Generalization Theorem

• Let $h \in \mathcal{H}$ with $|\mathcal{H}| = k$ (finite hypothesis class)
  $$h^* = \arg\min_h J_{\text{emp}}(h)$$
  (hypothesis with best training error)
  $$h_{\text{opt}} = \arg\min_h J(h)$$ (best hypothesis)

Fix $m$ and $\delta$, then $\epsilon = \sqrt{1/(2m) \log (2k/\delta)}$ and

$$J(h^*) \leq J(h_{\text{opt}}) + 2\epsilon$$
Review: learning theory setting

- \( x \in \mathcal{X} \) (input or instance)
- \( y = h(x), h \in \mathcal{H} \) (label or concept) (consider binary labels)
- \( d \in \{0, 1\} \) target outputs
- \( S = ((x(1), d(1), \ldots, x(m), d(m)) \) (sample drawn iid from some unknown distribution \( D \)), realizes a dichotomy
- \( h = \mathcal{A}(S, \mathcal{H}) \) takes a sample \( S \) and chooses a hypothesis, hypothesis is consistent with \( S \) if \( d(i) = y(i), 1 \leq i \leq m \)
VC dimension

- Consider function classes where each function labels each input as 1 or 0.
- A set of $m$ points is shattered by function class if the function class represents all $2^m$ possible labelings of the points.
- The VC dimension of a function class is the largest cardinality of points that is shattered by the function class. Example: linear threshold functions in Euclidean $n$ space has VC dimension of $n+1$.
- The VC dimension measures the complexity of the function class.
Growth functions and numbers

- Growth function: Let $X$ be a set of inputs. Let function $h$ be set of points where output label is 1.

$$\Pi_H(X) = \{ h \cap X : h \in \mathcal{H} \},$$

note that $\Pi_H(X) \subset \{0,1\}^X$ (power set). If equality, then $\mathcal{H}$ shatters $X$.

- Growth number: $\Pi_H(m) = \max_{|X|=m} |\Pi_H(X)|$

- VC dimension $VC(\mathcal{H}) = \max m$ such that

$$\Pi_H(m) = 2^m$$

If no number exists, then VC dimension is infinite.
Capabilities of Linear Threshold Functions

• Discussed three learning algorithms for linear threshold functions (LTF): PLA, SVM, LS SVM (FLDA)
• How can we describe capabilities of LTF? Given m points, how many dichotomies can homogenous LTF (HLTF) (zero threshold) realize?
• General position (GP): m points in $\mathbb{R}^n$ in GP if any subset of $k \leq \min(m,n)$ points are linearly independent.
Function Counting Theorem

Given $m$ points in $\mathbb{R}^n$ in GP there are $C(m,n)$ dichotomies that can be realized where

$$\Pi(m) = C(m,n) = 2 \sum_{k=0}^{n-1} \binom{m-1}{k}$$
FCT Proof

- \( C(m+1,n) = C(m,n) + C(m,n-1) \)
  - Given \( m \) points, add a point \( x^* \) in GP. Construct a hyperplane by projecting into null space of \( x^* \). For any dichotomy, \( x^* \) will either be ambiguous or not. Number of ambiguous points is \( C(m,n-1) \)

- Induction proof:
  - Base step: \( C(m,1) = C(1,n) = 2 \)
  - Induction step
Graphical representation of FCT proof

\[ C(m+1,n) = C(m,n) + C(m,n-1) \]
LTF Capacity

- HLTF capacity is n, LTF capacity is n+1.
- If points are not in GP capacity is less.
- Random capacity of HLTF is 2n.
- Higher capacity achieved by nonlinear threshold functions with capacity dependent on number of inputs.
- LTF can only realize a limited number of Boolean functions.
VC dimension examples

- Homogenous Linear Threshold Functions: $n$
- Linear Threshold Functions: $n+1$
- Quadratic Threshold Functions: $(n+1)(n+2)/2$
- One closed interval: $2$
- Closed intervals: $\infty$
- Axis aligned rectangles: $2n$
Growth function properties

• How does growth function, \( \Pi_H(m) \) depend on VC dimension, \( \text{VCD}(H) \)?

• Sauer’s Lemma: For any binary function class \( H \) let \( \text{VCD}(H) = d \), then

\[
\Pi_H(m) \leq \sum_{k=0}^{d} \binom{m}{k}
\]

• Corollary: If \( m \geq d \), then \( \Pi_H(m) \leq (em/d)^d \).
Bounds on Generalization Error

• For finite function classes with cardinality $k$ we have that
  \[ P \left( \bigcup_j \{ |J_{\text{emp}}(h_j) - J(h_j)| > \epsilon \} \right) \leq 2k \exp(-2m\epsilon^2) \]

• For function classes with finite $d = VCD(H)$ we have the following result known as the Vapnik Chervonenkis Inequality
  \[ P(\sup_h \{ |J_{\text{emp}}(h) - J(h)| > \epsilon \}) \leq 4 \Pi(2m) \exp(-m\epsilon^2/8) \]

• Using Sauer’s lemma we have that if $VCD(H) = d$, then can show that with probability $1-\delta$ that
  \[ J(h^*) \leq J(h_{\text{opt}}) + O(\sqrt{(d/m)\log(m/d) + (1/m)\log(1/\delta)}) \]
How does this help us with learning?

- If you have the right function class, then you can learn from training examples so that generalization error is small.
- However, this may be difficult to determine requiring using structural risk minimization (can be time consuming).
- Learning aids
  - Validation
  - Regularization
3 Validation methods

Learn on training data and test on validation data

- **Given data divide into training data and validation data**
- **Cross-validation:**
  - Divide data into k equal parts
  - Train on k-1 parts and do validation on other part
  - Repeat until every part serves as validation data
  - Average over all k learning epochs
- **Leave one out validation (expensive)**