CSC311 Introduction to Machine Learning

Vectorizing The Backpropagation Algorithm

Alice Gao and Marina Tawfik

Learning Outcomes

By the end of this lecture, students should be able to

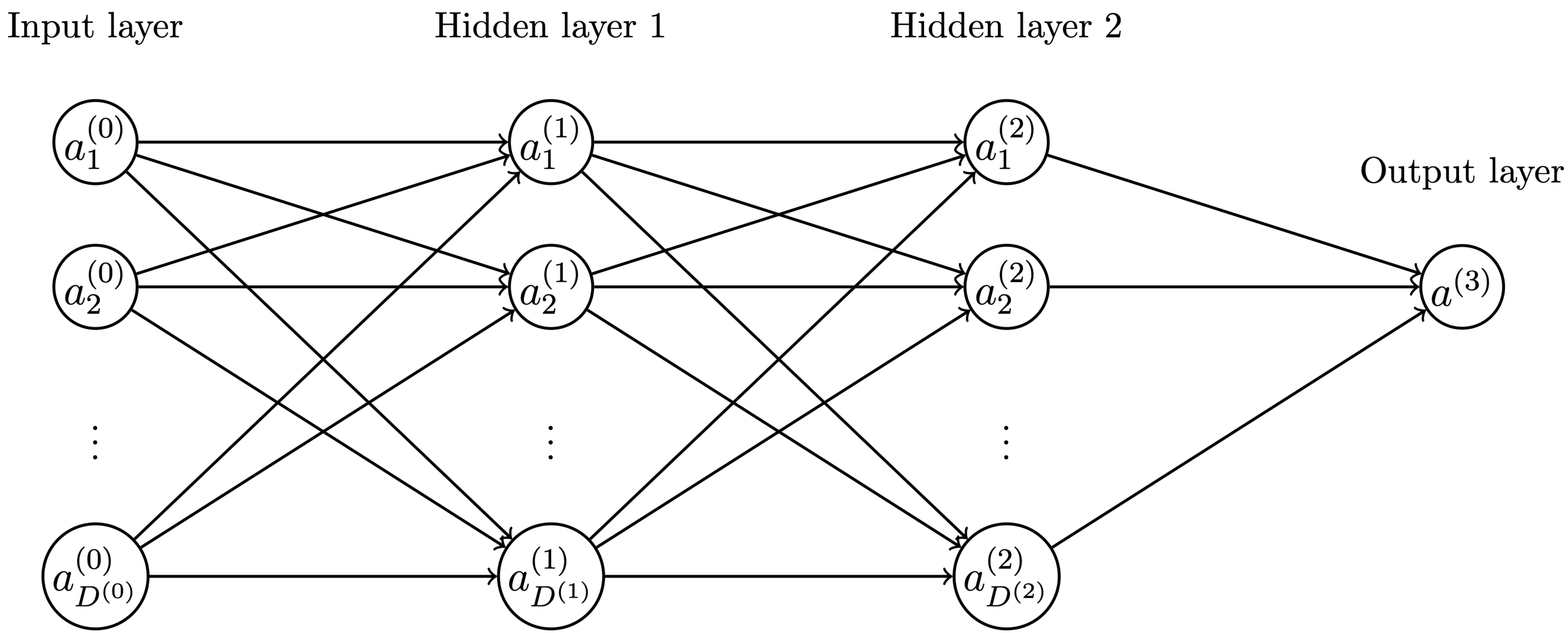
- Derive vectorized expressions for quantities in the forward pass of the backpropagation algorithm for a small neural network.
- Derive vectorized expressions for gradients with respect to activations, pre-activations, and weights in the backward pass of the backpropagation algorithm for a small neural network.

Outline

- Review of 3-Layer Neural Network
- Vectorizing Forward Pass
- Vectorizing Backward Pass

Review: 3-Layer Neural Network

3-Layer Neural Network



Notation and Definitions

L: number of layers

$\mathbf{a}^{(0)}$: input to the model ($\mathbf{x}$)

t: target output

C: loss function $C(a^{(L)}, t)$

For each layer m where $m \in [1, L]$

$D^{(m)}$: dimensionality of layer m

$\mathbf{W}^{(\mathrm{m})}$: weight matrix for layer m, $\mathbb{R}^{D^{(m)} \times D^{(m-1)}}$

weights going into the mth layer

output input

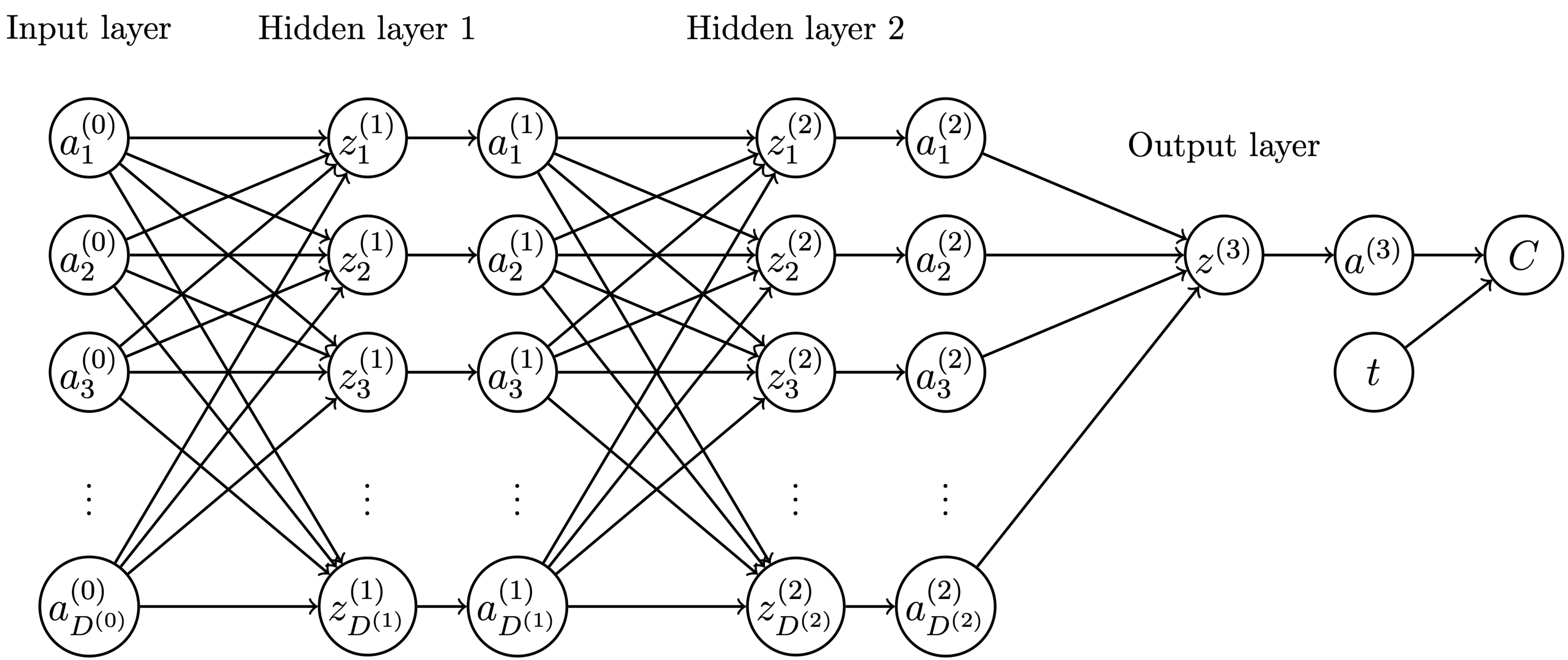
$\mathbf{b}^{(\mathrm{m})}$: bias vector for layer m, $\mathbb{R}^{D^{(m)}}$

$\sigma^{(m)}$: non-linearity for layer m

$\mathbf{z}^{(\mathrm{m})}$: pre-activations for layer m $\mathbb{R}^{D^{(m)}}$

$\mathbf{a}^{(\mathrm{m})}$: activations for layer m $\mathbb{R}^{D^{(m)}}$

Computation Graph



Vectorizing Forward Pass

Forward Pass Computations

layer 1:

$$z_i^{(1)} = \sum_{j=1}^{D^{(0)}} W_{ij}^{(1)} a_j^{(0)} + b_i^{(1)}$$

$$a_i^{(1)} = \sigma^{(1)}\left(z_i^{(1)}\right)$$

layer 2:

$$z_i^{(2)} = \sum_{j=1}^{D^{(1)}} W_{ij}^{(2)} a_j^{(1)} + b_i^{(2)}$$

$$a_i^{(2)} = \sigma^{(2)}\left(z_i^{(2)}\right)$$

layer 3:

$$z^{(3)} = \sum_{j=1}^{D^{(2)}} w_j^{(3)} a_j^{(2)} + b^{(3)}$$

$$a^{(3)} = \sigma^{(3)}\left(z^{(3)}\right)$$

$$\mathrm{C} = C(a^{(3)}, t)$$

Vectorizing Forward Pass Computations

Non-Vectorized

For each $m = 1, \ldots, L$,

$$z_i^{(m+1)} = \left(\sum_j W_{ij}^{(m+1)} a_j^{(m)} \right) + b_i^{(m+1)}$$

$$a_i^{(m+1)} = \sigma^{(m+1)} \left(z_i^{(m+1)} \right)$$

For output layer L

$$C = C(a^{(L)}, t)$$

Vectorized

For each $m = 1, \ldots, L$,

For output layer L

$$C = C(a^{(L)}, t)$$

Result of Vectorizing Weighted Sum

$W^{(m+1)}$: $D^{(m+1)} \times D^{(m)}$ $\vec{a}^{(m)}$: $D^{(m)} \times 1$

What is the vectorized expression for the following?

$$\begin{bmatrix} z_1^{(m+1)} \\ z_2^{(m+1)} \\ \vdots \\ z_{D^{(m+1)}}^{(m+1)} \end{bmatrix} \leftarrow z_i^{(m+1)} = \sum_{j=1}^{D^{(m)}} W_{ij}^{(m+1)} a_j^{(m)}, \qquad i = 1, \dots, D^{(m+1)}$$

dim: $D^{(m+1)} \times 1$

$$W^{(m+1)} = \begin{bmatrix} w_{11} & w_{12} & \cdots \\ & & \\ & & \end{bmatrix}$$

all the weights going into the 1st unit in the (m+1) th layer

output × input: $D^{(m+1)} \times D^{(m)}$

(A) $\left(\mathbf{W}^{(\mathrm{m}+1)}\right)^{\mathsf{T}} \mathbf{a}^{(\mathrm{m})}$	(C) $\mathbf{W}^{(\mathrm{m}+1)} \mathbf{a}^{(\mathrm{m})}$
(B) $\left(\mathbf{a}^{(\mathrm{m})}\right)^{\mathsf{T}} \mathbf{W}^{(\mathrm{m}+1)}$	(D) $\mathbf{a}^{(\mathrm{m})} \mathbf{W}^{(\mathrm{m}+1)}$

Solution: Result of Vectorizing Weighted Sum

What is the vectorized expression for the following?

$$\sum_{j=1}^{D^{(m)}} W_{ij}^{(m+1)} a_j^{(m)}, \qquad i = 1, \dots, D^{(m+1)}$$

(A) $\left(\mathbf{W}^{(\mathrm{m}+1)}\right)^{\mathsf{T}} \mathbf{a}^{(\mathrm{m})}$	(C) $\mathbf{W}^{(\mathrm{m}+1)} \mathbf{a}^{(\mathrm{m})}$ **(Correct)**
(B) $\left(\mathbf{a}^{(\mathrm{m})}\right)^{\mathsf{T}} \mathbf{W}^{(\mathrm{m}+1)}$	(D) $\mathbf{a}^{(\mathrm{m})} \mathbf{W}^{(\mathrm{m}+1)}$

Recall the dimensions $\mathbf{W}^{(m+1)} \in \mathbb{R}^{D^{(m+1)} \times D^{(m)}}$, $\mathbf{a}^{(m)} \in \mathbb{R}^{D^{(m)} \times 1}$

Vectorizing Forward Pass Computations

Non-Vectorized

For each $m = 1, \dots, L$,

$$z_i^{(m+1)} = \sum_j W_{ij}^{(m+1)} a_j^{(m)} + b_i^{(m+1)}$$

$$a_i^{(m+1)} = \sigma^{(m+1)}\left(z_i^{(m+1)}\right)$$

For output layer L

$$C = C(a^{(L)}, t)$$

Vectorized

For each $m = 1, \dots, L$,

$$\mathbf{z}^{(m+1)} = \mathbf{W}^{(m+1)}\mathbf{a}^{(m)} + \mathbf{b}^{(m+1)}$$

$$\mathbf{a}^{(m+1)} = \sigma^{(m+1)}\left(\mathbf{z}^{(m+1)}\right)$$

For output layer L

$$C = C(a^{(L)}, t)$$

Vectorizing Backward Pass

Backward Pass Computations

1. Compute gradients for output layer → this assumes a scalar output

$$\frac{\partial C}{\partial z^{(L)}} = \frac{\partial C}{\partial a^{(L)}} \cdot \sigma^{(L)'}\left(z^{(L)}\right)$$

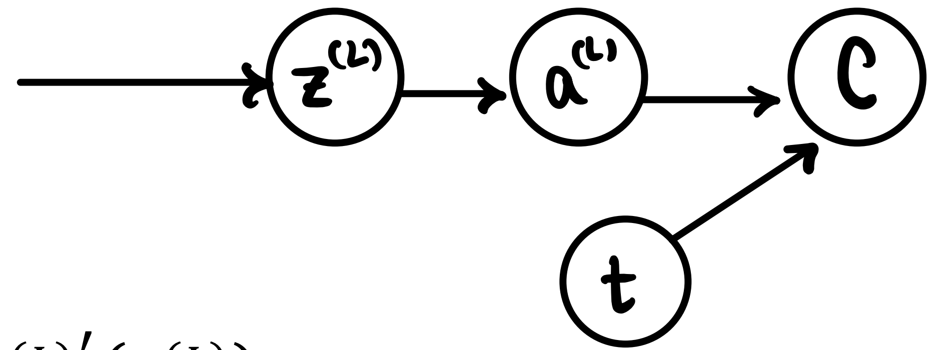
2. Compute gradients for each hidden layer recursively

$$\frac{\partial C}{\partial z_j^{(m)}} = \left(\sum_i \frac{\partial C}{\partial z_i^{(m+1)}} \cdot W_{ij}^{(m+1)}\right) \cdot \sigma^{(m)'}\left(z_j^{(m)}\right), \qquad j = 1, \dots, D^{(m)}$$

3. Compute gradients for the weights

$$\frac{\partial C}{\partial W_{ij}^{(m)}} = \frac{\partial C}{\partial z_i^{(m)}} \cdot a_j^{(m-1)}, \qquad i = 1, \dots, D^{(m)}, j = 1, \dots, D^{(m-1)}$$

Gradients for the Output Layer

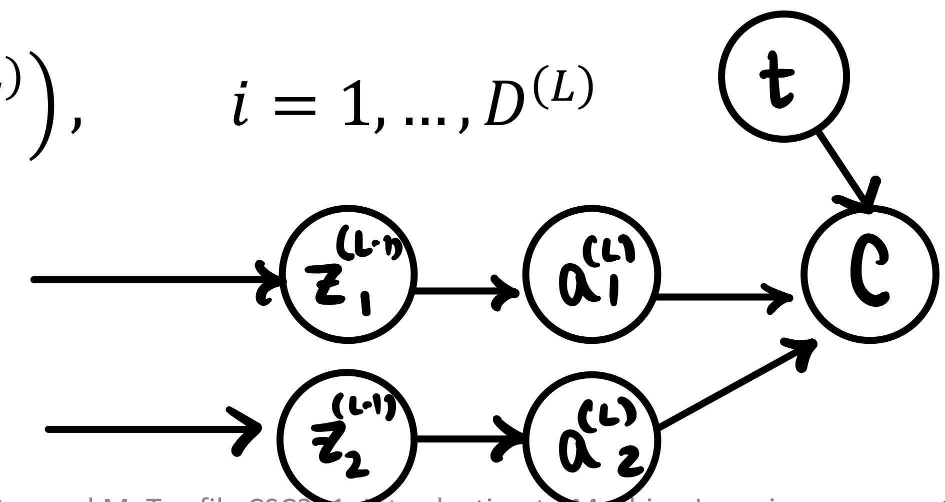


In our network, $z^{(L)}, a^{(L)}, t \in \mathbb{R}$ are scalars.

$$\frac{\partial C}{\partial z^{(L)}} = \frac{\partial C}{\partial a^{(L)}} \cdot \frac{\partial a^{(L)}}{\partial z^{(L)}} = \frac{\partial C}{\partial a^{(L)}} \cdot \sigma^{(L)'}\left(z^{(L)}\right)$$

In general, $\mathbf{z}^{(L)}, \mathbf{a}^{(L)}, \mathbf{t} \in \mathbb{R}^{D^{(L)} \times 1}$ can be vectors. e.g. multi-class classification

$$\frac{\partial C}{\partial z_i^{(L)}} = \frac{\partial C}{\partial a_i^{(L)}} \cdot \frac{\partial a_i^{(L)}}{\partial z_i^{(L)}} = \frac{\partial C}{\partial a_i^{(L)}} \cdot \sigma^{(L)'}\left(z_i^{(L)}\right), \qquad i = 1, \ldots, D^{(L)}$$



Exercise 1: Gradients for the Output Layer

What is the vectorized expression for the following?

$$\frac{\partial C}{\partial z_i^{(L)}} = \frac{\partial C}{\partial a_i^{(L)}} \cdot \sigma^{(L)\prime}\left(z_i^{(L)}\right), \qquad i = 1, \ldots, D^{(L)}$$

(A) $\left(\nabla_{\mathbf{a}^{(L)}} C\right)^{\top} \left(\sigma^{(L)\prime}\left(\mathbf{z}^{(L)}\right)\right)$

(B) $\left(\sigma^{(L)\prime}\left(\mathbf{z}^{(L)}\right)\right)^{\top} \left(\nabla_{\mathbf{a}^{(L)}} C\right)$

(C) $\left(\nabla_{\mathbf{a}^{(L)}} C\right) \odot \left(\sigma^{(L)\prime}\left(\mathbf{z}^{(L)}\right)\right)$

(D) $\left(\sigma^{(L)\prime}\left(\mathbf{z}^{(L)}\right)\right) \left(\nabla_{\mathbf{a}^{(L)}} C\right)^{\top}$

(E) $\left(\nabla_{\mathbf{a}^{(L)}} C\right) \left(\sigma^{(L)\prime}\left(\mathbf{z}^{(L)}\right)\right)^{\top}$

Solution 1: Gradients for the Output Layer

What is the vectorized expression for the following?

$$\frac{\partial C}{\partial z_i^{(L)}} = \frac{\partial C}{\partial a_i^{(L)}} \cdot \sigma^{(L)'}\left(z_i^{(L)}\right), \qquad i = 1, \dots, D^{(L)}$$

(A) $\left(\nabla_{\mathbf{a}^{(L)}} C\right)^{\mathsf{T}} \left(\sigma^{(L)'}\left(\mathbf{z}^{(L)}\right)\right)$

(D) $\left(\sigma^{(L)'}\left(\mathbf{z}^{(L)}\right)\right) \left(\nabla_{\mathbf{a}^{(L)}} C\right)^{\mathsf{T}}$

(B) $\left(\sigma^{(L)'}\left(\mathbf{z}^{(L)}\right)\right)^{\mathsf{T}} \left(\nabla_{\mathbf{a}^{(L)}} C\right)$

(E) $\left(\nabla_{\mathbf{a}^{(L)}} C\right) \left(\sigma^{(L)'}\left(\mathbf{z}^{(L)}\right)\right)^{\mathsf{T}}$

(C) $\left(\nabla_{\mathbf{a}^{(L)}} C\right) \odot \left(\sigma^{(L)'}\left(\mathbf{z}^{(L)}\right)\right)$ **(Correct)**

Recall the dimensions: $\nabla_{\mathbf{a}^{(L)}} C, \sigma^{(L)'}\left(\mathbf{z}^{(L)}\right), \nabla_{\mathbf{z}^{(L)}} C \in \mathbb{R}^{D^{(L)} \times 1}$

Vectorizing Gradients for the Output Layer

Non-vectorized:

$$\frac{\partial C}{\partial z_i^{(L)}} = \frac{\partial C}{\partial a_i^{(L)}} \cdot \frac{\partial a_i^{(L)}}{\partial z_i^{(L)}} = \frac{\partial C}{\partial a_i^{(L)}} \cdot \sigma^{(L)\prime}\left(z_i^{(L)}\right), \qquad i = 1, \dots, D^{(L)}$$

Vectorized:

Solution: Vectorizing Gradients for the Output Layer

Non-vectorized:

$$\frac{\partial C}{\partial z_i^{(L)}} = \frac{\partial C}{\partial a_i^{(L)}} \cdot \frac{\partial a_i^{(L)}}{\partial z_i^{(L)}} = \frac{\partial C}{\partial a_i^{(L)}} \cdot \sigma^{(L)'}\left(z_i^{(L)}\right), \qquad i = 1, \dots, D^{(L)}$$

Vectorized:

$$\nabla_{\mathbf{z}^{(L)}} \mathrm{C} = \left(\nabla_{\mathbf{a}^{(L)}} \mathrm{C}\right) \odot \left(\sigma^{(L)'}\left(\mathbf{z}^{(L)}\right)\right)$$

Backward Pass Computations

1. Compute gradients for output layer

$$\nabla_{\mathbf{z}^{(L)}} \mathrm{C} = \left(\nabla_{\mathbf{a}^{(L)}} \mathrm{C}\right) \odot \left(\sigma^{(L)\prime}\left(\mathbf{z}^{(L)}\right)\right)$$

2. Compute gradients for each hidden layer recursively

$$\frac{\partial C}{\partial z_j^{(m)}} = \left(\sum_i \frac{\partial C}{\partial z_i^{(m+1)}} \cdot W_{ij}^{(m+1)}\right) \cdot \sigma^{(m)\prime}\left(z_j^{(m)}\right), \qquad j = 1, \ldots, D^{(m)}$$

3. Compute gradients for the weights

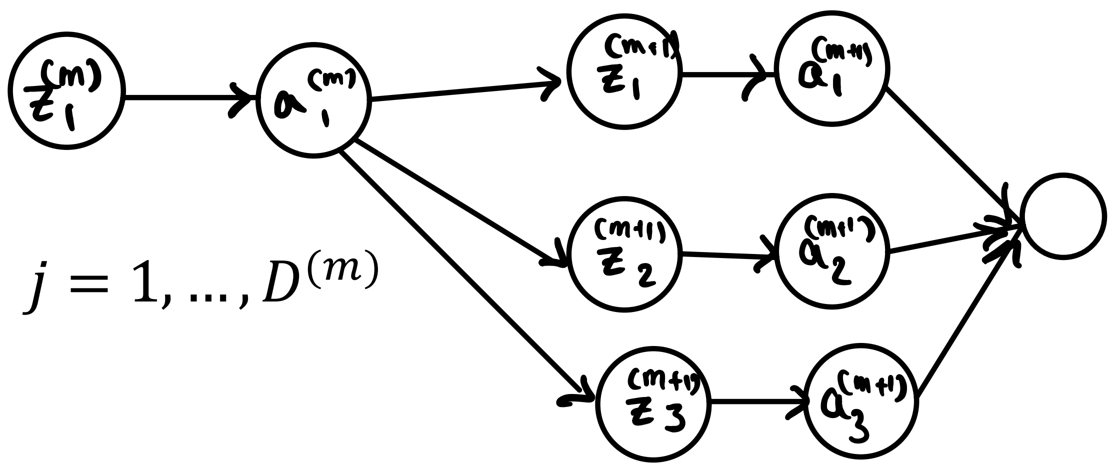
$$\frac{\partial C}{\partial W_{ij}^{(m)}} = \frac{\partial C}{\partial z_i^{(m)}} \cdot a_j^{(m-1)}, \qquad i = 1, \ldots, D^{(m)}, j = 1, \ldots, D^{(m-1)}$$

Vectorizing the Recursive Step

$$\frac{\partial C}{\partial z_j^{(m)}} = \left(\sum_{i=1}^{D^{(m+1)}} \left(\frac{\partial C}{\partial z_i^{(m+1)}} \cdot W_{ij}^{(m+1)} \right) \right) \cdot \sigma^{(m)\prime}\left(z_j^{(m)}\right), \qquad j = 1, \ldots, D^{(m)}$$

1. Vectorizing

$$\frac{\partial C}{\partial a_j^{(m)}} = \sum_i \left(\frac{\partial C}{\partial z_i^{(m+1)}} \cdot W_{ij}^{(m+1)} \right), \qquad j = 1, \ldots, D^{(m)}$$



2. Vectorizing

$$\frac{\partial C}{\partial z_j^{(m)}} = \frac{\partial C}{\partial a_j^{(m)}} \cdot \sigma^{(m)\prime}\left(z_j^{(m)}\right), \qquad j = 1, \ldots, D^{(m)}$$

Exercise 2: Loss Derivative with respect to Activations

What is the vectorized expression for the following?

$$\frac{\partial C}{\partial a_j^{(m)}} = \sum_{i=1}^{D^{(m+1)}} \left(\frac{\partial C}{\partial z_i^{(m+1)}} \cdot W_{ij}^{(m+1)} \right), \qquad j = 1, \ldots, D^{(m)}$$

$\nabla_{\vec{a}^{(m)}} C$ $D^{(m)} \times 1$ $\nabla_{\vec{z}^{(m+1)}} C$ $D^{(m+1)} \times 1$ $W^{(m+1)}$ $D^{(m+1)} \times D^{(m)}$

(A) $(\mathbf{W}^{(m+1)})^{\top} (\nabla_{\mathbf{z}^{(m+1)}} C)$	(C) $\mathbf{W}^{(m+1)} (\nabla_{\mathbf{z}^{(m+1)}} C)$
(B) $(\nabla_{\mathbf{z}^{(m+1)}} C)^{\top} \mathbf{W}^{(m+1)}$	(D) $(\nabla_{\mathbf{z}^{(m+1)}} C) \mathbf{W}^{(m+1)}$

UNIVERSITY OF TORONTO

Solution 2: Loss Derivative with respect to Activations

What is the vectorized expression for the following?

$$\frac{\partial C}{\partial a_j^{(m)}} = \sum_{i=1}^{D^{(m+1)}} \left(\frac{\partial C}{\partial z_i^{(m+1)}} \cdot W_{ij}^{(m+1)} \right), \qquad j = 1, \dots, D^{(m)}$$

(A) $(\mathbf{W}^{(\mathrm{m+1})})^{\mathsf{T}} \, (\nabla_{\mathbf{z}^{(\mathrm{m+1})}} \mathrm{C})$ **(Correct)**	(C) $\mathbf{W}^{(\mathrm{m+1})} \, (\nabla_{\mathbf{z}^{(\mathrm{m+1})}} \mathrm{C})$
(B) $(\nabla_{\mathbf{z}^{(\mathrm{m+1})}} \mathrm{C})^{\mathsf{T}} \, \mathbf{W}^{(\mathrm{m+1})}$	(D) $(\nabla_{\mathbf{z}^{(\mathrm{m+1})}} \mathrm{C}) \, \mathbf{W}^{(\mathrm{m+1})}$

Recall the dimensions: $\mathbf{W}^{(\mathrm{m+1})} \in \mathbb{R}^{D^{(m+1)} \times D^{(m)}}$, $\nabla_{\mathbf{z}^{(\mathrm{m+1})}} \mathrm{C} \in \mathbb{R}^{D^{(m+1)} \times 1}$

Exercise 3: Loss Derivative with respect to Pre-activations

What is the vectorized expression for the following?

$$\frac{\partial C}{\partial z_j^{(m)}} = \frac{\partial C}{\partial a_j^{(m)}} \cdot \sigma^{(m)'}\left(z_j^{(m)}\right), \qquad i = 1 \ldots, D^{(m)}$$

(A) $\left(\nabla_{\mathbf{a}^{(m)}} C\right)^{\mathsf{T}} \left(\sigma^{(m)'}\left(\mathbf{z}^{(m)}\right)\right)$	(D) $\left(\sigma^{(m)'}\left(\mathbf{z}^{(m)}\right)\right)\left(\nabla_{\mathbf{a}^{(m)}} C\right)^{\mathsf{T}}$
(B) $\left(\sigma^{(m)'}\left(\mathbf{z}^{(m)}\right)\right)^{\mathsf{T}} \left(\nabla_{\mathbf{a}^{(m)}} C\right)$	(E) $\left(\nabla_{\mathbf{a}^{(m)}} C\right)\left(\sigma^{(m)'}\left(\mathbf{z}^{(m)}\right)\right)^{\mathsf{T}}$
(C) $\left(\nabla_{\mathbf{a}^{(m)}} C\right) \odot \left(\sigma^{(m)'}\left(\mathbf{z}^{(m)}\right)\right)$	

Solution 3: Loss Derivative with respect to Pre-activations

What is the vectorized expression for the following?

$$\frac{\partial C}{\partial z_j^{(m)}} = \frac{\partial C}{\partial a_j^{(m)}} \cdot \sigma^{(m)\prime}\left(z_j^{(m)}\right), \qquad i = 1 \dots, D^{(m)}$$

(A) $\left(\nabla_{\mathbf{a}^{(\mathrm{m})}} \mathrm{C}\right)^{\mathrm{T}} \left(\sigma^{(m)\prime}\left(\mathbf{z}^{(\mathrm{m})}\right)\right)$	(D) $\left(\sigma^{(\mathrm{m})\prime}\left(\mathbf{z}^{(\mathrm{m})}\right)\right) \left(\nabla_{\mathbf{a}^{(\mathrm{m})}} \mathrm{C}\right)^{\mathrm{T}}$
(B) $\left(\sigma^{(\mathrm{m})\prime}\left(\mathbf{z}^{(\mathrm{m})}\right)\right)^{\mathrm{T}} \left(\nabla_{\mathbf{a}^{(\mathrm{m})}} \mathrm{C}\right)$	(E) $\left(\nabla_{\mathbf{a}^{(\mathrm{m})}} \mathrm{C}\right) \left(\sigma^{(\mathrm{m})\prime}\left(\mathbf{z}^{(\mathrm{m})}\right)\right)^{\mathrm{T}}$
(C) $\left(\nabla_{\mathbf{a}^{(\mathrm{m})}} \mathrm{C}\right) \odot \left(\sigma^{(m)\prime}\left(\mathbf{z}^{(\mathrm{m})}\right)\right)$ **(Correct)**	

Recall the dimensions: $\nabla_{\mathbf{a}^{(\mathrm{m})}} \mathrm{C}, \sigma^{(m)\prime}\left(\mathbf{z}^{(\mathrm{m})}\right) \in \mathbb{R}^{D^{(m)} \times 1}$

UNIVERSITY OF TORONTO

Vectorizing the Recursive Step

Non-vectorized:

$$\frac{\partial C}{\partial z_j^{(m)}} = \left(\sum_{i=1}^{D^{(m+1)}} \left(\frac{\partial C}{\partial z_i^{(m+1)}} \cdot W_{ij}^{(m+1)} \right) \right) \cdot \sigma^{(m)\prime}\left(z_j^{(m)}\right), \qquad i = 1 \ldots, D^{(m)}$$

Vectorized:

Vectorizing the Recursive Step

Non-vectorized:

$$\frac{\partial C}{\partial z_j^{(m)}} = \left(\sum_{i=1}^{D^{(m+1)}} \left(\frac{\partial C}{\partial z_i^{(m+1)}} \cdot W_{ij}^{(m+1)} \right) \right) \cdot \sigma^{(m)'}\left(z_j^{(m)}\right), \qquad i = 1 \ldots, D^{(m)}$$

Vectorized:

$$\nabla_{\mathbf{z}^{(m)}} \mathrm{C} = \left(\left(\mathbf{W}^{(\mathrm{m}+1)}\right)^{\top} \left(\nabla_{\mathbf{z}^{(\mathrm{m}+1)}} \mathrm{C}\right) \right) \odot \left(\sigma^{(m)'}\left(\mathbf{z}^{(\mathrm{m})}\right) \right)$$

Backward Pass Computations

1. Compute gradients for output layer

$$\nabla_{\mathbf{z}^{(L)}} \mathrm{C} = \left(\nabla_{\mathbf{a}^{(L)}} \mathrm{C}\right) \odot \left(\sigma^{(L)'}\left(\mathbf{z}^{(L)}\right)\right)$$

2. Compute gradients for each hidden layer recursively

$$\nabla_{\mathbf{z}^{(m)}} \mathrm{C} = \left(\left(\mathbf{W}^{(\mathrm{m}+1)}\right)^{\top} \left(\nabla_{\mathbf{z}^{(\mathrm{m}+1)}} \mathrm{C}\right)\right) \odot \left(\sigma^{(m)'}\left(\mathbf{z}^{(\mathrm{m})}\right)\right)$$

3. Compute gradients for the weights

$$\frac{\partial \mathrm{C}}{\partial W_{ij}^{(m)}} = \frac{\partial \mathrm{C}}{\partial z_i^{(m)}} \cdot \frac{\partial z_i^{(m)}}{\partial W_{ij}^{(m)}} = \frac{\partial \mathrm{C}}{\partial z_i^{(m)}} \cdot a_j^{(m-1)}$$

Exercise 4: Loss Derivative with respect to Weights

What is the vectorized expression for the following?

$$\frac{\partial C}{\partial W_{ij}^{(m)}} = \frac{\partial C}{\partial z_i^{(m)}} \cdot a_j^{(m-1)}, \qquad i = 1, \dots, D^{(m)}, j = 1, \dots, D^{(m-1)}$$

(A) $\left(\nabla_{\mathbf{z}^{(m)}} C\right)^{\top}\left(\mathbf{a}^{(m-1)}\right)$	(D) $\left(\mathbf{a}^{(m-1)}\right)\left(\nabla_{\mathbf{z}^{(m)}} C\right)^{\top}$
(B) $\left(\mathbf{a}^{(m-1)}\right)^{\top}\left(\nabla_{\mathbf{z}^{(m)}} C\right)$	(E) $\left(\nabla_{\mathbf{z}^{(m)}} C\right)\left(\mathbf{a}^{(m-1)}\right)^{\top}$
(C) $\left(\nabla_{\mathbf{z}^{(m)}} C\right) \odot \left(\mathbf{a}^{(m-1)}\right)$	

$\begin{bmatrix} \frac{\partial C}{\partial z_1^{(m)}} \\ \frac{\partial C}{\partial z_2^{(m)}} \\ \vdots \end{bmatrix}$ $\nabla_{\vec{z}^{(m)}} C$ $D^{(m)} \times 1$

outer products

$\vec{a}^{(m-1)}$ $D^{(m-1)} \times 1$

$D^{(m)} \times D^{(m-1)}$

$\begin{bmatrix} \frac{\partial C}{\partial w_{11}} & \frac{\partial C}{\partial w_{12}} & \cdots & \frac{\partial C}{\partial w_{1D^{(m-1)}}} \\ \vdots & \vdots & & \vdots \end{bmatrix}$

output x input

weights into the 1st unit in the mth layer

Solution 4: Loss Derivative with respect to Weights

What is the vectorized expression for the following?

$$\frac{\partial C}{\partial W_{ij}^{(m)}} = \frac{\partial C}{\partial z_i^{(m)}} \cdot a_j^{(m-1)}, \qquad i = 1, \dots, D^{(m)}, j = 1, \dots, D^{(m-1)}$$

(A) $(\nabla_{\mathbf{z}^{(m)}} C)^{\top}(\mathbf{a}^{(m-1)})$	(D) $(\mathbf{a}^{(m-1)})(\nabla_{\mathbf{z}^{(m)}} C)^{\top}$
(B) $(\mathbf{a}^{(m-1)})^{\top}(\nabla_{\mathbf{z}^{(m)}} C)$	(E) $(\nabla_{\mathbf{z}^{(m)}} C)(\mathbf{a}^{(m-1)})^{\top}$ **(Correct)**
(C) $(\nabla_{\mathbf{z}^{(m)}} C) \odot (\mathbf{a}^{(m-1)})$	

Recall the dimensions: $\nabla_{\mathbf{z}^{(m)}} C \in \mathbb{R}^{D^{(m)} \times 1}$, $\mathrm{a}^{(m-1)} \in \mathbb{R}^{D^{(m-1)} \times 1}$, $\nabla_{\mathbf{W}^{(m)}} C \in \mathbb{R}^{D^{(m)} \times D^{(m-1)}}$

Vectorizing the Gradients for the Weights

Non-vectorized:

$$\frac{\partial C}{\partial W_{ij}^{(m)}} = \frac{\partial C}{\partial z_i^{(m)}} \cdot \frac{\partial z_i^{(m)}}{\partial W_{ij}^{(m)}} = \frac{\partial C}{\partial z_i^{(m)}} \cdot a_j^{(m-1)}, \qquad i = 1, \ldots, D^{(m)}, j = 1, \ldots, D^{(m-1)}$$

Vectorized:

Solution: Vectorizing the Gradients for the Weights

Non-vectorized:

$$\frac{\partial C}{\partial W_{ij}^{(m)}} = \frac{\partial C}{\partial z_i^{(m)}} \cdot \frac{\partial z_i^{(m)}}{\partial W_{ij}^{(m)}} = \frac{\partial C}{\partial z_i^{(m)}} \cdot a_j^{(m-1)}, \qquad i = 1, \dots, D^{(m)}, j = 1, \dots, D^{(m-1)}$$

Vectorized:

$$\nabla_{\mathbf{W}^{(m)}} C = \left(\nabla_{\mathbf{z}^{(m)}} C\right)\left(\mathbf{a}^{(m-1)}\right)^{\top}$$

UNIVERSITY OF TORONTO

Backward Pass Computations

1. Compute gradients for output layer

$$\nabla_{\mathbf{z}^{(L)}} \mathrm{C} = \left(\nabla_{\mathbf{a}^{(L)}} \mathrm{C}\right) \odot \left(\sigma^{(L)\prime}\left(\mathbf{z}^{(L)}\right)\right)$$

2. Compute gradients for each hidden layer recursively

$$\nabla_{\mathbf{z}^{m}} \mathrm{C} = \left(\left(\mathbf{W}^{(\mathrm{m}+1)}\right)^{\mathrm{T}} \left(\nabla_{\mathbf{z}^{(\mathrm{m}+1)}} \mathrm{C}\right)\right) \odot \left(\sigma^{(m)\prime}\left(\mathbf{z}^{(\mathrm{m})}\right)\right)$$

3. Compute gradients for the weights

$$\nabla_{\mathbf{W}^{(m)}} \mathrm{C} = \left(\nabla_{\mathbf{z}^{(\mathrm{m})}} \mathrm{C}\right)\left(\mathbf{a}^{(m-1)}\right)^{\mathrm{T}}$$