ML HW 7 - Deep Learning

Thursday, January 14, 2021 12:41 PM

Problem 3: In machine learning you often come across problems which contain the following quantity

$$y = \log \sum_{i=1}^{N} e^{x_i} \leftarrow$$

For example if we want to calculate the log-likelihood of neural network with a softmax output we get this quantity due to the normalization constant. If you try to calculate it naively, you will quickly encounter underflows or overflows, depending on the scale of x_i. Despite working in log-space, the limited precision of computers is not enough and the result will be ∞ or $-\infty$.

To combat this issue we typically use the following identity:

$$y = \log \sum_{i=1}^{N} e^{x_i} = a + \log \sum_{i=1}^{N} e^{x_i - a}$$

for an arbitrary a. This means, you can shift the center of the exponential sum. A typical value is setting a to the maximum ($a = \max_i x_i$), which forces the greatest value to be zero and even if the other values would underflow, you get a reasonable result.

Your task is to show that the identity holds.

This is called the *log-sum-exp trick* and is often used in practice.

$$
\begin{aligned}
y &= \log \sum_{i=1}^{N} e^{x_i} \\
e^y &= \sum_{i=1}^{N} e^{x_i} \qquad \mid \exp(\cdot) \\
e^{-a} e^y &= e^{-a} \sum_{i=1}^{N} e^{x_i} \qquad \mid e^{-a} \\
e^{y-a} &= \sum_{i=1}^{N} e^{-a} e^{x_i} \\
y - a &= \log \sum_{i=1}^{N} e^{x_i - a} \\
y &= a + \log \sum_{i=1}^{N} e^{x_i - a}
\end{aligned}
$$

$$e^a \cdot e^b = e^{a+b}$$

Problem 4: Similar to the previous exercise we can compute the output of the softmax function $\pi_i = \frac{e^{x_i}}{\sum_{i=1}^{N} e^{x_i}}$ in a numerically stable way by shifting by an arbitrary constant a:

$$\frac{e^{x_i}}{\sum_{i=1}^{N} e^{x_i}} = \frac{e^{x_i - a}}{\sum_{i=1}^{N} e^{x_i - a}}$$

often chosen $a = \max_i x_i$. Show that the above identity holds.

For some arbitrary constant C we have

$$\frac{e^{x_i}}{\sum_{i=1}^{N} e^{x_i}} = \frac{C e^{x_i}}{C \sum_{i=1}^{N} e^{x_i}} = \frac{e^{x_i + \log(C)}}{\sum_{i=1}^{N} e^{x_i + \log(C)}}$$

Since C is just an arbitrary constant, we can replace $\log(C) = -a$ and get $\frac{e^{x_i - a}}{\sum_{i=1}^{N} e^{x_i - a}}$.

$$C = e^{-a}$$

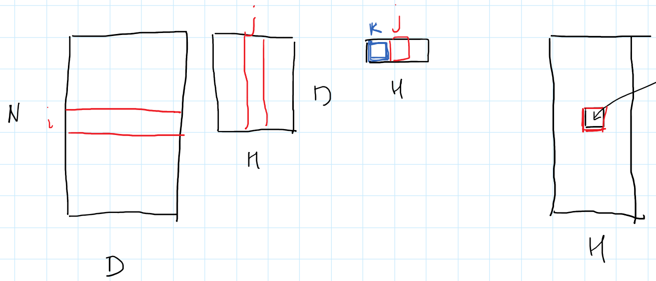
$$C \cdot e^a = e^{\log C} \cdot e^a = e^{\log C + a}$$

forward:

X [N,D] $\quad W$ [D,H] $\quad b$ [H] $\longmapsto A$ [N,H]

x_i [D] $\qquad a_i = W^T x_i + b$ [H]

E



$$A_{ij} = x_{i1} \cdot W_{1j} + x_{i2} \cdot W_{2j} + \dots + x_{iD} W_{Dj} + b_j$$
$$= X[i,:] @ W[:,j] + b[j]$$

$$\boxed{A = X @ W + b} \leftarrow \boxed{\text{forward}}$$

[N,H] [H]

$\frac{\partial E}{\partial b}$ [1,H]

backward

d_out [N,H] $\longmapsto$ d_inputs [N,D] $\quad$ d_weight [D,H] $\quad$ d_bias [H]

$$\text{d\_out}[i,j] = \frac{\partial E}{\partial A_{ij}} \qquad \text{d\_inputs}[k,l] = \frac{\partial E}{\partial X_{kl}} \qquad \text{d\_weight}[k,l] = \frac{\partial E}{\partial W_{kl}} \qquad \text{d\_bias}[k] = \frac{\partial E}{\partial b_k}$$

$\dots = \frac{\partial E}{\dots} \sum^{N} \sum^{H} \frac{\partial E}{\dots} \frac{\partial A_{ij}}{\dots} \sum^{D} \frac{\partial E}{\dots} \mathbb{I}(j=k) = \sum^{N} \frac{\partial E}{\dots}$

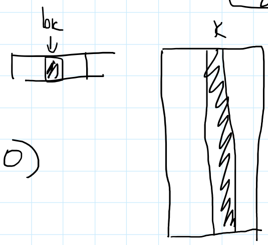
$$d\_bias[k] = \frac{\partial E}{\partial b_k} = \sum_{i=1}^{N}\sum_{j=1}^{H} \frac{\partial E}{\partial A_{ij}} \cdot \frac{\partial A_{ij}}{\partial b_k} = \sum_{i}\sum_{j=1}^{D} \frac{\partial E}{\partial A_{ij}} \cdot \mathbb{I}(j=k) = \boxed{\sum_{i=1}^{N} \frac{\partial E}{\partial A_{ik}}}$$

$$\frac{\partial A_{ij}}{\partial b_k} = \begin{cases} 0 & \text{if } j \neq k \\ 1 & \text{if } j = k \end{cases}$$

$$\cancel{\frac{\partial E}{\partial A_{i1}} \cdot \mathbb{I}(k=1)} + \boxed{\frac{\partial E}{\partial A_{i2}} \cdot \mathbb{I}(k=2)} + \cancel{\ldots}$$

$$d\_out = \left(\frac{\partial E}{\partial A}\right)^T$$

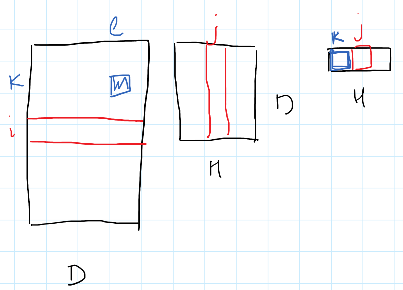
$d\_bias[k] =$

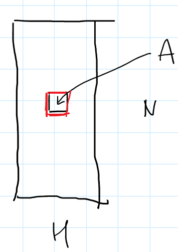


d_bias = np.sum(d_out, axis=0)

forward:

$$X \;[N,D] \quad W \;[D,H] \quad b \;[H] \longmapsto A \;[N,H] \qquad x_i \;[D] \qquad a_i = W^T x_i + b \;[H]$$





E

$$A_{ij} = \boxed{X_{i1} \cdot W_{1j}} + X_{i2} \cdot W_{2j} + \ldots + X_{iD} W_{Dj} + b_j$$

$$= X[i,:] @ W[:,j] + b[j]$$

$$d\_inputs[k,l] = \frac{\partial E}{\partial X_{kl}} = \sum_{i=1}^{N}\sum_{j=1}^{H} \frac{\partial E}{\partial A_{ij}} \cdot \frac{\partial A_{ij}}{\partial X_{kl}} = \sum_{i=1}^{N}\sum_{j} \frac{\partial E}{\partial A_{ij}} \cdot W_{lj} \cdot \mathbb{I}(i=k) = \sum_{j=1}^{H} \frac{\partial E}{\partial A_{kj}} \cdot W_{lj}$$

[N,D]

$$\frac{\partial A_{ij}}{\partial X_{i\boxed{l}}} = W_{lj}$$

$$\frac{\partial A_{ij}}{\partial X_{kl}} = \begin{cases} 0 & \text{if } i \neq k \\ W_{lj} & \text{if } i = k \end{cases}$$

$$= W_{lj} \cdot \mathbb{I}(i=k)$$

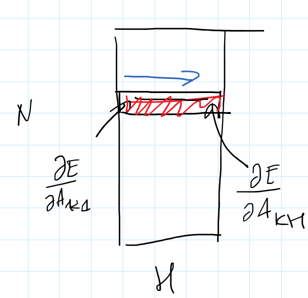
$$\frac{\partial E}{\partial X_{kl}} = \boxed{\sum_{j=1}^{H} \frac{\partial E}{\partial A_{kj}} \cdot W_{lj}}$$

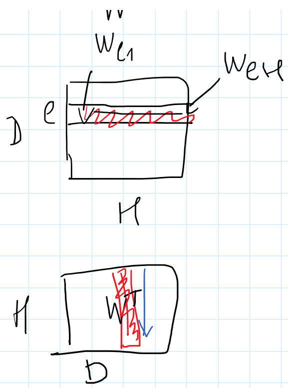
$$d\_out = (\partial E/\partial A)^T$$

W

W_{l1} W_{lH}

d_out = (∂E/∂A)





$$\text{d\_out}[k,:] @ W[l,:] = \boxed{\frac{\partial E}{\partial X_{kl}}}$$

$$\underset{[N,H]}{\text{d\_out}} @ \underset{[H,D]}{W.T} \mapsto \underset{[N,D]}{\text{d\_input}}$$

$$y = \begin{cases} x & \text{if } x \geq 0 \\ 0 & \text{if } x < 0 \end{cases} \iff y = x \cdot \mathbb{1}(x \geq 0) =$$

$$\frac{\partial y}{\partial x} = \begin{cases} 1 & \text{if } x \geq 0 \\ 0 & \text{if } x < 0 \end{cases} \iff \frac{\partial y}{\partial x} = \mathbb{1}(x \geq 0)$$

$$\frac{\partial E}{\partial x} = \frac{\partial E}{\partial y} \frac{\partial y}{\partial x} = \overbrace{\boxed{\frac{\partial E}{\partial y}}}^{\text{d\_out}} \overbrace{\boxed{\mathbb{1}(x \geq 0)}}^{(x \geq 0)}$$

Cross Ent $(y, A) \mapsto E$

forward:

y (labels) $[N,K]$

A (logits) $[N,K]$

$$\begin{bmatrix} 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \end{bmatrix}$$

$$E = \frac{1}{N} \sum_{i=1}^{N} \sum_{k=1}^{K} y_{ik} \cdot \log \boxed{P_{ik}}$$

$p(y_i = k \mid x_i)$

$$= \frac{1}{N} \sum_{i,k} y_{ik} \cdot \log \boxed{\frac{e^{A_{ik}}}{\sum_{c=1}^{K} e^{A_{ic}}}}$$

softmax $A[i,:]$

$$\log \frac{a}{b} = \log a - \log b$$

$$= \frac{1}{N} \sum_{i,k} y_{ik} \left(A_{ik} - \underbrace{\log\left(\sum_c e^{A_{ic}}\right)}_{\text{log-sum-exp}} \right)$$

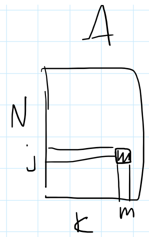
$$\boxed{\log P_{ik}}$$

$A[i,:]$

$$= \left(\boxed{A_{ik} - m_i} - \log\left(\sum_c e^{\boxed{A_{ic} - m_i}}\right) \right)$$

$$m_i = \max A_{ic}$$

$$\frac{1}{N} \text{np.sum}(\underset{[N,K]}{y \odot \log P})$$



$$m_i = \max_c A_{ic}$$

backward:

$$\underset{[\,]}{d\_out} \mapsto \underset{[N,K]}{d\_logits}[j,m] = \frac{\partial E}{\partial A_{jm}}$$

$$E = \frac{1}{N}\sum_{i=1}^{N}\sum_{k=1}^{K} y_{ik}\left(A_{ik} - \log\left(\sum_{c=1}^{K} e^{A_{ic}}\right)\right)$$

$$= \frac{1}{N}\sum_{i=1}^{N}\left[\left(\sum_{k=1}^{K} y_{ik}\cdot A_{ik}\right) - \left(\sum_{k=1}^{K} y_{ik}\cdot \log\left(\sum_c e^{A_{ic}}\right)\right)\right]$$

$$\boxed{E = \frac{1}{N}\sum_i\left[\left(\sum_k y_{ik}A_{ik}\right) - \log\left(\sum_c e^{A_{ic}}\right)\cdot \sum_k y_{ik}\right]}$$

$|01000|$

$$\boxed{\frac{\partial E}{\partial A_{jm}}} = \frac{1}{N}\frac{\partial}{\partial A_{jm}}\left[\left(\sum_k y_{jk}A_{jk}\right) - \log\left(\sum_c e^{A_{jc}}\right)\right]$$

$$= \frac{1}{N}\left[y_{jm} - \frac{\partial}{\partial A_{jm}}\log\left(\sum_c e^{A_{jc}}\right)\right]$$

$$\frac{\partial}{\partial A_{jm}}\log\left(e^{A_{j1}} + e^{A_{j2}} + \dots + e^{A_{jK}}\right)$$

$$\frac{1}{\sum_c e^{A_{jc}}}\cdot e^{A_{jm}} = \frac{e^{A_{jm}}}{\sum_c e^{A_{jc}}}$$

$$= \frac{1}{N}\left[y_{jm} - \frac{e^{A_{jm}}}{\sum_c e^{A_{jc}}}\right]$$

$$\boxed{= \frac{1}{N}\left[y_{jm} - P_{jm}\right]}$$

$$d\_logits = \frac{1}{N}(Y - P)$$

$$= \frac{1}{N}\cdot(\text{labels} - \text{probas})$$