**HW to Chapter 8 “Initialization and Training Sets”**

**Non-programming Assignment**

1. To which values initialize parameters (W, b) in a neural network and why?

Parameters W (weights) and b (biases) are typically initialized to small random values. Specifically, weights are often initialized using schemes like **Xavier initialization** or **He initialization** to avoid issues with exploding or vanishing gradients. Biases are commonly initialized to zero. Proper initialization ensures that the network can efficiently learn by maintaining the variance of the activations throughout the layers and preventing gradients from becoming too large or too small during backpropagation​.

1. Describe the problem of exploding and vanishing gradients?

**Exploding gradients** occur when gradients grow exponentially as they are propagated backward through deep layers, leading to instability and making training difficult. **Vanishing gradients** happen when gradients shrink as they move backward, making weight updates too small and causing slow or stalled training. Both problems hinder the training of deep networks, especially when using inappropriate initialization methods​​.

1. What is Xavier initialization?

**Xavier initialization** (also known as Glorot initialization) sets weights such that the variance of the activations is the same across every layer. It works by initializing the weights from a normal distribution with a mean of 0 and a variance of 1/nin​, where nin​ is the number of input units to the layer. This helps to prevent vanishing or exploding gradients by ensuring stable activations and gradients across layers​.

1. Describe training, validation, and testing data sets and explain their role and why all they are needed.

The **training set** is used to fit the model and adjust parameters. The **validation set** is used to fine-tune the model’s hyperparameters and assess performance during training. The **test set** is reserved for the final evaluation of the model's performance on unseen data. Each set plays a critical role in ensuring the model generalizes well to new data and doesn’t overfit the training data

1. What is training epoch?

An **epoch** is one complete cycle through the entire training dataset. During each epoch, the model processes all training samples once. Multiple epochs are typically needed to achieve convergence​.

1. How to distribute training, validation, and testing sets?

Common distributions are 70-80% of the data for training, 10-15% for validation, and 10-15% for testing. In cases of larger datasets, the training set might take a larger proportion, while smaller datasets require careful balancing of these ratios.

1. What is data augmentation and why may it needed?

**Data augmentation** involves artificially creating new training data by applying transformations like flipping, rotating, or scaling to the original data. This increases the diversity of the dataset, helping to improve generalization and reduce overfitting. It’s particularly useful when there’s a shortage of training data.