Last time we proved the Pessimistic VC inequality:

$$\mathbb{P}\left(\sup_C \left|\frac{1}{n}\sum_{i=1}^n I(X_i \in C) - \mathbb{P}(C)\right| \geq t\right) \leq 4\left(\frac{2en}{V}\right)^V e^{-\frac{nt^2}{8}},$$

which can be rewritten with

$$t = \sqrt{\frac{8}{n}\left(\log 4 + V\log\frac{2en}{V} + u\right)}$$

as

$$\mathbb{P}\left(\sup_C \left|\frac{1}{n}\sum_{i=1}^n I(X_i \in C) - \mathbb{P}(C)\right| \leq \sqrt{\frac{8}{n}\left(\log 4 + V\log\frac{2en}{V} + u\right)}\right) \geq 1 - e^{-u}.$$

Hence, the rate is $\sqrt{\frac{V\log n}{n}}$. In this lecture we will prove Optimistic VC inequality, which will improve on this rate when $\mathbb{P}(C)$ is small.

As before, we have pairs (X_i, Y_i), $Y_i = \pm 1$. These examples are labeled according to some unknown C_0 such that $Y = 1$ if $X = C_0$ and $Y = 0$ if $X \notin C_0$.

Let $\mathcal{C} = \{C : C \subseteq \mathcal{X}\}$, a set of classifiers. C makes a mistake if

$$X \in C \setminus C_0 \cup C_0 \setminus C = C \triangle C_0.$$

Similarly to last lecture, we can derive bounds on

$$\sup_C \left|\frac{1}{n}\sum_{i=1}^n I(X_i \in C\triangle C_0) - \mathbb{P}(C\triangle C_0)\right|,$$

where $\mathbb{P}(C\triangle C_0)$ is the generalization error.

Let $\mathcal{C}' = \{C\triangle C_0 : C \in \mathcal{C}\}$. One can prove that $VC(\mathcal{C}') \leq VC(\mathcal{C})$ and $\triangle_n(C', X_1, \ldots, X_n) \leq \triangle_n(C, X_1, \ldots, X_n)$.

By Hoeffding-Chernoff, if $\mathbb{P}(C) \leq \frac{1}{2}$,

$$\mathbb{P}\left(\mathbb{P}(C) - \frac{1}{n}\sum_{i=1}^n I(X_i \in C) \leq \sqrt{\frac{2\mathbb{P}(C)\,t}{n}}\right) \geq 1 - e^{-t}.$$

Theorem 11.1 (Optimistic VC inequality)**.**

$$\mathbb{P}\left(\sup_C \frac{\mathbb{P}(C) - \frac{1}{n}\sum_{i=1}^n I(X_i \in C)}{\sqrt{\mathbb{P}(C)}} \geq t\right) \leq 4\left(\frac{2en}{V}\right)^V e^{-\frac{nt^2}{4}}.$$

Proof. Let C be fixed. Then

$$\mathbb{P}_{(X_i')}\left(\frac{1}{n}\sum_{i=1}^{n} I(X_i' \in C) \geq \mathbb{P}(C)\right) \geq \frac{1}{4}$$

whenever $\mathbb{P}(C) \geq \frac{1}{n}$. Indeed, $\mathbb{P}(C) \geq \frac{1}{n}$ since $\sum_{i=1}^{n} I(X_i' \in C) \geq n\mathbb{P}(C) \geq 1$. Otherwise $\mathbb{P}\left(\sum_{i=1}^{n} I(X_i' \in C) = 0\right) = \prod_{i=1}^{n} \mathbb{P}(X_i' \notin C) = (1-\mathbb{P}(C))^n$ can be as close to 0 as we want.

Similarly to the proof of the previous lecture, let

$$(X_i) \in \left\{\sup_C \frac{\mathbb{P}(C) - \frac{1}{n}\sum_{i=1}^{n} I(X_i \in C)}{\sqrt{\mathbb{P}(C)}} \geq t\right\}.$$

Hence, there exists C_X such that

$$\frac{\mathbb{P}(C) - \frac{1}{n}\sum_{i=1}^{n} I(X_i \in C)}{\sqrt{\mathbb{P}(C)}} \geq t.$$

Exercise 1. *Show that if*

$$\frac{\mathbb{P}(C_X) - \frac{1}{n}\sum_{i=1}^{n} I(X_i \in C_X)}{\sqrt{\mathbb{P}(C_X)}} \geq t$$

and

$$\frac{1}{n}\sum_{i=1}^{n} I(X_i' \in C_X) \geq \mathbb{P}(C_X),$$

then

$$\frac{\frac{1}{n}\sum_{i=1}^{n} I(X_i' \in C_X) - \frac{1}{n}\sum_{i=1}^{n} I(X_i \in C_X)}{\sqrt{\frac{1}{n}\sum_{i=1}^{n} I(X_i \in C_X) + \frac{1}{n}\sum_{i=1}^{n} I(X_i' \in C_X)}} \geq \frac{t}{\sqrt{2}}.$$

Hint: use the fact that $\phi(s) = \frac{s-a}{\sqrt{s}} = \sqrt{s} - \frac{s}{\sqrt{s}}$ *is increasing in* s.

From the above exercise it follows that

$$\begin{aligned}\frac{1}{4} &\leq \mathbb{P}_{(X_i')}\left(\frac{1}{n}\sum_{i=1}^{n} I(X_i' \in C_X) \geq \mathbb{P}(C_X)\right) \\ &\leq \mathbb{P}_{(X_i')}\left(\frac{\frac{1}{n}\sum_{i=1}^{n} I(X_i' \in C_X) - \frac{1}{n}\sum_{i=1}^{n} I(X_i \in C_X)}{\sqrt{\frac{1}{n}\sum_{i=1}^{n} I(X_i \in C_X) + \frac{1}{n}\sum_{i=1}^{n} I(X_i' \in C_X)}} \geq t\right)\end{aligned}$$

Since indicator is $0, 1$-valued,

$$
\begin{aligned}
&\frac{1}{4} I\left(\sup_C \frac{\mathbb{P}(C_X) - \frac{1}{n}\sum_{i=1}^n I(X_i \in C_X)}{\sqrt{\mathbb{P}(C_X)}} \geq t\right) \\
&\leq \mathbb{P}_{(X_i')}\left(\frac{\frac{1}{n}\sum_{i=1}^n I(X_i' \in C_X) - \frac{1}{n}\sum_{i=1}^n I(X_i \in C_X)}{\sqrt{\frac{1}{n}\sum_{i=1}^n I(X_i \in C_X) + \frac{1}{n}\sum_{i=1}^n I(X_i' \in C_X)}} \geq t\right) \\
&\leq \mathbb{P}_{(X_i')}\left(\sup_C \frac{\frac{1}{n}\sum_{i=1}^n I(X_i' \in C) - \frac{1}{n}\sum_{i=1}^n I(X_i \in C)}{\sqrt{\frac{1}{n}\sum_{i=1}^n I(X_i \in C) + \frac{1}{n}\sum_{i=1}^n I(X_i' \in C)}} \geq t\right).
\end{aligned}
$$

Hence,

$$
\begin{aligned}
&\frac{1}{4} \mathbb{P}\left(\sup_C \frac{\mathbb{P}(C_X) - \frac{1}{n}\sum_{i=1}^n I(X_i \in C_X)}{\sqrt{\mathbb{P}(C_X)}} \geq t\right) \\
&\leq \mathbb{P}\left(\sup_C \frac{\frac{1}{n}\sum_{i=1}^n I(X_i' \in C) - \frac{1}{n}\sum_{i=1}^n I(X_i \in C)}{\sqrt{\frac{1}{n}\sum_{i=1}^n I(X_i \in C) + \frac{1}{n}\sum_{i=1}^n I(X_i' \in C)}} \geq t\right) \\
&= \mathbb{E}\mathbb{P}_\varepsilon\left(\sup_C \frac{\frac{1}{n}\sum_{i=1}^n \varepsilon_i \left(I(X_i' \in C) - I(X_i \in C)\right)}{\sqrt{\frac{1}{n}\sum_{i=1}^n I(X_i \in C) + \frac{1}{n}\sum_{i=1}^n I(X_i' \in C)}} \geq t\right)
\end{aligned}
$$

There exist $C_1, \ldots, C_N$, with $N \leq \triangle_{2n}(\mathcal{C}, X_1, \ldots, X_n, X_1', \ldots, X_n')$. Therefore,

$$
\begin{aligned}
&\mathbb{E}\mathbb{P}_\varepsilon\left(\sup_C \frac{\frac{1}{n}\sum_{i=1}^n \varepsilon_i \left(I(X_i' \in C) - I(X_i \in C)\right)}{\sqrt{\frac{1}{n}\sum_{i=1}^n I(X_i \in C) + \frac{1}{n}\sum_{i=1}^n I(X_i' \in C)}} \geq t\right) \\
&= \mathbb{E}\mathbb{P}_\varepsilon\left(\bigcup_{k \leq N}\left\{\frac{\frac{1}{n}\sum_{i=1}^n \varepsilon_i \left(I(X_i' \in C_k) - I(X_i \in C_k)\right)}{\sqrt{\frac{1}{n}\sum_{i=1}^n I(X_i \in C_k) + \frac{1}{n}\sum_{i=1}^n I(X_i' \in C_k)}} \geq t\right\}\right) \\
&\leq \mathbb{E}\sum_{k=1}^N \mathbb{P}_\varepsilon\left(\frac{\frac{1}{n}\sum_{i=1}^n \varepsilon_i \left(I(X_i' \in C_k) - I(X_i \in C_k)\right)}{\sqrt{\frac{1}{n}\sum_{i=1}^n I(X_i \in C_k) + \frac{1}{n}\sum_{i=1}^n I(X_i' \in C_k)}} \geq t\right) \\
&\leq \mathbb{E}\sum_{k=1}^N \mathbb{P}_\varepsilon\left(\frac{1}{n}\sum_{i=1}^n \varepsilon_i \left(I(X_i' \in C_k) - I(X_i \in C_k)\right) \geq t\sqrt{\frac{1}{n}\sum_{i=1}^n I(X_i \in C_k) + \frac{1}{n}\sum_{i=1}^n I(X_i' \in C_k)}\right)
\end{aligned}
$$

The last expression can be upper-bounded by Hoeffding's inequality as follows:

$$\mathbb{E}\sum_{k=1}^{N}\mathbb{P}_{\varepsilon}\left(\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i\left(I(X_i'\in C_k)-I(X_i\in C_k)\right)\geq t\sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(I(X_i\in C_k)+I(X_i'\in C_k)\right)}\right)$$

$$\leq \mathbb{E}\sum_{k=1}^{N}\exp\left(-\frac{t^2\frac{1}{n}\sum_{i=1}^{n}\left(I(X_i\in C_k)+I(X_i'\in C_k)\right)}{\frac{1}{n^2}2\sum\left(I(X_i'\in C_k)-I(X_i\in C_k)\right)^2}\right)$$

since upper sum in the exponent is bigger than the lower sum (compare term-by-term)

$$\leq \mathbb{E}\sum_{k=1}^{N}e^{-\frac{nt^2}{2}}\leq\left(\frac{2en}{V}\right)^{V}e^{-\frac{nt^2}{2}}.$$

$\square$