Theorem 1. *Let $\mathcal{H}$ be a class and let τ_{H} be its growth function. Then, for every distribution $\mathcal{D}$ and every $\delta \in (0,1)$, with probability of at least $1-\delta$ (over repeated sampling of training data $S \sim \mathcal{D}^m$), we have for a loss functions in the range $[0,c]$,*

$$\boxed{\forall h \in \mathcal{H} \,:\, |L_{\mathcal{D}}(h) - L_S(h)| \leq c\sqrt{\frac{8\log\left(\tau_{\mathcal{H}}(2m)\;4/\delta\right)}{m}}}$$

Proof. Lets recall the triangle inequality for absolute values, i.e.,

$$|a-b| \geq |a| - |b| \tag{1}$$

Also, recall that we can write the generalization error as

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

meaning that we can replace the term $L_{\mathcal{D}}(h)$ in Theorem 1 by that term without problems. Lets assume now we have $S, S' \sim \mathcal{D}^m$, i.e., two random samples from $\mathcal{D}$ of size m. Using the Eq. (1), we can write (with a, b being the terms in the triangle inequality)

$$|\underbrace{(L_{\mathcal{D}}(h) - L_S(h))}_{a} - \underbrace{(L_{\mathcal{D}}(h) - L_{S'}(h))}_{b}| \geq |(L_{\mathcal{D}}(h) - L_S(h)| - |(L_{\mathcal{D}}(h) - L_{S'}(h)|$$

and get

$$|L_{S'}(h) - L_S(h)| \geq |(L_{\mathcal{D}}(h) - L_S(h)| - |(L_{\mathcal{D}}(h) - L_{S'}(h)|$$

Now, if

1. $|(L_{\mathcal{D}}(h) - L_S(h)| > \epsilon$, and
2. $|(L_{\mathcal{D}}(h) - L_{S'}(h)| < \epsilon/2$,

we get

$$|L_{S'}(h) - L_S(h)| > \epsilon/2$$

Alternatively, this could be written in terms of indicator functions, 1_A, as

$$1_{|(L_{\mathcal{D}}(h)-L_S(h)|>\epsilon} \cdot \underbrace{1_{|(L_{\mathcal{D}}(h)-L_{S'}(h)|<\epsilon/2}}_{\text{Term 1}} \leq \underbrace{1_{|L_{S'}(h)-L_S(h)|>\epsilon/2}}_{\text{Term 2}} \tag{2}$$

Term 1. If we take the expectation w.r.t. S', we get, by the Hoeffding inequality

$$\begin{aligned} \mathbb{E}_{S'}[1_{|(L_{\mathcal{D}}(h)-L_{S'}(h)|<\epsilon/2}] &= \mathbb{P}_{S'}[|(L_{\mathcal{D}}(h) - L_{S'}(h)| < \epsilon/2] \\ &\geq 1 - 2e^{-\epsilon^2 n/(2c^2)} \end{aligned} \tag{3}$$

This is easy to see as

$$L_{S'}(h) = {}^1\!/_m \sum_i l(h, z_i') = \sum_i l(h, z_i')\,{}^1\!/_m \tag{4}$$

where each term in the sum is an i.i.d. random variable within the range $[0, c/m]$ (remember that the loss is within $[0, c]$). The Hoeffding inequality then reads as

$$\mathbb{P}[|S_n - \mathbb{E}[S_n]| \geq t) \leq e^{-2t^2/(\sum_i (b_i - a_i)^2)}$$

for a sum, S_n, of i.i.d. RVs where, in our case,

$$\sum_{i=1}^{m}(c/m - 0) = mc^2/m^2 = c^2/m$$

Term 2. Taking the expectation w.r.t. S', we get

$$\begin{aligned}\mathbb{E}_{S'}[1_{|(L_{S'}(h)-L_S(h)|>\epsilon/2}] &= \mathbb{P}_{S'}[|(L_{S'}(h)-L_S(h)|>\epsilon/2] \\ &\leq \mathbb{P}_{S'}[\exists h \in \mathcal{H} : |(L_{S'}(h)-L_S(h)|>\epsilon/2]\end{aligned} \tag{5}$$

To simplify Eq. (2) via Eqs. (3) and (5), we make the following assumption. In particular, we let

$$n \geq 4c^2\epsilon^{-2}\log(2) \ . \tag{6}$$

At equality, Eq. (3) the RHS simplifies to

$$1-2e^{-\epsilon^2(4c^2\epsilon^{-2}\log(2))/(2c^2)} = 1-2e^{-2\log(2)} = 1-2e^{-\log(2^2)} = 1-1/2 = 1/2$$

So, if $n \geq 4c^2\epsilon^{-2}\log(2)$, then obtain that

$$\mathbb{E}_{S'}[1_{|(L_{\mathcal{D}}(h)-L_{S'}(h)|<\epsilon/2]} \geq 1/2$$

Overall, we have, at the moment

$$\boxed{1_{|L_{\mathcal{D}}(h)-L_S(h)|>\epsilon} \leq 2\mathbb{P}_{S'}[\exists h \in \mathcal{H} : |L_{S'}(h)-L_S(h)|>\epsilon/2]} \tag{7}$$

Note that this inequality holds for **all** $h \in \mathcal{H}$. So, we might as well write

$$1_{\exists h\in\mathcal{H}:|L_{\mathcal{D}}(h)-L_S(h)|>\epsilon} \leq 2\mathbb{P}_{S'}[\exists h \in \mathcal{H} : |(L_{S'}(h)-L_S(h)|>\epsilon/2]$$

Upon taken expectations on both sides w.r.t. S, we get

$$\mathbb{P}_S[\exists h \in \mathcal{H} : |L_{\mathcal{D}}(h)-L_S(h)|>\epsilon] \leq 2\mathbb{P}_{S',S}[\exists h \in \mathcal{H} : |L_{S'}(h)-L_S(h)|>\epsilon/2]$$

Insight 1. The key thing to note now is that the RHS only involves empirical terms, in particular,

$$L_{S'}(h) = {}^1\!/_m \sum_{i=1}^{m} l(h, z_i') \quad \text{and} \quad L_S(h) = {}^1\!/_m \sum_{i=1}^{m} l(h, z_i)$$

Lets take a step back now and think about the fact that $z_i' \sim \mathcal{D}$ and $z_i \sim \mathcal{D}$. These samples are all drawn i.i.d. from $\mathcal{D}$. So, it does not matter if we switch z_i' with z_i. The only thing that changes if we do one such switch is that $(l(h, z_i') - l(h, z_i))$ changes to $-(l(h, z_i') - l(h, z_i))$. This is easy to see, e.g., in case of the $0 - 1$ loss:

$l(h, z_i')$	$l(h, z_i)$	$(l(h, z_i') - l(h, z_i))$
0	0	0
0	1	-1
0	0	0
1	0	1

In fact, a multiplication by -1 in one of the summation terms of

$${}^1\!/_m \sum_{i=1}^{m} [l(h, z_i') - l(h, z_i)]$$

amounts to a switch of z_i and z_i'. This allows us to write

$$\begin{aligned}
\mathbb{P}_S[\exists h \in \mathcal{H} \,:\, |L_\mathcal{D}(h) - L_S(h)| > \epsilon] &\leq 2\mathbb{P}_{S',S}[\exists h \in \mathcal{H} \,:\, |(L_{S'}(h) - L_S(h)| > \epsilon/2] \\
&= \mathbb{P}_{S',S}[\exists h \in \mathcal{H} \,:\, 1/m|\sum_i l(h, z_i') - l(h, z_i)| > \epsilon/2] \\
&= \mathbb{E}_{S',S}[1_{\exists h \in \mathcal{H} \,:\, 1/m|\sum_i l(h,z_i') - l(h,z_i)| > \epsilon/2}] \\
&= \mathbb{E}_{S',S}[\mathbb{E}_\sigma[1_{\exists h \in \mathcal{H} \,:\, 1/m|\sum_i (l(h,z_i') - l(h,z_i))\sigma_i| > \epsilon/2}]] \\
&= \mathbb{E}_{S',S}[\mathbb{P}_\sigma[\exists h \in \mathcal{H} \,:\, 1/m|\sum_i (l(h, z_i') - l(h, z_i))\sigma_i| > \epsilon/2]]
\end{aligned} \tag{8}$$

with σ_i drawn i.i.d. from a discrete uniform distribution on $\{-1, +1\}$. Such variables are also called Rademacher variables.

Insight 2. Again, since the RHS in Eq. (8) only involves empirical quantities, it also implies that every function from $\mathcal{H}$ is evaluated on, at most, $2m$ points. Combining samples from S, S' into C (which is of size $2m$), we can restrict $h \in \mathcal{H}$ to $h \in \mathcal{H}_C$, where

$\mathcal{H}_C$ is the restriction of H to C. Since C is finite, $\mathcal{H}_C$ is finite and we can work with our usual union bound. The question will obviously be, how fast, $|\mathcal{H}_C|$ will grow.

To bound the last term on the RHS of Eq. (8), note the "exists" (∃) statement translates into a union bound (once we switch to $\mathcal{H}_C$), i.e.,

$$\mathbb{P}_\sigma[\exists h \in \mathcal{H} \,:\, 1/m|\sum_i (l(h, z_i') - l(h, z_i))\sigma_i| > \epsilon/2] \leq \sum_{h \in \mathcal{H}_C} \mathbb{P}_\sigma[|\sum_i k_i/m \ \sigma_i| > \epsilon/2]$$

with

$$k_i \ := (l(h, z_i') - l(h, z_i))/m$$

Note that the k_i are constant in

$$\mathbb{P}_\sigma[|\sum_i k_i/m \ \sigma_i| > \epsilon/2] = \mathbb{E}_\sigma[1_{|\sum_i k_i/m \ \sigma_i| > \epsilon/2}]$$

So, effectively, $\sum_i {}^1\!/\!_m k_i \sigma_i$ is a sum over i.i.d. random variables. Also, $\mathbb{E}_\sigma[\sum_i {}^1\!/\!_m k_i \sigma_i] = 0$. Hence, we have a classic case for the Hoeffding inequality. Since each loss term is assumed to be in the range $[0, c]$, the difference computed in k_i is in $[-c, c]$ and dividing by m gives $[-c/m, c/m]$. We obtain

$$\begin{aligned} \mathbb{P}_\sigma[\exists h \in \mathcal{H} \,:\, 1/m|\sum_i (l(h, z_i') - l(h, z_i))\sigma_i| > \epsilon/2] &\leq \sum_{h \in \mathcal{H}_C} \mathbb{P}_\sigma[|\sum_i k_i/m \ \sigma_i| > \epsilon/2] \\ &\leq |\mathcal{H}_C| 2 e^{-m\epsilon^2/(8c^2)} \end{aligned} \tag{9}$$

and, overall

$$\begin{aligned} \mathbb{P}_S[\exists h \in \mathcal{H} \,:\, |L_\mathcal{D}(h) - L_S(h)| > \epsilon] &\leq 2\mathbb{E}_{S',S}[|\mathcal{H}_C| 2 e^{-m\epsilon^2/(8c^2)}] \\ &= 4\mathbb{E}_{S',S}[|\mathcal{H}_C| e^{-m\epsilon^2/(8c^2)}] \\ &\leq 4\tau_\mathcal{H}(2m) e^{-m\epsilon^2/(8c^2)} \end{aligned} \tag{10}$$

where the last equality results from the fact that we have replaced $|\mathcal{H}_C|$ by the growth function $\tau_{\mathcal{H}}$ evaluated on $2m$ samples (remember $|C| = 2m$) which does no longer depend on the choice of S, S' but only on their size. Setting the RHS equal to δ concludes the proof of Theorem 1. Also note that our assumption on m then boils down to

$$\delta \leq 2\sqrt{2}\tau_{\mathcal{H}}(2m)$$

which is always guaranteed as $\delta \in (0, 1)$.

Remark on 0-1 loss. In that case, we have $c = 1$. The result of Theorem 1 is slightly different from the book/slides, where the constants differ and the sup is bounded. However, as our result holds for all $h \in \mathcal{H}$, it holds for $\sup_{h \in \mathcal{H}}$ as well. It's now also fairly easy to show that if the VC dimension is finite, we obtain uniform convergence and hence APAC learnability. □