Modern Topics on Statistical Learning Theory Spring 2023

Lecture 7 — Implicit Regularization

Prof. Qi Lei *Scribe: Haoxu Huang, Xin Peng, Zihan Wang*

1 Recap

Neural Tangent Kernel (NTK): Defined as $x \mapsto \theta^T \phi(x)$, where θ is over-parameterized and $\phi(x)$ is high dimensional. This can easily lead to overfitting under traditional statistical learning framework. This leads to our motivation of understanding generalization.

• Uniform Convergence

Generalization Gap $\leq \sqrt{\frac{\mathcal{G}(\Theta)}{n}}$, where $\mathcal{G}(\Theta)$ is complexity of function and can be arbitrarily large.

This needs some kind of boundedness in the function class with explicit regularization defined as $\min_{\|\theta\| \leq R} \mathcal{L}(\theta) \iff \min_{\theta} \mathcal{L}(\theta) + \lambda \|\theta\|$, which has same effect as controlling $\|\theta\| \leq R$.

• (Zhang et al. 2017, [3]) Understand Deep Learning Requires Rethinking Generalization

This paper found that Neural Network (NN) can fit random labels with $Train\ Error = 0$. However, generalization is terrible in this case with $Test\ Error \simeq Random\ Guess$. This means that neural networks are able to capture the remaining signal in the data, while at the same time fit the noisy part using brute-force.

To understand this, we study Implicit Regularization, which states that Gradient Descend (GD) performs some implicit regularization to find better global minimum hence generalizing well. Geometrically, it can be illustrated as Figure 1, where GD helps to find flatter minimum and hence the difference between train and test loss is smaller.

Figure 1: Train vs. Test Loss. The generalization error is larger for sharper minima.

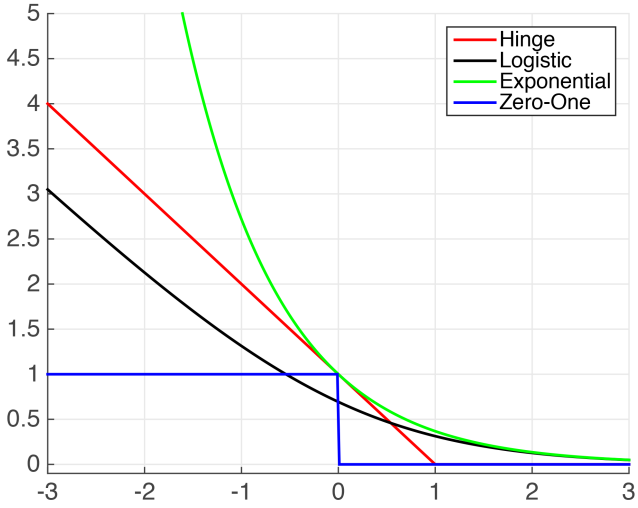

Figure 2: Loss Functions

2 Implicit Regularization in Linear and Overparametrized Setting

Given n equations and m unknowns and a function $F(\theta, X) = y, y \in \mathbb{R}^n$ and $\theta \in \mathbb{R}^m$, we say the function is overparametrized when $m > n$. For example, consider $\min_\theta \mathcal{L}(\theta) = \frac{1}{2}\|F(\theta, X) - y\|^2$, $n = 2$, $m = 3$.

• (Soudry et al. 2018, [2]) The Implicit Bias of Gradient Descent on Separable Data

Consider a setting where the target function $f_\theta(X) = \theta^\top X$ and data $\{(x_i, y_i)_{i=1}^n\}$, where $f_\theta : \mathbb{R} \to \mathbb{R}, x_i \in \mathbb{R}^d, y_i \in \{-1, 1\}$ for the problem is 1) binary classification, 2) linear separable 3) without bias term. Then, we can examine the effect on different loss functions as shown in Figure 2, where exponential loss is $\ell(u) = e^{-u}$, logistic loss is $\ell(u) = \log(1+e^{-u})$ for $u := yf_\theta(x)$. Notice that 0-1 loss or hinge loss will not show explicit effect because they have too small loss on correct classification.

For $\mathcal{L}(\theta) = \sum_{i=1}^n \ell(y_i\theta^\top x_i)$, taking $z_i = y_i x_i$ to have $\sum_{i=1}^n \ell(\theta^\top z_i)$. This now gives the optimization objective $\min_\theta \sum_{i=1}^n \ell(\theta^T z_i)$.

Formally, we have

Assumption 1. *The dataset is linearly separable:* $\exists \theta_*$ *such that* $\forall n : \theta_*^\top x_n > 0$.

Assumption 2. $\ell(u)$ *is a positive, differentiable, monotonically decreasing to zero, (so* $\forall u : \ell(u) > 0, \ell'(u) < 0, \lim_{u\to\infty} \ell'(u) = 0$*), a* β*-smooth function, i.e. its derivative is* β*-Lipschitz, and* $\limsup_{u\to-\infty} \ell'(u) < 0$.

Assumption 2 includes many common loss functions, including the logistic, exp-loss and probit losses. Assumption 2 implies that $\mathcal{L}(\theta)$ is a $\beta\sigma_{\max}^2$-smooth function, where $\sigma_{\max}(X)$ is the maximal singular value of the data matrix $X \in \mathbb{R}^{d\times N}$.

Under these conditions, the infimum of the optimization problem is zero, but it is not attained at any finite θ. In fact, no finite critical point θ exists. Since data is linearly separable, there exists θ^*,

such that $z_i^\top \theta^* > 0$ for all i. Then

$$\theta^{*\top} \nabla \mathcal{L}(\theta) = \sum_{i=1}^{n} \ell'(\theta^\top z_i) \cdot z_i^T \theta^* < 0$$

since $\ell'(\theta^\top z_i) < 0$ according to Assumption 2. Therefore, $\nabla \mathcal{L}(\theta) \neq 0$ and it further implies that there are no finite critical point for $\mathcal{L}$. However, GD on a smooth loss converge to critical point (with proper stepsize) so $\|\theta(t)\| \to \infty$. It is also direct that $\lim_{t\to\infty} \nabla \mathcal{L}(\theta(t)) \to 0$ and $\ell'(\theta(t); x_i) \to 0$ when $t \to 0$. Then our **main question** becomes what direction will the weight converges, or formally, what is $\lim_{t\to\infty} \frac{\theta(t)}{\|\theta(t)\|}$.

3 Hard SVM

Before we answer the our main question, we first introduce *hard SVM* problem of separable data. For a hard SVM problem, we are going to find the solution with max margin (as shown in Figure 3). Formally, we solve the following optimization problem:

$$\min \|\theta\|^2, \quad s.t.\ \theta^\top z_i \geq 1 \text{ for any } i. \tag{1}$$

For those samples satisfying $\theta_{\text{SVM}}^\top z_i = 1$, we call them *support vectors* (the samples on the dashed lines in Figure 3) and denote by $x_i \in \text{SV}$. Based on support vectors, we can define the *margin* by $\gamma := \frac{\theta_{\text{SVM}}}{\|\theta_{SVM}\|}^T z_i$, for $x_i \in \text{SV}$. The KKT condition for the hard SVM problem is as follows:

$$\theta_{\text{SVM}} = \sum_{i=1}^{n} \alpha_i x_i = \sum_{x_i \in SV} \alpha_i x_i \ \ (\alpha_i = 0, \ \ if \ \ x_i \notin SV). \tag{2}$$

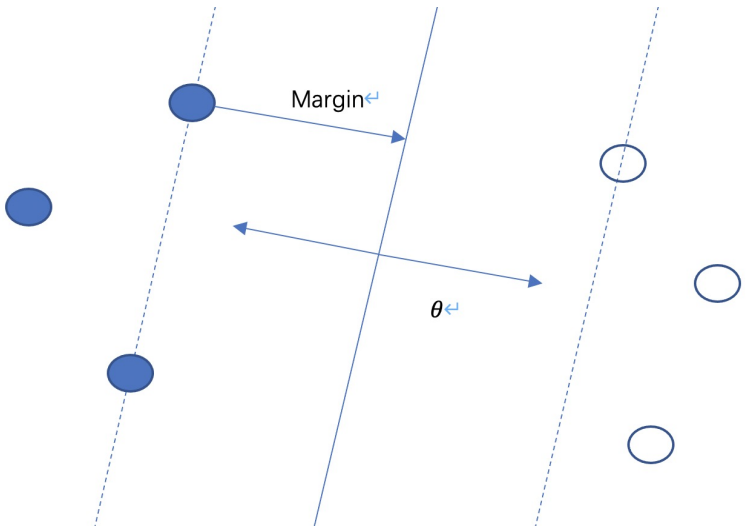

Figure 3: Hard SVM. The solid line is the solution and support vectors are listed are the dashed lines.

In the next section, we will show that if we optimize the problem

$$\min_{\theta} \sum_{i=1}^{n} \ell(\theta^\top z_i)$$

with gradient descent

$$\begin{aligned}\theta(t+1) &\leftarrow \theta(t) - \eta \nabla \mathcal{L}(\theta(t)) \\ &= \theta(t) - \eta \sum_{i=1}^{n} \ell'(\theta(t)^\top z_i) z_i\end{aligned}$$

the direction of the weight will converge to the hard SVM solution

$$\frac{\theta(t)}{\|\theta(t)\|} \to \frac{\theta_{\mathrm{SVM}}}{\|\theta_{\mathrm{SVM}}\|}. \tag{3}$$

4 Main Results

Before we come to our first result, we need some additional assumptions to the loss function $\ell(\theta)$.

Assumption 3. *The loss function ℓ has exponential tail.*

For example, the exponential loss and the logistic loss satisfy this assumption.

Theorem 1 (Theorem 3 from [2])**.** *With Assumption 1-3, small enough step size η and any starting point $\theta(0)$, gradient descent behave as:*

$$\theta(t) = \theta_{\mathrm{SVM}} \log(t) + \rho(t), \tag{4}$$

where the residue ρ grows at most as $\|\rho(t)\| = O(\log \log t)$, and so

$$\lim_{t\to\infty} \frac{\theta(t)}{\|\theta(t)\|} = \frac{\theta_{\mathrm{SVM}}}{\|\theta_{\mathrm{SVM}}\|}. \tag{5}$$

Proof Sketch. For simplicity, we assume that the loss function is $\ell(u) = e^{-u}$ and $\theta(t)^\top z_i \to \infty$ for any i. We also assume that $\frac{\theta(t)}{\|\theta(t)\|} \to \theta_\infty$ for some θ_∞. Then we can write the weight as $\theta(t) = g(t)\theta_\infty + \rho(t)$, where $\lim_{t\to\infty} \frac{\rho(t)}{g(t)} = 0$. Thus, the gradient can be written as:

$$\begin{aligned}-\nabla \mathcal{L}(\theta) &= \sum_{i=1}^{n} \exp(-\theta(t)^\top z_i) z_i \\ &\approx \sum_{i=1}^{n} \exp(-g(t)\theta_\infty^\top z_i) z_i,\end{aligned} \tag{6}$$

where we omitted $\rho(t)$. Note that $g(t) \to \infty$, then only those samples with smallest $\theta_\infty^\top z_i$ will contribute to the gradient. The samples $\mathrm{argmin}_i \theta_\infty^\top z_i$ are exactly the support vectors. Then the negative gradient is dominated by these vectors:

$$-\nabla \mathcal{L}(\theta(t)) \approx \sum_{i\in \mathrm{SV}} \alpha_i' z_i \tag{7}$$

as $t \to \infty$. Thus, the limit θ_∞ will also be dominated by the support vectors:

$$\theta_\infty = \sum_{i \in \mathrm{SV}} \alpha_i'' z_i. \tag{8}$$

We denote the normalized version of θ_∞ by $\hat{\theta} := \frac{\theta_\infty}{\min_i \infty^\top z_i}$. Then we have $\hat{\theta}^\top z_i = 1$ for any $i \in \mathrm{SV}$ and $\hat{\theta} = \sum_{i=1}^n \alpha_i z_i$, where $\alpha_i = 0$ when $i \notin \mathrm{SV}$ and $\alpha_i \neq 0$ when $i \in \mathrm{SV}$. Note that this is precisely the KKT condition of hard-SVM problem. Thus, $\hat{\theta} = \theta_{\mathrm{SVM}}$ and therefore, $\lim_{t\to\infty} \frac{\theta(t)}{\|\theta(t)\|} = \frac{\theta_{\mathrm{SVM}}}{\|\theta_{\mathrm{SVM}}\|}$. □

Theorem 1 implies the following results on the rates of convergence:

Theorem 2 (Theorem 5 from [2])**.** *Under the assumptions of Theorem 1, the normalized weight vector converges to the normalized max-margin vector as*

$$\left\| \frac{\theta(t)}{\|\theta(t)\|} - \frac{\theta_{\mathrm{SVM}}}{\|\theta_{\mathrm{SVM}}\|} \right\| = O\left(\frac{\log\log t}{\log t}\right) \tag{9}$$

while the loss decreases as

$$\mathcal{L}(\theta(t)) = O\left(\frac{1}{t}\right). \tag{10}$$

The difference between the $O(\frac{\log\log t}{\log t})$ and $O(\frac{1}{t})$ implies that the convergence of the loss is much faster than the convergence of weight $\theta(t)$ towards θ_{SVM}. So the continually training with gradient descent improves the generalization, even when the loss is very small.

5 Implicit Bias of GD on Neural Networks

We define a function f to be homogeneous if there exists α such that $f(c\theta) = c^\alpha f(\theta)$ for any $c > 0$. For example, a ReLU neural network f is homogeneous since $f(c\theta) = c^L f(\theta)$, where L is the number of layers.

For a homogeneous neural network, we can define the *normalized margin* similar to linear model:

$$\gamma(\theta) = \min_{i \in [n]} y_i f(\frac{\theta}{\|\theta\|}; x_i). \tag{11}$$

The implicit bias of gradient descent (or gradient flow) on homogeneous neural network is that the normalized margin is approximately monotone increasing during the optimization process. Thus, the normalized weight is roughly converging towards the direction that maximize the normalized margin. A more precise (but still informal) statement is as follows:

Theorem 3 (Theorem 4.1 and 4.2 from [1])**.** *Under certain assumptions, there exists an approximation function $\hat{\gamma}(\theta)$ close to the normalized margin $\gamma(\theta)$ and t_0, such that gradient descent (or gradient flow) satisfies*

1. *For a.e. $t > t_0$, $\hat{\gamma}(\theta(t))$ is monotone increasing;*
2. *$|\hat{\gamma}(\theta(t)) - \gamma(\theta(t))| \to 0$, when $t \to \infty$.*

References

[1] K. Lyu and J. Li. Gradient descent maximizes the margin of homogeneous neural networks. In *International Conference on Learning Representations*, 2020.

[2] D. Soudry, E. Hoffer, M. S. Nacson, S. Gunasekar, and N. Srebro. The implicit bias of gradient descent on separable data. *The Journal of Machine Learning Research*, 19(1):2822–2878, 2018.

[3] C. Zhang, S. Bengio, M. Hardt, B. Recht, and O. Vinyals. Understanding deep learning requires rethinking generalization. In *International Conference on Learning Representations*, 2017.