CSE 599 Theoretical Deep Learning Homework 2

February 21, 2021

- Deadline: Mar. 1st. No late homework.
- Include your name and UW NetID in your submission.
- Homework must be typed. You can use any typesetting software you wish (latex, markdown, ms word, etc).
- You may discuss assignments with others, but you must write down the solutions by yourself.

1 Implicit regularization of gradient descent on over-parameterized linear regression (6 points)

Consider a linear regression problem:

$$\min_{w \in \mathbb{R}^d} L(w) = \frac{1}{2n} \sum_{i=1}^{n} \left(x_i^\top w - y_i \right)^2$$

where $x_i \in \mathbb{R}^d$ is the input and $y_i \in \mathbb{R}$ is the label. We assume $d \geq n$ (the over-parameterized regime). Let $X = [x_1, \ldots, x_n] \in \mathbb{R}^{d \times n}$ and assume $\operatorname{rank}(X) = n$ (the least eigenvalue of X is strictly positive). We solve this linear regression problem via gradient flow with $w_0 = 0$,

$$\frac{dw_t}{dt} = -\nabla L(w_t).$$

We will show w_t converges to the solution of the following optimization problem

$$\begin{aligned} &\min \|w\|_2^2 \\ &\text{such that } y_i = x_i^\top w, \ \forall i = 1, \ldots, n. \end{aligned} \tag{1}$$

1. (2 points) Show $L(w_t) \to 0$ as $t \to \infty$.
2. (2 points) Show w is always in the span of $(x_1, \ldots, x_n)$.
3. (2 points) Use these two properties to argue w_t will converge to the solution of the optimization problem (1).

2 ReLU networks are piecewise linear in their input (9 points)

Consider a H-layer neural network with ReLU activation function

$$f(x,w) = W_{H+1}\sigma\left(W_H\sigma\left(\cdots W_2\sigma(W_1x)\right)\right)$$

where $w = (W_1, \ldots, W_H, W_{H+1})$ and $\sigma(\cdot)$ is ReLU activation function. $W_1 \in \mathbb{R}^{m\times d}$, $W_h \in \mathbb{R}^{m\times m}$ for $h = 1, \ldots, H$, and $W_{H+1} \in \mathbb{R}^m$. Given an input x, the per-layer outputs can be defined recursively as

$$\begin{aligned}
x_0(x) =& x\\
x_h(x) =& \sigma\left(W_hx_{h-1}(x)\right), \text{ for } h = 1, \ldots, H\\
x_{H+1}(x) =& W_{H+1}x_H(x).
\end{aligned}$$

We also define the activation vectors and matrices

$$\begin{aligned}
a_1(x) &= \mathbf{1}\left[W_1x_0(x) \geq 0\right],\\
a_h(x) &= \mathbf{1}\left[W_hx_{h-1}(x) \geq 0\right], \text{ for } h = 1, \ldots, H\\
A_h(x) &= \mathsf{diag}\left(a_h(x)\right)
\end{aligned}$$

where $\mathbf{1}[\cdot]$ is the indicator function and $\mathsf{diag}(\cdot)$ transforms a vector to diagonal matrix of the appropriate dimension. Note we have $x_h(x) = A_h(x)W_hx_{h-1}(x)$ for $h = 1, \ldots, H$.

1. (4 points) Fix activation patterns $\mathbf{a}' = (a_1', \ldots, a_H') \in \{1,0\}^{m\times H}$, and consider those inputs $x \in \mathbb{R}^d$ with these activations:

$$S_{\mathbf{a}'} \triangleq \left\{x \in \mathbb{R}^d : a_h(x) = a_h', h \in \{1, \ldots, H\}\right\}$$

   Prove that restricted to $S_{\mathbf{a}'}$, f is a linear function.

2. (5 points) Prove that $\mathbb{R}^d$ can be partitioned into finitely many regions (number of regions can be exponential in m, d and H) such that f is a (potentially different) linear function over each region.

3 A nice property of positive homogeneity (10 points)

Suppose $f : \mathbb{R}^d \to R$ is locally Lipschitz and positively homogeneity of degree L. We will prove that for any given $x \in \mathbb{R}^d$, for $s \in \partial f(x)$, we have $\langle s, x\rangle = Lf(x)$. Here $\partial f(x)$ is Clarke Differential.

1. (2 Points) Show that when $x = 0$, and $s \in \partial f(x)$, we have $\langle s, x\rangle = Lf(x)$.

2. (5 Points) Show for all $x \neq 0$ such that $\nabla f(x)$ exists, $\langle \nabla f(x), x\rangle = Lf(x)$.
   Hint: You can use the following basic property about gradient:

$$\lim_{\delta\to 0}\frac{f(x+\delta x) - f(x) - \langle\nabla f(x), \delta x\rangle}{\delta} = 0.$$

3. (3 Points) Using the definition of Clarke Differential to show that for any given $x \in \mathbb{R}^d$, for $s \in \partial f(x)$, we have $\langle s, x\rangle = Lf(x)$.