# Artificial Intelligence for Tree Failure Identification and Risk Quantification
## Introductory Meeting

**NARS Lab**

UMass Amherst
College of Engineering

December 2, 2020

# Outline

1. Convolutional NNs

2. Methods and Results

# The convolutional neural network (CNN)

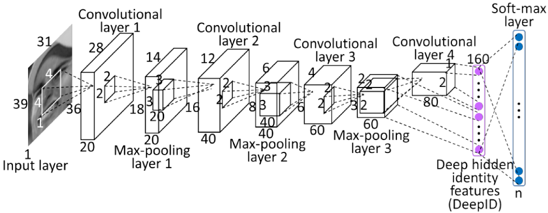
- Motivated by the image recognition process of the brain's visual cortex.
- Groundbreaking study on cats revealed the importance of *local receptive fields* for activating neurons in the visual cortex. (Hubel & Wiesel, 1958; 1959)
- Earliest neural network for image recognitron introduced: *neocognitron* (Fukushima, 1980)
- Milestone: introduction of *LeNet-5* architecture for handwritten digit recognition (Yann LeCun et al., 1998)

# The convolutional neural network (CNN)

- Motivated by the image recognition process of the brain's visual cortex.
- Groundbreaking study on cats revealed the importance of *local receptive fields* for activating neurons in the visual cortex. (Hubel & Wiesel, 1958; 1959)
- Earliest neural network for image recognitron introduced: *neocognitron* (Fukushima, 1980)
- Milestone: introduction of *LeNet-5* architecture for handwritten digit recognition (Yann LeCun et al., 1998)

# Building blocks of a CNN

- **Input layer:** the image to be classified
- **Convolutional layer:** represents the action of a filter transmitting signals (features) from various portions (receptive fields) of the preceding layer. The size of the receptive field is specified by the *convolutional kernel*. Each layer can have multiple feature maps representing different filters.
- **Pooling layer:** subsamples signals from preceding layer to reduce dimensionality and extract dominant features (subsample space determined by kernel size)
- **Dense layer:** neuron outputs are flattened and fully connected
- **Output layer:** neurons equal to number of classes; with softmax activation



# Training hyperparameters in a CNN

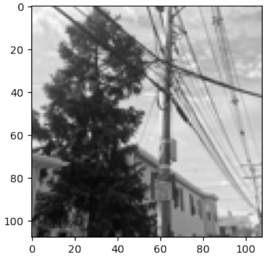
Several decisions must be made in selecting hyperparameters for training a CNN.

- Number of convolutional layers and feature maps in each layer
- Convolutional kernel size
- Stride length (spacing of filters)
- Choice of pooling function (max, average, etc)
- Number of dense layers
- Activation function in each layer (ReLU, tanh, etc)

Various high-performing architectures have been developed in recent years that can be adapted for other problems.
Training the network involves finding the weights for the various layers (using mini-batch gradient descent or other variants)

# Data preprocessing

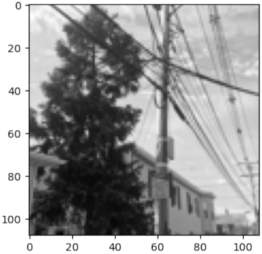
- Labeled 458 images serially and by class:
  - Probable images saved: 56
  - Possible images saved: 80
  - Improbable images saved: 322
- Cropped images to $3024 \times 3024$
- Downsampled to $108 \times 108$ (using standard cubic interpolation algorithm)




- Randomly split images into training and validation sets (80:20 ratio)
  - Size of validation sample: 76 images

# Data preprocessing

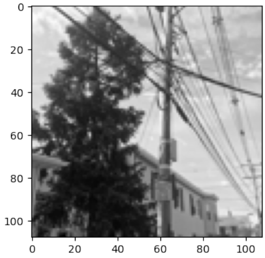
- Labeled 458 images serially and by class:
  - Probable images saved: 56
  - Possible images saved: 80
  - Improbable images saved: 322
- Cropped images to $3024 \times 3024$
- Downsampled to $108 \times 108$ (using standard cubic interpolation algorithm)




- Randomly split images into training and validation sets (80:20 ratio)
  - Size of validation sample: [illegible] images

# Data preprocessing

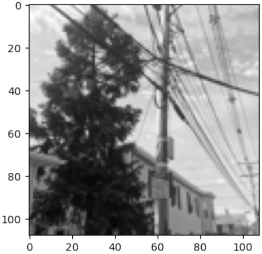
- Labeled 458 images serially and by class:
  - Probable images saved: 56
  - Possible images saved: 80
  - Improbable images saved: 322
- Cropped images to $3024 \times 3024$
- Downsampled to $108 \times 108$ (using standard cubic interpolation algorithm)




- Randomly split images into training and validation sets (80:20 ratio)
  - Size of validation sample: [illegible] images

# Data preprocessing

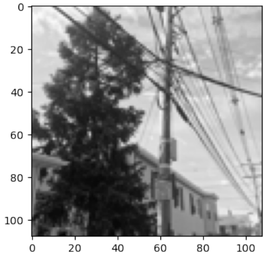
- Labeled 458 images serially and by class:
  - Probable images saved: 56
  - Possible images saved: 80
  - Improbable images saved: 322
- Cropped images to $3024 \times 3024$
- Downsampled to $108 \times 108$ (using standard cubic interpolation algorithm)




- Randomly split images into training and validation sets (80:20 ratio)
  - Size of validation sample: 76 images

# Data preprocessing

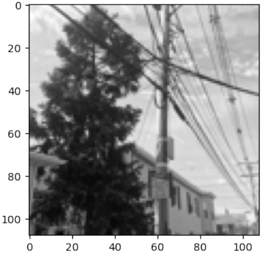
- Labeled 458 images serially and by class:
  - Probable images saved: 56
  - Possible images saved: 80
  - Improbable images saved: 322
- Cropped images to $3024 \times 3024$
- Downsampled to $108 \times 108$ (using standard cubic interpolation algorithm)




- Randomly split images into training and validation sets (80:20 ratio)
  - Size of validation sample: 76 images

# Data preprocessing

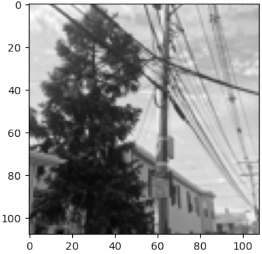
- Labeled 458 images serially and by class:
  - Probable images saved: 56
  - Possible images saved: 80
  - Improbable images saved: 322
- Cropped images to $3024 \times 3024$
- Downsampled to $108 \times 108$ (using standard cubic interpolation algorithm)




- Randomly split images into training and validation sets (80:20 ratio)
  - Size of validation sample: 75 images

# Data preprocessing

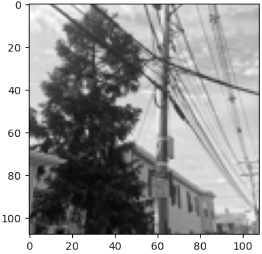
- Labeled 458 images serially and by class:
  - Probable images saved: 56
  - Possible images saved: 80
  - Improbable images saved: 322
- Cropped images to $3024 \times 3024$
- Downsampled to $108 \times 108$ (using standard cubic interpolation algorithm)




- Randomly split images into training and validation sets (80:20 ratio)
  - Size of validation sample: 76 images

# Data preprocessing

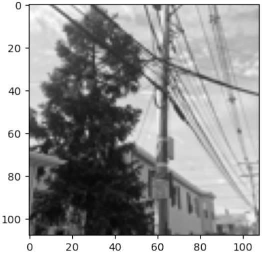
- Labeled 458 images serially and by class:
  - Probable images saved: 56
  - Possible images saved: 80
  - Improbable images saved: 322
- Cropped images to $3024 \times 3024$
- Downsampled to $108 \times 108$ (using standard cubic interpolation algorithm)




- Randomly split images into training and validation sets (80:20 ratio)
  - Size of validation sample: 75 images

# Data preprocessing

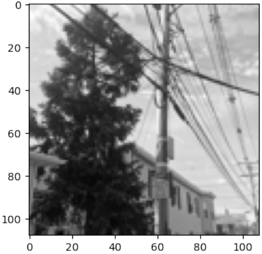
- Labeled 458 images serially and by class:
  - Probable images saved: 56
  - Possible images saved: 80
  - Improbable images saved: 322
- Cropped images to $3024 \times 3024$
- Downsampled to $108 \times 108$ (using standard cubic interpolation algorithm)




- Randomly split images into training and validation sets (80:20 ratio)
  - Size of validation sample: 75 images

# Data preprocessing

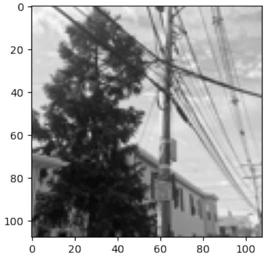
- Labeled 458 images serially and by class:
  - Probable images saved: 56
  - Possible images saved: 80
  - Improbable images saved: 322
- Cropped images to 3024 $\times$ 3024
- Downsampled to 108 $\times$ 108 (using standard cubic interpolation algorithm)




- Randomly split images into training and validation sets (80:20 ratio)
  - Size of validation sample: 76 images

# Data preprocessing

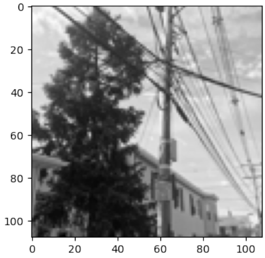
- Labeled 458 images serially and by class:
  - Probable images saved: 56
  - Possible images saved: 80
  - Improbable images saved: 322
- Cropped images to $3024 \times 3024$
- Downsampled to $108 \times 108$ (using standard cubic interpolation algorithm)




- Randomly split images into training and validation sets (80:20 ratio)
  - Size of validation sample: 76 images

# Data preprocessing

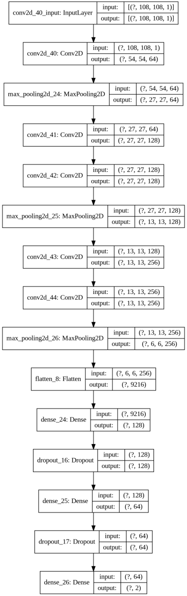
- Labeled 458 images serially and by class:
  - Probable images saved: 56
  - Possible images saved: 80
  - Improbable images saved: 322
- Cropped images to 3024 × 3024
- Downsampled to 108 × 108 (using standard cubic interpolation algorithm)




- Randomly split images into training and validation sets (80:20 ratio)
  - Size of validation sample: 76 images

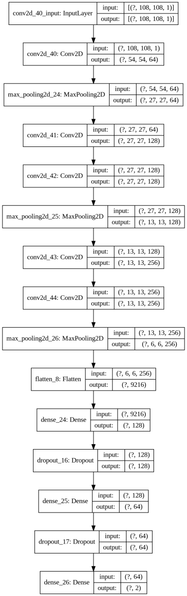
# CNN Model 1

- Binary image classifier
- Simple architecture:
  - 5 convolutional layers
  - 2 hidden dense layers
  - 1 output layer
- Input layer: $108 \times 108 \times 1$ tensor
- Output layer: $2 \times 1$ vector of probabilities (of the input belonging to either of the classes)



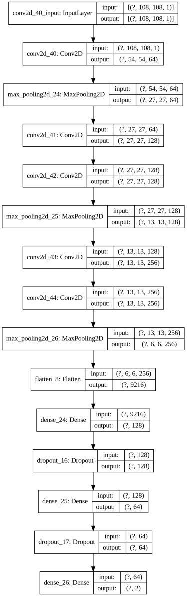
# CNN Model 1

- Binary image classifier
- Simple architecture:
  - 5 convolutional layers
  - 2 hidden dense layers
  - 1 output layer
- Input layer: $108 \times 108 \times 1$ tensor
- Output layer: $2 \times 1$ vector of probabilities (of the input belonging to either of the classes)



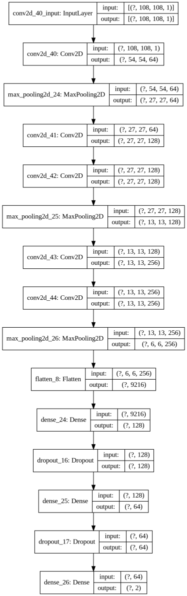
# CNN Model 1

- Binary image classifier
- Simple architecture:
  - 5 convolutional layers
  - 2 hidden dense layers
  - 1 output layer
- Input layer: $108 \times 108 \times 1$ tensor
- Output layer: $2 \times 1$ vector of probabilities (of the input belonging to either of the classes)



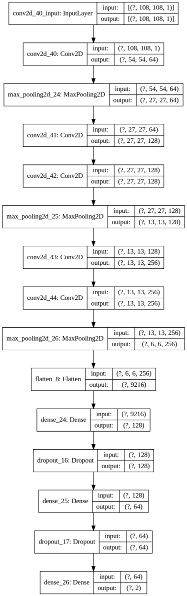
# CNN Model 1

- Binary image classifier
- Simple architecture:
  - 5 convolutional layers
  - 2 hidden dense layers
  - 1 output layer
- Input layer: $108 \times 108 \times 1$ tensor
- Output layer: $2 \times 1$ vector of probabilities (of the input belonging to either of the classes)



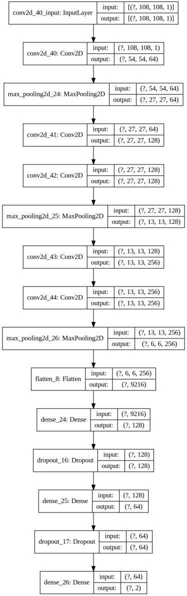
# CNN Model 1

- Binary image classifier
- Simple architecture:
  - 5 convolutional layers
  - 2 hidden dense layers
  - 1 output layer
- Input layer: $108 \times 108 \times 1$ tensor
- Output layer: $2 \times 1$ vector of probabilities (of the input belonging to either of the classes)



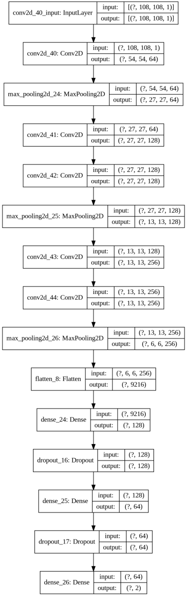
# CNN Model 1

- Binary image classifier
- Simple architecture:
  - 5 convolutional layers
  - 2 hidden dense layers
  - 1 output layer
- Input layer: $108 \times 108 \times 1$ tensor
- Output layer: $2 \times 1$ vector of probabilities (of the input belonging to either of the classes)



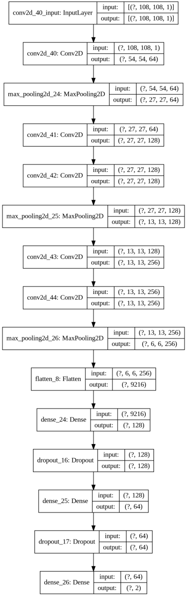
# CNN Model 1

- Binary image classifier
- Simple architecture:
    - 5 convolutional layers
    - 2 hidden dense layers
    - 1 output layer
- Input layer: $108 \times 108 \times 1$ tensor
- Output layer: $2 \times 1$ vector of probabilities (of the input belonging to either of the classes)



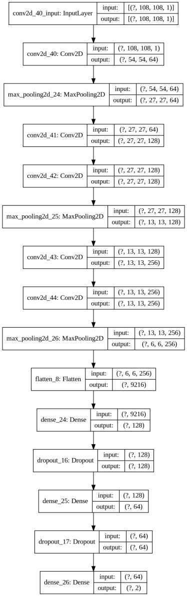
# CNN Model 1

- Binary image classifier
- Simple architecture:
  - 5 convolutional layers
  - 2 hidden dense layers
  - 1 output layer
- Input layer: $108 \times 108 \times 1$ tensor
- Output layer: $2 \times 1$ vector of probabilities (of the input belonging to either of the classes)



# CNN Model 1

- Binary image classifier
- Simple architecture:
  - 5 convolutional layers
  - 2 hidden dense layers
  - 1 output layer
- Input layer: $108 \times 108 \times 1$ tensor
- Output layer: $2 \times 1$ vector of probabilities (of the input belonging to either of the classes)



# Training hyperparameters and decisions

- **Optimizer:** algorithm used for finding optimal CNN weights (usually a variant of stochastic gradient descent)
- **Batch size:** number of images used in each iteration to compute the gradients and update CNN weights
- **Epochs:** number of sweeps through all the training observations for CNN learning
- **Loss function:** quantifies the error in a prediction compared to the observed label

# Training hyperparameters and decisions

- **Optimizer:** algorithm used for finding optimal CNN weights (usually a variant of stochastic gradient descent)
- **Batch size:** number of images used in each iteration to compute the gradients and update CNN weights
- **Epochs:** number of sweeps through all the training observations for CNN learning
- **Loss function:** quantifies the error in a prediction compared to the observed label

# Training hyperparameters and decisions

- **Optimizer:** algorithm used for finding optimal CNN weights (usually a variant of stochastic gradient descent)
- **Batch size:** number of images used in each iteration to compute the gradients and update CNN weights
- **Epochs:** number of sweeps through all the training observations for CNN learning
- **Loss function:** quantifies the error in a prediction compared to the observed label

# Training hyperparameters and decisions

- **Optimizer:** algorithm used for finding optimal CNN weights (usually a variant of stochastic gradient descent)
- **Batch size:** number of images used in each iteration to compute the gradients and update CNN weights
- **Epochs:** number of sweeps through all the training observations for CNN learning
- **Loss function:** quantifies the error in a prediction compared to the observed label

# Results

Current results are based on Model 1 using monochromatic images, 4 epochs (algorithm converges fast) and 32 samples per batch.

- Optimizer used: Adam (variant of SGD)
- All performance metrics are based on validation set:
  - Precision: $\frac{TP}{TP+FP}$
  - Recall: $\frac{TP}{TP+FN}$

| **Class Labels** | **Image Size** | **Prec.** | **Recall** | **Acc.** |
|---|---|---|---|---|
| {(Probable + Possible), Improbable} | $108 \times 108$ | | | 65.2 |
| {Probable, Possible, Improbable} | $108 \times 108$ | | | 72.8 |
| {Probable, Improbable} | $108 \times 108$ | 87.9 | 100 | 86.8 |
| {Probable, Improbable} | $224 \times 224$ | 87.9 | 100 | 86.8 |

# Results

Current results are based on Model 1 using monochromatic images, 4 epochs (algorithm converges fast) and 32 samples per batch.

- Optimizer used: Adam (variant of SGD)
- All performance metrics are based on validation set:
  - Precision: $\frac{TP}{TP+FP}$
  - Recall: $\frac{TP}{TP+FN}$

| Class Labels | Image Size | Prec. | Recall | Acc. |
|---|---|---|---|---|
| {(Probable + Possible), Improbable} | 108 × 108 | | | 65.2 |
| {Probable, Possible, Improbable} | 108 × 108 | | | 72.8 |
| {Probable, Improbable} | 108 × 108 | 87.9 | 100 | 86.8 |
| {Probable, Improbable} | 224 × 224 | 87.9 | 100 | 86.8 |

# Results

Current results are based on Model 1 using monochromatic images, 4 epochs (algorithm converges fast) and 32 samples per batch.

- Optimizer used: Adam (variant of SGD)
- All performance metrics are based on validation set:
  - Precision: $\frac{TP}{TP+FP}$
  - Recall: $\frac{TP}{TP+FN}$

| Class Labels | Image Size | Prec. | Recall | Acc. |
|---|---|---|---|---|
| $\{$(Probable + Possible), Improbable$\}$ | $108 \times 108$ | | | 65.2 |
| $\{$Probable, Possible, Improbable$\}$ | $108 \times 108$ | | | 72.8 |
| $\{$Probable, Improbable$\}$ | $108 \times 108$ | 87.9 | 100 | 86.8 |
| $\{$Probable, Improbable$\}$ | $224 \times 224$ | 87.9 | 100 | 86.8 |

# Results

Current results are based on Model 1 using monochromatic images, 4 epochs (algorithm converges fast) and 32 samples per batch.

- Optimizer used: Adam (variant of SGD)
- All performance metrics are based on validation set:
  - Precision: $\frac{TP}{TP+FP}$
  - Recall: $\frac{TP}{TP+FN}$

| Class Labels | Image Size | Prec. | Recall | Acc. |
|---|---|---|---|---|
| $\{$(Probable + Possible), Improbable$\}$ | $108 \times 108$ | | | 65.2 |
| $\{$Probable, Possible, Improbable$\}$ | $108 \times 108$ | | | 72.8 |
| $\{$Probable, Improbable$\}$ | $108 \times 108$ | 87.9 | 100 | 86.8 |
| $\{$Probable, Improbable$\}$ | $224 \times 224$ | 87.9 | 100 | 86.8 |

# Results

Current results are based on Model 1 using monochromatic images, 4 epochs (algorithm converges fast) and 32 samples per batch.

- Optimizer used: Adam (variant of SGD)
- All performance metrics are based on validation set:
  - Precision: $\frac{TP}{TP+FP}$
  - Recall: $\frac{TP}{TP+FN}$

| Class Labels | Image Size | Prec. | Recall | Acc. |
|---|---|---|---|---|
| {(Probable + Possible), Improbable} | 108 × 108 | | | 65.2 |
| {Probable, Possible, Improbable} | 108 × 108 | | | 72.8 |
| {Probable, Improbable} | 108 × 108 | 87.9 | 100 | 86.8 |
| {Probable, Improbable} | 224 × 224 | 87.9 | 100 | 86.8 |

# Results

Current results are based on Model 1 using monochromatic images, 4 epochs (algorithm converges fast) and 32 samples per batch.

- Optimizer used: Adam (variant of SGD)
- All performance metrics are based on validation set:
  - Precision: $\frac{TP}{TP+FP}$
  - Recall: $\frac{TP}{TP+FN}$

| Class Labels | Image Size | Prec. | Recall | Acc. |
|---|---|---|---|---|
| {(Probable + Possible), Improbable} | $108 \times 108$ | | | 65.2 |
| {Probable, Possible, Improbable} | $108 \times 108$ | | | 72.8 |
| {Probable, Improbable} | $108 \times 108$ | 87.9 | 100 | 86.8 |
| {Probable, Improbable} | $224 \times 224$ | 87.9 | 100 | 86.8 |

# Results

Current results are based on Model 1 using monochromatic images, 4 epochs (algorithm converges fast) and 32 samples per batch.

- Optimizer used: Adam (variant of SGD)
- All performance metrics are based on validation set:
  - Precision: $\frac{TP}{TP+FP}$
  - Recall: $\frac{TP}{TP+FN}$

| Class Labels | Image Size | Prec. | Recall | Acc. |
|---|---|---|---|---|
| {(Probable + Possible), Improbable} | $108 \times 108$ | | | 65.2 |
| {Probable, Possible, Improbable} | $108 \times 108$ | | | 72.8 |
| {Probable, Improbable} | $108 \times 108$ | 87.9 | 100 | 86.8 |
| {Probable, Improbable} | $224 \times 224$ | 87.9 | 100 | 86.8 |

# Results

Current results are based on Model 1 using monochromatic images, 4 epochs (algorithm converges fast) and 32 samples per batch.

- Optimizer used: Adam (variant of SGD)
- All performance metrics are based on validation set:
  - Precision: $\frac{TP}{TP+FP}$
  - Recall: $\frac{TP}{TP+FN}$

| Class Labels | Image Size | Prec. | Recall | Acc. |
|---|---|---|---|---|
| {(Probable + Possible), Improbable} | $108 \times 108$ | | | 65.2 |
| {Probable, Possible, Improbable} | $108 \times 108$ | | | 72.8 |
| {Probable, Improbable} | $108 \times 108$ | 87.9 | 100 | 86.8 |
| {Probable, Improbable} | $224 \times 224$ | 87.9 | 100 | 86.8 |

# Results

Current results are based on Model 1 using monochromatic images, 4 epochs (algorithm converges fast) and 32 samples per batch.

- Optimizer used: Adam (variant of SGD)
- All performance metrics are based on validation set:
  - Precision: $\frac{TP}{TP+FP}$
  - Recall: $\frac{TP}{TP+FN}$

| Class Labels | Image Size | Prec. | Recall | Acc. |
|---|---|---|---|---|
| {(Probable + Possible), Improbable} | $108 \times 108$ | | | 65.2 |
| {Probable, Possible, Improbable} | $108 \times 108$ | | | 72.8 |
| {Probable, Improbable} | $108 \times 108$ | 87.9 | 100 | 86.8 |
| {Probable, Improbable} | $224 \times 224$ | 87.9 | 100 | 86.8 |

# Results

Current results are based on Model 1 using monochromatic images, 4 epochs (algorithm converges fast) and 32 samples per batch.

- Optimizer used: Adam (variant of SGD)
- All performance metrics are based on validation set:
  - Precision: $\frac{TP}{TP+FP}$
  - Recall: $\frac{TP}{TP+FN}$

| Class Labels | Image Size | Prec. | Recall | Acc. |
|---|---|---|---|---|
| {(Probable + Possible), Improbable} | 108 × 108 | | | 65.2 |
| {Probable, Possible, Improbable} | 108 × 108 | | | 72.8 |
| {Probable, Improbable} | 108 × 108 | 87.9 | 100 | 86.8 |
| {Probable, Improbable} | 224 × 224 | 87.9 | 100 | 86.8 |

# Results

Current results are based on Model 1 using monochromatic images, 4 epochs (algorithm converges fast) and 32 samples per batch.

- Optimizer used: Adam (variant of SGD)
- All performance metrics are based on validation set:
  - Precision: $\frac{TP}{TP+FP}$
  - Recall: $\frac{TP}{TP+FN}$

| Class Labels | Image Size | Prec. | Recall | Acc. |
|---|---|---|---|---|
| {(Probable + Possible), Improbable} | $108 \times 108$ | | | 65.2 |
| {Probable, Possible, Improbable} | $108 \times 108$ | | | 72.8 |
| {Probable, Improbable} | $108 \times 108$ | 87.9 | 100 | 86.8 |
| {Probable, Improbable} | $224 \times 224$ | 87.9 | 100 | 86.8 |

# Results

Current results are based on Model 1 using monochromatic images, 4 epochs (algorithm converges fast) and 32 samples per batch.

- Optimizer used: Adam (variant of SGD)
- All performance metrics are based on validation set:
  - Precision: $\frac{TP}{TP+FP}$
  - Recall: $\frac{TP}{TP+FN}$

| **Class Labels** | **Image Size** | **Prec.** | **Recall** | **Acc.** |
|---|---|---|---|---|
| {(Probable + Possible), Improbable} | $108 \times 108$ | | | 65.2 |
| {Probable, Possible, Improbable} | $108 \times 108$ | | | 72.8 |
| {Probable, Improbable} | $108 \times 108$ | 87.9 | 100 | 86.8 |
| {Probable, Improbable} | $224 \times 224$ | 87.9 | 100 | 86.8 |

# Results summary

- We obtained the best performance from a binary classification of Probable and Improbable images
- Increasing pixel size (108 to 224) did not improve results

# Results summary

- We obtained the best performance from a binary classification of Probable and Improbable images
- Increasing pixel size (108 to 224) did not improve results

# Next steps

- Implement a high-performance CNN architecture (usually deeper is better). Candidates:
  - ResNet50 (50 layers)
  - GoogLeNet (22 layers)
- Investigate effects of using 3 channels (full-color images) in training
- Explore data augmentation procedures
- Longer term: perform inferential interpretation of relevant features in trained CNN and map to physical relationships
- Compare trained assessemnts (professional opinion)

# Next steps

- Implement a high-performance CNN architecture (usually deeper is better). Candidates:
  - ResNet50 (50 layers)
  - GoogLeNet (22 layers)
- Investigate effects of using 3 channels (full-color images) in training
- Explore data augmentation procedures
- Longer term: perform inferential interpretation of relevant features in trained CNN and map to physical relationships
- Compare trained assessemnts (professional opinion)

# Next steps

- Implement a high-performance CNN architecture (usually deeper is better). Candidates:
  - ResNet50 (50 layers)
  - GoogLeNet (22 layers)
- Investigate effects of using 3 channels (full-color images) in training
- Explore data augmentation procedures
- Longer term: perform inferential interpretation of relevant features in trained CNN and map to physical relationships
- Compare trained assessemnts (professional opinion)

# Next steps

- Implement a high-performance CNN architecture (usually deeper is better). Candidates:
  - ResNet50 (50 layers)
  - GoogLeNet (22 layers)
- Investigate effects of using 3 channels (full-color images) in training
- Explore data augmentation procedures
- Longer term: perform inferential interpretation of relevant features in trained CNN and map to physical relationships
- Compare trained assessemnts (professional opinion)

# Next steps

- Implement a high-performance CNN architecture (usually deeper is better). Candidates:
  - ResNet50 (50 layers)
  - GoogLeNet (22 layers)
- Investigate effects of using 3 channels (full-color images) in training
- Explore data augmentation procedures
- Longer term: perform inferential interpretation of relevant features in trained CNN and map to physical relationships
- Compare trained assessemnts (professional opinion)

# Next steps

- Implement a high-performance CNN architecture (usually deeper is better). Candidates:
  - ResNet50 (50 layers)
  - GoogLeNet (22 layers)
- Investigate effects of using 3 channels (full-color images) in training
- Explore data augmentation procedures
- Longer term: perform inferential interpretation of relevant features in trained CNN and map to physical relationships
- Compare trained assessemnts (professional opinion)

# Next steps

- Implement a high-performance CNN architecture (usually deeper is better). Candidates:
  - ResNet50 (50 layers)
  - GoogLeNet (22 layers)
- Investigate effects of using 3 channels (full-color images) in training
- Explore data augmentation procedures
- Longer term: perform inferential interpretation of relevant features in trained CNN and map to physical relationships
- Compare trained assessemnts (professional opinion)

# Next steps

- Implement a high-performance CNN architecture (usually deeper is better). Candidates:
  - ResNet50 (50 layers)
  - GoogLeNet (22 layers)
- Investigate effects of using 3 channels (full-color images) in training
- Explore data augmentation procedures
- Longer term: perform inferential interpretation of relevant features in trained CNN and map to physical relationships
- Compare trained assessemnts (professional opinion)

# Equation summary: outer layer

At the outer layer $L$ (without indexing by neuron):

$$z^{(L)} = w^{(L)} \times a^{(L-1)} + b^{(L)} \tag{1}$$

$$a^{(L)} = \sigma(z^{(L)}) \tag{2}$$

$$C = (a^{(L)} - y)^2 \tag{3}$$

The gradient of the cost function with respect to $w^{(L)}$ is:

$$\frac{\partial C}{\partial w^{(L)}} = \frac{\partial C}{\partial a^{(L)}} \frac{\partial a^{(L)}}{\partial z^{(L)}} \frac{\partial z^{(L)}}{\partial w^{(L)}} = 2 \left( a^{(L)} - y \right) \sigma' \left( z^{(L)} \right) a^{(L-1)} \tag{4}$$

Thus, we see that this gradient depends on the activation from the previous layer $a^{(L-1)}$. Also wrt to the bias:

$$\frac{\partial C}{\partial b^{(L)}} = \frac{\partial C}{\partial a^{(L)}} \frac{\partial a^{(L)}}{\partial z^{(L)}} \frac{\partial z^{(L)}}{\partial b^{(L)}} = 2 \left( a^{(L)} - y \right) \sigma'(z^{(L)})(1) \tag{5}$$

# Equation summary: outer layer

At the outer layer $L$ (without indexing by neuron):

$$z^{(L)} = w^{(L)} \times a^{(L-1)} + b^{(L)} \tag{1}$$
$$a^{(L)} = \sigma(z^{(L)}) \tag{2}$$
$$C = (a^{(L)} - y)^2 \tag{3}$$

The gradient of the cost function with respect to $w^{(L)}$ is:

$$\frac{\partial C}{\partial w^{(L)}} = \frac{\partial C}{\partial a^{(L)}} \frac{\partial a^{(L)}}{\partial z^{(L)}} \frac{\partial z^{(L)}}{\partial w^{(L)}} = 2\left(a^{(L)} - y\right)\sigma'\left(z^{(L)}\right)a^{(L-1)} \tag{4}$$

Thus, we see that this gradient depends on the activation from the previous layer $a^{(L-1)}$. Also wrt to the bias:

$$\frac{\partial C}{\partial b^{(L)}} = \frac{\partial C}{\partial a^{(L)}} \frac{\partial a^{(L)}}{\partial z^{(L)}} \frac{\partial z^{(L)}}{\partial b^{(L)}} = 2\left(a^{(L)} - y\right)\sigma'(z^{(L)})(1) \tag{5}$$

# Equation summary: outer layer

At the outer layer $L$ (without indexing by neuron):

$$z^{(L)} = w^{(L)} \times a^{(L-1)} + b^{(L)} \tag{1}$$

$$a^{(L)} = \sigma(z^{(L)}) \tag{2}$$

$$C = (a^{(L)} - y)^2 \tag{3}$$

The gradient of the cost function with respect to $w^{(L)}$ is:

$$\frac{\partial C}{\partial w^{(L)}} = \frac{\partial C}{\partial a^{(L)}} \frac{\partial a^{(L)}}{\partial z^{(L)}} \frac{\partial z^{(L)}}{\partial w^{(L)}} = 2\left(a^{(L)} - y\right) \sigma'\left(z^{(L)}\right) a^{(L-1)} \tag{4}$$

Thus, we see that this gradient depends on the activation from the previous layer $a^{(L-1)}$. Also wrt to the bias:

$$\frac{\partial C}{\partial b^{(L)}} = \frac{\partial C}{\partial a^{(L)}} \frac{\partial a^{(L)}}{\partial z^{(L)}} \frac{\partial z^{(L)}}{\partial b^{(L)}} = 2\left(a^{(L)} - y\right) \sigma'(z^{(L)})(1) \tag{5}$$

# Equation summary: outer layer

At the outer layer $L$ (without indexing by neuron):

$$z^{(L)} = w^{(L)} \times a^{(L-1)} + b^{(L)} \tag{1}$$

$$a^{(L)} = \sigma(z^{(L)}) \tag{2}$$

$$C = (a^{(L)} - y)^2 \tag{3}$$

The gradient of the cost function with respect to $w^{(L)}$ is:

$$\frac{\partial C}{\partial w^{(L)}} = \frac{\partial C}{\partial a^{(L)}} \frac{\partial a^{(L)}}{\partial z^{(L)}} \frac{\partial z^{(L)}}{\partial w^{(L)}} = 2\left(a^{(L)} - y\right) \sigma'\left(z^{(L)}\right) a^{(L-1)} \tag{4}$$

Thus, we see that this gradient depends on the activation from the previous layer $a^{(L-1)}$. Also wrt to the bias:

$$\frac{\partial C}{\partial b^{(L)}} = \frac{\partial C}{\partial a^{(L)}} \frac{\partial a^{(L)}}{\partial z^{(L)}} \frac{\partial z^{(L)}}{\partial b^{(L)}} = 2\left(a^{(L)} - y\right) \sigma'(z^{(L)})(1) \tag{5}$$

# Equation summary: outer layer

At the outer layer $L$ (without indexing by neuron):

$$z^{(L)} = w^{(L)} \times a^{(L-1)} + b^{(L)} \tag{1}$$

$$a^{(L)} = \sigma(z^{(L)}) \tag{2}$$

$$C = (a^{(L)} - y)^2 \tag{3}$$

The gradient of the cost function with respect to $w^{(L)}$ is:

$$\frac{\partial C}{\partial w^{(L)}} = \frac{\partial C}{\partial a^{(L)}} \frac{\partial a^{(L)}}{\partial z^{(L)}} \frac{\partial z^{(L)}}{\partial w^{(L)}} = 2\left(a^{(L)} - y\right) \sigma'\left(z^{(L)}\right) a^{(L-1)} \tag{4}$$

Thus, we see that this gradient depends on the activation from the previous layer $a^{(L-1)}$. Also wrt to the bias:

$$\frac{\partial C}{\partial b^{(L)}} = \frac{\partial C}{\partial a^{(L)}} \frac{\partial a^{(L)}}{\partial z^{(L)}} \frac{\partial z^{(L)}}{\partial b^{(L)}} = 2\left(a^{(L)} - y\right) \sigma'(z^{(L)})(1) \tag{5}$$

# Equation summary: outer layer

At the outer layer $L$ (without indexing by neuron):

$$z^{(L)} = w^{(L)} \times a^{(L-1)} + b^{(L)} \tag{1}$$

$$a^{(L)} = \sigma(z^{(L)}) \tag{2}$$

$$C = (a^{(L)} - y)^2 \tag{3}$$

The gradient of the cost function with respect to $w^{(L)}$ is:

$$\frac{\partial C}{\partial w^{(L)}} = \frac{\partial C}{\partial a^{(L)}} \frac{\partial a^{(L)}}{\partial z^{(L)}} \frac{\partial z^{(L)}}{\partial w^{(L)}} = 2\left(a^{(L)} - y\right) \sigma'\left(z^{(L)}\right) a^{(L-1)} \tag{4}$$

Thus, we see that this gradient depends on the activation from the previous layer $a^{(L-1)}$. Also wrt to the bias:

$$\frac{\partial C}{\partial b^{(L)}} = \frac{\partial C}{\partial a^{(L)}} \frac{\partial a^{(L)}}{\partial z^{(L)}} \frac{\partial z^{(L)}}{\partial b^{(L)}} = 2\left(a^{(L)} - y\right) \sigma'(z^{(L)})(1) \tag{5}$$

# Equation summary: outer layer

At the outer layer $L$ (without indexing by neuron):

$$z^{(L)} = w^{(L)} \times a^{(L-1)} + b^{(L)} \tag{1}$$

$$a^{(L)} = \sigma(z^{(L)}) \tag{2}$$

$$C = (a^{(L)} - y)^2 \tag{3}$$

The gradient of the cost function with respect to $w^{(L)}$ is:

$$\frac{\partial C}{\partial w^{(L)}} = \frac{\partial C}{\partial a^{(L)}} \frac{\partial a^{(L)}}{\partial z^{(L)}} \frac{\partial z^{(L)}}{\partial w^{(L)}} = 2\left(a^{(L)} - y\right) \sigma'\left(z^{(L)}\right) a^{(L-1)} \tag{4}$$

Thus, we see that this gradient depends on the activation from the previous layer $a^{(L-1)}$. Also wrt to the bias:

$$\frac{\partial C}{\partial b^{(L)}} = \frac{\partial C}{\partial a^{(L)}} \frac{\partial a^{(L)}}{\partial z^{(L)}} \frac{\partial z^{(L)}}{\partial b^{(L)}} = 2\left(a^{(L)} - y\right) \sigma'(z^{(L)})(1) \tag{5}$$

# Equation summary: outer layer

At the outer layer $L$ (without indexing by neuron):

$$z^{(L)} = w^{(L)} \times a^{(L-1)} + b^{(L)} \tag{1}$$

$$a^{(L)} = \sigma(z^{(L)}) \tag{2}$$

$$C = (a^{(L)} - y)^2 \tag{3}$$

The gradient of the cost function with respect to $w^{(L)}$ is:

$$\frac{\partial C}{\partial w^{(L)}} = \frac{\partial C}{\partial a^{(L)}} \frac{\partial a^{(L)}}{\partial z^{(L)}} \frac{\partial z^{(L)}}{\partial w^{(L)}} = 2\left(a^{(L)} - y\right)\sigma'\left(z^{(L)}\right)a^{(L-1)} \tag{4}$$

Thus, we see that this gradient depends on the activation from the previous layer $a^{(L-1)}$. Also wrt to the bias:

$$\frac{\partial C}{\partial b^{(L)}} = \frac{\partial C}{\partial a^{(L)}} \frac{\partial a^{(L)}}{\partial z^{(L)}} \frac{\partial z^{(L)}}{\partial b^{(L)}} = 2\left(a^{(L)} - y\right)\sigma'(z^{(L)})(1) \tag{5}$$

# Equation summary: outer layer

At the outer layer $L$ (without indexing by neuron):

$$z^{(L)} = w^{(L)} \times a^{(L-1)} + b^{(L)} \tag{1}$$

$$a^{(L)} = \sigma(z^{(L)}) \tag{2}$$

$$C = (a^{(L)} - y)^2 \tag{3}$$

The gradient of the cost function with respect to $w^{(L)}$ is:

$$\frac{\partial C}{\partial w^{(L)}} = \frac{\partial C}{\partial a^{(L)}} \frac{\partial a^{(L)}}{\partial z^{(L)}} \frac{\partial z^{(L)}}{\partial w^{(L)}} = 2\left(a^{(L)} - y\right) \sigma'\left(z^{(L)}\right) a^{(L-1)} \tag{4}$$

Thus, we see that this gradient depends on the activation from the previous layer $a^{(L-1)}$. Also wrt to the bias:

$$\frac{\partial C}{\partial b^{(L)}} = \frac{\partial C}{\partial a^{(L)}} \frac{\partial a^{(L)}}{\partial z^{(L)}} \frac{\partial z^{(L)}}{\partial b^{(L)}} = 2\left(a^{(L)} - y\right) \sigma'(z^{(L)})(1) \tag{5}$$

# Equation summary: outer layer

At the outer layer $L$ (without indexing by neuron):

$$z^{(L)} = w^{(L)} \times a^{(L-1)} + b^{(L)} \tag{1}$$

$$a^{(L)} = \sigma(z^{(L)}) \tag{2}$$

$$C = (a^{(L)} - y)^2 \tag{3}$$

The gradient of the cost function with respect to $w^{(L)}$ is:

$$\frac{\partial C}{\partial w^{(L)}} = \frac{\partial C}{\partial a^{(L)}} \frac{\partial a^{(L)}}{\partial z^{(L)}} \frac{\partial z^{(L)}}{\partial w^{(L)}} = 2\left(a^{(L)} - y\right) \sigma' \left(z^{(L)}\right) a^{(L-1)} \tag{4}$$

Thus, we see that this gradient depends on the activation from the previous layer $a^{(L-1)}$. Also wrt to the bias:

$$\frac{\partial C}{\partial b^{(L)}} = \frac{\partial C}{\partial a^{(L)}} \frac{\partial a^{(L)}}{\partial z^{(L)}} \frac{\partial z^{(L)}}{\partial b^{(L)}} = 2\left(a^{(L)} - y\right) \sigma'(z^{(L)})(1) \tag{5}$$

# Equation summary: outer layer

At the outer layer $L$ (without indexing by neuron):

$$z^{(L)} = w^{(L)} \times a^{(L-1)} + b^{(L)} \tag{1}$$

$$a^{(L)} = \sigma(z^{(L)}) \tag{2}$$

$$C = (a^{(L)} - y)^2 \tag{3}$$

The gradient of the cost function with respect to $w^{(L)}$ is:

$$\frac{\partial C}{\partial w^{(L)}} = \frac{\partial C}{\partial a^{(L)}} \frac{\partial a^{(L)}}{\partial z^{(L)}} \frac{\partial z^{(L)}}{\partial w^{(L)}} = 2\left(a^{(L)} - y\right)\sigma'\left(z^{(L)}\right)a^{(L-1)} \tag{4}$$

Thus, we see that this gradient depends on the activation from the previous layer $a^{(L-1)}$. Also wrt to the bias:

$$\frac{\partial C}{\partial b^{(L)}} = \frac{\partial C}{\partial a^{(L)}} \frac{\partial a^{(L)}}{\partial z^{(L)}} \frac{\partial z^{(L)}}{\partial b^{(L)}} = 2\left(a^{(L)} - y\right)\sigma'(z^{(L)})(1) \tag{5}$$

# Equation summary: outer layer

At the outer layer $L$ (without indexing by neuron):

$$z^{(L)} = w^{(L)} \times a^{(L-1)} + b^{(L)} \tag{1}$$

$$a^{(L)} = \sigma(z^{(L)}) \tag{2}$$

$$C = (a^{(L)} - y)^2 \tag{3}$$

The gradient of the cost function with respect to $w^{(L)}$ is:

$$\frac{\partial C}{\partial w^{(L)}} = \frac{\partial C}{\partial a^{(L)}} \frac{\partial a^{(L)}}{\partial z^{(L)}} \frac{\partial z^{(L)}}{\partial w^{(L)}} = 2\left(a^{(L)} - y\right) \sigma' \left(z^{(L)}\right) a^{(L-1)} \tag{4}$$

Thus, we see that this gradient depends on the activation from the previous layer $a^{(L-1)}$. Also wrt to the bias:

$$\frac{\partial C}{\partial b^{(L)}} = \frac{\partial C}{\partial a^{(L)}} \frac{\partial a^{(L)}}{\partial z^{(L)}} \frac{\partial z^{(L)}}{\partial b^{(L)}} = 2\left(a^{(L)} - y\right) \sigma'(z^{(L)})(1) \tag{5}$$

# Equation summary: outer layer

At the outer layer $L$ (without indexing by neuron):

$$z^{(L)} = w^{(L)} \times a^{(L-1)} + b^{(L)} \tag{1}$$

$$a^{(L)} = \sigma(z^{(L)}) \tag{2}$$

$$C = (a^{(L)} - y)^2 \tag{3}$$

The gradient of the cost function with respect to $w^{(L)}$ is:

$$\frac{\partial C}{\partial w^{(L)}} = \frac{\partial C}{\partial a^{(L)}} \frac{\partial a^{(L)}}{\partial z^{(L)}} \frac{\partial z^{(L)}}{\partial w^{(L)}} = 2\left(a^{(L)} - y\right)\sigma'\left(z^{(L)}\right)a^{(L-1)} \tag{4}$$

Thus, we see that this gradient depends on the activation from the previous layer $a^{(L-1)}$. Also wrt to the bias:

$$\frac{\partial C}{\partial b^{(L)}} = \frac{\partial C}{\partial a^{(L)}} \frac{\partial a^{(L)}}{\partial z^{(L)}} \frac{\partial z^{(L)}}{\partial b^{(L)}} = 2\left(a^{(L)} - y\right)\sigma'(z^{(L)})(1) \tag{5}$$

# Updating weights

We can then update the weights for the last layer for the next iteration $r+1$:

$$w^{(L),r+1} = w^{(L),r} - \eta \frac{\partial C}{\partial w^{(L)}} \tag{6}$$

$$b^{(L),r+1} = b^{(L),r} - \eta \frac{\partial C}{\partial b^{(L)}} \tag{7}$$

To update the weights for layer $L-1$, we need to find the gradients $\frac{\partial C}{\partial w^{(L-1)}}$ and $\frac{\partial C}{\partial b^{(L-1)}}$.

Using the chain rule again, we write:

$$\frac{\partial C}{\partial w^{(L-1)}} = \frac{\partial C}{\partial a^{(L-1)}} \frac{\partial a^{(L-1)}}{\partial z^{(L-1)}} \frac{\partial z^{(L-1)}}{\partial w^{(L-1)}} \tag{8}$$

$$\frac{\partial C}{\partial b^{(L-1)}} = \frac{\partial C}{\partial a^{(L-1)}} \frac{\partial a^{(L-1)}}{\partial z^{(L-1)}} \frac{\partial z^{(L-1)}}{\partial b^{(L-1)}} \tag{9}$$

# Updating weights

We can then update the weights for the last layer for the next iteration $r + 1$:

$$w^{(L),r+1} = w^{(L),r} - \eta \frac{\partial C}{\partial w^{(L)}} \tag{6}$$

$$b^{(L),r+1} = b^{(L),r} - \eta \frac{\partial C}{\partial b^{(L)}} \tag{7}$$

To update the weights for layer $L - 1$, we need to find the gradients $\frac{\partial C}{\partial w^{(L-1)}}$ and $\frac{\partial C}{\partial b^{(L-1)}}$.

Using the chain rule again, we write:

$$\frac{\partial C}{\partial w^{(L-1)}} = \frac{\partial C}{\partial a^{(L-1)}} \frac{\partial a^{(L-1)}}{\partial z^{(L-1)}} \frac{\partial z^{(L-1)}}{\partial w^{(L-1)}} \tag{8}$$

$$\frac{\partial C}{\partial b^{(L-1)}} = \frac{\partial C}{\partial a^{(L-1)}} \frac{\partial a^{(L-1)}}{\partial z^{(L-1)}} \frac{\partial z^{(L-1)}}{\partial b^{(L-1)}} \tag{9}$$

# Updating weights

We can then update the weights for the last layer for the next iteration $r + 1$:

$$w^{(L),r+1} = w^{(L),r} - \eta \frac{\partial C}{\partial w^{(L)}} \tag{6}$$

$$b^{(L),r+1} = b^{(L),r} - \eta \frac{\partial C}{\partial b^{(L)}} \tag{7}$$

To update the weights for layer $L - 1$, we need to find the gradients $\frac{\partial C}{\partial w^{(L-1)}}$ and $\frac{\partial C}{\partial b^{(L-1)}}$.

Using the chain rule again, we write:

$$\frac{\partial C}{\partial w^{(L-1)}} = \frac{\partial C}{\partial a^{(L-1)}} \frac{\partial a^{(L-1)}}{\partial z^{(L-1)}} \frac{\partial z^{(L-1)}}{\partial w^{(L-1)}} \tag{8}$$

$$\frac{\partial C}{\partial b^{(L-1)}} = \frac{\partial C}{\partial a^{(L-1)}} \frac{\partial a^{(L-1)}}{\partial z^{(L-1)}} \frac{\partial z^{(L-1)}}{\partial b^{(L-1)}} \tag{9}$$

# Updating weights

We can then update the weights for the last layer for the next iteration $r+1$:

$$w^{(L),r+1} = w^{(L),r} - \eta \frac{\partial C}{\partial w^{(L)}} \tag{6}$$

$$b^{(L),r+1} = b^{(L),r} - \eta \frac{\partial C}{\partial b^{(L)}} \tag{7}$$

To update the weights for layer $L-1$, we need to find the gradients $\frac{\partial C}{\partial w^{(L-1)}}$ and $\frac{\partial C}{\partial b^{(L-1)}}$.

Using the chain rule again, we write:

$$\frac{\partial C}{\partial w^{(L-1)}} = \frac{\partial C}{\partial a^{(L-1)}} \frac{\partial a^{(L-1)}}{\partial z^{(L-1)}} \frac{\partial z^{(L-1)}}{\partial w^{(L-1)}} \tag{8}$$

$$\frac{\partial C}{\partial b^{(L-1)}} = \frac{\partial C}{\partial a^{(L-1)}} \frac{\partial a^{(L-1)}}{\partial z^{(L-1)}} \frac{\partial z^{(L-1)}}{\partial b^{(L-1)}} \tag{9}$$

# Updating weights

We can then update the weights for the last layer for the next iteration $r+1$:

$$w^{(L),r+1} = w^{(L),r} - \eta \frac{\partial C}{\partial w^{(L)}} \tag{6}$$

$$b^{(L),r+1} = b^{(L),r} - \eta \frac{\partial C}{\partial b^{(L)}} \tag{7}$$

To update the weights for layer $L-1$, we need to find the gradients $\frac{\partial C}{\partial w^{(L-1)}}$ and $\frac{\partial C}{\partial b^{(L-1)}}$.

Using the chain rule again, we write:

$$\frac{\partial C}{\partial w^{(L-1)}} = \frac{\partial C}{\partial a^{(L-1)}} \frac{\partial a^{(L-1)}}{\partial z^{(L-1)}} \frac{\partial z^{(L-1)}}{\partial w^{(L-1)}} \tag{8}$$

$$\frac{\partial C}{\partial b^{(L-1)}} = \frac{\partial C}{\partial a^{(L-1)}} \frac{\partial a^{(L-1)}}{\partial z^{(L-1)}} \frac{\partial z^{(L-1)}}{\partial b^{(L-1)}} \tag{9}$$

# Updating weights

We can then update the weights for the last layer for the next iteration $r + 1$:

$$w^{(L),r+1} = w^{(L),r} - \eta \frac{\partial C}{\partial w^{(L)}} \tag{6}$$

$$b^{(L),r+1} = b^{(L),r} - \eta \frac{\partial C}{\partial b^{(L)}} \tag{7}$$

To update the weights for layer $L - 1$, we need to find the gradients $\frac{\partial C}{\partial w^{(L-1)}}$ and $\frac{\partial C}{\partial b^{(L-1)}}$.

Using the chain rule again, we write:

$$\frac{\partial C}{\partial w^{(L-1)}} = \frac{\partial C}{\partial a^{(L-1)}} \frac{\partial a^{(L-1)}}{\partial z^{(L-1)}} \frac{\partial z^{(L-1)}}{\partial w^{(L-1)}} \tag{8}$$

$$\frac{\partial C}{\partial b^{(L-1)}} = \frac{\partial C}{\partial a^{(L-1)}} \frac{\partial a^{(L-1)}}{\partial z^{(L-1)}} \frac{\partial z^{(L-1)}}{\partial b^{(L-1)}} \tag{9}$$

# Updating weights

We can then update the weights for the last layer for the next iteration $r+1$:

$$w^{(L),r+1} = w^{(L),r} - \eta \frac{\partial C}{\partial w^{(L)}} \tag{6}$$

$$b^{(L),r+1} = b^{(L),r} - \eta \frac{\partial C}{\partial b^{(L)}} \tag{7}$$

To update the weights for layer $L-1$, we need to find the gradients $\frac{\partial C}{\partial w^{(L-1)}}$ and $\frac{\partial C}{\partial b^{(L-1)}}$.

Using the chain rule again, we write:

$$\frac{\partial C}{\partial w^{(L-1)}} = \frac{\partial C}{\partial a^{(L-1)}} \frac{\partial a^{(L-1)}}{\partial z^{(L-1)}} \frac{\partial z^{(L-1)}}{\partial w^{(L-1)}} \tag{8}$$

$$\frac{\partial C}{\partial b^{(L-1)}} = \frac{\partial C}{\partial a^{(L-1)}} \frac{\partial a^{(L-1)}}{\partial z^{(L-1)}} \frac{\partial z^{(L-1)}}{\partial b^{(L-1)}} \tag{9}$$

# Updating weights

We can then update the weights for the last layer for the next iteration $r + 1$:

$$w^{(L),r+1} = w^{(L),r} - \eta \frac{\partial C}{\partial w^{(L)}} \tag{6}$$

$$b^{(L),r+1} = b^{(L),r} - \eta \frac{\partial C}{\partial b^{(L)}} \tag{7}$$

To update the weights for layer $L - 1$, we need to find the gradients $\frac{\partial C}{\partial w^{(L-1)}}$ and $\frac{\partial C}{\partial b^{(L-1)}}$.

Using the chain rule again, we write:

$$\frac{\partial C}{\partial w^{(L-1)}} = \frac{\partial C}{\partial a^{(L-1)}} \frac{\partial a^{(L-1)}}{\partial z^{(L-1)}} \frac{\partial z^{(L-1)}}{\partial w^{(L-1)}} \tag{8}$$

$$\frac{\partial C}{\partial b^{(L-1)}} = \frac{\partial C}{\partial a^{(L-1)}} \frac{\partial a^{(L-1)}}{\partial z^{(L-1)}} \frac{\partial z^{(L-1)}}{\partial b^{(L-1)}} \tag{9}$$

# Backward pass

But we recall that $C$ is not *explicitly* dependent on $a^{(L-1)}$ as $C = (a^{(L)} - y)^2$. However, it is *implicitly* dependent, since

$$C \propto a^{(L)}, \tag{10}$$

$$a^{(L)} \propto z^{(L)} \tag{11}$$

and

$$z^{(L)} \propto a^{(L-1)} \tag{12}$$

So, we use the chain rule to expand $\frac{\partial C}{\partial a^{(L-1)}}$ as follows:

$$\frac{\partial C}{\partial a^{(L-1)}} = \frac{\partial C}{\partial a^{(L)}} \frac{\partial a^{(L)}}{\partial z^{(L)}} \frac{\partial z^{(L)}}{\partial a^{(L-1)}} \tag{13}$$

# Backward pass

But we recall that $C$ is not *explicitly* dependent on $a^{(L-1)}$ as $C = (a^{(L)} - y)^2$.
However, it is *implicitly* dependent, since

$$C \propto a^{(L)}, \tag{10}$$

$$a^{(L)} \propto z^{(L)} \tag{11}$$

and

$$z^{(L)} \propto a^{(L-1)} \tag{12}$$

So, we use the chain rule to expand $\frac{\partial C}{\partial a^{(L-1)}}$ as follows:

$$\frac{\partial C}{\partial a^{(L-1)}} = \frac{\partial C}{\partial a^{(L)}} \frac{\partial a^{(L)}}{\partial z^{(L)}} \frac{\partial z^{(L)}}{\partial a^{(L-1)}} \tag{13}$$

# Backward pass

But we recall that $C$ is not *explicitly* dependent on $a^{(L-1)}$ as $C = (a^{(L)} - y)^2$. However, it is *implicitly* dependent, since

$$C \propto a^{(L)}, \tag{10}$$

$$a^{(L)} \propto z^{(L)} \tag{11}$$

and

$$z^{(L)} \propto a^{(L-1)} \tag{12}$$

So, we use the chain rule to expand $\frac{\partial C}{\partial a^{(L-1)}}$ as follows:

$$\frac{\partial C}{\partial a^{(L-1)}} = \frac{\partial C}{\partial a^{(L)}} \frac{\partial a^{(L)}}{\partial z^{(L)}} \frac{\partial z^{(L)}}{\partial a^{(L-1)}} \tag{13}$$

# Backward pass

But we recall that $C$ is not *explicitly* dependent on $a^{(L-1)}$ as $C = (a^{(L)} - y)^2$. However, it is *implicitly* dependent, since

$$C \propto a^{(L)}, \tag{10}$$

$$a^{(L)} \propto z^{(L)} \tag{11}$$

and

$$z^{(L)} \propto a^{(L-1)} \tag{12}$$

So, we use the chain rule to expand $\frac{\partial C}{\partial a^{(L-1)}}$ as follows:

$$\frac{\partial C}{\partial a^{(L-1)}} = \frac{\partial C}{\partial a^{(L)}} \frac{\partial a^{(L)}}{\partial z^{(L)}} \frac{\partial z^{(L)}}{\partial a^{(L-1)}} \tag{13}$$

# Backward pass

But we recall that $C$ is not *explicitly* dependent on $a^{(L-1)}$ as $C = (a^{(L)} - y)^2$. However, it is *implicitly* dependent, since

$$C \propto a^{(L)}, \tag{10}$$

$$a^{(L)} \propto z^{(L)} \tag{11}$$

and

$$z^{(L)} \propto a^{(L-1)} \tag{12}$$

So, we use the chain rule to expand $\frac{\partial C}{\partial a^{(L-1)}}$ as follows:

$$\frac{\partial C}{\partial a^{(L-1)}} = \frac{\partial C}{\partial a^{(L)}} \frac{\partial a^{(L)}}{\partial z^{(L)}} \frac{\partial z^{(L)}}{\partial a^{(L-1)}} \tag{13}$$

# Backward pass

But we recall that $C$ is not *explicitly* dependent on $a^{(L-1)}$ as $C = (a^{(L)} - y)^2$. However, it is *implicitly* dependent, since

$$C \propto a^{(L)}, \tag{10}$$

$$a^{(L)} \propto z^{(L)} \tag{11}$$

and

$$z^{(L)} \propto a^{(L-1)} \tag{12}$$

So, we use the chain rule to expand $\frac{\partial C}{\partial a^{(L-1)}}$ as follows:

$$\frac{\partial C}{\partial a^{(L-1)}} = \frac{\partial C}{\partial a^{(L)}} \frac{\partial a^{(L)}}{\partial z^{(L)}} \frac{\partial z^{(L)}}{\partial a^{(L-1)}} \tag{13}$$

# Backward pass

But we recall that $C$ is not *explicitly* dependent on $a^{(L-1)}$ as $C = (a^{(L)} - y)^2$. However, it is *implicitly* dependent, since

$$C \propto a^{(L)}, \tag{10}$$

$$a^{(L)} \propto z^{(L)} \tag{11}$$

and

$$z^{(L)} \propto a^{(L-1)} \tag{12}$$

So, we use the chain rule to expand $\frac{\partial C}{\partial a^{(L-1)}}$ as follows:

$$\frac{\partial C}{\partial a^{(L-1)}} = \frac{\partial C}{\partial a^{(L)}} \frac{\partial a^{(L)}}{\partial z^{(L)}} \frac{\partial z^{(L)}}{\partial a^{(L-1)}} \tag{13}$$

# Backward pass

But we recall that $C$ is not *explicitly* dependent on $a^{(L-1)}$ as $C = (a^{(L)} - y)^2$. However, it is *implicitly* dependent, since

$$C \propto a^{(L)}, \tag{10}$$

$$a^{(L)} \propto z^{(L)} \tag{11}$$

and

$$z^{(L)} \propto a^{(L-1)} \tag{12}$$

So, we use the chain rule to expand $\frac{\partial C}{\partial a^{(L-1)}}$ as follows:

$$\frac{\partial C}{\partial a^{(L-1)}} = \frac{\partial C}{\partial a^{(L)}} \frac{\partial a^{(L)}}{\partial z^{(L)}} \frac{\partial z^{(L)}}{\partial a^{(L-1)}} \tag{13}$$

# Backward pass

But we recall that $C$ is not *explicitly* dependent on $a^{(L-1)}$ as $C = (a^{(L)} - y)^2$. However, it is *implicitly* dependent, since

$$C \propto a^{(L)}, \tag{10}$$

$$a^{(L)} \propto z^{(L)} \tag{11}$$

and

$$z^{(L)} \propto a^{(L-1)} \tag{12}$$

So, we use the chain rule to expand $\frac{\partial C}{\partial a^{(L-1)}}$ as follows:

$$\frac{\partial C}{\partial a^{(L-1)}} = \frac{\partial C}{\partial a^{(L)}} \frac{\partial a^{(L)}}{\partial z^{(L)}} \frac{\partial z^{(L)}}{\partial a^{(L-1)}} \tag{13}$$

# Backward pass

But we recall that $C$ is not *explicitly* dependent on $a^{(L-1)}$ as $C = (a^{(L)} - y)^2$. However, it is *implicitly* dependent, since

$$C \propto a^{(L)}, \tag{10}$$

$$a^{(L)} \propto z^{(L)} \tag{11}$$

and

$$z^{(L)} \propto a^{(L-1)} \tag{12}$$

So, we use the chain rule to expand $\frac{\partial C}{\partial a^{(L-1)}}$ as follows:

$$\frac{\partial C}{\partial a^{(L-1)}} = \frac{\partial C}{\partial a^{(L)}} \frac{\partial a^{(L)}}{\partial z^{(L)}} \frac{\partial z^{(L)}}{\partial a^{(L-1)}} \tag{13}$$

# Backward pass (cont.)

We can then expand the cost function gradient wrt to weights for layer $L-1$ as:

$$\frac{\partial C}{\partial w^{(L-1)}} = \frac{\partial C}{\partial a^{(L-1)}} \frac{\partial a^{(L-1)}}{\partial z^{(L-1)}} \frac{\partial z^{(L-1)}}{\partial w^{(L-1)}} = \frac{\partial C}{\partial a^{(L)}} \frac{\partial a^{(L)}}{\partial z^{(L)}} \frac{\partial z^{(L)}}{\partial a^{(L-1)}} \frac{\partial a^{(L-1)}}{\partial z^{(L-1)}} \frac{\partial z^{(L-1)}}{\partial w^{(L-1)}} \tag{14}$$

$$\frac{\partial C}{\partial b^{(L-1)}} = \frac{\partial C}{\partial a^{(L-1)}} \frac{\partial a^{(L-1)}}{\partial z^{(L-1)}} \frac{\partial z^{(L-1)}}{\partial b^{(L-1)}} = \frac{\partial C}{\partial a^{(L)}} \frac{\partial a^{(L)}}{\partial z^{(L)}} \frac{\partial z^{(L)}}{\partial a^{(L-1)}} \frac{\partial a^{(L-1)}}{\partial z^{(L-1)}} \frac{\partial z^{(L-1)}}{\partial b^{(L-1)}} \tag{15}$$

Once these gradients are computed, we update the weights for the $r+1$th iteration using:

$$w^{(L-1),r+1} = w^{(L-1),r} - \eta \frac{\partial C}{\partial w^{(L-1)}} \tag{16}$$

$$b^{(L-1),r+1} = b^{(L-1),r} - \eta \frac{\partial C}{\partial b^{(L-1)}} \tag{17}$$

# Backward pass (cont.)

We can then expand the cost function gradient wrt to weights for layer $L-1$ as:

$$\frac{\partial C}{\partial w^{(L-1)}} = \frac{\partial C}{\partial a^{(L-1)}} \frac{\partial a^{(L-1)}}{\partial z^{(L-1)}} \frac{\partial z^{(L-1)}}{\partial w^{(L-1)}} = \frac{\partial C}{\partial a^{(L)}} \frac{\partial a^{(L)}}{\partial z^{(L)}} \frac{\partial z^{(L)}}{\partial a^{(L-1)}} \frac{\partial a^{(L-1)}}{\partial z^{(L-1)}} \frac{\partial z^{(L-1)}}{\partial w^{(L-1)}} \tag{14}$$

$$\frac{\partial C}{\partial b^{(L-1)}} = \frac{\partial C}{\partial a^{(L-1)}} \frac{\partial a^{(L-1)}}{\partial z^{(L-1)}} \frac{\partial z^{(L-1)}}{\partial b^{(L-1)}} = \frac{\partial C}{\partial a^{(L)}} \frac{\partial a^{(L)}}{\partial z^{(L)}} \frac{\partial z^{(L)}}{\partial a^{(L-1)}} \frac{\partial a^{(L-1)}}{\partial z^{(L-1)}} \frac{\partial z^{(L-1)}}{\partial b^{(L-1)}} \tag{15}$$

Once these gradients are computed, we update the weights for the $r+1$th iteration using:

$$w^{(L-1),r+1} = w^{(L-1),r} - \eta \frac{\partial C}{\partial w^{(L-1)}} \tag{16}$$

$$b^{(L-1),r+1} = b^{(L-1),r} - \eta \frac{\partial C}{\partial b^{(L-1)}} \tag{17}$$

# Backward pass (cont.)

We can then expand the cost function gradient wrt to weights for layer $L-1$ as:

$$\frac{\partial C}{\partial w^{(L-1)}} = \frac{\partial C}{\partial a^{(L-1)}} \frac{\partial a^{(L-1)}}{\partial z^{(L-1)}} \frac{\partial z^{(L-1)}}{\partial w^{(L-1)}} = \frac{\partial C}{\partial a^{(L)}} \frac{\partial a^{(L)}}{\partial z^{(L)}} \frac{\partial z^{(L)}}{\partial a^{(L-1)}} \frac{\partial a^{(L-1)}}{\partial z^{(L-1)}} \frac{\partial z^{(L-1)}}{\partial w^{(L-1)}} \tag{14}$$

$$\frac{\partial C}{\partial b^{(L-1)}} = \frac{\partial C}{\partial a^{(L-1)}} \frac{\partial a^{(L-1)}}{\partial z^{(L-1)}} \frac{\partial z^{(L-1)}}{\partial b^{(L-1)}} = \frac{\partial C}{\partial a^{(L)}} \frac{\partial a^{(L)}}{\partial z^{(L)}} \frac{\partial z^{(L)}}{\partial a^{(L-1)}} \frac{\partial a^{(L-1)}}{\partial z^{(L-1)}} \frac{\partial z^{(L-1)}}{\partial b^{(L-1)}} \tag{15}$$

Once these gradients are computed, we update the weights for the $r+1$th iteration using:

$$w^{(L-1),r+1} = w^{(L-1),r} - \eta \frac{\partial C}{\partial w^{(L-1)}} \tag{16}$$

$$b^{(L-1),r+1} = b^{(L-1),r} - \eta \frac{\partial C}{\partial b^{(L-1)}} \tag{17}$$

# Backward pass (cont.)

We can then expand the cost function gradient wrt to weights for layer $L-1$ as:

$$\frac{\partial C}{\partial w^{(L-1)}} = \frac{\partial C}{\partial a^{(L-1)}} \frac{\partial a^{(L-1)}}{\partial z^{(L-1)}} \frac{\partial z^{(L-1)}}{\partial w^{(L-1)}} = \frac{\partial C}{\partial a^{(L)}} \frac{\partial a^{(L)}}{\partial z^{(L)}} \frac{\partial z^{(L)}}{\partial a^{(L-1)}} \frac{\partial a^{(L-1)}}{\partial z^{(L-1)}} \frac{\partial z^{(L-1)}}{\partial w^{(L-1)}} \tag{14}$$

$$\frac{\partial C}{\partial b^{(L-1)}} = \frac{\partial C}{\partial a^{(L-1)}} \frac{\partial a^{(L-1)}}{\partial z^{(L-1)}} \frac{\partial z^{(L-1)}}{\partial b^{(L-1)}} = \frac{\partial C}{\partial a^{(L)}} \frac{\partial a^{(L)}}{\partial z^{(L)}} \frac{\partial z^{(L)}}{\partial a^{(L-1)}} \frac{\partial a^{(L-1)}}{\partial z^{(L-1)}} \frac{\partial z^{(L-1)}}{\partial b^{(L-1)}} \tag{15}$$

Once these gradients are computed, we update the weights for the $r+1$th iteration using:

$$w^{(L-1),r+1} = w^{(L-1),r} - \eta \frac{\partial C}{\partial w^{(L-1)}} \tag{16}$$

$$b^{(L-1),r+1} = b^{(L-1),r} - \eta \frac{\partial C}{\partial b^{(L-1)}} \tag{17}$$

# Backward pass (cont.)

We can then expand the cost function gradient wrt to weights for layer $L-1$ as:

$$\frac{\partial C}{\partial w^{(L-1)}} = \frac{\partial C}{\partial a^{(L-1)}} \frac{\partial a^{(L-1)}}{\partial z^{(L-1)}} \frac{\partial z^{(L-1)}}{\partial w^{(L-1)}} = \frac{\partial C}{\partial a^{(L)}} \frac{\partial a^{(L)}}{\partial z^{(L)}} \frac{\partial z^{(L)}}{\partial a^{(L-1)}} \frac{\partial a^{(L-1)}}{\partial z^{(L-1)}} \frac{\partial z^{(L-1)}}{\partial w^{(L-1)}} \tag{14}$$

$$\frac{\partial C}{\partial b^{(L-1)}} = \frac{\partial C}{\partial a^{(L-1)}} \frac{\partial a^{(L-1)}}{\partial z^{(L-1)}} \frac{\partial z^{(L-1)}}{\partial b^{(L-1)}} = \frac{\partial C}{\partial a^{(L)}} \frac{\partial a^{(L)}}{\partial z^{(L)}} \frac{\partial z^{(L)}}{\partial a^{(L-1)}} \frac{\partial a^{(L-1)}}{\partial z^{(L-1)}} \frac{\partial z^{(L-1)}}{\partial b^{(L-1)}} \tag{15}$$

Once these gradients are computed, we update the weights for the $r+1$th iteration using:

$$w^{(L-1),r+1} = w^{(L-1),r} - \eta \frac{\partial C}{\partial w^{(L-1)}} \tag{16}$$

$$b^{(L-1),r+1} = b^{(L-1),r} - \eta \frac{\partial C}{\partial b^{(L-1)}} \tag{17}$$

# Backward pass (cont.)

We can then expand the cost function gradient wrt to weights for layer $L-1$ as:

$$\frac{\partial C}{\partial w^{(L-1)}} = \frac{\partial C}{\partial a^{(L-1)}} \frac{\partial a^{(L-1)}}{\partial z^{(L-1)}} \frac{\partial z^{(L-1)}}{\partial w^{(L-1)}} = \frac{\partial C}{\partial a^{(L)}} \frac{\partial a^{(L)}}{\partial z^{(L)}} \frac{\partial z^{(L)}}{\partial a^{(L-1)}} \frac{\partial a^{(L-1)}}{\partial z^{(L-1)}} \frac{\partial z^{(L-1)}}{\partial w^{(L-1)}} \tag{14}$$

$$\frac{\partial C}{\partial b^{(L-1)}} = \frac{\partial C}{\partial a^{(L-1)}} \frac{\partial a^{(L-1)}}{\partial z^{(L-1)}} \frac{\partial z^{(L-1)}}{\partial b^{(L-1)}} = \frac{\partial C}{\partial a^{(L)}} \frac{\partial a^{(L)}}{\partial z^{(L)}} \frac{\partial z^{(L)}}{\partial a^{(L-1)}} \frac{\partial a^{(L-1)}}{\partial z^{(L-1)}} \frac{\partial z^{(L-1)}}{\partial b^{(L-1)}} \tag{15}$$

Once these gradients are computed, we update the weights for the $r+1$th iteration using:

$$w^{(L-1),r+1} = w^{(L-1),r} - \eta \frac{\partial C}{\partial w^{(L-1)}} \tag{16}$$

$$b^{(L-1),r+1} = b^{(L-1),r} - \eta \frac{\partial C}{\partial b^{(L-1)}} \tag{17}$$

# Backward pass (cont.)

We can then expand the cost function gradient wrt to weights for layer $L-1$ as:

$$\frac{\partial C}{\partial w^{(L-1)}} = \frac{\partial C}{\partial a^{(L-1)}} \frac{\partial a^{(L-1)}}{\partial z^{(L-1)}} \frac{\partial z^{(L-1)}}{\partial w^{(L-1)}} = \frac{\partial C}{\partial a^{(L)}} \frac{\partial a^{(L)}}{\partial z^{(L)}} \frac{\partial z^{(L)}}{\partial a^{(L-1)}} \frac{\partial a^{(L-1)}}{\partial z^{(L-1)}} \frac{\partial z^{(L-1)}}{\partial w^{(L-1)}} \tag{14}$$

$$\frac{\partial C}{\partial b^{(L-1)}} = \frac{\partial C}{\partial a^{(L-1)}} \frac{\partial a^{(L-1)}}{\partial z^{(L-1)}} \frac{\partial z^{(L-1)}}{\partial b^{(L-1)}} = \frac{\partial C}{\partial a^{(L)}} \frac{\partial a^{(L)}}{\partial z^{(L)}} \frac{\partial z^{(L)}}{\partial a^{(L-1)}} \frac{\partial a^{(L-1)}}{\partial z^{(L-1)}} \frac{\partial z^{(L-1)}}{\partial b^{(L-1)}} \tag{15}$$

Once these gradients are computed, we update the weights for the $r+1$th iteration using:

$$w^{(L-1),r+1} = w^{(L-1),r} - \eta \frac{\partial C}{\partial w^{(L-1)}} \tag{16}$$

$$b^{(L-1),r+1} = b^{(L-1),r} - \eta \frac{\partial C}{\partial b^{(L-1)}} \tag{17}$$

# Backward pass (cont.)

We can then expand the cost function gradient wrt to weights for layer $L-1$ as:

$$\frac{\partial C}{\partial w^{(L-1)}} = \frac{\partial C}{\partial a^{(L-1)}} \frac{\partial a^{(L-1)}}{\partial z^{(L-1)}} \frac{\partial z^{(L-1)}}{\partial w^{(L-1)}} = \frac{\partial C}{\partial a^{(L)}} \frac{\partial a^{(L)}}{\partial z^{(L)}} \frac{\partial z^{(L)}}{\partial a^{(L-1)}} \frac{\partial a^{(L-1)}}{\partial z^{(L-1)}} \frac{\partial z^{(L-1)}}{\partial w^{(L-1)}} \tag{14}$$

$$\frac{\partial C}{\partial b^{(L-1)}} = \frac{\partial C}{\partial a^{(L-1)}} \frac{\partial a^{(L-1)}}{\partial z^{(L-1)}} \frac{\partial z^{(L-1)}}{\partial b^{(L-1)}} = \frac{\partial C}{\partial a^{(L)}} \frac{\partial a^{(L)}}{\partial z^{(L)}} \frac{\partial z^{(L)}}{\partial a^{(L-1)}} \frac{\partial a^{(L-1)}}{\partial z^{(L-1)}} \frac{\partial z^{(L-1)}}{\partial b^{(L-1)}} \tag{15}$$

Once these gradients are computed, we update the weights for the $r+1$th iteration using:

$$w^{(L-1),r+1} = w^{(L-1),r} - \eta \frac{\partial C}{\partial w^{(L-1)}} \tag{16}$$

$$b^{(L-1),r+1} = b^{(L-1),r} - \eta \frac{\partial C}{\partial b^{(L-1)}} \tag{17}$$

# Backward pass (cont.)

We can then expand the cost function gradient wrt to weights for layer $L-1$ as:

$$\frac{\partial C}{\partial w^{(L-1)}} = \frac{\partial C}{\partial a^{(L-1)}} \frac{\partial a^{(L-1)}}{\partial z^{(L-1)}} \frac{\partial z^{(L-1)}}{\partial w^{(L-1)}} = \frac{\partial C}{\partial a^{(L)}} \frac{\partial a^{(L)}}{\partial z^{(L)}} \frac{\partial z^{(L)}}{\partial a^{(L-1)}} \frac{\partial a^{(L-1)}}{\partial z^{(L-1)}} \frac{\partial z^{(L-1)}}{\partial w^{(L-1)}} \tag{14}$$

$$\frac{\partial C}{\partial b^{(L-1)}} = \frac{\partial C}{\partial a^{(L-1)}} \frac{\partial a^{(L-1)}}{\partial z^{(L-1)}} \frac{\partial z^{(L-1)}}{\partial b^{(L-1)}} = \frac{\partial C}{\partial a^{(L)}} \frac{\partial a^{(L)}}{\partial z^{(L)}} \frac{\partial z^{(L)}}{\partial a^{(L-1)}} \frac{\partial a^{(L-1)}}{\partial z^{(L-1)}} \frac{\partial z^{(L-1)}}{\partial b^{(L-1)}} \tag{15}$$

Once these gradients are computed, we update the weights for the $r+1$th iteration using:

$$w^{(L-1),r+1} = w^{(L-1),r} - \eta \frac{\partial C}{\partial w^{(L-1)}} \tag{16}$$

$$b^{(L-1),r+1} = b^{(L-1),r} - \eta \frac{\partial C}{\partial b^{(L-1)}} \tag{17}$$

# Backward pass (cont.)

We can then expand the cost function gradient wrt to weights for layer $L-1$ as:

$$\frac{\partial C}{\partial w^{(L-1)}} = \frac{\partial C}{\partial a^{(L-1)}} \frac{\partial a^{(L-1)}}{\partial z^{(L-1)}} \frac{\partial z^{(L-1)}}{\partial w^{(L-1)}} = \frac{\partial C}{\partial a^{(L)}} \frac{\partial a^{(L)}}{\partial z^{(L)}} \frac{\partial z^{(L)}}{\partial a^{(L-1)}} \frac{\partial a^{(L-1)}}{\partial z^{(L-1)}} \frac{\partial z^{(L-1)}}{\partial w^{(L-1)}} \tag{14}$$

$$\frac{\partial C}{\partial b^{(L-1)}} = \frac{\partial C}{\partial a^{(L-1)}} \frac{\partial a^{(L-1)}}{\partial z^{(L-1)}} \frac{\partial z^{(L-1)}}{\partial b^{(L-1)}} = \frac{\partial C}{\partial a^{(L)}} \frac{\partial a^{(L)}}{\partial z^{(L)}} \frac{\partial z^{(L)}}{\partial a^{(L-1)}} \frac{\partial a^{(L-1)}}{\partial z^{(L-1)}} \frac{\partial z^{(L-1)}}{\partial b^{(L-1)}} \tag{15}$$

Once these gradients are computed, we update the weights for the $r+1$th iteration using:

$$w^{(L-1),r+1} = w^{(L-1),r} - \eta \frac{\partial C}{\partial w^{(L-1)}} \tag{16}$$

$$b^{(L-1),r+1} = b^{(L-1),r} - \eta \frac{\partial C}{\partial b^{(L-1)}} \tag{17}$$

# Backward pass (cont.)

We can then expand the cost function gradient wrt to weights for layer $L-1$ as:

$$\frac{\partial C}{\partial w^{(L-1)}} = \frac{\partial C}{\partial a^{(L-1)}} \frac{\partial a^{(L-1)}}{\partial z^{(L-1)}} \frac{\partial z^{(L-1)}}{\partial w^{(L-1)}} = \frac{\partial C}{\partial a^{(L)}} \frac{\partial a^{(L)}}{\partial z^{(L)}} \frac{\partial z^{(L)}}{\partial a^{(L-1)}} \frac{\partial a^{(L-1)}}{\partial z^{(L-1)}} \frac{\partial z^{(L-1)}}{\partial w^{(L-1)}} \tag{14}$$

$$\frac{\partial C}{\partial b^{(L-1)}} = \frac{\partial C}{\partial a^{(L-1)}} \frac{\partial a^{(L-1)}}{\partial z^{(L-1)}} \frac{\partial z^{(L-1)}}{\partial b^{(L-1)}} = \frac{\partial C}{\partial a^{(L)}} \frac{\partial a^{(L)}}{\partial z^{(L)}} \frac{\partial z^{(L)}}{\partial a^{(L-1)}} \frac{\partial a^{(L-1)}}{\partial z^{(L-1)}} \frac{\partial z^{(L-1)}}{\partial b^{(L-1)}} \tag{15}$$

Once these gradients are computed, we update the weights for the $r+1$th iteration using:

$$w^{(L-1),r+1} = w^{(L-1),r} - \eta \frac{\partial C}{\partial w^{(L-1)}} \tag{16}$$

$$b^{(L-1),r+1} = b^{(L-1),r} - \eta \frac{\partial C}{\partial b^{(L-1)}} \tag{17}$$

# Backward pass (cont.)

We can then expand the cost function gradient wrt to weights for layer $L-1$ as:

$$\frac{\partial C}{\partial w^{(L-1)}} = \frac{\partial C}{\partial a^{(L-1)}} \frac{\partial a^{(L-1)}}{\partial z^{(L-1)}} \frac{\partial z^{(L-1)}}{\partial w^{(L-1)}} = \frac{\partial C}{\partial a^{(L)}} \frac{\partial a^{(L)}}{\partial z^{(L)}} \frac{\partial z^{(L)}}{\partial a^{(L-1)}} \frac{\partial a^{(L-1)}}{\partial z^{(L-1)}} \frac{\partial z^{(L-1)}}{\partial w^{(L-1)}} \tag{14}$$

$$\frac{\partial C}{\partial b^{(L-1)}} = \frac{\partial C}{\partial a^{(L-1)}} \frac{\partial a^{(L-1)}}{\partial z^{(L-1)}} \frac{\partial z^{(L-1)}}{\partial b^{(L-1)}} = \frac{\partial C}{\partial a^{(L)}} \frac{\partial a^{(L)}}{\partial z^{(L)}} \frac{\partial z^{(L)}}{\partial a^{(L-1)}} \frac{\partial a^{(L-1)}}{\partial z^{(L-1)}} \frac{\partial z^{(L-1)}}{\partial b^{(L-1)}} \tag{15}$$

Once these gradients are computed, we update the weights for the $r+1$th iteration using:

$$w^{(L-1),r+1} = w^{(L-1),r} - \eta \frac{\partial C}{\partial w^{(L-1)}} \tag{16}$$

$$b^{(L-1),r+1} = b^{(L-1),r} - \eta \frac{\partial C}{\partial b^{(L-1)}} \tag{17}$$

# Summary: forward pass

1. Initialize weights and biases: $w^{(l),0}$, $b^{(l),0}$
2. Perform forward pass to compute activations:

$$z^{(l),0} = w^{(l),0} \times a^{(l-1),0} + b^{(l),0} \tag{18}$$

$$a^{(l)} = \sigma(z^{(l),0}) \tag{19}$$

At output layer:

$$z^{(L),0} = w^{(L),0} \times a^{(L-1),0} + b^{(L),0} \tag{20}$$

$$a^{(L),0} = \sigma(z^{(L),0}) \tag{21}$$

$$C = (a^{(L),0} - y)^2 \tag{22}$$

# Summary: forward pass

1. Initialize weights and biases: $w^{(l),0}$, $b^{(l),0}$
2. Perform forward pass to compute activations:

$$z^{(l),0} = w^{(l),0} \times a^{(l-1),0} + b^{(l),0} \tag{18}$$
$$a^{(l)} = \sigma(z^{(l),0}) \tag{19}$$

At output layer:

$$z^{(L),0} = w^{(L),0} \times a^{(L-1),0} + b^{(L),0} \tag{20}$$
$$a^{(L),0} = \sigma(z^{(L),0}) \tag{21}$$
$$C = (a^{(L),0} - y)^2 \tag{22}$$

# Summary: forward pass

1. Initialize weights and biases: $w^{(l),0}$, $b^{(l),0}$
2. Perform forward pass to compute activations:

$$z^{(l),0} = w^{(l),0} \times a^{(l-1),0} + b^{(l),0} \tag{18}$$

$$a^{(l)} = \sigma(z^{(l),0}) \tag{19}$$

At output layer:

$$z^{(L),0} = w^{(L),0} \times a^{(L-1),0} + b^{(L),0} \tag{20}$$

$$a^{(L),0} = \sigma(z^{(L),0}) \tag{21}$$

$$C = (a^{(L),0} - y)^2 \tag{22}$$

# Summary: forward pass

1. Initialize weights and biases: $w^{(l),0}$, $b^{(l),0}$
2. Perform forward pass to compute activations:

$$z^{(l),0} = w^{(l),0} \times a^{(l-1),0} + b^{(l),0} \tag{18}$$

$$a^{(l)} = \sigma(z^{(l),0}) \tag{19}$$

At output layer:

$$z^{(L),0} = w^{(L),0} \times a^{(L-1),0} + b^{(L),0} \tag{20}$$

$$a^{(L),0} = \sigma(z^{(L),0}) \tag{21}$$

$$C = (a^{(L),0} - y)^2 \tag{22}$$

# Summary: forward pass

1. Initialize weights and biases: $w^{(l),0}$, $b^{(l),0}$
2. Perform forward pass to compute activations:

$$z^{(l),0} = w^{(l),0} \times a^{(l-1),0} + b^{(l),0} \tag{18}$$
$$a^{(l)} = \sigma(z^{(l),0}) \tag{19}$$

At output layer:

$$z^{(L),0} = w^{(L),0} \times a^{(L-1),0} + b^{(L),0} \tag{20}$$
$$a^{(L),0} = \sigma(z^{(L),0}) \tag{21}$$
$$C = (a^{(L),0} - y)^2 \tag{22}$$

# Summary: forward pass

1. Initialize weights and biases: $w^{(l),0}$, $b^{(l),0}$
2. Perform forward pass to compute activations:

$$z^{(l),0} = w^{(l),0} \times a^{(l-1),0} + b^{(l),0} \tag{18}$$

$$a^{(l)} = \sigma(z^{(l),0}) \tag{19}$$

At output layer:

$$z^{(L),0} = w^{(L),0} \times a^{(L-1),0} + b^{(L),0} \tag{20}$$

$$a^{(L),0} = \sigma(z^{(L),0}) \tag{21}$$

$$C = (a^{(L),0} - y)^2 \tag{22}$$

# Summary: backward pass—outer layer

3. Backward pass, outer layer ($L$):
    1. Compute gradients:

$$\frac{\partial C}{\partial w^{(L)}} = \frac{\partial C}{\partial a^{(L)}} \frac{\partial a^{(L)}}{\partial z^{(L)}} \frac{\partial z^{(L)}}{\partial w^{(L)}} \tag{23}$$

$$\frac{\partial C}{\partial b^{(L)}} = \frac{\partial C}{\partial a^{(L)}} \frac{\partial a^{(L)}}{\partial z^{(L)}} \frac{\partial z^{(L)}}{\partial b^{(L)}} \tag{24}$$

    2. Update weights:

$$w^{(L),1} = w^{(L),0} - \eta \frac{\partial C^0}{\partial w^{(L)}} \tag{25}$$

$$b^{(L),1} = b^{(L),0} - \eta \frac{\partial C^0}{\partial b^{(L)}} \tag{26}$$

# Summary: backward pass—outer layer

3. Backward pass, outer layer ($L$):
   1. Compute gradients:

$$\frac{\partial C}{\partial w^{(L)}} = \frac{\partial C}{\partial a^{(L)}} \frac{\partial a^{(L)}}{\partial z^{(L)}} \frac{\partial z^{(L)}}{\partial w^{(L)}} \tag{23}$$

$$\frac{\partial C}{\partial b^{(L)}} = \frac{\partial C}{\partial a^{(L)}} \frac{\partial a^{(L)}}{\partial z^{(L)}} \frac{\partial z^{(L)}}{\partial b^{(L)}} \tag{24}$$

   2. Update weights:

$$w^{(L),1} = w^{(L),0} - \eta \frac{\partial C^0}{\partial w^{(L)}} \tag{25}$$

$$b^{(L),1} = b^{(L),0} - \eta \frac{\partial C^0}{\partial b^{(L)}} \tag{26}$$

# Summary: backward pass—outer layer

3. Backward pass, outer layer ($L$):
   1. Compute gradients:

$$\frac{\partial C}{\partial w^{(L)}} = \frac{\partial C}{\partial a^{(L)}} \frac{\partial a^{(L)}}{\partial z^{(L)}} \frac{\partial z^{(L)}}{\partial w^{(L)}} \tag{23}$$

$$\frac{\partial C}{\partial b^{(L)}} = \frac{\partial C}{\partial a^{(L)}} \frac{\partial a^{(L)}}{\partial z^{(L)}} \frac{\partial z^{(L)}}{\partial b^{(L)}} \tag{24}$$

   2. Update weights:

$$w^{(L),1} = w^{(L),0} - \eta \frac{\partial C^0}{\partial w^{(L)}} \tag{25}$$

$$b^{(L),1} = b^{(L),0} - \eta \frac{\partial C^0}{\partial b^{(L)}} \tag{26}$$

# Summary: backward pass—outer layer

3. Backward pass, outer layer ($L$):
    1. Compute gradients:

$$\frac{\partial C}{\partial w^{(L)}} = \frac{\partial C}{\partial a^{(L)}} \frac{\partial a^{(L)}}{\partial z^{(L)}} \frac{\partial z^{(L)}}{\partial w^{(L)}} \tag{23}$$

$$\frac{\partial C}{\partial b^{(L)}} = \frac{\partial C}{\partial a^{(L)}} \frac{\partial a^{(L)}}{\partial z^{(L)}} \frac{\partial z^{(L)}}{\partial b^{(L)}} \tag{24}$$

    2. Update weights:

$$w^{(L),1} = w^{(L),0} - \eta \frac{\partial C^0}{\partial w^{(L)}} \tag{25}$$

$$b^{(L),1} = b^{(L),0} - \eta \frac{\partial C^0}{\partial b^{(L)}} \tag{26}$$

# Summary: backward pass—outer layer

3. Backward pass, outer layer ($L$):
   1. Compute gradients:

$$\frac{\partial C}{\partial w^{(L)}} = \frac{\partial C}{\partial a^{(L)}} \frac{\partial a^{(L)}}{\partial z^{(L)}} \frac{\partial z^{(L)}}{\partial w^{(L)}} \tag{23}$$

$$\frac{\partial C}{\partial b^{(L)}} = \frac{\partial C}{\partial a^{(L)}} \frac{\partial a^{(L)}}{\partial z^{(L)}} \frac{\partial z^{(L)}}{\partial b^{(L)}} \tag{24}$$

   2. Update weights:

$$w^{(L),1} = w^{(L),0} - \eta \frac{\partial C^0}{\partial w^{(L)}} \tag{25}$$

$$b^{(L),1} = b^{(L),0} - \eta \frac{\partial C^0}{\partial b^{(L)}} \tag{26}$$

# Summary: backward pass—outer layer

3. Backward pass, outer layer ($L$):
   1. Compute gradients:

$$\frac{\partial C}{\partial w^{(L)}} = \frac{\partial C}{\partial a^{(L)}} \frac{\partial a^{(L)}}{\partial z^{(L)}} \frac{\partial z^{(L)}}{\partial w^{(L)}} \tag{23}$$

$$\frac{\partial C}{\partial b^{(L)}} = \frac{\partial C}{\partial a^{(L)}} \frac{\partial a^{(L)}}{\partial z^{(L)}} \frac{\partial z^{(L)}}{\partial b^{(L)}} \tag{24}$$

   2. Update weights:

$$w^{(L),1} = w^{(L),0} - \eta \frac{\partial C^0}{\partial w^{(L)}} \tag{25}$$

$$b^{(L),1} = b^{(L),0} - \eta \frac{\partial C^0}{\partial b^{(L)}} \tag{26}$$

# Summary: backward pass—outer layer

3. Backward pass, outer layer ($L$):
   1. Compute gradients:

$$\frac{\partial C}{\partial w^{(L)}} = \frac{\partial C}{\partial a^{(L)}} \frac{\partial a^{(L)}}{\partial z^{(L)}} \frac{\partial z^{(L)}}{\partial w^{(L)}} \tag{23}$$

$$\frac{\partial C}{\partial b^{(L)}} = \frac{\partial C}{\partial a^{(L)}} \frac{\partial a^{(L)}}{\partial z^{(L)}} \frac{\partial z^{(L)}}{\partial b^{(L)}} \tag{24}$$

   2. Update weights:

$$w^{(L),1} = w^{(L),0} - \eta \frac{\partial C^0}{\partial w^{(L)}} \tag{25}$$

$$b^{(L),1} = b^{(L),0} - \eta \frac{\partial C^0}{\partial b^{(L)}} \tag{26}$$

# Summary: backward pass—outer layer

3. Backward pass, outer layer ($L$):
    1. Compute gradients:

$$\frac{\partial C}{\partial w^{(L)}} = \frac{\partial C}{\partial a^{(L)}} \frac{\partial a^{(L)}}{\partial z^{(L)}} \frac{\partial z^{(L)}}{\partial w^{(L)}} \tag{23}$$

$$\frac{\partial C}{\partial b^{(L)}} = \frac{\partial C}{\partial a^{(L)}} \frac{\partial a^{(L)}}{\partial z^{(L)}} \frac{\partial z^{(L)}}{\partial b^{(L)}} \tag{24}$$

    2. Update weights:

$$w^{(L),1} = w^{(L),0} - \eta \frac{\partial C^0}{\partial w^{(L)}} \tag{25}$$

$$b^{(L),1} = b^{(L),0} - \eta \frac{\partial C^0}{\partial b^{(L)}} \tag{26}$$

# Summary: backward pass—outer layer

3. Backward pass, outer layer ($L$):
    1. Compute gradients:

$$\frac{\partial C}{\partial w^{(L)}} = \frac{\partial C}{\partial a^{(L)}} \frac{\partial a^{(L)}}{\partial z^{(L)}} \frac{\partial z^{(L)}}{\partial w^{(L)}} \tag{23}$$

$$\frac{\partial C}{\partial b^{(L)}} = \frac{\partial C}{\partial a^{(L)}} \frac{\partial a^{(L)}}{\partial z^{(L)}} \frac{\partial z^{(L)}}{\partial b^{(L)}} \tag{24}$$

    2. Update weights:

$$w^{(L),1} = w^{(L),0} - \eta \frac{\partial C^0}{\partial w^{(L)}} \tag{25}$$

$$b^{(L),1} = b^{(L),0} - \eta \frac{\partial C^0}{\partial b^{(L)}} \tag{26}$$

# Summary: backward pass—outer layer

3. Backward pass, outer layer ($L$):
   1. Compute gradients:

$$\frac{\partial C}{\partial w^{(L)}} = \frac{\partial C}{\partial a^{(L)}} \frac{\partial a^{(L)}}{\partial z^{(L)}} \frac{\partial z^{(L)}}{\partial w^{(L)}} \tag{23}$$

$$\frac{\partial C}{\partial b^{(L)}} = \frac{\partial C}{\partial a^{(L)}} \frac{\partial a^{(L)}}{\partial z^{(L)}} \frac{\partial z^{(L)}}{\partial b^{(L)}} \tag{24}$$

   2. Update weights:

$$w^{(L),1} = w^{(L),0} - \eta \frac{\partial C^0}{\partial w^{(L)}} \tag{25}$$

$$b^{(L),1} = b^{(L),0} - \eta \frac{\partial C^0}{\partial b^{(L)}} \tag{26}$$

# Summary: backward pass—outer layer

3. Backward pass, outer layer ($L$):
   1. Compute gradients:

$$\frac{\partial C}{\partial w^{(L)}} = \frac{\partial C}{\partial a^{(L)}} \frac{\partial a^{(L)}}{\partial z^{(L)}} \frac{\partial z^{(L)}}{\partial w^{(L)}} \tag{23}$$

$$\frac{\partial C}{\partial b^{(L)}} = \frac{\partial C}{\partial a^{(L)}} \frac{\partial a^{(L)}}{\partial z^{(L)}} \frac{\partial z^{(L)}}{\partial b^{(L)}} \tag{24}$$

   2. Update weights:

$$w^{(L),1} = w^{(L),0} - \eta \frac{\partial C^0}{\partial w^{(L)}} \tag{25}$$

$$b^{(L),1} = b^{(L),0} - \eta \frac{\partial C^0}{\partial b^{(L)}} \tag{26}$$

# Summary: backward pass—outer layer

3. Backward pass, outer layer ($L$):
   1. Compute gradients:

$$\frac{\partial C}{\partial w^{(L)}} = \frac{\partial C}{\partial a^{(L)}} \frac{\partial a^{(L)}}{\partial z^{(L)}} \frac{\partial z^{(L)}}{\partial w^{(L)}} \tag{23}$$

$$\frac{\partial C}{\partial b^{(L)}} = \frac{\partial C}{\partial a^{(L)}} \frac{\partial a^{(L)}}{\partial z^{(L)}} \frac{\partial z^{(L)}}{\partial b^{(L)}} \tag{24}$$

   2. Update weights:

$$w^{(L),1} = w^{(L),0} - \eta \frac{\partial C^0}{\partial w^{(L)}} \tag{25}$$

$$b^{(L),1} = b^{(L),0} - \eta \frac{\partial C^0}{\partial b^{(L)}} \tag{26}$$

# Summary: backward pass—last hidden layer

3. Backward pass, layer $(L-1)$:
   3. Compute gradients:

$$\frac{\partial C}{\partial w^{(L-1)}} = \frac{\partial C}{\partial a^{(L)}} \frac{\partial a^{(L)}}{\partial z^{(L)}} \frac{\partial z^{(L)}}{\partial a^{(L-1)}} \frac{\partial a^{(L-1)}}{\partial z^{(L-1)}} \frac{\partial z^{(L-1)}}{\partial w^{(L-1)}} \tag{27}$$

$$\frac{\partial C}{\partial b^{(L-1)}} = \frac{\partial C}{\partial a^{(L)}} \frac{\partial a^{(L)}}{\partial z^{(L)}} \frac{\partial z^{(L)}}{\partial a^{(L-1)}} \frac{\partial a^{(L-1)}}{\partial z^{(L-1)}} \frac{\partial z^{(L-1)}}{\partial b^{(L-1)}} \tag{28}$$

   4. Update weights:

$$w^{(L-1),1} = w^{(L-1),0} - \eta \frac{\partial C}{\partial w^{(L-1)}} \tag{29}$$

$$b^{(L-1),1} = b^{(L-1),0} - \eta \frac{\partial C}{\partial b^{(L-1)}} \tag{30}$$

# Summary: backward pass—last hidden layer

3. Backward pass, layer $(L-1)$:
   3. Compute gradients:

$$\frac{\partial C}{\partial w^{(L-1)}} = \frac{\partial C}{\partial a^{(L)}} \frac{\partial a^{(L)}}{\partial z^{(L)}} \frac{\partial z^{(L)}}{\partial a^{(L-1)}} \frac{\partial a^{(L-1)}}{\partial z^{(L-1)}} \frac{\partial z^{(L-1)}}{\partial w^{(L-1)}} \tag{27}$$

$$\frac{\partial C}{\partial b^{(L-1)}} = \frac{\partial C}{\partial a^{(L)}} \frac{\partial a^{(L)}}{\partial z^{(L)}} \frac{\partial z^{(L)}}{\partial a^{(L-1)}} \frac{\partial a^{(L-1)}}{\partial z^{(L-1)}} \frac{\partial z^{(L-1)}}{\partial b^{(L-1)}} \tag{28}$$

   4. Update weights:

$$w^{(L-1),1} = w^{(L-1),0} - \eta \frac{\partial C}{\partial w^{(L-1)}} \tag{29}$$

$$b^{(L-1),1} = b^{(L-1),0} - \eta \frac{\partial C}{\partial b^{(L-1)}} \tag{30}$$

# Summary: backward pass—last hidden layer

3. Backward pass, layer $(L-1)$:
    3. Compute gradients:

$$\frac{\partial C}{\partial w^{(L-1)}} = \frac{\partial C}{\partial a^{(L)}} \frac{\partial a^{(L)}}{\partial z^{(L)}} \frac{\partial z^{(L)}}{\partial a^{(L-1)}} \frac{\partial a^{(L-1)}}{\partial z^{(L-1)}} \frac{\partial z^{(L-1)}}{\partial w^{(L-1)}} \tag{27}$$

$$\frac{\partial C}{\partial b^{(L-1)}} = \frac{\partial C}{\partial a^{(L)}} \frac{\partial a^{(L)}}{\partial z^{(L)}} \frac{\partial z^{(L)}}{\partial a^{(L-1)}} \frac{\partial a^{(L-1)}}{\partial z^{(L-1)}} \frac{\partial z^{(L-1)}}{\partial b^{(L-1)}} \tag{28}$$

    4. Update weights:

$$w^{(L-1),1} = w^{(L-1),0} - \eta \frac{\partial C}{\partial w^{(L-1)}} \tag{29}$$

$$b^{(L-1),1} = b^{(L-1),0} - \eta \frac{\partial C}{\partial b^{(L-1)}} \tag{30}$$

# Summary: backward pass—last hidden layer

3. Backward pass, layer $(L-1)$:
   3. Compute gradients:

$$\frac{\partial C}{\partial w^{(L-1)}} = \frac{\partial C}{\partial a^{(L)}} \frac{\partial a^{(L)}}{\partial z^{(L)}} \frac{\partial z^{(L)}}{\partial a^{(L-1)}} \frac{\partial a^{(L-1)}}{\partial z^{(L-1)}} \frac{\partial z^{(L-1)}}{\partial w^{(L-1)}} \tag{27}$$

$$\frac{\partial C}{\partial b^{(L-1)}} = \frac{\partial C}{\partial a^{(L)}} \frac{\partial a^{(L)}}{\partial z^{(L)}} \frac{\partial z^{(L)}}{\partial a^{(L-1)}} \frac{\partial a^{(L-1)}}{\partial z^{(L-1)}} \frac{\partial z^{(L-1)}}{\partial b^{(L-1)}} \tag{28}$$

   4. Update weights:

$$w^{(L-1),1} = w^{(L-1),0} - \eta \frac{\partial C}{\partial w^{(L-1)}} \tag{29}$$

$$b^{(L-1),1} = b^{(L-1),0} - \eta \frac{\partial C}{\partial b^{(L-1)}} \tag{30}$$

# Summary: backward pass—last hidden layer

3. Backward pass, layer $(L-1)$:
   3. Compute gradients:

$$\frac{\partial C}{\partial w^{(L-1)}} = \frac{\partial C}{\partial a^{(L)}} \frac{\partial a^{(L)}}{\partial z^{(L)}} \frac{\partial z^{(L)}}{\partial a^{(L-1)}} \frac{\partial a^{(L-1)}}{\partial z^{(L-1)}} \frac{\partial z^{(L-1)}}{\partial w^{(L-1)}} \tag{27}$$

$$\frac{\partial C}{\partial b^{(L-1)}} = \frac{\partial C}{\partial a^{(L)}} \frac{\partial a^{(L)}}{\partial z^{(L)}} \frac{\partial z^{(L)}}{\partial a^{(L-1)}} \frac{\partial a^{(L-1)}}{\partial z^{(L-1)}} \frac{\partial z^{(L-1)}}{\partial b^{(L-1)}} \tag{28}$$

   4. Update weights:

$$w^{(L-1),1} = w^{(L-1),0} - \eta \frac{\partial C}{\partial w^{(L-1)}} \tag{29}$$

$$b^{(L-1),1} = b^{(L-1),0} - \eta \frac{\partial C}{\partial b^{(L-1)}} \tag{30}$$

# Summary: backward pass—last hidden layer

3. Backward pass, layer $(L-1)$:
   3. Compute gradients:

$$\frac{\partial C}{\partial w^{(L-1)}} = \frac{\partial C}{\partial a^{(L)}} \frac{\partial a^{(L)}}{\partial z^{(L)}} \frac{\partial z^{(L)}}{\partial a^{(L-1)}} \frac{\partial a^{(L-1)}}{\partial z^{(L-1)}} \frac{\partial z^{(L-1)}}{\partial w^{(L-1)}} \tag{27}$$

$$\frac{\partial C}{\partial b^{(L-1)}} = \frac{\partial C}{\partial a^{(L)}} \frac{\partial a^{(L)}}{\partial z^{(L)}} \frac{\partial z^{(L)}}{\partial a^{(L-1)}} \frac{\partial a^{(L-1)}}{\partial z^{(L-1)}} \frac{\partial z^{(L-1)}}{\partial b^{(L-1)}} \tag{28}$$

   4. Update weights:

$$w^{(L-1),1} = w^{(L-1),0} - \eta \frac{\partial C}{\partial w^{(L-1)}} \tag{29}$$

$$b^{(L-1),1} = b^{(L-1),0} - \eta \frac{\partial C}{\partial b^{(L-1)}} \tag{30}$$

# Summary: backward pass—last hidden layer

3. Backward pass, layer $(L-1)$:
   3. Compute gradients:

$$\frac{\partial C}{\partial w^{(L-1)}} = \frac{\partial C}{\partial a^{(L)}} \frac{\partial a^{(L)}}{\partial z^{(L)}} \frac{\partial z^{(L)}}{\partial a^{(L-1)}} \frac{\partial a^{(L-1)}}{\partial z^{(L-1)}} \frac{\partial z^{(L-1)}}{\partial w^{(L-1)}} \tag{27}$$

$$\frac{\partial C}{\partial b^{(L-1)}} = \frac{\partial C}{\partial a^{(L)}} \frac{\partial a^{(L)}}{\partial z^{(L)}} \frac{\partial z^{(L)}}{\partial a^{(L-1)}} \frac{\partial a^{(L-1)}}{\partial z^{(L-1)}} \frac{\partial z^{(L-1)}}{\partial b^{(L-1)}} \tag{28}$$

   4. Update weights:

$$w^{(L-1),1} = w^{(L-1),0} - \eta \frac{\partial C}{\partial w^{(L-1)}} \tag{29}$$

$$b^{(L-1),1} = b^{(L-1),0} - \eta \frac{\partial C}{\partial b^{(L-1)}} \tag{30}$$

# Summary: backward pass—last hidden layer

3. Backward pass, layer $(L-1)$:
   3. Compute gradients:

$$\frac{\partial C}{\partial w^{(L-1)}} = \frac{\partial C}{\partial a^{(L)}} \frac{\partial a^{(L)}}{\partial z^{(L)}} \frac{\partial z^{(L)}}{\partial a^{(L-1)}} \frac{\partial a^{(L-1)}}{\partial z^{(L-1)}} \frac{\partial z^{(L-1)}}{\partial w^{(L-1)}} \tag{27}$$

$$\frac{\partial C}{\partial b^{(L-1)}} = \frac{\partial C}{\partial a^{(L)}} \frac{\partial a^{(L)}}{\partial z^{(L)}} \frac{\partial z^{(L)}}{\partial a^{(L-1)}} \frac{\partial a^{(L-1)}}{\partial z^{(L-1)}} \frac{\partial z^{(L-1)}}{\partial b^{(L-1)}} \tag{28}$$

   4. Update weights:

$$w^{(L-1),1} = w^{(L-1),0} - \eta \frac{\partial C}{\partial w^{(L-1)}} \tag{29}$$

$$b^{(L-1),1} = b^{(L-1),0} - \eta \frac{\partial C}{\partial b^{(L-1)}} \tag{30}$$

# Summary: backward pass—last hidden layer

3. Backward pass, layer $(L-1)$:
   3. Compute gradients:

$$\frac{\partial C}{\partial w^{(L-1)}} = \frac{\partial C}{\partial a^{(L)}} \frac{\partial a^{(L)}}{\partial z^{(L)}} \frac{\partial z^{(L)}}{\partial a^{(L-1)}} \frac{\partial a^{(L-1)}}{\partial z^{(L-1)}} \frac{\partial z^{(L-1)}}{\partial w^{(L-1)}} \tag{27}$$

$$\frac{\partial C}{\partial b^{(L-1)}} = \frac{\partial C}{\partial a^{(L)}} \frac{\partial a^{(L)}}{\partial z^{(L)}} \frac{\partial z^{(L)}}{\partial a^{(L-1)}} \frac{\partial a^{(L-1)}}{\partial z^{(L-1)}} \frac{\partial z^{(L-1)}}{\partial b^{(L-1)}} \tag{28}$$

   4. Update weights:

$$w^{(L-1),1} = w^{(L-1),0} - \eta \frac{\partial C}{\partial w^{(L-1)}} \tag{29}$$

$$b^{(L-1),1} = b^{(L-1),0} - \eta \frac{\partial C}{\partial b^{(L-1)}} \tag{30}$$

# Summary: backward pass—last hidden layer

3. Backward pass, layer $(L-1)$:
   3. Compute gradients:

$$\frac{\partial C}{\partial w^{(L-1)}} = \frac{\partial C}{\partial a^{(L)}} \frac{\partial a^{(L)}}{\partial z^{(L)}} \frac{\partial z^{(L)}}{\partial a^{(L-1)}} \frac{\partial a^{(L-1)}}{\partial z^{(L-1)}} \frac{\partial z^{(L-1)}}{\partial w^{(L-1)}} \tag{27}$$

$$\frac{\partial C}{\partial b^{(L-1)}} = \frac{\partial C}{\partial a^{(L)}} \frac{\partial a^{(L)}}{\partial z^{(L)}} \frac{\partial z^{(L)}}{\partial a^{(L-1)}} \frac{\partial a^{(L-1)}}{\partial z^{(L-1)}} \frac{\partial z^{(L-1)}}{\partial b^{(L-1)}} \tag{28}$$

   4. Update weights:

$$w^{(L-1),1} = w^{(L-1),0} - \eta \frac{\partial C}{\partial w^{(L-1)}} \tag{29}$$

$$b^{(L-1),1} = b^{(L-1),0} - \eta \frac{\partial C}{\partial b^{(L-1)}} \tag{30}$$

# Summary: backward pass—last hidden layer

3. Backward pass, layer $(L-1)$:
   3. Compute gradients:

$$\frac{\partial C}{\partial w^{(L-1)}} = \frac{\partial C}{\partial a^{(L)}} \frac{\partial a^{(L)}}{\partial z^{(L)}} \frac{\partial z^{(L)}}{\partial a^{(L-1)}} \frac{\partial a^{(L-1)}}{\partial z^{(L-1)}} \frac{\partial z^{(L-1)}}{\partial w^{(L-1)}} \tag{27}$$

$$\frac{\partial C}{\partial b^{(L-1)}} = \frac{\partial C}{\partial a^{(L)}} \frac{\partial a^{(L)}}{\partial z^{(L)}} \frac{\partial z^{(L)}}{\partial a^{(L-1)}} \frac{\partial a^{(L-1)}}{\partial z^{(L-1)}} \frac{\partial z^{(L-1)}}{\partial b^{(L-1)}} \tag{28}$$

   4. Update weights:

$$w^{(L-1),1} = w^{(L-1),0} - \eta \frac{\partial C}{\partial w^{(L-1)}} \tag{29}$$

$$b^{(L-1),1} = b^{(L-1),0} - \eta \frac{\partial C}{\partial b^{(L-1)}} \tag{30}$$

# Summary: backward pass—last hidden layer

3. Backward pass, layer $(L-1)$:
   3. Compute gradients:

$$\frac{\partial C}{\partial w^{(L-1)}} = \frac{\partial C}{\partial a^{(L)}} \frac{\partial a^{(L)}}{\partial z^{(L)}} \frac{\partial z^{(L)}}{\partial a^{(L-1)}} \frac{\partial a^{(L-1)}}{\partial z^{(L-1)}} \frac{\partial z^{(L-1)}}{\partial w^{(L-1)}} \tag{27}$$

$$\frac{\partial C}{\partial b^{(L-1)}} = \frac{\partial C}{\partial a^{(L)}} \frac{\partial a^{(L)}}{\partial z^{(L)}} \frac{\partial z^{(L)}}{\partial a^{(L-1)}} \frac{\partial a^{(L-1)}}{\partial z^{(L-1)}} \frac{\partial z^{(L-1)}}{\partial b^{(L-1)}} \tag{28}$$

   4. Update weights:

$$w^{(L-1),1} = w^{(L-1),0} - \eta \frac{\partial C}{\partial w^{(L-1)}} \tag{29}$$

$$b^{(L-1),1} = b^{(L-1),0} - \eta \frac{\partial C}{\partial b^{(L-1)}} \tag{30}$$

# Summary: backward pass—last hidden layer

3. Backward pass, layer $(L-1)$:
   3. Compute gradients:

$$\frac{\partial C}{\partial w^{(L-1)}} = \frac{\partial C}{\partial a^{(L)}} \frac{\partial a^{(L)}}{\partial z^{(L)}} \frac{\partial z^{(L)}}{\partial a^{(L-1)}} \frac{\partial a^{(L-1)}}{\partial z^{(L-1)}} \frac{\partial z^{(L-1)}}{\partial w^{(L-1)}} \tag{27}$$

$$\frac{\partial C}{\partial b^{(L-1)}} = \frac{\partial C}{\partial a^{(L)}} \frac{\partial a^{(L)}}{\partial z^{(L)}} \frac{\partial z^{(L)}}{\partial a^{(L-1)}} \frac{\partial a^{(L-1)}}{\partial z^{(L-1)}} \frac{\partial z^{(L-1)}}{\partial b^{(L-1)}} \tag{28}$$

   4. Update weights:

$$w^{(L-1),1} = w^{(L-1),0} - \eta \frac{\partial C}{\partial w^{(L-1)}} \tag{29}$$

$$b^{(L-1),1} = b^{(L-1),0} - \eta \frac{\partial C}{\partial b^{(L-1)}} \tag{30}$$

# Summary: backward pass—last hidden layer

3. Backward pass, layer $(L-1)$:
   3. Compute gradients:

$$\frac{\partial C}{\partial w^{(L-1)}} = \frac{\partial C}{\partial a^{(L)}} \frac{\partial a^{(L)}}{\partial z^{(L)}} \frac{\partial z^{(L)}}{\partial a^{(L-1)}} \frac{\partial a^{(L-1)}}{\partial z^{(L-1)}} \frac{\partial z^{(L-1)}}{\partial w^{(L-1)}} \tag{27}$$

$$\frac{\partial C}{\partial b^{(L-1)}} = \frac{\partial C}{\partial a^{(L)}} \frac{\partial a^{(L)}}{\partial z^{(L)}} \frac{\partial z^{(L)}}{\partial a^{(L-1)}} \frac{\partial a^{(L-1)}}{\partial z^{(L-1)}} \frac{\partial z^{(L-1)}}{\partial b^{(L-1)}} \tag{28}$$

   4. Update weights:

$$w^{(L-1),1} = w^{(L-1),0} - \eta \frac{\partial C}{\partial w^{(L-1)}} \tag{29}$$

$$b^{(L-1),1} = b^{(L-1),0} - \eta \frac{\partial C}{\partial b^{(L-1)}} \tag{30}$$

# Summary: backward pass—second-to-last hidden layer

3. Backward pass, layer $(L-2)$:

   5. Compute gradients:

$$\frac{\partial C}{\partial w^{(L-2)}} = \frac{\partial C}{\partial a^{(L)}} \frac{\partial a^{(L)}}{\partial z^{(L)}} \frac{\partial z^{(L)}}{\partial a^{(L-1)}} \frac{\partial a^{(L-1)}}{\partial z^{(L-1)}} \frac{\partial z^{(L-1)}}{\partial a^{(L-2)}} \frac{\partial a^{(L-2)}}{\partial z^{(L-2)}} \frac{\partial z^{(L-2)}}{\partial w^{(L-2)}} \tag{31}$$

$$\frac{\partial C}{\partial b^{(L-2)}} = \frac{\partial C}{\partial a^{(L)}} \frac{\partial a^{(L)}}{\partial z^{(L)}} \frac{\partial z^{(L)}}{\partial a^{(L-1)}} \frac{\partial a^{(L-1)}}{\partial z^{(L-1)}} \frac{\partial z^{(L-1)}}{\partial a^{(L-2)}} \frac{\partial a^{(L-2)}}{\partial z^{(L-2)}} \frac{\partial z^{(L-2)}}{\partial b^{(L-2)}} \tag{32}$$

$$\tag{33}$$

   6. Update weights:

$$w^{(L-2),1} = w^{(L-2),0} - \eta \frac{\partial C}{\partial w^{(L-2)}} \tag{34}$$

$$b^{(L-2),1} = b^{(L-2),0} - \eta \frac{\partial C}{\partial b^{(L-2)}} \tag{35}$$

# Summary: backward pass—second-to-last hidden layer

3. Backward pass, layer $(L-2)$:

   5. Compute gradients:

$$\frac{\partial C}{\partial w^{(L-2)}} = \frac{\partial C}{\partial a^{(L)}} \frac{\partial a^{(L)}}{\partial z^{(L)}} \frac{\partial z^{(L)}}{\partial a^{(L-1)}} \frac{\partial a^{(L-1)}}{\partial z^{(L-1)}} \frac{\partial z^{(L-1)}}{\partial a^{(L-2)}} \frac{\partial a^{(L-2)}}{\partial z^{(L-2)}} \frac{\partial z^{(L-2)}}{\partial w^{(L-2)}} \tag{31}$$

$$\frac{\partial C}{\partial b^{(L-2)}} = \frac{\partial C}{\partial a^{(L)}} \frac{\partial a^{(L)}}{\partial z^{(L)}} \frac{\partial z^{(L)}}{\partial a^{(L-1)}} \frac{\partial a^{(L-1)}}{\partial z^{(L-1)}} \frac{\partial z^{(L-1)}}{\partial a^{(L-2)}} \frac{\partial a^{(L-2)}}{\partial z^{(L-2)}} \frac{\partial z^{(L-2)}}{\partial b^{(L-2)}} \tag{32}$$

(33)

   6. Update weights:

$$w^{(L-2),1} = w^{(L-2),0} - \eta \frac{\partial C}{\partial w^{(L-2)}} \tag{34}$$

$$b^{(L-2),1} = b^{(L-2),0} - \eta \frac{\partial C}{\partial b^{(L-2)}} \tag{35}$$

# Summary: backward pass—second-to-last hidden layer

3. Backward pass, layer $(L-2)$:
   5. Compute gradients:

$$\frac{\partial C}{\partial w^{(L-2)}} = \frac{\partial C}{\partial a^{(L)}} \frac{\partial a^{(L)}}{\partial z^{(L)}} \frac{\partial z^{(L)}}{\partial a^{(L-1)}} \frac{\partial a^{(L-1)}}{\partial z^{(L-1)}} \frac{\partial z^{(L-1)}}{\partial a^{(L-2)}} \frac{\partial a^{(L-2)}}{\partial z^{(L-2)}} \frac{\partial z^{(L-2)}}{\partial w^{(L-2)}} \tag{31}$$

$$\frac{\partial C}{\partial b^{(L-2)}} = \frac{\partial C}{\partial a^{(L)}} \frac{\partial a^{(L)}}{\partial z^{(L)}} \frac{\partial z^{(L)}}{\partial a^{(L-1)}} \frac{\partial a^{(L-1)}}{\partial z^{(L-1)}} \frac{\partial z^{(L-1)}}{\partial a^{(L-2)}} \frac{\partial a^{(L-2)}}{\partial z^{(L-2)}} \frac{\partial z^{(L-2)}}{\partial b^{(L-2)}} \tag{32}$$

$$\tag{33}$$

   6. Update weights:

$$w^{(L-2),1} = w^{(L-2),0} - \eta \frac{\partial C}{\partial w^{(L-2)}} \tag{34}$$

$$b^{(L-2),1} = b^{(L-2),0} - \eta \frac{\partial C}{\partial b^{(L-2)}} \tag{35}$$

# Summary: backward pass—second-to-last hidden layer

3. Backward pass, layer $(L-2)$:
   5. Compute gradients:

$$\frac{\partial C}{\partial w^{(L-2)}} = \frac{\partial C}{\partial a^{(L)}} \frac{\partial a^{(L)}}{\partial z^{(L)}} \frac{\partial z^{(L)}}{\partial a^{(L-1)}} \frac{\partial a^{(L-1)}}{\partial z^{(L-1)}} \frac{\partial z^{(L-1)}}{\partial a^{(L-2)}} \frac{\partial a^{(L-2)}}{\partial z^{(L-2)}} \frac{\partial z^{(L-2)}}{\partial w^{(L-2)}} \tag{31}$$

$$\frac{\partial C}{\partial b^{(L-2)}} = \frac{\partial C}{\partial a^{(L)}} \frac{\partial a^{(L)}}{\partial z^{(L)}} \frac{\partial z^{(L)}}{\partial a^{(L-1)}} \frac{\partial a^{(L-1)}}{\partial z^{(L-1)}} \frac{\partial z^{(L-1)}}{\partial a^{(L-2)}} \frac{\partial a^{(L-2)}}{\partial z^{(L-2)}} \frac{\partial z^{(L-2)}}{\partial b^{(L-2)}} \tag{32}$$

(33)

   6. Update weights:

$$w^{(L-2),1} = w^{(L-2),0} - \eta \frac{\partial C}{\partial w^{(L-2)}} \tag{34}$$

$$b^{(L-2),1} = b^{(L-2),0} - \eta \frac{\partial C}{\partial b^{(L-2)}} \tag{35}$$

# Summary: backward pass—second-to-last hidden layer

3. Backward pass, layer $(L-2)$:
   5. Compute gradients:

$$\frac{\partial C}{\partial w^{(L-2)}} = \frac{\partial C}{\partial a^{(L)}} \frac{\partial a^{(L)}}{\partial z^{(L)}} \frac{\partial z^{(L)}}{\partial a^{(L-1)}} \frac{\partial a^{(L-1)}}{\partial z^{(L-1)}} \frac{\partial z^{(L-1)}}{\partial a^{(L-2)}} \frac{\partial a^{(L-2)}}{\partial z^{(L-2)}} \frac{\partial z^{(L-2)}}{\partial w^{(L-2)}} \tag{31}$$

$$\frac{\partial C}{\partial b^{(L-2)}} = \frac{\partial C}{\partial a^{(L)}} \frac{\partial a^{(L)}}{\partial z^{(L)}} \frac{\partial z^{(L)}}{\partial a^{(L-1)}} \frac{\partial a^{(L-1)}}{\partial z^{(L-1)}} \frac{\partial z^{(L-1)}}{\partial a^{(L-2)}} \frac{\partial a^{(L-2)}}{\partial z^{(L-2)}} \frac{\partial z^{(L-2)}}{\partial b^{(L-2)}} \tag{32}$$

(33)

   6. Update weights:

$$w^{(L-2),1} = w^{(L-2),0} - \eta \frac{\partial C}{\partial w^{(L-2)}} \tag{34}$$

$$b^{(L-2),1} = b^{(L-2),0} - \eta \frac{\partial C}{\partial b^{(L-2)}} \tag{35}$$

# Summary: backward pass—second-to-last hidden layer

3. Backward pass, layer $(L-2)$:
   5. Compute gradients:

$$\frac{\partial C}{\partial w^{(L-2)}} = \frac{\partial C}{\partial a^{(L)}} \frac{\partial a^{(L)}}{\partial z^{(L)}} \frac{\partial z^{(L)}}{\partial a^{(L-1)}} \frac{\partial a^{(L-1)}}{\partial z^{(L-1)}} \frac{\partial z^{(L-1)}}{\partial a^{(L-2)}} \frac{\partial a^{(L-2)}}{\partial z^{(L-2)}} \frac{\partial z^{(L-2)}}{\partial w^{(L-2)}} \tag{31}$$

$$\frac{\partial C}{\partial b^{(L-2)}} = \frac{\partial C}{\partial a^{(L)}} \frac{\partial a^{(L)}}{\partial z^{(L)}} \frac{\partial z^{(L)}}{\partial a^{(L-1)}} \frac{\partial a^{(L-1)}}{\partial z^{(L-1)}} \frac{\partial z^{(L-1)}}{\partial a^{(L-2)}} \frac{\partial a^{(L-2)}}{\partial z^{(L-2)}} \frac{\partial z^{(L-2)}}{\partial b^{(L-2)}} \tag{32}$$

(33)

   6. Update weights:

$$w^{(L-2),1} = w^{(L-2),0} - \eta \frac{\partial C}{\partial w^{(L-2)}} \tag{34}$$

$$b^{(L-2),1} = b^{(L-2),0} - \eta \frac{\partial C}{\partial b^{(L-2)}} \tag{35}$$

# Summary: backward pass—second-to-last hidden layer

3. Backward pass, layer $(L-2)$:
   5. Compute gradients:

$$\frac{\partial C}{\partial w^{(L-2)}} = \frac{\partial C}{\partial a^{(L)}} \frac{\partial a^{(L)}}{\partial z^{(L)}} \frac{\partial z^{(L)}}{\partial a^{(L-1)}} \frac{\partial a^{(L-1)}}{\partial z^{(L-1)}} \frac{\partial z^{(L-1)}}{\partial a^{(L-2)}} \frac{\partial a^{(L-2)}}{\partial z^{(L-2)}} \frac{\partial z^{(L-2)}}{\partial w^{(L-2)}} \tag{31}$$

$$\frac{\partial C}{\partial b^{(L-2)}} = \frac{\partial C}{\partial a^{(L)}} \frac{\partial a^{(L)}}{\partial z^{(L)}} \frac{\partial z^{(L)}}{\partial a^{(L-1)}} \frac{\partial a^{(L-1)}}{\partial z^{(L-1)}} \frac{\partial z^{(L-1)}}{\partial a^{(L-2)}} \frac{\partial a^{(L-2)}}{\partial z^{(L-2)}} \frac{\partial z^{(L-2)}}{\partial b^{(L-2)}} \tag{32}$$

(33)

   6. Update weights:

$$w^{(L-2),1} = w^{(L-2),0} - \eta \frac{\partial C}{\partial w^{(L-2)}} \tag{34}$$

$$b^{(L-2),1} = b^{(L-2),0} - \eta \frac{\partial C}{\partial b^{(L-2)}} \tag{35}$$

# Summary: backward pass—second-to-last hidden layer

3. Backward pass, layer $(L-2)$:
   5. Compute gradients:

$$\frac{\partial C}{\partial w^{(L-2)}} = \frac{\partial C}{\partial a^{(L)}} \frac{\partial a^{(L)}}{\partial z^{(L)}} \frac{\partial z^{(L)}}{\partial a^{(L-1)}} \frac{\partial a^{(L-1)}}{\partial z^{(L-1)}} \frac{\partial z^{(L-1)}}{\partial a^{(L-2)}} \frac{\partial a^{(L-2)}}{\partial z^{(L-2)}} \frac{\partial z^{(L-2)}}{\partial w^{(L-2)}} \tag{31}$$

$$\frac{\partial C}{\partial b^{(L-2)}} = \frac{\partial C}{\partial a^{(L)}} \frac{\partial a^{(L)}}{\partial z^{(L)}} \frac{\partial z^{(L)}}{\partial a^{(L-1)}} \frac{\partial a^{(L-1)}}{\partial z^{(L-1)}} \frac{\partial z^{(L-1)}}{\partial a^{(L-2)}} \frac{\partial a^{(L-2)}}{\partial z^{(L-2)}} \frac{\partial z^{(L-2)}}{\partial b^{(L-2)}} \tag{32}$$

(33)

   6. Update weights:

$$w^{(L-2),1} = w^{(L-2),0} - \eta \frac{\partial C}{\partial w^{(L-2)}} \tag{34}$$

$$b^{(L-2),1} = b^{(L-2),0} - \eta \frac{\partial C}{\partial b^{(L-2)}} \tag{35}$$

# Summary: backward pass—second-to-last hidden layer

3. Backward pass, layer $(L-2)$:
   5. Compute gradients:

$$\frac{\partial C}{\partial w^{(L-2)}} = \frac{\partial C}{\partial a^{(L)}} \frac{\partial a^{(L)}}{\partial z^{(L)}} \frac{\partial z^{(L)}}{\partial a^{(L-1)}} \frac{\partial a^{(L-1)}}{\partial z^{(L-1)}} \frac{\partial z^{(L-1)}}{\partial a^{(L-2)}} \frac{\partial a^{(L-2)}}{\partial z^{(L-2)}} \frac{\partial z^{(L-2)}}{\partial w^{(L-2)}} \tag{31}$$

$$\frac{\partial C}{\partial b^{(L-2)}} = \frac{\partial C}{\partial a^{(L)}} \frac{\partial a^{(L)}}{\partial z^{(L)}} \frac{\partial z^{(L)}}{\partial a^{(L-1)}} \frac{\partial a^{(L-1)}}{\partial z^{(L-1)}} \frac{\partial z^{(L-1)}}{\partial a^{(L-2)}} \frac{\partial a^{(L-2)}}{\partial z^{(L-2)}} \frac{\partial z^{(L-2)}}{\partial b^{(L-2)}} \tag{32}$$

$$ \tag{33}$$

   6. Update weights:

$$w^{(L-2),1} = w^{(L-2),0} - \eta \frac{\partial C}{\partial w^{(L-2)}} \tag{34}$$

$$b^{(L-2),1} = b^{(L-2),0} - \eta \frac{\partial C}{\partial b^{(L-2)}} \tag{35}$$

# Summary: backward pass—second-to-last hidden layer

3. Backward pass, layer $(L-2)$:
   5. Compute gradients:

$$\frac{\partial C}{\partial w^{(L-2)}} = \frac{\partial C}{\partial a^{(L)}} \frac{\partial a^{(L)}}{\partial z^{(L)}} \frac{\partial z^{(L)}}{\partial a^{(L-1)}} \frac{\partial a^{(L-1)}}{\partial z^{(L-1)}} \frac{\partial z^{(L-1)}}{\partial a^{(L-2)}} \frac{\partial a^{(L-2)}}{\partial z^{(L-2)}} \frac{\partial z^{(L-2)}}{\partial w^{(L-2)}} \tag{31}$$

$$\frac{\partial C}{\partial b^{(L-2)}} = \frac{\partial C}{\partial a^{(L)}} \frac{\partial a^{(L)}}{\partial z^{(L)}} \frac{\partial z^{(L)}}{\partial a^{(L-1)}} \frac{\partial a^{(L-1)}}{\partial z^{(L-1)}} \frac{\partial z^{(L-1)}}{\partial a^{(L-2)}} \frac{\partial a^{(L-2)}}{\partial z^{(L-2)}} \frac{\partial z^{(L-2)}}{\partial b^{(L-2)}} \tag{32}$$

(33)

   6. Update weights:

$$w^{(L-2),1} = w^{(L-2),0} - \eta \frac{\partial C}{\partial w^{(L-2)}} \tag{34}$$

$$b^{(L-2),1} = b^{(L-2),0} - \eta \frac{\partial C}{\partial b^{(L-2)}} \tag{35}$$

# Summary: backward pass—second-to-last hidden layer

3. Backward pass, layer $(L-2)$:
   5. Compute gradients:

$$\frac{\partial C}{\partial w^{(L-2)}} = \frac{\partial C}{\partial a^{(L)}} \frac{\partial a^{(L)}}{\partial z^{(L)}} \frac{\partial z^{(L)}}{\partial a^{(L-1)}} \frac{\partial a^{(L-1)}}{\partial z^{(L-1)}} \frac{\partial z^{(L-1)}}{\partial a^{(L-2)}} \frac{\partial a^{(L-2)}}{\partial z^{(L-2)}} \frac{\partial z^{(L-2)}}{\partial w^{(L-2)}} \tag{31}$$

$$\frac{\partial C}{\partial b^{(L-2)}} = \frac{\partial C}{\partial a^{(L)}} \frac{\partial a^{(L)}}{\partial z^{(L)}} \frac{\partial z^{(L)}}{\partial a^{(L-1)}} \frac{\partial a^{(L-1)}}{\partial z^{(L-1)}} \frac{\partial z^{(L-1)}}{\partial a^{(L-2)}} \frac{\partial a^{(L-2)}}{\partial z^{(L-2)}} \frac{\partial z^{(L-2)}}{\partial b^{(L-2)}} \tag{32}$$

$$\tag{33}$$

   6. Update weights:

$$w^{(L-2),1} = w^{(L-2),0} - \eta \frac{\partial C}{\partial w^{(L-2)}} \tag{34}$$

$$b^{(L-2),1} = b^{(L-2),0} - \eta \frac{\partial C}{\partial b^{(L-2)}} \tag{35}$$

# Summary: backward pass—second-to-last hidden layer

3. Backward pass, layer $(L-2)$:
    5. Compute gradients:

$$\frac{\partial C}{\partial w^{(L-2)}} = \frac{\partial C}{\partial a^{(L)}} \frac{\partial a^{(L)}}{\partial z^{(L)}} \frac{\partial z^{(L)}}{\partial a^{(L-1)}} \frac{\partial a^{(L-1)}}{\partial z^{(L-1)}} \frac{\partial z^{(L-1)}}{\partial a^{(L-2)}} \frac{\partial a^{(L-2)}}{\partial z^{(L-2)}} \frac{\partial z^{(L-2)}}{\partial w^{(L-2)}} \tag{31}$$

$$\frac{\partial C}{\partial b^{(L-2)}} = \frac{\partial C}{\partial a^{(L)}} \frac{\partial a^{(L)}}{\partial z^{(L)}} \frac{\partial z^{(L)}}{\partial a^{(L-1)}} \frac{\partial a^{(L-1)}}{\partial z^{(L-1)}} \frac{\partial z^{(L-1)}}{\partial a^{(L-2)}} \frac{\partial a^{(L-2)}}{\partial z^{(L-2)}} \frac{\partial z^{(L-2)}}{\partial b^{(L-2)}} \tag{32}$$

$$\tag{33}$$

    6. Update weights:

$$w^{(L-2),1} = w^{(L-2),0} - \eta \frac{\partial C}{\partial w^{(L-2)}} \tag{34}$$

$$b^{(L-2),1} = b^{(L-2),0} - \eta \frac{\partial C}{\partial b^{(L-2)}} \tag{35}$$

# Summary: backward pass—second-to-last hidden layer

3. Backward pass, layer $(L-2)$:
   5. Compute gradients:

$$\frac{\partial C}{\partial w^{(L-2)}} = \frac{\partial C}{\partial a^{(L)}} \frac{\partial a^{(L)}}{\partial z^{(L)}} \frac{\partial z^{(L)}}{\partial a^{(L-1)}} \frac{\partial a^{(L-1)}}{\partial z^{(L-1)}} \frac{\partial z^{(L-1)}}{\partial a^{(L-2)}} \frac{\partial a^{(L-2)}}{\partial z^{(L-2)}} \frac{\partial z^{(L-2)}}{\partial w^{(L-2)}} \tag{31}$$

$$\frac{\partial C}{\partial b^{(L-2)}} = \frac{\partial C}{\partial a^{(L)}} \frac{\partial a^{(L)}}{\partial z^{(L)}} \frac{\partial z^{(L)}}{\partial a^{(L-1)}} \frac{\partial a^{(L-1)}}{\partial z^{(L-1)}} \frac{\partial z^{(L-1)}}{\partial a^{(L-2)}} \frac{\partial a^{(L-2)}}{\partial z^{(L-2)}} \frac{\partial z^{(L-2)}}{\partial b^{(L-2)}} \tag{32}$$

(33)

   6. Update weights:

$$w^{(L-2),1} = w^{(L-2),0} - \eta \frac{\partial C}{\partial w^{(L-2)}} \tag{34}$$

$$b^{(L-2),1} = b^{(L-2),0} - \eta \frac{\partial C}{\partial b^{(L-2)}} \tag{35}$$

# Summary: backward pass—second-to-last hidden layer

3. Backward pass, layer $(L-2)$:
   5. Compute gradients:

$$\frac{\partial C}{\partial w^{(L-2)}} = \frac{\partial C}{\partial a^{(L)}} \frac{\partial a^{(L)}}{\partial z^{(L)}} \frac{\partial z^{(L)}}{\partial a^{(L-1)}} \frac{\partial a^{(L-1)}}{\partial z^{(L-1)}} \frac{\partial z^{(L-1)}}{\partial a^{(L-2)}} \frac{\partial a^{(L-2)}}{\partial z^{(L-2)}} \frac{\partial z^{(L-2)}}{\partial w^{(L-2)}} \tag{31}$$

$$\frac{\partial C}{\partial b^{(L-2)}} = \frac{\partial C}{\partial a^{(L)}} \frac{\partial a^{(L)}}{\partial z^{(L)}} \frac{\partial z^{(L)}}{\partial a^{(L-1)}} \frac{\partial a^{(L-1)}}{\partial z^{(L-1)}} \frac{\partial z^{(L-1)}}{\partial a^{(L-2)}} \frac{\partial a^{(L-2)}}{\partial z^{(L-2)}} \frac{\partial z^{(L-2)}}{\partial b^{(L-2)}} \tag{32}$$

$$\tag{33}$$

   6. Update weights:

$$w^{(L-2),1} = w^{(L-2),0} - \eta \frac{\partial C}{\partial w^{(L-2)}} \tag{34}$$

$$b^{(L-2),1} = b^{(L-2),0} - \eta \frac{\partial C}{\partial b^{(L-2)}} \tag{35}$$

# Summary: backward pass—second-to-last hidden layer

3. Backward pass, layer $(L-2)$:
   5. Compute gradients:

$$\frac{\partial C}{\partial w^{(L-2)}} = \frac{\partial C}{\partial a^{(L)}} \frac{\partial a^{(L)}}{\partial z^{(L)}} \frac{\partial z^{(L)}}{\partial a^{(L-1)}} \frac{\partial a^{(L-1)}}{\partial z^{(L-1)}} \frac{\partial z^{(L-1)}}{\partial a^{(L-2)}} \frac{\partial a^{(L-2)}}{\partial z^{(L-2)}} \frac{\partial z^{(L-2)}}{\partial w^{(L-2)}} \tag{31}$$

$$\frac{\partial C}{\partial b^{(L-2)}} = \frac{\partial C}{\partial a^{(L)}} \frac{\partial a^{(L)}}{\partial z^{(L)}} \frac{\partial z^{(L)}}{\partial a^{(L-1)}} \frac{\partial a^{(L-1)}}{\partial z^{(L-1)}} \frac{\partial z^{(L-1)}}{\partial a^{(L-2)}} \frac{\partial a^{(L-2)}}{\partial z^{(L-2)}} \frac{\partial z^{(L-2)}}{\partial b^{(L-2)}} \tag{32}$$

(33)

   6. Update weights:

$$w^{(L-2),1} = w^{(L-2),0} - \eta \frac{\partial C}{\partial w^{(L-2)}} \tag{34}$$

$$b^{(L-2),1} = b^{(L-2),0} - \eta \frac{\partial C}{\partial b^{(L-2)}} \tag{35}$$

# Summary: backward pass—second-to-last hidden layer

3. Backward pass, layer $(L-2)$:
   5. Compute gradients:

$$\frac{\partial C}{\partial w^{(L-2)}} = \frac{\partial C}{\partial a^{(L)}} \frac{\partial a^{(L)}}{\partial z^{(L)}} \frac{\partial z^{(L)}}{\partial a^{(L-1)}} \frac{\partial a^{(L-1)}}{\partial z^{(L-1)}} \frac{\partial z^{(L-1)}}{\partial a^{(L-2)}} \frac{\partial a^{(L-2)}}{\partial z^{(L-2)}} \frac{\partial z^{(L-2)}}{\partial w^{(L-2)}} \tag{31}$$

$$\frac{\partial C}{\partial b^{(L-2)}} = \frac{\partial C}{\partial a^{(L)}} \frac{\partial a^{(L)}}{\partial z^{(L)}} \frac{\partial z^{(L)}}{\partial a^{(L-1)}} \frac{\partial a^{(L-1)}}{\partial z^{(L-1)}} \frac{\partial z^{(L-1)}}{\partial a^{(L-2)}} \frac{\partial a^{(L-2)}}{\partial z^{(L-2)}} \frac{\partial z^{(L-2)}}{\partial b^{(L-2)}} \tag{32}$$

(33)

   6. Update weights:

$$w^{(L-2),1} = w^{(L-2),0} - \eta \frac{\partial C}{\partial w^{(L-2)}} \tag{34}$$

$$b^{(L-2),1} = b^{(L-2),0} - \eta \frac{\partial C}{\partial b^{(L-2)}} \tag{35}$$

# Summary: backward pass—second-to-last hidden layer

3. Backward pass, layer $(L-2)$:
   5. Compute gradients:

$$\frac{\partial C}{\partial w^{(L-2)}} = \frac{\partial C}{\partial a^{(L)}} \frac{\partial a^{(L)}}{\partial z^{(L)}} \frac{\partial z^{(L)}}{\partial a^{(L-1)}} \frac{\partial a^{(L-1)}}{\partial z^{(L-1)}} \frac{\partial z^{(L-1)}}{\partial a^{(L-2)}} \frac{\partial a^{(L-2)}}{\partial z^{(L-2)}} \frac{\partial z^{(L-2)}}{\partial w^{(L-2)}} \tag{31}$$

$$\frac{\partial C}{\partial b^{(L-2)}} = \frac{\partial C}{\partial a^{(L)}} \frac{\partial a^{(L)}}{\partial z^{(L)}} \frac{\partial z^{(L)}}{\partial a^{(L-1)}} \frac{\partial a^{(L-1)}}{\partial z^{(L-1)}} \frac{\partial z^{(L-1)}}{\partial a^{(L-2)}} \frac{\partial a^{(L-2)}}{\partial z^{(L-2)}} \frac{\partial z^{(L-2)}}{\partial b^{(L-2)}} \tag{32}$$

(33)

   6. Update weights:

$$w^{(L-2),1} = w^{(L-2),0} - \eta \frac{\partial C}{\partial w^{(L-2)}} \tag{34}$$

$$b^{(L-2),1} = b^{(L-2),0} - \eta \frac{\partial C}{\partial b^{(L-2)}} \tag{35}$$

# Summary: backward pass—second-to-last hidden layer

3. Backward pass, layer $(L-2)$:
   5. Compute gradients:

$$\frac{\partial C}{\partial w^{(L-2)}} = \frac{\partial C}{\partial a^{(L)}} \frac{\partial a^{(L)}}{\partial z^{(L)}} \frac{\partial z^{(L)}}{\partial a^{(L-1)}} \frac{\partial a^{(L-1)}}{\partial z^{(L-1)}} \frac{\partial z^{(L-1)}}{\partial a^{(L-2)}} \frac{\partial a^{(L-2)}}{\partial z^{(L-2)}} \frac{\partial z^{(L-2)}}{\partial w^{(L-2)}} \tag{31}$$

$$\frac{\partial C}{\partial b^{(L-2)}} = \frac{\partial C}{\partial a^{(L)}} \frac{\partial a^{(L)}}{\partial z^{(L)}} \frac{\partial z^{(L)}}{\partial a^{(L-1)}} \frac{\partial a^{(L-1)}}{\partial z^{(L-1)}} \frac{\partial z^{(L-1)}}{\partial a^{(L-2)}} \frac{\partial a^{(L-2)}}{\partial z^{(L-2)}} \frac{\partial z^{(L-2)}}{\partial b^{(L-2)}} \tag{32}$$

(33)

   6. Update weights:

$$w^{(L-2),1} = w^{(L-2),0} - \eta \frac{\partial C}{\partial w^{(L-2)}} \tag{34}$$

$$b^{(L-2),1} = b^{(L-2),0} - \eta \frac{\partial C}{\partial b^{(L-2)}} \tag{35}$$

# Summary: backward pass—second-to-last hidden layer

3. Backward pass, layer $(L-2)$:
   5. Compute gradients:

$$\frac{\partial C}{\partial w^{(L-2)}} = \frac{\partial C}{\partial a^{(L)}} \frac{\partial a^{(L)}}{\partial z^{(L)}} \frac{\partial z^{(L)}}{\partial a^{(L-1)}} \frac{\partial a^{(L-1)}}{\partial z^{(L-1)}} \frac{\partial z^{(L-1)}}{\partial a^{(L-2)}} \frac{\partial a^{(L-2)}}{\partial z^{(L-2)}} \frac{\partial z^{(L-2)}}{\partial w^{(L-2)}} \tag{31}$$

$$\frac{\partial C}{\partial b^{(L-2)}} = \frac{\partial C}{\partial a^{(L)}} \frac{\partial a^{(L)}}{\partial z^{(L)}} \frac{\partial z^{(L)}}{\partial a^{(L-1)}} \frac{\partial a^{(L-1)}}{\partial z^{(L-1)}} \frac{\partial z^{(L-1)}}{\partial a^{(L-2)}} \frac{\partial a^{(L-2)}}{\partial z^{(L-2)}} \frac{\partial z^{(L-2)}}{\partial b^{(L-2)}} \tag{32}$$

(33)

   6. Update weights:

$$w^{(L-2),1} = w^{(L-2),0} - \eta \frac{\partial C}{\partial w^{(L-2)}} \tag{34}$$

$$b^{(L-2),1} = b^{(L-2),0} - \eta \frac{\partial C}{\partial b^{(L-2)}} \tag{35}$$

# Summary: backward pass—first hidden layer

3. Backward pass, layer (1):

   5. Compute gradients:

$$\frac{\partial C}{\partial w^{(1)}} = \frac{\partial C}{\partial a^{(L)}} \frac{\partial a^{(L)}}{\partial z^{(L)}} \frac{\partial z^{(L)}}{\partial a^{(L-1)}} \frac{\partial a^{(L-1)}}{\partial z^{(L-1)}} \cdots \frac{\partial z^{(2)}}{\partial a^{(1)}} \frac{\partial a^{(1)}}{\partial z^{(1)}} \frac{\partial z^{(1)}}{\partial w^{(1)}} \tag{36}$$

$$\frac{\partial C}{\partial b^{(1)}} = \frac{\partial C}{\partial a^{(L)}} \frac{\partial a^{(L)}}{\partial z^{(L)}} \frac{\partial z^{(L)}}{\partial a^{(L-1)}} \frac{\partial a^{(L-1)}}{\partial z^{(L-1)}} \cdots \frac{\partial z^{(2)}}{\partial a^{(1)}} \frac{\partial a^{(1)}}{\partial z^{(1)}} \frac{\partial z^{(1)}}{\partial b^{(1)}} \tag{37}$$

(38)

   6. Update weights:

$$w^{(1),1} = w^{(1),0} - \eta \frac{\partial C}{\partial w^{(1)}} \tag{39}$$

$$b^{(1),1} = b^{(1),0} - \eta \frac{\partial C}{\partial b^{(1)}} \tag{40}$$

# Summary: backward pass—first hidden layer

3. Backward pass, layer (1):

   5. Compute gradients:

$$\frac{\partial C}{\partial w^{(1)}} = \frac{\partial C}{\partial a^{(L)}}\frac{\partial a^{(L)}}{\partial z^{(L)}}\frac{\partial z^{(L)}}{\partial a^{(L-1)}}\frac{\partial a^{(L-1)}}{\partial z^{(L-1)}}\cdots\frac{\partial z^{(2)}}{\partial a^{(1)}}\frac{\partial a^{(1)}}{\partial z^{(1)}}\frac{\partial z^{(1)}}{\partial w^{(1)}} \tag{36}$$

$$\frac{\partial C}{\partial b^{(1)}} = \frac{\partial C}{\partial a^{(L)}}\frac{\partial a^{(L)}}{\partial z^{(L)}}\frac{\partial z^{(L)}}{\partial a^{(L-1)}}\frac{\partial a^{(L-1)}}{\partial z^{(L-1)}}\cdots\frac{\partial z^{(2)}}{\partial a^{(1)}}\frac{\partial a^{(1)}}{\partial z^{(1)}}\frac{\partial z^{(1)}}{\partial b^{(1)}} \tag{37}$$

(38)

   6. Update weights:

$$w^{(1),1} = w^{(1),0} - \eta\frac{\partial C}{\partial w^{(1)}} \tag{39}$$

$$b^{(1),1} = b^{(1),0} - \eta\frac{\partial C}{\partial b^{(1)}} \tag{40}$$

# Summary: backward pass—first hidden layer

3. Backward pass, layer (1):
   5. Compute gradients:

$$\frac{\partial C}{\partial w^{(1)}} = \frac{\partial C}{\partial a^{(L)}} \frac{\partial a^{(L)}}{\partial z^{(L)}} \frac{\partial z^{(L)}}{\partial a^{(L-1)}} \frac{\partial a^{(L-1)}}{\partial z^{(L-1)}} \cdots \frac{\partial z^{(2)}}{\partial a^{(1)}} \frac{\partial a^{(1)}}{\partial z^{(1)}} \frac{\partial z^{(1)}}{\partial w^{(1)}} \tag{36}$$

$$\frac{\partial C}{\partial b^{(1)}} = \frac{\partial C}{\partial a^{(L)}} \frac{\partial a^{(L)}}{\partial z^{(L)}} \frac{\partial z^{(L)}}{\partial a^{(L-1)}} \frac{\partial a^{(L-1)}}{\partial z^{(L-1)}} \cdots \frac{\partial z^{(2)}}{\partial a^{(1)}} \frac{\partial a^{(1)}}{\partial z^{(1)}} \frac{\partial z^{(1)}}{\partial b^{(1)}} \tag{37}$$

(38)

   6. Update weights:

$$w^{(1),1} = w^{(1),0} - \eta \frac{\partial C}{\partial w^{(1)}} \tag{39}$$

$$b^{(1),1} = b^{(1),0} - \eta \frac{\partial C}{\partial b^{(1)}} \tag{40}$$

# Summary: backward pass—first hidden layer

3. Backward pass, layer (1):
   5. Compute gradients:

$$\frac{\partial C}{\partial w^{(1)}} = \frac{\partial C}{\partial a^{(L)}} \frac{\partial a^{(L)}}{\partial z^{(L)}} \frac{\partial z^{(L)}}{\partial a^{(L-1)}} \frac{\partial a^{(L-1)}}{\partial z^{(L-1)}} \cdots \frac{\partial z^{(2)}}{\partial a^{(1)}} \frac{\partial a^{(1)}}{\partial z^{(1)}} \frac{\partial z^{(1)}}{\partial w^{(1)}} \tag{36}$$

$$\frac{\partial C}{\partial b^{(1)}} = \frac{\partial C}{\partial a^{(L)}} \frac{\partial a^{(L)}}{\partial z^{(L)}} \frac{\partial z^{(L)}}{\partial a^{(L-1)}} \frac{\partial a^{(L-1)}}{\partial z^{(L-1)}} \cdots \frac{\partial z^{(2)}}{\partial a^{(1)}} \frac{\partial a^{(1)}}{\partial z^{(1)}} \frac{\partial z^{(1)}}{\partial b^{(1)}} \tag{37}$$

(38)

   6. Update weights:

$$w^{(1),1} = w^{(1),0} - \eta \frac{\partial C}{\partial w^{(1)}} \tag{39}$$

$$b^{(1),1} = b^{(1),0} - \eta \frac{\partial C}{\partial b^{(1)}} \tag{40}$$

# Summary: backward pass—first hidden layer

3. Backward pass, layer (1):
   5. Compute gradients:

$$\frac{\partial C}{\partial w^{(1)}} = \frac{\partial C}{\partial a^{(L)}} \frac{\partial a^{(L)}}{\partial z^{(L)}} \frac{\partial z^{(L)}}{\partial a^{(L-1)}} \frac{\partial a^{(L-1)}}{\partial z^{(L-1)}} \cdots \frac{\partial z^{(2)}}{\partial a^{(1)}} \frac{\partial a^{(1)}}{\partial z^{(1)}} \frac{\partial z^{(1)}}{\partial w^{(1)}} \tag{36}$$

$$\frac{\partial C}{\partial b^{(1)}} = \frac{\partial C}{\partial a^{(L)}} \frac{\partial a^{(L)}}{\partial z^{(L)}} \frac{\partial z^{(L)}}{\partial a^{(L-1)}} \frac{\partial a^{(L-1)}}{\partial z^{(L-1)}} \cdots \frac{\partial z^{(2)}}{\partial a^{(1)}} \frac{\partial a^{(1)}}{\partial z^{(1)}} \frac{\partial z^{(1)}}{\partial b^{(1)}} \tag{37}$$

$$\tag{38}$$

   6. Update weights:

$$w^{(1),1} = w^{(1),0} - \eta \frac{\partial C}{\partial w^{(1)}} \tag{39}$$

$$b^{(1),1} = b^{(1),0} - \eta \frac{\partial C}{\partial b^{(1)}} \tag{40}$$

# Summary: backward pass—first hidden layer

3. Backward pass, layer (1):
   5. Compute gradients:

$$\frac{\partial C}{\partial w^{(1)}} = \frac{\partial C}{\partial a^{(L)}} \frac{\partial a^{(L)}}{\partial z^{(L)}} \frac{\partial z^{(L)}}{\partial a^{(L-1)}} \frac{\partial a^{(L-1)}}{\partial z^{(L-1)}} \cdots \frac{\partial z^{(2)}}{\partial a^{(1)}} \frac{\partial a^{(1)}}{\partial z^{(1)}} \frac{\partial z^{(1)}}{\partial w^{(1)}} \tag{36}$$

$$\frac{\partial C}{\partial b^{(1)}} = \frac{\partial C}{\partial a^{(L)}} \frac{\partial a^{(L)}}{\partial z^{(L)}} \frac{\partial z^{(L)}}{\partial a^{(L-1)}} \frac{\partial a^{(L-1)}}{\partial z^{(L-1)}} \cdots \frac{\partial z^{(2)}}{\partial a^{(1)}} \frac{\partial a^{(1)}}{\partial z^{(1)}} \frac{\partial z^{(1)}}{\partial b^{(1)}} \tag{37}$$

(38)

   6. Update weights:

$$w^{(1),1} = w^{(1),0} - \eta \frac{\partial C}{\partial w^{(1)}} \tag{39}$$

$$b^{(1),1} = b^{(1),0} - \eta \frac{\partial C}{\partial b^{(1)}} \tag{40}$$

# Summary: backward pass—first hidden layer

3. Backward pass, layer (1):
   5. Compute gradients:

$$\frac{\partial C}{\partial w^{(1)}} = \frac{\partial C}{\partial a^{(L)}} \frac{\partial a^{(L)}}{\partial z^{(L)}} \frac{\partial z^{(L)}}{\partial a^{(L-1)}} \frac{\partial a^{(L-1)}}{\partial z^{(L-1)}} \cdots \frac{\partial z^{(2)}}{\partial a^{(1)}} \frac{\partial a^{(1)}}{\partial z^{(1)}} \frac{\partial z^{(1)}}{\partial w^{(1)}} \tag{36}$$

$$\frac{\partial C}{\partial b^{(1)}} = \frac{\partial C}{\partial a^{(L)}} \frac{\partial a^{(L)}}{\partial z^{(L)}} \frac{\partial z^{(L)}}{\partial a^{(L-1)}} \frac{\partial a^{(L-1)}}{\partial z^{(L-1)}} \cdots \frac{\partial z^{(2)}}{\partial a^{(1)}} \frac{\partial a^{(1)}}{\partial z^{(1)}} \frac{\partial z^{(1)}}{\partial b^{(1)}} \tag{37}$$

(38)

   6. Update weights:

$$w^{(1),1} = w^{(1),0} - \eta \frac{\partial C}{\partial w^{(1)}} \tag{39}$$

$$b^{(1),1} = b^{(1),0} - \eta \frac{\partial C}{\partial b^{(1)}} \tag{40}$$

# Summary: backward pass—first hidden layer

3. Backward pass, layer (1):
   5. Compute gradients:

$$\frac{\partial C}{\partial w^{(1)}} = \frac{\partial C}{\partial a^{(L)}} \frac{\partial a^{(L)}}{\partial z^{(L)}} \frac{\partial z^{(L)}}{\partial a^{(L-1)}} \frac{\partial a^{(L-1)}}{\partial z^{(L-1)}} \cdots \frac{\partial z^{(2)}}{\partial a^{(1)}} \frac{\partial a^{(1)}}{\partial z^{(1)}} \frac{\partial z^{(1)}}{\partial w^{(1)}} \tag{36}$$

$$\frac{\partial C}{\partial b^{(1)}} = \frac{\partial C}{\partial a^{(L)}} \frac{\partial a^{(L)}}{\partial z^{(L)}} \frac{\partial z^{(L)}}{\partial a^{(L-1)}} \frac{\partial a^{(L-1)}}{\partial z^{(L-1)}} \cdots \frac{\partial z^{(2)}}{\partial a^{(1)}} \frac{\partial a^{(1)}}{\partial z^{(1)}} \frac{\partial z^{(1)}}{\partial b^{(1)}} \tag{37}$$

(38)

   6. Update weights:

$$w^{(1),1} = w^{(1),0} - \eta \frac{\partial C}{\partial w^{(1)}} \tag{39}$$

$$b^{(1),1} = b^{(1),0} - \eta \frac{\partial C}{\partial b^{(1)}} \tag{40}$$

# Summary: backward pass—first hidden layer

3. Backward pass, layer (1):
   5. Compute gradients:

$$\frac{\partial C}{\partial w^{(1)}} = \frac{\partial C}{\partial a^{(L)}} \frac{\partial a^{(L)}}{\partial z^{(L)}} \frac{\partial z^{(L)}}{\partial a^{(L-1)}} \frac{\partial a^{(L-1)}}{\partial z^{(L-1)}} \cdots \frac{\partial z^{(2)}}{\partial a^{(1)}} \frac{\partial a^{(1)}}{\partial z^{(1)}} \frac{\partial z^{(1)}}{\partial w^{(1)}} \tag{36}$$

$$\frac{\partial C}{\partial b^{(1)}} = \frac{\partial C}{\partial a^{(L)}} \frac{\partial a^{(L)}}{\partial z^{(L)}} \frac{\partial z^{(L)}}{\partial a^{(L-1)}} \frac{\partial a^{(L-1)}}{\partial z^{(L-1)}} \cdots \frac{\partial z^{(2)}}{\partial a^{(1)}} \frac{\partial a^{(1)}}{\partial z^{(1)}} \frac{\partial z^{(1)}}{\partial b^{(1)}} \tag{37}$$

(38)

   6. Update weights:

$$w^{(1),1} = w^{(1),0} - \eta \frac{\partial C}{\partial w^{(1)}} \tag{39}$$

$$b^{(1),1} = b^{(1),0} - \eta \frac{\partial C}{\partial b^{(1)}} \tag{40}$$

# Summary: backward pass—first hidden layer

3. Backward pass, layer (1):
   5. Compute gradients:

$$\frac{\partial C}{\partial w^{(1)}} = \frac{\partial C}{\partial a^{(L)}} \frac{\partial a^{(L)}}{\partial z^{(L)}} \frac{\partial z^{(L)}}{\partial a^{(L-1)}} \frac{\partial a^{(L-1)}}{\partial z^{(L-1)}} \cdots \frac{\partial z^{(2)}}{\partial a^{(1)}} \frac{\partial a^{(1)}}{\partial z^{(1)}} \frac{\partial z^{(1)}}{\partial w^{(1)}} \tag{36}$$

$$\frac{\partial C}{\partial b^{(1)}} = \frac{\partial C}{\partial a^{(L)}} \frac{\partial a^{(L)}}{\partial z^{(L)}} \frac{\partial z^{(L)}}{\partial a^{(L-1)}} \frac{\partial a^{(L-1)}}{\partial z^{(L-1)}} \cdots \frac{\partial z^{(2)}}{\partial a^{(1)}} \frac{\partial a^{(1)}}{\partial z^{(1)}} \frac{\partial z^{(1)}}{\partial b^{(1)}} \tag{37}$$

(38)

   6. Update weights:

$$w^{(1),1} = w^{(1),0} - \eta \frac{\partial C}{\partial w^{(1)}} \tag{39}$$

$$b^{(1),1} = b^{(1),0} - \eta \frac{\partial C}{\partial b^{(1)}} \tag{40}$$

# Summary: backward pass—first hidden layer

3. Backward pass, layer (1):
   5. Compute gradients:

$$\frac{\partial C}{\partial w^{(1)}} = \frac{\partial C}{\partial a^{(L)}} \frac{\partial a^{(L)}}{\partial z^{(L)}} \frac{\partial z^{(L)}}{\partial a^{(L-1)}} \frac{\partial a^{(L-1)}}{\partial z^{(L-1)}} \cdots \frac{\partial z^{(2)}}{\partial a^{(1)}} \frac{\partial a^{(1)}}{\partial z^{(1)}} \frac{\partial z^{(1)}}{\partial w^{(1)}} \tag{36}$$

$$\frac{\partial C}{\partial b^{(1)}} = \frac{\partial C}{\partial a^{(L)}} \frac{\partial a^{(L)}}{\partial z^{(L)}} \frac{\partial z^{(L)}}{\partial a^{(L-1)}} \frac{\partial a^{(L-1)}}{\partial z^{(L-1)}} \cdots \frac{\partial z^{(2)}}{\partial a^{(1)}} \frac{\partial a^{(1)}}{\partial z^{(1)}} \frac{\partial z^{(1)}}{\partial b^{(1)}} \tag{37}$$

(38)

   6. Update weights:

$$w^{(1),1} = w^{(1),0} - \eta \frac{\partial C}{\partial w^{(1)}} \tag{39}$$

$$b^{(1),1} = b^{(1),0} - \eta \frac{\partial C}{\partial b^{(1)}} \tag{40}$$

# Summary: backward pass—first hidden layer

3. Backward pass, layer (1):
   5. Compute gradients:

$$\frac{\partial C}{\partial w^{(1)}} = \frac{\partial C}{\partial a^{(L)}} \frac{\partial a^{(L)}}{\partial z^{(L)}} \frac{\partial z^{(L)}}{\partial a^{(L-1)}} \frac{\partial a^{(L-1)}}{\partial z^{(L-1)}} \cdots \frac{\partial z^{(2)}}{\partial a^{(1)}} \frac{\partial a^{(1)}}{\partial z^{(1)}} \frac{\partial z^{(1)}}{\partial w^{(1)}} \tag{36}$$

$$\frac{\partial C}{\partial b^{(1)}} = \frac{\partial C}{\partial a^{(L)}} \frac{\partial a^{(L)}}{\partial z^{(L)}} \frac{\partial z^{(L)}}{\partial a^{(L-1)}} \frac{\partial a^{(L-1)}}{\partial z^{(L-1)}} \cdots \frac{\partial z^{(2)}}{\partial a^{(1)}} \frac{\partial a^{(1)}}{\partial z^{(1)}} \frac{\partial z^{(1)}}{\partial b^{(1)}} \tag{37}$$

(38)

   6. Update weights:

$$w^{(1),1} = w^{(1),0} - \eta \frac{\partial C}{\partial w^{(1)}} \tag{39}$$

$$b^{(1),1} = b^{(1),0} - \eta \frac{\partial C}{\partial b^{(1)}} \tag{40}$$

# Summary: backward pass—first hidden layer

3. Backward pass, layer (1):
   5. Compute gradients:

$$\frac{\partial C}{\partial w^{(1)}} = \frac{\partial C}{\partial a^{(L)}} \frac{\partial a^{(L)}}{\partial z^{(L)}} \frac{\partial z^{(L)}}{\partial a^{(L-1)}} \frac{\partial a^{(L-1)}}{\partial z^{(L-1)}} \cdots \frac{\partial z^{(2)}}{\partial a^{(1)}} \frac{\partial a^{(1)}}{\partial z^{(1)}} \frac{\partial z^{(1)}}{\partial w^{(1)}} \tag{36}$$

$$\frac{\partial C}{\partial b^{(1)}} = \frac{\partial C}{\partial a^{(L)}} \frac{\partial a^{(L)}}{\partial z^{(L)}} \frac{\partial z^{(L)}}{\partial a^{(L-1)}} \frac{\partial a^{(L-1)}}{\partial z^{(L-1)}} \cdots \frac{\partial z^{(2)}}{\partial a^{(1)}} \frac{\partial a^{(1)}}{\partial z^{(1)}} \frac{\partial z^{(1)}}{\partial b^{(1)}} \tag{37}$$

(38)

   6. Update weights:

$$w^{(1),1} = w^{(1),0} - \eta \frac{\partial C}{\partial w^{(1)}} \tag{39}$$

$$b^{(1),1} = b^{(1),0} - \eta \frac{\partial C}{\partial b^{(1)}} \tag{40}$$

# Summary: backward pass—first hidden layer

3. Backward pass, layer (1):
   5. Compute gradients:

$$\frac{\partial C}{\partial w^{(1)}} = \frac{\partial C}{\partial a^{(L)}} \frac{\partial a^{(L)}}{\partial z^{(L)}} \frac{\partial z^{(L)}}{\partial a^{(L-1)}} \frac{\partial a^{(L-1)}}{\partial z^{(L-1)}} \cdots \frac{\partial z^{(2)}}{\partial a^{(1)}} \frac{\partial a^{(1)}}{\partial z^{(1)}} \frac{\partial z^{(1)}}{\partial w^{(1)}} \tag{36}$$

$$\frac{\partial C}{\partial b^{(1)}} = \frac{\partial C}{\partial a^{(L)}} \frac{\partial a^{(L)}}{\partial z^{(L)}} \frac{\partial z^{(L)}}{\partial a^{(L-1)}} \frac{\partial a^{(L-1)}}{\partial z^{(L-1)}} \cdots \frac{\partial z^{(2)}}{\partial a^{(1)}} \frac{\partial a^{(1)}}{\partial z^{(1)}} \frac{\partial z^{(1)}}{\partial b^{(1)}} \tag{37}$$

(38)

   6. Update weights:

$$w^{(1),1} = w^{(1),0} - \eta \frac{\partial C}{\partial w^{(1)}} \tag{39}$$

$$b^{(1),1} = b^{(1),0} - \eta \frac{\partial C}{\partial b^{(1)}} \tag{40}$$

# Summary: backward pass—first hidden layer

3. Backward pass, layer (1):
   5. Compute gradients:

$$\frac{\partial C}{\partial w^{(1)}} = \frac{\partial C}{\partial a^{(L)}} \frac{\partial a^{(L)}}{\partial z^{(L)}} \frac{\partial z^{(L)}}{\partial a^{(L-1)}} \frac{\partial a^{(L-1)}}{\partial z^{(L-1)}} \cdots \frac{\partial z^{(2)}}{\partial a^{(1)}} \frac{\partial a^{(1)}}{\partial z^{(1)}} \frac{\partial z^{(1)}}{\partial w^{(1)}} \quad (36)$$

$$\frac{\partial C}{\partial b^{(1)}} = \frac{\partial C}{\partial a^{(L)}} \frac{\partial a^{(L)}}{\partial z^{(L)}} \frac{\partial z^{(L)}}{\partial a^{(L-1)}} \frac{\partial a^{(L-1)}}{\partial z^{(L-1)}} \cdots \frac{\partial z^{(2)}}{\partial a^{(1)}} \frac{\partial a^{(1)}}{\partial z^{(1)}} \frac{\partial z^{(1)}}{\partial b^{(1)}} \quad (37)$$

(38)

   6. Update weights:

$$w^{(1),1} = w^{(1),0} - \eta \frac{\partial C}{\partial w^{(1)}} \quad (39)$$

$$b^{(1),1} = b^{(1),0} - \eta \frac{\partial C}{\partial b^{(1)}} \quad (40)$$

# Summary: backward pass—first hidden layer

3. Backward pass, layer (1):
   5. Compute gradients:

$$\frac{\partial C}{\partial w^{(1)}} = \frac{\partial C}{\partial a^{(L)}} \frac{\partial a^{(L)}}{\partial z^{(L)}} \frac{\partial z^{(L)}}{\partial a^{(L-1)}} \frac{\partial a^{(L-1)}}{\partial z^{(L-1)}} \cdots \frac{\partial z^{(2)}}{\partial a^{(1)}} \frac{\partial a^{(1)}}{\partial z^{(1)}} \frac{\partial z^{(1)}}{\partial w^{(1)}} \tag{36}$$

$$\frac{\partial C}{\partial b^{(1)}} = \frac{\partial C}{\partial a^{(L)}} \frac{\partial a^{(L)}}{\partial z^{(L)}} \frac{\partial z^{(L)}}{\partial a^{(L-1)}} \frac{\partial a^{(L-1)}}{\partial z^{(L-1)}} \cdots \frac{\partial z^{(2)}}{\partial a^{(1)}} \frac{\partial a^{(1)}}{\partial z^{(1)}} \frac{\partial z^{(1)}}{\partial b^{(1)}} \tag{37}$$

(38)

   6. Update weights:

$$w^{(1),1} = w^{(1),0} - \eta \frac{\partial C}{\partial w^{(1)}} \tag{39}$$

$$b^{(1),1} = b^{(1),0} - \eta \frac{\partial C}{\partial b^{(1)}} \tag{40}$$

# Summary: backward pass—first hidden layer

3. Backward pass, layer (1):
   5. Compute gradients:

$$\frac{\partial C}{\partial w^{(1)}} = \frac{\partial C}{\partial a^{(L)}} \frac{\partial a^{(L)}}{\partial z^{(L)}} \frac{\partial z^{(L)}}{\partial a^{(L-1)}} \frac{\partial a^{(L-1)}}{\partial z^{(L-1)}} \cdots \frac{\partial z^{(2)}}{\partial a^{(1)}} \frac{\partial a^{(1)}}{\partial z^{(1)}} \frac{\partial z^{(1)}}{\partial w^{(1)}} \tag{36}$$

$$\frac{\partial C}{\partial b^{(1)}} = \frac{\partial C}{\partial a^{(L)}} \frac{\partial a^{(L)}}{\partial z^{(L)}} \frac{\partial z^{(L)}}{\partial a^{(L-1)}} \frac{\partial a^{(L-1)}}{\partial z^{(L-1)}} \cdots \frac{\partial z^{(2)}}{\partial a^{(1)}} \frac{\partial a^{(1)}}{\partial z^{(1)}} \frac{\partial z^{(1)}}{\partial b^{(1)}} \tag{37}$$

(38)

   6. Update weights:

$$w^{(1),1} = w^{(1),0} - \eta \frac{\partial C}{\partial w^{(1)}} \tag{39}$$

$$b^{(1),1} = b^{(1),0} - \eta \frac{\partial C}{\partial b^{(1)}} \tag{40}$$

# Summary: backward pass—first hidden layer

3. Backward pass, layer (1):
   5. Compute gradients:

$$\frac{\partial C}{\partial w^{(1)}} = \frac{\partial C}{\partial a^{(L)}} \frac{\partial a^{(L)}}{\partial z^{(L)}} \frac{\partial z^{(L)}}{\partial a^{(L-1)}} \frac{\partial a^{(L-1)}}{\partial z^{(L-1)}} \cdots \frac{\partial z^{(2)}}{\partial a^{(1)}} \frac{\partial a^{(1)}}{\partial z^{(1)}} \frac{\partial z^{(1)}}{\partial w^{(1)}} \tag{36}$$

$$\frac{\partial C}{\partial b^{(1)}} = \frac{\partial C}{\partial a^{(L)}} \frac{\partial a^{(L)}}{\partial z^{(L)}} \frac{\partial z^{(L)}}{\partial a^{(L-1)}} \frac{\partial a^{(L-1)}}{\partial z^{(L-1)}} \cdots \frac{\partial z^{(2)}}{\partial a^{(1)}} \frac{\partial a^{(1)}}{\partial z^{(1)}} \frac{\partial z^{(1)}}{\partial b^{(1)}} \tag{37}$$

(38)

   6. Update weights:

$$w^{(1),1} = w^{(1),0} - \eta \frac{\partial C}{\partial w^{(1)}} \tag{39}$$

$$b^{(1),1} = b^{(1),0} - \eta \frac{\partial C}{\partial b^{(1)}} \tag{40}$$

# Summary: backward pass—first hidden layer

3. Backward pass, layer (1):
   5. Compute gradients:

$$\frac{\partial C}{\partial w^{(1)}} = \frac{\partial C}{\partial a^{(L)}} \frac{\partial a^{(L)}}{\partial z^{(L)}} \frac{\partial z^{(L)}}{\partial a^{(L-1)}} \frac{\partial a^{(L-1)}}{\partial z^{(L-1)}} \cdots \frac{\partial z^{(2)}}{\partial a^{(1)}} \frac{\partial a^{(1)}}{\partial z^{(1)}} \frac{\partial z^{(1)}}{\partial w^{(1)}} \tag{36}$$

$$\frac{\partial C}{\partial b^{(1)}} = \frac{\partial C}{\partial a^{(L)}} \frac{\partial a^{(L)}}{\partial z^{(L)}} \frac{\partial z^{(L)}}{\partial a^{(L-1)}} \frac{\partial a^{(L-1)}}{\partial z^{(L-1)}} \cdots \frac{\partial z^{(2)}}{\partial a^{(1)}} \frac{\partial a^{(1)}}{\partial z^{(1)}} \frac{\partial z^{(1)}}{\partial b^{(1)}} \tag{37}$$

$$\tag{38}$$

   6. Update weights:

$$w^{(1),1} = w^{(1),0} - \eta \frac{\partial C}{\partial w^{(1)}} \tag{39}$$

$$b^{(1),1} = b^{(1),0} - \eta \frac{\partial C}{\partial b^{(1)}} \tag{40}$$