We will show $\boxed{6 \rightarrow 1}$ in the *Fundamendal Theorem of Learning* from the lecture.

Theorem 1. *Let $\mathcal{H}$ be a hypothesis class. If $\mathrm{VC}(\mathcal{H}) = d < \infty$. Then, $\mathcal{H}$ has the uniform convergence property.*

Proof. In particular, we will show an upper bound for

$$\mathbb{E}_{S\sim\mathcal{D}^m}[\sup_{h\in\mathcal{H}} |L_{\mathcal{D}}(h) - L_S(h)|] \tag{1}$$

As $|L_{\mathcal{D}}(h) - L_S(h)|$ is a non-negative random variable, we can use Markov's inequality to obtain the uniform convergence property, i.e.,

$$\forall h \in \mathcal{H}: \quad \mathbb{P}(|L_{\mathcal{D}}(h) - L_S(h)| \leq \epsilon) \geq 1 - \delta \ .$$

First, we remember that

$$L_{\mathcal{D}}(h) = \mathbb{E}_{S\sim\mathcal{D}^m}[L_S(h)],$$

i.e., the generalization error is the expectation of the empirical error. Rephrasing Eq. (1), we get

$$\mathbb{E}_{S\sim\mathcal{D}^m}[\sup_{h\in\mathcal{H}} |L_{\mathcal{D}}(h) - L_S(h)|] = \mathbb{E}_{S\sim\mathcal{D}^m}[\sup_{h\in\mathcal{H}} |\mathbb{E}_{S'\sim\mathcal{D}^m}[L_{S'}(h)] - L_S(h)|] \ .$$

Now, by looking at $\mathbb{E}_{S'\sim\mathcal{D}^m}[L_{S'}(h)] - L_S(h)$, we realize that $L_S(h)$ could equally be combined into

$$\mathbb{E}_{S'\sim\mathcal{D}^m}[L_{S'}(h) - L_S(h)] \ ,$$

as the expectation is taken over S' and $L_S(h)$ is constant in this context. Further, we know that $|\cdot|$ is convex and we can use Jensen's inequality to get

$$\mathbb{E}_{S'\sim\mathcal{D}^m}|(L_{S'}(h) - L_S(h))| \leq |\mathbb{E}_{S'\sim\mathcal{D}^m}[L_{S'}(h) - L_S(h)]|.$$

Based on this result, we now study

$$\sup_h |\mathbb{E}_{S'\sim\mathcal{D}^m}[L_{S'}(h) - L_S(h)]|$$

To bound this, we add a little remark.

Remark 2. *Let $f(X, y)$ be a function of a random variable X and some scalar $y \in \mathbb{R}$. It obviously holds that*

$$f(X, y) \leq \sup_y f(X, y)$$

Upon taking the expectation over X, we obtain

$$\mathbb{E}_X f(X, y) \leq \mathbb{E}_X \sup_y f(X, y) \ .$$

As the right-hand side is the smallest upper bound for the left-hand side, taking the supremum over y on the left-hand side still holds, i.e.,

$$\sup_y \mathbb{E}_X f(X, y) \leq \mathbb{E}_X \sup_y f(X, y) \ .$$

Now, by Remark 2, we get

$$\sup_h |\mathbb{E}_{S' \sim \mathcal{D}^m}[L_{S'}(h) - L_S(h)]| \leq \mathbb{E}_{S' \sim \mathcal{D}^m}[\sup_h |(L_{S'}(h) - L_S(h))|]$$

Upon combination with the outer expectation over S from Eq. (1), we obtain

$$\mathbb{E}_{S \sim \mathcal{D}^m}[\sup_{h \in \mathcal{H}} |L_{\mathcal{D}}(h) - L_S(h)|] \leq \mathbb{E}_{S,S' \sim \mathcal{D}^m}[\sup_h |(L_{S'}(h) - L_S(h))|] \tag{2}$$

as our first intermediate result. We can now expand the term on the right-hand side of Eq. (2) by our definition of the empirical loss, i.e.,

$$\mathbb{E}_{S,S' \sim \mathcal{D}^m}[\sup_h |(L_{S'}(h) - L_S(h))|] = \mathbb{E}_{S,S' \sim \mathcal{D}^m}\left[\sup_h \left|\left(\frac{1}{m}\sum_i l(h, z_i') - \frac{1}{m}\sum_i l(h, z_i)\right)\right|\right]$$

As we work with 0-1 loss functions $l : \mathcal{H} \times Z \to \{0, 1\}$, $l(h, z_i)$ and $l(h, z_i')$ can only take values 0 or 1. Also, S and S' contain samples that are drawn independently and identically distributed (iid). Consequently, we can switch z_i and z_i'. This only leads to a switch of sign for the term $l(h, z_i') - l(h, z_i)$. In other words, multiplying $l(h, z_i') - l(h, z_i)$ by $\sigma_i = -1$, i.e.,

$$\sigma_i(l(h, z_i') - l(h, z_i))$$

corresponds to one such switch. If we multiply each difference with some $\sigma_i \in \{-1, +1\}$ this corresponds to a switch (if $\sigma_i = -1$) or no switch (if $\sigma_i = 1$). Either way, nothing changes. So, we could just draw a random bipolar sequence $\sigma = (\sigma_1, \ldots, \sigma_m)$ from a uniform distribution U^m on $\{-1, +1\}$, written as $\sigma \sim U^m$, and take the expectation, i.e.,

$$\mathbb{E}_{\sigma \sim U^m}\mathbb{E}_{S,S' \sim \mathcal{D}^m}\left[\sup_h \left|\left(\frac{1}{m}\sum_i \sigma_i(l(h, z_i') - l(h, z_i))\right)\right|\right]$$

Next, by the Fubini theorem, we can switch the expectations to get

$$\mathbb{E}_{S,S' \sim \mathcal{D}^m}\mathbb{E}_{\sigma \sim U^m}\left[\sup_h \left|\left(\frac{1}{m}\sum_i \sigma_i(l(h, z_i') - l(h, z_i))\right)\right|\right]$$

Lets fix S and S' for now and combine S and S' into a set C that contains the elements from both sets. C has size $2m$. When restricting $\mathcal{H}$ to C, we see that $\mathcal{H}_C$ is finite. Consequently, the previous equation simplifies to

$$\mathbb{E}_{\sigma \sim U^m}\left[\max_{h \in \mathcal{H}_C} \left|\left(\frac{1}{m}\sum_i \sigma_i(l(h, z_i') - l(h, z_i))\right)\right|\right]$$

Note that the summation inside is still from $1 \ldots m$, but the z_i' and z_i are from different sets (which we just fixed).

Now, lets additionally fix some $h \in \mathcal{H}_C$. We obtain

$$\mathbb{E}_{\sigma \sim U^m}\left[\left|\left(\frac{1}{m}\sum_i \sigma_i(l(h, z_i') - l(h, z_i))\right)\right|\right]$$

This looks very much like something where we could use the Hoeffding inequality. The Hoeffding inequality bounds the absolute deviation of the empirical mean of random variables from its expected values. Our random variables are $\gamma_i := \sigma_i(l(h, z_i') - l(h, z_i))$ in that case. We also know that the expectation over σ is 0 (as we draw from a uniform distribution on $\{-1, +1\}$). Further, the random variables can take on values in $[-1, +1] = [a, b]$, so $b - a = 2$. We get

$$\mathbb{P}_{\sigma \sim U^m}[|\overline{\gamma} - \underbrace{\mathbb{E}[\overline{\gamma}]}_{=0}| > \rho] = \mathbb{P}_{\sigma \sim U^m}[|\overline{\gamma}| > \rho] \leq 2e^{-2mt^2/(b-a)^2} = 2e^{-mt^2/2}$$

Now, if h is not fixed, we have to take care of $\max_{h \in \mathcal{H}_C}$. This can easily be done via the union bound to obtain

$$\mathbb{P}_{\sigma \sim U^m}[\max_{h \in \mathcal{H}_C} |\overline{\gamma}| > \rho] \leq 2|\mathcal{H}_C|e^{-mt^2/2}$$

However, we need to bound

$$\mathbb{E}_{\sigma \sim U^m}[|\gamma_h|]$$

for which we need an additional lemma.

Lemma 1. *Let X be a random variable and $x' \in \mathbb{R}$. Assume that there exists $a > 0$ and $b \geq e$ s.t. for all $t \geq 0$ we have*

$$\mathbb{P}[|X - x'| > t] \leq 2be^{-t^2/a^2}$$

Then,

$$\mathbb{E}[|X - x'|] \leq a\left(2 + \sqrt{\log(b)}\right)$$

In our case, we have $x' = 0$ and we assume that $|\mathcal{H}_C| \geq e$. Also, we have that

$$m/2 = \frac{1}{a^2} \Rightarrow a = \sqrt{\frac{2}{m}}$$

Upon combination, we obtain

$$\mathbb{E}_{\sigma \sim U^m}[\max_{h \in \mathcal{H}_C} |\gamma_h|] \leq \sqrt{\frac{2}{m}}\left(2 + \sqrt{\log(|\mathcal{H}_C)|}\right)$$

Remember that we fixed S and S'. However, as the right-hand side is independent of S and S', we obtain from Eq. (2)

$$\mathbb{E}_{S \sim \mathcal{D}^m}[\sup_{h \in \mathcal{H}} |L_{\mathcal{D}}(h) - L_S(h)|] \leq \frac{4 + 2\sqrt{\log(|\mathcal{H}_C)|)}}{\sqrt{2m}} = \frac{4 + 2\sqrt{\log(\tau_{\mathcal{H}}(2m))}}{\sqrt{2m}}$$

□

Please note that $2m$ inside the growth function: remember that we limited $\mathcal{H}$ to C which was of size $2m$. From Markov's inequality, we know that for a non-negative random variable Z,

$$\mathbb{P}[Z > a] \leq \frac{\mathbb{E}[Z]}{a}$$

holds. Now, as $|L_{\mathcal{D}}(h) - L_S(h)|$ is obviously non-negative, we get that for every $\delta \in (0, 1)$

$$\mathbb{P}[\sup_h |L_{\mathcal{D}}(h) - L_S(h)| > \delta] \leq \frac{4 + 2\sqrt{\log(\tau_{\mathcal{H}}(2m))}}{\delta\sqrt{2m}}$$

Or, reformulated

$$\mathbb{P}[\sup_h |L_{\mathcal{D}}(h) - L_S(h)| > \frac{4 + 2\sqrt{\log(\tau_{\mathcal{H}}(2m))}}{\delta\sqrt{2m}}] \leq \frac{\frac{4+2\sqrt{\log(\tau_{\mathcal{H}}(2m))}}{\sqrt{2m}}}{\frac{4+2\sqrt{\log(\tau_{\mathcal{H}}(2m))}}{\delta\sqrt{2m}}} = \delta$$

Equivalently,

$$\mathbb{P}_{S\sim\mathcal{D}^m}\left[\sup_h |L_{\mathcal{D}}(h) - L_S(h)| < \frac{2 + \sqrt{\log(\tau_{\mathcal{H}}(2m))}}{\delta\sqrt{2m}}\right] \geq 1 - \delta, \quad ,$$

which means that probability of at least $1-\delta$ over the choice of S, we have that $|L_{\mathcal{D}}(h) - L_S(h)|$ is bounded by $(2 + \cdots)/\delta\sqrt{2m}$.

Effectively, what remains to be shown for $\boxed{6 \to 1}$ to hold is that if the VC dimension is finite, uniform convergence holds. In other words, we need an expression for the *sample complexity function* $m_{\mathcal{H}}^{UC}(m)$. We could assume that $m > d$ as, in that case, the Sauer lemma states

$$\tau_{\mathcal{H}}(2m) \leq \left(\frac{2em}{d}\right)^d \quad .$$

Now, set this in for $\tau_{\mathcal{H}}(2m)$ and express the right hand side of the inequality in terms of m such that it is at most ϵ. This is left as an exercise, but requires a little bit of algebraic manipulations.

In summary, we say that $\mathcal{H}$ has *small effective size* if $\tau_{\mathcal{H}}$ grows polynomially, i.e., when $m > d$. As we have shown, in that case we get uniform convergence, because we can use Sauer's lemma (which only works for $\text{VC}(\mathcal{H}) < \infty$).