SUMMARY of math framework of ML (last class)

- \( P \): an unknown distribution on \( X \times Y \);
- We see training data: \((x_1, y_1), \ldots, (x_n, y_n) \sim P \) iid.
  \[ \text{Goal: oracle } h: X \to Y: \quad h(x) = y \]
- Choose a hypothesis class \( \mathcal{H} \) (functions \( X \to Y \))
- \( \text{Vapnik, Risk, a.k.a. "test error"} \)
  \[ R(h) := \mathbb{E}[L(h(x), y)] \]
  \[ h^* := \text{argmin}_{h \in \mathcal{H}} R(h). \quad \text{Not computable} \]
- Empirical risk a.k.a. training error
  \[ R_n(h) := \frac{1}{n} \sum_{i=1}^{n} L(h(x_i), y_i). \quad h_n^* := \text{argmin}_{h \in \mathcal{H}} R_n(h). \quad \text{Computable} \]

**ERM algorithm:**

1. Training: for input data \((x_1, y_1), \ldots, (x_n, y_n)\); compute \( h_n^* \).
2. Prediction: on query \( X \), output \( h_n^*(x) \) “oracle”

\[
\text{Generalization error} \\
R(h^*) \leq R_n(h^*) + 2 \sup_{h \in \mathcal{H}} \left| R_n(h) - R(h) \right|
\]
- test error of ERM
- best possible error (with no data)
  for a given class \( \mathcal{H} \)

**Proof**

\[
R(h_n^*) \leq R_n(h_n^*) + \varepsilon \quad (h_n^* \in \mathcal{H}) \\
\leq R_n(h^*) + \varepsilon \quad (h^* = \text{minimizer of } R_n) \\
\leq R(h^*) + 2\varepsilon \quad (h^* \in \mathcal{H}) \quad \square
\]

Next time, add a lemma:

if \( \|f-g\|_{\infty} < \varepsilon \) and \( x^*, y^* \) are minimizers of \( f, g \),
then
\[ |f(x^*) - g(y^*)| < 2\varepsilon \]
Then apply this lemma
for \( f = R_n, g = R, h_n^*, h^* \).
\[ E = \sup_{h \in H} \left| \frac{1}{n} \sum_{i=1}^{n} Z_i(h) - E Z(h) \right| \]
where \( Z_i(h) = \ell(h(x_i), y_i) \) are iid r.v.s.

"empirical process" \[ \overline{Z}_i(h) \]

- For binary classification, \( \ell(\cdot, \cdot) \in \{0,1\} \Rightarrow |Z_i(h)| \leq 1 

\[ P \left\{ \left| \frac{1}{n} \sum_{i=1}^{n} Z_i(h) \right| > t \right\} = P \left\{ \left| \frac{1}{n} \sum_{i=1}^{n} Z_i(h) \right| > tn \right\} \leq 2 \exp \left( -\frac{t^2 n}{2} \right) \]

multiply both sides by \( \frac{1}{n} \) to scale like in CLT

General Hoeffding inequality (Lec.5)

- Union Bound:

\[ P \left\{ \sup_{h \in H} \left| \frac{1}{n} \sum_{i=1}^{n} Z_i(h) \right| > t \right\} \leq \sum_{h \in H} P \left\{ \left| \frac{1}{n} \sum_{i=1}^{n} Z_i(h) \right| > t \right\} \]

\[ \leq |H| \exp \left( -\frac{t^2 n}{2} \right) = 2 \exp \left( \log |H| - \frac{t^2 n}{2} \right) \]

\[ \Rightarrow \text{We proved:} \]

\[ t = C \sqrt{\frac{\log |H|}{n}} \]

**THM** (Generalization bound) If the hypothesis class \( H \) is finite, \( R(h_n^*) \leq R(h^*) + C \sqrt{\frac{\log |H|}{n}} \) with prob. \( \geq 0.99 \).

- Good: logarithmic in \( |H| \)
- Bad: most hypothesis classes are infinite.

Can \( \log |H| \) be replaced by some "complexity" of \( H \)?

Yes: VC dimension.

Hence the ERM algorithm generalizes well from \( n \sim \log |H| \) training data points.
**VC DIMENSION**

- Heuristically: \( \text{vc}(\mathcal{H}) = \text{largest } \#(\text{data } \mathcal{H} \text{ overfits}) \)

\[ \text{i.e functions } h : X \rightarrow \{0, 1\} \]

**Def** Let \( \mathcal{H} \) be any collection of Boolean functions on a set \( X \). We say that \( \mathcal{H} \) overfits, or "shatters" a subset \( \{x_1, \ldots, x_d\} \subset X \) if \( \forall \) labels \( y_1, \ldots, y_d \in \{0, 1\} \) \( \exists h \in \mathcal{H} \) such that

\[ h(x_i) = y_i \quad \forall i = 1, \ldots, d. \]

The \textit{VC dimension} of \( \mathcal{H} \), denoted \( \text{vc}(\mathcal{H}) \), is the maximal size of a subset \( \mathcal{H} \) shatters.

**Examples**

1. \( \mathcal{H} = \{1\} \) has \( \text{vc}(\mathcal{H}) = 0 \): it can't shatter even one point \( x_i \) since \( h(x_i) = 1 \).

2. Half-lines \( \mathcal{H} = \{1_{(-\infty, a]} : a \in \mathbb{R}\} \)

   \[ \text{vc}(\mathcal{H}) = 1 \]

**Prof.**

\( \mathcal{H} \) can shatter some 1-point set \( \{x_1\} \), but can't shatter any 2-point set \( \{x_1, x_2\} \)

\[ x_1 \quad \{0, 1\} \quad x_2 \]

**HW:** \( \{1_{(-\infty, a]} : a \in (b, +\infty)\} \)

3. Intervals: \( \mathcal{H} = \{1_{[a, b]} : a \leq b\} \)

   \[ \text{vc}(\mathcal{H}) = 2 \]

**Prof.**

\( \mathcal{H} \) can shatter some 2-pt set \( \{x_1, x_2\} \), but can't shatter any 3-point set \( \{x_1, x_2, x_3\} \)

\[ x_1 \quad x_2 \quad x_3 \]
4. Half-planes in \( \mathbb{R}^2 \):

\[ \mathcal{H} = \left\{ \mathbf{1} a_1 x_1 + a_2 x_2 + b \right\} : \ a_1, a_2, b \in \mathbb{R} \right\} \]

\[ \text{VC} (\mathcal{H}) = 3 \]

If can shatter some 3-pt set \( \{x_1, x_2, x_3\} \),

but can't shatter any 4-point set \( \{x_1, x_2, x_3, x_4\} \):

If 4-point set is like this, or like this:

"Convex position"  "non-convex position"

In either case, \( \exists \) label assignment that is impossible to realize.