Assignment 3

Peter LORENZ

November 29, 2016

1 Compute the gradient $\nabla L(p)$ of the Loss Function.

It is given: $\phi(u) = \frac{1}{1+e^u}$, and $e^c = (p; z^c, t^c) = h(z^c, p) - t^c$, where $h(z^c, (w^s, b^s))$, $s = 1, 2$, $c = 1 \dots C$.

$$
\begin{aligned}
L(p; z, t) &= \frac{1}{C} \sum_{c=1}^{C} \frac{1}{2} \left\| e^c(p; z^c, t^c) \right\|^2 \\
&= \left\| b^2 + w^2 \varphi(b^1 + w^1 + z^c) - t^c \right\|^2 \\
&= \left\| y - t^c \right\|^2 \\
&= \frac{1}{2} \sqrt{(b_1^2 + w_1^2 \varphi(b^1 + w^1 z^c) - t_1^c)^2 + (b_2^2 + w_2^2 \varphi(b^1 + w^1 z^c))^2)}^2 \\
&= \frac{1}{2} \left(b_1^2 + \begin{pmatrix} w_{11}^2 & w_{12}^2 & w_{13}^2 & w_{14}^2 \end{pmatrix} \begin{pmatrix} \varphi_1 \\ \varphi_2 \\ \varphi_3 \\ \varphi_4 \end{pmatrix} - t_1^2 \right)^2 + \\
&+ \frac{1}{2} \left(b_2^2 + \begin{pmatrix} w_{21}^2 & w_{22}^2 & w_{23}^2 & w_{24}^2 \end{pmatrix} \begin{pmatrix} \varphi_1 \\ \varphi_2 \\ \varphi_3 \\ \varphi_4 \end{pmatrix} - t_2^2 \right)^2 \\
&= \underline{\frac{1}{2} \left((e_1^c)^2 + (e_2^c)^2 \right)}
\end{aligned}
$$

Now, we have to make derivations $\nabla_b L$, $\nabla_{w_1} L$, $\nabla_{b_2} L$ and $\nabla_{w_2} L$:

$$
\begin{aligned}
\nabla_{b^2} L(p; z, t) &= \begin{bmatrix} \frac{\partial}{\partial b_1^2} L \\ \frac{\partial}{\partial b_2^2} L \end{bmatrix} = \begin{bmatrix} \frac{1}{2} \cdot 2 \cdot e_1^c \cdot 1 \\ \frac{1}{2} \cdot 2 \cdot e_2^c \cdot 1 \end{bmatrix} = \underline{\begin{bmatrix} e_1^c \\ e_2^c \end{bmatrix}} \\
\nabla_{w^2} L(p; z, t) &= \begin{bmatrix} \frac{\partial}{\partial b_{11}^2} L & \frac{\partial}{\partial b_{12}^2} L & \frac{\partial}{\partial b_{13}^2} L \\ \frac{\partial}{\partial b_{21}^2} L & \frac{\partial}{\partial b_{22}^2} L & \frac{\partial}{\partial b_{23}^2} L \end{bmatrix} = \begin{bmatrix} \frac{1}{2} \cdot 2 \cdot e_1^c \cdot \varphi_1 & \frac{1}{2} \cdot 2 \cdot e_1^c \cdot \varphi_2 & \frac{1}{2} \cdot 2 \cdot e_1^c \cdot \varphi_3 \\ \frac{1}{2} \cdot 2 \cdot e_2^c \cdot \varphi_1 & \frac{1}{2} \cdot 2 \cdot e_2^c \cdot \varphi_2 & \frac{1}{2} \cdot 2 \cdot e_2^c \cdot \varphi_3 \end{bmatrix} \\
&= \begin{bmatrix} e_1^c \cdot & \begin{pmatrix} \varphi_1 & \varphi_2 & \varphi_3 \end{pmatrix} \\ e_2^c \cdot & \begin{pmatrix} \varphi_1 & \varphi_2 & \varphi_3 \end{pmatrix} \end{bmatrix} = \begin{bmatrix} e_1^c & \cdot \varphi^T \\ e_2^c & \cdot \varphi^T \end{bmatrix} = \underline{e \cdot \varphi^T} \\
\nabla_{b^1} L(p; z, t) &= \begin{bmatrix} \frac{\partial}{\partial b_1^1} L \\ \frac{\partial}{\partial b_2^1} L \end{bmatrix} = \begin{bmatrix} e_1^c \cdot w_{11}^2 \cdot \varphi_1' & + & e_2^c \cdot w_{21}^2 \cdot \varphi_1' \\ e_1^c \cdot w_{12}^2 \cdot \varphi_2' & + & e_2^c \cdot w_{21}^2 \cdot \varphi_2' \\ e_1^c \cdot w_{13}^2 \cdot \varphi_3' & + & e_2^c \cdot w_{21}^2 \cdot \varphi_3' \\ e_1^c \cdot w_{14}^2 \cdot \varphi_4' & + & e_2^c \cdot w_{21}^2 \cdot \varphi_4' \end{bmatrix} = \\
&= \begin{bmatrix} e_1 \cdot w_{11}^2 & + & e_2 \cdot w_{21}^2 \\ e_1 \cdot w_{12}^2 & + & e_2 \cdot w_{21}^2 \\ e_1 \cdot w_{13}^2 & + & e_2 \cdot w_{21}^2 \\ e_1 \cdot w_{14}^2 & + & e_2 \cdot w_{21}^2 \end{bmatrix} \circ \begin{bmatrix} \varphi_1' \\ \varphi_2' \\ \varphi_3' \\ \varphi_4' \end{bmatrix} = \underline{\left[\left(w^2 \right)^T \cdot e^c \right] \circ \varphi'}
\end{aligned}
$$

$$\nabla_{w^1} L(p;z,t) = \begin{bmatrix} \frac{\partial}{\partial w^1_{11}}L & \frac{\partial}{\partial w^1_{12}}L & \frac{\partial}{\partial w^1_{13}}L \\ \frac{\partial}{\partial w^1_{21}}L & \frac{\partial}{\partial w^1_{22}}L & \frac{\partial}{\partial w^1_{23}}L \\ \frac{\partial}{\partial w^1_{31}}L & \frac{\partial}{\partial w^1_{32}}L & \frac{\partial}{\partial w^1_{33}}L \\ \frac{\partial}{\partial w^1_{41}}L & \frac{\partial}{\partial w^1_{42}}L & \frac{\partial}{\partial w^1_{43}}L \end{bmatrix} = \begin{bmatrix} (w_1 e)(\; z_1 \quad z_2 \quad z_3 \;) \\ (w_2 e)(\; z_1 \quad z_2 \quad z_3 \;) \\ (w_3 e)(\; z_1 \quad z_2 \quad z_3 \;) \\ (w_4 e)(\; z_1 \quad z_2 \quad z_3 \;) \end{bmatrix} \circ \varphi^{'} = \underline{\left(\left(\left(w^2\right) e\right) z^T\right) \circ \varphi^{'}}$$

Auxilery Calculation: (Please, do not try this at home!)

$$\begin{aligned}
\frac{\partial}{\partial w^1_{11}}L &= \frac{1}{2}\left(2\left(e_1 \cdot w^2_{11} \cdot \varphi^{'}_1 \cdot 1 \cdot z^c_1\right) + 2 \cdot \left(e_2 \cdot w^2_{21} \cdot \varphi^{'}_1 \cdot 1 \cdot z^c_1\right)\right) = \\
&\quad e_1 \cdot w^2_{11} \cdot \varphi^{'}_1 \cdot z^c_1 + e_2 \cdot w^2_{21} \cdot \varphi^{'}_1 \cdot z^c_1 = \left[e_1 \cdot w^2_{11} + e_2 \cdot w^2_{21}\right] \cdot \varphi^{'}_1 z^c_1
\end{aligned}$$

$$\begin{aligned}
\frac{\partial}{\partial w^1_{12}}L &= e_1 \cdot w^2_{11} \cdot \varphi^{'}_1 \cdot z^c_2 + e_2 \cdot w^2_{21} \cdot \varphi^{'}_1 \cdot z^c_2 = \left[e_1 \cdot w^2_{11} + e_2 \cdot w^2_{21}\right] \cdot \varphi^{'}_1 z^c_2 \\
\frac{\partial}{\partial w^1_{13}}L &= e_1 \cdot w^2_{11} \cdot \varphi^{'}_1 \cdot z^c_3 + e_2 \cdot w^2_{21} \cdot \varphi^{'}_1 \cdot z^c_3 = \left[e_1 \cdot w^2_{11} + e_2 \cdot w^2_{21}\right] \cdot \varphi^{'}_1 z^c_3
\end{aligned}$$

$$\begin{aligned}
\frac{\partial}{\partial w^1_{21}}L &= e_1 \cdot w^2_{12} \cdot \varphi^{'}_2 \cdot z^c_1 + e_2 \cdot w^2_{22} \cdot \varphi^{'}_2 \cdot z^c_1 = \left[e_1 \cdot w^2_{12} + e_2 \cdot w^2_{22}\right] \cdot \varphi^{'}_2 z^c_1 \\
\frac{\partial}{\partial w^1_{22}}L &= e_1 \cdot w^2_{12} \cdot \varphi^{'}_2 \cdot z^c_2 + e_2 \cdot w^2_{22} \cdot \varphi^{'}_2 \cdot z^c_2 = \left[e_1 \cdot w^2_{12} + e_2 \cdot w^2_{22}\right] \cdot \varphi^{'}_2 z^c_2 \\
\frac{\partial}{\partial w^1_{23}}L &= e_1 \cdot w^2_{12} \cdot \varphi^{'}_2 \cdot z^c_3 + e_2 \cdot w^2_{22} \cdot \varphi^{'}_2 \cdot z^c_3 = \left[e_1 \cdot w^2_{12} + e_2 \cdot w^2_{22}\right] \cdot \varphi^{'}_2 z^c_3
\end{aligned}$$

$$\begin{aligned}
\frac{\partial}{\partial w^1_{31}}L &= e_1 \cdot w^2_{13} \cdot \varphi^{'}_3 \cdot z^c_1 + e_2 \cdot w^2_{23} \cdot \varphi^{'}_3 \cdot z^c_1 = \left[e_1 \cdot w^2_{13} + e_2 \cdot w^2_{23}\right] \cdot \varphi^{'}_3 z^c_1 \\
\frac{\partial}{\partial w^1_{32}}L &= e_1 \cdot w^2_{13} \cdot \varphi^{'}_3 \cdot z^c_2 + e_2 \cdot w^2_{23} \cdot \varphi^{'}_3 \cdot z^c_2 = \left[e_1 \cdot w^2_{13} + e_2 \cdot w^2_{23}\right] \cdot \varphi^{'}_3 z^c_2 \\
\frac{\partial}{\partial w^1_{33}}L &= e_1 \cdot w^2_{13} \cdot \varphi^{'}_3 \cdot z^c_3 + e_2 \cdot w^2_{23} \cdot \varphi^{'}_3 \cdot z^c_3 = \left[e_1 \cdot w^2_{13} + e_2 \cdot w^2_{23}\right] \cdot \varphi^{'}_3 z^c_3
\end{aligned}$$

$$\begin{aligned}
\frac{\partial}{\partial w^1_{41}}L &= e_1 \cdot w^2_{14} \cdot \varphi^{'}_4 \cdot z^c_1 + e_2 \cdot w^2_{24} \cdot \varphi^{'}_4 \cdot z^c_1 = \left[e_1 \cdot w^2_{14} + e_2 \cdot w^2_{24}\right] \cdot \varphi^{'}_4 z^c_1 \\
\frac{\partial}{\partial w^1_{42}}L &= e_1 \cdot w^2_{14} \cdot \varphi^{'}_4 \cdot z^c_2 + e_2 \cdot w^2_{24} \cdot \varphi^{'}_4 \cdot z^c_2 = \left[e_1 \cdot w^2_{14} + e_2 \cdot w^2_{24}\right] \cdot \varphi^{'}_4 z^c_2 \\
\frac{\partial}{\partial w^1_{43}}L &= e_1 \cdot w^2_{14} \cdot \varphi^{'}_4 \cdot z^c_3 + e_2 \cdot w^2_{24} \cdot \varphi^{'}_4 \cdot z^c_3 = \left[e_1 \cdot w^2_{14} + e_2 \cdot w^2_{24}\right] \cdot \varphi^{'}_4 z^c_3
\end{aligned}$$

$\circ$ This symbol means: `https://en.wikipedia.org/wiki/Hadamard_product_(matrices)`

2 What could be the stopping Criteria?

I have implemented two kind of stopping criterion:

1. current_error - train_error $< 10e-8$
2. $\left\|d^k - d^{k+1}\right\|^2 = \left\|\alpha \cdot d^k\right\|^2 < 10e-8$ (this one was suggested in the UE hour)

Note: $10e-8$ is just a placeholder and can be variable according to each count of hidden neurons.

3 Report the loss (eq. 4) on the training set as well as on the test set.

I saved the loss in txt files, in the directory "report/simple_algo/".
Simple Algorithm:

Hidden Units	Training	Test Set
3	0.09290453992	0.124218595812
5	0.09019591330	0.141352647565
15	0.10264507422	0.123082317606
25	0.16950097952	0.287452065718

The average error of the training set shows up a linear increasing and the test set shows an increasing of the error, as well.

4 Investigate the different possibilities for computing the descent direction d^k. Which one performs better?

The simple algorithm shows up little difference between training and testing error, except of the last line, where it seems to be overfitted. The gauss newton method shows more or less little error differences, but the last line is more close to each other. I cannot really state that which one performs better, but the gauss newton converges faster.

5 Run the better algorithm with different numbers of units in the hidden layer (e.g. 3, 5, 15, 25) (hidden units). What do you find?

Hidden Units	Training	Test Set
3	0.17912691741	0.222413742791
5	0.15996384268	0.156167594273
15	0.18566526823	0.143968071313
25	0.25700683902	0.286701007958

Compared to the simple algorithm the gauss newton algorithm does not show linear increasing and decreasing.