Lecture 4: BackPropagation (BackP. Makemore Manually)

X: #182625 elements of size 3
training

[0, 0, 0], [. . .]
[0, 0, 25], [. . Y]

⋮

↓ construct minibatch of size 32

Xb : #32

↓ embedding

dimensionality (10)

C : ⋮

vocab size (27)

⋮

emb

↓ Linear layer (= emb . W1 + b1)

hpreBn

↓ Batch normalization (= hpreBn . bngain + bnbias)

hpreAct

↓ tanh (non linearity)

h

Lecture 4: BackPropagation (BackP. Makemore Manually)

X: #182625 elements of size 3
training
[0, 0, 0], [. . .]
[0, 0, 25], [. . Y]

↓ construct minibatch of size 32

Xb : #32

↓ embedding

C : ---- dimensionality (10)

vocab size (27)

emb

↓ Linear layer (= emb . W1 + b1)

hprebn

↓ Batch normalization (= hprebn . bngain + bnbias)

hpreact

↓ tanh (non linearity)

h

h

↓ Linear layer 2 ($= h \cdot W2 + b2$)

logits

↓ Numerical stability ($logits - max(logits)$)

norm_logits

↓ Softmax (normalize = $\frac{exp(logits)}{sum(counts)}$)

probs

↓ log(probs)

loss

---

dloss

dlogprobs

How does logprobs influence the loss?

loss = -logprobs[range(n), Yb].mean()

where n = batch_size = 32

Yb = array of correct indices

logprobs.size = (32, 27)

iterating down the rows of logprobs with range(n), we pluck out the index at the column specified by Yb

$Yb = [8, 14, 15 \ldots]$

In 0^{th} row, we take 8^{th} column

1^{st} row, " 14^{th} "

logprobs [range(n), Yb] contains the log probabilities of the correct next character in a sequence. Size = 32 (batch_size)

Yb (answer) = 27

logprobs	[1]	[2]	
	0'1	(0'3)	0'2 ...
↑	(0'3)	0'1	0'17 ...
to	0'05	0'6	0'24 0
↓ (= 32)			0
		0	

loss = - logprobs[0, 6], logprobs[1, 17], logprobs[2, 4] .... mean()

Size of logprobs and dlogprobs are the same an equal to [32, 27]

logits → [32, 27]

logit_maxes → [32, 1]

logits = h @ W2 + b2

[32, 64] [64, 27] [27]

Due to broadcasting in Python, b2 will replicate and become a [32, 27] (previously it was considered as [1, 27]

Toy example

$$d = a \cdot b + c$$

[3×4] [4×2] [2]

$$= \begin{bmatrix} a_{11} & a_{12} & a_{13} & a_{14} \\ a_{21} & a_{22} & a_{23} & a_{24} \\ a_{31} & a_{32} & a_{33} & a_{34} \end{bmatrix} \cdot \begin{bmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \\ b_{31} & b_{32} \\ b_{41} & b_{42} \end{bmatrix} + \begin{bmatrix} c_1 \\ c_2 \end{bmatrix}$$

$$= \begin{bmatrix} a_{11} \cdot b_{11} + a_{12} \cdot b_{21} + a_{13} \cdot b_{31} + a_{14} \cdot b_{41} & a_{11} \cdot b_{12} + \dots \\ a_{21} \cdot b_{11} + a_{22} \cdot b_{21} + a_{23} \cdot b_{31} \dots & a_{21} \cdot b_{12} + \dots \\ a_{31} \cdot b_{11} + a_{32} \cdot b_{21} + \dots & a_{31} \cdot b_{12} + \dots \end{bmatrix}$$

[3×2]

$$= \begin{bmatrix} d_{11} & d_{12} \\ d_{21} & d_{22} \\ d_{31} & d_{32} \end{bmatrix}$$

$$d_{11} = a_{11} \cdot b_{11} + a_{12} \cdot b_{21} + a_{13} \cdot b_{31} + a_{14} \cdot b_{41} + c_1$$

$$d_{12} = a_{11} \cdot b_{12} + a_{12} \cdot b_{22} + a_{13} \cdot b_{32} + a_{14} \cdot b_{42} + c_2$$

$$d_{21} = a_{21} \cdot b_{11} + a_{22} \cdot b_{21} + a_{23} \cdot b_{31} + a_{24} \cdot b_{41} + c_1$$

$$d_{22} = a_{21} \cdot b_{12} + a_{22} \cdot b_{22} + a_{23} \cdot b_{32} + a_{24} \cdot b_{42} + c_2$$

$$d_{31} = a_{31} \cdot b_{11} + a_{32} \cdot b_{21} + a_{33} \cdot b_{31} + a_{34} \cdot b_{41} + c_1$$

$$d_{32} = a_{31} \cdot b_{12} + a_{32} \cdot b_{22} + a_{33} \cdot b_{32} + a_{34} \cdot b_{42} + c_2$$

$$\frac{\partial L}{\partial a_{11}} = \frac{\partial L}{\partial d_{11}} \cdot \underbrace{\left(\frac{\partial d_{11}}{\partial a_{11}}\right)}_{b_{11}} + \frac{\partial L}{\partial d_{12}} \cdot \underbrace{\frac{\partial d_{12}}{\partial a_{11}}}_{b_{12}}$$

$$\frac{\partial L}{\partial a_{12}} = \frac{\partial L}{\partial d_{11}} \cdot b_{21} + \frac{\partial L}{\partial d_{12}} \cdot b_{22}$$

$$\frac{\partial L}{\partial a_{13}} = \frac{\partial L}{\partial d_{11}} \cdot b_{31} + \frac{\partial L}{\partial d_{12}} \cdot b_{32}$$

$$\frac{\partial L}{\partial a_{14}} = \frac{\partial L}{\partial d_{11}} \cdot b_{41} + \frac{\partial L}{\partial d_{12}} \cdot b_{42}$$

$$\frac{\partial L}{\partial a_{21}} = \frac{\partial L}{\partial d_{21}} \cdot b_{11} + \frac{\partial L}{\partial d_{22}} \cdot b_{12}$$

$$\frac{\partial L}{\partial a_{22}} = \frac{\partial L}{\partial d_{21}} \cdot b_{21} + \frac{\partial L}{\partial d_{22}} \cdot b_{22}$$

$$\frac{\partial L}{\partial a_{23}} = \frac{\partial L}{\partial d_{21}} \cdot b_{31} + \frac{\partial L}{\partial d_{22}} \cdot b_{32}$$

$$\frac{\partial L}{\partial a_{24}} = \frac{\partial L}{\partial d_{21}} \cdot b_{41} + \frac{\partial L}{\partial d_{22}} \cdot b_{42}$$

$$\frac{\partial L}{\partial a_{31}} = \frac{\partial L}{\partial d_{31}} \cdot b_{11} + \frac{\partial L}{\partial d_{32}} \cdot b_{12}$$

$$\frac{\partial L}{\partial a_{32}} = \frac{\partial L}{\partial d_{31}} \cdot b_{21} + \frac{\partial L}{\partial d_{32}} \cdot b_{22}$$

$$\frac{\partial L}{\partial a_{33}} = \frac{\partial L}{\partial d_{31}} \cdot b_{31} + \frac{\partial L}{\partial d_{32}} \cdot b_{32}$$

$$\frac{\partial L}{\partial a_{34}} = \frac{\partial L}{\partial d_{31}} \cdot b_{41} + \frac{\partial L}{\partial d_{32}} \cdot b_{42}$$

$$\frac{\partial L}{\partial a} = \begin{bmatrix} \frac{\partial L}{\partial a_{11}} & \frac{\partial L}{\partial a_{12}} & \frac{\partial L}{\partial a_{13}} & \frac{\partial L}{\partial a_{14}} \\ \frac{\partial L}{\partial a_{21}} & \frac{\partial L}{\partial a_{22}} & \frac{\partial L}{\partial a_{23}} & \frac{\partial L}{\partial a_{24}} \\ \frac{\partial L}{\partial a_{31}} & \frac{\partial L}{\partial a_{32}} & \frac{\partial L}{\partial a_{33}} & \frac{\partial L}{\partial a_{34}} \end{bmatrix}$$

$$= \begin{bmatrix} \frac{\partial L}{\partial d_{11}} & \frac{\partial L}{\partial d_{12}} \\ \frac{\partial L}{\partial d_{21}} & \frac{\partial L}{\partial d_{22}} \\ \frac{\partial L}{\partial d_{31}} & \frac{\partial L}{\partial d_{32}} \end{bmatrix} \cdot \begin{bmatrix} b_{11} & b_{21} & b_{31} & b_{41} \\ b_{12} & b_{22} & b_{32} & b_{42} \end{bmatrix}$$

$$= \frac{\partial L}{\partial d} @ b^T$$

$$\frac{\partial L}{\partial b} = \cdots = a^T @ \frac{\partial L}{\partial d}$$

$$\frac{\partial L}{\partial c_1} = \frac{\partial L}{\partial d_{11}} \cdot 1 + \frac{\partial L}{\partial d_{21}} \cdot 1 + \frac{\partial L}{\partial d_{31}} \cdot 1$$

$$\frac{\partial L}{\partial c_2} = \frac{\partial L}{\partial d_{12}} \cdot 1 + \frac{\partial L}{\partial d_{22}} \cdot 1 + \frac{\partial L}{\partial d_{32}} \cdot 1$$

$$\frac{\partial L}{\partial c} = \begin{bmatrix} \frac{\partial L}{\partial c_1} \\ \frac{\partial L}{\partial c_2} \end{bmatrix} = \begin{bmatrix} \frac{\partial L}{\partial d_{11}} + \frac{\partial L}{\partial d_{21}} + \frac{\partial L}{\partial d_{31}} \\ \frac{\partial L}{\partial d_{12}} + \frac{\partial L}{\partial d_{22}} + \frac{\partial L}{\partial d_{32}} \end{bmatrix}$$

$$= \frac{\partial L}{\partial d} \cdot \text{sum}(0)$$

sum across columns of $\frac{\partial L}{\partial d}$

Backpropagate through line

emb = C[Xb], where

Xb = batch containing 32 examples, size [32x3]

C = matrix 27x10 embedding

emb = C[Xb] → size [32 x 3 x 10]

embcat = emb.view(emb.shape[0], -1) concatenate

32

embcat.shape = 32 x 30

↑ the shape of -1 is inferred from other dimensions

emb: 32 examples, 3 characters and each of them has 10 dimensional embedding

This was achieved by looking at the lookup table C (27 possible characters). We look at the rows specified by tensor Xb

10-dim

Xb = [1, 1, 4],
[18, 14, 1],
[0, 3, 1],
⋮

These integers specify which row of C we want to pluck out

emb = C[[1, 1, 4]], (3 x 10)
C[[18, 14, 1]], (3 x 10)
⋮

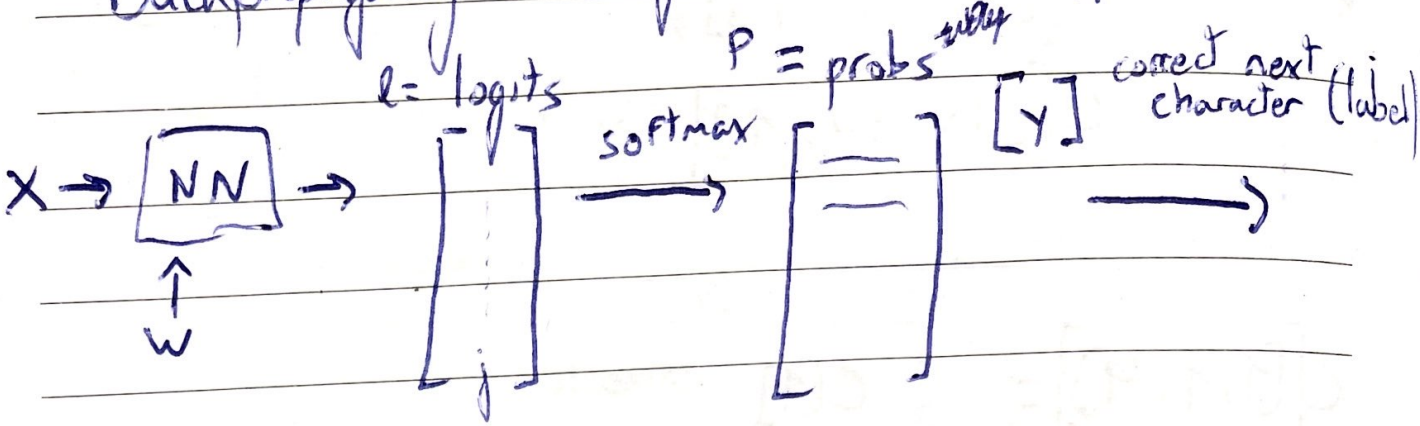
C[[1, 1, 4]] = [C[1], (size 10)
(3 x 10) C[1], (size 10)
C[4]]

demb contains the gradients of the plucked rows

which row of C did the 10-dimensional embeddings of emb come from? Deposit this into dC

Some rows of C are used multiple times, so the gradient have to be added

Backpropagating through cross entropy function



→ probs [y] $\xrightarrow{-\log}$ $-\log$ probs [y] $\xrightarrow{\text{avg}}$ loss

$$\text{loss} = -\log p_y$$

$$= -\log \frac{e^{l_y}}{\sum_j e^{l_j}}$$

$$p_i = \frac{e^{l_i}}{\sum_j e^{l_j}} \quad \text{(softmax)}$$

$\frac{d\,\text{loss}}{d\,l_i}$ ≡ derivative of the loss w.r.t. the ith logit

$$= \frac{d}{dl_i}\left[-\log \frac{e^{l_y}}{\sum_j e^{l_j}}\right]$$

l_y indexed with specific label y

$\frac{d}{dx}\log x = \frac{1}{x}$

$$= -\frac{\sum_j e^{l_j}}{e^{l_y}} \cdot \frac{d}{dl_i}\left[\frac{e^{l_y}}{\sum_j e^{l_j}}\right]$$

product rule → $(u \cdot v)' = u' \cdot v + u \cdot v'$

power rule → $(x^n)' = n\,x^{n-1}$

if $i \neq y$

$u = e^{l_y}$ $\quad v = \frac{1}{\sum_j e^{l_j}}$

$$\frac{d\text{loss}}{dl_i} = \frac{-\sum_j e^{l_j}}{e^{l_y}} \left[\underbrace{0}_{u'} \cdot \underbrace{\frac{1}{\sum_j e^{l_j}}}_{v} + \underbrace{-}_{u} \cdot \underbrace{\frac{e^{l_y} \cdot e^{l_i}}{\left(\sum_j e^{l_j}\right)^2}}_{v'} \right]$$

since $\frac{du}{dl_i} = 0$

$$\frac{dv}{dl_i} = \frac{d}{dl_i}\left(\frac{1}{\sum_j e^{l_j}}\right)$$

$$= \frac{d}{dl_i}\left(\frac{1}{e^{l_1} + e^{l_2} + \ldots + e^{l_i} + \ldots e^{l_j}}\right)$$

$$\left(\text{recall } \frac{d}{dx}\left(\frac{1}{x}\right) = -\frac{1}{x^2}\right)$$

$$= -\frac{1}{\left(\sum_j e^{l_j}\right)^2} \cdot \frac{d}{dl_i}\left(\sum_j e^{l_j}\right)$$

$$\left(\text{recall } \frac{d}{dx} e^x = e^x\right)$$

$$= \frac{-1}{\left(\sum_j e^{l_j}\right)^2} \cdot \frac{d}{dl_i}\left(0 + 0 + \ldots e^{l_i} + 0 + 0\right)$$

$$= -\frac{e^{l_i}}{\left(\sum_j e^{l_j}\right)^2}$$

if $i \neq y$

$$\frac{d\text{loss}}{dl_i} = -\frac{\sum_j e^{l_j}}{e^{l_y}} \left[0 \cdot \frac{1}{\sum_j e^{l_j}} + -\frac{e^{l_y} \cdot e^{l_i}}{\left(\sum_j e^{l_j}\right)^2} \right]$$

$$= + \frac{\left(\sum_j e^{l_j}\right) \cdot e^{l_y} \cdot e^{l_i}}{e^{l_y} \cdot \left(\sum_j e^{l_j}\right)^2} = \frac{e^{l_i}}{\sum_j e^{l_j}} = \boxed{P_i}$$

softmax

if $i = y$

$u' \cdot v$ + $u \cdot v'$

$$\frac{d\text{loss}}{dl_i} = -\frac{\sum_j e^{l_j}}{e^{l_y}} \left[\frac{e^{l_y}}{\sum_j e^{l_j}} + -\frac{e^{l_y} \cdot e^{l_i}}{\left(\sum_j e^{l_j}\right)^2} \right]$$

$$\left(\text{since } \frac{du}{dl_i} = e^{l_y} \right)$$

$$= -\frac{\sum_j e^{l_j}}{e^{l_y}} \left[\frac{\sum_j e^{l_j} \cdot e^{l_y} + -e^{l_y} \cdot e^{l_i}}{\left(\sum_j e^{l_j}\right)^2} \right]$$

$$= -\frac{\sum_j e^{l_j} - e^{l_i}}{\sum_j e^{l_j}} = \frac{e^{l_i}}{\sum_j e^{l_j}} - 1 = \boxed{P_i - 1}$$

Backpropagating through batch Normalization in a single line

Calculate ~~$\frac{dh_{preact}}{dh_{prebn}}$~~ $\frac{dh_{prebn}}{dh_{preact}}$ given dh_{preact}

$$h_{prebn} = x_i \qquad h_{preact} = y_i$$

batch Norm paper

$$\mu = \frac{1}{m}\sum_{i=1}^{m} x_i$$

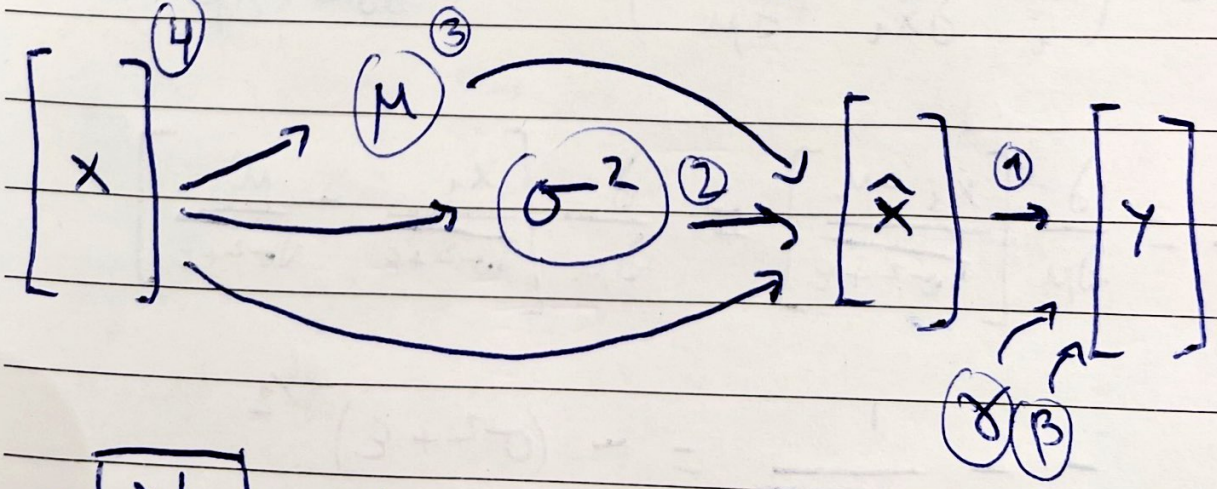
$$\sigma^2 = \frac{1}{m-1}\sum_{i=1}^{m}(x_i - \mu)^2$$

Bessel's correction

$$\hat{x}_i = \frac{x_i - \mu}{\sqrt{\sigma^2 + \varepsilon}}$$

$$y_i = \gamma \hat{x}_i + \beta$$

we have $\frac{\partial Loss}{\partial y_i}$; we want $\frac{\partial Loss}{\partial x_i}$



① $$\frac{\partial L}{\partial \hat{x}_i} = \frac{\partial L}{\partial y_i} \cdot \gamma$$

② $\frac{\partial L}{\partial \sigma^2} = \sum_i \frac{\partial L}{\partial \hat{x}_i} \frac{\partial \hat{x}_i}{\partial \sigma^2}$ sigma^2 propagates through many $\hat{x}_i$

$$= \sum_i \frac{\partial L}{\partial y_i} \cdot \gamma \cdot \frac{\partial}{\partial \sigma^2} \frac{x_i - \mu}{\sqrt{\sigma^2 + \varepsilon}}$$

$$= \gamma \cdot \sum_i \frac{\partial}{\partial \sigma^2} \left[(x_i - \mu) \cdot (\sigma^2 + \varepsilon)^{-1/2} \right] \cdot \frac{\partial L}{\partial y_i}$$

$$= \gamma \cdot \sum_i \frac{-1}{2} \cdot (x_i - \mu) (\sigma^2 + \varepsilon)^{-3/2} \cdot \frac{\partial L}{\partial y_i}$$

$$\boxed{\frac{\partial L}{\partial \sigma^2}} = -\frac{1}{2} \gamma \sum_i \frac{dL}{\partial y_i} (x_i - \mu) (\sigma^2 + \varepsilon)^{-3/2}$$

③ $\frac{\partial L}{\partial \mu}$? Multiple μ influence $\hat{x}_i$
Only one single μ influences σ^2
} #33 μ (batch+1)

$$\frac{\partial L}{\partial \mu} = \left(\sum_i \frac{\partial L}{\partial \hat{x}_i} \frac{\partial \hat{x}_i}{\partial \mu} \right) + \left(\frac{\partial L}{\partial \sigma^2} \frac{\partial \sigma^2}{\partial \mu} \right)$$

$$\frac{\partial \hat{x}_i}{\partial \mu} = \frac{\partial}{\partial \mu} \left[\frac{x_i - \mu}{\sqrt{\sigma^2 + \varepsilon}} \right] = \frac{\partial}{\partial \mu} \left[\frac{x_i}{\sqrt{\sigma^2 + \varepsilon}} - \frac{\mu}{\sqrt{\sigma^2 + \varepsilon}} \right]$$

$$= -\frac{1}{\sqrt{\sigma^2 + \varepsilon}} = -(\sigma^2 + \varepsilon)^{-1/2}$$

$$\frac{\partial \sigma^2}{\partial \mu} = \frac{\partial}{\partial \mu}\left[\frac{1}{m-1}\sum_{i=1}^{m}(x_i-\mu)^2\right]$$

$$= \frac{2}{m-1}\sum_{i=1}^{m}(x_i-\mu)\cdot -1$$

$$= \frac{-2}{m-1}\sum x_i - \sum \mu$$

$$= \frac{-2}{m-1}\left(\underbrace{m\mu} - m\mu\right); \text{ since } \mu = \frac{1}{m}\sum x_i$$

$$= \frac{-2}{m-1}\cdot 0 = 0$$

in BatchNorm it holds that μ = average

$$\boxed{\frac{\partial L}{\partial \mu} = \sum_i \frac{\partial L}{\partial y_i}\cdot \gamma \cdot -(\sigma^2+\varepsilon)^{-1/2} + 0}$$

$\left(\frac{\partial L}{\partial \sigma^2}\frac{\partial \sigma^2}{\partial \mu} = 0\right)$

(4) ~~x is contributed~~

The gradient of x is the sum of 3 parts: gradient coming from μ, from σ^2 and from a single x_i. $\hat{x}_i = \frac{x_i-\mu}{\sqrt{\sigma^2+\varepsilon}}$ There is a bijective relation between x_i and $\hat{x}_i$, despite x and $\hat{x}$ being 32 ~~size~~ sized vectors (batch size)

(4) $$\frac{\partial L}{\partial x_i} = \frac{\partial L}{\partial \hat{x}_i}\underbrace{\frac{\partial \hat{x}_i}{\partial x_i}}_{?} + \frac{\partial L}{\partial \mu}\underbrace{\frac{\partial \mu}{\partial x_i}}_{?} + \frac{\partial L}{\partial \sigma^2}\underbrace{\frac{\partial \sigma^2}{\partial x_i}}_{?}$$

$$\frac{\partial \hat{x}_i}{\partial x_i} = \frac{\partial}{\partial x_i}\left[\frac{x_i - \mu}{\sqrt{\sigma^2+\varepsilon}}\right] = \frac{\partial}{\partial x_i}\left[\frac{x_i}{\sqrt{\sigma^2+\varepsilon}} - \frac{\mu}{\sqrt{\sigma^2+\varepsilon}}\right]$$

$$= \frac{1}{\sqrt{\sigma^2+\varepsilon}}$$

$$\frac{\partial \mu}{\partial x_i} = \frac{\partial}{\partial x_i}\left[\frac{1}{m}\sum x_i\right] = \frac{1}{m}$$

$$\frac{\partial \sigma^2}{\partial x_i} = \frac{\partial}{\partial x_i}\left[\frac{1}{m-1}\sum (x_i - \mu)^2\right] = \frac{2}{m-1}(x_i - \mu)\cdot 1$$

Join

~~Add~~ everything

~~Wrong~~

$$\frac{\partial L}{\partial x_i} = \left(\frac{\partial L}{\partial y_i} \cdot \gamma\right) \cdot (\sigma^2+\varepsilon)^{-1/2} +$$

$$+ \left(-\sum_j \frac{\partial L}{\partial y_i} \cdot \gamma \cdot (\sigma^2+\varepsilon)^{-1/2}\right) \cdot \frac{1}{m} +$$

$$+ \left(\frac{-1}{2} \gamma \sum_j \frac{\partial L}{\partial y_j} (x_j-\mu)(\sigma^2+\varepsilon)^{-3/2}\right) \cdot \frac{2}{m-1}(x_i-\mu)$$

$$= (\sigma^2+\varepsilon)^{-1/2}\left(\frac{\partial L}{\partial y_i}\gamma\right) + (\sigma^2+\varepsilon)^{-1/2}\left(\frac{-1}{m}\sum_j \frac{\partial L}{\partial y_j}\cdot\gamma\right) +$$

$$+ (\sigma^2+\varepsilon)^{-1/2}\left(-\gamma \sum_j \frac{\partial L}{\partial y_i} \frac{(x_j-\mu)}{\sqrt{\sigma^2+\varepsilon}}\right) \cdot \left(\frac{1}{m-1}\frac{x_i-\mu}{\sqrt{\sigma^2+\varepsilon}}\right)$$

$$\left(\text{recall } \hat{x}_i = \frac{x_i-\mu}{\sqrt{\sigma^2+\varepsilon}}\right)$$

$$= \frac{(\sigma^2+\varepsilon)^{-1/2}}{m}\left[m\gamma\frac{\partial L}{\partial y_i} - \sum_j \frac{\partial L}{\partial y_j}\gamma - \gamma\left(\sum_j \frac{\partial L}{\partial y_j}\hat{x}_j\right)\left(\frac{m}{m-1}\hat{x}_i\right)\right]$$

$$= \frac{\gamma(\sigma^2+\varepsilon)^{-1/2}}{m}\left[m\frac{\partial L}{\partial y_i} - \sum_j \frac{\partial L}{\partial y_i} - \frac{m}{m-1}\hat{x}_i \sum_j \frac{\partial L}{\partial y_j}\hat{x}_j\right]$$

already have them from previous steps in backprop.

come from forward pass