We are interested in bounding

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

In Lecture 7 we hinted at Symmetrization as a way to deal with the unknown $\mathbb{P}(C)$.

Lemma 10.1. *[Symmetrization] If* $t\geq\sqrt{\frac{2}{n}}$*, then*

$$\mathbb{P}\left(\sup_{C\in\mathcal{C}}\left|\frac{1}{n}\sum_{i=1}^{n} I(X_i\in C)-\mathbb{P}(C)\right|\geq t\right)\leq 2\mathbb{P}\left(\sup_{C\in\mathcal{C}}\left|\frac{1}{n}\sum_{i=1}^{n} I(X_i\in C)-\frac{1}{n}\sum_{i=1}^{n} I(X_i'\in C)\right|\geq t/2\right).$$

Proof. Suppose the event

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

occurs. Let $X=(X_1,\ldots,X_n)\in\{\sup_{C\in\mathcal{C}}\left|\frac{1}{n}\sum_{i=1}^{n} I(X_i\in C)-\mathbb{P}(C)\right|\geq t\}$. Then

$$\exists C_X \;\text{ such that }\; \left|\frac{1}{n}\sum_{i=1}^{n} I(X_i\in C_X)-\mathbb{P}(C_X)\right|\geq t.$$

For a fixed C,

$$\mathbb{P}_{X'}\left(\left|\frac{1}{n}\sum_{i=1}^{n} I(X_i'\in C)-\mathbb{P}(C)\right|\geq t/2\right)=\mathbb{P}\left(\left(\frac{1}{n}\sum_{i=1}^{n} I(X_i'\in C)-\mathbb{P}(C)\right)^2\geq t^2/4\right)$$

$$\leq \text{ (by Chebyshev's Ineq) } \frac{4\mathbb{E}\left(\frac{1}{n}\sum_{i=1}^{n} I(X_i'\in C)-\mathbb{P}(C)\right)^2}{t^2}$$

$$=\frac{4}{n^2t^2}\sum_{i,j}\mathbb{E}(I(X_i'\in C)-\mathbb{P}(C))(I(X_j'\in C)-\mathbb{P}(C))$$

$$=\frac{4}{n^2t^2}\sum_{i=1}^{n}\mathbb{E}(I(X_i'\in C)-\mathbb{P}(C))^2=\frac{4n\mathbb{P}(C)(1-\mathbb{P}(C)}{n^2t^2}\leq\frac{1}{nt^2}\leq\frac{1}{2}$$

since we chose $t\geq\sqrt{\frac{2}{n}}$.

So,

$$\mathbb{P}_{X'}\left(\left|\frac{1}{n}\sum_{i=1}^{n} I(X_i'\in C_X)-\mathbb{P}(C_X)\right|\leq t/2\middle|\exists C_X\right)\geq 1/2$$

if $t\geq\sqrt{2/n}$. Assume that the event

$$\left|\frac{1}{n}\sum_{i=1}^{n} I(X_i'\in C_X)-\mathbb{P}(C_X)\right|\leq t/2$$

occurs. Recall that

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

Hence, it must be that

$$\left|\frac{1}{n}\sum_{i=1}^{n} I(X_i\in C_X)-\frac{1}{n}\sum_{i=1}^{n} I(X_i'\in C_X)\right|\geq t/2.$$

We conclude

$$\begin{aligned}\frac{1}{2} &\le \mathbb{P}_{X'}\left(\left|\frac{1}{n}\sum_{i=1}^n I(X_i' \in C_X) - \mathbb{P}(C_X)\right| \le t/2 \middle| \exists C_X\right) \\ &\le \mathbb{P}_{X'}\left(\left|\frac{1}{n}\sum_{i=1}^n I(X_i \in C_X) - \frac{1}{n}\sum_{i=1}^n I(X_i' \in C_X)\right| \ge t/2 \middle| \exists C_X\right) \\ &\quad \mathbb{P}_{X'}\left(\sup_{C\in\mathcal{C}}\left|\frac{1}{n}\sum_{i=1}^n I(X_i \in C) - \frac{1}{n}\sum_{i=1}^n I(X_i' \in C)\right| \ge t/2 \middle| \exists C_X\right).\end{aligned}$$

Since indicators are $0, 1$-valued,

$$\frac{1}{2} I\left(\underbrace{\sup_{C\in\mathcal{C}}\left|\frac{1}{n}\sum_{i=1}^n I(X_i \in C) - \mathbb{P}(C)\right| \ge t}_{\exists C_X}\right)$$

$$\le \mathbb{P}_{X'}\left(\sup_{C\in\mathcal{C}}\left|\frac{1}{n}\sum_{i=1}^n I(X_i \in C) - \frac{1}{n}\sum_{i=1}^n I(X_i' \in C)\right| \ge t/2 \middle| \exists C_X\right) \cdot I(\exists C_X)$$

$$\le \mathbb{P}_{X,X'}\left(\sup_{C\in\mathcal{C}}\left|\frac{1}{n}\sum_{i=1}^n I(X_i \in C) - \frac{1}{n}\sum_{i=1}^n I(X_i' \in C)\right| \ge t/2\right).$$

Now, take expectation with respect to X_i's to obtain

$$\mathbb{P}_X\left(\sup_{C\in\mathcal{C}}\left|\frac{1}{n}\sum_{i=1}^n I(X_i \in C) - \mathbb{P}(C)\right| \ge t\right)$$

$$\le 2 \cdot \mathbb{P}_{X,X'}\left(\sup_{C\in\mathcal{C}}\left|\frac{1}{n}\sum_{i=1}^n I(X_i \in C) - \frac{1}{n}\sum_{i=1}^n I(X_i' \in C)\right| \ge t/2\right).$$

□

Theorem 10.1. *If $VC(\mathcal{C}) = V$, then*

$$\mathbb{P}\left(\sup_{C\in\mathcal{C}}\left|\frac{1}{n}\sum_{i=1}^n I(X_i \in C) - \mathbb{P}(C)\right| \ge t\right) \le 4\left(\frac{2en}{V}\right)^V e^{-\frac{nt^2}{8}}.$$

Proof.

$$2\mathbb{P}\left(\sup_{C\in\mathcal{C}}\left|\frac{1}{n}\sum_{i=1}^n I(X_i \in C) - \frac{1}{n}\sum_{i=1}^n I(X_i' \in C)\right| \ge t/2\right)$$

$$= 2\mathbb{P}\left(\sup_{C\in\mathcal{C}}\left|\frac{1}{n}\sum_{i=1}^n \varepsilon_i\left(I(X_i \in C) - I(X_i' \in C)\right)\right| \ge t/2\right)$$

$$= 2\mathbb{E}_{X,X'}\mathbb{P}_\varepsilon\left(\sup_{C\in\mathcal{C}}\left|\frac{1}{n}\sum_{i=1}^n \varepsilon_i\left(I(X_i \in C) - I(X_i' \in C)\right)\right| \ge t/2\right).$$

The first equality is due to the fact that X_i and X_i' are i.i.d., and so switching their names (i.e. introducing random signs ε_i, $\mathbb{P}(\varepsilon_i = \pm 1) = 1/2$) does not have any effect. In the last line, it's important to see that the probability is taken with respect to ε_i's, while X_i and X_i''s are fixed.

By Sauer's lemma,

$$\triangle_{2n}(\mathcal{C}, X_1, \ldots, X_n, X_1', \ldots, X_n') \le \left(\frac{2en}{V}\right)^V.$$

In other words, any class will be equivalent to one of $C_1, \ldots, C_N$ on the data, where $N \le \left(\frac{2en}{V}\right)^V$. Hence,

$$\begin{aligned}
& 2\mathbb{E}_{X,X'}\mathbb{P}_\varepsilon\left(\sup_{C\in\mathcal{C}}\left|\frac{1}{n}\sum_{i=1}^n \varepsilon_i\left(I(X_i\in C) - I(X_i'\in C)\right)\right| \ge t/2\right)\\
&= 2\mathbb{E}_{X,X'}\mathbb{P}_\varepsilon\left(\sup_{1\le k\le N}\left|\frac{1}{n}\sum_{i=1}^n \varepsilon_i\left(I(X_i\in C_k) - I(X_i'\in C_k)\right)\right| \ge t/2\right)\\
&= 2\mathbb{E}_{X,X'}\mathbb{P}_\varepsilon\left(\bigcup_{k=1}^N\left|\frac{1}{n}\sum_{i=1}^n \varepsilon_i\left(I(X_i\in C_k) - I(X_i'\in C_k)\right)\right| \ge t/2\right)\\
&\overset{\text{union bound}}{\le} 2\mathbb{E}\sum_{k=1}^N \mathbb{P}_\varepsilon\left(\left|\frac{1}{n}\sum_{i=1}^n \varepsilon_i\left(I(X_i\in C_k) - I(X_i'\in C_k)\right)\right| \ge t/2\right)\\
&\overset{\text{Hoeffding's inequality}}{\le} 2\mathbb{E}\sum_{k=1}^N 2\exp\left(-\frac{-n^2t^2}{8\sum_{i=1}^n\left(I(X_i\in C) - I(X_i'\in C)\right)^2}\right)\\
&\le 2\mathbb{E}\sum_{k=1}^N 2\exp\left(-\frac{-n^2t^2}{8n}\right) \le 2\left(\frac{2en}{V}\right)^V 2e^{-\frac{nt^2}{8}}.
\end{aligned}$$

$\square$