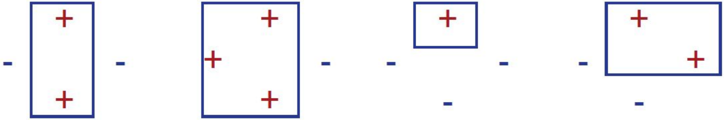
The Infinite Case: Rademacher Complexity and VC-dimension

Rationale: PAC provides no bounds for the infinite hypothesis class

$$m_{\mathcal{H}}(\epsilon,\delta) \leq m_{\mathcal{H}}^{UC}(\epsilon/2,\delta) \leq \left\lceil \frac{2\log(2|\mathcal{H}|/\delta)}{\epsilon^2} \right\rceil$$

Is it still possible to learn if |H| is infinite?

Είναι άραγε τόσο δύσκολο, ακόμα και για πολύ απλούς ταξινομητές όπως οι axis-aligned rectangles (που όμως έχει άπειρο |H|), να βρούμε ένα άνω φράγμα για το σφάλμα της γενίκευσης;



ΟΧΙ: έχουμε φράγματα για το σφάλμα της γενίκευσης και στην περίπτωση της απειρομελούς κλάσης υποθέσεων:

φράγμα Rademacher <= Growth <= VC-dimension (shattering dimension)

Setup

Nothing new ...

- Samples $S = ((x_1, y_1), \ldots, (x_m, y_m))$
- Labels $y_i = \{-1, +1\}$
- Hypothesis $h : X \rightarrow \{-1, +1\}$
- Training error: $\hat{R}(h) = \frac{1}{m} \sum_i^m \mathbb{1}\left[h(x_i) \neq y_i\right]$

An alternative derivation of training error

$$\hat{R}(h) = \frac{1}{m}\sum_{i}^{m} \mathbb{1}\left[h(x_i) \neq y_i\right] \tag{1}$$

(2)

(3)

(4)

An alternative derivation of training error

$$\hat{R}(h) = \frac{1}{m}\sum_{i}^{m} \mathbb{1}\left[h(x_i) \neq y_i\right] \tag{1}$$

$$= \frac{1}{m}\sum_{i}^{m} \begin{cases} 1 & \text{if } (h(x_i, y_i) == (1,-1) \text{ or } (-1,1) \\ 0 & (h(x_i, y_i) == (1,1) \text{ or } (-1,-1) \end{cases} \tag{2}$$

$$\tag{3}$$

$$\tag{4}$$

An alternative derivation of training error

$$
\begin{aligned}
\hat{R}(h) &= \frac{1}{m}\sum_{i}^{m} \mathbb{1}\left[h(x_i) \neq y_i\right] & (1)\\
&= \frac{1}{m}\sum_{i}^{m} \begin{cases} 1 & \text{if } \big(h(x_i, y_i\big) == \big(1, -1\big) \text{ or } \big(-1, 1\big) \\ 0 & \big(h(x_i, y_i\big) == \big(1, 1\big) \text{ or } \big(-1, -1\big) \end{cases} & (2)\\
&= \frac{1}{m}\sum_{i}^{m} \frac{1 - y_i h(x_i)}{2} & (3)\\
& & (4)
\end{aligned}
$$

An alternative derivation of training error

$$\hat{R}(h) = \frac{1}{m}\sum_{i}^{m} \mathbb{1}\left[h(x_i) \neq y_i\right] \tag{1}$$

$$= \frac{1}{m}\sum_{i}^{m} \begin{cases} 1 & \text{if } \left(h(x_i, y_i\right) == (1, -1) \text{ or } (-1, 1) \\ 0 & \left(h(x_i, y_i\right) == (1, 1) \text{ or } (-1, -1) \end{cases} \tag{2}$$

$$= \frac{1}{m}\sum_{i}^{m} \frac{1 - y_i h(x_i)}{2} \tag{3}$$

$$= \frac{1}{2} - \frac{1}{2m}\sum_{i}^{m} y_i h(x_i) \tag{4}$$

An alternative derivation of training error

$$\hat{R}(h) = \frac{1}{m}\sum_{i}^{m} \mathbb{1}\left[h(x_i) \neq y_i\right] \quad (1)$$

$$= \frac{1}{m}\sum_{i}^{m} \begin{cases} 1 & \text{if } \left(h(x_i, y_i\right) == \left(1, -1\right) \text{ or } \left(-1, 1\right) \\ 0 & \left(h(x_i, y_i\right) == \left(1, 1\right) \text{ or } \left(-1, -1\right) \end{cases} \quad (2)$$

$$= \frac{1}{m}\sum_{i}^{m} \frac{1 - y_i h(x_i)}{2} \quad (3)$$

$$= \frac{1}{2} - \frac{1}{2m}\sum_{i}^{m} y_i h(x_i) \quad (4)$$

Correlation between predictions and labels

An alternative derivation of training error

$$\hat{R}(h) = \frac{1}{m}\sum_{i}^{m} \mathbb{1}\left[h(x_i) \neq y_i\right] \tag{1}$$

$$= \frac{1}{m}\sum_{i}^{m} \begin{cases} 1 & \text{if } \left(h(x_i, y_i\right) == (1, -1) \text{ or } (-1, 1) \\ 0 & \left(h(x_i, y_i\right) == (1, 1) \text{ or } (-1, -1) \end{cases} \tag{2}$$

$$= \frac{1}{m}\sum_{i}^{m} \frac{1 - y_i h(x_i)}{2} \tag{3}$$

$$= \frac{1}{2} - \frac{1}{2m}\sum_{i}^{m} y_i h(x_i) \tag{4}$$

Minimizing training error is thus equivalent to maximizing correlation

$$\arg\max_{h} \frac{1}{m}\sum_{i}^{m} y_i h(x_i) \tag{5}$$

Playing with Correlation

Imagine where we replace true labels with *Rademacher random variables*

$$\sigma_i = \begin{cases} +1 & \text{with prob } .5 \\ -1 & \text{with prob } .5 \end{cases} \tag{6}$$

Imagine where we replace true labels with *Rademacher random variables*

$$\sigma_i = \begin{cases} +1 & \text{with prob } .5 \\ -1 & \text{with prob } .5 \end{cases} \tag{6}$$

This gives us Rademacher correlation—what's the best that a random classifier could do?

$$\hat{\mathscr{R}}_S(H) \equiv \mathbb{E}_\sigma \left[\max_{h \in H} \frac{1}{m} \sum_i^m \sigma_i h(x_i) \right] \tag{7}$$

Imagine where we replace true labels with *Rademacher random variables*

$$\sigma_i = \begin{cases} +1 & \text{with prob } .5 \\ -1 & \text{with prob } .5 \end{cases} \tag{6}$$

This gives us Rademacher correlation—what's the best that a random classifier could do?

$$\hat{\mathscr{R}}_S(H) \equiv \mathbb{E}_\sigma \left[\max_{h \in H} \frac{1}{m} \sum_i^m \sigma_i h(x_i) \right] \tag{7}$$

Notation: $\mathbb{E}_p[f] \equiv \sum_x p(x) f(x)$

Imagine where we replace true labels with *Rademacher random variables*

$$\sigma_i = \begin{cases} +1 & \text{with prob .5} \\ -1 & \text{with prob .5} \end{cases} \tag{6}$$

This gives us Rademacher correlation—what's the best that a random classifier could do?

$$\hat{\mathscr{R}}_S(H) \equiv \mathbb{E}_\sigma \left[\max_{h \in H} \frac{1}{m} \sum_i^m \sigma_i h(x_i) \right] \tag{7}$$

Note: Empirical Rademacher complexity is with respect to a sample.

Rademacher Extrema

- What are the maximum values of Rademacher correlation?

Rademacher Extrema

- What are the maximum values of Rademacher correlation?

$\|H\| = 1$	$\|H\| = 2^m$

Rademacher Extrema

- What are the maximum values of Rademacher correlation?

$|H| = 1$

$\mathbb{E}_\sigma \left[\max_{h \in H} \frac{1}{m} \sum_i^m \sigma_i h(x_i) \right]$

$|H| = 2^m$

Rademacher Extrema

- What are the maximum values of Rademacher correlation?

$|H| = 1$

$h(x_i)\mathbb{E}_\sigma\left[\frac{1}{m}\sum_i^m \sigma_i\right]$

$|H| = 2^m$

Rademacher Extrema

- What are the maximum values of Rademacher correlation?

$|H| = 1$

$$h(x_i)\mathbb{E}_\sigma\left[\frac{1}{m}\sum_i^m \sigma_i\right] = 0$$

$|H| = 2^m$

Rademacher Extrema

- What are the maximum values of Rademacher correlation?

$|H| = 1$

$h(x_i)\mathbb{E}_\sigma\left[\frac{1}{m}\sum_i^m \sigma_i\right] = 0$

$|H| = 2^m$

$\mathbb{E}_\sigma\left[\max_{h\in H}\frac{1}{m}\sum_i^m \sigma_i h(x_i)\right]$

Rademacher Extrema

- What are the maximum values of Rademacher correlation?

$|H| = 1$

$h(x_i)\mathbb{E}_\sigma \left[\frac{1}{m} \sum_i^m \sigma_i \right] = 0$

$|H| = 2^m$

$\frac{m}{m} = 1$

Rademacher Extrema

- What are the maximum values of Rademacher correlation?

$|H| = 1$

$$h(x_i)\mathbb{E}_\sigma\left[\frac{1}{m}\sum_i^m \sigma_i\right] = 0$$

$|H| = 2^m$

$$\frac{m}{m} = 1$$

- Rademacher correlation is larger for more complicated hypothesis space.
- What if you're right for stupid reasons?

Generalizing Rademacher Complexity

We can generalize Rademacher complexity to consider all sets of a particular size.

$$\mathscr{R}_m(H) = \mathbb{E}_{S \sim D^m}\left[\hat{\mathscr{R}}_S(H)\right] \tag{8}$$

Theorem

Convergence Bounds *Let F be a family of functions mapping from Z to $[0,1]$, and let sample $S = (z_1, \ldots, z_m)$ were $z_i \sim D$ for some distribution D over Z. Define $\mathbb{E}[f] \equiv \mathbb{E}_{z \sim D}[f(z)]$ and $\hat{\mathbb{E}}_S[f] \equiv \frac{1}{m}\sum_{i=1}^{m} f(z_i)$. With probability greater than $1-\delta$ for all $f \in F$:*

$$\mathbb{E}[f] \leq \hat{\mathbb{E}}_s[f] + 2\mathscr{R}_m(F) + \mathcal{O}\left(\sqrt{\frac{\ln\frac{1}{\delta}}{m}}\right) \quad (8)$$

Theorem

Convergence Bounds *Let F be a family of functions mapping from Z to* $[0,1]$*, and let sample $S = (z_1, \ldots, z_m)$ were $z_i \sim D$ for some distribution D over Z. Define $\mathbb{E}[f] \equiv \mathbb{E}_{z \sim D}[f(z)]$ and $\hat{\mathbb{E}}_S[f] \equiv \frac{1}{m}\sum_{i=1}^{m} f(z_i)$. With probability greater than $1-\delta$ for all $f \in F$:*

$$\mathbb{E}[f] \leq \hat{\mathbb{E}}_s[f] + 2\mathscr{R}_m(F) + \mathscr{O}\left(\sqrt{\frac{\ln \frac{1}{\delta}}{m}}\right) \tag{8}$$

f is a surrogate for the accuracy of a hypothesis (mathematically convenient)

Aside: McDiarmid's Inequality

If we have a function:

$$|f(x_1,\ldots,x_i,\ldots x_m) - f(x_1,\ldots,x_i',\ldots,x_m)| \leq c_i \tag{9}$$

then:

$$\Pr\left[f(x_1,\ldots,x_m) \geq \mathbb{E}\left[f(X_1,\ldots,X_m)\right] + \epsilon\right] \leq \exp\left\{\frac{-2\epsilon^2}{\sum_i^m c_i^2}\right\} \tag{10}$$

Aside: McDiarmid's Inequality

If we have a function:

$$|f(x_1,\ldots,x_i,\ldots x_m) - f(x_1,\ldots,x_i',\ldots,x_m)| \leq c_i \tag{9}$$

then:

$$\Pr\left[f(x_1,\ldots,x_m) \geq \mathbb{E}\left[f(X_1,\ldots,X_m)\right] + \epsilon\right] \leq \exp\left\{\frac{-2\epsilon^2}{\sum_i^m c_i^2}\right\} \tag{10}$$

Proof in Mohri (appendix D.7, p.442) (requires Martingale, constructing $V_k = \mathbb{E}\left[V | x_1 \ldots x_k\right] - \mathbb{E}\left[V | x_1 \ldots x_{k-1}\right]$).

Aside: McDiarmid's Inequality

Theorem D.8 (McDiarmid's inequality) *Let $X_1, \ldots, X_m \in \mathcal{X}^m$ be a set of $m \geq 1$ independent random variables and assume that there exist $c_1, \ldots, c_m > 0$ such that $f\colon \mathcal{X}^m \to \mathbb{R}$ satisfies the following conditions:*

$$\left|f(x_1, \ldots, x_i, \ldots, x_m) - f(x_1, \ldots, x_i', \ldots x_m)\right| \leq c_i\,, \tag{D.15}$$

for all $i \in [m]$ and any points $x_1, \ldots, x_m, x_i' \in \mathcal{X}$. Let $f(S)$ denote $f(X_1, \ldots, X_m)$, then, for all $\epsilon > 0$, the following inequalities hold:

$$\mathbb{P}[f(S) - \mathbb{E}[f(S)] \geq \epsilon] \leq \exp\left(\frac{-2\epsilon^2}{\sum_{i=1}^m c_i^2}\right) \tag{D.16}$$

$$\mathbb{P}[f(S) - \mathbb{E}[f(S)] \leq -\epsilon] \leq \exp\left(\frac{-2\epsilon^2}{\sum_{i=1}^m c_i^2}\right). \tag{D.17}$$

McDiarmid's inequality is used in several of the proofs in this book. It can be understood in terms of stability: if changing any of its argument affects f only in a limited way, then, its deviations from its mean can be exponentially bounded.

Εξαιτίας των δύο ανισοτήτων έχουμε το O(). Βλέμε θεώρημα 3.3 σε Mohri.

Aside: McDiarmid's Inequality

If we have a function:

$$|f(x_1, \ldots, x_i, \ldots x_m) - f(x_1, \ldots, x_i', \ldots, x_m)| \leq c_i \tag{9}$$

then:

$$\Pr\left[f(x_1, \ldots, x_m) \geq \mathbb{E}\left[f(X_1, \ldots, X_m)\right] + \epsilon\right] \leq \exp\left\{\frac{-2\epsilon^2}{\sum_i^m c_i^2}\right\} \tag{10}$$

Proof in Mohri (appendix D.7, p.442) (requires Martingale, constructing $V_k = \mathbb{E}\left[V \mid x_1 \ldots x_k\right] - \mathbb{E}\left[V \mid x_1 \ldots x_{k-1}\right]$).
What function do we care about for Rademacher complexity? Let's define

$$\Phi(S) = \sup_f \left(\mathbb{E}\left[f\right] - \hat{\mathbb{E}}_S\left[f\right]\right) = \sup_f \left(\mathbb{E}\left[f\right] - \frac{1}{m}\sum_i f(z_i)\right) \tag{11}$$

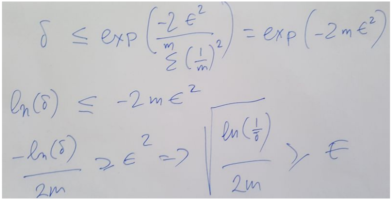
Step 1: Bounding divergence from true Expectation

Lemma

Moving to Expectation *With probability at least* $1 - \delta$,

$$\Phi(S) \leq \mathbb{E}_s[\Phi(S)] + \sqrt{\frac{\ln \frac{1}{\delta}}{2m}}$$

Since $f(z_1) \in [0,1]$, changing any z_i to z_i' in the training set will change $\frac{1}{m}\sum_i f(z_i)$ by at most $\frac{1}{m}$, so we can apply McDiarmid's inequality with $\epsilon = \sqrt{\frac{\ln \frac{1}{\delta}}{2m}}$ and $c_i = \frac{1}{m}$.



Step 2: Comparing two different empirical expectations

Define a ghost sample $S' = (z'_1, \dots, z'_m) \sim D$. How much can two samples from the same distribution vary?

Lemma

Two Different Samples

$$\mathbb{E}_S[\Phi(S)] = \mathbb{E}_S\left[\sup_f(\mathbb{E}[f] - \hat{\mathbb{E}}_S[f])\right] \quad (12)$$

$$(13)$$

Step 2: Comparing two different empirical expectations

Define a ghost sample $S' = (z'_1, \dots, z'_m) \sim D$. How much can two samples from the same distribution vary?

Lemma

Two Different Samples

$$\mathbb{E}_S[\Phi(S)] = \mathbb{E}_S\left[\sup_f(\mathbb{E}[f] - \hat{\mathbb{E}}_S[f])\right] \tag{12}$$

$$= \mathbb{E}_S\left[\sup_{f \in F}(\mathbb{E}_{S'}\left[\hat{\mathbb{E}}_{S'}[f]\right] - \hat{\mathbb{E}}_S[f])\right] \tag{13}$$

$$\tag{14}$$

The expectation is equal to the expectation of the empirical expectation of all sets S'

Step 2: Comparing two different empirical expectations

Define a ghost sample $S' = (z'_1, \dots, z'_m) \sim D$. How much can two samples from the same distribution vary?

Lemma

Two Different Samples

$$\mathbb{E}_S[\Phi(S)] = \mathbb{E}_S\left[\sup_f(\mathbb{E}[f] - \hat{\mathbb{E}}_S[f])\right] \tag{12}$$

$$= \mathbb{E}_S\left[\sup_{f\in F}(\mathbb{E}_{S'}\left[\hat{\mathbb{E}}_{S'}[f]\right] - \hat{\mathbb{E}}_S[f])\right] \tag{13}$$

$$= \mathbb{E}_S\left[\sup_{f\in F}(\mathbb{E}_{S'}\left[\hat{\mathbb{E}}_{S'}[f] - \hat{\mathbb{E}}_S[f]\right])\right] \tag{14}$$

(15)

S and S' are distinct random variables, so we can move inside the expectation

Step 2: Comparing two different empirical expectations

Define a ghost sample $S' = (z'_1, \ldots, z'_m) \sim D$. How much can two samples from the same distribution vary?

Lemma

Two Different Samples

$$\mathbb{E}_S[\Phi(S)] = \mathbb{E}_S\left[\sup_f(\mathbb{E}[f] - \hat{\mathbb{E}}_S[f])\right] \tag{12}$$

$$= \mathbb{E}_S\left[\sup_{f \in F}(\mathbb{E}_{S'}\left[\hat{\mathbb{E}}_{S'}[f] - \hat{\mathbb{E}}_S[f]\right])\right] \tag{13}$$

$$\leq \mathbb{E}_{S,S'}\left[\sup_f(\hat{\mathbb{E}}_{S'}[f] - \hat{\mathbb{E}}_S[f])\right] \tag{14}$$

The expectation of a max over some function is at least the max of that expectation over that function

Step 3: Adding in Rademacher Variables

From S, S' we'll create T, T' by swapping elements between S and S' with probability .5. This is still idependent, identically distributed (iid) from D. They have the same distribution:

$$\hat{\mathbb{E}}_{S'}[f] - \hat{\mathbb{E}}_{S}[f] \sim \hat{\mathbb{E}}_{T'}[f] - \hat{\mathbb{E}}_{T}[f] \tag{15}$$

Step 3: Adding in Rademacher Variables

From S, S' we'll create T, T' by swapping elements between S and S' with probability .5. This is still idependent, identically distributed (iid) from D. They have the same distribution:

$$\hat{\mathbb{E}}_{S'}[f] - \hat{\mathbb{E}}_S[f] \sim \hat{\mathbb{E}}_{T'}[f] - \hat{\mathbb{E}}_T[f] \tag{15}$$

Let's introduce σ_i:

$$\hat{\mathbb{E}}_{T'}[f] - \hat{\mathbb{E}}_T[f] = \frac{1}{m} \begin{cases} f(z_i) - f(z_i') \text{ with prob } .5 \\ f(z_i') - f(z_i) \text{ with prob } .5 \end{cases} \tag{16}$$

$$= \frac{1}{m} \sum_i \sigma_i (f(z_i') - f(z_i)) \tag{17}$$

Step 3: Adding in Rademacher Variables

From S, S' we'll create T, T' by swapping elements between S and S' with probability .5. This is still idependent, identically distributed (iid) from D. They have the same distribution:

$$\hat{\mathbb{E}}_{S'}[f] - \hat{\mathbb{E}}_{S}[f] \sim \hat{\mathbb{E}}_{T'}[f] - \hat{\mathbb{E}}_{T}[f] \tag{15}$$

Let's introduce σ_i:

$$\hat{\mathbb{E}}_{T'}[f] - \hat{\mathbb{E}}_{T}[f] = \frac{1}{m}\begin{cases} f(z_i) - f(z_i') \text{ with prob .5} \\ f(z_i') - f(z_i) \text{ with prob .5} \end{cases} \tag{16}$$

$$= \frac{1}{m}\sum_i \sigma_i (f(z_i') - f(z_i)) \tag{17}$$

Thus:
$\mathbb{E}_{S,S'}\left[\sup_{f\in F}\left(\hat{\mathbb{E}}_{S'}[f] - \hat{\mathbb{E}}_{S}[f]\right)\right] = \mathbb{E}_{S,S',\sigma}\left[\sup_{f\in F}\left(\sum_i \sigma_i (f(z_i') - f(z_i))\right)\right]$.

Step 4: Making These Rademacher Complexities

Before, we had $\mathbb{E}_{S,S',\sigma}\left[\sup_{f \in F} \sum_i \sigma_i (f(z_i') - f(z_i))\right]$

Before, we had $\mathbb{E}_{S,S',\sigma}\left[\sup_{f\in F}\sum_i \sigma_i (f(z_i') - f(z_i))\right]$

$$\leq \mathbb{E}_{S,S',\sigma}\left[\sup_{f\in F}\sum_i \sigma_i f(z_i') + \sup_{f\in F}\sum_i (-\sigma_i) f(z_i)\right] \tag{18}$$

(19)

Taking the sup jointly must be less than or equal the individual sup.

Step 4: Making These Rademacher Complexities

Before, we had $\mathbb{E}_{S,S',\sigma}\left[\sup_{f\in F}\sum_i \sigma_i(f(z_i') - f(z_i))\right]$

$$\leq \mathbb{E}_{S,S',\sigma}\left[\sup_{f\in F}\sum_i \sigma_i f(z_i') + \sup_{f\in F}\sum_i (-\sigma_i) f(z_i)\right] \tag{18}$$

$$\leq \mathbb{E}_{S,S',\sigma}\left[\sup_{f\in F}\sum_i \sigma_i f(z_i')\right] + \mathbb{E}_{S,S',\sigma}\left[\sup_{f\in F}\sum_i (-\sigma_i) f(z_i)\right] \tag{19}$$

(20)

Linearity

Step 4: Making These Rademacher Complexities

Before, we had $\mathbb{E}_{S,S',\sigma}\left[\sup_{f\in F}\sum_i \sigma_i(f(z'_i) - f(z_i))\right]$

$$\leq \mathbb{E}_{S,S',\sigma}\left[\sup_{f\in F}\sum_i \sigma_i f(z'_i) + \sup_{f\in F}\sum_i (-\sigma_i) f(z_i)\right] \tag{18}$$

$$\leq \mathbb{E}_{S,S',\sigma}\left[\sup_{f\in F}\sum_i \sigma_i f(z'_i)\right] + \mathbb{E}_{S,S',\sigma}\left[\sup_{f\in F}\sum_i (-\sigma_i) f(z_i)\right] \tag{19}$$

$$= \mathscr{R}_m(F) + \mathscr{R}_m(F) \tag{20}$$

Definition

Putting the Pieces Together

With probability $\geq 1-\delta$:

$$\Phi(S) \leq \mathbb{E}_S[\Phi(S)] + \sqrt{\frac{\ln \frac{1}{\delta}}{2m}} \tag{21}$$

Step 1

Putting the Pieces Together

With probability $\geq 1 - \delta$:

$$\sup_f \left(\mathbb{E}[f] - \hat{\mathbb{E}}_S[f] \right) \leq \mathbb{E}_S[\Phi(S)] + \sqrt{\frac{\ln \frac{1}{\delta}}{2m}} \tag{21}$$

Definition of Φ

Putting the Pieces Together

With probability $\geq 1-\delta$:

$$\mathbb{E}\left[f\right]-\hat{\mathbb{E}}_S\left[f\right] \leq \mathbb{E}_S\left[\Phi(S)\right]+\sqrt{\frac{\ln\frac{1}{\delta}}{2m}} \tag{21}$$

Drop the sup, still true

Putting the Pieces Together

With probability $\geq 1 - \delta$:

$$\mathbb{E}\left[f\right] - \hat{\mathbb{E}}_S\left[f\right] \leq \mathbb{E}_{S,S'}\left[\sup_f \left(\hat{\mathbb{E}}_{S'}\left[f\right] - \hat{\mathbb{E}}_S\left[f\right]\right)\right] + \sqrt{\frac{\ln\frac{1}{\delta}}{2m}} \qquad (21)$$

Step 2

Putting the Pieces Together

With probability $\geq 1-\delta$:

$$\mathbb{E}[f] - \hat{\mathbb{E}}_S[f] \leq \mathbb{E}_{S,S',\sigma}\left[\sup_{f\in F}\left(\sum_i \sigma_i(f(z_i') - f(z_i))\right)\right] + \sqrt{\frac{\ln\frac{1}{\delta}}{2m}} \qquad (21)$$

Step 3

Putting the Pieces Together

With probability $\geq 1 - \delta$:

$$\mathbb{E}[f] - \hat{\mathbb{E}}_S[f] \leq 2\mathscr{R}_m(F) + \sqrt{\frac{\ln \frac{1}{\delta}}{2m}} \tag{21}$$

Step 4

Putting the Pieces Together

With probability $\geq 1-\delta$:

$$\mathbb{E}[f] - \hat{\mathbb{E}}_S[f] \leq 2\mathscr{R}_m(F) + \sqrt{\frac{\ln\frac{1}{\delta}}{2m}} \tag{21}$$

Recall that $\hat{\mathscr{R}}_S(F) \equiv \mathbb{E}_\sigma\left[\sup_f \frac{1}{m}\sum_i \sigma_i f(z_i)\right]$, so we apply McDiarmid's inequality again (because $f \in [0,1]$):

$$\mathscr{R}_m(F) \leq \hat{\mathscr{R}}_S(F) + \sqrt{\frac{\ln\frac{1}{\delta}}{2m}} \tag{22}$$

Putting the Pieces Together

With probability $\geq 1-\delta$:

$$\mathbb{E}[f] - \hat{\mathbb{E}}_S[f] \leq 2\mathscr{R}_m(F) + \sqrt{\frac{\ln\frac{1}{\delta}}{2m}} \tag{21}$$

Recall that $\hat{\mathscr{R}}_S(F) \equiv \mathbb{E}_\sigma\left[\sup_f \frac{1}{m}\sum_i \sigma_i f(z_i)\right]$, so we apply McDiarmid's inequality again (because $f \in [0,1]$):

$$\mathscr{R}_m(F) \leq \hat{\mathscr{R}}_S(F) + \sqrt{\frac{\ln\frac{1}{\delta}}{2m}} \tag{22}$$

Putting the two together:

$$\mathbb{E}[f] \leq \hat{\mathbb{E}}_s[f] + 2\mathscr{R}_S(F) + \mathscr{O}\left(\sqrt{\frac{\ln\frac{1}{\delta}}{m}}\right) \tag{23}$$

What about hypothesis classes?

Define:

$$Z \equiv X \times \{-1, +1\} \tag{24}$$

$$f_h(x, y) \equiv \mathbb{1}\left[h(x) \neq y\right] \tag{25}$$

$$F_H \equiv \{f_h : h \in H\} \tag{26}$$

What about hypothesis classes?

Define:

$$Z \equiv X \times \{-1, +1\} \tag{24}$$

$$f_h(x, y) \equiv \mathbb{1}\left[h(x) \neq y\right] \tag{25}$$

$$F_H \equiv \{f_h : h \in H\} \tag{26}$$

We can use this to create expressions for generalization and empirical error:

$$R(h) = \mathbb{E}_{(x,y)\sim D}\left[\mathbb{1}\left[h(x) \neq y\right]\right] = \mathbb{E}\left[f_h\right] \tag{27}$$

$$\hat{R}(h) = \frac{1}{m}\sum_i \mathbb{1}\left[h(x_i) \neq y\right] = \hat{\mathbb{E}}_S\left[f_h\right] \tag{28}$$

What about hypothesis classes?

Define:

$$Z \equiv X \times \{-1, +1\} \tag{24}$$

$$f_h(x, y) \equiv \mathbb{1}\left[h(x) \neq y\right] \tag{25}$$

$$F_H \equiv \{f_h : h \in H\} \tag{26}$$

We can use this to create expressions for generalization and empirical error:

$$R(h) = \mathbb{E}_{(x,y)\sim D}\left[\mathbb{1}\left[h(x) \neq y\right]\right] = \mathbb{E}\left[f_h\right] \tag{27}$$

$$\hat{R}(h) = \frac{1}{m}\sum_i \mathbb{1}\left[h(x_i) \neq y\right] = \hat{\mathbb{E}}_S\left[f_h\right] \tag{28}$$

We can plug this into our theorem!

Generalization bounds

- We started with expectations

$$\mathbb{E}[f] \leq \hat{\mathbb{E}}_S[f] + 2\hat{\mathscr{R}}_S(F) + \mathscr{O}\left(\sqrt{\frac{\ln \frac{1}{\delta}}{m}}\right) \tag{29}$$

- We also had our definition of the generalization and empirical error:

$$R(h) = \mathbb{E}_{(x,y)\sim D}[\mathbb{1}[h(x) \neq y]] = \mathbb{E}[f_h] \quad \hat{R}(h) = \frac{1}{m}\sum_i \mathbb{1}[h(x_i) \neq y] = \hat{\mathbb{E}}_S[f_h]$$

$$\hat{\mathcal{R}}_S(F_H) = \frac{1}{2}\hat{\mathcal{R}}_S(H) \quad (30)$$

$$\widehat{\mathfrak{R}}_S(\mathcal{G}) = \frac{1}{2}\widehat{\mathfrak{R}}_{S_\mathcal{X}}(\mathcal{H}). \quad (3.16)$$

Proof: For any sample $S = ((x_1, y_1), \ldots, (x_m, y_m))$ of elements in $\mathcal{X} \times \{-1, +1\}$, by definition, the empirical Rademacher complexity of $\mathcal{G}$ can be written as:

$$\begin{aligned}
\widehat{\mathfrak{R}}_S(\mathcal{G}) &= \mathop{\mathbb{E}}_{\boldsymbol{\sigma}}\left[\sup_{h\in\mathcal{H}} \frac{1}{m}\sum_{i=1}^{m} \sigma_i 1_{h(x_i)\neq y_i}\right] \\
&= \mathop{\mathbb{E}}_{\boldsymbol{\sigma}}\left[\sup_{h\in\mathcal{H}} \frac{1}{m}\sum_{i=1}^{m} \sigma_i \tfrac{1-y_i h(x_i)}{2}\right] \\
&= \frac{1}{2}\mathop{\mathbb{E}}_{\boldsymbol{\sigma}}\left[\sup_{h\in\mathcal{H}} \frac{1}{m}\sum_{i=1}^{m} -\sigma_i y_i h(x_i)\right] \\
&= \frac{1}{2}\mathop{\mathbb{E}}_{\boldsymbol{\sigma}}\left[\sup_{h\in\mathcal{H}} \frac{1}{m}\sum_{i=1}^{m} \sigma_i h(x_i)\right] = \frac{1}{2}\widehat{\mathfrak{R}}_{S_\mathcal{X}}(\mathcal{H}),
\end{aligned}$$

where we used the fact that $1_{h(x_i)\neq y_i} = (1 - y_i h(x_i))/2$ and the fact that for a fixed $y_i \in \{-1, +1\}$, σ_i and $-y_i\sigma_i$ are distributed in the same way. □

Generalization bounds

- We started with expectations

$$\mathbb{E}[f] \leq \hat{\mathbb{E}}_S[f] + 2\hat{\mathscr{R}}_S(F) + \mathscr{O}\left(\sqrt{\frac{\ln \frac{1}{\delta}}{m}}\right) \tag{29}$$

- We also had our definition of the generalization and empirical error:

$$R(h) = \mathbb{E}_{(x,y)\sim D}[\mathbb{1}[h(x) \neq y]] = \mathbb{E}[f_h] \quad \hat{R}(h) = \frac{1}{m}\sum_i \mathbb{1}[h(x_i) \neq y] = \hat{\mathbb{E}}_S[f_h]$$

- Combined with the previous result:

$$\hat{\mathscr{R}}_S(F_H) = \frac{1}{2}\hat{\mathscr{R}}_S(H) \tag{30}$$

- All together:

$$R(h) \leq \hat{R}(h) + \hat{\mathscr{R}}_S(H) + \mathscr{O}\left(\sqrt{\frac{\log \frac{1}{\delta}}{m}}\right) \tag{31}$$

Wrapup

- Interaction of data, complexity, and accuracy
- Still very theoretical
- Next up: How to evaluate generalizability of specific hypothesis classes

Recap

- Rademacher complexity provides nice guarantees

$$R(h) \leq \hat{R}(h) + \hat{\mathscr{R}}_S(H) + \mathscr{O}\left(\sqrt{\frac{\log \frac{1}{\delta}}{2m}}\right) \quad (32)$$

- But in practice hard to compute for real hypothesis classes
- Is there a relationship with simpler combinatorial measures?

Growth Function

Define the **growth function** $\Pi_H : \mathbb{N} \rightarrow \mathbb{N}$ for a hypothesis set H as:

$$\forall m \in \mathbb{N}, \Pi_H(m) \equiv \max_{\{x_1,\ldots,x_m\} \in X} \left| \{ (h(x_1), \ldots, h(x_m) : h \in H \} \right| \tag{33}$$

Define the **growth function** $\Pi_H : \mathbb{N} \rightarrow \mathbb{N}$ for a hypothesis set H as:

$$\forall m \in \mathbb{N}, \Pi_H(m) \equiv \max_{\{x_1,\ldots,x_m\} \in X} \left| \{(h(x_1), \ldots, h(x_m) : h \in H\} \right| \tag{33}$$

i.e., the number of ways m points can be classified using H.

Rademacher Complexity vs. Growth Function

If G is a function taking values in $\{-1, +1\}$, then

$$\mathscr{R}_m(G) \leq \sqrt{\frac{2 \ln \Pi_G(m)}{m}} \tag{34}$$

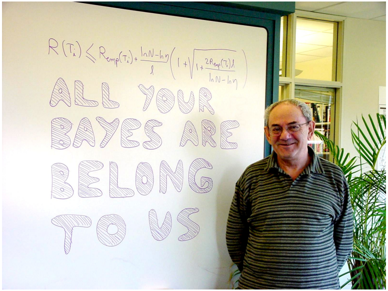
Uses Masart's lemma (Theorem 3.7)

Corollary 3.9 (Growth function generalization bound) *Let $\mathcal{H}$ be a family of functions taking values in $\{-1, +1\}$. Then, for any $\delta > 0$, with probability at least $1 - \delta$, for any $h \in \mathcal{H}$,*

$$R(h) \leq \widehat{R}_S(h) + \sqrt{\frac{2 \log \Pi_{\mathcal{H}}(m)}{m}} + \sqrt{\frac{\log \frac{1}{\delta}}{2m}}. \tag{3.22}$$

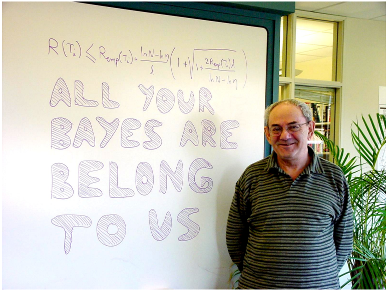
Not very convenient since it requires computing $\Pi_H(m), \forall m$

Vapnik-Chervonenkis Dimension



$$\mathrm{VC}(H) \equiv \max\{m : \Pi_H(m) = 2^m\} \tag{35}$$

Vapnik-Chervonenkis Dimension



$$\mathrm{VC}(H) \equiv \max\{m : \Pi_H(m) = 2^m\} \tag{35}$$

The size of the largest set that can be fully shattered (θρυμματιστεί) by H.

Entropy Properties of a Decision Rule Class with ML abilities - Alexey Chervonenkis lecture

VC Dimension for Hypotheses

- Need upper and lower bounds
- Lower bound: example
- Upper bound: Prove that no set of $d + 1$ points can be shattered by H (harder)

Intervals

What is the VC dimension of $[a, b]$ intervals on the real line.

Intervals

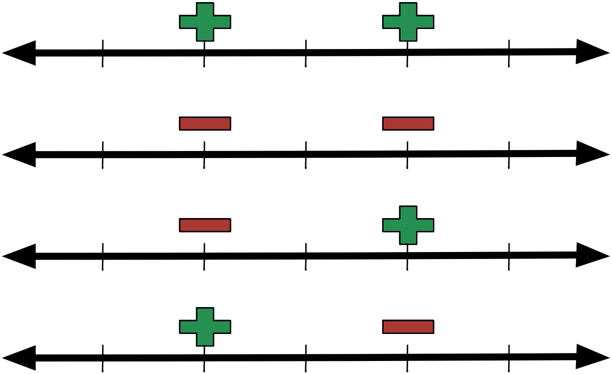
What is the VC dimension of $[a, b]$ intervals on the real line.

- What about two points?

Intervals

What is the VC dimension of $[a, b]$ intervals on the real line.

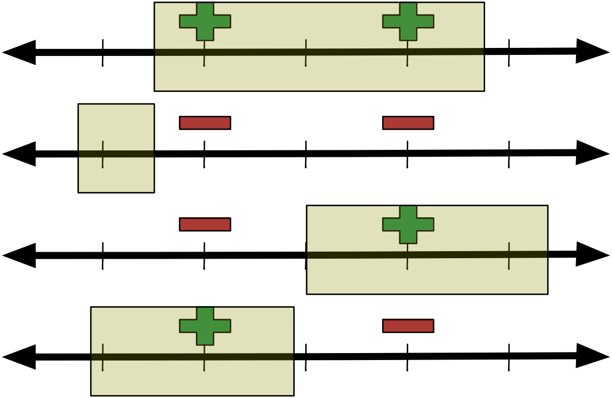
- What about two points?



Intervals

What is the VC dimension of $[a, b]$ intervals on the real line.

- What about two points?



Intervals

What is the VC dimension of $[a, b]$ intervals on the real line.

- Two points can be perfectly classified, so VC dimension ≥ 2

Intervals

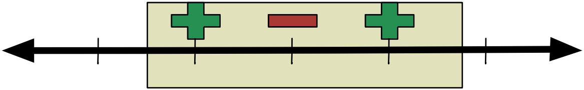
What is the VC dimension of $[a, b]$ intervals on the real line.

- Two points can be perfectly classified, so VC dimension ≥ 2
- What about three points?

Intervals

What is the VC dimension of $[a, b]$ intervals on the real line.

- Two points can be perfectly classified, so VC dimension ≥ 2
- What about three points?



Intervals

What is the VC dimension of $[a, b]$ intervals on the real line.

- Two points can be perfectly classified, so VC dimension ≥ 2
- What about three points?
- **No set** of three points can be shattered

Intervals

What is the VC dimension of $[a, b]$ intervals on the real line.

- Two points can be perfectly classified, so VC dimension ≥ 2
- What about three points?
- **No set** of three points can be shattered
- Thus, VC dimension of intervals is 2

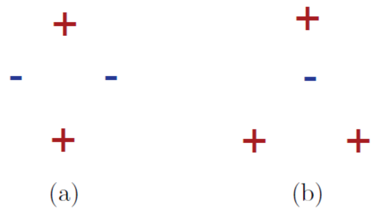


Figure 3.2
Unrealizable dichotomies for four points using hyperplanes in $\mathbb{R}^2$. (a) All four points lie on the convex hull. (b) Three points lie on the convex hull while the remaining point is interior.

Axis-aligned-rectangles

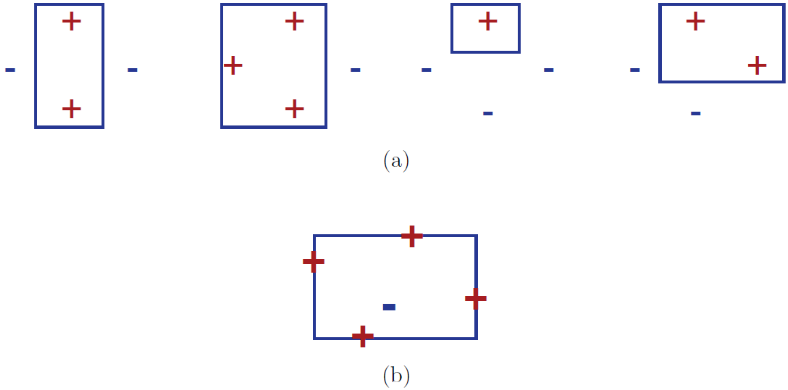


Figure 3.3

VC-dimension of axis-aligned rectangles. (a) Examples of realizable dichotomies for four points in a diamond pattern. (b) No sample of five points can be realized if the interior point and the remaining points have opposite labels.

Sine Functions

- Consider hypothesis that classifies points on a line as either being above or below a sine wave

$$\{t \rightarrow \sin(\omega x) : \omega \in \mathbb{R}\} \tag{36}$$

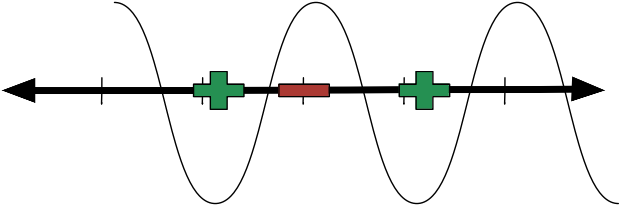
- Can you shatter three points?

Sine Functions

- Consider hypothesis that classifies points on a line as either being above or below a sine wave

$$\{t \rightarrow \sin(\omega x) : \omega \in \mathbb{R}\} \tag{36}$$

- Can you shatter three points?



Sine Functions

- Consider hypothesis that classifies points on a line as either being above or below a sine wave

$$\{t \rightarrow \sin(\omega x) : \omega \in \mathbb{R}\} \tag{36}$$

- Can you shatter four points?

Sine Functions

- Consider hypothesis that classifies points on a line as either being above or below a sine wave

$$\{t \rightarrow \sin(\omega x) : \omega \in \mathbb{R}\} \tag{36}$$

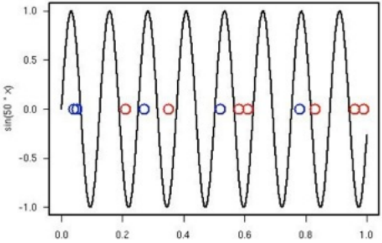
- How many points can you shatter?

Sine Functions

- Consider hypothesis that classifies points on a line as either being above or below a sine wave

$$\{t \rightarrow \sin(\omega x) : \omega \in \mathbb{R}\} \tag{36}$$

- Thus, VC dim of sine on line is ∞



Connecting VC with growth function

VC dimension obviously encodes the complexity of a hypothesis class, but we want to connect that to Rademacher complexity and the growth function so we can prove generalization bounds.

Connecting VC with growth function

VC dimension obviously encodes the complexity of a hypothesis class, but we want to connect that to Rademacher complexity and the growth function so we can prove generalization bounds.

Theorem

Sauer's Lemma *Let H be a hypothesis set with VC dimension d. Then* $\forall m \in \mathbb{N}$

$$\Pi_H(m) \leq \sum_{i=0}^{d} \binom{m}{i} \equiv \Phi_d(m) \tag{37}$$

VC dimension obviously encodes the complexity of a hypothesis class, but we want to connect that to Rademacher complexity and the growth function so we can prove generalization bounds.

Theorem

Sauer's Lemma *Let H be a hypothesis set with VC dimension d. Then* $\forall m \in \mathbb{N}$

$$\Pi_H(m) \leq \sum_{i=0}^{d} \binom{m}{i} \equiv \Phi_d(m) \tag{37}$$

This is good because the sum when multiplied out becomes $\binom{m}{i} = \frac{m \cdot (m-1) \dots}{i!} = \mathcal{O}\left(m^d\right)$. When we plug this into the learning error limits: $\log(\Pi_H(2m)) = \log(\mathcal{O}\left(m^d\right)) = \mathcal{O}\left(d \log m\right)$.

Sauer's Lemma

Definition. Growth Function:

$$\Pi_F(n) = \max\{|F_{|s} \ : \ s \subseteq \mathcal{X}, |s| = n\}$$

Definition. VC dimension

$$d_{\text{VC}}(F) = \max\{|s| : \ s \subseteq \mathcal{X}, f \text{ shatters } s\}$$

Here, we say that a family of binary functions F shatters a set $\mathcal{S} \in \mathcal{X}$ if $F_{|\mathcal{S}} = 2^{|\mathcal{S}|}$.

Theorem 2.1. Sauer's Lemma: If $F \subseteq \{\pm 1\}^{\mathcal{X}}$ and $d_{VC} = d$, then $\Pi_F(n) \le \sum_{i=0}^{d} \binom{n}{i}$. And for $n \ge d$, $\Pi_F(n) \le \left(\frac{en}{d}\right)^d$

That means: if $d_{VC}(F)$ is ∞, we always get exponential growth function; however, if $d_{VC}(F) = d$ is finite, the growth function increases exponentially up to d and polynomially for $n > d$.

Proof. Fix $(x_1, \ldots, x_n) \in \mathcal{X}$, and consider a table containing the values of functions in the class $F_{|x_1^n}$ restricted to the sample. For instance, consider the following example:

	x_1	x_2	x_3	x_4	x_5
f_1	-	+	-	+	+
f_2	+	-	-	+	+
f_3	+	+	+	-	+
f_4	-	+	+	-	-
f_5	-	-	-	+	-

Each row is one possible evaluation of the functions in F on the fixed sample, and the cardinality of $F_{|x_1^n}$ equals to the number of rows. We transform the table by "shifting" columns.

Definition. shifting column i: for each row, replace a "+" in column i with a "-" unless it would produce a row that is already in the table.

After applying the shifting operation in order from x_1 to x_5, we get the table($F^*_{|x_1^n}$):

	x_1	x_2	x_3	x_4	x_5
f_1	-	+	-	-	-
f_2	-	-	-	+	+
f_3	-	-	-	-	+
f_4	-	-	-	-	-
f_5	-	-	-	+	-

Observations:

(1) Size of the table unchanged, because the rows in $F^*_{|x_1^n}$ are still distinct;

(2) The table $F^*_{|x_1^n}$ exhibits "closed below" property, i.e., for each row containing a "+", replacing that "+" with a "-" produces another row in the table.

(3) $d_{VC}(F^*_{|x_1^n}) \le d_{VC}(F_{|x_1^n})$. To see this, consider the application of the shifting operation to a single column, and notice that if F^* (after shifting) shatters a subset of columns, then so does F (before shifting).

Therefore,

(3) and (2) $\Rightarrow$ F^* can not have more than d "+"'s in a row. Hence, #row of $F^* \le \sum_{i=0}^{d} \binom{n}{i}$;

(1) $\Rightarrow$ $|F_{|x_1^n}| \le \sum_{i=1}^{d} \binom{n}{i}$

Wait a minute . . .

Is this combinatorial expression really $\mathcal{O}(m^d)$?

$$
\begin{aligned}
\sum_{i=0}^{d}\binom{m}{i} &\leq \sum_{i=0}^{d}\binom{m}{i}\left(\frac{m}{d}\right)^{d-i} && \frac{m}{d} \geq 1 \\
&\leq \sum_{i=0}^{m}\binom{m}{i}\left(\frac{m}{d}\right)^{d-i} && \text{προσθέτω θετικούς όρους} \\
&= \left(\frac{m}{d}\right)^{d}\sum_{i=0}^{m}\binom{m}{i}\left(\frac{d}{m}\right)^{i} && (1+x)^m = \sum_{i=0}^{m}\binom{m}{i}x^i \\
&= \left(\frac{m}{d}\right)^{d}\left(1+\frac{d}{m}\right)^{m} \leq \left(\frac{m}{d}\right)^{d} e^{d}. && (1-x) \leq e^{-x}
\end{aligned}
$$

Combining our previous generalization results with Sauer's lemma, we have that for a hypothesis class H with VC dimension d, for any $\delta > 0$ with probability at least $1-\delta$, for any $h \in H$,

$$R(h) \leq \hat{R}(h) + \sqrt{\frac{2d \log \frac{em}{d}}{m}} + \sqrt{\frac{\log \frac{1}{\delta}}{2m}} \tag{43}$$

Whew!

- Infinite hypothesis class is PAC-learnable iff it has finite VC dimension
- We're now going to see if we can find an algorithm that has good VC dimension
- And works well in practice . . . Support Vector Machines