial L}{\partial a_6}\frac{\partial a_6}{\partial n_6}w_{26} \\
&= \left(\frac{\partial L}{\partial a_7}\frac{\partial a_7}{\partial n_7}w_{47} + \frac{\partial L}{\partial a_8}\frac{\partial a_8}{\partial n_8}w_{48} + \frac{\partial L}{\partial a_9}\frac{\partial a_9}{\partial n_7}w_{49}\right)\frac{\partial a_4}{\partial n_4}w_{24} \\
&+ \left(\frac{\partial L}{\partial a_7}\frac{\partial a_7}{\partial n_7}w_{57} + \frac{\partial L}{\partial a_8}\frac{\partial a_8}{\partial n_8}w_{58} + \frac{\partial L}{\partial a_9}\frac{\partial a_9}{\partial n_7}w_{59}\right)\frac{\partial a_5}{\partial n_5}w_{25} \\
&+ \left(\frac{\partial L}{\partial a_7}\frac{\partial a_7}{\partial n_7}w_{67} + \frac{\partial L}{\partial a_8}\frac{\partial a_8}{\partial n_8}w_{68} + \frac{\partial L}{\partial a_9}\frac{\partial a_9}{\partial n_7}w_{69}\right)\frac{\partial a_6}{\partial n_6}w_{26}. \\
\frac{\partial L}{\partial a_3} &= \frac{\partial L}{\partial a_4}\frac{\partial a_4}{\partial n_4}w_{34} + \frac{\partial L}{\partial a_5}\frac{\partial a_5}{\partial n_5}w_{35} + \frac{\partial L}{\partial a_6}\frac{\partial a_6}{\partial n_6}w_{36} \\
&= \left(\frac{\partial L}{\partial a_7}\frac{\partial a_7}{\partial n_7}w_{47} + \frac{\partial L}{\partial a_8}\frac{\partial a_8}{\partial n_8}w_{48} + \frac{\partial L}{\partial a_9}\frac{\partial a_9}{\partial n_7}w_{49}\right)\frac{\partial a_4}{\partial n_4}w_{34} \\
&+ \left(\frac{\partial L}{\partial a_7}\frac{\partial a_7}{\partial n_7}w_{57} + \frac{\partial L}{\partial a_8}\frac{\partial a_8}{\partial n_8}w_{58} + \frac{\partial L}{\partial a_9}\frac{\partial a_9}{\partial n_7}w_{59}\right)\frac{\partial a_5}{\partial n_5}w_{35} \\
&+ \left(\frac{\partial L}{\partial a_7}\frac{\partial a_7}{\partial n_7}w_{67} + \frac{\partial L}{\partial a_8}\frac{\partial a_8}{\partial n_8}w_{68} + \frac{\partial L}{\partial a_9}\frac{\partial a_9}{\partial n_7}w_{69}\right)\frac{\partial a_6}{\partial n_6}w_{36}.
\end{aligned}
\tag{31}
$$