CS 839: Theoretical Foundations of Deep Learning **Spring 2022**

Lecture 8 Implicit Regularization III

Instructor: Yingyu Liang *Date: Feb 17^{th}, 2022* *Scriber: Chenghui Li*

1 Overview

In this course, we will see an asymptotic result for the Gradient Descent algorithm for logistic loss. Afterwards, we will see a rate result of Gradient Descent given that the step size is fixed with finite number of iterations.

2 Setup

Let's review some basic setups of the logistic regression and gradient descent.

Assume $\{x_i, y_i\}_{i=1}^n$ linearly separable, and $\|x_i\|_2 \leq 1$. Let $z_i = x_i y_i$, and consider the exponential loss:

$$L(w) = \frac{1}{n}\sum_{i=1}^{n} \exp(-w^\top z_i) \tag{1}$$

Consider gradient descent with any initialization w_0, we do the update as follow:

$$w_{t+1} = w_t - \eta_t \nabla L(w_t) \tag{2}$$

where η_t such that $0 < \eta_t \leq \min\{\eta_+, \frac{1}{L(w_t)}\}$ such that $0 < \eta_+ < +\infty$. When $\eta_t \to 0$, then this is gradient flow in a continuum regime, but can be hard to quantify under a discrete step size.

We want to show the following theorem in the course. This theorem tells us that minimizing exponential loss is equivalent to maximize the margin.

Theorem 1. Let $\{x_i, y_i\}_{i=1}^n$ be any linearly separable dataset. Let $l(\widehat{y}, y) = \exp(-\widehat{y}y)$ be the exponential loss. Suppose $\|x_i\|_2 \leq 1$, the step size is bounded $\eta_t \leq \min\{\eta_+, \frac{1}{L(w_t)}\}$ where $0 < \eta_+ < +\infty$, and we use an arbitrary initialization w_0, then the iterate w_t of gradient decent satisfies,

$$\lim_{t\to\infty} \min_{1\leq i\leq n} \frac{w_t^\top z_i}{\|w_t\|_2} = \max_{w} \min_{1\leq i\leq n} \frac{w^\top z_i}{\|w\|_2} := \gamma > 0.$$

The following lemmas can be easily proved by using what the loss function L is.

Lemma 2.

$$\|\nabla L(w)\|_2 \geq \gamma L(w), \forall w \tag{3}$$

Basically, Lemma 2 can be interpreted as if w is a bad solution for $L(w)$, then it has somewhere to go.

Lemma 3. The following properties of $L(w_t)$ and $\nabla L(w_t)$ hold:

(A) $\sum_{t=0}^{\infty} \eta_t \|\nabla L(w_t)\|_2^2 < \infty$.

(B) w_t converges to a global minimum, i.e., $L(w_t) \to 0$ and hence $\forall i, w_t^\top z_i \to \infty$ for any i.

(C) $\sum_{t=0}^{\infty} \eta_t \|\nabla L(w_t)\| = \infty$.

Lemma 4. If $\eta_t \leq \sqrt{2}/L(w_t)$, then $L(w_{t+1}) \leq L(w_t)$.

The following claim can also be easily shown by plugging L.

Claim 5.

$$\begin{aligned} \nabla L(w) &= -\frac{1}{n}\sum_{i=1}^{n} \exp(-w^\top z_i) z_i \\ \nabla^2 L(w) &= \frac{1}{n}\sum_{i=1}^{n} \exp(-w^\top z_i) z_i z_i^\top \end{aligned} \tag{4}$$

With the claim and lemmas in hand, we are now ready to show Theorem 1.

3 The proof of Theorem 1

First consider the unnormalized margin $\min_i w_{t+t}^\top z_i$. Basically, we will look at the approximation:

$$L(w_{t+1}) \leq L(w_t) + \langle \nabla L(w_t), w_t - w_{t+1} \rangle + \frac{1}{2} \sup_{\beta \in (0,1)} (w_{t+1} - w_t)^\top \nabla^2 L(w^\beta)(w_{t+1} - w_t) \tag{5}$$

where w^β is a linear combination between w_t and w_{t+1}. Notice that the above inequality is in fact equality for some $\beta \in (0,1)$, while we only need the upper bound.

By using $\|z\| \leq 1$, we can easily show $v^\top \nabla^2 L(w) v \leq \|v\|^2 L(w)$ by expanding left hand side and using (4).

Notice that by using what w_{t+1} is and the fact that $v^\top \nabla^2 L(w) v \leq \|v\|^2 L(w)$, we can see that

$$\begin{aligned} &L(w_t) + \langle \nabla L(w_t), w_t - w_{t+1} \rangle + \frac{1}{2} \sup_{\beta \in (0,1)} (w_{t+1} - w_t)^\top \nabla^2 L(w^\beta)(w_{t+1} - w_t) \\ &\leq L(w_t) - \eta_t \|\nabla L(w_t)\|^2 + \frac{1}{2}\eta_t^2 \|\nabla L(w_t)\|^2 L(w_t) \\ &= L(w_t) - \eta_t \gamma_t^2 + \frac{1}{2}\eta_t^2 L(w_t) \gamma_t^2 \\ &\leq L(w_t) \exp[-\frac{\eta_t \gamma_t^2}{L(w_t)} + \frac{1}{2}\eta^2 \gamma_t^2] \end{aligned} \tag{6}$$

where we denote $\|\nabla L(w_t)\|_2$ to be γ_t and we used $\exp(z) \geq z - 1$ for $z \in \mathbb{R}$ in the last inequality.

So, by combining (5) and (6), we have

$$L(w_{t+1}) \leq L(w_0) \exp\left(-\sum_{0 \leq s \leq t} \frac{\eta_s \gamma_s^2}{L(w_s)} + \sum_{0 \leq s \leq t} \frac{\eta_s^2 \gamma_s^2}{2} \right) \tag{7}$$

On the other hand, we have

$$L(w_{t+1}) = \frac{1}{n}\sum_{i=1}^{n}\exp(-w_{t+1}^\top z_i) \geq \frac{1}{n}\max_i \exp(-w_{t+1}^\top z_i) \tag{8}$$

So, by combining the above two equations, we have

$$\min_{1\leq i\leq n} w_{t+1}^\top z_i \geq \sum_{0\leq s\leq t}\frac{\eta_s\gamma_s^2}{L(w_s)} - \sum_{0\leq s\leq t}\frac{\eta_s\gamma_s^2}{2} - \log(nL(w_0)) \tag{9}$$

$$= \sum_{0\leq s\leq t}\frac{\eta_s\gamma_s^2}{L(w_s)} + \gamma\|w_0\| - \sum_{0\leq s\leq t}\frac{\eta_s\gamma_s^2}{2} - \log(nL(w_0)) - \gamma\|w_0\|. \tag{10}$$

Now consider the norm of the iterate. By using how gradient descent works, we have

$$\|w_{t+1}\| = \|w_0 - \sum_{0\leq s\leq t}\eta_s\nabla L(w_s)\| \leq \|w_0\| + \sum_{0\leq s\leq t}\eta_s\gamma_s \tag{11}$$

Recall that $\gamma_s = \|\nabla L(w_s)\| \geq \gamma L(w_s)$ by Lemma 2. Then we have

$$\frac{\sum_{0\leq s\leq t}\frac{\eta_s\gamma_s^2}{L(w_s)} + \gamma\|w_0\|}{\|w_0\| + \sum_{0\leq s\leq t}\eta_s\gamma_s} \geq \frac{\gamma\sum_{0\leq s\leq t}\eta_s\gamma_s + \gamma\|w_0\|}{\|w_0\| + \sum_{0\leq s\leq t}\eta_s\gamma_s} = \gamma. \tag{12}$$

Furthermore, by Lemma 3(A), we know that $\sum_{0\leq s\leq t}\frac{\eta_s\gamma_s^2}{2} < +\infty$; by Lemma 3(B), $\|w_{t+1}\| \to +\infty$. So

$$\frac{-\sum_{0\leq s\leq t}\frac{\eta_s\gamma_s^2}{2} - \log(nL(w_0)) - \gamma\|w_0\|}{\|w_{t+1}\|} \to 0. \tag{13}$$

Also, by definition of γ,

$$\frac{w_{t+1}^\top z_i}{\|w_{t+1}\|} \leq \gamma. \tag{14}$$

Combining (12)(13)(14), we have

$$\frac{w_{t+1}^\top z_i}{\|w_{t+1}\|} \to \gamma.$$

when $t \to \infty$. This completes the proof. This gives us a consistency result for Gradient Descent.

4 A stronger result

The above result only holds when $n \to \infty$, what the convergence result is about, and the then Theorem 6 is to analyze the rate with some additional assumptions. This will give us a result of the margin under a finite number of iterations circumstance.

Theorem 6. In the same setting as in Theorem 1, and further set $\eta_t = \eta = \frac{1}{L(w_0)}$. Then $\min_i \frac{w_t^\top z_i}{\|w_t\|} = \max_w \min_i \frac{w^\top z_i}{\|w\|_2} - O(\frac{1}{\log t})$.

Proof. Following the proof in Theorem 1, we arrive at

$$\min_{1\le i\le n} \frac{w_{t+1}^\top z_i}{\|w_{t+1}\|} \ge \frac{\sum_{0\le s\le t} \frac{\eta_s\gamma_s^2}{L(w_s)} + \gamma\|w_0\|}{\|w_{t+1}\|} - \frac{\sum_{0\le s\le t} \frac{\eta_s\gamma_s^2}{2} + \log(nL(w_0)) + \gamma\|w_0\|}{\|w_{t+1}\|}. \tag{15}$$

We also know the first term is lower bounded by γ and $\sum_{0\le s\le t} \frac{\eta_s\gamma_s^2}{2} < \infty$. So we only need to show that $\|w_t\|_2 = \Omega(\log t)$.

We have derived

$$L(w_{t+1}) \le L(w_t) - \eta_t\gamma_t^2 + \frac{1}{2}\eta_t^2\gamma_t^2 L(w_t) \le L(w_t) - \frac{1}{2}\eta\gamma_t^2 \le L(w_t) - \frac{1}{2}\eta\gamma^2 L(w_t)^2. \tag{16}$$

If we simplify the notation by denoting $L(w_t)$ to be a_t and $c^2 = \frac{1}{2}\eta\gamma^2$, then the above result can be concluded as

$$a_{t+1} \le a_t - c^2 a_t^2. \tag{17}$$

Then, by solving this induction,

$$a_{t+1} \le \frac{1}{\frac{1}{a_0} + \frac{(t+1)c^2}{1-c^2 a_0}} \tag{18}$$

By using the fact that

$$0 \le c^2 a_0 = \frac{1}{2}\eta\gamma^2 L(w_0) = \frac{1}{2}\gamma^2 \le \frac{1}{2}$$

we have

$$\frac{c^2}{1-c^2 a_0} \ge c^2 \tag{19}$$

Then, by combining (18) and (19), we have

$$a_{t+1} \le \frac{1}{(t+1)c^2} = \frac{2}{(t+1)\eta\gamma^2}. \tag{20}$$

Then for $\forall i$,

$$\frac{1}{n}\exp(-w_{t+1}^\top z_i) \le L(w_{t+1}) \le \frac{2}{(t+1)\eta\gamma^2}. \tag{21}$$

This leads to

$$\|w_{t+1}\| \ge w_{t+1}^\top z_i \ge \log\frac{(t+1)\eta\gamma^2}{2n}$$

This shows the claim in the beginning of the proof.

Combining all of the above, and we can conclude the result. □

Remark 7. Theorem 6 only holds when constraining the step size because we only know when the step size is large enough and then we can know the rate. Theorem 1 holds for the case that η_t is bounded above, but it might come to a continuum regime. If η_t is very small, then GD will converge to a gradient flow case, which is the continuous limit of gradient descent. In this case, it is impossible to talk about the rate. Theorem 6 considers the discrete case and analyzes the rate of margins.