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

Lecture 5 Implicit Regularization I

Instructor: Yingyu Liang *Date:* *Scriber: Zhenmei Shi*

1 Overview

In previous lecture, we covered the approximation power of neural networks. Recall that we decomposed the risk into three parts: approximation, estimation/generalization and optimization. We also have a conjecture that the optimization has some implicit regularization effect that restricts the learning dynamics to a subset of the whole hypothesis class. In this lecture, we will start to study implicit regularization of gradient decent optimization in training dynamic of neural networks.

2 Basics in Optimization

2.1 Optimization Problem

Let $S = \{(\mathbf{x}_i, y_i)\}_{i=1}^n$ be the training set, each $(\mathbf{x}_i, y_i)$ is a pair of training data point with $\mathbf{x}_i$ being the feature vector and y_i being the label, and $f_{\mathbf{w}}$ is a neural network function parameterized by $\mathbf{w}$. Let $L(\mathbf{w})$ be the training loss of some neural network parameterized by $\mathbf{w}$:

$$L(\mathbf{w}) = \frac{1}{n}\sum_{i=1}^{n} l(f_{\mathbf{w}}(\mathbf{x}_i), y).$$

(Usually, $L(\mathbf{w})$ denotes the expected loss and $L_S(\mathbf{w})$ denote the training loss. Here for simplicity, we let $L(\mathbf{w})$ denote the training loss.)

2.2 Gradient Descent

Gradient descent is one of the simplest algorithms to solve the optimization problem.

1. Initially, we set the parameter to $\mathbf{w}_0$.
2. Then at the t-th step, we iteratively update the parameter $\mathbf{w}_{t+1} = \mathbf{w}_t - \eta_t \nabla L(\mathbf{w}_t)$ for $t = 0, 1, \ldots$

Specifically, in each iteration, we walk along the negation of the gradient direction, with a step size of $\eta_t > 0$ (a.k.a. learning rate in deep learning).

2.3 Gradient Flow (gradient descent with infinitesimal step size)

Sometimes we want a smoothness property, and we can analyze the gradient flow instead in this case. It can be viewed as a gradient descent update with infinitesimal step size, i.e., we let $\eta_t \to 0$. Gradient flow can be written as an ordinary differential equation (ODE):

1. $\mathbf{w}_0 = \mathbf{w}(0)$.
2. $t \in \mathbb{R}_{\geqslant 0}$, $\frac{d\mathbf{w}(t)}{dt} := \dot{\mathbf{w}}(t) = -\nabla L(\mathbf{w}(t))$.

where we take the derivative of $\mathbf{w}$ w.r.t. time index t, denoted as $\dot{\mathbf{w}}(t)$, while on the right hand side ∇L is the gradient of the training loss w.r.t. the weight parameter (not w.r.t. the time).

3 Implicit Bias/Regularization

We want to analyze the implicit property of the training when we use gradient descent to optimize some objective function. It is called implicit bias or implicit regularization. In this lecture, we will focus on linear regression and logistic regression first.

3.1 Linear Regression

Let $S = \{(\mathbf{x}_i, y_i)\}_{i=1}^n$ is the 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 i.e., there are an infinite number of optimal solutions. Each row of $\mathbf{X}$ is a data point.

$$L(\mathbf{w}) = \frac{1}{n}\sum_{i=1}^{n} \frac{1}{2}(\langle \mathbf{x}_i, \mathbf{w} \rangle - y_i)^2 = \frac{1}{2n}\|\mathbf{X}\mathbf{w} - \mathbf{y}\|_2^2.$$

Here we use square loss which is convex. Due to convexity, if we have small enough learning rate, we can guarantee convergence to global minimum by convex optimization for linear regression.

We use gradient descent to solve the above linear regression problem. Without loss of generality, we initialize $\mathbf{w}_0 \in \mathbb{R}^d$ to the $\mathbf{0}$ vector, and thereby the update rule is,

$$\mathbf{w}_{t+1} = \mathbf{w}_t - \eta_t \nabla L(\mathbf{w}_t) = \mathbf{w}_{t+1} = \mathbf{w}_t - \frac{\eta_t}{n}\mathbf{X}^\top(\mathbf{X}\mathbf{w}_t - \mathbf{y}).$$

We can see that $\mathbf{X}^\top(\mathbf{X}\mathbf{w}_t - \mathbf{y})$ is a linear combination of the data points in $\mathbb{R}^d$. We denote it as $\mathbf{X}^\top(\mathbf{X}\mathbf{w}_t - \mathbf{y}) \in \mathrm{lin}\{\mathbf{x}_i\}_{i=1}^n$. Thus, we have the following lemma.

Lemma 1. The iterates of GD lie in the span of data points, i.e., $\mathbf{w}_t \in \mathrm{lin}\{\mathbf{x}_i\}_{i=1}^n$ for $t = 0, 1, 2, \ldots$

Proof. Can be easily proved with mathematical induction. □

So we can see that the GD converges to some solution in the linear span of the training data. Now let's consider the structure of the solutions in this linear span.

Lemma 2. There is a unique solution in $\mathrm{lin}\{\mathbf{x}_i\}_{i=1}^n$ to the linear regression problem.

Proof. It is easy to see the existence, e.g., $\mathbf{X}^\dagger y$ is such a solution where $\mathbf{X}^\dagger$ is the pseudo-inverse of $\mathbf{X}$.

The uniqueness can be proved by contradiction. Suppose $\mathbf{w}_1^* = \mathbf{X}^\top \alpha_1$ and $\mathbf{w}_2^* = \mathbf{X}^\top \alpha_2$ are both in $\mathrm{lin}\{\mathbf{x}_i\}_{i=1}^n$ and solutions to the linear regression problem. Because $n < d$, we have zero training loss, $\mathbf{y} = \mathbf{X}\mathbf{w}_1^* = \mathbf{X}\mathbf{w}_2^*$. Thus,

$$\mathbf{0} = \mathbf{X}(\mathbf{w}_1^* - \mathbf{w}_2^*) = \mathbf{X}\mathbf{X}^\top(\alpha_1 - \alpha_2).$$

Because $\mathbf{X}\mathbf{X}^\top \in \mathbb{R}^{n\times n}$ is full rank, $\alpha_1 = \alpha_2$. □

This unique solution can be further characterized.

Lemma 3. The minimum norm solution is in $\mathrm{lin}\{\mathbf{x}_i\}_{i=1}^n$: If $\widehat{\mathbf{w}} = \mathrm{argmin}_{\mathbf{w}} \frac{1}{2}\|\mathbf{w}\|^2$ s.t. $\mathbf{X}\mathbf{w} = \mathbf{y}$, then $\widehat{\mathbf{w}} \in \mathrm{lin}\{\mathbf{x}_i\}_{i=1}^n$.

Proof. We decompose $\widehat{\mathbf{w}}$ into the space of $\mathrm{lin}\{\mathbf{x}_i\}_{i=1}^n$, and the complement space of it. Namely, $\widehat{\mathbf{w}} = \widehat{\mathbf{w}}_1 + \widehat{\mathbf{w}}_\perp$, where $\widehat{\mathbf{w}}_1 \in \mathrm{lin}\{\mathbf{x}_i\}_{i=1}^n$, $\widehat{\mathbf{w}}_\perp \notin \mathrm{lin}\{\mathbf{x}_i\}_{i=1}^n$. Note that,

$$\|\widehat{\mathbf{w}}\|_2^2 = \|\widehat{\mathbf{w}}_1\|_2^2 + \|\widehat{\mathbf{w}}_\perp\|_2^2 \geqslant \|\widehat{\mathbf{w}}_1\|_2^2.$$

Note that $\mathbf{X}\widehat{\mathbf{w}}_\perp = 0$, $\widehat{\mathbf{w}}_1$ satisfies,

$$\mathbf{X}\widehat{\mathbf{w}}_1 = \mathbf{X}(\widehat{\mathbf{w}} - \widehat{\mathbf{w}}_\perp) = \mathbf{X}\widehat{\mathbf{w}} = 0,$$

However, we know that $\widehat{\mathbf{w}}$ is the minimum norm solution, so we must have,

$$\widehat{\mathbf{w}} = \widehat{\mathbf{w}}_1 \in \mathrm{lin}\{\mathbf{x}_i\}_{i=1}^n.$$

This completes the proof. □

Thus, we can conclude that gradient descent with $\mathbf{w}_0 = \mathbf{0}$ converges to the minimum norm solution to the linear regression problem.

3.2 Logistic Regression

Let $S = \{(\mathbf{x}_i, y_i)\}_{i=1}^n$ is the 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 us consider the 0-1 loss and exponential loss in a binary classification problem with $y \in \{+1, -1\}$.

$$\begin{aligned} L_{\text{0-1}}(\mathbf{w}) &= \frac{1}{n}\sum_{i=1}^n \mathbb{I}\{\mathrm{sign}(\mathbf{w}^\top \mathbf{x}_i) \neq y_i\} = \frac{1}{n}\sum_{i=1}^n \mathbb{I}\{-y_i \mathbf{w}^\top \mathbf{x}_i \geqslant 0\} \\ L_{\exp}(\mathbf{w}) &= \frac{1}{n}\sum_{i=1}^n \exp(-y_i \mathbf{w}^\top \mathbf{x}_i). \end{aligned}$$

Let $\widehat{y}_i = \mathbf{w}^\top \mathbf{x}_i$ denote the output of the model. Normally, when $\widehat{y}_i > 0$ the predicted label is $+1$ and when $\widehat{y}_i < 0$ the predicted label is -1.

The 0-1 loss function is non-convex and non-differentiable (as the blue line shows in Figure 1). Thus it is computationally difficult (NP hard) to directly minimize this 0-1 training loss.

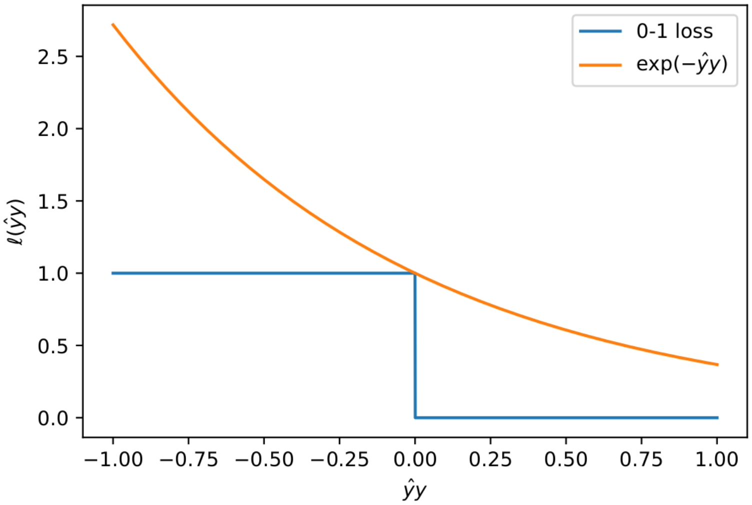

Figure 1: 0-1 loss and exponential loss.

In practice we usually use a surrogate loss, such as an exponential loss, logistic loss or hinge loss, which are upper bounds of the 0-1 loss. We use the exponential loss here (the orange line in Figure 1), and 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 4. 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 5 (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 6 (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 7. Assume that for any $i \in \mathbb{S}$, $\alpha_i > 0$.

We have the following theorem.

Theorem 8. We consider the overparameterized setting with linearly separable training data $(\mathbf{X}, \mathbf{y})$ satisfying Assumptions 4 and 7. 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. We are going to show that

$$\mathbf{w}(t) = \widehat{\mathbf{w}} \log t + \phi(t),$$

where $\widehat{\mathbf{w}}$ is the maximum margin classifier, and $\phi(t) \in \mathbb{R}^d$ is some residual. $\widehat{\mathbf{w}} \log t$ will dominate (much larger than $\phi(t)$).

Define

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

where $\widetilde{\mathbf{w}}$ is a vector satisfying,

$$\forall i \in \mathbb{S}, \alpha_i > 0 \text{ we have } \widetilde{\mathbf{w}} \text{ s.t. } \exp(-\widetilde{\mathbf{w}}^\top \mathbf{z}_i) = \alpha_i.$$

Note that overparameterization guarantees its existence.

Then by Assumption 7,

$$\widehat{\mathbf{w}} = \sum_{i=1}^{n} \alpha_i \mathbf{z}_i = \sum_{\alpha_i \neq 0} \alpha_i \mathbf{z}_i = \sum_{i \in \mathbb{S}} \exp(-\widetilde{\mathbf{w}}^\top \mathbf{z}_i) \mathbf{z}_i.$$

We consider $\|\mathbf{r}(t)\|_2^2$. By ODE we have,

$$\begin{aligned}
\frac{1}{2}\frac{d\|\mathbf{r}(t)\|^2}{dt} &= \mathbf{r}(t)\cdot\dot{\mathbf{r}}(t) && (3)\\
&= \left(-\nabla L(\mathbf{w}(t)) - \frac{1}{t}\widehat{\mathbf{w}}\right)^\top \mathbf{r}(t) && (4)\\
&= \sum_{i=1}^{n}\exp(-\mathbf{z}_i^\top\mathbf{w}(t))\mathbf{z}_i^\top\mathbf{r}(t) - \frac{1}{t}\widehat{\mathbf{w}}^\top\mathbf{r}(t) && (5)\\
&= \sum_{i\in\mathbb{S}}\exp(-\mathbf{z}_i^\top\mathbf{w}(t))\mathbf{z}_i^\top\mathbf{r}(t) - \frac{1}{t}\widehat{\mathbf{w}}^\top\mathbf{r}(t) + \sum_{i\notin\mathbb{S}}\exp(-\mathbf{z}_i^\top\mathbf{w}(t))\mathbf{z}_i^\top\mathbf{r}(t). && (6)
\end{aligned}$$

By (2), we have $\mathbf{w}(t) = \widehat{\mathbf{w}}\log t + \widetilde{\mathbf{w}} + \mathbf{r}(t)$. By (1), we have $\mathbf{z}_i^\top\widehat{\mathbf{w}} = 1$ for $i \in \mathbb{S}$. We will show that the first two terms $\leqslant 0$,

$$\begin{aligned}
&\sum_{i\in\mathbb{S}}\exp(-\mathbf{z}_i^\top\mathbf{w}(t))\mathbf{z}_i^\top\mathbf{r}(t) - \frac{1}{t}\widehat{\mathbf{w}}^\top\mathbf{r}(t)\\
=&\sum_{i\in\mathbb{S}}\exp(-\mathbf{z}_i^\top\widehat{\mathbf{w}}\log t - \mathbf{z}_i^\top\widetilde{\mathbf{w}} - \mathbf{z}_i^\top\mathbf{r}(t))\mathbf{z}_i^\top\mathbf{r}(t) - \frac{1}{t}\widehat{\mathbf{w}}^\top\mathbf{r}(t)\\
=&\sum_{i\in\mathbb{S}}\exp(-\log t - \mathbf{z}_i^\top\widetilde{\mathbf{w}} - \mathbf{z}_i^\top\mathbf{r}(t))\mathbf{z}_i^\top\mathbf{r}(t) - \frac{1}{t}\sum_{i\in\mathbb{S}}\exp(-\mathbf{z}_i^\top\widetilde{\mathbf{w}})\mathbf{z}_i^\top\mathbf{r}(t)\\
=&\frac{1}{t}\sum_{i\in\mathbb{S}}\exp(-\mathbf{z}_i^\top\widetilde{\mathbf{w}})[\exp(-\mathbf{z}_i^\top\mathbf{r}(t)) - 1]\mathbf{z}_i^\top\mathbf{r}(t)\\
\leqslant&0
\end{aligned}$$

The last inequality comes from $(\exp(-r) - 1)r \leqslant 0$. Next lecture, we will handle the third term in (6) and continue to finish the proof. □