Let $\mathcal{H}$ be a class of "simple" functions (VC-subgraph, perceptrons). Define recursively

$$\mathcal{H}_{i+1} = \left\{\sigma\left(\sum \alpha_j h_j\right) : \ h_j \in \mathcal{H}_i, \ \alpha_j \in \mathbb{R}\right\}$$

where σ is sigmoid function such that $\sigma(0) = 0$ and $|\sigma(s) - \sigma(t)| \le L|s - t|$, $-1 \le \sigma \le 1$.

Example:

$$\sigma(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}}.$$

Assume we have data $(x_1, y_1), \ldots, (x_n, y_n)$, $-1 \le y_i \le 1$. We can minimize

$$\frac{1}{n}\sum_{i=1}^n (y_i - h(x_i))^2$$

over $\mathcal{H}_k$, where k is the number of layers.

Define $\mathcal{L}(y, h(x)) = (y - h(x))^2$, $0 \le \mathcal{L}(y, h(x)) \le 4$. We want to bound $\mathbb{E}\mathcal{L}(y, h(x))$.

From the previous lectures,

$$\sup \left|\mathbb{E}\mathcal{L}(y, h(x)) - \frac{1}{n}\sum_{i=1}^n \mathcal{L}(y_i, h(x_i))\right| \le 2\mathbb{E}\sup\left|\frac{1}{n}\sum_{i=1}^n \varepsilon_i \mathcal{L}(y_i, h(x_i))\right| + 4\sqrt{\frac{2t}{n}}$$

with probability at least $1 - e^{-t}$.

Define

$$\mathcal{H}_{i+1}(A_1, \ldots, A_{i+1}) = \left\{\sigma\left(\sum \alpha_j h_j\right) : \ \sum |\alpha_j| \le A_{i+1}, \ h_j \in \mathcal{H}_i\right\}.$$

For now, assume bounds A_i on sum of weights (although this is not true in practice, so we will take union bound later).

Theorem 28.1.

$$\mathbb{E}\sup_{h\in\mathcal{H}_k(A_1,\ldots,A_k)}\left|\frac{1}{n}\sum_{i=1}^n \varepsilon_i \mathcal{L}(y_i, h(x_i))\right| \le 8\prod_{j=1}^k (2L\cdot A_j)\cdot\mathbb{E}\sup_{h\in\mathcal{H}}\left|\frac{1}{n}\sum_{i=1}^n \varepsilon_i h(x_i)\right| + \frac{8}{\sqrt{n}}.$$

Proof. Since $-2 \le y - h(x) \le 2$, $\frac{(y-h(x))^2}{4} : [-2, 2] \mapsto \mathbb{R}$ is a contraction because largest derivative of s^2 on $[-2, 2]$ is 4. Hence,

$$\begin{aligned}
\mathbb{E}\sup_{h\in\mathcal{H}_k(A_1,\ldots,A_k)}\left|\frac{1}{n}\sum_{i=1}^n \varepsilon_i (y_i - h(x_i))^2\right| &= \mathbb{E}\mathbb{E}_\varepsilon \sup_{h\in\mathcal{H}_k(A_1,\ldots,A_k)}\left|\frac{1}{n}\sum_{i=1}^n \varepsilon_i (y_i - h(x_i))^2\right| \\
&= 4\mathbb{E}\mathbb{E}_\varepsilon \sup_{h\in\mathcal{H}_k(A_1,\ldots,A_k)}\left|\frac{1}{n}\sum_{i=1}^n \varepsilon_i \frac{(y_i - h(x_i))^2}{4}\right| \\
&\le 8\mathbb{E}\mathbb{E}_\varepsilon \sup_{h\in\mathcal{H}_k(A_1,\ldots,A_k)}\left|\frac{1}{n}\sum_{i=1}^n \varepsilon_i (y_i - h(x_i))\right| \\
&\le 8\mathbb{E}\left|\frac{1}{n}\sum_{i=1}^n \varepsilon_i y_i\right| + 8\mathbb{E}\sup_{h\in\mathcal{H}_k(A_1,\ldots,A_k)}\left|\frac{1}{n}\sum_{i=1}^n \varepsilon_i h(x_i)\right|
\end{aligned}$$

Furthermore,

$$\begin{aligned}
\mathbb{E}\left|\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i y_i\right| &\le \left(\mathbb{E}\left(\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i y_i\right)^2\right)^{1/2} \\
&= \left(\mathbb{E}\sum_{i=1}^{n}\frac{1}{n^2}\varepsilon_i^2 y_i^2\right)^{1/2} \\
&= \left(\frac{1}{n}\mathbb{E}y_1^2\right)^{1/2} \le \sqrt{\frac{1}{n}}
\end{aligned}$$

Using the fact that σ/L is a contraction,

$$\begin{aligned}
\mathbb{E}_\varepsilon \sup_{h\in\mathcal{H}_k(A_1,\ldots,A_k)}\left|\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i\sigma\left(\sum\alpha_j h_j(x_i)\right)\right| &= L\mathbb{E}_\varepsilon\sup_h\left|\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i\frac{\sigma}{L}\left(\sum\alpha_j h_j(x_i)\right)\right| \\
&\le 2L\mathbb{E}_\varepsilon\sup_h\left|\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i\left(\sum\alpha_j h_j(x_i)\right)\right| \\
&= 2L\mathbb{E}_\varepsilon\sup_h\left|\frac{1}{n}\sum_j\alpha_j\left(\sum_{i=1}^{n}\varepsilon_i h_j(x_i)\right)\right| \\
&= 2L\mathbb{E}_\varepsilon\sup_h\left|\frac{\sum|\alpha_j|}{n}\sum_j\alpha_j'\left(\sum_{i=1}^{n}\varepsilon_i h_j(x_i)\right)\right|
\end{aligned}$$

where $\alpha_j' = \frac{\alpha_j}{\sum_j|\alpha_j|}$. Since $\sum_j|\alpha_j|\le A_k$ for the layer k,

$$\begin{aligned}
&2L\mathbb{E}_\varepsilon\sup_{h\in\mathcal{H}_k(A_1,\ldots,A_k)}\left|\frac{\sum|\alpha_j|}{n}\sum_j\alpha_j'\left(\sum_{i=1}^{n}\varepsilon_i h_j(x_i)\right)\right| \\
&\le 2LA_k\mathbb{E}_\varepsilon\sup_{h\in\mathcal{H}_k(A_1,\ldots,A_k)}\left|\frac{1}{n}\sum_j\alpha_j'\left(\sum_{i=1}^{n}\varepsilon_i h_j(x_i)\right)\right| \\
&= 2LA_k\mathbb{E}_\varepsilon\sup_{h\in\mathcal{H}_{k-1}(A_1,\ldots,A_{k-1})}\left|\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i h_j(x_i)\right|
\end{aligned}$$

The last equality holds because $\sup|\sum\lambda_j s_j| = \max_j|s_j|$, i.e. max is attained at one of the vertices.

By induction,

$$\mathbb{E}\sup_{h\in\mathcal{H}_k(A_1,\ldots,A_k)}\left|\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i(y_i-h(x_i))^2\right| \le 8\prod_{j=1}^{k}(2LA_j)\cdot\mathbb{E}\sup_{h\in\mathcal{H}}\left|\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i h(x_i)\right| + \frac{8}{\sqrt{n}},$$

where $\mathcal{H}$ is the class of simple classifiers. $\square$