Theory of Generalization and VC Dimension
Review

- **Hoeffding’s inequality**
  \[ P[|E_{in} - E_{out}| > \varepsilon] \leq 2M e^{-2\varepsilon^2n} \]

- **The growth function** for a hypothesis set \( \mathcal{H} \)
  is the maximum number of dichotomies (patterns) we can get on \( n \) data points.
  \[ m_{\mathcal{H}}(n) \leq 2^n \]
  Examples:
  - \( m_{\mathcal{H}}(n) = n+1 \) positive rays
  - \( m_{\mathcal{H}}(n) = \frac{1}{2} n^2 + \frac{1}{2} n + 1 \) positive intervals
  - \( m_{\mathcal{H}}(n) = 2^n \) convex sets

- **The break point** for a hypothesis set \( \mathcal{H} \)
  is the value of \( n \) (# of data points) for which we fail to get all possible dichotomies.
Outline

• Bounding $m_H(n)$
• Vapnik – Chervonenkis inequality
• VC dimension - the definition
• Integrating the VC dimension
• Generalization bounds
Bounding $m_\mathcal{H}(n)$

- It is proven that:
  - if there is a break point $k$ for a hypothesis set $\mathcal{H}$
  - then $m_\mathcal{H}(n)$ is polynomial.

- **Theorem**
  $$m_\mathcal{H}(n) \leq \sum_{i=0}^{k-1} \binom{n}{i}$$
  maximum power is $n^{k-1}$

**Note:**  
1. The growth function $m_\mathcal{H}(n)$ is either $2^n$ or polynomial, there is nothing in between.  
2. For a given hypothesis set $\mathcal{H}$, the break point $k$ is fixed, and does not grow with $n$. 
Three examples

\[ m_{\mathcal{H}}(n) \leq \sum_{i=0}^{k-1} \binom{n}{i} \]

- \( \mathcal{H} \) is positive rays: (break point \( k = 2 \))
  \[ m_{\mathcal{H}}(n) = n + 1 \leq n + 1 \]

- \( \mathcal{H} \) is positive intervals: (break point \( k = 3 \))
  \[ m_{\mathcal{H}}(n) = \frac{1}{2} n^2 + \frac{1}{2} n + 1 \leq \frac{1}{2} n^2 + \frac{1}{2} n + 1 \]

- \( \mathcal{H} \) is 2D perceptrons: (break point \( k = 4 \))
  \[ m_{\mathcal{H}}(n) = \ ? \leq \frac{1}{6} n^3 + \frac{5}{6} n + 1 \]
What we want?

• Instead of

$$P[|E_{in} - E_{out}| > \varepsilon] \leq 2 \quad M \quad e^{-2\varepsilon^2n}$$

can we have?

$$P[|E_{in} - E_{out}| > \varepsilon] \leq 2 \quad m_{\mathcal{H}}(n) \quad e^{-2\varepsilon^2n}$$

• It is proven that

not quite, but rather

$$P[|E_{in} - E_{out}| > \varepsilon] \leq 4 \quad m_{\mathcal{H}}(2n) \quad e^{-\frac{1}{8}\varepsilon^2n}$$

The Vapnik – Chervonenkis Inequality
holds true for any hypothesis set that has a break point
Definition of VC dimension

• The VC dimension of a hypothesis set \( \mathcal{H} \), denoted by \( d_{VC}(\mathcal{H}) \), is

  the largest value of \( n \) for which \( m_{\mathcal{H}}(n) = 2^n \),
  in other words, “the most points \( \mathcal{H} \) can shatter”

  \[
  n \leq d_{VC}(\mathcal{H}) \implies \mathcal{H} \text{ can shatter } n \text{ points}
  \]

  \[
  k > d_{VC}(\mathcal{H}) \implies k \text{ is a break point for } \mathcal{H}
  \]
The growth function

- In terms of a break point $k$:

$$m_{\mathcal{H}}(n) \leq \sum_{i=0}^{k-1} \binom{n}{i}$$

- In terms of the VC dimension $d_{VC}$:

$$m_{\mathcal{H}}(n) \leq \sum_{i=0}^{d_{VC}} \binom{n}{i}$$

maximum power is $n^{d_{VC}}$
examples

- $\mathcal{H}$ is positive rays:
  \[ d_{VC} = 1 \]

- $\mathcal{H}$ is 2D perceptrons:
  \[ d_{VC} = 3 \]

- $\mathcal{H}$ is convex sets:
  \[ d_{VC} = \infty \]
VC dimension and learning

The main result:

If $d_{VC}(\mathcal{H})$ is finite $\Rightarrow g \in \mathcal{H}$ will generalize

- Independent of the learning algorithm
- Independent of the input distribution
- Independent of the target function
VC dimension of perceptrons

For two dimensional perceptron \((d = 2)\), \(d_{VC} = 3\)

In general,

**Theorem:**

For a \(d\) - dimensional perceptron, the VC dimension is

\[
d_{VC} = d + 1
\]

what is \(d + 1\) in the perceptron?

It is the number of parameters \(w_0, w_1, \ldots, w_d\)
Putting it together

• For a hypothesis set $\mathcal{H}$, the existence of a finite $d_{VC}$ means that **learning is feasible** (i.e. generalization is possible).
  (finite $d_{VC} \Rightarrow$ Existence of a break point $\Rightarrow$ polynomial growth function)

• The value of $d_{VC}$ (or break point) tells us **the resources needed** to achieve a certain performance.

• The larger $d_{VC}$, the more complex the hypothesis set.

• Infinite $d_{VC} \Rightarrow$ no break point for $\mathcal{H} \Rightarrow m_{\mathcal{H}}(n) = 2^n$ for any $n$, and the hypothesis set shatters every set of points (e.g. Convex sets).
  – good for fitting the data set, but
  – bad for generalization.
Interpreting the VC dimension

- Two questions to answer:
  1. What does VC dimension signify?
     - “Degrees of freedom”
  2. How to apply the value of VC dimension in practice?
     - “Number of data points needed”
1. Degrees of freedom

Parameters create degrees of freedom.

- # of parameters:
  ‘analog’ degrees of freedom

- $d_{vc}$:
  equivalent ‘binary’ degrees of freedom
The usual suspects

• Positive rays ($d_{VC} = 1$):

\[ h(x) = -1 \hspace{2cm} a \hspace{1cm} h(x) = +1 \]

• Positive intervals ($d_{VC} = 2$):

\[ h(x) = -1 \hspace{2cm} a \hspace{1cm} h(x) = +1 \hspace{1cm} b \hspace{1cm} h(x) = -1 \]
Not just parameters

- Parameters may not contribute degrees of freedom:

  \[ x \rightarrow \quad \rightarrow \quad \rightarrow \rightarrow \rightarrow y \]

- \( d_{VC} \) measures the effective number of parameters (rather than the raw number of parameters)
2. Number of data points needed

- Two performance quantities in the VC inequality that we would like to be small:

\[ P[|E_{in}(g) - E_{out}(g)| > \varepsilon] \leq 4m_{\mathcal{H}} (2n)e^{\frac{1}{8} \varepsilon^2 n} \]

- In a particular situation, we want certain \( \varepsilon \) (approximation) and \( \delta \) (probability), for this how many examples do I need?

  How does \( n \) depend on \( d_{\text{VC}} \)?

- Let us look at \( n^d e^{-n} \)
Fix $n^d e^{-n} = \text{small value}$

How does $n$ change with $d$?

A rule of thumb to get a reasonable generalization:

$$n \geq 10 \cdot d_{\text{VC}}$$
Rearranging things

- Start from the VC inequality (from \( \varepsilon \) to \( \delta \))

\[
P[|E_{\text{out}}(g) - E_{\text{in}}(g)| > \varepsilon] \leq 4m_{\mathcal{H}}(2n) e^{\frac{1}{8} \varepsilon^2 n}
\]

- Get \( \varepsilon \) in terms of \( \delta \) (from \( \delta \) to \( \varepsilon \)):

\[
\delta = 4m_{\mathcal{H}}(2n) e^{\frac{1}{8} \varepsilon^2 n} \quad \Rightarrow \quad \varepsilon = \sqrt{\frac{8}{n} \ln \frac{4m_{\mathcal{H}}(2n)}{\delta}}
\]

- With probability \( \geq 1 - \delta \), \( |E_{\text{out}} - E_{\text{in}}| \leq \Omega(n, \mathcal{H}, \delta) \)
Generalization bound

With probability $\geq 1 - \delta$, $|E_{\text{out}} - E_{\text{in}}| \leq \Omega(n, H, \delta)$

$$E_{\text{out}} - E_{\text{in}} \leq \Omega$$

This is called the generalization error

\[ \downarrow \]

With probability $\geq 1 - \delta$, $E_{\text{out}} \leq E_{\text{in}} + \Omega$