EE6550 Machine Learning

Lecture Two – Rademacher Complexity and VC-Dimension

Chung-Chin Lu

Department of Electrical Engineering

National Tsing Hua University

February 20, 2017

Motivation

- Is efficient learning from a finite sample possible when the hypothesis set $\mathcal{H}$ is infinite?
- Are there useful measures of complexity for infinite hypothesis sets?

The Contents of This Lecture

- Rademacher complexity
- Growth function
- VC-dimension
- Lower bounds

Definition of Empirical Rademacher Complexity

- $\mathcal{G}$: a family of measurable functions from a set $\mathscr{Z}$ to $[a, b]$.
  - $(\mathscr{Z}, \tilde{\mathcal{F}}, \tilde{P})$: a probability space associated with the set $\mathscr{Z}$.
- $\tilde{S} = (z_1, z_2, \ldots, z_m)$: a sample of m elements drawn i.i.d. from $\mathscr{Z}$ according to the probability distribution $\tilde{P}$.
- $\sigma_1, \sigma_2, \ldots, \sigma_m$: m binary variables taking values in $\{-1, +1\}$, called Rademacher variables.

The empirical Rademacher complexity of $\mathcal{G}$ w.r.t. the sample $\tilde{S}$ is defined as

$$\begin{aligned}\hat{\mathfrak{R}}_{\tilde{S}}(\mathcal{G}) &= \frac{1}{2^m} \sum_{\sigma_1, \sigma_2, \ldots, \sigma_m \in \{-1,+1\}} \sup_{g \in \mathcal{G}} \frac{1}{m} \sum_{i=1}^{m} \sigma_i g(z_i) \\ &= \mathop{E}_{\sigma_1, \sigma_2, \ldots, \sigma_m} \left[\sup_{g \in \mathcal{G}} \frac{1}{m} \sum_{i=1}^{m} \sigma_i g(z_i) \right].\end{aligned}$$

Interpretation of Empirical Rademacher Complexity

The empirical Rademacher complexity can be re-written as

$$\hat{\mathfrak{R}}_{\tilde{S}}(\mathcal{G}) = \frac{1}{2^m} \sum_{\boldsymbol{\sigma} \in \{-1,+1\}^m} \sup_{g \in \mathcal{G}} \frac{\boldsymbol{\sigma} \cdot \boldsymbol{g}_{\tilde{S}}}{m} = \underset{\boldsymbol{\sigma}}{E} \left[\sup_{g \in \mathcal{G}} \frac{\boldsymbol{\sigma} \cdot \boldsymbol{g}_{\tilde{S}}}{m} \right],$$

- $\boldsymbol{g}_{\tilde{S}} = [g(z_1), g(z_2), \ldots, g(z_m)]^T$: the vector of values taken by function g over the sample $\tilde{S}$;
- $\boldsymbol{\sigma} = [\sigma_1, \sigma_2, \ldots, \sigma_m]^T$: the vector of Rademacher variables;
- $\frac{\boldsymbol{\sigma} \cdot \boldsymbol{g}_{\tilde{S}}}{m}$: the correlation between $\boldsymbol{g}_{\tilde{S}}$ and $\boldsymbol{\sigma}$;
- $\sup_{g \in \mathcal{G}} \frac{\boldsymbol{\sigma} \cdot \boldsymbol{g}_{\tilde{S}}}{m}$: a measure of how well the function class G correlates with a pattern of $\boldsymbol{\sigma}$ over the sample $\tilde{S}$;

- $\hat{\mathfrak{R}}_{\tilde{S}}(\mathcal{G})$: a measure of how well the function class G correlates with $\boldsymbol{\sigma}$ on $\tilde{S}$, averaged over all possible patterns of $\boldsymbol{\sigma}$;
  - The set of all possible patterns of $\boldsymbol{\sigma}$ is the same as the ensemble of outcomes of a white binary noise vector;
- $\hat{\mathfrak{R}}_{\tilde{S}}(\mathcal{G})$ describes the richness of the family G: richer or more complex families G can generate more vectors $\boldsymbol{g}_{\tilde{S}}$ and thus better correlate with white binary noise vector, on average.

Definition of Rademacher Complexity

- $\mathcal{G}$: a family of measurable functions from a set $\mathscr{Z}$ to $[a, b]$.
  - $(\mathscr{Z}, \tilde{\mathcal{F}}, \tilde{P})$: a probability space associated with the set $\mathscr{Z}$.
- $\tilde{S} = (z_1, z_2, \ldots, z_m)$: a sample of m elements drawn i.i.d. from $\mathscr{Z}$ according to the probability distribution $\tilde{P}$.
  - $\boldsymbol{g}_{\tilde{S}} = [g(z_1), g(z_2), \ldots, g(z_m)]^T$: the vector of values by the function $g \in \mathcal{G}$ on $\tilde{S}$.
- $\boldsymbol{\sigma} = [\sigma_1, \sigma_2, \ldots, \sigma_m]^T$: the vector of m Rademacher variables.

The Rademacher complexity of the family $\mathcal{G}$ is defined as

$$\mathfrak{R}_m(\mathcal{G}) = \underset{\tilde{S} \sim \tilde{P}_m}{E} [\hat{\mathfrak{R}}_{\tilde{S}}(\mathcal{G})] = \underset{\tilde{S} \sim \tilde{P}_m}{E} \left[\frac{1}{2^m} \sum_{\boldsymbol{\sigma} \in \{-1,+1\}^m} \sup_{g \in \mathcal{G}} \frac{\boldsymbol{\sigma} \cdot \boldsymbol{g}_{\tilde{S}}}{m} \right].$$

McDiarmid's Inequality

- $(\mathscr{Z}, \tilde{\mathcal{F}}, \tilde{P})$: a probability space.
- $\tilde{S} = (z_1, z_2, \ldots, z_m)$: a sample of m elements drawn i.i.d. from $\mathscr{Z}$ according to the probability distribution $\tilde{P}$.
- $f : \mathscr{Z}^m \to \mathbb{R}$: a measurable function for which there exist $c_1, c_2, \ldots, c_m > 0$ such that

$$|f(z_1, \ldots, z_i, \ldots, z_m) - f(z_1, \ldots, z_i', \ldots, z_m)| \leq c_i$$

for all $i \in [1, m]$ and $z_1, z_1', \ldots, z_i, z_i', \ldots, z_m, z_m'$ in $\mathscr{Z}$.

Then for any $\epsilon > 0$, we have

$$\tilde{P}_m \left(f(\tilde{S}) - \underset{\tilde{S} \sim \tilde{P}_m}{E} [f(\tilde{S})] > \epsilon \right) < e^{-\frac{2\epsilon^2}{\sum_{i=1}^m c_i^2}},$$

$$\tilde{P}_m \left(f(\tilde{S}) - \underset{\tilde{S} \sim \tilde{P}_m}{E} [f(\tilde{S})] < -\epsilon \right) < e^{-\frac{2\epsilon^2}{\sum_{i=1}^m c_i^2}}.$$

Proof. For a proof, please see pp. 371-373 of the *Foundation of Machine Learning* textbook. □

Confidence Interval of $\mathfrak{R}_m(\mathcal{G})$

- $\mathcal{G}$: a family of measurable functions from a set $\mathscr{Z}$ to $[a, b]$.
  - $(\mathscr{Z}, \tilde{\mathcal{F}}, \tilde{P})$: a probability space associated with the set $\mathscr{Z}$.
- $\tilde{S} = (z_1, \ldots, z_i, \ldots, z_m)$: a sample of m elements drawn i.i.d. from $\mathscr{Z}$ according to $\tilde{P}$.
- $\hat{\mathfrak{R}}_{\tilde{S}}(\mathcal{G}) = \frac{1}{2^m} \sum_{\boldsymbol{\sigma} \in \{-1,+1\}^m} \sup_{g \in \mathcal{G}} \frac{\boldsymbol{\sigma} \cdot \boldsymbol{g}_{\tilde{S}}}{m}$: empirical Rademacher complexity of the family $\mathcal{G}$ w.r.t. S.
  - $\underset{\tilde{S} \sim \tilde{P}_m}{E} [\hat{\mathfrak{R}}_{\tilde{S}}(\mathcal{G})] = \mathfrak{R}_m(\mathcal{G})$.
- $\tilde{S}' = (z_1, \ldots, z_i', \ldots, z_m)$: a sample different from $\tilde{S}$ exactly in the ith point.

Now we have

$$\begin{aligned}
\hat{\mathfrak{R}}_{\tilde{S}}(\mathcal{G}) - \hat{\mathfrak{R}}_{\tilde{S}'}(\mathcal{G}) &= \frac{1}{2^m} \sum_{\boldsymbol{\sigma} \in \{-1,+1\}^m} \left(\sup_{g \in \mathcal{G}} \frac{\boldsymbol{\sigma} \cdot \boldsymbol{g}_{\tilde{S}}}{m} - \sup_{g \in \mathcal{G}} \frac{\boldsymbol{\sigma} \cdot \boldsymbol{g}_{\tilde{S}'}}{m} \right) \\
&\leq \frac{1}{2^m} \sum_{\boldsymbol{\sigma} \in \{-1,+1\}^m} \sup_{g \in \mathcal{G}} \left(\frac{\boldsymbol{\sigma} \cdot \boldsymbol{g}_{\tilde{S}}}{m} - \frac{\boldsymbol{\sigma} \cdot \boldsymbol{g}_{\tilde{S}'}}{m} \right) \\
&= \frac{1}{2^m} \sum_{\boldsymbol{\sigma} \in \{-1,+1\}^m} \sup_{g \in \mathcal{G}} \frac{\sigma_i (g(z_i) - g(z_i'))}{m} \\
&= \frac{1}{2} \sum_{\sigma_i \in \{-1,+1\}} \sup_{g \in \mathcal{G}} \frac{\sigma_i (g(z_i) - g(z_i'))}{m} \leq \frac{(b-a)}{m}.
\end{aligned}$$

Similarly,

$$\hat{\mathfrak{R}}_{\tilde{S}'}(\mathcal{G}) - \hat{\mathfrak{R}}_{\tilde{S}}(\mathcal{G}) \leq \frac{(b-a)}{m}$$

so that

$$\left| \hat{\mathfrak{R}}_{\tilde{S}}(\mathcal{G}) - \hat{\mathfrak{R}}_{\tilde{S}'}(\mathcal{G}) \right| \leq \frac{(b-a)}{m}.$$

As a measurable function from $\mathscr{Z}^m$ to $\mathbb{R}$, $\hat{\mathfrak{R}}_{\tilde{S}}(\mathcal{G})$ satisfies the hypotheses of McDiarmid's inequality with $c_i = (b-a)/m$ for all $i \in [1, m]$ so that for any $\epsilon > 0$, we have

$$\begin{aligned} \tilde{P}_m \left(\hat{\mathfrak{R}}_{\tilde{S}}(\mathcal{G}) - \mathfrak{R}_m(\mathcal{G}) > \epsilon \right) &< e^{-\frac{2\epsilon^2 m}{(b-a)^2}}, \\ \tilde{P}_m \left(\hat{\mathfrak{R}}_{\tilde{S}}(\mathcal{G}) - \mathfrak{R}_m(\mathcal{G}) < -\epsilon \right) &< e^{-\frac{2\epsilon^2 m}{(b-a)^2}}. \end{aligned}$$

By letting $\frac{\delta}{2} = e^{-\frac{2\epsilon^2 m}{(b-a)^2}}$, we have $\epsilon = (b-a)\sqrt{\frac{\ln \frac{2}{\delta}}{2m}}$. Equivalently, for any $\delta > 0$, with probability at least $1 - \delta/2$,

$$\mathfrak{R}_m(\mathcal{G}) \leq \hat{\mathfrak{R}}_{\tilde{S}}(\mathcal{G}) + (b-a)\sqrt{\frac{\ln \frac{2}{\delta}}{2m}}$$

and with probability at least $1 - \delta/2$,

$$\hat{\mathfrak{R}}_{\tilde{S}}(\mathcal{G}) - (b-a)\sqrt{\frac{\ln \frac{2}{\delta}}{2m}} \leq \mathfrak{R}_m(\mathcal{G}).$$

Finally, with probability at least $1 - \delta$,

$$\hat{\mathfrak{R}}_{\tilde{S}}(\mathcal{G}) - (b-a)\sqrt{\frac{\ln \frac{2}{\delta}}{2m}} \leq \mathfrak{R}_m(\mathcal{G}) \leq \hat{\mathfrak{R}}_{\tilde{S}}(\mathcal{G}) + (b-a)\sqrt{\frac{\ln \frac{2}{\delta}}{2m}}.$$

Rademacher Complexity Bound

Theorem 3.1: Let

- $(\mathscr{Z}, \tilde{\mathcal{F}}, \tilde{P})$: a probability space.
- $\mathcal{G}$: a family of measurable functions from $\mathscr{Z}$ to $[a, b]$.
- $\tilde{S} = (z_1, z_2, \ldots, z_m)$: a sample of m elements drawn i.i.d. from $\mathscr{Z}$ according to the probability distribution $\tilde{P}$.

For any $\delta > 0$, with probability at least $1 - \delta$, each of the following holds for all g in $\mathcal{G}$:

$$\begin{aligned} \underset{z\sim\tilde{P}}{E}[g(z)] &\leq \frac{1}{m}\sum_{i=1}^{m} g(z_i) + 2\mathfrak{R}_m(\mathcal{G}) + (b-a)\sqrt{\frac{\ln\frac{1}{\delta}}{2m}} \\ \underset{z\sim\tilde{P}}{E}[g(z)] &\leq \frac{1}{m}\sum_{i=1}^{m} g(z_i) + 2\hat{\mathfrak{R}}_{\tilde{S}}(\mathcal{G}) + 3(b-a)\sqrt{\frac{\ln\frac{2}{\delta}}{2m}}. \end{aligned}$$

Proof of Theorem 3.1

- $(\mathscr{Z}, \tilde{\mathcal{F}}, \tilde{P})$: a probability space.
- $\tilde{S} = (z_1, z_2, \ldots, z_m)$: a sample of m elements drawn i.i.d. from $\mathscr{Z}$ according to the probability distribution $\tilde{P}$.
- g: a member of the family $\mathcal{G}$.
- $\hat{A}_{\tilde{S}}(g) = \frac{1}{m}\sum_{i=1}^{m} g(z_i)$: the empirical average of g on $\tilde{S}$.
  - $\underset{\tilde{S}\sim\tilde{P}_m}{E}[\hat{A}_{\tilde{S}}(g)] = \underset{z\sim\tilde{P}}{E}[g(z)]$.
- $\Phi(\tilde{S}) \triangleq \sup_{g\in\mathcal{G}} \left(\underset{z\sim\tilde{P}}{E}[g(z)] - \hat{A}_{\tilde{S}}(g) \right)$: a measurable function from $\mathscr{Z}^m$ to $\mathbb{R}$ such that for any $i \in [1, m]$ and any sample $\tilde{S}' = (z_1, \ldots, z'_i, \ldots, z_m)$ which differs from the sample $\tilde{S}$ exactly in the ith points, we have

$$
\begin{aligned}
& \Phi(\tilde{S}') - \Phi(\tilde{S}) \\
= \; & \sup_{g \in \mathcal{G}} \left(\underset{z \sim \tilde{P}}{E}[g(z)] - \hat{A}_{\tilde{S}'}(g) \right) - \sup_{g \in \mathcal{G}} \left(\underset{z \sim \tilde{P}}{E}[g(z)] - \hat{A}_{\tilde{S}}(g) \right) \\
\leq \; & \sup_{g \in \mathcal{G}} \left(\left(\underset{z \sim \tilde{P}}{E}[g(z)] - \hat{A}_{\tilde{S}'}(g) \right) - \left(\underset{z \sim \tilde{P}}{E}[g(z)] - \hat{A}_{\tilde{S}}(g) \right) \right) \\
= \; & \sup_{g \in \mathcal{G}} \left(\hat{A}_{\tilde{S}}(g) - \hat{A}_{\tilde{S}'}(g) \right) \\
= \; & \sup_{g \in \mathcal{G}} \frac{g(z_i) - g(z_i')}{m} \leq \frac{b-a}{m}.
\end{aligned}
$$

Similarly,

$$
\Phi(\tilde{S}) - \Phi(\tilde{S}') \leq \frac{b-a}{m}
$$

so that

$$
\left| \Phi(\tilde{S}) - \Phi(\tilde{S}') \right| \leq \frac{b-a}{m}.
$$

- $\Phi(\tilde{S})$ satisfies the hypotheses of the McDiarmid's inequality with $c_i = (b-a)/m$ for all $i \in [1,m]$ and then for any $\epsilon > 0$, we have

$$\tilde{P}_m \left(\Phi(\tilde{S}) - \underset{\tilde{S} \sim \tilde{P}_m}{E} [\Phi(\tilde{S})] > \epsilon \right) < e^{-\frac{2\epsilon^2 m}{(b-a)^2}}$$

- By letting $\delta/2 = e^{-\frac{2\epsilon^2 m}{(b-a)^2}}$, we have $\epsilon = (b-a)\sqrt{\frac{\ln \frac{2}{\delta}}{2m}}$.
- Thus for any $\delta > 0$, with probability at least $1 - \delta/2$,

$$\Phi(\tilde{S}) \leq \underset{\tilde{S} \sim \tilde{P}_m}{E} [\Phi(\tilde{S})] + (b-a)\sqrt{\frac{\ln \frac{2}{\delta}}{2m}}.$$

- Next we bound the expectation $\underset{\tilde{S} \sim \tilde{P}_m}{E} [\Phi(\tilde{S})]$ of $\Phi(\tilde{S})$ as follows:

$$
\begin{aligned}
\underset{\tilde{S}\sim\tilde{P}_m}{E}[\Phi(\tilde{S})] &= \underset{\tilde{S}\sim\tilde{P}_m}{E}\left[\sup_{g\in\mathcal{G}}\left(\underset{z\sim\tilde{P}}{E}[g(z)] - \hat{A}_{\tilde{S}}(g)\right)\right] \\
&= \underset{\tilde{S}\sim\tilde{P}_m}{E}\left[\sup_{g\in\mathcal{G}}\left(\underset{\tilde{S}''\sim\tilde{P}_m}{E}[\hat{A}_{\tilde{S}''}(g)] - \hat{A}_{\tilde{S}}(g)\right)\right] \\
&\qquad \tilde{S}'' \text{ is drawn independent of } \tilde{S} \\
&= \underset{\tilde{S}\sim\tilde{P}_m}{E}\left[\sup_{g\in\mathcal{G}}\left(\underset{\tilde{S}''\sim\tilde{P}_m}{E}\left[\hat{A}_{\tilde{S}''}(g) - \hat{A}_{\tilde{S}}(g)\right]\right)\right] \\
&\leq \underset{(\tilde{S},\tilde{S}'')\sim\tilde{P}_{2m}}{E}\left[\sup_{g\in\mathcal{G}}\left(\hat{A}_{\tilde{S}''}(g) - \hat{A}_{\tilde{S}}(g)\right)\right] \\
&= \underset{(\tilde{S},\tilde{S}'')\sim\tilde{P}_{2m}}{E}\left[\sup_{g\in\mathcal{G}}\frac{1}{m}\sum_{i=1}^{m}(g(z_i'') - g(z_i))\right].
\end{aligned}
$$

- $\sigma_1, \sigma_2, \ldots, \sigma_m$: m Rademacher variables. For each pattern of the m Rademacher variables, we have

$$\begin{aligned} & \underset{(\tilde{S},\tilde{S}'')\sim\tilde{P}_{2m}}{E}\left[\sup_{g\in\mathcal{G}}\frac{1}{m}\sum_{i=1}^{m}\sigma_i(g(z_i'')-g(z_i))\right] \\ = & \underset{(\tilde{S},\tilde{S}'')\sim\tilde{P}_{2m}}{E}\left[\sup_{g\in\mathcal{G}}\frac{1}{m}\sum_{i=1}^{m}(g(z_i'')-g(z_i))\right] \end{aligned}$$

  by swapping z_i'' and z_i between $\tilde{S}''$ and $\tilde{S}$ if $\sigma_i = -1$ and all z_i's and z_i'''s drawn i.i.d from $\mathcal{Z}$ according to $\tilde{P}$. Thus we have

$$\begin{aligned} & \underset{(\tilde{S},\tilde{S}'')\sim\tilde{P}_{2m}}{E}\left[\sup_{g\in\mathcal{G}}\frac{1}{m}\sum_{i=1}^{m}(g(z_i'')-g(z_i))\right] \\ = & \underset{(\tilde{S},\tilde{S}'')\sim\tilde{P}_{2m}}{E}\left[\frac{1}{2^m}\sum_{\sigma_1,\sigma_2,\ldots,\sigma_m\in\{-1,+1\}}\sup_{g\in\mathcal{G}}\frac{1}{m}\sum_{i=1}^{m}\sigma_i(g(z_i'')-g(z_i))\right] \end{aligned}$$

- Since

$$\begin{aligned} & \sup_{g\in\mathcal{G}} \frac{1}{m}\sum_{i=1}^{m} \sigma_i (g(z_i'') - g(z_i)) \\ \leq \quad & \sup_{g\in\mathcal{G}} \frac{1}{m}\sum_{i=1}^{m} \sigma_i g(z_i'') + \sup_{g\in\mathcal{G}} \frac{1}{m}\sum_{i=1}^{m} -\sigma_i g(z_i''), \end{aligned}$$

  we have

$$
\begin{aligned}
& \underset{(\tilde{S},\tilde{S}'')\sim\tilde{P}_{2m}}{E}\left[\frac{1}{2^m}\sum_{\sigma_1,\sigma_2,\ldots,\sigma_m\in\{-1,+1\}}\sup_{g\in\mathcal{G}}\frac{1}{m}\sum_{i=1}^{m}\sigma_i(g(z_i'')-g(z_i))\right] \\
\leq\ & \underset{(\tilde{S},\tilde{S}'')\sim\tilde{P}_{2m}}{E}\left[\frac{1}{2^m}\sum_{\sigma_1,\sigma_2,\ldots,\sigma_m\in\{-1,+1\}}\sup_{g\in\mathcal{G}}\frac{1}{m}\sum_{i=1}^{m}\sigma_i g(z_i'')\right] \\
& + \underset{(\tilde{S},\tilde{S}'')\sim\tilde{P}_{2m}}{E}\left[\frac{1}{2^m}\sum_{\sigma_1,\sigma_2,\ldots,\sigma_m\in\{-1,+1\}}\sup_{g\in\mathcal{G}}\frac{1}{m}\sum_{i=1}^{m}-\sigma_i g(z_i)\right] \\
=\ & \underset{\tilde{S}''\sim\tilde{P}_m}{E}\left[\frac{1}{2^m}\sum_{\sigma_1,\sigma_2,\ldots,\sigma_m\in\{-1,+1\}}\sup_{g\in\mathcal{G}}\frac{1}{m}\sum_{i=1}^{m}\sigma_i g(z_i'')\right] \\
& + \underset{\tilde{S}\sim\tilde{P}_m}{E}\left[\frac{1}{2^m}\sum_{\sigma_1,\sigma_2,\ldots,\sigma_m\in\{-1,+1\}}\sup_{g\in\mathcal{G}}\frac{1}{m}\sum_{i=1}^{m}\sigma_i g(z_i)\right]
\end{aligned}
$$

- An upper bound of $\underset{\tilde{S}\sim\tilde{P}_m}{E}[\Phi(\tilde{S})]$ is:

$$\begin{aligned}\underset{\tilde{S}\sim\tilde{P}_m}{E}[\Phi(\tilde{S})] &\leq 2\underset{\tilde{S}\sim\tilde{P}_m}{E}\left[\frac{1}{2^m}\sum_{\sigma_1,\sigma_2,\ldots,\sigma_m\in\{-1,+1\}}\sup_{g\in\mathcal{G}}\frac{1}{m}\sum_{i=1}^{m}\sigma_i g(z_i)\right] \\ &= 2\mathfrak{R}_m(\mathcal{G}).\end{aligned}$$

- Now for any $\delta > 0$, with probability at least $1-\delta$, we have

$$\sup_{g\in\mathcal{G}}\left(\underset{z\sim\tilde{P}}{E}[g(z)] - \hat{A}_{\tilde{S}}(g)\right) \leq 2\mathfrak{R}_m(\mathcal{G}) + (b-a)\sqrt{\frac{\ln\frac{1}{\delta}}{2m}},$$

equivalently for all $g \in \mathcal{G}$,

$$\underset{z\sim\tilde{P}}{E}[g(z)] \leq \frac{1}{m}\sum_{i=1}^{m} g(z_i) + 2\mathfrak{R}_m(\mathcal{G}) + (b-a)\sqrt{\frac{\ln\frac{1}{\delta}}{2m}}. \quad (1)$$

- Also from the study of confidence interval, with probability at

least $1 - \delta/2$,

$$\mathfrak{R}_m(\mathcal{G}) \leq \hat{\mathfrak{R}}_{\tilde{S}}(\mathcal{G}) + (b-a)\sqrt{\frac{\ln \frac{2}{\delta}}{2m}}. \tag{2}$$

- Combining (1) and (2) we conclude that with probability at least $1 - \delta$, we have

$$\underset{z \sim \tilde{P}}{E}[g(z)] \leq \frac{1}{m}\sum_{i=1}^{m} g(z_i) + 2\hat{\mathfrak{R}}_{\tilde{S}}(\mathcal{G}) + 3(b-a)\sqrt{\frac{\ln \frac{2}{\delta}}{2m}}$$

for all $g \in \mathcal{G}$. □

Empirical Rademacher Complexity of a Hypothesis Set $\mathcal{H}$

Lemma 3.1: Let

- $\mathscr{I}$: the input space of all possible items ω, associated with a probability space $(\mathscr{I}, \mathcal{F}, P)$.
- $\mathscr{Y} = \mathscr{Y}' = \{-1, +1\}$: a binary label (output) space.
- $L(y', y) = 1_{y' \neq y}$: 0-1 loss function.
- $h : \mathscr{I} \to \{-1, +1\}$: a hypothesis in the hypothesis set $\mathcal{H}$.
- $g_h : \mathscr{I} \times \{-1, +1\} \to [0, 1]$: the error function associated with h, defined as $g_h(\omega, y) \triangleq L(h(\omega), y)$.
- $\mathcal{G} = \{g_h \mid h \in \mathcal{H}\}$: the family of error functions associated to $\mathcal{H}$ for the zero-one loss.

- $\mathscr{Z} = \mathscr{I} \times \{-1, +1\}$: the input set of error functions g_h associated with a probability space $(\mathscr{Z}, \tilde{\mathcal{F}}, \tilde{P})$ where $\tilde{P}$ is an extension of P from on $\mathcal{F}$ to on $\tilde{\mathcal{F}} = \mathcal{F} \times 2^{\{-1,+1\}}$.
- $\tilde{S} = (z_1, z_2, \ldots, z_m)$: a sample of m labeled items $z_i = (\omega_i, y_i)$ drawn i.i.d. from $\mathscr{Z}$ according to $\tilde{P}$.
- $S = (\omega_1, \omega_2, \ldots, \omega_m)$: associated with the labeled sample $\tilde{S}$, a sample of m items drawn i.i.d. from $\mathscr{I}$ according to P.

Then we have

$$\hat{\mathfrak{R}}_{\tilde{S}}(\mathcal{G}) = \frac{1}{2}\hat{\mathfrak{R}}_S(\mathcal{H}).$$

Proof.

$$
\begin{aligned}
& \hat{\mathfrak{R}}_{\tilde{S}}(\mathcal{G}) \\
= \quad & \frac{1}{2^m} \sum_{\sigma_1,\sigma_2,\ldots,\sigma_m \in \{-1,+1\}} \sup_{g_h \in \mathcal{G}} \frac{1}{m} \sum_{i=1}^{m} \sigma_i g_h((\omega_i, y_i)) \\
= \quad & \frac{1}{2^m} \sum_{\sigma_1,\sigma_2,\ldots,\sigma_m \in \{-1,+1\}} \sup_{h \in \mathcal{H}} \frac{1}{m} \sum_{i=1}^{m} \sigma_i \frac{1 - y_i h(\omega_i)}{2} \\
= \quad & \frac{1}{2^m} \sum_{\sigma_1,\sigma_2,\ldots,\sigma_m \in \{-1,+1\}} \left(\frac{1}{m} \sum_{i=1}^{m} \frac{\sigma_i}{2} + \sup_{h \in \mathcal{H}} \frac{1}{m} \sum_{i=1}^{m} -\frac{\sigma_i y_i h(\omega_i)}{2} \right) \\
= \quad & \frac{1}{2} \frac{1}{2^m} \sum_{\sigma_1,\sigma_2,\ldots,\sigma_m \in \{-1,+1\}} \sup_{h \in \mathcal{H}} \frac{1}{m} \sum_{i=1}^{m} -\sigma_i y_i h(\omega_i) \\
= \quad & \frac{1}{2} \frac{1}{2^m} \sum_{\sigma_1,\sigma_2,\ldots,\sigma_m \in \{-1,+1\}} \sup_{h \in \mathcal{H}} \frac{1}{m} \sum_{i=1}^{m} \sigma_i h(\omega_i) = \frac{1}{2} \hat{\mathfrak{R}}_S(\mathcal{H}).
\end{aligned}
$$

□

A Remark

- For all $m \geq 1$, $\mathfrak{R}_m(\mathcal{G}) = \frac{1}{2}\mathfrak{R}_m(\mathcal{H})$.

Rademacher Complexity Bound - Binary Classification

Theorem 3.2: Let

- $c : \mathscr{I} \to \{-1, +1\}$: a fixed but unknown target concept in the concept class $\mathcal{C}$.
- $\mathcal{H}$: the hypothesis set, consisting of possibly infinitely many hypotheses $h : \mathscr{I} \to \{-1, +1\}$.
- $S = (\omega_1, \ldots, \omega_m)$: a sample of size m drawn i.i.d. from the input space $\mathscr{I}$ according to an unknown distribution P.
- $\tilde{S} = ((\omega_1, c(\omega_1)), \ldots, (\omega_m, c(\omega_m))$: the labeled sample corresponding to S.
- $L(y', y) = 1_{y' \neq y}$: 0-1 loss function.
- $g_h : \mathscr{I} \times \{-1, +1\} \to [0, 1]$: the error function associated with h, defined as $g_h(\omega, y) \triangleq L(h(\omega), y)$.

- $\mathcal{G} = \{g_h \mid h \in \mathcal{H}\}$: the family of error functions associated to $\mathcal{H}$ for the zero-one loss.
- $\hat{A}_{\tilde{S}}(g_h) = \frac{1}{m}\sum_{i=1}^{m} g_h(\omega_i, c(\omega_i)) = \frac{1}{m}\sum_{i=1}^{m} L(h(\omega_i), c(\omega_i)) = \hat{R}_S(h)$.
- $\underset{z\sim\tilde{P}}{E}[g_h(z)] = \underset{\tilde{S}\sim\tilde{P}_m}{E}[\hat{A}_{\tilde{S}}(g_h)] = \underset{S\sim P_m}{E}[\hat{R}_S(h)] = R(h)$.

For any $\delta > 0$, with probability at least $1-\delta$, each of the following holds for all h in $\mathcal{H}$:

$$R(h) \leq \hat{R}_S(h) + \mathfrak{R}_m(\mathcal{H}) + \sqrt{\frac{\ln\frac{1}{\delta}}{2m}},$$

$$R(h) \leq \hat{R}_S(h) + \hat{\mathfrak{R}}_S(\mathcal{H}) + 3\sqrt{\frac{\ln\frac{2}{\delta}}{2m}}.$$

Remarks

- The second uniform generalization bound in Theorem 3.2 is data-dependent and can be informative if $\hat{\mathfrak{R}}_S(\mathcal{H})$ could be computed.
- Note that

$$\begin{aligned}
\hat{\mathfrak{R}}_S(\mathcal{H}) &= \frac{1}{2^m} \sum_{\sigma_1,\sigma_2,\ldots,\sigma_m \in \{-1,+1\}} \sup_{h\in\mathcal{H}} \frac{1}{m} \sum_{i=1}^{m} \sigma_i h(\omega_i) \\
&= \frac{1}{2^m} \sum_{\sigma_1,\sigma_2,\ldots,\sigma_m \in \{-1,+1\}} \sup_{h\in\mathcal{H}} \frac{1}{m} \sum_{i=1}^{m} -\sigma_i h(\omega_i) \\
&= \frac{1}{2^m} \sum_{\sigma_1,\sigma_2,\ldots,\sigma_m \in \{-1,+1\}} -\inf_{h\in\mathcal{H}} \frac{1}{m} \sum_{i=1}^{m} \sigma_i h(\omega_i),
\end{aligned}$$

  which indicates that the computation of $\hat{\mathfrak{R}}_S(\mathcal{H})$ is equivalent to solving 2^m empirical risk minimization problems.

The Contents of This Lecture

- Rademacher complexity
- Growth function
- VC-dimension
- Lower bounds

Definition: Growth Function

- $\mathscr{I}$: the input space of all possible items ω.
- $\mathscr{Y}'$: the output space.
- $\mathcal{H}$: the hypothesis set, consisting of possibly infinitely many hypotheses $h : \mathscr{I} \to \mathscr{Y}'$.

The growth function $\Pi_{\mathcal{H}} : \mathbb{N} \to \mathbb{N}$ for a hypothesis set $\mathcal{H}$ is defined by:

$$\forall\, m \in \mathbb{N}, \Pi_{\mathcal{H}}(m) = \sup_{\omega_1,\ldots,\omega_m \in \mathscr{I}} \left|\{(h(\omega_1,\ldots,h(\omega_m)) \mid h \in \mathcal{H}\}\right|.$$

Remarks

- When the output space $\mathscr{Y}'$ is finite, the sup operation can be replaced by the max operation.
- $\Pi_{\mathcal{H}}(m)$ is the maximum number of distinct ways in which m points can be classified using hypotheses in $\mathcal{H}$.
- This provides another measure of the richness of the hypothesis set $\mathcal{H}$.
- However, unlike the Rademacher complexity, this measure does not depend on the distribution, it is purely combinatorial.

Massart's Lemma

Theorem 3.3: Let

- $A \subseteq \mathbb{R}^m$: a finite set of m-dimensional real vectors.
- $r \triangleq \max_{\boldsymbol{x} \in A} ||\boldsymbol{x}||_2$.

Then

$$\begin{aligned} & \underset{\boldsymbol{\sigma}}{E}\left[\frac{1}{m}\max_{\boldsymbol{x}\in A}\boldsymbol{\sigma}\cdot\boldsymbol{x}\right] \\ = & \frac{1}{2^m}\sum_{\sigma_1,\sigma_2,\ldots,\sigma_m\in\{-1,+1\}}\max_{(x_1,\ldots,x_m)\in A}\frac{1}{m}\sum_{i=1}^{m}\sigma_i x_i \\ \leq & \frac{r\sqrt{2\ln|A|}}{m}. \end{aligned}$$

Proof. For any $t > 0$, $f(x) = e^{tx}$ is a convex function on $\mathbb{R}$. Thus

we have

$$
\begin{aligned}
& e^{t\left(\sum_{\sigma_1,\sigma_2,\ldots,\sigma_m\in\{-1,+1\}} \frac{1}{2^m} \max_{(x_1,\ldots,x_m)\in A} \sum_{i=1}^m \sigma_i x_i\right)} \\
\leq & \sum_{\sigma_1,\sigma_2,\ldots,\sigma_m\in\{-1,+1\}} \frac{1}{2^m} e^{t\left(\max_{(x_1,\ldots,x_m)\in A} \sum_{i=1}^m \sigma_i x_i\right)} \\
= & \frac{1}{2^m} \sum_{\sigma_1,\sigma_2,\ldots,\sigma_m\in\{-1,+1\}} \max_{(x_1,\ldots,x_m)\in A} e^{t\left(\sum_{i=1}^m \sigma_i x_i\right)} \\
& \text{since } f(x) = e^{tx} \text{ is monotone increasing for } t > 0 \\
\leq & \frac{1}{2^m} \sum_{\sigma_1,\sigma_2,\ldots,\sigma_m\in\{-1,+1\}} \sum_{(x_1,\ldots,x_m)\in A} \prod_{i=1}^m e^{t\sigma_i x_i} \\
= & \sum_{\boldsymbol{x}\in A} \left(\frac{1}{2} \sum_{\sigma_1\in\{-1,+1\}} e^{t\sigma_1 x_1}\right) \cdots \left(\frac{1}{2} \sum_{\sigma_m\in\{-1,+1\}} e^{t\sigma_m x_m}\right).
\end{aligned}
$$

By regarding $X = x_i\sigma_i$ as a r.v. such that the two possible values x_i and $-x_i$ occur equally probably, we have $E[X] = 0$ and by

applying Azuma-Hoeffding lemma,

$$\frac{1}{2}\sum_{\sigma_i\in\{-1,+1\}} e^{t\sigma_i x_i} = E\left[e^{tX}\right] \leq e^{t^2(2x_i)^2/8}.$$

Now we have

$$\begin{aligned} & e^{t\left(\sum_{\sigma_1,\sigma_2,\ldots,\sigma_m\in\{-1,+1\}} \frac{1}{2^m} \max_{\boldsymbol{x}\in A} \sum_{i=1}^m \sigma_i x_i\right)} \\ \leq & \sum_{\boldsymbol{x}\in A}\prod_{i=1}^m e^{t^2x_i^2/2} = \sum_{\boldsymbol{x}\in A} e^{t^2\|\boldsymbol{x}\|_2^2/2} \leq |A|e^{t^2r^2/2}, \end{aligned}$$

which implies that

$$\frac{1}{2^m}\sum_{\sigma_1,\sigma_2,\ldots,\sigma_m\in\{-1,+1\}} \max_{\boldsymbol{x}\in A}\sum_{i=1}^m \sigma_i x_i \leq \frac{\ln|A|}{t} + \frac{tr^2}{2} \;\forall\; t>0.$$

By choosing $t = \frac{\sqrt{2\ln|A|}}{r}$ which minimizes the upper bound, we

have

$$\frac{1}{2^m} \sum_{\sigma_1,\sigma_2,\ldots,\sigma_m \in \{-1,+1\}} \max_{\boldsymbol{x} \in A} \sum_{i=1}^{m} \sigma_i x_i \leq r\sqrt{2 \ln |A|}$$

and then

$$\frac{1}{2^m} \sum_{\sigma_1,\sigma_2,\ldots,\sigma_m \in \{-1,+1\}} \max_{\boldsymbol{x} \in A} \frac{1}{m} \sum_{i=1}^{m} \sigma_i x_i \leq \frac{r\sqrt{2 \ln |A|}}{m}.$$

□

Growth Function Bound on Emp. Rademacher Complexity

Corollary 3.1: Let

- $\mathcal{G}$: a family of measurable functions from a set $\mathscr{Z}$ to $\{-1, +1\}$.
  - $(\mathscr{Z}, \tilde{\mathcal{F}}, \tilde{P})$: a probability space associated with the set $\mathscr{Z}$.
- $\tilde{S} = (z_1, z_2, \ldots, z_m)$: a sample of m elements drawn i.i.d. from $\mathscr{Z}$ according to the probability distribution $\tilde{P}$.
  - $\boldsymbol{g}_{\tilde{S}} = [g(z_1), g(z_2), \ldots, g(z_m)]^T$: the vector of values by the function $g \in \mathcal{G}$ on $\tilde{S}$.

Then we have

$$\hat{\mathfrak{R}}_{\tilde{S}}(\mathcal{G}) \leq \sqrt{\frac{2 \ln \Pi_{\mathcal{G}}(m)}{m}}.$$

Proof. Since each function $g \in \mathcal{G}$ takes values in $\{-1, +1\}$,

$$A_{m,\tilde{S}} \triangleq \{\boldsymbol{g}_{\tilde{S}} \mid g \in \mathcal{G}\} \subseteq \{-1, +1\}^m \subseteq \mathbb{R}^m$$

is a finite set with

$$r_{m,\tilde{S}} \triangleq \max_{\boldsymbol{x} \in A_{m,\tilde{S}}} \|\boldsymbol{x}\|_2 = \sqrt{m}.$$

Then we have

$$\begin{aligned}
\hat{\mathfrak{R}}_{\tilde{S}}(\mathcal{G}) &= \frac{1}{2^m} \sum_{\boldsymbol{\sigma} \in \{-1,+1\}^m} \sup_{g \in \mathcal{G}} \frac{\boldsymbol{\sigma} \cdot \boldsymbol{g}_{\tilde{S}}}{m} = \frac{1}{2^m} \sum_{\boldsymbol{\sigma} \in \{-1,+1\}^m} \sup_{\boldsymbol{x} \in A_{m,\tilde{S}}} \frac{\boldsymbol{\sigma} \cdot \boldsymbol{x}}{m} \\
&\leq \frac{r_{m,\tilde{S}} \sqrt{2 \ln |A_{m,\tilde{S}}|}}{m}.
\end{aligned}$$

by Massart's lemma. Since

$$|A_{m,\tilde{S}}| \leq \sup_{\tilde{S} \in \mathcal{Z}^m} |A_{m,\tilde{S}}| = \Pi_{\mathcal{G}}(m)$$

by the definition of growth function $\Pi_{\mathcal{G}}(m)$ of $\mathcal{G}$, we have

$$\hat{\mathfrak{R}}_{\tilde{S}}(\mathcal{G}) \leq \frac{\sqrt{m}\sqrt{2\ln \Pi_{\mathcal{G}}(m)}}{m} = \sqrt{\frac{2\ln \Pi_{\mathcal{G}}(m)}{m}}.$$

□

- By taking expectation, we have $\mathfrak{R}_m(\mathcal{G}) \leq \sqrt{\frac{2\ln \Pi_{\mathcal{G}}(m)}{m}}$.

Growth Function Generalization Bound - Binary Classification

Corollary 3.2: Let

- $c : \mathscr{I} \to \{-1, +1\}$: a fixed but unknown target concept in the concept class $\mathcal{C}$.
- $\mathcal{H}$: the hypothesis set, consisting of possibly infinitely many hypotheses $h : \mathscr{I} \to \{-1, +1\}$.
- $S = (\omega_1, \ldots, \omega_m)$: a sample of size m drawn i.i.d. from the input space $\mathscr{I}$ according to an unknown distribution P.
- $L(y', y) = 1_{y' \neq y}$: 0-1 loss function.

For any $\delta > 0$, with probability at least $1 - \delta$, we have

$$\forall\ h \in \mathcal{H}, \quad R(h) \leq \hat{R}_S(h) + \sqrt{\frac{2 \ln \Pi_{\mathcal{H}}(m)}{m}} + \sqrt{\frac{\ln \frac{1}{\delta}}{2m}}.$$

The Contents of This Lecture

- Rademacher complexity
- Growth function
- VC-dimension
- Lower bounds

Dichotomy and Shattering for Binary Classification

- $\mathscr{I}$: the input space of all possible items ω.
- $\mathscr{Y}' = \mathscr{Y} = \{-1, +1\}$: the output, label space.
- $\mathcal{H}$: a hypothesis set, consisting of possibly infinitely many hypotheses $h : \mathscr{I} \to \mathscr{Y}$.
- Dichotomy: a dichotomy of a set S of items is one of the possible ways of labeling the items of S.
  - Two hypotheses h and h' in $\mathcal{H}$ realize the same dichotomy of a set S if $h(\omega) = h'(\omega)$ for all $\omega \in S$.
- Shattering: a set S of $m \geq 1$ points is said to be shattered by a hypothesis set $\mathcal{H}$ when $\mathcal{H}$ realizes all possible dichotomies of S, that is when $\Pi_{\mathcal{H}}(m) = 2^m$.

Definition: Vapnik-Chervonenkis(VC)-Dimension

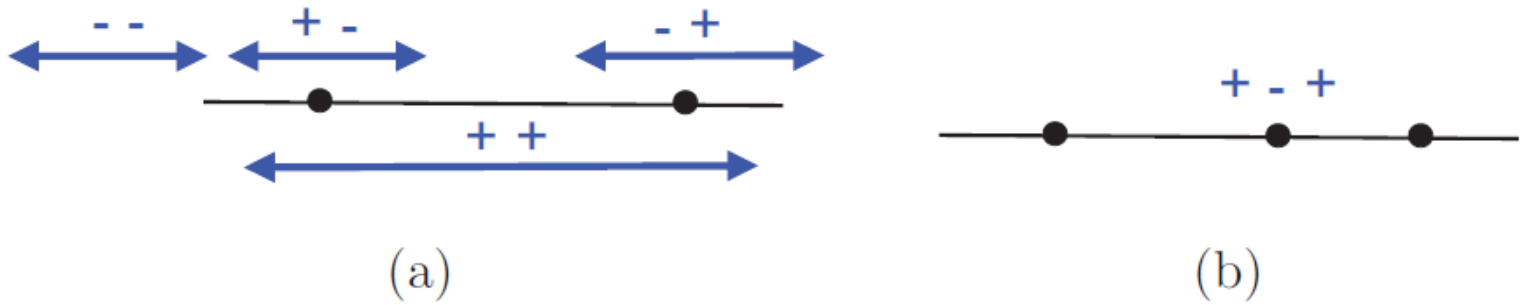
- $\mathscr{I}$: the input space of all possible items ω.
- $\mathscr{Y}' = \mathscr{Y} = \{-1, +1\}$: the output, label space.
- $\mathcal{H}$: the hypothesis set, consisting of possibly infinitely many hypotheses $h : \mathscr{I} \to \mathscr{Y}$.

The VC-dimension of a hypothesis set $\mathcal{H}$ is the size of the largest set of items that can be fully shattered by $\mathcal{H}$:

$$\mathrm{VC\,dim}(\mathcal{H}) \triangleq \sup\{m \geq 1 \mid \Pi_{\mathcal{H}}(m) = 2^m\}.$$

Example: Intervals on the Real Line

- $\mathbb{R}$: the input space.
- $\mathscr{Y}' = \mathscr{Y} = \{-1, +1\}$: the output, label space.
- $\mathcal{H}$: the family of all intervals on $\mathbb{R}$.
- $\mathrm{VC}\dim(\mathcal{H}) \geq 2$: for any two points on the real line, all four dichotomies $(-,-), (-,+), (+,-), (+,+)$ can be realized by $\mathcal{H}$.
- $\mathrm{VC}\dim(\mathcal{H}) < 3$: for any three points on the real line, the dichotomy $(+,-,+)$ cannot be realized by $\mathcal{H}$.
- $\mathrm{VC}\dim(\mathcal{H}) = 2$.



VC-dimension of intervals on the real line. (a) Any two points can be shattered. (b) No sample of three points can be shattered as the $(+, -, +)$ labeling cannot be realized.

Example: Hyperplanes in $\mathbb{R}^2$

- $\mathbb{R}^2$: the input space.
- $\mathscr{Y}' = \mathscr{Y} = \{-1, +1\}$: the output, label space.
- $\mathcal{H}$: the family of all hyperplanes in $\mathbb{R}^2$.
- $\mathrm{VC}\dim(\mathcal{H}) \geq 2$: any two points can be shattered.
- $\mathrm{VC}\dim(\mathcal{H}) < 3$: any three points cannot be shattered.
  - Collinear: the dichotomy $(+, -, +)$ cannot be realized.
  - Non-collinear: the dichotomy $(+, +, +)$ cannot be realized.
- $\mathrm{VC}\dim(\mathcal{H}) = 2$.

Example: Hyperplanes in $\mathbb{R}^d$

- $\mathbb{R}^d$: the input space.
- $\mathscr{Y}' = \mathscr{Y} = \{-1, +1\}$: the output, label space.
- $\mathcal{H}$: the family of all hyperplanes in $\mathbb{R}^d$.
- $\mathrm{VC}\dim(\mathcal{H}) \geq d$: the d standard unit vectors $\boldsymbol{e}_i = (0, \ldots, 0, 1, 0, \ldots, 0), 1 \leq i \leq d$, in $\mathbb{R}^d$ can be shattered.

  Proof. Let $(w_1, w_2, \ldots, w_d)$, $w_i \in \{-1, +1\}$, be an arbitrary dichotomy of the d points $\boldsymbol{e}_i, 1 \leq i \leq d$. Define a hyperplane in $\mathbb{R}^d$:

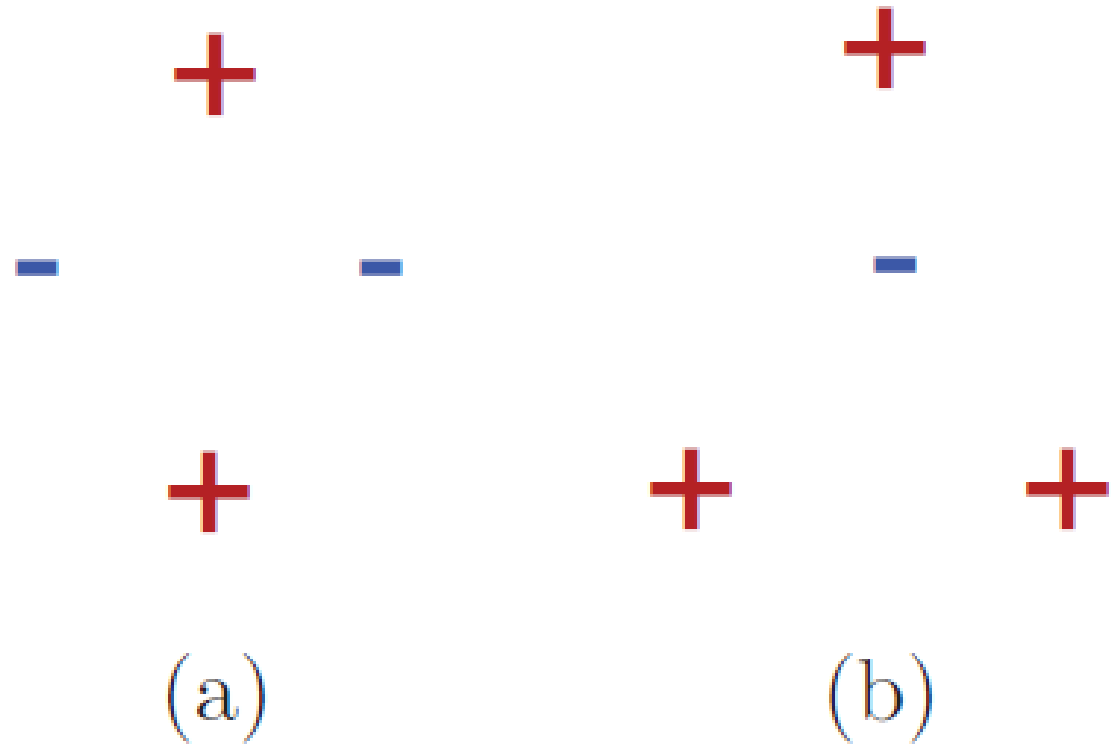
  $$w_1 x_1 + w_2 x_2 + \cdots + w_d x_d = 1$$

  where $x_1, \ldots, x_d$ are variables in $\mathbb{R}$. It can be seen that $\boldsymbol{e}_i$ is in that hyperplane if and only if $w_i = +1$. Thus the dichotomy $(w_1, w_2, \ldots, w_d)$ can be realized by that hyperplane. $\square$

- $\mathrm{VC}\dim(\mathcal{H}) < d+1$: any $d+1$ points $\boldsymbol{x}_i, 0 \leq i \leq d$, cannot be shattered.
  - $(\boldsymbol{x}_i - \boldsymbol{x}_0), 1 \leq i \leq d$, are linearly dependent: assume $(\boldsymbol{x}_j - \boldsymbol{x}_0)$ is a linear combination of all other $(\boldsymbol{x}_i - \boldsymbol{x}_0)$, $1 \leq i \leq d$, $i \neq j$, the dichotomy $(\underbrace{+, \ldots, +}_{j}, -, +, \ldots, +)$ cannot be realized.
  - $(\boldsymbol{x}_i - \boldsymbol{x}_0), 1 \leq i \leq d$, are linearly independent: the all $+$ dichotomy $(+, \ldots, +)$ cannot be realized, since if there is a hyperplane containing all $d+1$ points, the dimension of this hyperplane is d, which is a contradiction.
- $\mathrm{VC}\dim(\mathcal{H}) = d$.

Example: Halfspaces in $\mathbb{R}^2$

- $\mathbb{R}^2$: the input space.
- $\mathscr{Y}' = \mathscr{Y} = \{-1, +1\}$: the output, label space.
- $\mathcal{H}$: the family of all halfspaces in $\mathbb{R}^2$.
- $\mathrm{VC}\dim(\mathcal{H}) \geq 3$: any noncollinear three points can be shattered.
- $\mathrm{VC}\dim(\mathcal{H}) < 4$: any four points cannot be shattered.
  - The four points lying on the convex hull defined by the four points: the dichotomy $(+, -, +, -)$ cannot be realized.
  - Three of the four points lie on the convex hull and the remaining point is internal: the dichotomy $(+, +, +, -)$, where the $-$ label is for the internal point, cannot be realized.
- $\mathrm{VC}\dim(\mathcal{H}) = 3$.



Unrealizable dichotomy for four points using halfspaces 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.

Example: Halfspaces in $\mathbb{R}^d$

- $\mathbb{R}^d$: the input space.
- $\mathscr{Y}' = \mathscr{Y} = \{-1, +1\}$: the output, label space.
- $\mathcal{H}$: the family of all halfspaces in $\mathbb{R}^d$.
- $\mathrm{VC\,dim}(\mathcal{H}) \geq d + 1$: the origin $\mathbf{0} = (0, \ldots, 0)$ and the standard unit vectors $\boldsymbol{e}_i = (0, \ldots, 0, 1, 0, \ldots, 0)$ can be shattered.

  Proof. Let $(w_0, w_1, w_2, \ldots, w_d)$, $w_i \in \{-1, +1\}$, be an arbitrary dichotomy of the $d + 1$ points $\mathbf{0}$ and $\boldsymbol{e}_i, 1 \leq i \leq d$. Define a hyperplane in $\mathbb{R}^d$:

$$w_1 x_1 + w_2 x_2 + \cdots + w_d x_d + w_0/2 = 0$$

  where $x_1, \ldots, x_d$ are variables in $\mathbb{R}$. It can be seen that the (positive or nonnegative) halfspace defined by this hyperplane, i.e., a hypothesis, labels $\mathbf{0}$ with w_0 and $\boldsymbol{e}_i$ with w_i, $1 \leq i \leq d$. $\square$

- $\mathrm{VC}\dim(\mathcal{H}) < d+2$: any $d+2$ points cannot be shattered.

  Proof. By Radon's theorem, a set X of $d+2$ points in $\mathbb{R}^d$ can be partitioned into two sets X_1 and X_2 such that their convex hulls intersect. Suppose that points in X_1 can be labeled with $+$ and points in X_2 with $-$ by a halfspace (a hypothesis in $\mathcal{H}$) defined by a hyperplane, i.e., X_1 and X_2 are separated by the hyperplane, their convex hulls are also separated by that hyperplane, which is a contradiction. Thus, no hypothesis in $\mathcal{H}$ can label points in X_1 and in X_2 in such a way. This conclude that any $d+2$ points cannot be shattered. $\square$

- $\mathrm{VC}\dim(\mathcal{H}) = d+1$.

Radon's Theorem

Theorem 3.4: Let

- $X = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_{d+2}\}$: a set of $d+2$ points in $\mathbb{R}^d$.

There is a partition $\{X_1, X_2\}$ of X, i.e., $X = X_1 \cup X_2$ and $X_1 \cap X_2 = \emptyset$, such that the convex hulls of X_1 and X_2 intersect.

Proof. Consider the system of $d+1$ homogeneous linear equations with $d+2$ unknowns $\alpha_i, 1 \leq i \leq d+2$,:

$$\begin{bmatrix} \boldsymbol{x}_1 & \boldsymbol{x}_2 & \cdots & \boldsymbol{x}_{d+2} \\ 1 & 1 & \cdots & 1 \end{bmatrix} \begin{bmatrix} \alpha_1 \\ \alpha_2 \\ \vdots \\ \alpha_{d+2} \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}.$$

Since the rank of the coefficient matrix is less than $d+2$, there

exists a non-trivial solution, says $\beta_1, \beta_2, \ldots, \beta_{d+2}$, not all zeros. Since $\sum_{i=1}^{d+2} \beta_i = 0$, both

$$I_1 = \{i \in [1, d+2] \mid \beta_i \geq 0\} \quad \text{and} \quad I_2 = \{i \in [1, d+2] \mid \beta_i < 0\}$$

are non-empty sets and $\sum_{i \in I_1} \beta_i = \sum_{i \in I_2} -\beta_i$. Now we have

$$\sum_{i \in I_1} \beta_i \boldsymbol{x}_i = \sum_{i \in I_2} -\beta_i \boldsymbol{x}_i$$

and then

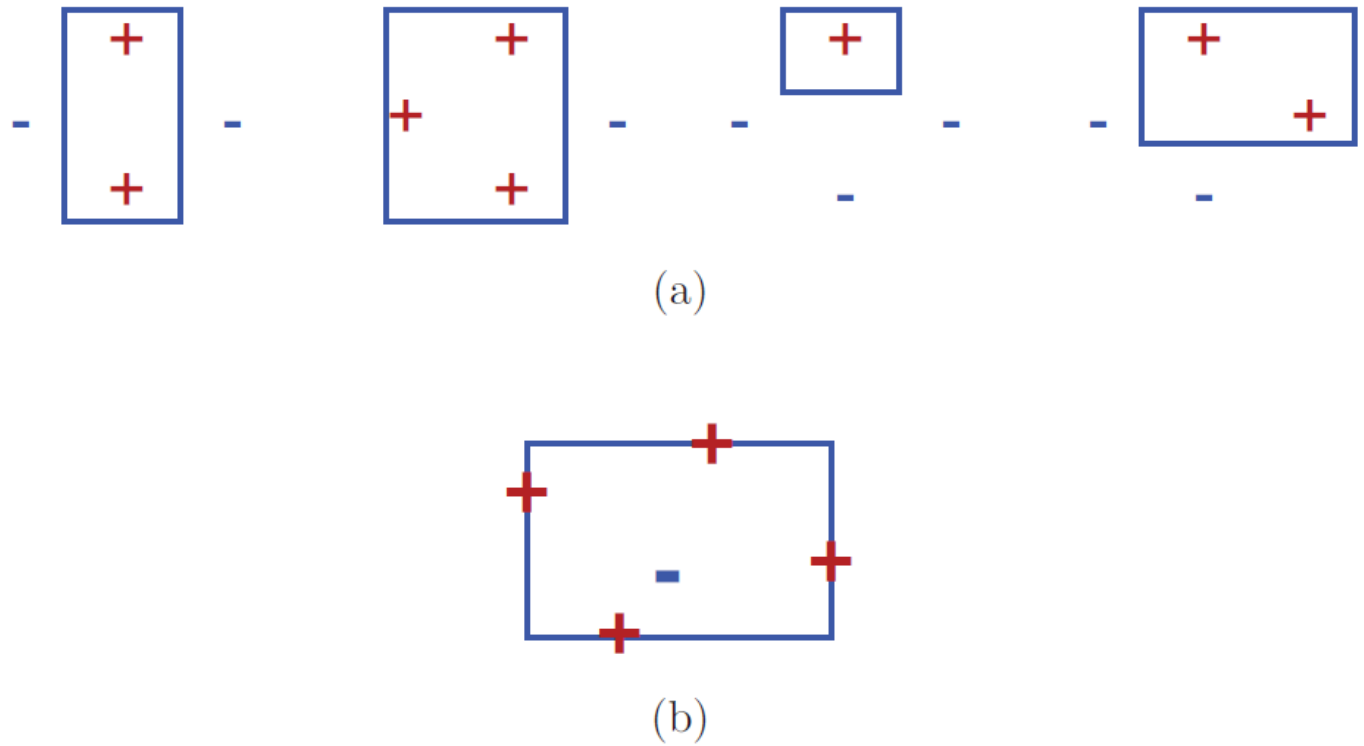
$$\sum_{i \in I_1} \frac{\beta_i}{\sum_{j \in I_1} \beta_j} \boldsymbol{x}_i = \sum_{i \in I_2} \frac{-\beta_i}{\sum_{j \in I_2} -\beta_j} \boldsymbol{x}_i.$$

The point in the left-hand side is in the convex hull of $X_1 = \{\boldsymbol{x}_i\}_{i \in I_1}$ and the point in the right-hand side is in the convex hull of $X_2 = \{\boldsymbol{x}_i\}_{i \in I_2}$. This completes the proof. □

Example: Axis-Aligned Rectangular Areas in $\mathbb{R}^2$

- $\mathbb{R}^2$: input space; $\mathscr{Y} = \{-1, +1\}$: label space.
- $\mathcal{H}$: the family of all axis-aligned rectangular areas in $\mathbb{R}^2$.
- $\mathrm{VC}\dim(\mathcal{H}) \geq 4$: any four points in a diamond pattern can be shattered.
- $\mathrm{VC}\dim(\mathcal{H}) < 5$: any five points cannot be shattered.

  Proof. For any five distinct points, we construct the minimal axis-aligned rectangular area containing these points. At most four of the five points determine this rectangular area. Then the dichotomy giving positive labels to the determining points and negative labels to the remaining points cannot be realized by any axis-aligned rectangular area. □
- $\mathrm{VC}\dim(\mathcal{H}) = 4$.



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.

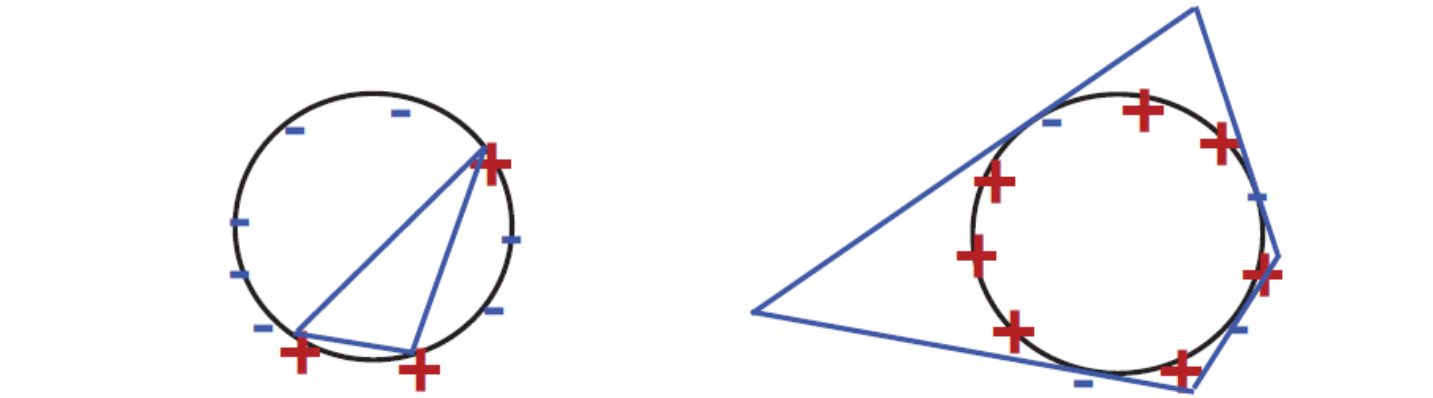
Example: Convex d-gons in $\mathbb{R}^2$, $d \geq 3$

- $\mathbb{R}^2$: input space; $\mathscr{Y} = \{-1, +1\}$: label space.
- $\mathcal{H}$: the family of all convex d-gons (including all interior points) in $\mathbb{R}^2$.
- Fact: choosing points on a circle maximizes the number of dichotomies realized by $\mathcal{H}$.
- $\text{VC}\dim(\mathcal{H}) \geq 2d + 1$: any $2d + 1$ points on a circle can be shattered.
  - If there are more negative than positive labels, then the points with the positive labels are used as the d-gon's vertices.

- If there are more positive than negative labels, then secant lines of the circle very close to and parallel to the tangent lines at the negatively labeled points such that only one negatively labeled point is in one side serve as the edges of the d-gon.

- $\mathrm{VC}\dim(\mathcal{H}) < 2d+2$: any $2d+2$ points on a circle cannot be shattered since the alternating pattern $(+1,-1,\ldots,+1,-1)$ cannot be realized by any d-gon.

- $\mathrm{VC}\dim(\mathcal{H}) = 2d+1$.



(a) (b)

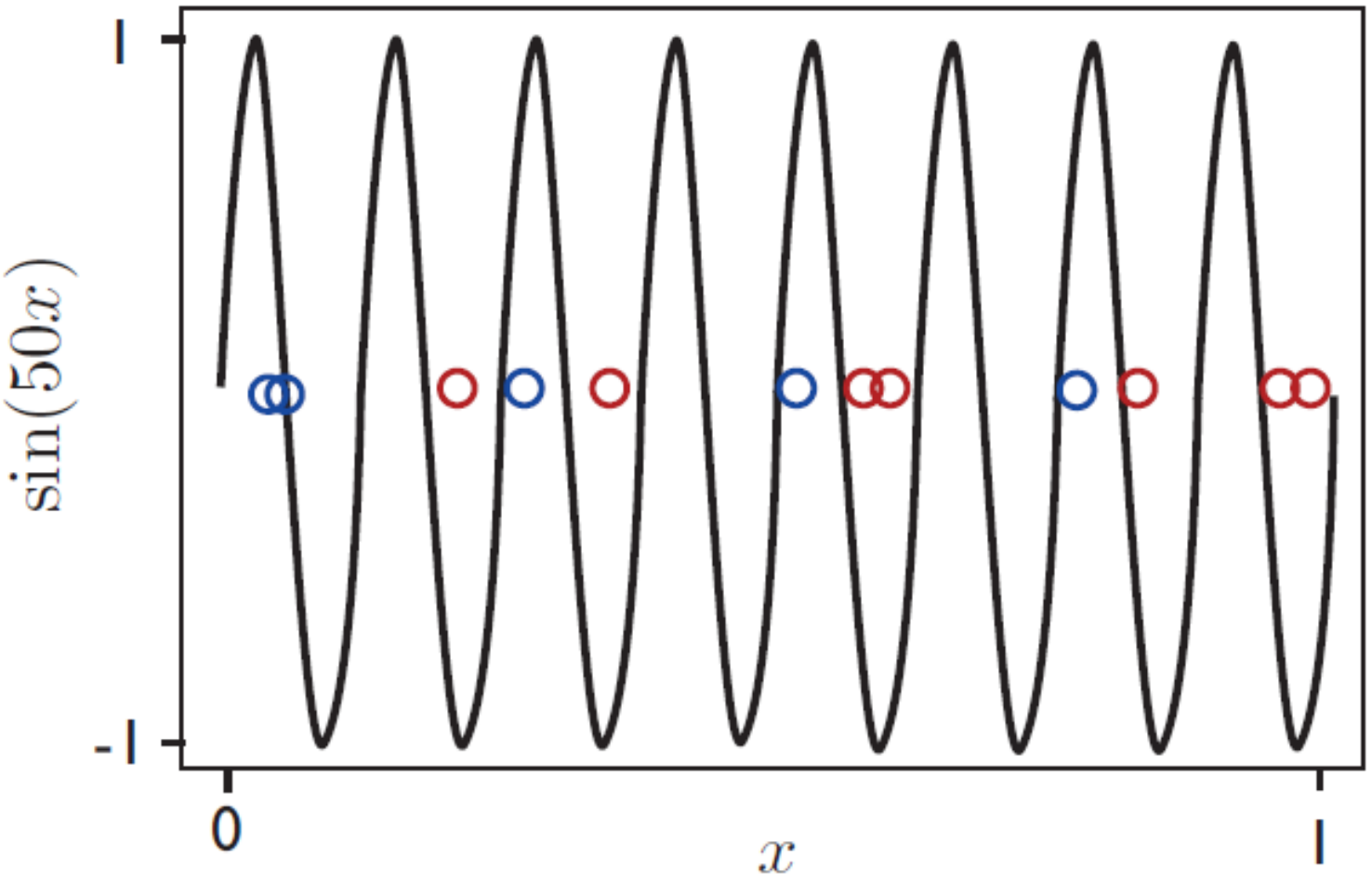
Convex d-gons in the plane can shatter $2d + 1$ points. (a) d-gon construction when there are more negative labels. (b) d-gon construction when there are more positive labels.

Example: Convex polygons in $\mathbb{R}^2$

- $\mathbb{R}^2$: input space; $\mathscr{Y} = \{-1, +1\}$: label space.
- $\mathcal{H}$: the family of all convex polygons (including all interior points) in $\mathbb{R}^2$.
- $\mathrm{VC}\dim(\mathcal{H}) = +\infty$.

Example: Sine Functions on $\mathbb{R}$

- $\mathbb{R}$: input space; $\mathscr{Y} = \{-1, +1\}$: label space.
- $\mathcal{H}$: the family of all sine functions: $t \mapsto \sin(\omega t)$, $\omega \in \mathbb{R}$, where a point t on the real line is labeled with $+1$ if $\sin(\omega t) \leq 0$ and with -1 if $\sin(\omega t) > 0$.
- $\mathrm{VC}\dim(\mathcal{H}) = +\infty$ (Exercise 3.12).



An example of sine function (with $\omega = 50$) used for classification.

Sauer's Lemma - VC-Dimension and Growth Function

Theorem 3.5: Let

- $\mathscr{I}$: an arbitrary input space;
- $\mathcal{H}$: a set of hypotheses from $\mathscr{I}$ to $\{-1, +1\}$ with

$$\mathrm{VC\,dim}(\mathcal{H}) = d < +\infty.$$

Then for all $m \geq 1$,

$$\Pi_{\mathcal{H}}(m) \leq \sum_{i=0}^{d} C_i^m,$$

where $C_i^j \triangleq 0$ if $i > j$.

Proof.

- If $\mathrm{VC}\dim(\mathcal{H}) = d = 0$, then $\mathcal{H}$ must be a singleton and then $1 = \Pi_{\mathcal{H}}(m) \leq \sum_0^d C_i^m = \sum_0^0 C_i^m = 1$ for all $m \geq 1$.
- If $\mathrm{VC}\dim(\mathcal{H}) = d > 0$, there is a set of d distinct items in $\mathscr{I}$ fully shattered by $\mathcal{H}$. It is clear that any of its subset can be shattered by $\mathcal{H}$ so that

$$\Pi_{\mathcal{H}}(m) = 2^m = \sum_{i=0}^{m} C_i^m = \sum_{i=0}^{d} C_i^m \ \forall\ 1 \leq m \leq d.$$

- The proof will be by induction on $m + d$.
- For $m + d = 1$, we must have $m = 1, d = 0$ and the theorem is true.
- Assume that $\Pi_{\mathcal{H}}(m) \leq \sum_{i=0}^{d} C_i^m$ for all $1 \leq m + d \leq k$ with $k \geq 1$ for any input space $\mathscr{I}$ and any hypothesis set $\mathcal{H}$ with $\mathrm{VC}\dim(\mathcal{H}) = d$.

- Consider $m + d = k + 1$. We may assume $1 \leq d < m$ since other cases are done.
- $S = \{\omega_1, \ldots, \omega_m\}$: a set of m distinct items in $\mathscr{I}$ such that $\Pi_{\mathcal{H}}(m) = |\{(h(\omega_1), \ldots, h(\omega_m)) \mid h \in \mathcal{H}\}| = |\{h_{|S} \mid h \in \mathcal{H}\}|$.
- $S' \triangleq \{\omega_1, \ldots, \omega_{m-1}\}$.
- $\mathcal{H}'$: a subset of $\mathcal{H}$ and $h' \in \mathcal{H}'$ if and only if there is an $h \in \mathcal{H}$ such that $h_{|S'} = h'_{|S'}$ but $h(\omega_m) \neq h'(\omega_m)$.
- $\mathcal{H}_{|S}, \mathcal{H}_{|S'}, \mathcal{H}'_{|S'}$: the set of all hypotheses in $\mathcal{H}, \mathcal{H}, \mathcal{H}'$ restricted to S, S', S' respectively.

- $\Pi_{\mathcal{H}}(m) = |\mathcal{H}_{|S}| = |\mathcal{H}_{|S'}| + |\mathcal{H}'_{|S'}|$.

	ω_1	ω_2	$\cdots$	ω_{m-1}	ω_m
h'	-1	$+1$	$\cdots$	-1	-1
h	-1	$+1$	$\cdots$	-1	$+1$
h_1	$+1$	$+1$	$\cdots$	-1	$+1$
h_2	-1	-1	$\cdots$	$+1$	-1
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$

- $\mathscr{I}' \triangleq S' = \{\omega_1, \ldots, \omega_{m-1}\}$: a new input space.
- $\mathcal{H}_{|S'}, \mathcal{H}'_{|S'}$: sets of hypotheses from input space $\mathscr{I}'$ to $\{-1, +1\}$.
- $d_1 \triangleq \mathrm{VC\,dim}(\mathcal{H}_{|S'}) \leq \mathrm{VC\,dim}(\mathcal{H}) = d$: any subset of S' which is shattered by $\mathcal{H}_{|S'}$ is also shattered by $\mathcal{H}$.
  - Since $d_1 + m - 1 \leq d + m - 1 = k$,

$$|\mathcal{H}_{|S'}| = \Pi_{\mathcal{H}_{|S'}}(m-1) \leq \sum_{i=0}^{d_1} C_i^{m-1} \leq \sum_{i=0}^{d} C_i^{m-1}.$$

- $d_2 \triangleq \mathrm{VC\,dim}(\mathcal{H}'_{|S'}) \leq \mathrm{VC\,dim}(\mathcal{H}) - 1 = d - 1$: if a subset Z of S' is shattered by $\mathcal{H}'_{|S'}$, then $Z \cup \{\omega_m\}$ is shattered by $\mathcal{H}'$ and then by $\mathcal{H}$.
  - Since $d_2 + m - 1 \leq d - 1 + m - 1 \leq k$,

$$|\mathcal{H}'_{|S'}| = \Pi_{\mathcal{H}'_{|S'}}(m-1) \leq \sum_{i=1}^{d_2} C_i^{m-1} \leq \sum_{i=0}^{d-1} C_i^{m-1}.$$

- $$\begin{aligned}
\Pi_{\mathcal{H}}(m) &= |\mathcal{H}_{|S'}| + |\mathcal{H}'_{|S'}| \\
&\leq \sum_{i=0}^{d} C_i^{m-1} + \sum_{i=0}^{d-1} C_i^{m-1} \\
&= \sum_{i=0}^{d} C_i^{m-1} + \sum_{i=1}^{d} C_{i-1}^{m-1} \\
&= C_0^{m-1} + \sum_{i=1}^{d} (C_i^{m-1} + C_{i-1}^{m-1}) \\
&= C_0^m + \sum_{i=1}^{d} C_i^m \\
&= \sum_{i=0}^{d} C_i^m .
\end{aligned}$$

- The proof is completed by induction on $d + m$. □

The Order of Growth Function

Corollary 3.3: Let

- $\mathscr{I}$: an arbitrary input space;
- $\mathcal{H}$: a set of hypotheses from $\mathscr{I}$ to $\{-1, +1\}$ with

$$\mathrm{VC}\dim(\mathcal{H}) = d < +\infty.$$

Then for all $m \geq d$,

$$\Pi_{\mathcal{H}}(m) \leq \left(\frac{em}{d}\right)^d = O(m^d).$$

Proof. By Sauer's lemma, we have

$$\begin{aligned}
\Pi_{\mathcal{H}}(m) &\leq \sum_{i=0}^{d} C_i^m \\
&\leq \sum_{i=0}^{d} C_i^m \left(\frac{m}{d}\right)^{d-i} \text{ since } m \geq d \\
&\leq \left(\frac{m}{d}\right)^{d} \sum_{i=0}^{m} C_i^m \left(\frac{d}{m}\right)^{i} \\
&= \left(\frac{m}{d}\right)^{d} \left(1+\frac{d}{m}\right)^{m} \\
&\leq \left(\frac{m}{d}\right)^{d} e^{\frac{d}{m}m} \text{ since } 1+x \leq e^{x} \ \forall \ x \in \mathbb{R} \\
&= \left(\frac{em}{d}\right)^{d} = O(m^d).
\end{aligned}$$

□

Remarks

- There are only two types of growth functions.
  - $\mathrm{VC}\dim(\mathcal{H}) = d < +\infty$: $\Pi_{\mathcal{H}}(m) = O(m^d)$.
  - $\mathrm{VC}\dim(\mathcal{H}) = +\infty$: $\Pi_{\mathcal{H}}(m) = 2^m$.

VC-Dimension Generalization Bound - Binary Classification

Corollary 3.4: Let

- $c : \mathscr{I} \to \{-1, +1\}$: a fixed but unknown target concept in the concept class $\mathcal{C}$.
- $\mathcal{H}$: a set of possibly infinitely many hypotheses $h : \mathscr{I} \to \{-1, +1\}$ with 0-1 loss function $L(y', y) = 1_{y' \neq y}$.
- $\text{VCdim}(\mathcal{H}) = d < +\infty$.
- $S = (\omega_1, \ldots, \omega_m)$: a sample of size m drawn i.i.d. from the input space $\mathscr{I}$ according to an unknown distribution P.

For any $\delta > 0$, with probability at least $1 - \delta$, we have

$$\forall\ h \in \mathcal{H}, \quad R(h) \leq \hat{R}_S(h) + \sqrt{\frac{2d \ln \frac{em}{d}}{m}} + \sqrt{\frac{\ln \frac{1}{\delta}}{2m}}.$$

Remarks

- The VC-dimension generalization bound can be written as

$$R(h) \leq \hat{R}_S(h) + O\left(\sqrt{\frac{\ln(m/d)}{(m/d)}}\right),$$

  which emphasizes the importance of the ratio m/d for generalization.

- The corollary provides another instance of Occam's razor principle where simplicity is measured in terms of smaller VC-dimension.

The Contents of This Lecture

- Rademacher complexity
- Growth function
- VC-dimension
- Lower bounds

Lower Bound - Realizable Case

Theorem 3.6: Let

- $\mathcal{H}$: a set of possibly infinitely many hypotheses $h : \mathscr{I} \to \{-1, +1\}$ with 0-1 loss function $L(y', y) = 1_{y' \neq y}$.
- $\text{VCdim}(\mathcal{H}) = d \geq 1$.
- $S = (\omega_1, \ldots, \omega_m)$: a labeled sample of size m drawn i.i.d. from the input space $\mathscr{I}$ according to an unknown distribution P.
- $\mathbb{A}$: a learning algorithm which returns $h_S \in \mathcal{H}$ to approximate an unknown target concept $c \in C$ on the labeled sample S, $h_S = \mathbb{A}(S; c, \mathcal{H})$.

Then there exist a distribution P over $\mathscr{I}$ and a target function c in $\mathcal{H}$ such that

$$P_m\left(R(h_S; c) > \frac{d-1}{32m}\right) \geq \frac{1}{100}.$$