Chapter 8 Review Notes

1 Stability In Formal Manner

Given a L2-regularised Objective function $F(\mathbf{w}, S) = \sum_{i \in S} l(z^i; \mathbf{w}) + \lambda ||w||^2$ where S is set of training examples and l is any loss function.
Claim: Perturbing one training example K does not change the values of optimal parameters **w*** to large extent w.r.t objective function F(**w**,S) given that training set S is very large.
Assumptions:

- l is convex function i.e. $eig\left(\frac{\partial^2 l}{\partial \mathbf{w}^2}\right) \geq 0$
- $\left|\frac{\partial l}{\partial \mathbf{w}}\right| < B$ where B is any bounding number
- $0 \leq \Lambda_{min} < eig\left[\frac{\partial^2 l}{\partial \mathbf{w}^2}\right] < \Lambda_{max}$

Let's write this claim formally:

$$||w^*(S \cup K) - w^*(S)|| = O(\frac{1}{\lambda T}) \tag{1}$$

Proof: We can write following using taylor series expansion:

$$\begin{aligned} F(\mathbf{w*}(S \cup K), S) = F(\mathbf{w*}(S), S) + \left(\frac{\partial F}{\partial \mathbf{w}}\right)^T (\mathbf{w*}(S \cup K) - \mathbf{w*}(S)) \\ + \frac{1}{2}(\mathbf{w*}(S \cup K) - \mathbf{w*}(S))^T \left[\frac{\partial^2 F}{\partial \mathbf{w}^2}\right] (\mathbf{w*}(S \cup K) - \mathbf{w*}(S)) \end{aligned} \tag{2}$$

Here, the first derivate term is zero and second derivate term will be as following:

$$\min \frac{\partial^2 F}{\partial \mathbf{w}^2} = \min \sum_{i \in S} \left[2\lambda + \frac{\partial^2 l}{\partial \mathbf{w}^2}\right] = 2\lambda |S| \tag{3}$$

Using this (2) and (3), we get:

$$F(\mathbf{w*}(S \cup K), S) - F(\mathbf{w*}(S), S) \geq \lambda |S|.||\mathbf{w*}(S \cup K) - \mathbf{w*}(S)||^2 \tag{4}$$

Now,

$$
\begin{aligned}
&F(\mathbf{w}^*(S \cup K), S) - F(\mathbf{w}^*(S), S) \\
&\quad = F(\mathbf{w}^*(S \cup K), S \cup K) - F(\mathbf{w}^*(S), S \cup K) + F(\mathbf{w}^*(S), K) - F(\mathbf{w}^*(S \cup K), K) \\
&\quad \leq F(\mathbf{w}^*(S), K) - F(\mathbf{w}^*(S \cup K), K) \\
&\quad \text{As the difference of first two term in the equation is negative for minimization function.} \\
&\quad \leq (2\lambda \mathbf{w}_{max} + B\sqrt{d}).||\mathbf{w}^*(S) - \mathbf{w}^*(S \cup K)||_2 \\
&\quad \text{using taylor series and assumption 2}
\end{aligned}
\tag{5}
$$

Using equation (4) and (5):

$$
\begin{aligned}
&\lambda|S|.||\mathbf{w}^*(S \cup K) - \mathbf{w}^*(S)||^2 \leq (2\lambda \mathbf{w}_{max} + B\sqrt{d}).||\mathbf{w}^*(S) - \mathbf{w}^*(S \cup K)|| \\
&\implies ||\mathbf{w}^*(S) - \mathbf{w}^*(S \cup K)|| \leq \frac{2\lambda \mathbf{w}_{max} + B\sqrt{d}}{\lambda|S|}
\end{aligned}
\tag{6}
$$

Hence Proved the claim

Now, Let us find the difference in objective function on perturbing a single point from dataset.

$$
\begin{aligned}
F(\mathbf{w}^*(S \cup K), S \cup K) - F(\mathbf{w}^*(S), S) &\leq F(\mathbf{w}^*(S), S \cup K) - F(\mathbf{w}^*(S), S) \\
&\quad \text{As objective is minimum when we consider S U K} \\
&= F(\mathbf{w}^*(S), K)
\end{aligned}
\tag{7}
$$

So, we can say that F($\mathbf{w}^*$(S),S) is not growing with S as the above term does not depend on the size of the training set.

Additional Reading:

- chapter 13: Regularization and Stability: Understanding Machine Learning Textbook by Shai Ben-David and Shai Shalev-Shwartz