Learning theory proves statistical bounds on the sample complexity of function classes. The fundamental theorem says that VC dimension characterizes sample complexity, and that ERM algorithms are optimal for PAC learning.

VC dimension completely characterizes the growth of the shattering number. *Sauer's lemma* says that when $|\mathcal{X}|$ is below the VC dimension, $\mathcal{H}$ shatters all points and the growth function is exponential; whereas above the VC dimension, the shattering number grows slowly (polynomially).

The shattering number measures the effective size of $\mathcal{H}$ on some $\mathcal{X}$. If $\mathcal{H}$ can label $\mathcal{X}$ in many different ways, then the search problem is difficult, and the sample complexity will be high. In the extreme case, $\mathcal{H}$ shatters all points and any labeling is possible, which means that it is impossible to learn from a training subset how h^* will behave on the rest of the points. The *no free lunch theorem* formalizes this notion, and we can show that if $|\mathcal{X}| = \infty$ and $VC(\mathcal{H}) = \infty$, no learning is possible at all. On the other hand, if the shattering number is small, then we only need to observe a few points to have a good idea of how h^* will behave on all of $\mathcal{X}$.

Textbooks: Understanding Machine Learning, Elements of Statistical Learning

Risk is expected loss, e.g. $R(h) = P_{x\sim D}(h(x) \neq y)$ for the 0-1 loss

Generalization gap $\Delta = \sup_f |\hat{R} - R|$

For finite $\mathcal{F}$, $\Delta \leq \sqrt{\frac{\ln|\mathcal{F}|+\ln 2/\delta}{2n}}$
Pf. Hoeffding's (bdd 1) + UB: $P(\Delta \geq t) \leq 2|\mathcal{F}| \exp(-2nt^2) = \delta$

VC theorem. $\Delta \leq O\left(\sqrt{\frac{|VC(\mathcal{F})\log n|+\ln 1/\delta}{n}}\right)$

For **realizable** $\mathcal{H}$, $\mathcal{H}$ is **PAC learnable** iff $\forall\varepsilon,\delta > 0$, $\exists n$ s.t. $P_{X\sim D}(R(\hat{h}) \geq \varepsilon) \leq \delta$

Thm. Finite, realizable $\mathcal{H}$ is PAC learnable for $n \geq \frac{1}{\varepsilon}\log\frac{|\mathcal{H}|}{\delta}$ with $\hat{h}$ from ERM.

Pf. Let $\mathcal{H}_\varepsilon = \{h | R(h) > \varepsilon\}$.

$$\begin{aligned} P(R(\hat{h}) \le \varepsilon) &= P\left(\bigcup_{h\in\mathcal{H}_\varepsilon} \{X | \hat{R}_X(h) = 0\}\right) \\ &\le \sum_{h\in\mathcal{H}_\varepsilon} P(\hat{R}_X(h) = 0) \\ &\le \sum_{h\in\mathcal{H}_\varepsilon} (1-\varepsilon)^n \\ &\le |\mathcal{H}|(1-\varepsilon)^n \\ &\le |\mathcal{H}|e^{-n\varepsilon} \end{aligned}$$

E.g. For a finite domain $\mathcal{X}$ and the full class of $2^{|\mathcal{X}|}$ hypotheses $\mathcal{H} : \mathcal{X} \to \{-1, 1\}$, $P(R(\hat{h}) \ge \frac{1}{8}) \le \frac{1}{7}$ with at least $8\log(7|\mathcal{H}|) \approx 5|\mathcal{X}|$ samples.

VC dimension

The *restriction* of $\mathcal{H}$ onto X $\mathcal{H}_X = \{\,(h(X_1), \ldots, h(X_m)) \,|\, h \in \mathcal{H}\}$ is the set of label vectors that can be generated by hypotheses in $\mathcal{H}$

The *shattering number* $|\mathcal{H}_X|$ is the number of possible labelings of X
$|\mathcal{H}_X| \le 2^{|X|}$

$\mathcal{H}$ **shatters** X if $|\mathcal{H}_X| = 2^{|X|}$ i.e. $\mathcal{H}$ can generate any possible labeling of X

VC dimension of $\mathcal{H}$ is the maximum size of a set that can be shattered by $\mathcal{H}$

- *VC of hyperplane* is $\le d + 1$ by Radon's thm: for any $d + 2$ points, $\exists$ a partition into two sets s.t. convex hulls overlap.
- *VC of NN* is $N \log N$, for N edges. Concatenation $\mathcal{F}_1 \times \mathcal{F}_2$ i.e. Cartesian product at most $\Pi_{\mathcal{F}_1}(m)\Pi_{\mathcal{F}_2}(m)$. Composition same since $|\cup_{y=f_1(x)} f_2(y)| \le \sum_x |f_2(f_1(x))|$. Network is composition of concatenations of hyperplanes, each $\Pi \le m^{d_i-1}$. Thus total $\Pi \le m^N$. Shattered $2^m \le m^N$.

Growth function $\tau_{\mathcal{H}}(m) = \max_{|X|=m} |\mathcal{H}_X|$

Sauer's lemma (UML page 74).

$$|\mathcal{H}_X| \le \sum_{i=0}^{VC(\mathcal{H})} \binom{|X|}{i}$$

$VC(\mathcal{H}) = O(m^{VC(\mathcal{H})})$ since for $m > VC(\mathcal{H})$, $|\mathcal{H}_X| \le (\frac{em}{VC(\mathcal{H})})^{VC(\mathcal{H})}$.

Pf.

$$|\mathcal{H}_X| \le |\{B | B \subseteq X, \mathcal{H} \text{ shatters } B\}| \le \sum_{n=0}^{VC(\mathcal{H})} \binom{|X|}{n}$$

The second inequality adds up the number of ways to choose of set of size $n \le VC(\mathcal{H})$, since any larger set cannot be shattered by $\mathcal{H}$.

For the first inequality, we induct on $m = |X|$. (Base case: if $\mathcal{H}$ is empty, the inequality holds. For nonempty $\mathcal{H}$ and $m = 1$, either $|\mathcal{H}_X| = 1$ and $\mathcal{H}$ only shatters the empty set, or $|\mathcal{H}_X| = 2$ and $\mathcal{H}$ shatters the empty set and X.)

Induction: let $X' = \{X_2, \ldots, X_m\}$. We list the unique elements of $\mathcal{H}_X$:

$$\begin{array}{c} h_1(X_1), h_1(X_2), \ldots, h_1(X_m) \\ h_2(X_1), h_2(X_2), \ldots, h_2(X_m) \\ \vdots \\ h_k(X_1), \underbrace{h_k(X_2), \ldots, h_k(X_m)} \end{array}$$

Ignoring the first column, the table includes all elements of $\mathcal{H}_{X'}$. In addition, some elements appear twice if $\mathcal{H}_X$ generates both $(0, y_2, \ldots, y_m)$ and $(1, y_2, \ldots, y_m)$. Let $\mathcal{H}'$ be the set of hypotheses corresponding to these rows, so $|\mathcal{H}_X| = |\mathcal{H}_{X'}| + |\mathcal{H}'_{X'}|$. By the induction hypothesis,

$$\begin{aligned} |\mathcal{H}_{X'}| &\le |\{B | B \subseteq X', \mathcal{H} \text{ shatters } B\}| \\ |\mathcal{H}'_{X'}| &\le |\{B | B \subseteq X', \mathcal{H}' \text{ shatters } B\}| \\ &= |\{B \cup \{X_1\} | B \subseteq X', \mathcal{H}' \text{ shatters } B \cup \{X_1\}\}| \\ &\le |\{B \cup \{X_1\} | B \subseteq X', \mathcal{H} \text{ shatters } B \cup \{X_1\}\}| \end{aligned}$$

Finally, since subsets of X either exclude or include X_1,

$$|\mathcal{H}_X| = |\mathcal{H}_{X'}| + |\mathcal{H}'_{X'}| \le |\{B | B \subseteq X, \mathcal{H} \text{ shatters } B\}|$$

Alternative proof via *shifting*. We can change any 1 entry in the table above to a 0 unless it would produce a row that is already in the table. We repeat this until no more entries can be shifted, producing a new table $\tilde{\mathcal{H}}_X$. All possible shifts exist in $\tilde{\mathcal{H}}_X$, so if a row contains columns with 1s, then those columns are shattered by $\tilde{\mathcal{H}}$. Thus $|\mathcal{H}_X| = |\tilde{\mathcal{H}}_X| \le \{B | B \subseteq X, \tilde{\mathcal{H}} \text{ shatters } B\}$. Finally, $VC(\tilde{\mathcal{H}}) \le VC(\mathcal{H})$

because if some columns were shattered after a shift then they were shattered before the shift.

No free lunch theorem (UML page 61). If the hypothesis space $\mathcal{H}$ is unconstrained, then for any learning algorithm given at most $\frac{1}{2}|\mathcal{X}|$ training samples, the output $\hat{h}$ has significant risk:

$$P\left(R(\hat{h}) \geq \tfrac{1}{8}\right) \geq \tfrac{1}{7}$$

Cor. Any $\mathcal{H}$ with infinite VC dimension is not PAC learnable.

Pf. Let $|\mathcal{X}| = 2n$. Choose $h^* : \mathcal{X} \to \{-1, 1\}$ uniformly at random from the 2^{2n} possibilities, and choose the test distribution $\mathcal{D}$ uniformly at random over $\mathcal{X}$. The learning algorithm has at least a $\frac{1}{2}$ chance of being tested on an unseen point U_i, and in that case it can't do better than chance:

$$R_{\mathcal{D}}(\hat{h}) \geq \tfrac{1}{2} R_U(\hat{h}) = \tfrac{1}{2} \cdot \tfrac{1}{2} = \tfrac{1}{4}$$

Finally, note that $R(\hat{h}) \leq 1$, so $1 - R(\hat{h})$ is nonnegative, so by Markov's

$$P\left(R(\hat{h}) \leq \tfrac{1}{8}\right) = P\left(1 - R(\hat{h}) \geq \tfrac{7}{8}\right) \leq \tfrac{E[1-R(\hat{h})]}{7/8} \leq \tfrac{3/4}{7/8} = \tfrac{6}{7} \Rightarrow P\left(R(\hat{h}) \geq \tfrac{1}{8}\right) \geq \tfrac{1}{7}$$

FTSLT. $\mathcal{H}$ is PAC learnable iff $\mathcal{H}$ has finite VC dimension, with sample complexity

$$\tfrac{1}{\varepsilon}(d + \log \tfrac{1}{\delta}) \lesssim n^* \lesssim \tfrac{1}{\varepsilon}(d \log \tfrac{1}{\varepsilon} + \log \tfrac{1}{\delta})$$

where the bound on the right is achieved via ERM.

E.g. Threshold $\mathcal{H} = \{I(x > a) | a \in \mathbb{R}\}$ is PAC learnable.

Pf. Let a be the threshold for h^*, and define a^-, a^+ s.t. $P(x \in (a^-, a)) = \varepsilon$.

$P(\hat{h} \geq \varepsilon) = P(\{x | \min x_i < a^-\} \cup \{x | \max x_i > a^+\})$
$\leq P(\{x | \min x_i < a^-\}) + P(\{\max x_i > a^+\}) \leq 2(1-\varepsilon)^n \leq 2e^{-n\varepsilon} = \delta$

Neural Network Expressiveness

Universal approx. For Lipschitz f on $[0,1]^d$, 3-layer NN with $O((\frac{L}{\varepsilon})^d)$ neurons has $\int |f - \hat{f}|\, dx \leq \varepsilon$

NN approx. Exists 2-layer NN with $O(\frac{C}{\varepsilon})$ sigmoid neurons s.t. $\sup_x |G_0(x) - g_w(x)| \leq \varepsilon$

Depth. $\exists$ ReLU NN $[0,1] \to [0,1]$ of $O(L^2)$ depth s.t. any NN of depth L and 2^L nodes has $\int |f - g|\, dx \geq \frac{1}{32}$