Multi-Class Classification
Solution

• Traditional Method: 1-vs-other method
  • Too slow. If we have n-classes, we need to train n models
  • Performance is not great, because the sample size is different for positive and negative classes

• Multiple Neurons
  • Use n output neuron to correspond n classes.
  • Easy, fast, and robust
  • Problem: how to model the probability? The values in the neural network can be negative or greater than 1.
Softmax: normalized exponential

Input: vector of reals
Output: probability distribution

softmax([1,2,7,3,2]):

Calculate $e^x$: [2.72, 7.39, 1096.63, 20.09, 7.39]
Calculate sum(e^x): 2.72+7.39+1096.63+20.09+7.39 = 1134.22
Normalize: $e^x$/sum(e^x) = [0.002, 0.007, 0.967, 0.017, 0.007]

Result is a vector of reals.
A Simple Example

Here, we will go over a simple 2-layer neural network (no bias).
Mini-batch for Machine Learning

• Let’s say we have $S$ images of size 28x28
• We use a matrix to represent data.

When $s$ is very large RAM overload
• In HW4, we will use a batch size of 128
Neural Network Easy Example

Input:

\[
X_{in} = \begin{bmatrix}
3 & 2 & 4 \\
\cdot & \cdot & \cdot \\
\cdot & \cdot & \cdot
\end{bmatrix}
\]

First pixel: 3
Second pixel: 2
Third pixel: 4

First Batch

Bsz = 3
Neural Network Easy Example

First pixel 3
Second pixel 2
Third pixel 4
Input

Output Layer with Softmax

A
B
C
D
Neural Network Easy Example

First pixel
- 3

Second pixel
- 2

Third pixel
- 4

Input
- 3
- 2
- 4

$X_{in}$

A

B

1-st Layer (ReLU)

$w_1$

\[
\begin{pmatrix}
1 & 0.5 \\
0.1 & 1 \\
-2.3 & -0.5
\end{pmatrix}
\]

Graph showing the ReLU function $f(x)$
Neural Network Easy Example

First pixel
3

Second pixel
2

Third pixel
4

Input
3 2 4

$X_{in}$ . . .

1-st Layer (ReLU)

$w_1$

Output Layer with Softmax

$w_2$
First pixel

Second pixel

Third pixel

\[ \text{Output}_A = 3 \times 1 + 2 \times 0.1 + 4 \times (-2.3) = -6.0 \]

\[ \text{Output}_A = 0 \text{ (after RELU)} \]
Neural Network Easy Example

Input: $X_{in} = \begin{bmatrix} 3 & 2 & 4 \end{bmatrix}$

$A = 0$

$B = 3 \times 0.5 + 2 \times 1 + 4 \times -0.5 = 1.5$

1-st Layer (ReLU):

$w_1 = \begin{bmatrix} 1 & 0.5 \\ 0.1 & 1 \\ -2.3 & -0.5 \end{bmatrix}$

$o_1 = \begin{bmatrix} 0 \\\ 1.5 \\\ \cdot \\\ \cdot \end{bmatrix}$
Neural Network Easy Example

\[ \text{output}_C = \frac{e^{0.15}}{e^{0.15} + e^{-0.3}} \approx \frac{1.16}{1.16 + 0.74} = 0.61 \]

\[ \text{output}_D = \frac{e^{-0.3}}{e^{0.15} + e^{-0.3}} \approx \frac{0.74}{1.16 + 0.74} = 0.39 \]
Neural Network Easy Example

Input

1-st Layer (ReLU)

Output Layer with Softmax

0.61 0.39

0.15

A

C

B

D

3

2

4

0.1

-2.3

1

0.5

1.5

-0.5

1

-2.1

-0.2

1.5

-0.3

0.7

0.1

0.1
Neural Network Easy Example

First pixel

Second pixel

Third pixel

Input

$X_{in}$

$3$  $2$  $4$

$\cdot$  $\cdot$  $\cdot$

$w_1$

$0.1$  $1$  $0.5$

$1$  $0.1$  $-2.3$

$1$  $0.5$  $-0.5$

$1$  $0.5$  $1.5$

$1$  $0.1$  $-2.3$

$1$  $0.1$  $-0.5$

$w_2$

$0.7$  $-2.1$

$0.1$  $-0.2$

$0.61$  $0.39$

Output Layer with Softmax
Neural Network Easy Example

Input:

$X_{in} = [3, 2, 4, \ldots, \ldots, \ldots]$  

1-st Layer (ReLU):

$w_1 = \begin{bmatrix} 1 & 0.1 & 0.5 & -2.3 & -0.5 \end{bmatrix}$,  
\begin{align*}
o_1 &= \begin{bmatrix} 0 & 1.5 \end{bmatrix}
\end{align*}$

Output Layer with Softmax:

$w_2 = \begin{bmatrix} 0.7 & -2.1 & 0.1 & -0.2 \end{bmatrix}$,  
\begin{align*}
o_2 &= \begin{bmatrix} 0.61 & 0.39 \end{bmatrix}
\end{align*}$

Ground Truth:

$\Delta = 1$,  
$\Delta o_2 = 0$
Neural Network Easy Example

\[ \Delta = 1 - 0.61 = 0.39 \]

\[ \Delta = 0 - 0.39 = -0.39 \]
What to do with this error?
What to do with this error?

Backpropagate it to update the weights!

How?
A Computational Graph Perspective

\[ f(x, y, z) = (x + y)z \]
A Computational Graph Perspective

Backpropagation: a simple example

\[ f(x, y, z) = (x + y)z \]
A Computational Graph Perspective

Backpropagation: a simple example

\[ f(x, y, z) = (x + y)z \]

e.g. \( x = -2, y = 5, z = -4 \)

Figure taken from Ranjay Krishna’s course CSE493
A Computational Graph Perspective

Backpropagation: a simple example

\[ f(x, y, z) = (x + y)z \]

e.g. \( x = -2, y = 5, z = -4 \)

\[ q = x + y \quad \frac{\partial q}{\partial x} = 1, \quad \frac{\partial q}{\partial y} = 1 \]

Figure taken from Ranjay Krishna’s course CSE493
A Computational Graph Perspective

Backpropagation: a simple example

\[ f(x, y, z) = (x + y)z \]

e.g. \( x = -2, y = 5, z = -4 \)

\[
q = x + y \quad \frac{\partial q}{\partial x} = 1, \quad \frac{\partial q}{\partial y} = 1 
\]

\[
f = qz \quad \frac{\partial f}{\partial q} = z, \quad \frac{\partial f}{\partial z} = q
\]

Figure taken from Ranjay Krishna’s course CSE493
A Computational Graph Perspective

Backpropagation: a simple example

\[ f(x, y, z) = (x + y)z \]

e.g. \( x = -2, \ y = 5, \ z = -4 \)

\[ q = x + y \quad \frac{\partial q}{\partial x} = 1, \quad \frac{\partial q}{\partial y} = 1 \]

\[ f = qz \quad \frac{\partial f}{\partial q} = z, \quad \frac{\partial f}{\partial z} = q \]

Want: \( \frac{\partial f}{\partial x}, \frac{\partial f}{\partial y}, \frac{\partial f}{\partial z} \)

Figure taken from Ranjay Krishna’s course CSE493
A Computational Graph Perspective

Backpropagation: a simple example

\[ f(x, y, z) = (x + y)z \]

e.g. \( x = -2, y = 5, z = -4 \)

\[ q = x + y \]
\[ \frac{\partial q}{\partial x} = 1, \quad \frac{\partial q}{\partial y} = 1 \]

\[ f = qz \]
\[ \frac{\partial f}{\partial q} = z, \quad \frac{\partial f}{\partial z} = q \]

Want:
\[ \frac{\partial f}{\partial x}, \quad \frac{\partial f}{\partial y}, \quad \frac{\partial f}{\partial z} \]

Figure taken from Ranjay Krishna’s course CSE493
A Computational Graph Perspective

Backpropagation: a simple example

\[ f(x, y, z) = (x + y)z \]

e.g. \( x = -2, y = 5, z = -4 \)

\[ q = x + y \quad \frac{\partial q}{\partial x} = 1, \quad \frac{\partial q}{\partial y} = 1 \]

\[ f = qz \quad \frac{\partial f}{\partial q} = z, \quad \frac{\partial f}{\partial z} = q \]

Want: \( \frac{\partial f}{\partial x}, \frac{\partial f}{\partial y}, \frac{\partial f}{\partial z} \)

Figure taken from Ranjay Krishna’s course CSE493
A Computational Graph Perspective

Backpropagation: a simple example

\[ f(x, y, z) = (x + y)z \]

e.g. \( x = -2, y = 5, z = -4 \)

\[ q = x + y \quad \frac{\partial q}{\partial x} = 1, \quad \frac{\partial q}{\partial y} = 1 \]

\[ f = qz \quad \frac{\partial f}{\partial q} = z, \quad \frac{\partial f}{\partial z} = q \]

Want: \( \frac{\partial f}{\partial x}, \frac{\partial f}{\partial y}, \frac{\partial f}{\partial z} \)

Figure taken from Ranjay Krishna’s course CSE493
A Computational Graph Perspective

Backpropagation: a simple example

\[ f(x, y, z) = (x + y)z \]

e.g. \( x = -2, y = 5, z = -4 \)

\[ q = x + y \quad \frac{\partial q}{\partial x} = 1, \quad \frac{\partial q}{\partial y} = 1 \]

\[ f = qz \quad \frac{\partial f}{\partial q} = z, \quad \frac{\partial f}{\partial z} = q \]

Want: \( \frac{\partial f}{\partial x}, \frac{\partial f}{\partial y}, \frac{\partial f}{\partial z} \)

Figure taken from Ranjay Krishna’s course CSE493
A Computational Graph Perspective

Backpropagation: a simple example

\[ f(x, y, z) = (x + y)z \]

e.g. \( x = -2, y = 5, z = -4 \)

\[ q = x + y \]
\[ \frac{\partial q}{\partial x} = 1, \frac{\partial q}{\partial y} = 1 \]

\[ f = qz \]
\[ \frac{\partial f}{\partial q} = z, \frac{\partial f}{\partial z} = q \]

Want: \[ \frac{\partial f}{\partial x}, \frac{\partial f}{\partial y}, \frac{\partial f}{\partial z} \]

Figure taken from Ranjay Krishna’s course CSE493
A Computational Graph Perspective

Backpropagation: a simple example

\[ f(x, y, z) = (x + y)z \]

e.g. \( x = -2, y = 5, z = -4 \)

\[
q = x + y \quad \frac{\partial q}{\partial x} = 1, \quad \frac{\partial q}{\partial y} = 1
\]

\[ f = qz \quad \frac{\partial f}{\partial q} = z, \quad \frac{\partial f}{\partial z} = q \]

Want: \( \frac{\partial f}{\partial x}, \frac{\partial f}{\partial y}, \frac{\partial f}{\partial z} \)

Figure taken from Ranjay Krishna’s course CSE493
A Computational Graph Perspective

Backpropagation: a simple example

\[ f(x, y, z) = (x + y)z \]

e.g. \( x = -2, y = 5, z = -4 \)

\[ q = x + y \quad \frac{\partial q}{\partial x} = 1, \quad \frac{\partial q}{\partial y} = 1 \]

\[ f = qz \quad \frac{\partial f}{\partial q} = z, \quad \frac{\partial f}{\partial z} = q \]

Want: \( \frac{\partial f}{\partial x}, \frac{\partial f}{\partial y}, \frac{\partial f}{\partial z} \)

Chain rule:

\[ \frac{\partial f}{\partial y} = \frac{\partial f}{\partial q} \frac{\partial q}{\partial y} \]

Figure taken from Ranjay Krishna’s course CSE493
A Computational Graph Perspective

Backpropagation: a simple example

\[ f(x, y, z) = (x + y)z \]

e.g. \( x = -2, \ y = 5, \ z = -4 \)

\[ q = x + y \quad \frac{\partial q}{\partial x} = 1, \ \frac{\partial q}{\partial y} = 1 \]

\[ f = qz \quad \frac{\partial f}{\partial q} = z, \ \frac{\partial f}{\partial z} = q \]

Want: \( \frac{\partial f}{\partial x}, \ \frac{\partial f}{\partial y}, \ \frac{\partial f}{\partial z} \)

Figure taken from Ranjay Krishna’s course CSE493
A Computational Graph Perspective

Backpropagation: a simple example

\[ f(x, y, z) = (x + y)z \]

e.g. \( x = -2, y = 5, z = -4 \)

\[ q = x + y \quad \frac{\partial q}{\partial x} = 1, \quad \frac{\partial q}{\partial y} = 1 \]

\[ f = qz \quad \frac{\partial f}{\partial q} = z, \quad \frac{\partial f}{\partial z} = q \]

Want: \( \frac{\partial f}{\partial x}, \frac{\partial f}{\partial y}, \frac{\partial f}{\partial z} \)

Chain rule:

\[ \frac{\partial f}{\partial x} = \frac{\partial f}{\partial q} \frac{\partial q}{\partial x} \]

Figure taken from Ranjay Krishna’s course CSE493
A Computational Graph Perspective

Backpropagation: a simple example

\[ f(x, y, z) = (x + y)z \]

e.g. \( x = -2, y = 5, z = -4 \)

\[ q = x + y \quad \frac{\partial q}{\partial x} = 1, \quad \frac{\partial q}{\partial y} = 1 \]

\[ f = qz \quad \frac{\partial f}{\partial q} = z, \quad \frac{\partial f}{\partial z} = q \]

Want: \( \frac{\partial f}{\partial x}, \frac{\partial f}{\partial y}, \frac{\partial f}{\partial z} \)

Chain rule:

\[ \frac{\partial f}{\partial x} = \frac{\partial f}{\partial q} \cdot \frac{\partial q}{\partial x} \]

Figure taken from Ranjay Krishna’s course CSE493
One more example

\[ f(w, x) = \frac{1}{1 + e^{-(w_0 x_0 + w_1 x_1 + w_2)}} \]

Figure taken from Ranjay Krishna’s course CSE493
One more example

Figure taken from Ranjay Krishna’s course CSE493
One more example

\[ f(x) = e^x \quad \rightarrow \quad \frac{df}{dx} = e^x \]
\[ f_a(x) = ax \quad \rightarrow \quad \frac{df}{dx} = a \]
\[ f_c(x) = c + x \quad \rightarrow \quad \frac{df}{dx} = 1 \]

Figure taken from Ranjay Krishna’s course CSE493
One more example

\[ f(x) = e^x \quad \rightarrow \quad \frac{df}{dx} = e^x \]
\[ f_a(x) = ax \quad \rightarrow \quad \frac{df}{dx} = a \]
\[ f_c(x) = c + x \quad \rightarrow \quad \frac{df}{dx} = 1 \]

Figure taken from Ranjay Krishna’s course CSE493
One more example

\[ f(x) = e^x \quad \rightarrow \quad \frac{df}{dx} = e^x \]

\[ f_a(x) = ax \quad \rightarrow \quad \frac{df}{dx} = a \]

\[ f(x) = \frac{1}{x} \quad \rightarrow \quad \frac{df}{dx} = -\frac{1}{x^2} \]

\[ f_c(x) = c + x \quad \rightarrow \quad \frac{df}{dx} = 1 \]
One more example

\[ f(x) = e^x \quad \rightarrow \quad \frac{df}{dx} = e^x \]

\[ f_a(x) = ax \quad \rightarrow \quad \frac{df}{dx} = a \]

\[ f_c(x) = c + x \quad \rightarrow \quad \frac{df}{dx} = 1 \]

\[ \left( -0.53 \right) \left( 1 \right) = -0.53 \]

Figure taken from Ranjay Krishna’s course CSE493
One more example

$f(x) = e^x \quad \rightarrow \quad \frac{df}{dx} = e^x$

$f_a(x) = ax \quad \rightarrow \quad \frac{df}{dx} = a$

$f(x) = \frac{1}{x} \quad \rightarrow \quad \frac{df}{dx} = -\frac{1}{x^2}$

$f_c(x) = c + x \quad \rightarrow \quad \frac{df}{dx} = 1$

Figure taken from Ranjay Krishna’s course CSE493
One more example

\[ \text{[upstream gradient]} \times \text{[local gradient]} \]

\[ \begin{align*}
  w0: \ [0.2] \times \ [-1] &= -0.2 \\
  x0: \ [0.2] \times \ [2] &= 0.4
\end{align*} \]

\[
\begin{align*}
  f(x) &= e^x & \rightarrow & & \frac{df}{dx} = e^x \\
  f_a(x) &= ax & \rightarrow & & \frac{df}{dx} = a \\
  f_c(x) &= c + x & \rightarrow & & \frac{df}{dx} = 1
\end{align*}
\]

Figure taken from Ranjay Krishna’s course CSE493
One more example

Another example:

$$f(w, x) = \frac{1}{1 + e^{-(w_0 x_0 + w_1 x_1 + w_2)}}$$

Sigmoid function

$$\sigma(x) = \frac{1}{1 + e^{-x}}$$

Computational graph representation may not be unique. Choose one where local gradients at each node can be easily expressed!

Figure taken from Ranjay Krishna’s course CSE493
One more example

Another example:

\[ f(w, x) = \frac{1}{1 + e^{-(w_0 x_0 + w_1 x_1 + w_2)}} \]

Sigmoid function

\[ \sigma(x) = \frac{1}{1 + e^{-x}} \]

Computational graph representation may not be unique. Choose one where local gradients at each node can be easily expressed!

Sigmoid local gradient:

\[ \frac{d\sigma(x)}{dx} = \frac{e^{-x}}{(1 + e^{-x})^2} = \left( \frac{1 + e^{-x} - 1}{1 + e^{-x}} \right) \left( \frac{1}{1 + e^{-x}} \right) = (1 - \sigma(x)) \sigma(x) \]

Figure taken from Ranjay Krishna’s course CSE493
One more example

Another example:

$$f(w, x) = \frac{1}{1 + e^{-(w_0x_0 + w_1x_1 + w_2)}}$$

Sigmoid function

$$\sigma(x) = \frac{1}{1 + e^{-x}}$$

Computational graph representation may not be unique. Choose one where local gradients at each node can be easily expressed!

Sigmoid local gradient:

$$\frac{d\sigma(x)}{dx} = \frac{e^{-x}}{(1 + e^{-x})^2} = \left(\frac{1 + e^{-x} - 1}{1 + e^{-x}}\right) \left(\frac{1}{1 + e^{-x}}\right) = (1 - \sigma(x))\sigma(x)$$

[upstream gradient] x [local gradient]

$$[1.00] \times [(1 - 0.73)(0.73)] = 0.2$$

Figure taken from Ranjay Krishna’s course CSE493
So far: backprop with scalars, let’s go back to our initial example of vectors
Back to our Neural Network Example

Δ = 1 - 0.61 = 0.39

Δ = 0 - 0.39 = -0.39

Ground Truth

Δ"o_2"
Backpropagation [Cont.]

1-st Layer (ReLU)

Input

3 2 4

\( X_{in} \)

\( w_1 \)

1 0.1 -2.3

Output Layer with Softmax

\( w_2 \)

0.7 -2.1

0.1 -0.2

\( o_1 \)

0 1.5

\( o_2 \)

0.61 0.39

\( g'(o_2) \circ \Delta o_2 \)

\( \Delta w_2 = o_1^T g'(o_2) \circ \Delta o_2 \)

\( \Delta o_1 = g'(o_2) \circ \Delta o_2 w_2^T \)
Backpropagation [Cont.]

\[ g'(o_1) \cdot \Delta o_1 \]

\[ \Delta w_1 = o_0^T g'(o_1) \cdot \Delta o_1 \]

\[ \begin{array}{c|c}
0 & 0.117 \\
. & . \\
. & . \\
. & . \\
\end{array} \]

\[ \begin{array}{c|c}
0 & 0.351 \\
0 & 0.234 \\
0 & 0.468 \\
\end{array} \]

Input

\[ X_{in} \]

\[ 3 \quad 2 \quad 4 \quad . \quad . \quad . \quad . \]

Output Layer with Softmax

\[ g'(o_2) \cdot \Delta o_2 \]

\[ \Delta w_2 = o_1^T g'(o_2) \cdot \Delta o_2 \]

\[ \Delta o_1 = g'(o_2) \cdot \Delta o_2 w_2^T \]

1-st Layer (ReLU)

\[ \begin{array}{c|c|c}
1 & 0.5 \\
0.1 & 1 \\
-2.3 & -0.5 \\
\end{array} \]

\[ \begin{array}{c|c}
0 & 1.5 \\
. & . \\
. & . \\
\end{array} \]

\[ \begin{array}{c|c|c}
0.7 & -2.1 \\
0.1 & -0.2 \\
. & . \\
\end{array} \]

\[ \begin{array}{c|c}
0.61 & 0.39 \\
. & . \\
. & . \\
\end{array} \]
Now we backpropagate with a learning rate of 0.1

\[
w_1 = w_1 + \alpha \Delta w_1
\]

\[
w_2 = w_2 + \alpha \Delta w_2
\]
Think: What will happen if we go forward again?

The final output is closer to the actual label.
Tricks for Training Neural Networks
Problem: Under and Overfitting

Underfitting: model not powerful enough, too much bias

Overfitting: model too powerful, fits to noise, doesn’t generalize well
**Weight decay**: neural network regularization

We want the weights to be close to 0.

Let $L$ be the “loss” function; (e.g. $L = |y - g(in)|$, $L = (y - g(in))^2$, etc.)

$\lambda$ is a regularization parameter (for decay)
   - Higher: more penalty for large weights, less powerful model
   - Lower: less penalty, more overfitting

Before:
   \[
   \Delta w_t = -\frac{\partial}{\partial w_t} L(w_t)
   \]

   \[
   w_{t+1} = w_t + \alpha \Delta w_t
   \]

Now:
   \[
   w_{t+1} = w_t - \alpha \left[ \frac{\partial}{\partial w_t} L(w_t) + \lambda w_t \right] = w_t - \alpha [\Delta w_t + \lambda w_t]
   \]

   \[
   = w_t - \alpha \frac{\partial}{\partial w_t} L(w_t) - \alpha \lambda w_t = w_t + \alpha \Delta w_t - \alpha \lambda w_t
   \]

We use $\Delta w_t$ to represent the weight gradient for timepoint $t$ (the current step).

Subtract a little bit of weight every iteration
Momentum: speeding up SGD

If we keep moving in same direction we should move further every round

Before:
\[ \Delta w_t = -\frac{\partial}{\partial w_t} L(w_t) \]

Now:
\[ \Delta w_t = -\frac{\partial}{\partial w_t} L(w_t) + m\Delta w_{t-1} \]
\[ w_{t+1} = w_t + \alpha \Delta w_t \]

Side effect: smooths out updates if gradient is in different directions

\[ \Delta w_{t-1} \] represent the gradient calculated in the previous step.
NN updates with weight decay and momentum

\[ \Delta w'_t = -\frac{\partial}{\partial w_t} L(w_t) - \lambda w_t + m\Delta w'_{t-1} \]

- **Gradient of loss**
- **Weight decay**
- **Momentum**

\[ w_{t+1} = w_t + \alpha \Delta w'_t \]

- **Learning rate**
Dropout

Randomly eliminate some nodes during training
Activations
Linear Activation

- \[ g(x) = x \]
- \[ g'(x) = 1 \]

- Only offers linear effects.
- For a 2-layer NN with linear activations for both layers.

\[ f(X) = g(g(Xw_1)w_2) = Xw_1w_2 = Xw \]

- Not so great, need Non-Linear activations to learn more complex data distribution.
Logistic Activation

- \[ g(x) = \frac{1}{1 + e^{-x}} \]
- \[ g'(x) = g(x)g(1 - x) \]
- Aka Sigmoid function (S-shape)
- Used in Logistic regression.
- The result is in range (0, 1),
- It can represent probability.
- A special case of logistic growth (population model).
Do you see the problem with Sigmoidal activations?
Do you see the problem with Sigmoidal activations?

Vanishing gradients!
ReLU Activation

- \( g(x) = \max(0, x) \)
- \( g'(x) = 1_{g(x)>0} \)
- Rectified linear unit
- **Fast!** In backpropagation, 1 when positive, 0 otherwise.
- Optimizes important (positive) values and ignore the others.
- Analog to neurons
- Information loss is small (other neurons will carry information)
Visualization with ReLU

https://www.youtube.com/channel/UCYO_jab_esuFRV4b17AJtAw
Why ReLU provides non-linearity?

\[ f(x) = \begin{cases} x, & x > 0 \\ 0, & \text{else} \end{cases} \]

\[ \frac{d}{dx} f(x) = \begin{cases} 1, & f(x) > 0 \\ 0, & \text{else} \end{cases} \]
Why ReLU provides non-linearity?

The function is not linear if it does not satisfy the superposition principles: 1) additivity, 2) homogeneity

1) $F(x_1 + x_2) = F(x_1) + F(x_2)$
2) $F(ax) = aF(x)$
Why ReLU provides non-linearity?

Taken from https://www.blog.dailyydoseofds.com/p/a-visual-and-intuitive-guide-to-what#:~:text=The%20core%20point%20to%20understand%2C+of%20non%20linear%20curve.
The Dying ReLU Problem

\[ f'(x) = g(x) = 1, \ x \geq 0 \]
\[ = 0, \ x < 0 \]
LeakyReLU Activation

- No information loss (compared to ReLU)
- Solves “dying ReLU” problem (i.e. all neurons output 0)
- Similar to ReLU, pays less attention to less important neurons
- Not always better than ReLU
Do you see any other issues with ReLU?

\[ f(x) = \begin{cases} x, & x > 0 \\ 0, & \text{else} \end{cases} \]

\[ \frac{d}{dx} f(x) = \begin{cases} 1, & f(x) > 0 \\ 0, & \text{else} \end{cases} \]
Do you see any other issues with ReLU?

- Exploding gradients: Exact opposite phenomena we had with Sigmoid (vanishing gradients)
Do you see any other issues with ReLU?

- Exploding gradients: Exact opposite phenomena we had with Sigmoid (vanishing gradients)
  - Do gradient clipping, or layer-wise normalization
Homework 4
Neural Network
MNIST: Handwriting recognition

50,000 images of handwriting
Input: 28 x 28 x 1 (grayscale)  784 pixel values
Numbers 0-9  10 classes

Train a linear softmax model
> 95% accuracy
Functions You need to Code

Functions You need to Code (*classifier.c*)

```c
void activate_matrix(matrix m, ACTIVATION a);
void gradient_matrix(matrix m, ACTIVATION a, matrix d);
matrix forward_layer(layer *l, matrix in);
matrix backward_layer(layer *l, matrix delta);
void update_layer(layer *l, double rate, double momentum, double decay);
```

Run Experiments and Write a Report (*hw4.pdf*)

Play around with tryhw4.py file, and answer the questions.

Save your question to a PDF file and submit to Canvas for grading.
typedef enum{LINEAR, LOGISTIC, RELU, LRELU, SOFTMAX} ACTIVATION;

typedef struct {
    matrix in;         // Saved input to a layer
    matrix w;         // Current weights for a layer
    matrix dw;        // Current weight updates
    matrix v;         // Past weight updates (for use with momentum)
    matrix out;       // Saved output from the layer
    ACTIVATION activation; // Activation the layer uses
} layer;

typedef struct {
    layer *layers;
    int n;
} model;
Useful Matrix manipulation functions (matrix.c)

matrix matrix_mult_matrix (matrix a, matrix b);
matrix transpose_matrix (matrix m);
matrix axpy_matrix (double a, matrix x, matrix y); // a * x + y
Forward Pass in Homework

**forward_model**
- Input: model m, data X
- forward_layer
- Output

**forward_layer**
- Input: layer l, data in
- $X = \text{in} \times l->w$
- activation_matrix
  - $(X, l->activation)$
- Output
Backward Pass in Homework

**backward_model**

- Input: model m, matrix d

**backward_layer**

- Input: layer l, matrix delta
- Calculate $\Delta w$
- Calculate $\Delta o$
- $\Delta o$
Weight Update in Homework

**update_model**

Input: model m, learning rate $\alpha$, decay $\lambda$, momentum $m$

**update_layer**

Input: layer l, learning rate $\alpha$, decay $\lambda$, momentum $m$

$$\Delta w'_{t-1}$$ represent the regularized gradient from the previous step. In the code, we use "l->v" to store this value.

$$\Delta w' = \Delta w - \lambda w + m \Delta w'_{t-1}$$

$$w = w + \alpha \Delta w'$$
void activate_matrix(matrix m, ACTIVATION a)
{
    for(i = 0; i < m.rows; ++i){
        double sum = 0;
        for(j = 0; j < m.cols; ++j){
            double x = m.data[i][j];
            if(a == LOGISTIC){
                // TODO m.data[i][j] should equals 1 / (1 + exp(-x));
            } else if (a == RELU){
                // TODO m.data[i][j] should equals x if x > 0; otherwise, it should equal 0
            } else if (a == LRELU){
                // TODO m.data[i][j] should equals x if x > 0; otherwise, it should equal 0.1 * x.
            } else if (a == SOFTMAX){
                // TODO m.data[i][j] should equals exp(x) here, and we will normalize it later.
            }
            sum += m.data[i][j];
        }
        if (a == SOFTMAX) {
            // TODO: have to normalize by sum if we are using SOFTMAX
            // for all the possible j, we should normalize it as m.data[i][j] /= sum;
        }
    }
}
**TODO**

```c
void gradient_matrix(matrix m, ACTIVATION a, matrix d)
```

*Calculate \( g'(m) \times d \), and store in-place to matrix \( d \).
The matrix “\( m \)” is the output of a layer, and matrix “\( d \)” is the \( \Delta \) of output.*

```c
int i, j;
for (i = 0; i < m.rows; ++i){
    for (j = 0; j < m.cols; ++j){
        double x = m.data[i][j];
        // TODO: multiply the correct element of d by the gradient
        // if a is SOFTMAX or a is LINEAR, we should do nothing (multiply by 1)
        // if a is LOGISTIC, d.data[i][j] should times x * (1.0 - x);
        // if a is RELU and x <= 0, d.data[i][j] should be zero
        // if a is LRELU and x <= 0, d.data[i][j] should multiple 0.1
    }
}
```
TODO matrix forward_layer(layer *l, matrix in)

Given the input data "in" and layer "l", calculate the output data.

l->in = in;  // Save the input for backpropagation

// TODO: multiply input by weights and apply activation function.
// Calculate out = in * l->w (note: matrix multiplication here)
// Then, apply activate_matrix function to out with l->activation

free_matrix(l->out); // free the old output
l->out = out;        // Save the current output for gradient calculation
return out;
TODO matrix backward_layer(layer *l, matrix delta)

// delta is Δout
// TODO: modify it in place to be “g'(out) * delta” out with // gradient_matrix function.
// You can use gradient_matrix function with “l->out” and “l->activation” to “delta”

// TODO: then calculate dL/dw and save it in l->dw
free_matrix(l->dw);

// Calculate xt as the transpose matrix of “l->in”
// Calculate dw as xt times delta (matrix multiplication)
// free matrix xt to avoid memory leak
l->dw = dw;

// TODO: finally, calculate dL/dx and return it. (Similar to 1.4.2. Care memory leak)
// Calculate dx = delta * (l->w)^T, where * is matrix multiplication and ^T is matrix transpose
return dx;
void update_layer(layer *l, double rate, double momentum, double decay)

// Calculate Δw_t = dL/dw_t - λw_t + mΔw_{t-1}
// save it to l->v
// Note that You can use axpy_matrix to perform the matrix summation/subtraction

// Update l->w
// l->w = rate * l->v + l->w

Note the multiplication and summation in this slides all mean matrix multiplication or matrix summation.
Functions You Need to Know before Experiments

For simplicity, we already filled the following functions for you. You should read and understand these functions (classifier.c) before running experiments.

layer make_layer(int input, int output, ACTIVATION activation)
matrix forward_model(model m, matrix X)
void backward_model(model m, matrix dL)
void update_model(model m, double rate, double momentum, double decay)
double accuracy_model(model m, data d)
double cross_entropy_loss(matrix y, matrix p)
void train_model(model m, data d, int batch, int iters, double rate, double momentum, double decay)
Get the Data

1. Download, Unzip, and Prepare the MNIST Dataset

wget https://pjreddie.com/media/files/mnist_train.tar.gz
tar xzf mnist_train.tar.gz
tar xzf mnist_test.tar.gz
find train -name *.png > mnist.train
find test -name *.png > mnist.test

2. Download, Unzip, and Prepare the CIFAR-10 Dataset

tar xzf cifar.tgz
find cifar/train -name *.png > cifar.train
find cifar/test -name *.png > cifar.test
Experiments (Write Your Answers to hw4.pdf)

1. Coding and Data prepare
2. MNIST Experiments
   1. Linear Softmax Model (1-layer)
      1. Run the basic model
      2. Tune the learning rate
      3. Tune the decay
   2. Neural Network (2-layer NNs and 3-layer NNs)
      1. Find the best activation
      2. Tune the learning rate
      3. Tune the decay
      4. Tune the decay for 3-layer Neural Network
3. Experiments for CIFAR-10
   1. Neural Network (3-layer NNs)
      1. Tune the learning rate and decay
How to write this in code

Backprop Implementation:
“Flat” code

Forward pass:
Compute output

```
def f(w0, x0, w1, x1, w2):
    s0 = w0 * x0
    s1 = w1 * x1
    s2 = s0 + s1
    s3 = s2 + w2
    L = sigmoid(s3)
```

Backward pass:
Compute grads

```
grad_L = 1.0
grad_s3 = grad_L * (1 - L) * L
grad_w2 = grad_s3
grad_s2 = grad_s3
grad_s0 = grad_s2
grad_s1 = grad_s2
grad_w1 = grad_s1 * x1
grad_x1 = grad_s1 * w1
grad_w0 = grad_s0 * x0
grad_x0 = grad_s0 * w0
```
Back to our Neural Network Example

\[ \Delta w_t = -\frac{\partial E}{\partial w_t} = o_t^T g'(o_2) \circ \Delta o_2 \]

\[ w_{t+1} = w_t + \alpha \Delta w_t \]

Assume \( g'(.) = 1 \)

“\( \circ \)” represents elementwise multiplication for matrix
Neural Network Easy Example

\[ g'(o_2) \cdot \Delta o_2 \]

\[
\begin{array}{c|c}
0.39 & -0.39 \\
\vdots & \vdots \\
0.585 & -0.585 \\
\end{array}
\]

\[ \Delta w_2 = o_1^T g'(o_2) \cdot \Delta o_2 \]

\[
\begin{array}{c|c}
0 & 0 \\
1.092 & 0.117 \\
\vdots & \vdots \\
\end{array}
\]

\[ \Delta o_1 = g'(o_2) \cdot \Delta o_2 w_2^T \]

“\( \circ \)” represents elementwise multiplication for matrix

\[
\Delta w_t = -\frac{\partial E}{\partial w_t} = o_1^T g'(o_2) \cdot \Delta o_2
\]

\[ w_{t+1} = w_t + \alpha \Delta w_t \]

Assume \( g'(.) = 1 \)