**Deep Learning** Winter 2025–2026

# Homework 1 — Backpropagation

*Naaman Kopty* naamankoptyta@gmail.com

**Due:** 03.12.2025

## Questions 1–2: (70 Points)

Programming assignment — See attached notebook in Moodle.

## Question 1: (15 Points)

Let the classifier logits be $z \in \mathbb{R}^C$. The softmax outputs are

$$p_i = \frac{e^{z_i}}{\sum_{j=1}^{C} e^{z_j}}, \qquad i = 1, \dots, C. \tag{1}$$

For a one-hot label vector $y \in \{0,1\}^C$ with $\sum_i y_i = 1$, the cross-entropy loss is

$$L = -\sum_{i=1}^{C} y_i \log p_i. \tag{2}$$

(a) Derive the gradient $\frac{\partial L}{\partial z_i}$ for each $i \in \{1, \dots, C\}$. Write your result in simplest closed form.

(b) Show why this derivative simplifies to: $\frac{\partial L}{\partial z_i} = p_i - y_i$.

(c) Using your result in (a), explain succinctly why the softmax+cross-entropy combination provides strong learning signals (non-vanishing gradients) even when the model is confidently wrong.

*Hint:* You may find it helpful to use the softmax Jacobian identity

$$\frac{\partial p_k}{\partial z_i} = \begin{cases} p_i(1-p_i), & i = k, \\ -\,p_i\,p_k, & i \neq k. \end{cases} \tag{3}$$

## Question 2: (15 Points)

Consider a mini-batch $\{z_1, \dots, z_m\} \subset \mathbb{R}$. Batch Normalization (BN) computes

$$\mu = \frac{1}{m}\sum_{i=1}^{m} z_i, \qquad \sigma^2 = \frac{1}{m}\sum_{i=1}^{m}(z_i - \mu)^2, \tag{4}$$

then normalizes and applies an affine transform

$$\hat{z}_i = \frac{z_i - \mu}{\sqrt{\sigma^2 + \epsilon}}, \qquad y_i = \gamma \hat{z}_i + \beta, \tag{5}$$

where $\gamma, \beta \in \mathbb{R}$ are learnable parameters and $\epsilon > 0$ is a small constant.

Let $dy_i := \frac{\partial L}{\partial y_i}$.

(a) Derive the gradients with respect to the affine parameters: $\frac{\partial L}{\partial \gamma}$ and $\frac{\partial L}{\partial \beta}$.

(b) Express $\frac{\partial L}{\partial z_i}$ in terms of $dy_j$, $\gamma$, $\hat{z}_j$, $\mu$, $\sigma^2$, and $m$.

(c) Briefly explain (2–3 sentences) how BN can improve gradient flow and mitigate vanishing/exploding gradients in deep networks.

*Note:* You may assume the batch statistics $(\mu, \sigma^2)$ are computed over the same set used in the backward pass.

Good Luck!