CSE516/ECE559: Theories of Deep Learning
Problem Sheet 3

Dr. Vinayak Abrol

October 22, 2024

Questions entail computational experiments can be attempted in programming language of your choice with the link to your code and results.

1. **Trainability**

   (a) Train a 100 layer FC network with ReLU and ELU activation on MNIST/FMNIST where network weights and bias are initialized on EOC.
   ReLU: $\sigma_b^2, \sigma_w^2 = 0, 2$, normalize inputs to have $q^* = 0.5$.
   ELU: numerically compute σ_b^2, σ_w^2 corresponding to $q^* = 0.5$, and normalize inputs to $q^* = 0.5$ before training.
   Now repeat and compare ELU network with initialization corresponding to $q^* = 0.2$

2. **Adversarial attacks for neural networks**
   Adversarial examples are intentionally designed illusions, where such inputs to learned models cause the model to make a mistake. Mathematically, given a point $\mathbf{x} \in \Omega$ drawn from class y, a scalar $\epsilon > 0$, and a metric d, we say that $\mathbf{x}$ admits an adversarial example in the metric d if there exists a point $\mathbf{x}^* \in \Omega$ with $Class(\mathbf{x}^*) \neq y$, and $d(\mathbf{x}, \mathbf{x}^*) \leq \epsilon$. In practice d is chosen as ℓ^p-norms with ℓ^∞ being the most popular choice, which limits the absolute change that can be made to any one dimension of $\mathbf{x}$.

   (a) Task1: One Layer Net: Consider the neural net defined as $\hat{y} = SM(\mathbf{W}\mathbf{x})$ trained with the cross-entropy loss $L(\mathbf{x}, y)$, where SM denotes softmax activation. Let $\mathbf{x}^*$ be the adversarial image of x resulting from fast gradient sign method (FGSM) attacks[1] with constant ϵ. Prove that $\forall \epsilon > 0$ we have $L(\mathbf{x}^*, y) \geq L(\mathbf{x}, y)$

   (b) Task2: Two Layer Net: Consider the neural net defined as $\hat{y} = SM(\mathbf{V}\sigma\mathbf{W}\mathbf{x})$ trained with the cross-entropy loss $L(\mathbf{x}, y)$, where $\mathbf{V}, \mathbf{W}$ are weights, SM denotes softmax activation and σ is ReLU activation. Suppose every element of $\mathbf{W}\mathbf{x}$ is non-zero, if $\epsilon < \frac{|\mathbf{W}\mathbf{x}|_{min}}{\|\mathbf{W}\|_\infty}$, then prove that $L(\mathbf{x}^*, y) \geq L(\mathbf{x}, y)$,
   given the fact that for $j = 1, 2, \ldots; sign(\mathbf{W}\mathbf{x})_j = sign(\mathbf{W}\mathbf{x}^*)_j$

[1] https://arxiv.org/pdf/1412.6572.pdf