Advanced Studies In Mathematics Exercise

Hwijae Son

May 21, 2024

1. (Python) Consider the univariate function

$$f(x) = (x-2)\cos(4x).$$

Let $f_\theta(x)$ be a 3-layer MLP with the sigmoid activation function $\sigma(x) = \frac{1}{1+e^{-x}}$. Given data x_i generated as i.i.d. standard normal distribution and corresponding labels $y_i = f(x_i)$ for $i = 1, \ldots, 1000$, define the loss function

$$\mathcal{L}(\theta) = \frac{1}{2000}\sum_{i=1}^{1000}(f_\theta(x_i) - y_i)^2.$$

Use PyTorch to train f_θ with SGD. Use layer width $(1, 64, 64, 1)$, total epochs $K = 1000$, stepsize $\eta = 0.1$, batch size $B = 128$. Plot the final trained function and the true function.

2. In the previous problem, how many trainable parameters are in the 3-layer MLP?

3. Repeat the problem 1 with noisy label $y_i = f(x_i) + \epsilon$, where $\epsilon \sim \mathcal{N}(0, 0.5)$.

4. Prove the followings.

(a) The ReLU activation $\sigma(x) = \max\{0, x\}$ is idempotent, i.e.,

$$\sigma(\sigma(x)) = \sigma(x), \forall x \in \mathbb{R}.$$

(b) The *softplus* function $\sigma(x) = \log(1+e^x)$ is considered a smooth alternative of ReLU. Show that the *softplus* has Lipschitz continuous derivatives while ReLU does not.

(c) Let $\sigma(x) = \frac{1}{1+e^{-x}}$ be the sigmoid function and let $\rho(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}}$ be the hyperbolic tangent function. Show that the two activation functions are equivalent in the sense that MLPs built with them are equivalent: given $L > 1$, $A_1, \ldots A_L, b_1, \ldots b_L$ there are $C_1, \ldots C_L, d_1, \ldots d_L$ such that a tanh network with weights C_i, biases d_i represents identical mapping with the sigmoid network with weights A_i, biases b_i and vice versa.

5. Consider the 2-layer neural network

$$f_\theta(x) = v^T\sigma(wx + b) = \sum_{j=1}^{p} v_j \sigma(w_j x + b_j),$$

where $x \in \mathbb{R}, w, b, v \in \mathbb{R}^p$, and $\sigma(x) = \max\{0, x\}$. We train the network with the standard MSE loss function with SGD and data $\{(x_i, y_i)\}_{i=1}^N$. Assume the $j-$th ReLU output is "dead" at initialization in the sense that $w_j^0 x_i + b_j^0 < 0$ for all $i = 1, 2, \ldots, N$. Show that $j-$th ReLU output remains dead throughout the training.

6. The leaky ReLU activation is defined as

$$\sigma(x) = \begin{cases} x & \text{for } x \geq 0, \\ \alpha x & \text{otherwise,} \end{cases}$$

where α is a fixed parameter often set to $\alpha = 0.01$. Show that if leaky ReLU is used in the previous problem, the ReLU node no longer be dead.