1 APPENDIX

1.1 Softmax proof

Proof. Let c be a scalar constant such that $log(c) = -max(x_1, ..., x_n)$

$$
\begin{aligned}
softmax(x_i) &= \frac{e^{x_i}}{\sum_{j=1}^{n} e^{x_j}} && (1)\\
&= \frac{c * e^{x_i}}{c * \sum_{j=1}^{n} e^{x_j}} && (2)\\
&= \frac{c * e^{x_i}}{\sum_{j=1}^{n} c * e^{x_j}} && (3)\\
&= \frac{e^{log(c)} * e^{x_i}}{\sum_{j=1}^{n} e^{log(c)} * e^{x_j}} && \text{(By property } c = e^{log(c)}\text{)}\\
&= \frac{e^{(log(c))+x_i}}{\sum_{j=1}^{n} e^{(log(c))+x_j}} && \text{(By property } e^a * e^b = e^{a+b}\text{)}\\
&= \frac{e^{-max(x)+x_i}}{\sum_{j=1}^{n} e^{-max(x)+x_j}} && \square
\end{aligned}
$$

1.2 LogSoftmax proof

Logsoftmax is a canonical example of a numerically unstable formula that is commonly used in DL. Prior literature shows how to rewrite the formula to obtain a numerically stable solution, but does not provide proof that the two formulas are mathematically equivalent. To our best knowledge, this is the fist comprehensive proof which shows that step by step.

LogSoftmax performs softmax followed by the logarithm function and therefore, outputs log probabilities. LogSoftmax is defined as:

$$logsoftmax(\vec{x})_i = \frac{log(e^{x_i})}{\sum_{j=1}^{n} e^{x_j}} \tag{4}$$

This mathematical formula is numerically unstable and should be implemented as:

$$logsoftmax(\vec{x})_i = x_i - max(\vec{x}) - log(\sum_{j=1}^{n} e^{x_j - max(\vec{x})}) \tag{5}$$

These two equations are mathematically equivalent, which can be proved utilizing the identity:

$$log(\sum_{j=1}^{n} e^{x_j}) = max(\vec{x}) + log(\sum_{j=1}^{n} e^{x_j - max(\vec{x})}) \tag{6}$$

We first prove the correctness of the identity and then the mathematical equivalence of the numerically stable and unstable logsoftmax formulas.

Proof. Let c be a scalar constant such that $c = max(x_1, ..., x_n)$

$$log(\sum_{j=1}^{n} e^{x_j}) = max(\vec{x}) + log(\sum_{j=1}^{n} e^{x_j - max(\vec{x})}) \tag{7}$$

$$= c + log(\sum_{j=1}^{n} e^{x_j - c}) \tag{8}$$

$$= c + log(\sum_{j=1}^{n} e^{x_j} * e^{-c}) \qquad \text{By property } e^{a-b} = e^a * e^b \tag{9}$$

$$= c + log(e^{-c}) + log(\sum_{j=1}^{n} e^{x_j}) \qquad \text{By property } log(ab) = log(a) + log(b) \tag{10}$$

$$= c + (-c * log(e)) + log(\sum_{j=1}^{n} e^{x_j}) \qquad \text{By property } log(a^b = b * log(a) \tag{11}$$

$$= c + (-c(1)) + log(\sum_{j=1}^{n} e^{x_j}) \qquad \text{By property } log(e) = 1 \tag{12}$$

$$= c - c + log(\sum_{j=1}^{n} e^{x_j}) \tag{13}$$

$$= log(\sum_{j=1}^{n} e^{x_j}) \tag{14}$$

□

Proof.

$$logsoftmax(\vec{x})_i = \frac{log(e^{x_i})}{\sum_{j=1}^{n}(e^{x_j})} \tag{15}$$

$$= log(e^{x_i}) - log(\sum_{j=1}^{n} e^{x_j}) \qquad \text{By property } log(a/b) = log(a) - log(b) \tag{16}$$

$$= x_i * log(e) - log(\sum_{j=1}^{n} e^{x_j}) \qquad \text{By property } log(a^b) = b * log(a) \tag{17}$$

$$= x_i * (1) - log(\sum_{j=1}^{n} e^{x_j}) \qquad \text{By property } log(e) = 1 \tag{18}$$

$$= x_i - (max(\vec{x}) + log(\sum_{j=1}^{n} e^{x_j - max(\vec{x})})) \qquad \text{By identity in Equation 6} \tag{19}$$

$$= x_i - max(\vec{x}) - log(\sum_{j=1}^{n} e^{x_j - max(\vec{x})}) \tag{20}$$

□