In this lecture, we give another example of margin-sparsity bound involved with mixture-of-experts type of models. Let $\mathcal{H}$ be a set of functions $h_i : \mathcal{X} \to [-1,+1]$ with finite VC dimension. Let $C_1, \ldots, C_m$ be partitions of $\mathcal{H}$ into $m$ clusters $\mathcal{H} = \bigcup_{i=1}^m C_i$. The elements in the convex hull $\text{conv}\mathcal{H}$ takes the form $f = \sum_{i=1}^T \lambda_i h_i = \sum_{c \in \{C_1, \ldots, C_m\}} \alpha_c \sum_{h \in \mathcal{H}} \lambda_h^c \cdot h$, where $T \gg m$, $\sum_i \lambda_i = 1$, $\alpha_c = \sum_{h \in \mathcal{H}} \lambda_h$, and $\lambda_h^c = \lambda_h/\alpha_c$ for $h \in c$. We can approximate $f$ by $g$ as follows. For each cluster $c$, let $\{Y_k^c\}_{k=1,\ldots,N}$ be random variables such that $\forall h \in c \subset \mathcal{H}$, we have $\mathbb{P}(Y_k^c = h) = \lambda_h^c$. Then $\mathbb{E}Y_k^c = \sum_{h \in \mathcal{H}} \lambda_h^c \cdot h$. Let $Z_k = \sum_c \alpha_c Y_k^c$ and $g = \sum_c \alpha_c \frac{1}{N} \sum_{k=1}^N Y_k^c = \frac{1}{N} \sum_{k=1}^N Z_k$. Then $\mathbb{E}Z_k = \mathbb{E}g = f$. We define $\sigma_c^2 \triangleq \text{var}(Z_k) = \sum_c \alpha_c^2 \text{var}(Y_k^c)$, where $\text{var}(Y_k^c) = \|Y_k^c - \mathbb{E}Y_k^c\|^2 = \sum_{h \in \mathcal{H}} \lambda_h^c (h - \mathbb{E}Y_k^c)^2$. (If we define $\{Y_k\}_{k=1,\ldots,N}$ be random variables such that $\forall h \in \mathcal{H}$, $\mathbb{P}(Y_k = h) = \lambda_h$, and define $g = \frac{1}{N} \sum_{k=1}^N Y_k$, we might get much larger $\text{var}(Y_k)$).

\begin{align*}
\text{C}_2 & \quad \cdots \quad \text{h} \\
\text{C}_1 & \quad \cdots \quad \text{C}_m
\end{align*}

Recall that a classifier takes the form $y = \text{sign}(f(x))$ and a classification error corresponds to $yf(x) < 0$. We can bound the error by

$$\mathbb{P}(yf(x) < 0) \leq \mathbb{P}(yg \leq \delta) + \mathbb{P}(\sigma_c^2 > r) + \mathbb{P}(yg > \delta|yf(x) \leq 0, \sigma_c^2 < r).$$

The third term on the right side of inequality 24.1 can be bounded in the following way,

$$\mathbb{P}(yg > \delta|yf(x) \leq 0, \sigma_c^2 < r) = \mathbb{P}\left(\frac{1}{N} \sum_{k=1}^N (yZ_k - \mathbb{E}yZ_k) > \delta - yf(x) | yf(x) \leq 0, \sigma_c^2 < r\right)$$

$$\leq \mathbb{P}\left(\frac{1}{N} \sum_{k=1}^N (yZ_k - \mathbb{E}yZ_k) > \delta | yf(x) \leq 0, \sigma_c^2 < r\right)$$

$$\leq \exp\left(-\frac{N^2 \delta^2}{2N\sigma_c^2 + \frac{2}{3}N\delta \cdot 2}\right),\text{Bernstein’s inequality}$$

$$\leq \exp\left(-\min\left(\frac{N^2 \delta^2}{4N\sigma_c^2}, \frac{N^2 \delta^2}{\frac{4}{3}N\delta}\right)\right)$$

$$\leq \exp\left(-\frac{N\delta^2}{4r}\right),\text{for } r \text{ small enough}$$

$$\leq \frac{1}{n}.$$  

(24.2)

As a result, $\forall N \geq \frac{10}{d^2} \log n$, inequality 24.2 is satisfied.

To bound the first term on the right side of inequality 24.1, we note that $\mathbb{E}Y_1,\ldots,Y_N \mathbb{P}(yg \leq \delta) \leq \mathbb{E}Y_1,\ldots,Y_N \mathbb{E}\phi_\delta(yg)$ and $\mathbb{E}_n \phi_\delta(yg) \leq \mathbb{P}_n(yg \leq 2\delta)$ for some $\phi_\delta$. 

61
Any realization of \( g = \sum_{k=1}^{N_m} Z_k \), where \( N_m \) depends on the number of clusters \( (C_1, \ldots, C_m) \), is a linear combination of \( h \in \mathcal{H} \), and \( g \in \text{conv}_N \mathcal{H} \). According to lemma 20.2,

\[
(E \phi_\delta(yg) - E_n \phi_\delta(yg)) / \sqrt{E \phi_\delta(yg)} \leq K \left( \sqrt{V N_m \log \frac{n}{\delta} / n} + \frac{u}{n} \right)
\]

with probability at least \( 1 - e^{-u} \). Using a technique developed earlier in this course, and taking the union bound over all \( m, \delta \), we get, with probability at least \( 1 - Ke^{-u} \),

\[
\mathbb{P}(yg \leq \delta) \leq K \inf_{m,\delta} \left( \mathbb{P}_n(yg \leq 2\delta) + \frac{V \cdot N_m}{n} \log \frac{n}{\delta} + \frac{u}{n} \right).
\]

(Since \( \mathbb{E} \mathbb{P}_n(yg \leq 2\delta) \leq \mathbb{E} \mathbb{P}_n(yf(x) \leq 3\delta) + \mathbb{E} \mathbb{P}_n(\sigma_c^2 \geq r) + \frac{1}{n} \) with appropriate choice of \( N \), based on the same reasoning as inequality 24.1, we can also control \( \mathbb{P}_n(yg \leq 2\delta) \) by \( \mathbb{P}_n(yf \leq 3\delta) \) and \( \mathbb{P}_n(\sigma_c^2 \geq r) \) probabilistically).

To bound the second term on the right side of inequality 24.1, we approximate \( \sigma_c^2 \) by

\[
\sigma_N^2 = \frac{1}{N} \sum_{k=1}^{N} \frac{1}{2} \left( Z_k^{(1)} - Z_k^{(2)} \right)^2 \text{ where } Z_k^{(1)} \text{ and } Z_k^{(2)} \text{ are independent copies of } Z_k.
\]

We have

\[
\mathbb{E}_{Y_{1,\ldots,N}} \sigma_N^2 = \sigma_c^2
\]

\[
\text{var}_{Y_{1,\ldots,N}} \frac{1}{2} \left( Z_k^{(1)} - Z_k^{(2)} \right)^2 = \frac{1}{4} \text{var} \left( Z_k^{(1)} - Z_k^{(2)} \right)^2
\]

\[
\leq \frac{1}{4} \mathbb{E} \left( Z_k^{(1)} - Z_k^{(2)} \right)^4
\]

\[
\leq \mathbb{E} \left( Z_k^{(1)} - Z_k^{(2)} \right)^2
\]

\[
= \frac{2}{\sigma_c^2}
\]

\[
\text{var}_{Y_{1,\ldots,N}} \sigma_N^2 \leq 2 \cdot \sigma_c^2.
\]
We start with

\[ P_{Y_1, \ldots, N} (\sigma_c^2 \geq 4r) \leq P_{Y_1, \ldots, N} (\sigma_N^2 \geq 3r) + P_{Y_1, \ldots, N} (\sigma_N^2 \geq 4r | \sigma_N^2 \leq 3r) \]

\[ \leq E_{Y_1, \ldots, N} \phi_r (\sigma_N^2 \geq 3r) + \frac{1}{n} \]

with appropriate choice of \( N \), following the same line of reasoning as in inequality 24.1. We note that

\[ P_{Y_1, \ldots, N} (\sigma_N^2 \geq 3r) \leq E_{Y_1, \ldots, N} \phi_r (\sigma_N^2), \]

and

\[ E_n \phi_\delta (\sigma_N^2) \leq P_n (\sigma_N^2 \geq 2r) \]

for some \( \phi_\delta \).

Since

\[ \sigma_N^2 \in \left\{ \frac{1}{2N} \sum_{k=1}^{N} \left( \sum_c \alpha_c (h_{k,c}^{(1)} - h_{k,c}^{(2)}) \right)^2 : h_{k,c}^{(1)}, h_{k,c}^{(2)} \in \mathcal{H} \right\} \subset \text{conv}_{N_m} \{ h_i \cdot h_j : h_i, h_j \in \mathcal{H} \}, \]

and \( \log D(\{ h_i \cdot h_j : h_i, h_j \in \mathcal{H} \}, \varepsilon) \leq KV \log \frac{2}{r} \) by the assumption of our problem, we have \( \log D(\text{conv}_{N_m} \{ h_i \cdot h_j : h_i, h_j \in \mathcal{H} \}, \varepsilon) \leq KV \cdot N_m \cdot \log \frac{2}{r} \) by the VC inequality, and

\[ \frac{(E\phi_r (\sigma_N^2)) - E_n \phi_r (\sigma_N^2)}{\sqrt{E\phi_r (\sigma_N^2)}} \leq K \left( \sqrt{V \cdot N_m \log \frac{n}{r} / n + \sqrt{\frac{u}{n}}} \right) \]

with probability at least \( 1 - e^{-u} \). Using a technique developed earlier in this course, and taking the union bound over all \( m, \delta, r \), with probability at least \( 1 - Ke^{-u} \),

\[ P(\sigma_c^2 \geq 4r) \leq K \inf_{m, \delta, r} \left( P_n (\sigma_N^2 \geq 2r) + \frac{1}{n} + \frac{V \cdot N_m \log \frac{n}{\delta} + \frac{u}{n}}{n} \right) \]

As a result, with probability at least \( 1 - Ke^{-u} \), we have

\[ P(yf(x) \leq 0) \leq K \inf_{r, \delta, m} \left( P_n (yg \leq 2 \cdot \delta) + P_n (\sigma_N^2 \geq r) + \frac{V \cdot \min(r_m / \delta^2, N_m)}{n} \log \frac{n}{\delta} \log n + \frac{u}{n} \right) \]

for all \( f \in \text{conv}\mathcal{H} \).