MATH 6490. Nonlinear Optimization in Machine Learning.
Assignment 1.

There are 3 problems, each problem is worth 5 points, total is 15 points.

1. Consider the following one–hidden layer Neural Network with $2k$ hidden units. The network parameters are $W \in \mathbb{R}^{2k\times d}$ and $\boldsymbol{v} \in \mathbb{R}^{2k}$, which we denote jointly by $\mathcal{W} = (W, \boldsymbol{v})$. The network output is given by the function $g_{\mathcal{W}} : \mathbb{R}^d \to \mathbb{R}$ defined as

$$g_{\mathcal{W}}(\boldsymbol{x}) = \boldsymbol{v}^T\sigma(W\boldsymbol{x}) \ , \ \boldsymbol{x} \in \mathbb{R}^d \ ,$$

where σ is the ReLU activation function applied element–wise, such that element–wise $\sigma(z) = \max(z, 0)$.

Consider a set of binary classification training data $S = \{(\boldsymbol{x}_1, y_1), ..., (\boldsymbol{x}_n, y_n)\}$ where $\boldsymbol{x}_i \in \mathbb{R}^d$ and $y_i = \pm 1$. We define the empirical loss over S to be the mean hinge–loss

$$L_S(\mathcal{W}) = \frac{1}{n}\sum_{i=1}^{n} \max(1 - y_i g_{\mathcal{W}}(\boldsymbol{x}_i), 0) \ .$$

Let $n = 1$, $k = 1$ and $\boldsymbol{v} = \begin{pmatrix} 1 \\ -1 \end{pmatrix}$, show that the network output is given by the function

$$g_{\mathcal{W}}(\boldsymbol{x}) = \sigma(\langle \boldsymbol{w}, \boldsymbol{x}\rangle) - \sigma(\langle \boldsymbol{u}, \boldsymbol{x}\rangle)$$

for $\boldsymbol{w}, \boldsymbol{u} \in \mathbb{R}^d$. Suppose $y_1 = -1$, then show that the loss function takes the form

$$L_S(\boldsymbol{w}, \boldsymbol{u}) = \max(1 + (\sigma(\langle \boldsymbol{w}, \boldsymbol{x}\rangle) - \sigma(\langle \boldsymbol{u}, \boldsymbol{x}\rangle)), 0) \ .$$

2. Continuing the example in problem 1, set $\boldsymbol{w}_1 = \boldsymbol{w}_2 = \boldsymbol{u}_1 = \boldsymbol{x}$ and $\boldsymbol{u}_2 = -\boldsymbol{x}$. Set $0 < \|\boldsymbol{x}\|^2 < 1$. Show that

$$L_S\left(\frac{\boldsymbol{w}_1 + \boldsymbol{w}_2}{2}, \frac{\boldsymbol{u}_1 + \boldsymbol{u}_2}{2}\right) > \frac{1}{2}\left(L_S(\boldsymbol{w}_1, \boldsymbol{u}_1) + L_S(\boldsymbol{w}_2, \boldsymbol{u}_2)\right) \ .$$

Thus the loss function $L_S(W)$ in this case is not convex.

3. Show that if f is continuously differentiable and convex in the sense that for any $x, y \in \mathrm{dom}(f)$ and all $\alpha \in [0, 1]$ we have

$$f((1-\alpha)x + \alpha y) \le (1-\alpha)f(x) + \alpha f(y) \ ,$$

then for any $x, y \in \mathrm{dom}(f)$ we have

$$f(y) \ge f(x) + (\nabla f(x))^T(y - x) \ .$$

(Hint: First consider for some small $\alpha > 0$ that $z_\alpha = (1-\alpha)x + \alpha y$ and work out the Taylor expansion $f(z_\alpha) = f(x) + (\nabla f(x))^T(z_\alpha - x) + O(|z_\alpha - x|^2)$. Make use of convexity, we obtain that $f(z_\alpha) \leq (1-\alpha)f(x) + \alpha f(y)$, so that $\alpha f(y) - \alpha f(x) \geq (\nabla f(x))^T(z_\alpha - x) + O(|z_\alpha - x|^2)$. Divide by α on both sides we obtain that $f(y) \geq f(x) + (\nabla f(x))^T \dfrac{z_\alpha - x}{\alpha} + O\left(\dfrac{|z_\alpha - x|^2}{\alpha}\right)$. Show that $\dfrac{z_\alpha - x}{\alpha} = y - x$.)