Homework-4

Question 1

The Adam method uses a recursive method for computing running averages:

$$\overline{X}_0 = 0, \quad \overline{X_t} = \beta \overline{X}_{t-1} + (1-\beta) X_t, \quad \hat{X}_t = \frac{\overline{X}_t}{1-\beta^t}$$

1. Show that if $\beta = 0$ then $\hat{X}_t = X_t$ for all t.

2. Show that if $\beta \to 1$ then $\hat{X}_t \to \frac{1}{t}\sum_{i=1}^{t} X_i$ for all t.

Hint: Use the explicit formula for $\overline{X_t}$, form the limit, and solve it using, for example, L'Hopital's rule.

Question 2

The Adam technique for accelerating back propagation was specified in terms of the following parameters: λ, α, ϵ, β_1, β_2. The standard back propagation algorithm was specified in terms of the learning-rate parameter ϵ. Show how to select the parameters of Adam so that the result is as close approximation to standard back propagation as possible.

Question 3

Consider the following techniques that may be used in a feed-forward neural network training/testing model.

A. l_2 regularization with the regularization parameter λ.

B. Dropouts controlled by the drop-probability parameter p.

C. Number of nodes in a layer.

D. Number of layers.

E. Stochastic steepest descent optimizer with the learning rate parameter ϵ.

F. Adam optimizer with the parameters α, ϵ, β_1, β_2.

You are asked to determine how these techniques affect the training and the quality of the learned model. In all cases your answer should be based on the theory and **not** on results you may have observed in your experiments. In each case your selection should be **the most appropriate** among the given choices.

1. The running time of a single training iteration (one random batch).

A1. Using l_2 regularization would: increase / descrease / no-effect / impossible-to-tell .

A2. Increasing λ would: increase / descrease / no-effect / impossible-to-tell .

B1. Using dropouts would: increase / descrease / no-effect / impossible-to-tell .

B2. Increasing p would: increase / descrease / no-effect / impossible-to-tell .

C. Increasing number of layer nodes would: increase / descrease / no-effect / impossible-to-tell .

D. Increasing number of layers would: increase / descrease / no-effect / impossible-to-tell .

E. Increasing ϵ would: increase / descrease / no-effect / impossible-to-tell .

F. Increasing α would: increase / descrease / no-effect / impossible-to-tell .

2. The running time of the learned model on a single testing example.

A1. Using l_2 regularization would: increase / descrease / no-effect / impossible-to-tell .

A2. Increasing λ would: increase / descrease / no-effect / impossible-to-tell .

B1. Using dropouts would: increase / descrease / no-effect / impossible-to-tell .

B2. Increasing p would: increase / descrease / no-effect / impossible-to-tell .

C. Increasing number of layer nodes would: increase / descrease / no-effect / impossible-to-tell .

D. Increasing number of layers would: increase / descrease / no-effect / impossible-to-tell .

E. Increasing ϵ would: increase / descrease / no-effect / impossible-to-tell .

F. Increasing α would: increase / descrease / no-effect / impossible-to-tell .

3. The accuracy of the learned model on the **training** data.

A1. Using l_2 regularization would: increase / descrease / no-effect / impossible-to-tell .

A2. Increasing λ would: increase / descrease / no-effect / impossible-to-tell .

B1. Using dropouts would: increase / descrease / no-effect / impossible-to-tell .

B2. Increasing p would: increase / descrease / no-effect / impossible-to-tell .

C. Increasing number of layer nodes would: increase / descrease / no-effect / impossible-to-tell .

D. Increasing number of layers would: increase / descrease / no-effect / impossible-to-tell .

E. Increasing ϵ would: increase / descrease / no-effect / impossible-to-tell .

F. Increasing α would: increase / descrease / no-effect / impossible-to-tell .

4. Speed of convergence on the **training** data.

A1. Using l_2 regularization would: increase / descrease / no-effect / impossible-to-tell .

A2. Increasing λ would: increase / descrease / no-effect / impossible-to-tell .

B1. Using dropouts would: increase / descrease / no-effect / impossible-to-tell .

B2. Increasing p would: increase / descrease / no-effect / impossible-to-tell .

C. Increasing number of layer nodes would: increase / descrease / no-effect / impossible-to-tell .

D. Increasing number of layers would: increase / descrease / no-effect / impossible-to-tell .

E. Increasing ϵ would: increase / descrease / no-effect / impossible-to-tell .

F. Increasing α would: increase / descrease / no-effect / impossible-to-tell .

5. The accuracy of the learned model on **testing** data.

A1. Using l_2 regularization would: increase / descrease / no-effect / impossible-to-tell .

A2. Increasing λ would: increase / descrease / no-effect / impossible-to-tell .

B1. Using dropouts would: increase / descrease / no-effect / impossible-to-tell .

B2. Increasing p would: increase / descrease / no-effect / impossible-to-tell .

C. Increasing number of layer nodes would: increase / descrease / no-effect / impossible-to-tell .

D. Increasing number of layers would: increase / descrease / no-effect / impossible-to-tell .

E. Increasing ϵ would: increase / descrease / no-effect / impossible-to-tell .

F. Increasing α would: increase / descrease / no-effect / impossible-to-tell .