CS 839: Theoretical Foundations of Deep Learning **Spring 2023**

Lecture 10 Implicit Regularization for Neural Networks

Instructor: Yingyu Liang *Date:* *Scriber: Matthew Zurek*

1 Overview

Continuing the theme of implicit regularization from the previous four lectures, we use the tools developed in the last lecture to show that a non-smooth version of gradient flow (using the Clarke subdifferential) yields non-decreasing "soft" margin for homogeneous predictors on separable data. In particular this applies to neural networks since ReLU networks are homogeneous. To perform this analysis, we first prove a useful lemma about Clarke subdifferentials of homogeneous functions and generalize margin beyond linear classifiers.

2 Review

First we recall the definition of the Clarke subdifferential from last lecture.

Definition 1. For a locally-Lipschitz function $f : \mathcal{X} \to \mathbb{R}$, the Clarke subdifferential of f at $w \in \mathcal{X}$ is

$$\partial f(w) = \operatorname{conv}\{s : \exists (w_n)_n \text{ such that } w_n \to w, \nabla f(w_n) \to s\}.$$

3 Subdifferential for Homogeneous Functions

Motivated by the observation last lecture that an L-hidden layer ReLU neural network is L-homogeneous, we prove the following lemma.

Lemma 2. If $f : \mathbb{R}^d \to \mathbb{R}$ is locally-Lipschitz and L-homogeneous, then $\forall w \in \mathbb{R}^d$ and $\forall s \in \partial f(w)$, we have

$$\langle s, w \rangle = Lf(w).$$

Proof. First, if $w = 0$ then this is trivial, since $f(0) = 0^L f(0) = 0$ by L-positive homogeneity. Now we handle $w \neq 0$. Let $D = \{w : f \text{ is differentiable at } w\}$. (This is almost everywhere by local-Lipschitzness and Radamacher's theorem.) If $w \in D \setminus \{0\}$, then

$$\begin{aligned}
0 &= \lim_{\delta \downarrow 0} \frac{f(w + \delta w) - f(w) - \langle \nabla f(w), \delta w \rangle}{\delta \|w\|} \\
&= \lim_{\delta \downarrow 0} \frac{\left((1+\delta)^L - 1\right) f(w)}{\delta \|w\|} - \frac{\langle \nabla f(w), w \rangle}{\|w\|} \\
&= \frac{Lf(w)}{\|w\|} - \frac{\langle \nabla f(w), w \rangle}{\|w\|}
\end{aligned}$$

where we used L-positive homogeneity and the fact that $\lim_{\delta \downarrow 0} \frac{\left((1+\delta)^L - 1\right)}{\delta}$ is the (right) derivative of $z \mapsto z^L$ at 1. Now we can rearrange to conclude that $\langle \nabla f(w), w \rangle = Lf(w)$. This concludes this case, because since f is differentiable at w, $\partial f(w) = \{\nabla f(w)\}$.

Now we handle the case that $w \notin D \setminus \{0\}$ in two steps. Let $s \in \partial f(w)$ be such that there exists a sequence $(w_n) \to w$ such that $\nabla f(w_n) \to s$ (not all $s \in \partial f(w)$ are of this form so this is only the first step). Then since all $(w_n)_n$ are contained in D, for each n it holds from previous cases that $Lf(w_n) - \langle \nabla f(w_n), w_n \rangle = 0$. Then by continuity of f and the inner product, as well as the fact that $\nabla f(w_n) \to s$, we may take the limit to conclude that $Lf(w) - \langle s, w \rangle = 0$ as desired. Finally, all $s \in \partial f(w)$ are by definition convex combinations of vectors $s_1, \dots, s_k$ which are handled by the previous step, and thus writing $s = \sum_{i=1}^k \alpha_i s_i$ where $\sum_{i=1}^k \alpha_i = 1$, using the result from the previous step we have that

$$\begin{aligned} \langle s, w \rangle &= \sum_{i=1}^k \alpha_i \langle s_i, w \rangle \\ &= \sum_{i=1}^k \alpha_i L f(w) \\ &= Lf(w). \end{aligned}$$

□

4 Margin of Homogeneous Predictors

Now we move towards our main result on implicit regularization for neural networks. We will show for L-homogeneous predictors that a "soft" version of the margin is non-decreasing along the (non-smooth analogue of) gradient flow. Before we can do so, we first generalize the margin beyond linear classifiers.

Definition 3. For an L-homogeneous predictor $f(\cdot; w)$ we define the margin on a single point (x_i, y_i) as

$$m_i(w) = y_i f(x_i; w).$$

The (overall) margin of $f(\cdot; w)$ is

$$\gamma(w) = \min_i m_i \left(\frac{w}{\|w\|} \right) = \min_i \frac{m_i(w)}{\|w\|^L}.$$

Note that if f is a linear predictor then we recover the same definition as we have seen before. The margin of the maximum-margin predictor is

$$\overline{\gamma} = \max_{w : \|w\| = 1} \gamma(w).$$

Instead of analyzing this "hard" version of margin, we will analyze the soft margin. For a (non-averaged) loss $\mathcal{L}(w) = \sum_{i=1}^n \ell(y_i f(x_i; w))$ where $\ell(\cdot)$ is monotonic, we define the soft margin as

$$\widetilde{\gamma}(w) = \frac{\ell^{-1}(\mathcal{L}(w))}{\|w\|^L}.$$

In the sequel we will focus on the exponential loss $\ell(z) = \exp(-z)$. In this case the soft margin becomes

$$\widetilde{\gamma}(w) = \frac{-\ln \sum_{i=1}^{n} \exp(-y_i f(x_i; w))}{\|w\|^L} = \frac{-\ln \sum_{i=1}^{n} \exp(-m_i(w))}{\|w\|^L}.$$

Our separability assumption on the dataset will be that there exists w such that $\widetilde{\gamma}(w) > 0$.

5 Main Result

We will analyze the flow given by the differential inclusion equation $\dot{w}(t) \in -\partial \ln \sum_{i=1}^{n} \exp(-m_i(w(t)))$, but first we prove a final useful lemma.

Lemma 4. For all $w \in \mathbb{R}^d$, if $v \in -\partial \ln \sum_{i=1}^{n} \exp(-m_i(w))$ and if the chain rule holds, then

$$-L \ln \sum_{i=1}^{n} \exp(-m_i(w)) \leq \langle v, w \rangle.$$

Proof. Fix such a v. Then by the chain rule, for each $i = 1, \ldots, n$ there exists $v_i \in \partial m_i(w)$ such that

$$v = \sum_{i=1}^{n} \frac{\exp(-m_i(w)) v_i}{\sum_{j=1}^{n} \exp(-m_j(w))}.$$

Then we can calculate

$$\begin{aligned}
\langle v, w \rangle &= \sum_{i=1}^{n} \frac{\exp(-m_i(w))}{\sum_{j=1}^{n} \exp(-m_j(w))} \langle v_i, w \rangle \\
&= \sum_{i=1}^{n} \frac{\exp(-m_i(w))}{\sum_{j=1}^{n} \exp(-m_j(w))} L m_i(w) \\
&= \sum_{i=1}^{n} \frac{\exp(-m_i(w))}{\sum_{j=1}^{n} \exp(-m_j(w))} (-L \ln(\exp(-m_i(w)))) \\
&\geq \sum_{i=1}^{n} \frac{\exp(-m_i(w))}{\sum_{j=1}^{n} \exp(-m_j(w))} \left(-L \ln \left(\sum_{k=1}^{n} \exp(-m_k(w)) \right) \right) \\
&= -L \ln \left(\sum_{k=1}^{n} \exp(-m_k(w)) \right) \sum_{i=1}^{n} \frac{\exp(-m_i(w))}{\sum_{j=1}^{n} \exp(-m_j(w))} \\
&= -L \ln \left(\sum_{k=1}^{n} \exp(-m_k(w)) \right)
\end{aligned}$$

where the second step made use of lemma 2 and the inequality used the fact that $-\ln$ is monotonically decreasing. □

Theorem 5. For the flow with $w(0) = 0$, $\dot{w}(t) \in -\partial \ln \sum_{i=1}^{n} \exp(-m_i(w(t)))$, assuming that the chain rule holds for almost all $t \geq 0$ and assuming that there exists t_0 such that $\widetilde{\gamma}(w(t_0)) > 0$, then $\widetilde{\gamma}(w(t))$ is non-decreasing for $t \geq t_0$.

Proof. For convenience let $\widetilde{\gamma}(t) = \widetilde{\gamma}(w(t))$. Appealing to the fundamental theorem of calculus, we want to show that $\frac{d}{dt}\widetilde{\gamma}(t) \geq 0 \ \forall t \geq t_0$. Fix an arbitrary $t \geq t_0$ and define

$$u(t) = -\ln \sum_{i=1}^{n} \exp(-m_i(w(t))), \qquad v(t) = \|w(t)\|^L,$$

so that

$$\widetilde{\gamma}(t) = \frac{u(t)}{v(t)}.$$

Then

$$\frac{d}{dt}\widetilde{\gamma}(t) = \frac{\dot{u}(t)v(t) - u(t)\dot{v}(t)}{v(t)^2}.$$

Note that when $\widetilde{\gamma}(t) > 0$ we must have $w \neq 0$ so $v(t) > 0$. Now we analyze both $\dot{u}(t)$ and $\dot{v}(t)$. Since $\dot{w}(t) \in \partial u(t)$ and we assume the chain rule holds, we have for almost all t that

$$\begin{aligned}\dot{u}(t) &= \|\dot{w}(t)\|^2 \\ &\geq \|\dot{w}(t)\| \left\langle \frac{w(t)}{\|w(t)\|}, \dot{w}(t) \right\rangle \\ &\geq \frac{Lu(t)\|\dot{w}(t)\|}{\|w(t)\|}\end{aligned}$$

where the first inequality was by Cauchy-Schwarz and the second was by lemma 4. Next, again using Cauchy-Schwarz

$$\begin{aligned}\dot{v}(t) &= L\|w(t)\|^{L-1} \left\langle \frac{w(t)}{\|w(t)\|}, \dot{w}(t) \right\rangle \\ &\leq L\|w(t)\|^{L-1}\|\dot{w}(t)\|.\end{aligned}$$

Using these upper and lower bounds we have that

$$\begin{aligned}\dot{u}(t)v(t) - u(t)\dot{v}(t) &\geq \frac{Lu(t)\|\dot{w}(t)\|}{\|w(t)\|}v(t) - u(t)L\|w(t)\|^{L-1}\|\dot{w}(t)\| \\ &= u(t)L\|w(t)\|^{L-1}\|\dot{w}(t)\| - u(t)L\|w(t)\|^{L-1}\|\dot{w}(t)\| \\ &= 0\end{aligned}$$

(where $v(t) > 0$ as explained above and also $u(t) > 0$ because $\widetilde{\gamma}(t) > 0$). □