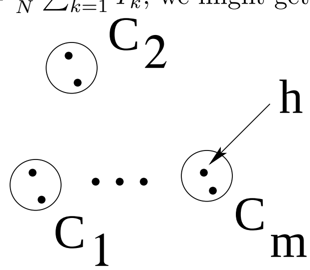
In this lecture, we give another example of margin-sparsity bound involved with mixture-of-experts type of models. Let $\mathcal{H}$ be a set of functions $h_i : \mathcal{X} \to [-1,+1]$ with finite VC dimension. Let $C_1, \cdots, C_m$ be partitions of $\mathcal{H}$ into m clusters $\mathcal{H} = \bigcup_{i=1}^m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



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

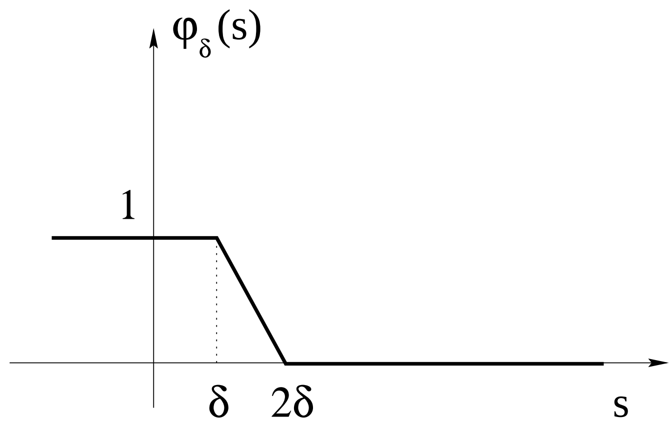
$$\mathbb{P}(yf(x) < 0) \le \mathbb{P}(yg \le \delta) + \mathbb{P}(\sigma_c^2 > r) + \mathbb{P}(yg > \delta | yf(x) \le 0, \sigma_c^2 < r). \tag{24.1}$$

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

$$\begin{aligned}
\mathbb{P}(yg > \delta | yf(x) \le 0, \sigma_c^2 < r) &= \mathbb{P}\left(\frac{1}{N}\sum_{k=1}^N (yZ_k - \mathbb{E}yZ_k) > \delta - yf(x) | yf(x) \le 0, \sigma_c^2 < r\right) \\
&\le \mathbb{P}\left(\frac{1}{N}\sum_{k=1}^N (yZ_k - \mathbb{E}yZ_k) > \delta | yf(x) \le 0, \sigma_c^2 < r\right) \\
&\le \exp\left(-\frac{N^2\delta^2}{2N\sigma_c^2 + \frac{2}{3}N\delta\cdot 2}\right) \text{ ,Bernstein's inequality} \\
&\le \exp\left(-\min\left(\frac{N^2\delta^2}{4N\sigma_c^2}, \frac{N^2\delta^2}{\frac{8}{3}N\delta}\right)\right) \\
&\le \exp\left(-\frac{N\delta^2}{4r}\right) \text{ , for } r \text{ small enough} \\
&\stackrel{set}{\le} \frac{1}{n}. \qquad (24.2)
\end{aligned}$$

As a result, $\forall N \ge \frac{4\cdot r}{\delta^2}\log n$, inequality 24.2 is satisfied.

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



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

$$(\mathbb{E}\phi_\delta(yg) - \mathbb{E}_n\phi_\delta(yg))\,/\sqrt{\mathbb{E}\phi_\delta(yg)} \quad \leq \quad K\left(\sqrt{VN_m \log\frac{n}{\delta}/n} + \sqrt{u/n}\right)$$

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

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

(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).

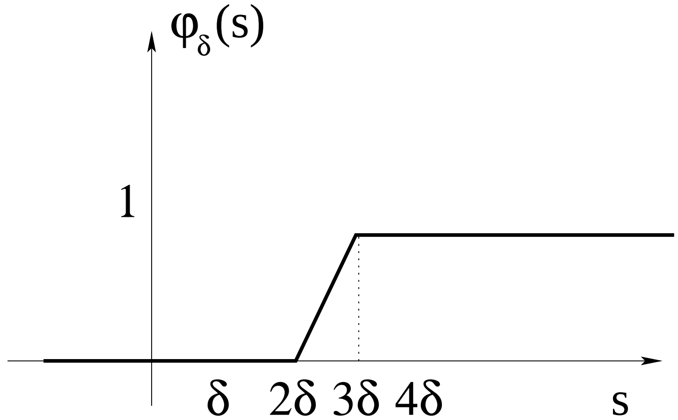
To bound the second term on the right side of inequality 24.1, we approximate σ_c^2 by $\sigma_N^2 = \frac{1}{N}\sum_{k=1}^{N}\frac{1}{2}\left(Z_k^{(1)} - Z_k^{(2)}\right)^2$ where $Z_k^{(1)}$ and $Z_k^{(2)}$ are independent copies of Z_k . We have

$$\begin{aligned}
\mathbb{E}_{Y_{1,\cdots,N}^{(1,2)}}\sigma_N^2 &= \sigma_c^2 \\
\text{var}_{Y_{1,\cdots,N}^{(1,2)}}\frac{1}{2}\left(Z_k^{(1)} - Z_k^{(2)}\right)^2 &= \frac{1}{4}\text{var}\left(Z_k^{(1)} - Z_k^{(2)}\right)^2 \\
&\leq \frac{1}{4}\mathbb{E}\left(Z_k^{(1)} - Z_k^{(2)}\right)^4 \\
&\quad \left(-1 \leq Z_k^{(1)}, Z_k^{(2)} \leq 1 \text{ ,and } \left(Z_k^{(1)} - Z_k^{(2)}\right)^2 \leq 4\right) \\
&\leq \mathbb{E}\left(Z_k^{(1)} - Z_k^{(2)}\right)^2 \\
&= 2\sigma_c^2 \\
\text{var}_{Y_{1,\cdots,N}^{(1,2)}}\sigma_N^2 &\leq 2 \cdot \sigma_c^2.
\end{aligned}$$

We start with

$$\begin{aligned}\mathbb{P}_{Y_{1,\cdots,N}}(\sigma_c^2 \geq 4r) &\leq \mathbb{P}_{Y^{(1,2)}_{1,\cdots,N}}(\sigma_N^2 \geq 3r) + \mathbb{P}_{Y^{(1,2)}_{1,\cdots,N}}(\sigma_c^2 \geq 4r | \sigma_N^2 \leq 3r) \\ &\leq \mathbb{E}_{Y^{(1,2)}_{1,\cdots,N}} \phi_r\left(\sigma_N^2 \geq 3r\right) + \frac{1}{n}\end{aligned}$$

with appropriate choice of N, following the same line of reasoning as in inequality 24.1. We note that $\mathbb{P}_{Y_1,\cdots,Y_N}(\sigma_N^2 \geq 3r) \leq \mathbb{E}_{Y_1,\cdots,Y_N}\phi_r(\sigma_N^2)$, and $\mathbb{E}_n\phi_\delta(\sigma_N^2) \leq \mathbb{P}_n(\sigma_N^2 \geq 2r)$ for some ϕ_δ.



Since

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

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

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

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

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

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

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

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