Your Paper

You

November 18, 2023

Abstract

Your abstract.

1 Introduction

Remembering that

$$\frac{\partial g(M)}{\partial t} = \sum_{i,j} \frac{\partial g(M)}{\partial M_{i,j}} \frac{\partial M_{i,j}}{\partial t},$$

chain rule allows us to say the following:

$$\left[\frac{\partial L}{\partial W}\right]_{m,n} = \frac{\partial g(Y)}{\partial W_{m,n}}$$

$$\begin{aligned}
\left[\frac{\partial L}{\partial W}\right]_{m,n} &= \frac{\partial g(Y)}{\partial W_{m,n}} \qquad \text{because of } L = g(Y) && (1)\\
&= \sum_{i,j} \frac{\partial g(Y)}{\partial Y_{i,j}} \frac{\partial Y_{i,j}}{\partial W_{m,n}} \qquad \text{because of property above + chain rule} && (2)\\
&= \sum_{i,j} \frac{\partial g(Y)}{\partial Y_{i,j}} \frac{\partial}{\partial W_{m,n}} \left(\left[XW^T\right]_{i,j} + \cancel{B_{i,j}} \right) \qquad \text{Since } B_{i,j} \text{ is not a function of } W && (3)\\
& && (4)
\end{aligned}$$

Expanding $\left[XW^T\right]_{i,j}$ we get:

$$\begin{aligned}
\left[XW^T\right]_{i,j} &= \sum_k X_{i,k} W^T_{k,j} && (5)\\
&= \sum_k X_{i,k} W_{j,k} \qquad \text{flipping rows and columns for } W \text{ transpose} && (6)
\end{aligned}$$

Which now allows us to state:

$$\begin{aligned}
\frac{\partial \left[XW^T\right]_{i,j}}{\partial W_{m,n}} &= \sum_k X_{i,k} \frac{\partial W_{j,k}}{\partial W_{m,n}} && (7)\\
&= \cancel{\sum_k} X_{i,k} \delta_{j,m} \cancel{\delta_{k,n}} \qquad \delta_{k,n} = 1 \text{ iff } k = n && (8)\\
&= X_{i,n} \delta_{j,m} && (9)
\end{aligned}$$

Plugging in this result in the main equation above yields:

$$\begin{aligned}
&= \sum_{i,j} \frac{\partial g(Y)}{\partial Y_{i,j}} \frac{\partial}{\partial W_{m,n}} \left(\left[XW^T\right]_{i,j} \right) && (10)\\
&= \sum_{i,j} \frac{\partial g(Y)}{\partial Y_{i,j}} X_{i,n} \delta_{j,m} && (11)\\
&= \sum_i \cancel{\sum_j} \frac{\partial g(Y)}{\partial Y_{i,j}} X_{i,n} \cancel{\delta_{j,m}} \qquad \delta_{k,n} = 1 \text{ iff } j = m && (12)\\
&= \underbrace{\sum_i X_{i,n}}_{\text{loop over the rows}} \left[\frac{\partial L}{\partial Y}\right]_{i,m} && (13)\\
&= X^T \frac{\partial L}{\partial Y} && (14)
\end{aligned}$$

again, since $L = g(Y)$:

$$\frac{\partial L}{\partial b_k} = \sum_{i,j} \frac{\partial g(Y)}{\partial Y_{i,j}} \frac{\partial Y_{i,j}}{\partial b_k} \tag{1}$$

We can unpack $Y_{i,j}$ just like we did above:

$$\begin{aligned}
\frac{\partial Y_{i,j}}{\partial b_k} &= \frac{\partial}{\partial b_k}\left(\cancel{\left[XW^T\right]_{i,j}} + B_{i,j}\right) \qquad XW^T \text{ independent of } b_k && (2)\\
&= \frac{\partial b_j}{b_k} \qquad \text{due to } B_{i,j} = b_j && (3)\\
&= \delta_{j,k} && (4)
\end{aligned}$$

Plugging the above result into the main equation yields:

$$\begin{aligned}
\frac{\partial L}{\partial b_k} &= \sum_i \cancel{\sum_j} \frac{\partial g(Y)}{\partial Y_{i,j}} \cancel{\delta_{j,k}} \qquad \delta_{j,k} = 1 \text{ iff } j = k && (5)\\
&= \sum_i \frac{\partial L}{\partial Y_{i,k}} && (6)\\
\frac{\partial L}{\partial \boldsymbol{b}} &= \sum_i \frac{\partial L}{\partial \boldsymbol{y}_i} && (7)
\end{aligned}$$

Once again, since $L = g(Y)$:

$$\begin{aligned}
\left[\frac{\partial L}{\partial X}\right]_{m,n} &= \frac{\partial g(Y)}{\partial X_{m,n}} && (1)\\
&= \sum_{i,j} \frac{\partial g(Y)}{\partial Y_{i,j}} \frac{\partial Y_{i,j}}{\partial X_{m,n}} \qquad \text{because of property above + chain rule} && (2)\\
&= \sum_{i,j} \frac{\partial g(Y)}{\partial Y_{i,j}} \frac{\partial}{\partial X_{m,n}} \left(\left[XW^T\right]_{i,j} + \cancel{B_{i,j}}\right) \qquad \text{Since } B_{i,j} \text{ is not a function of } X && (3)
\end{aligned}$$

Expanding $\left[XW^T\right]_{i,j}$ we get:

$$\begin{aligned}
\left[XW^T\right]_{i,j} &= \sum_k X_{i,k} W^T_{k,j} && (4)\\
&= \sum_k X_{i,k} W_{j,k} \qquad \text{flipping rows and columns for } W \text{ transpose} && (5)
\end{aligned}$$

Which now allows us to state:

$$\begin{aligned}
\frac{\partial \left[XW^T\right]_{i,j}}{\partial X_{m,n}} &= \sum_k W X_{j,k} \frac{\partial X_{i,k}}{\partial X_{m,n}} && (6)\\
&= \cancel{\sum_k} W_{j,k} \delta_{i,m} \cancel{\delta_{k,n}} \qquad \delta_{k,n} = 1 \text{ iff } k = n && (7)\\
&= W_{j,k} \delta_{i,m} && (8)
\end{aligned}$$

Plugging in this result in the main equation above yields:

$$\begin{aligned}
&= \sum_{i,j} \frac{\partial g(Y)}{\partial Y_{i,j}} \frac{\partial}{\partial X_{m,n}} \left(\left[XW^T\right]_{i,j}\right) && (9)\\
&= \sum_{i,j} \frac{\partial g(Y)}{\partial Y_{i,j}} W_{j,k} \delta_{i,m} && (10)\\
&= \cancel{\sum_i} \sum_j \frac{\partial g(Y)}{\partial Y_{i,j}} W_{j,k} \cancel{\delta_{i,m}} \qquad \delta_{i,m} = 1 \text{ iff } i = m && (11)\\
&= \underbrace{\sum_j W_{j,k}}_{\text{loop over the columns}} \left[\frac{\partial L}{\partial Y}\right]_{m,j} && (12)\\
&= \frac{\partial L}{\partial Y} W && (13)
\end{aligned}$$

And the dimensions match due to $Y \in \mathbb{R}^{S\times N}$, resulting in $\frac{\partial L}{\partial Y} \in \mathbb{R}^{S\times N}$. Since $W^T \in \mathbb{R}^{M\times N}$, $W \in \mathbb{R}^{N\times M}$, thus the matrix product results in a $S \times M$ matrix, which is the original size of X.

Once again, since $L = g(Y)$ and $\boldsymbol{Y} = h(\boldsymbol{X}) \rightarrow Y_{i,j} = h(X_{i,j})$:

$$\left[\frac{\partial L}{\partial X}\right]_{m,n} = \frac{\partial g(Y)}{\partial X_{m,n}} = \frac{\partial g(h(X))}{\partial X_{m,n}} \tag{1}$$

$$= \sum_{i,j} \frac{\partial g(h(X))_{i,j}}{\partial h(X)_{i,j}} \frac{\partial h(X)_{i,j}}{\partial X_{i,j}} \frac{\partial X_{i,j}}{\partial X_{m,n}} \qquad \text{chain rule} \tag{2}$$

$$= \sum_{i,j} \frac{\partial g(h(X))_{i,j}}{\partial h(X)_{i,j}} \frac{\partial h(X)_{i,j}}{\partial X_{i,j}} \delta_{i,m}\delta_{j,n} \tag{3}$$

$$= \cancel{\sum_{i} \sum_{j}} \frac{\partial g(h(X))_{i,j}}{\partial h(X)_{i,j}} \frac{\partial h(X)_{i,j}}{\partial X_{i,j}} \cancel{\delta_{i,m}\delta_{j,n}} \qquad \delta_{i,m}, \delta_{j,n} = 1 \text{ iff } i = m, j = n \tag{4}$$

$$= \frac{\partial L}{\partial Y_{m,n}} \frac{\partial h(X)_{m,n}}{\partial X_{m,n}} \tag{5}$$

Thus, rewriting the result in matrix notation provides us with the result

$$\frac{\partial L}{\partial X} = \frac{\partial L}{\partial Y} \circ \frac{\partial h(X)}{\partial X},$$

where $\circ$ is the Hadamard product.

Since in a local minimum the Hessian is positive definite,

$$\mathbf{x}^\top \mathbf{H} \mathbf{x} > 0 \quad \forall \mathbf{x} \neq \mathbf{0}$$

By definition, an eigenvalue of $\mathbf{H}$ is the one scalar λ for which

$$\mathbf{H}\mathbf{x} = \lambda \mathbf{x}$$

Left multiplying both sides by $\mathbf{x}^\top$ we get

$$\mathbf{x}^\top \mathbf{H}\mathbf{x} = \mathbf{x}^\top \lambda \mathbf{x}$$

Since $\mathbf{x}^\top \mathbf{H}\mathbf{x}$ is greater than 0, focusing on the RHS we can see

$$\lambda \mathbf{x}^\top \mathbf{x}$$

and given that λ is a scalar, commutative, it follows that

$$= \lambda \|\mathbf{x}\|^2$$

The norm of a non-zero vector is greater than zero by definition, which leads us to the desired result of λ being greater than zero for the equality to hold.