Convolutional Neural Networks

Slides from Dr. Vlad Morariu
Deep learning

• Deep learning
  – multiple layer neural networks
  – learn features and classifiers directly ("end-to-end" training)

Artificial Neuron

Activation function:

\[ z = b + \sum_{i} w_i x_i \]

\[ y = f(z) \]

Activation function:

**Binary threshold**

\[ y = \begin{cases} 1 & \text{if } z \geq 0 \\ 0 & \text{otherwise} \end{cases} \]

**Sigmoid**

\[ y = \frac{1}{1 + e^{-z}} \]

**Rectified linear**

\[ y = \begin{cases} z & \text{if } z > 0 \\ 0 & \text{otherwise} \end{cases} \]
Single layer

- Input layer: $x_1, x_2, x_3, x_4$
- Single layer:
  - $y_1$: $f(w_1 \cdot x + b_1)$
  - $y_2$: $f(w_2 \cdot x + b_2)$
  - $y_3$: $f(w_3 \cdot x + b_3)$

Weights and biases:
- $w_{ij}$ for each connection
- $b_1, b_2, b_3$ for each neuron in the single layer
Multiple layers

```plaintext
Multiple layers

input

x_1

x_2

x_3

x_4

hidden layer

h_1

h_2

h_3

output layer

y_1

y_2

y_3

w_{111}
w_{111}
w_{112}
w_{112}
w_{113}
w_{113}
w_{114}
w_{114}
b_{11}
b_{11}

w_{121}
w_{121}
w_{122}
w_{122}
w_{123}
w_{123}
w_{124}
w_{124}
b_{12}
b_{12}

w_{131}
w_{131}
w_{132}
w_{132}
w_{133}
w_{133}
w_{134}
w_{134}
b_{13}
b_{13}

w_{211}
w_{211}
w_{212}
w_{212}
w_{213}
w_{213}
w_{214}
w_{214}
b_{21}
b_{21}

w_{221}
w_{221}
w_{222}
w_{222}
w_{223}
w_{223}
w_{224}
w_{224}
b_{22}
b_{22}

w_{231}
w_{231}
w_{232}
w_{232}
w_{233}
w_{233}
w_{234}
w_{234}
b_{23}
b_{23}

w_{31}
w_{31}
w_{32}
w_{32}
w_{33}
w_{33}
w_{34}
w_{34}
b_{31}
b_{31}
```

The diagram shows a neural network with multiple layers: the input layer, hidden layers, and the output layer.
Multiple layers

input layer

hidden layer

output layer
One layer (output)

What can the network represent with a threshold activation function?

\[ f(x) = \begin{cases} 1 & \text{if } w \cdot x + b > 0 \\ 0 & \text{otherwise} \end{cases} \]
Two layers (1 hidden + output)

What can the network represent with a threshold activation function?

\[ f(x) = \begin{cases} 1 & \text{if } x > 0 \\ 0 & \text{otherwise} \end{cases} \]
Three layers (2 hidden + output)

What can the network represent with a threshold activation function?
Training

Let \( \mathbf{x} = [x_1, \ldots, x_d]^T, \mathbf{y} = [y_1, \ldots, y_c]^T, \mathbf{W} = [\ldots w_{ljk} \ldots b_{lk} \ldots]^T \)
for all layers \( l \), neurons \( k \) in layer \( l \) and input \( j \) to layer \( k \).

The network computes \( \hat{\mathbf{y}} = f_{\text{net}}(\mathbf{x}, \mathbf{W}) \)

- Training data: \( \mathbf{X} = [\mathbf{x}^1, \ldots, \mathbf{x}^N], \mathbf{Y} = [\mathbf{y}^1, \ldots, \mathbf{y}^M] \)
- **Goal:** learn a \( \mathbf{W} \) so that \( \hat{\mathbf{y}}^n = f_{\text{net}}(\mathbf{x}^n, \mathbf{W}) \) is close to \( \mathbf{y}^n \)

Measure “closeness” by a **loss function**
- A value of 0 means we got the answer right
- A value > 0 means we got it wrong
- the higher the value, the more wrong
- e.g., \( \mathcal{L}(\hat{\mathbf{y}}, \mathbf{y}) = \frac{1}{2} (\hat{\mathbf{y}} - \mathbf{y})^2 \)

- Loss for dataset \( L = \sum_{n=1}^{N} \mathcal{L}(\hat{\mathbf{y}}^n, \mathbf{y}^n) \)
Gradient

In 1-dimension, the derivative of a function:

\[
\frac{df(x)}{dx} = \lim_{h \to 0} \frac{f(x + h) - f(x)}{h}
\]

For multidimensional input \( x = [x_1, \ldots, x_d]^\top \), the gradient of function \( f(x) \) consists of partial derivatives:

\[
\nabla f(x) = \left[ \frac{\partial f(x)}{\partial x_1}, \ldots, \frac{\partial f(x)}{\partial x_d} \right]^\top
\]

Why do we care about the gradient vector?
- It is the direction of largest increase in \( f(x) \).
- The negative gradient is the direction of largest decrease!

Any constraints on \( f(x) \)?
- It must be differentiable!

Slide credit: Modified from Fei-Fei Li, Andrej Karpathy, and Justin Johnson
Gradient descent

Until convergence

– Compute gradient
– Take a step in the negative gradient direction
  • Step size matters, and it is called “learning rate” in neural networks

Slide credit: Modified from Fei-Fei Li, Andrej Karpathy, and Justin Johnson
Numerical gradient

current W:

[0.34, -1.11, 0.78, 0.12, 0.55, 2.81, -3.1, -1.5, 0.33,...]

loss 1.25347

gradient dW:

### Numerical gradient

<table>
<thead>
<tr>
<th>current W:</th>
<th>( W + h ) (first dim):</th>
<th>gradient ( dW ):</th>
</tr>
</thead>
<tbody>
<tr>
<td>[0.34, -1.11, 0.78, 0.12, 0.55, 2.81, -3.1, -1.5, 0.33,...]</td>
<td>[0.34 + 0.0001, -1.11, 0.78, 0.12, 0.55, 2.81, -3.1, -1.5, 0.33,...]</td>
<td>[?, ?, ?, ?, ?, ?, ?, ?, ?,...]</td>
</tr>
<tr>
<td>loss 1.25347</td>
<td>loss 1.25322</td>
<td></td>
</tr>
</tbody>
</table>

Slide credit: Fei-Fei Li, Andrej Karpathy, and Justin Johnson
## Numerical gradient

<table>
<thead>
<tr>
<th>current W:</th>
<th>W + h (first dim):</th>
<th>gradient dW:</th>
</tr>
</thead>
<tbody>
<tr>
<td>[0.34, -1.11, 0.78, 0.12, 0.55, 2.81, -3.1, -1.5, 0.33,...]</td>
<td>[0.34 + 0.0001, -1.11, 0.78, 0.12, 0.55, 2.81, -3.1, -1.5, 0.33,...]</td>
<td>[-2.5, ?, ?, (1.25322 - 1.25347)/0.0001 = -2.5]</td>
</tr>
</tbody>
</table>

**loss 1.25347**

**loss 1.25322**

---

**slide credit:** Fei-Fei Li, Andrej Karpathy, and Justin Johnson
### Numerical gradient

<table>
<thead>
<tr>
<th>current $W$:</th>
<th>$W + h$ (second dim):</th>
<th>gradient $dW$:</th>
</tr>
</thead>
<tbody>
<tr>
<td>[0.34, -1.11, 0.78, 0.12, 0.55, 2.81, -3.1, -1.5, 0.33,...]</td>
<td>[0.34, -1.11 + 0.0001, 0.78, 0.12, 0.55, 2.81, -3.1, -1.5, 0.33,...]</td>
<td>[-2.5, ?, ?, ?, ?, ?, ?, ?,...]</td>
</tr>
</tbody>
</table>

**loss 1.25347**

**loss 1.25353**

---

Slide credit: Fei-Fei Li, Andrej Karpathy, and Justin Johnson
### Numerical gradient

<table>
<thead>
<tr>
<th>current W:</th>
<th>( W + h ) (second dim):</th>
<th>gradient ( dW ):</th>
</tr>
</thead>
<tbody>
<tr>
<td>[0.34, -1.11, 0.78, 0.12, 0.55, 2.81, -3.1, -1.5, 0.33,...]</td>
<td>[0.34, -1.11 + 0.0001, 0.78, 0.12, 0.55, 2.81, -3.1, -1.5, 0.33,...]</td>
<td>[-2.5, 0.6, ?, ?]</td>
</tr>
</tbody>
</table>

**loss 1.25347**

**loss 1.25353**

\[
\frac{df(x)}{dx} = \lim_{h \to 0} \frac{f(x + h) - f(x)}{h}
\]

\[
(1.25353 - 1.25347) / 0.0001 = 0.6
\]

Slide credit: Fei-Fei Li, Andrej Karpathy, and Justin Johnson
### Numerical gradient

<table>
<thead>
<tr>
<th>current W:</th>
<th>( \mathbf{W} + \mathbf{h} ) (third dim):</th>
<th>gradient ( d\mathbf{W} ):</th>
</tr>
</thead>
<tbody>
<tr>
<td>[0.34, -1.11, 0.78, 0.12, 0.55, 2.81, -3.1, -1.5, 0.33, ...]</td>
<td>[0.34, -1.11, 0.78 + 0.0001, 0.12, 0.55, 2.81, -3.1, -1.5, 0.33, ...]</td>
<td>[-2.5, 0.6, ?, ?, ?, ?, ?, ?, ?, ..., ]</td>
</tr>
<tr>
<td>loss 1.25347</td>
<td>loss 1.25347</td>
<td></td>
</tr>
</tbody>
</table>

Slide credit: Fei-Fei Li, Andrej Karpathy, and Justin Johnson
Numerical gradient

current W:  

W + h (third dim):

gradient dW:

\[
\begin{align*}
\text{current } W: & \quad [0.34, \\
& -1.11, \\
& 0.78, \\
& 0.12, \\
& 0.55, \\
& 2.81, \\
& -3.1, \\
& -1.5, \\
& 0.33, \ldots] \\
\text{loss } 1.25347
\end{align*}
\]

\[
\begin{align*}
\text{W + h (third dim):} & \quad [0.34, \\
& -1.11, \\
& 0.78 + 0.0001, \\
& 0.12, \\
& 0.55, \\
& 2.81, \\
& -3.1, \\
& -1.5, \\
& 0.33, \ldots] \\
\text{loss } 1.25347
\end{align*}
\]

\[
\begin{align*}
\text{gradient } dW: & \quad [-2.5, \\
& 0.6, \\
& 0, \\
& ?, \\
& ?, \\
& \ldots]
\end{align*}
\]

\[
(1.25347 - 1.25347)/0.0001 = 0
\]

\[
\frac{df(x)}{dx} = \lim_{h \to 0} \frac{f(x + h) - f(x)}{h}
\]
Numerical gradient is slow to compute!

Analytical gradient can be derived directly
• **Gradient descent**

  – Training data: $(x^1, y^1), (x^2, y^2), \ldots, (x^N, y^N)$

  – Goal: Finding the optimal parameters $W$, which minimize the loss, e.g., the quadratic loss*

  \[
  L = \frac{1}{2} \sum_{n} (y^n - \hat{y}^n)^2 \quad \text{where } \hat{y}^n = f(x^n, W)
  \]

  – Iteratively update the model parameters to decrease $L$ as

  \[
  w_i(t + 1) = w_i(t) - \epsilon \frac{\partial L}{\partial w_i}
  \]

* There are many other types loss functions

**Slide credit:** Bohyung Han
Single neuron gradient

\[ z = b + \sum_i w_i x_i \]
\[ \hat{y} = \frac{1}{1 + e^{-z}} \]
\[ L = \frac{1}{2} \sum_n (y^n - \hat{y}^n)^2 \]

Chain rule: If \( y = f(x), z = g(y) \), then
\[ \frac{dz}{dx} = \frac{dy}{dx} \frac{dz}{dy} \]

\[ \frac{\partial L}{\partial w_i} = \sum_n \frac{\partial \hat{y}^n}{\partial w_i} \frac{\partial L}{\partial \hat{y}^n} = \sum_n \frac{\partial z^n}{\partial w_i} \frac{d\hat{y}^n}{dz^n} \frac{\partial L}{\partial \hat{y}^n} = -\sum_n x_i^n \hat{y}^n (1 - \hat{y}^n) (y^n - \hat{y}^n) \]
Single neuron training

for $t = 1, ..., T$
\[ \hat{y}^n = f(x^n, w_t) \quad (n = 1, ..., N) \]
\[ \frac{\partial L}{\partial w_i} = - \sum_n x_i^n \hat{y}^n (1 - \hat{y}^n)(y^n - \hat{y}^n) \quad (i = 1, ..., d) \]
\[ w_{t+1} = w_t + \Delta w \]
endfor

Slide credit: Adapted from Bohyung Han
Multi-Layer: Backpropagation

\[
\frac{\partial L}{\partial z_j} = \frac{d\hat{y}_j}{dz_j} \frac{\partial L}{\partial \hat{y}_j}
\]

\[
\frac{\partial L}{\partial \hat{y}_i} = \sum_j \frac{d\hat{y}_i}{d\hat{y}_j} \frac{\partial L}{\partial z_j} = \sum_j w_{ij} \frac{\partial L}{\partial z_j} = \sum_j w_{ij} \frac{d\hat{y}_j}{dz_j} \frac{\partial L}{\partial \hat{y}_j}
\]

\[
\frac{\partial L}{\partial w_{ki}} = \sum_n \frac{\partial z_i^n}{\partial w_{ki}} \frac{d\hat{y}_i^n}{dz_i^n} \frac{\partial L}{\partial \hat{y}_i^n} = \sum_n \frac{\partial z_i^n}{\partial w_{ki}} \frac{d\hat{y}_i^n}{dz_i^n} \sum_j w_{ij} \frac{d\hat{y}_j^n}{dz_j^n} \frac{\partial L}{\partial \hat{y}_j^n}
\]

Slide credit: Bohyung Han
In practice

Analytical gradient is defined for each component of the architecture. Two passes per iteration:

• **Forward pass**: compute value of loss function (and intermediate neurons) given inputs

• **Backward pass**: propagate gradient of loss (error) backwards through the network using the chain rule
Analytical Gradient Example 1

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

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

Slide credit: Fei-Fei Li, Andrej Karpathy, and Justin Johnson
Analytical Gradient Example 1

\[ f(x, y, z) = (x + y)z \]
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: \( \frac{\partial f}{\partial x}, \ \frac{\partial f}{\partial y}, \ \frac{\partial f}{\partial z} \)
\[ f(x, y, z) = (x + y)z \]

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: \( \frac{\partial f}{\partial x}, \frac{\partial f}{\partial y}, \frac{\partial f}{\partial z} \)
Analytical Gradient Example 1

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

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: \( \frac{\partial f}{\partial x}, \frac{\partial f}{\partial y}, \frac{\partial f}{\partial z} \)
$$f(x, y, z) = (x + y)z$$
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: \[ \frac{\partial f}{\partial x}, \frac{\partial f}{\partial y}, \frac{\partial f}{\partial z} \]
Analytical Gradient Example 1

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

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: \( \frac{\partial f}{\partial x}, \ \frac{\partial f}{\partial y}, \ \frac{\partial f}{\partial z} \)
$f(x, y, z) = (x + y)z$

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: $\frac{\partial f}{\partial x}, \frac{\partial f}{\partial y}, \frac{\partial f}{\partial z}$
Analytical Gradient Example 1

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

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: \( \frac{\partial f}{\partial x}, \frac{\partial f}{\partial y}, \frac{\partial f}{\partial z} \)
$f(x, y, z) = (x + y)z$

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: $\frac{\partial f}{\partial x}, \frac{\partial f}{\partial y}, \frac{\partial f}{\partial z}$
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}

Chain rule:
\[ \frac{\partial f}{\partial y} = \frac{\partial f}{\partial q} \frac{\partial q}{\partial y} \]
Analytical Gradient Example 1

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

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: \( \frac{\partial f}{\partial x}, \frac{\partial f}{\partial y}, \frac{\partial f}{\partial z} \)
$f(x, y, z) = (x + y)z$

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: $\frac{\partial f}{\partial x}, \frac{\partial f}{\partial y}, \frac{\partial f}{\partial z}$

Chain rule:

$\frac{\partial f}{\partial x} = \frac{\partial f}{\partial q} \frac{\partial q}{\partial x}$
Analytical Gradient – One Neuron

activations

Slide credit: Fei-Fei Li, Andrej Karpathy, and Justin Johnson
Analytical Gradient – One Neuron

activations

\[
\frac{\partial z}{\partial x}
\]

\[
\frac{\partial z}{\partial y}
\]

“local gradient”

Slide credit: Fei-Fei Li, Andrej Karpathy, and Justin Johnson
Analytical Gradient – One Neuron

activations

\( \frac{\partial z}{\partial x} \)

\( \frac{\partial z}{\partial y} \)

“local gradient”

\( z \)

\( \frac{\partial L}{\partial z} \)

gradients

Slide credit: Fei-Fei Li, Andrej Karpathy, and Justin Johnson
Analytical Gradient – One Neuron

\[
\frac{\partial L}{\partial x} = \frac{\partial L}{\partial z} \frac{\partial z}{\partial x} + \frac{\partial L}{\partial y} \frac{\partial z}{\partial y}
\]

activations

“local gradient”

gradients

Slide credit: Fei-Fei Li, Andrej Karpathy, and Justin Johnson
Analytical Gradient – One Neuron

 activations

\[ \frac{\partial L}{\partial x} = \frac{\partial L}{\partial z} \frac{\partial z}{\partial x} \]

\[ \frac{\partial L}{\partial y} = \frac{\partial L}{\partial z} \frac{\partial z}{\partial y} \]

“local gradient”

 gradients

Slide credit: Fei-Fei Li, Andrej Karpathy, and Justin Johnson
Analytical Gradient – One Neuron

\[
\frac{\partial L}{\partial x} = \frac{\partial L}{\partial z} \frac{\partial z}{\partial x}
\]

\[
\frac{\partial L}{\partial y} = \frac{\partial L}{\partial z} \frac{\partial z}{\partial y}
\]

"local gradient"

Slide credit: Fei-Fei Li, Andrej Karpathy, and Justin Johnson
Another example:

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

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

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

Slide credit: Fei-Fei Li, Andrej Karpathy, and Justin Johnson
Another example:

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

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

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

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

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

Slide credit: Fei-Fei Li, Andrej Karpathy, and Justin Johnson
Another example: \[ f(w, x) = \frac{1}{1 + e^{-(w_0 x_0 + w_1 x_1 + w_2)}} \]

\[
\begin{align*}
f(x) &= e^x \\
f_a(x) &= ax
\end{align*}
\]

\[
\begin{align*}
\frac{df}{dx} &= e^x \\
\frac{df}{dx} &= a
\end{align*}
\]

\[
\begin{align*}
f(x) &= \frac{1}{x} \\
f_c(x) &= c + x
\end{align*}
\]

\[
\begin{align*}
\frac{df}{dx} &= -\frac{1}{x^2} \\
\frac{df}{dx} &= 1
\end{align*}
\]

Slide credit: Fei-Fei Li, Andrej Karpathy, and Justin Johnson
Another example:

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

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

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

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

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

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

\[
\begin{align*}
  f(x) &= e^x \\
  f_a(x) &= ax
\end{align*}
\]

\[
\begin{align*}
  \frac{df}{dx} &= e^x \\
  \frac{df}{dx} &= a
\end{align*}
\]

\[
\begin{align*}
  f(x) &= \frac{1}{x} \\
  f_c(x) &= c + x
\end{align*}
\]

\[
\begin{align*}
  \frac{df}{dx} &= -\frac{1}{x^2} \\
  \frac{df}{dx} &= 1
\end{align*}
\]

\[(1)(-0.53) = -0.53\]
Another example:

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

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

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

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

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

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

\[(e^{-1})(-0.53) = -0.20\]

\[
\begin{align*}
  f(x) &= e^x \\
  f_a(x) &= ax
\end{align*}
\]

\[
\begin{align*}
  \frac{df}{dx} &= e^x \\
  \frac{df}{dx} &= a
\end{align*}
\]

\[
\begin{align*}
  f(x) &= \frac{1}{x} \\
  f_c(x) &= c + x
\end{align*}
\]

\[
\begin{align*}
  \frac{df}{dx} &= -\frac{1}{x^2} \\
  \frac{df}{dx} &= 1
\end{align*}
\]
Another example: \( f(w, x) = \frac{1}{1 + e^{-(w_0 x_0 + w_1 x_1 + w_2)}} \)

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

\[
\begin{align*}
  f(x) &= \frac{1}{x} \\
  \frac{df}{dx} &= -\frac{1}{x^2} \\
  f_c(x) &= c + x \\
  \frac{df}{dx} &= 1
\end{align*}
\]
Another example:

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

\[ (-1) \times (-0.20) = 0.20 \]

\[
\begin{align*}
  f(x) &= e^x & \Rightarrow & \quad \frac{df}{dx} = e^x \\
  f_a(x) &= ax & \Rightarrow & \quad \frac{df}{dx} = a \\
  f_c(x) &= c + x & \Rightarrow & \quad \frac{df}{dx} = 1 \\
  f(x) &= \frac{1}{x} & \Rightarrow & \quad \frac{df}{dx} = -\frac{1}{x^2}
\end{align*}
\]

Slide credit: Fei-Fei Li, Andrej Karpathy, and Justin Johnson.
Another example:

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

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

\[ f_a(x) = ax \quad \rightarrow \quad \frac{df}{dx} = a \quad \rightarrow \quad f_c(x) = c + x \quad \rightarrow \quad \frac{df}{dx} = 1 \]
Another example:

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

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

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

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

[local gradient] x [its gradient]

\[ [1] \times [0.2] = 0.2 \]

\[ [1] \times [0.2] = 0.2 \quad \text{(both inputs!)} \]
Another example:

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

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

Slide credit: Fei-Fei Li, Andrej Karpathy, and Justin Johnson
Another example:

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

(local gradient) \(\times\) (its gradient)

\[ x_0: [2] \times [0.2] = 0.4 \]
\[ w_0: [-1] \times [0.2] = -0.2 \]

\[
\begin{align*}
 f(x) &= e^x \
 f_a(x) &= ax \\
 \frac{df}{dx} &= e^x \
 \frac{df}{dx} &= a \\
 f_c(x) &= c + x \\
 \frac{df}{dx} &= \frac{1}{x^2} \
 \frac{df}{dx} &= 1
\end{align*}
\]
Analytical Gradient – Example 2

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

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

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

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

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

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

(0.73) * (1 - 0.73) = 0.2

sigmoid function

sigmoid gate

Slide credit: Fei-Fei Li, Andrej Karpathy, and Justin Johnson
Analytical Gradient – Patterns

Patterns in backward flow

**add** gate: gradient distributor
**max** gate: gradient router
**mul** gate: gradient... “switcher”?
Gradients add at branches
Example of optimization progress while training a neural network.

(Loss over mini-batches goes down over time.)
Learning rate

The effects of step size (or “learning rate”)

Slide credit: Fei-Fei Li, Andrej Karpathy, and Justin Johnson
Multiple gradient update formulas

The effects of different update formulas

(image credits to Alec Radford)

Slide credit: Fei-Fei Li, Andrej Karpathy, and Justin Johnson