Question 2

Suppose we know that the implicit regularization of optimization tends to flat minima, but not of low norm. Instead, it attempts to produce a solution close to at least one of a finite set of points $\{\theta_1, \dots, \theta_k\}$.

Using PAC-BAYES derive a generalization bound that accounts for the latter implicit regularization.

Solution

For $1 \le i \le k$ we define P_i as the Gaussian distribution $N(\theta_i, \sigma^2 \cdot I)$ ($\sigma^2 \cdot I$ is the same variance to all $\{P_i\}_{i=1}^{k}$)

Now, the distribution Q will be $N(\hat{\theta}, \bar{\sigma}^2 \cdot I)$ where $\hat{\theta} \in \mathbb{R}^r$ are the params returned by the training algorithm and $\bar{\sigma}^2$ is some variance we fix in advance.

By the lemma we proved in question 1:

$$\forall 1 \le i \le k \quad KL(Q \| P_i) = \frac{1}{2}\left(r \cdot \frac{1}{\sigma^2} \cdot \bar{\sigma}^2 + \frac{1}{\sigma^2} \|\hat{\theta} - \theta_i\|^2 - r + r \ln(\sigma^2) - r \ln(\bar{\sigma}^2) \right)$$

As we explained in class, fixing $\hat{\theta}$ and minimizing over $\bar{\sigma}^2$ will yield to $\bar{\sigma}^2 = \sigma^2$ and $Q = N(\hat{\theta}, \sigma^2 \cdot I)$ and:

$$KL(Q \| P_i) = \frac{1}{2\sigma^2} \|\hat{\theta} - \theta_i\|^2$$

Let $\delta \in (0,1)$. By the theorem from class, w.p $\ge 1 - \frac{\delta}{k}$ over $S \sim D^m$:

$$L_D(Q) - L_S(Q) \le \sqrt{\frac{KL(Q \| P_i) + \ln\left(\frac{2m}{(\delta/k)}\right)}{2(m-1)}} = \sqrt{\frac{\frac{1}{2\sigma^2}\|\hat{\theta} - \theta_i\|^2 + \ln\left(\frac{2m}{(\frac{\delta}{k})}\right)}{2(m-1)}}$$

Thus, we've got for every $1 \le i \le k$:

w.p $\ge 1 - \frac{\delta}{k}$ over $S \sim D^m$:

$$L_D(Q) - L_S(Q) \le \sqrt{\frac{\frac{1}{2\sigma^2}\|\hat{\theta}-\theta_i\|^2 + \ln\left(\frac{2mk}{\delta}\right)}{2(m-1)}}$$

$$\Rightarrow \quad \Pr\left(L_D(Q) - L_S(Q) \ge \sqrt{\frac{\frac{1}{2\sigma^2}\|\hat{\theta}-\theta_i\|^2 + \ln\left(\frac{2mk}{\delta}\right)}{2(m-1)}} \right) \le \frac{\delta}{k}$$

$$\Rightarrow \quad \Pr\left(\exists i: L_D(Q) - L_S(Q) \ge \sqrt{\frac{\frac{1}{2\sigma^2}\|\hat{\theta}-\theta_i\|^2 + \ln\left(\frac{2mk}{\delta}\right)}{2(m-1)}} \right)$$

union bound

$$\le \sum_{i=1}^{k} \Pr\left(L_D(Q) - L_S(Q) \ge \sqrt{\frac{\frac{1}{2\sigma^2}\|\hat{\theta}-\theta_i\|^2 + \ln\left(\frac{2mk}{\delta}\right)}{2(m-1)}} \right)$$

$$\le \sum_{i=1}^{k} \frac{\delta}{k} = \delta$$

$\Rightarrow$ w.p $\ge 1 - \delta$ over $S \sim D^m$:

$$L_D(Q) - L_S(Q) \le \min_{1 \le i \le k} \sqrt{\frac{\frac{1}{2\sigma^2}\|\hat{\theta}-\theta_i\|^2 + \ln\left(\frac{2mk}{\delta}\right)}{2(m-1)}}$$

$$= \sqrt{\frac{\frac{1}{2\sigma^2} \cdot \min_{1 \le i \le k}\|\hat{\theta}-\theta_i\|^2 + \ln\left(\frac{2mk}{\delta}\right)}{2(m-1)}}$$

Since $\hat{\theta}$ tends to ~~[illegible]~~ be close to at least one of $\{\theta_i\}_{i=1}^{k}$ we'll get that $\min_{1 \le i \le k}\|\hat{\theta}-\theta_i\|^2$ will be with low value.

In addition by assumption the optimization tends to flat minima which means $L_S(Q)$ tends to be low. Thus, in total the generalization bound on $L_D(Q)$ tends to be low as required.

$$L_D(Q) \le L_S(Q) + \sqrt{\frac{\frac{1}{2\sigma^2} \min_{1 \le i \le k}\|\hat{\theta}-\theta_i\|^2 + \ln\left(\frac{2mk}{\delta}\right)}{2(m-1)}}$$

$\square$