Recap: Regression

- \( Z_i = (X_i, Y_i) \in \mathbb{R}^d \times \mathbb{R} \). \( \mathcal{A} \subseteq \mathcal{B} := \{ a : \mathbb{R}^d \rightarrow \mathbb{R} \} \). \( \ell(a, (x, y)) = \phi(a(x), y) \)

- Goal: \( \text{Rad}(\mathcal{A} \circ \{x_1, \ldots, x_n\}) \leq \frac{f(\text{dimension, complexity of } \mathcal{A})}{n^\alpha} \)

### SVM (Proposition 3.2)

Let \( \mathcal{A}_2 := \{ x \in \mathbb{R}^d \rightarrow w^\top x : \|w\|_2 \leq c \} \). Then

\[
\text{Rad}(\mathcal{A}_2 \circ \{x_1, \ldots, x_n\}) \leq \max_i \|x_i\|_\infty c \frac{\sqrt{d}}{\sqrt{n}}
\]

### Boosting (Proposition 3.6)

Let \( \mathcal{A}_\Delta := \{ x \in \mathbb{R}^d \rightarrow w^\top x : \|w\|_1 = c, w_1, \ldots, w_d \geq 0 \} \). Then

\[
\text{Rad}(\mathcal{A}_\Delta \circ \{x_1, \ldots, x_n\}) \leq \max_i \|x_i\|_\infty c \frac{\sqrt{2 \log d}}{\sqrt{n}}
\]

Difference between \( d \) and \( \log d \) related to difference between \( \ell_2 \) and \( \ell_1 \) ball, resp.
Today: Classification (binary)

- \( Z_i = (X_i, Y_i) \in \mathbb{R}^d \times \{-1, 1\} \)
- Admissible action set \( \mathcal{A} \subseteq \mathcal{B} := \{a : \mathbb{R}^d \rightarrow \{-1, 1\}\} \)
- Loss function \( \ell(a, (x, y)) = \phi(a(x), y) \), for \( \phi : \{-1, 1\}^2 \rightarrow \mathbb{R}_+ \)
- Today we consider \( \phi(\hat{y}, y) = 1_{\hat{y} \neq y} = (1 - y\hat{y})/2 \), a.k.a. the true loss

Recall. For regression we used:

\[
\text{(Proposition 3.1)}
\]

If the function \( \hat{y} \rightarrow \phi(\hat{y}, y) \) is \( \gamma \)-Lipschitz for any \( y \in \mathcal{Y} \), then

\[
\text{Rad}(\mathcal{L} \circ \{z_1, \ldots, z_n\}) \leq \gamma \text{Rad}(\mathcal{A} \circ \{x_1, \ldots, x_n\})
\]

For classification with the true loss we can use:

\[
\text{(Proposition 4.1)}
\]

If \( \phi \) is the true loss, then

\[
\text{Rad}(\mathcal{L} \circ \{z_1, \ldots, z_n\}) = \frac{1}{2} \text{Rad}(\mathcal{A} \circ \{x_1, \ldots, x_n\})
\]
Growth Function

- $\mathcal{A} \circ \{x_1, \ldots, x_n\} = \{(a(x_1), \ldots, a(x_n)) \in \{-1, 1\}^n : a \in \mathcal{A}\}$
- $|\mathcal{A} \circ \{x_1, \ldots, x_n\}| \leq 2^n$ even if the class $\mathcal{A}$ is infinite
- **Important:** It can grow polynomialsly with $n$

---

**Growth function (Definition 4.2)**

The *growth function* of $\mathcal{A}$ is defined as

$$n \in \mathbb{N} \rightarrow \tau_{\mathcal{A}}(n) := \sup_{x_1, \ldots, x_n \in \mathbb{R}^d} |\mathcal{A} \circ \{x_1, \ldots, x_n\}|$$

Max number of labelings of $n$ vectors that we can obtain using classifiers in $\mathcal{A}$

Yields “data-independent” bound on Rademacher complexity (Massart’s lemma)

---

**Proposition 4.3**

$$\text{Rad}(\mathcal{A} \circ \{x_1, \ldots, x_n\}) \leq \sqrt{\frac{2 \log \tau_{\mathcal{A}}(n)}{n}}$$

**Note:** To drive convergence to 0 as $n$ grows, we need $\tau_{\mathcal{A}}$ to grow **polynomially**.
Growth Function: Examples

- **Half spaces over the real line** $\mathcal{A} = \{a(x) = 21_{x \leq w - 1} : w \in \mathbb{R}\}$
  
  $\varepsilon$
  
  0000 ⋯ 0
  1000 ⋯ 0
  1100 ⋯ 0
  ⋮
  1111 ⋯ 1

  $\tau_{\mathcal{A}}(n) = n + 1$

- **Intervals over the real line** $\mathcal{A} = \{a(x) = 21_{w^- \leq x \leq w^+ - 1} : w^- \leq w^+\}$

  $\varepsilon$
  
  0000 ⋯ 00
  1000 ⋯ 00 0100 ⋯ 00 0010 ⋯ 00 ⋮ 0000 ⋯ 10 0000 ⋯ 01
  1100 ⋯ 00 0110 ⋯ 00 0011 ⋯ 00 ⋮ 0000 ⋯ 11

  $\tau_{\mathcal{A}}(n) = 1 + n(n + 1)/2$

**Problem:** not always easy to compute! **Solution:** VC dimension
VC Dimension

VC dimension (Definition 4.6)

\[ VC(A) := \max\{n \in \mathbb{N} : \tau_A(n) = 2^n\} \]

If \( \tau_A(n) = 2^n \) for all integer \( n \), then \( VC(A) = \infty \)

- **Half spaces over the real line** \( A = \{a(x) = 21_{x \leq w - 1} : w \in \mathbb{R}\} \)
  \[ \tau_A(n) = n + 1 \]
  \( \tau_A(1) = 2^1 \) and \( \tau_A(2) = 3 < 2^2 \implies VC(A) = 1 \)

- **Intervals over the real line** \( A = \{a(x) = 21_{w^- \leq x \leq w^+ - 1} : w^- \leq w^+\} \)
  \[ \tau_A(n) = 1 + n(n + 1)/2 \]
  \( \tau_A(2) = 2^2 \) and \( \tau_A(3) = 7 < 2^3 \implies VC(A) = 2 \)

**Key point:** We can compute the VC dimension without computing \( \tau_A \)

- **Sufficient** to find \( k \) such that \( \tau_A(k) = 2^k \) and \( \tau_A(k + 1) < 2^{k+1} \)
- **This can be done without computing \( \tau_A \). Sufficient** to:
  - Find distinct \( x_1, \ldots, x_k \) that are “shattered” by \( A \) \( \Rightarrow VC(A) \geq k \)
    (i.e., classifiers in \( A \) can assign all possible \( 2^k \) labelings to these points)
  - Show that no set of \( k + 1 \) points can be “shattered” by \( A \) \( \Rightarrow VC(A) < k + 1 \)
    (i.e., for any set of \( k + 1 \) points there is a label that can not be assigned)
Bounds using VC Dimension

If $\text{VC}(A)$ is finite, then $\tau_A$ eventually grows polynomially

**Sauer-Shelah's Lemma (Lemma 4.11)**

\[
\tau_A(n) \begin{cases} 
2^n & \text{if } n \leq \text{VC}(A) \\
\left(\frac{en}{\text{VC}(A)}\right)^{\text{VC}(A)} & \text{if } n > \text{VC}(A)
\end{cases}
\]

**Proposition 4.12**

For any $x_1, \ldots, x_n \in \mathbb{R}^d$ we have

\[
\text{Rad}(A \circ \{x_1, \ldots, x_n\}) \leq \sqrt{\frac{2 \text{VC}(A) \log(en/\text{VC}(A))}{n}}
\]

- This bound is “data-independent” as it holds for any $x_1, \ldots, x_n$ (as such, it does not allow to exploit the statistical nature of the data)
- We will remove the log-term using covering numbers and chaining
Covering and Packing Numbers

A pseudometric space \((S, \rho)\) is a set \(S\) and a function \(\rho : S \times S \rightarrow \mathbb{R}_+\) (called a pseudometric) such that, for any \(x, y, z \in S\) we have:

- \(\rho(x, y) = \rho(y, x)\) (symmetry)
- \(\rho(x, z) \leq \rho(x, y) + \rho(y, z)\) (triangle inequality)
- \(\rho(x, x) = 0\)

A metric space is obtained if one further assumes that \(\rho(x, y) = 0\) implies \(x = y\).

Covering and Packing Numbers (Definition 4.13)

Let \((S, \rho)\) be a pseudometric space, \(\varepsilon > 0\)

- The set \(C \subseteq S\) is a \(\varepsilon\)-cover of \((S, \rho)\) if for every \(x \in S\) there exists \(y \in C\) such that \(\rho(x, y) \leq \varepsilon\). The set \(C \subseteq S\) is a minimal \(\varepsilon\)-cover if there is no other \(\varepsilon\)-cover with lower cardinality. The cardinality of any minimal \(\varepsilon\)-cover is the \(\varepsilon\)-covering number, denoted by \(\text{Cov}(S, \rho, \varepsilon)\)

- The set \(P \subseteq S\) is a \(\varepsilon\)-packing of \((S, \rho)\) if for every \(x, x' \in P\) we have \(\rho(x, x') > \varepsilon\). The set \(P \subseteq S\) is a maximal \(\varepsilon\)-packing if there is no other \(\varepsilon\)-packing with greater cardinality. The cardinality of any maximal \(\varepsilon\)-packing is the \(\varepsilon\)-packing number, denoted by \(\text{Pack}(S, \rho, \varepsilon)\).
Covering and Packing Numbers. Properties

Duality (Proposition 4.14)

\[ \text{Cov}(S, \rho, \varepsilon) \leq \text{Pack}(S, \rho, \varepsilon) \leq \text{Cov}(S, \rho, \varepsilon/2) \]

Covering and packing numbers typically grow exponentially with the dimension (in so-called “Logarithmic metric entropy” spaces)

Bounded Balls (Proposition 4.15)

\[ B^d_r := \{ y \in \mathbb{R}^d : \| y \| \leq r \} \] be the \( d \)-dim. ball with radius \( r \geq 0 \). If \( \varepsilon \leq r \), then

\[ \left( \frac{r}{\varepsilon} \right)^d \leq \text{Cov}(B^d_r, \| \cdot \|, \varepsilon) \leq \text{Pack}(B^d_r, \| \cdot \|, \varepsilon) \leq \left( \frac{3r}{\varepsilon} \right)^d \]

Proof: Volume argument

Covering and packing numbers grow exponentially also w.r.t. the VC dimension. This, along with chaining, will allow us to remove the log-term in Prop. 4.12