RECAP:

(Informal Version of Occam's Razor Theorem): If we have a consistent learner for C using H. Then provided we draw $m \geq \frac{1}{\varepsilon}\left(\log|H_n| + \log\frac{1}{\delta}\right)$ examples, we have a PAC learning algorithm using L.

VC-Dimension: For $S \subseteq X$, $\Pi_C(S)$ is the set of all possible dichotomies of S realised by C.

Shattering: We say that a finite set S is shattered by C if $|\Pi_C(S)| = 2^{|S|}$.

VC-dim: The cardinality of the largest finite set shattered by C, if it exists. If arbitrarily large finite sets are shattered VC-dim$(C) = \infty$.

Growth Function:

$$\Pi_C(m) = \max\{|\Pi_C(S)| \mid |S| = m\}$$

If VC-dim$(C) = d < \infty$, then $\Pi_C(m) = 2^m$ for $m \leq d$.

Sauer's Lemma: $\Pi_C(m) \leq \left(\frac{me}{d}\right)^d$ for $m \geq d$.

Theorem: Let C be a concept class with VC-dim$(C) = d_c < \infty$. Let L be a consistent learner for C using H. Let VC-dim$(H) = d < \infty$. Then L is PAC-learner for C provided it is given m examples drawn from $EX(c, D)$, where

$$m \geq K_0\left(\frac{1}{\varepsilon}\log\frac{1}{\delta} + \boxed{\frac{d}{\varepsilon}}\log\frac{1}{\varepsilon}\right),$$ where K_0 is a universal constant.

(If L is an efficient consistent learner & H is polynomially evaluatable, the we get an efficient PAC-learner.)

Proof: Fix $c \in C$ to be the target concept; and let D be the distribution over X that generates the data.

Let $H \oplus c = \{ h \oplus c \mid h \in H \}$,

$$h \oplus c(x) = \begin{cases} 1 & \text{if } h(x) \neq c(x) \\ 0 & \text{otherwise} \end{cases}$$

$$err(h; c, D) = \mathbb{P}_{x \sim D}[h(x) \neq c(x)] = \mathbb{P}_{x \sim D}[h \oplus c(x) = 1]$$

Observation: VC-dim $(H \oplus c)$ = VC-dim(H). [exercise].

$$\Delta_{D,c,\varepsilon}(H) = \{ h \oplus c \in H \oplus c \mid \mathbb{P}_{x \sim D}[h \oplus c(x) = 1] \geq \varepsilon \}.$$ [Hypothesis that we need to eliminate]

$S \subseteq X$ is an ε-net for H (or $H \oplus c$) if $\forall h' \in \Delta_{D,c,\varepsilon}$

there exists $x \in S$, s.t. $h'(x) = 1$

if S was our training sample, and S is an ε-net, then if h is consistent with c on S, $err(h) \leq \varepsilon$.

Need to get a bound on the probability that a random set of size m drawn i.i.d from D fails to be an ε-net with respect to the target c.

$S_1 \sim D^m$: S_1 is of size m drawn i.i.d from D.
(can allow to be multisets)

A : $S_1 \sim D^m$ and S_1 fails to be an ε-net.

B : $S_1 \sim D^m$, $S_2 \sim D^m$, $\exists\, h' \in H \oplus c$, s.t.

(i) $|\{x \in S_1 \cup S_2 \mid h'(x) = 1\}| \geq \frac{\varepsilon m}{2}$ $\quad T$ (as defined below)

(ii) $|\{x \in S_1 \mid h''(x) = 1\}| = 0$ $\quad \forall\, h'' \in T$ ~~$\Pi_{H \oplus c}(S_1 \cup S_2)$~~

(a) $\mathbb{P}(B) \leq$ "small"

(b) $\mathbb{P}(B \mid A) \geq 1/2$

(c) $\mathbb{P}(B) \geq \mathbb{P}(B \mid A) \cdot \mathbb{P}(A)$

$\implies \mathbb{P}(A) \leq 2 \cdot \mathbb{P}(B) \leq$ "small".

$\underline{B}$: Let $T = \{h \in \Pi_{H \oplus c}(S_1 \cup S_2) \mid h(x) = 1$ for at least $\varepsilon m/2$ x in $S_1 \cup S_2\}$

S_1 are S_2 are identically distributed.

For any fixed $h \in T$,

C [if there are r black balls and $2m - r$ red balls, then in a random arrangement all r black balls appear in the second half; ($r \geq \varepsilon m/2$)

$$\mathbb{P}(C) = \frac{\binom{m}{r}}{\binom{2m}{r}} = \frac{m(m-1)\cdots(m-r+1)}{2m(2m-1)\cdots(2m-r+1)} \leq \left(\frac{1}{2}\right)^r = 2^{-r}$$

(union bound)

$\exists\, h \in T$, for which this happens $\leq |T| \cdot 2^{-r}$

$$\leq |\Pi_{H \oplus c}(S_1 \cup S_2)| \cdot 2^{-r}$$

$$\leq \left(\frac{2me}{d}\right)^d \cdot 2^{-\varepsilon m/2}$$

$$\mathbb{P}(B) \leq \left(\frac{2me}{d}\right)^d \cdot 2^{-\varepsilon m/2}$$

How large does m need to be for $\mathbb{P}(B) \leq \delta/2$

Check that this is the case for m in the statement of the Theorem.

A: S_1 fails ~~a~~ to be an ε-net.

Let $h_1 \in H$, $\mathbb{P}(h_1(x)=1) \geq \varepsilon$ but, $h_1(x) = 0 \quad \forall x \in S_1$.

Draw $S_2 \sim D^m$.

$\mathbb{P}[h_1(x) = 1] \geq \varepsilon$ for $x \sim D$

$\mathbb{E}\,|\{x \in S_2 \mid h_1(x) = 1\}| \geq \varepsilon m$ (by Linearity of expectation).

$\mathbb{P}\left(|\{x \in S_2 \mid h_1(x) = 1\}| < \frac{\varepsilon m}{2}\right) \leq \frac{1}{2}$ — Either use Chernoff bound or explicit calculation.

$Z = |\{x \in S_2 \mid h_1(x) = 1\}|$

$Z \sim \mathrm{Bin}(m, p)$ where $p = \mathbb{P}_{x\sim D}(h_1(x)=1) \geq \varepsilon$.

$\mathbb{P}(|\{x \in S_2 \mid h_1(x) = 1\}| \geq \varepsilon m/2) \geq \frac{1}{2}$.

if this happens the event B occurs.

$\therefore \mathbb{P}[B \mid A] \geq \frac{1}{2}$.

$$\left(\frac{2me}{d}\right)^d \cdot 2^{-\varepsilon m/2} \leq \delta/2$$

$$\Leftrightarrow \left(\frac{2me}{d}\right)^d \cdot \frac{2}{\delta} \leq 2^{\varepsilon m/2}$$

$$\frac{\varepsilon m}{2} \geq \log_2\left(\left(\frac{2me}{d}\right)^d \cdot \frac{2}{\delta}\right)$$

$$\Leftrightarrow m \geq \frac{2}{\varepsilon}\left(d \log_2\left(\frac{2me}{d}\right) + \log_2 \frac{2}{\delta}\right)$$

$\left(m \geq \frac{K}{\varepsilon}\left(d \log\frac{1}{\varepsilon} + \log\frac{1}{\delta}\right)\right.$ this will hold.