Let $\mathcal{F} \subset \{f \in [0,1]\}$ be a class of $[0,1]$ valued functions, $Z = \sup_f \left(\mathbb{E}f - \frac{1}{n}\sum_{i=1}^n f(x_i)\right)$, and $R = \sup_f \sum_f \frac{1}{n}\sum_{i=1}^n \epsilon_i f(x_i)$ for any given $x_i, \cdots, x_n$ where $\epsilon_1, \cdots \epsilon_n$ are Rademarcher random variables. For any $f \in \mathcal{F}$ unknown and to be estimated, the empirical error Z can be probabilistically bounded by R in the following way. Using the fact that $Z \leq 2R$ and by Martingale inequality, $\mathbb{P}\left(Z \leq \mathbb{E}Z + \sqrt{\frac{2u}{n}}\right) \geq 1 - e^{-u}$, and $\mathbb{P}\left(\mathbb{E}R \leq R + 2\sqrt{\frac{2u}{n}}\right) \geq 1 - e^{-u}$. Taking union bound, $\mathbb{P}\left(Z \leq R + 5\sqrt{\frac{2u}{n}}\right) \geq 1 - 2e^{-u}$. Taking union bound again over all $(n_k)_{k>1}$ and let $\epsilon = 5\sqrt{\frac{2u}{n}}$, $\mathbb{P}\left(\forall n \in (n_k)_{k\geq 1} \forall f \in \mathcal{F}, Z \leq 2R + \epsilon\right) \geq 1 - \exp\left(-\sum_k \frac{n_k \epsilon^2}{50}\right) \overset{\text{set}}{\geq} 1 - \delta$. Using big O notation, $n_k = \mathcal{O}\left(\frac{1}{\epsilon^2}\log\frac{1}{\delta^2}\right)$.

For voting algorithms, the candidate function to be estimated is a symmetric convex combination of some base functions $\mathcal{F} = \text{conv}\mathcal{H}$, where $\mathcal{H} \subset \{h \in [0,1]\}$. The trained classifier is $\text{sign}(f(x))$ where $f \in \mathcal{F}$ is our estimation, and the training error is $\mathbb{P}(yf(x))$. The training error can be bounded as the following,

$$
\begin{aligned}
\mathbb{P}(yf(x) < 0) &\leq \mathbb{E}\phi_\delta(yf(x)) \\
&\leq \mathbb{E}_n\phi_\delta(yf(x)) + \underbrace{\sup_{f\in\mathcal{F}}(\mathbb{E}\phi_\delta(yf(x)) - \frac{1}{n}\sum_{i=1}^n \phi_\delta(y_i f(x_i)))}_{Z} \\
&\underbrace{\leq}_{\text{with probability } 1-e^{-u}} \mathbb{E}_n\phi_\delta(yf(x)) + 2\cdot\mathbb{E}\underbrace{\sup_{f\in\mathcal{F}}(\frac{1}{n}\sum_{i=1}^n \epsilon_i\phi_\delta(y_i f(x_i)))}_{R} + \sqrt{\frac{2u}{n}} \\
&\underbrace{\leq}_{\text{contraction}} \mathbb{E}_n\phi_\delta(yf(x)) + \frac{2}{\delta}\mathbb{E}\sup_{f\in\mathcal{F}}\frac{1}{n}\sum_{i=1}^n \epsilon_i y_i f(x_i) + \sqrt{\frac{2u}{n}} \\
&= \mathbb{E}_n\phi_\delta(yf(x)) + \frac{2}{\delta}\mathbb{E}\sup_{f\in\mathcal{F}}\frac{1}{n}\sum_{i=1}^n \epsilon_i f(x_i) + \sqrt{\frac{2u}{n}} \\
&\leq \mathbb{P}_n(yf(x) < 0) + \frac{2}{\delta}\mathbb{E}\sup_{h\in\mathcal{H}}\frac{1}{n}\sum_{i=1}^n \epsilon_i h(x_i) + \sqrt{\frac{2u}{n}}.
\end{aligned}
$$

To bound the second term $(\mathbb{E}\sup_{h\in\mathcal{H}}\frac{1}{n}\sum_{i=1}^n \epsilon_i h(x_i))$ above, we will use the following fact.

Fact 27.1. *If $\mathbb{P}(\xi \geq a + b\cdot t) \leq \exp(-t^2)$, then $\mathbb{E}\xi \leq K\cdot(a+b)$ for some constant K.*

If $\mathcal{H}$ is a VC-subgraph class and V is its VC dimension, $D(\mathcal{H}, \epsilon, d_x) \leq K\left(\frac{1}{\epsilon}\right)^{2\cdot V}$ by D. Haussler. By Kolmogorov's chaining method (Lecture 14),

$$
\begin{aligned}
&= \mathbb{P}\left(\sup_h \frac{1}{n}\sum_{i=1}^n \epsilon_i h(x_i) \leq K\left(\frac{1}{n}\int_0^1 \log^{1/2} D(\mathcal{H}, \epsilon, d_x)d\epsilon + \sqrt{\frac{u}{n}}\right)\right) \\
&= \mathbb{P}\left(\sup_h \frac{1}{n}\sum_{i=1}^n \epsilon_i h(x_i) \leq K\left(\frac{1}{n}\int_0^1 \sqrt{V\log\frac{1}{\epsilon}}d\epsilon + \sqrt{\frac{u}{n}}\right)\right) \\
&\geq 1 - e^{-u}.
\end{aligned}
$$

Thus $\mathbb{E}\sup \frac{1}{n}\sum \epsilon_i h(x_i) \le K\left(\sqrt{\frac{V}{n}} + \sqrt{\frac{1}{n}}\right) \le K\sqrt{\frac{V}{n}}$, and

$$\mathbb{P}\left(\mathbb{P}(yf(x)<0) \le \mathbb{P}_n(yf(x)<0) + K\frac{1}{\delta}\sqrt{\frac{V}{n}} + \sqrt{\frac{2u}{n}}\right) \quad \ge \quad 1 - e^{-u}.$$

Recall our set up for Martingale inequalities. Let $Z = Z(x_1, \cdots, x_n)$ where $x_1, \cdots, x_n$ are independent random variables. We need to bound $Z - \mathbb{E}Z$. Since Z is not a sum of independent random variables, certain classical concentration inequalities is not applicable. But we can try to bound $Z - \mathbb{E}Z$ with certain form of Martingale inequalities.

$$\begin{aligned} Z - \mathbb{E}Z \quad &= \quad \underbrace{Z - \mathbb{E}_{x_1}(Z|x_2,\cdots,x_n)}_{d_1(x_1,\cdots,x_n)} + \underbrace{\mathbb{E}_{x_1}(Z|x_2,\cdots,x_n) - \mathbb{E}_{x_1,x_2}(Z|x_3,\cdots,x_n)}_{d_2(x_2,\cdots,x_n)} + \\ &\quad \cdots + \underbrace{\mathbb{E}_{x_1,\cdots,x_{n-1}}(Z|x_n) - \mathbb{E}_{x_1,\cdots,x_n}(Z)}_{d_n(x_n)} \end{aligned}$$

with the assumptions that $\mathbb{E}_{x_i} d_i = 0$, and $\|d_i\|_\infty \le c_i$.

We will give a generalized martingale inequality below. $\sum_{i=1}^n d_i = Z - \mathbb{E}Z$ where $d_i = d_i(x_i, \cdots, x_n)$, $\max_i \|d_i\|_\infty \le C$, $\sigma_i^2 = \sigma_i^2(x_{i+1}, \cdots, x_n) = \mathrm{var}(d_i)$, and $\mathbb{E}d_i = 0$. Take $\epsilon > 0$,

$$\begin{aligned} &\mathbb{P}(\sum_{i=1}^n d_i - \epsilon\sum_{i=1}^n \sigma_i^2 \ge t) \\ \le \quad & e^{-\lambda t}\mathbb{E}\exp(\sum_{i=1}^n \lambda(d_i - \epsilon\sigma_i^2)) \\ = \quad & e^{-\lambda t}\mathbb{E}\exp(\sum_{i=1}^{n-1} \lambda(d_i - \epsilon\sigma_i^2) \cdot \mathbb{E}\exp(\lambda d_n) \cdot \exp(\lambda\epsilon\sigma_n^2) \end{aligned}$$

The term $\exp(\lambda d_n)$ can be bounded in the following way.

$$\begin{aligned} & \mathbb{E}\exp(\lambda d_n) \\ \underbrace{=}_{\text{Taylor expansion}} \quad & \mathbb{E}\left(1 + \lambda d_n + \frac{\lambda^2}{2!}d_n^2 + \frac{\lambda^3}{3!}d_n^3 + \cdots\right) \\ \le \quad & 1 + \frac{\lambda^2}{2}\sigma_n^2 \cdot \left(1 + \frac{\lambda C}{3} + \frac{\lambda^2 C^2}{3\cdot 4} + \cdots\right) \\ \le \quad & \exp\left(\frac{\lambda^2\cdot\sigma_n^2}{2}\cdot\frac{1}{(1-\lambda C)}\right). \end{aligned}$$

Choose λ such that $\frac{\lambda^2}{2\cdot(1-\lambda C)} \le \lambda\epsilon$, we get $\lambda \le \frac{2\epsilon}{1+2\epsilon C}$, and $\mathbb{E}_{d_n}\exp(\lambda d_n)\cdot\exp(\lambda\epsilon\sigma_n^2) \le 1$. Iterate over $i = n, \cdots, 1$, we get

$$\mathbb{P}\left(\sum_{i=1}^n d_i - \epsilon\sum_{i=1}^n \sigma_i^2 \ge t\right) \quad \le \quad \exp\left(-\lambda\cdot t\right)$$

. Take $t = u/\lambda$, we get

$$\mathbb{P}\left(\sum_{i=1}^{n} d_i \geq \epsilon \sum_{i=1}^{n} \sigma_i^2 + \frac{u}{2\epsilon}(1 + 2\epsilon C)\right) \quad \leq \quad \exp\left(-u\right)$$

To minimize the sum $\epsilon \sum_{i=1}^{n} \sigma_i^2 + \frac{u}{2\epsilon}(1 + 2\epsilon C)$, we set its derivative to 0, and get $\epsilon = \sqrt{\frac{u}{2\sum \sigma_i^2}}$. Thus

$$\mathbb{P}\left(\sum d_i \geq 3\sqrt{u \sum_i \sigma_i^2/2} + Cu\right) \quad \leq \quad e^{-u}$$

. This inequality takes the form of the Bernstein's inequality.