STAT3007/7007 Deep Learning, Tutorial 8

2022 Semester 2

1. (Basic concepts of numerical optimization)

(a) What is a convex function? Is $f(x) = \sin(x)$ a convex function?

(b) Some special points/regions in the surface of a nonconvex function can cause significant difficulty for a gradient-based optimizer. Name two, and explain why they cause difficulty for a gradient-based optimizer.

2. (Initialization) In this question, we take a closer look at a random initialization strategy.

(a) Assume that the inputs $X_1, \ldots, X_d$ are all independently sampled from $N(0, 1)$, and all the parameters are independently sampled from $N(0, 1)$. We consider a tanh unit with weights $W_1, \ldots, W_d$ and computes $\tanh(\sum_i W_i X_i)$ (we ignore the bias term here). Show that $\mathbb{E}(\sum_i W_i X_i) = 0$ and $\mathrm{Var}(\sum_i W_i X_i) = d$.

(b) How does the distribution of the output of the tanh unit in (a) change as d increases? What is a typical value (or value range) of the output? How does the output distribution affect training the unit using gradient descent? Provide a justification (no formal proof is needed).