CS 839: Theoretical Foundations of Deep Learning		Spring 2023
Lecture 7 Implicit Regularization II		
Instructor: Yingyu Liang	*Date:*	*Scriber: Zhenmei Shi*

1 Overview

In this lecture, we will finish the proof of using gradient flow on logistic regression with exponential loss under overparameterized setting, which is defined in the last lecture. We will also introduce a new proof of implicit bias by replacing the gradient flow to the gradient descent.

2 Logistic Regression with Gradient Flow

Let $S = \{(\mathbf{x}_i, y_i)\}_{i=1}^n$ be a training set. $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_n]^\top \in \mathbb{R}^{n\times d}$, $\mathbf{y} = (y_1, \ldots, y_n)^\top \in \mathbb{R}^n$, where $n < d$ which is overparameterized setting and $\mathbf{X}$ is full rank. Let $\mathbf{z}_i = y_i\mathbf{x}_i$ and $\widehat{y}_i = \mathbf{w}^\top\mathbf{x}_i$ (the output of the model) be the score of classification. Normally, when $\widehat{y}_i > 0$ the predicted label is $+1$ and when $\widehat{y}_i < 0$ the predicted label is -1.

Let us consider exponential loss in a binary classification problem with $y \in \{+1, -1\}$

$$L_{\exp}(\mathbf{w}) = \sum_{i=1}^n \exp(-y_i\mathbf{w}^\top\mathbf{x}_i)$$

where we ignore the $\frac{1}{n}$ factor for simplicity. Consider gradient flow with exponential loss

$$\dot{\mathbf{w}}(t) = \frac{d\mathbf{w}(t)}{dt} = -\nabla L(\mathbf{w}(t)).$$

We will assume that perfect classification exists.

Assumption 1. Assume the training data is linear separable: there exists $\mathbf{w}$, such that $\forall i \in [n],\ y_i\mathbf{w}^\top\mathbf{x}_i \geqslant 1$.

One observation is that for exponential loss the optimal $\mathbf{w}$ is infinite far away. Thus, we are interested in the direction of $\mathbf{w}$.

Definition 2 (Convergence in direction)**.** If for some $\widehat{\mathbf{w}}$,

$$\lim_{t\to\infty} \frac{\mathbf{w}(t)}{\|\mathbf{w}(t)\|_2} = \frac{\widehat{\mathbf{w}}}{\|\widehat{\mathbf{w}}\|_2},$$

we say that the direction of $\mathbf{w}(t)$ will converge to the the direction of $\widehat{\mathbf{w}}$ as $t \to \infty$.

Definition 3 (Maximum margin solution)**.** $\widehat{\mathbf{w}}$ is the maximum margin solution if

$$\widehat{\mathbf{w}} = \operatorname*{argmin}_{\mathbf{w}} \frac{1}{2}\|\mathbf{w}\|^2$$
$$\text{s.t. } \forall i \in [n], y_i\mathbf{w}^\top\mathbf{x}_i \geqslant 1.$$

Note that in the above definition the margin will be $\frac{1}{\|\mathbf{w}\|}$, so minimum norm leads to maximum margin.

For simplicity, let

$$\mathbf{z}_i = y_i \mathbf{x}_i.$$

By KKT condition on the maximum margin solution, we have

$$\begin{aligned} L(\mathbf{w}, \alpha) &= \frac{1}{2}\|\mathbf{w}\|_2^2 + \sum_{i=1}^n \alpha_i [1 - \mathbf{w}^\top \mathbf{z}_i] \\ 0 &= \frac{\partial L(\mathbf{w}, \alpha)}{\partial \mathbf{w}} = \mathbf{w} + \sum_{i=1}^n \alpha_i [-\mathbf{z}_i] \\ \widehat{\mathbf{w}} &= \sum_{i=1}^n \alpha_i \mathbf{z}_i, \end{aligned}$$

where α_i is a Lagrangian multiplier. Furthermore, by complementary slackness,

$$\begin{cases} \forall i \in \mathbb{S}^c, & \widehat{\mathbf{w}}^\top \mathbf{z}_i > 1, \alpha_i = 0 \\ \forall i \in \mathbb{S}, & \widehat{\mathbf{w}}^\top \mathbf{z}_i = 1, \alpha_i \geqslant 0, \end{cases} \tag{1}$$

where $\mathbb{S} = \{i \in [n] : \widehat{\mathbf{w}}^\top \mathbf{z}_i = 1\}$ denotes the support vectors. For simplicity of the analysis, we will assume the following mild technical assumption:

Assumption 4. Assume that for any $i \in \mathbb{S}$, $\alpha_i > 0$.

Theorem 5. We consider the overparameterized setting with linearly separable training data $(\mathbf{X}, \mathbf{y})$ satisfying Assumptions 1 and 4. For exponential loss and and any initialization $\mathbf{w}(0)$, gradient flow with infinitesimal step satisfies $\mathbf{w}(t)$ converge to the direction of $\widehat{\mathbf{w}}$, where $\widehat{\mathbf{w}}$ is the maximum margin solution.

Proof. The intuition is that we would like to show that

$$\mathbf{w}(t) = \widehat{\mathbf{w}} s(t) + o(s(t)),$$

where $\widehat{\mathbf{w}}$ is the maximum margin classifier, and $\widehat{\mathbf{w}} s(t)$ will dominate (much larger than $o(s(t))$). If above equation is true, then we can get direction convergence. In later calculation, we will find that $s(t) = \log t$ and $o(s(t))$ related to $\widetilde{\mathbf{w}}$ which we defined before. This is the intuition about the definition of $\widetilde{\mathbf{w}}$ and the equation below.

$$\mathbf{w}(t) = \widehat{\mathbf{w}} \log t + \widetilde{\mathbf{w}} + \mathbf{r}(t), \tag{2}$$

where we need to prove $\|\mathbf{r}(t)\|_2^2 = o(\log t)$. By fundamental theorem of calculus,

$$\|\mathbf{r}(t_2)\|_2^2 = \|\mathbf{r}(t_1)\|_2^2 + \int_{t_1}^{t_2} \frac{d\|\mathbf{r}(t)\|_2^2}{dt} dt.$$

We only need to prove $\frac{d\|\mathbf{r}(t)\|_2^2}{dt} < \frac{1}{t}$. By ODE, in the last lecture, we showed,

$$\frac{1}{2}\frac{d\|\mathbf{r}(t)\|^2}{dt} = \mathbf{r}(t)\cdot\dot{\mathbf{r}}(t) \tag{3}$$

$$= \sum_{i\in\mathbb{S}} \exp(-\mathbf{z}_i^\top \mathbf{w}(t))\mathbf{z}_i^\top \mathbf{r}(t) - \frac{1}{t}\widehat{\mathbf{w}}^\top \mathbf{r}(t) + \sum_{i\notin\mathbb{S}} \exp(-\mathbf{z}_i^\top \mathbf{w}(t))\mathbf{z}_i^\top \mathbf{r}(t). \tag{4}$$

By (1) and (2), we showed that the first two terms $\leqslant 0$ in the last lecture,

$$\sum_{i\in\mathbb{S}} \exp(-\mathbf{z}_i^\top \mathbf{w}(t))\mathbf{z}_i^\top \mathbf{r}(t) - \frac{1}{t}\widehat{\mathbf{w}}^\top \mathbf{r}(t) \leqslant 0.$$

Now, we will prove the third term in (4) $< \frac{1}{t}$. Let $\theta = \min_{i\notin\mathbb{S}} \mathbf{z}_i^\top \widehat{\mathbf{w}}$, we have $\theta > 1$ by the definition of $\mathbb{S}^c$ in (1).

$$\begin{aligned}
\sum_{i\notin\mathbb{S}} \exp(-\mathbf{z}_i^\top \mathbf{w}(t))\mathbf{z}_i^\top \mathbf{r}(t) &= \sum_{i\notin\mathbb{S}} \exp(-\underbrace{\mathbf{z}_i^\top \widehat{\mathbf{w}}}_{\geqslant\theta>1}\log t - \mathbf{z}_i^\top \widetilde{\mathbf{w}} - \mathbf{z}_i^\top \mathbf{r}(t))\mathbf{z}_i^\top \mathbf{r}(t) \\
&\leqslant \frac{1}{t^\theta}\underbrace{\sum_{i\notin\mathbb{S}} \exp(-\mathbf{z}_i^\top \widetilde{\mathbf{w}})}_{\leqslant C}\underbrace{\exp(-\mathbf{z}_i^\top \mathbf{r}(t))\mathbf{z}_i^\top \mathbf{r}(t)}_{\leqslant 1} \\
&\leqslant \frac{C}{t^\theta},
\end{aligned}$$

where C is a constant that does not depend on t. Thus, we have

$$\|\mathbf{r}(t_2)\|_2^2 \leqslant \|\mathbf{r}(t_1)\|_2^2 + 2\int_{t_1}^{t_2} \frac{C}{t^\theta} dt < C'.$$

This shows that $\|\mathbf{r}(t)\|_2^2$ is bounded, and thus the residual $\mathbf{r}(t)$ is bounded. We finished the proof. □

3 Logistic Regression with Gradient Descent

We want to use gradient descent to minimize this $L(\mathbf{w}) = \frac{1}{n}\sum_{i=1}^n \exp(-y_i \mathbf{w}^\top \mathbf{x}_i)$. In each iteration, our update rule is:

$$\mathbf{w}_{t+1} = \mathbf{w}_t + \frac{\eta_t}{n}\sum_{i=1}^n \exp(-y_i \mathbf{w}^\top \mathbf{x}_i) y_i \mathbf{x}_i.$$

We will prove the following theorem.

Theorem 6. Let $\{\mathbf{x}_i, y_i\}_{i=1}^n$ be any linearly separable dataset. Let $l(\widehat{y}, y) = \exp(-\widehat{y}y)$ be the exponential loss. Suppose $\|\mathbf{x}_i\|_2 \leqslant 1$, the step size is bounded $\eta_t \leqslant \min\{\eta_+, \frac{1}{L(\mathbf{w}_t)}\}$ where $0 < \eta_+ < +\infty$, and we use an arbitrary initialization $\mathbf{w}_0$, then the iterate $\mathbf{w}_t$ of gradient decent satisfies,

$$\lim_{t\to\infty} \min_{i\in[n]} \frac{y_i \mathbf{w}_t^\top \mathbf{x}_i}{\|\mathbf{w}_t\|_2^2} = \max_{\|\mathbf{w}\|=1} \min_{i\in[n]} y_i \mathbf{w}^\top \mathbf{x}_i := \gamma.$$

Here, $\min_{i\in[n]} \frac{y_i \mathbf{w}_t^\top \mathbf{x}_i}{\|\mathbf{w}_t\|_2^2}$ represents the margin of classifier $\mathbf{w}_t$, and the right most term is the maximum margin. The high level idea we have the theorem is that we can do convex optimization on the margin based loss.

We first lay down some lemmas that will be useful in our proof.

Lemma 7. For any $\mathbf{w}$, we have $\|L(\mathbf{w})\| \geqslant \gamma L(\mathbf{w})$.

Above Lemma is also called the PL condition, which guarantees there is no bad local minimum in the loss landscape.

Lemma 8. The following properties of $L(\mathbf{w}_t)$ and $\nabla L(\mathbf{w}_t)$ hold:

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

(B) $\mathbf{w}_t$ converges to a global minimum i.e., $L(\mathbf{w}_t) \to 0$ and hence , $\mathbf{w}_t^\top \mathbf{z}_i \to \infty$ for any i.

(C) $\sum_{t=0}^{\infty} \eta_t \|\nabla L(\mathbf{w}_t)\| = \infty$

Lemma 9. If $\eta_t \leqslant \frac{\sqrt{2}}{L(\mathbf{w}_t)}$, we have $L(\mathbf{w}_{t+1}) \leqslant L(\mathbf{w}_t)$.

We will prove above lemmas in the homework. Now, we are ready to prove the main theorem.

Proof. By Taylor expansion, we have,

$$\begin{aligned}
L(\mathbf{w}_{t+1}) =& L(\mathbf{w}_t) + \langle \nabla L(\mathbf{w}_t), \mathbf{w}_{t+1} - \mathbf{w}_t \rangle \\
&+ \frac{1}{2} \sup_{\beta\in(0,1)} (\mathbf{w}_{t+1} - \mathbf{w}_t)^\top \nabla^2 L(\mathbf{w}_t + \beta(\mathbf{w}_{t+1} - \mathbf{w}_t))(\mathbf{w}_{t+1} - \mathbf{w}_t) \\
\leqslant& L(\mathbf{w}_t) - \eta_t \|\nabla L(\mathbf{w}_t)\|_2^2 + \frac{\eta_t^2}{2} \underbrace{\sup_{\beta\in(0,1)} L(\mathbf{w}_t - \beta\eta_t \nabla L(\mathbf{w}_t))}_{\leqslant L(\mathbf{w}_t) \text{ by Lemma 9}} \|\nabla L(\mathbf{w}_t)\|_2^2 \\
\leqslant& L(\mathbf{w}_t) - \eta_t \|\nabla L(\mathbf{w}_t)\|_2^2 + \frac{\eta_t^2}{2} L(\mathbf{w}_t) \|\nabla L(\mathbf{w}_t)\|_2^2,
\end{aligned}$$

where the first inequality is from $\mathbf{w}_{t+1} = \mathbf{w}_t - \eta_t \nabla L(\mathbf{w}_t)$ and Hessian matrix calculated by loss function. We will continue the proof in the next lecture. □