

# Problem Set 3 – Loss Functions and Fitting Models

DS542 – DL4DS

Spring, 2025

**Note:** Refer to the equations in the *Understanding Deep Learning* textbook to solve the following problems.

## Problem 5.9

Consider a multivariate regression problem in which we predict the height of an individual in meters and their weight in kilos from some data  $x$ . Here, the units take quite different values. What problems do you see this causing? Propose two solutions to these problems.

Scale Imbalance: The magnitude of weight is much larger than height. This can lead to the model giving disproportionate importance to weight when minimizing the loss function. This is solved with either: standardization of the values to a normal scale (feature scaling) or based on their respective min-max values (min max normalization). Can also use separate models for height and weight, then assign different weights to their respective loss terms.

## Problem 6.6

Which of the functions in Figure 6.11 from the book is convex? Justify your answer. Characterize each of the points 1–7 as (i) a local minimum, (ii) the global minimum, or (iii) neither.

(a): The loss function has multiple local minima and maxima, indicating non-convexity. (b): The function appears to be a monotonically decreasing curve with a single minimum, making it convex. (c): The function has multiple inflection points and local minima, suggesting non-convexity.

(a)

1. Point 1: Neither (it is a local maximum).
2. Point 2: Local minimum (it's a dip in the loss function).
3. Point 3: Neither (it is a local maximum).

(b)

1. Point 4: Neither (it is not a local minimum or maximum).
2. Point 5: Global minimum (it is the lowest point in the convex function).

(c)

1. Point 6: Local minimum (it is a dip in the function).
2. Point 7: Neither (it is not a local minimum).

## Problem 6.10

Show that the momentum term  $m_t$  (equation (6.11)) is an infinite weighted sum of the gradients at the previous iterations and derive an expression for the coefficients (weights) of that sum.

$$m_{t+1} = \beta m_t + (1 - \beta) \sum_{i \in B_t} \frac{\partial \ell_i[\phi_t]}{\partial \phi} \quad (1)$$

$$m_t = \beta m_{t-1} + (1 - \beta) \sum_{i \in B_{t-1}} \frac{\partial \ell_i[\phi_{t-1}]}{\partial \phi} \quad (2)$$

Substituting  $m_{t-1}$ :

$$m_t = \beta(\beta m_{t-2} + (1 - \beta) \sum_{i \in B_{t-2}} \frac{\partial \ell_i[\phi_{t-2}]}{\partial \phi}) + (1 - \beta) \sum_{i \in B_{t-1}} \frac{\partial \ell_i[\phi_{t-1}]}{\partial \phi} \quad (3)$$

$$m_t = \beta^2 m_{t-2} + (1 - \beta) \sum_{i \in B_{t-2}} \beta \frac{\partial \ell_i[\phi_{t-2}]}{\partial \phi} + (1 - \beta) \sum_{i \in B_{t-1}} \frac{\partial \ell_i[\phi_{t-1}]}{\partial \phi} \quad (4)$$

$$m_t = (1 - \beta) \sum_{k=0}^t \beta^k \sum_{i \in B_{t-k}} \frac{\partial \ell_i[\phi_{t-k}]}{\partial \phi} \quad (5)$$

As  $t \rightarrow \infty$ , we get an infinite weighted sum:

$$m_t = (1 - \beta) \sum_{k=0}^{\infty} \beta^k \sum_{i \in B_{t-k}} \frac{\partial \ell_i[\phi_{t-k}]}{\partial \phi} \quad (6)$$

The weight of each past gradient at iteration  $t - k$  is given by:

$$(1 - \beta) \beta^k \quad (7)$$

which shows that the contribution of past gradients decays exponentially as we move back in time.