Offline statistical learning: prediction

Given a batch of observations (images & labels) interested in predicting the label of a new image
Offline statistical learning: prediction

1. Observe training data $Z_1, \ldots, Z_n$ i.i.d. from unknown distribution
2. Choose action $A \in A \subseteq B$
3. Suffer an expected/population loss/risk $r(A)$, where
   \[ a \in B \rightarrow r(a) := \mathbb{E} \ell(a, Z) \]
   with $\ell$ is a prediction loss function and $Z$ is a new test data point

**Goal:** Minimize the estimation error defined by the following decomposition
\[
r(A) - \inf_{a \in B} r(a) = r(A) - \inf_{a \in A} r(a) + \inf_{a \in A} r(a) - \inf_{a \in B} r(a)
\]
- excess risk
- estimation error
- approximation error

as a function of $n$ and notions of “complexity” of the set $A$ of the function $\ell$

**Note:** Estimation/Approximation trade-off, a.k.a. complexity/bias
ERM and Uniform Learning

- A natural framework is given by the empirical risk minimization (ERM)

\[
a \in \mathcal{B} \longrightarrow R(a) := \frac{1}{n} \sum_{i=1}^{n} \ell(a, Z_i)
\]

- A natural algorithm is given by the minimizer of the ERM

\[A^* \in \arg\min_{a \in \mathcal{A}} R(a)\]

- **Uniform Learning:** The estimation error is bounded by

\[
\underbrace{r(A^*) - r(a^*)}_{\text{estimation error for ERM}} \leq \sup_{a \in \mathcal{A}} \{r(a) - R(a)\} + \sup_{a \in \mathcal{A}} \{R(a) - r(a)\}
\]

- **Statistics**

- Statistical Learning deals with bounding the Statistics term (Vapnik 1995)

- **Generalization Error:** \[r(a) - R(a) \approx \frac{1}{n^{(\text{test})}} \sum_{i=1}^{n^{(\text{test})}} \ell(a, Z_i^{(\text{test})}) - \frac{1}{n} \sum_{i=1}^{n} \ell(a, Z_i)\]
Goal: derive bounds in expectation

- Goal:
  \[ \mathbb{E} r(A^*) - r(a^*) \lesssim \frac{f(\text{dimension})}{n^{\alpha}} \]
  estimation error for ERM

- By uniform learning, it suffices to bound the suprema of random processes:
  \[ \mathbb{E} g(Z_1, \ldots, Z_n) \leq \frac{f(\text{dimension, complexity of } A)}{n^{\alpha}} \]

  with \( g(Z_1, \ldots, Z_n) = \sup_{a \in A} \{r(a) - R(a)\} = \sup_{a \in A} \left\{ \mathbb{E} \ell(a, Z) - \frac{1}{n} \sum_{i=1}^{n} \ell(a, Z_i) \right\} \)

- We aim to derive a uniform, non-asymptotic Law of Large Numbers

- In machine learning, dimension can be \( \gg 10^6 \), e.g., number of pixels

- Ideally, \( f(\text{dimension}) \ll \text{dimension} \), e.g., \( f(\text{dimension}) \sim \log(\text{dimension}) \)

- Ideally, \( \alpha = 1 \) (fast rate)
Hoeffding’s Lemma (Lemma 2.1)

Let \( X \) be a bounded random variable \( a \leq X - \mathbb{E}X \leq b \). Then, for any \( \lambda \in \mathbb{R} \),

\[
\mathbb{E} e^{\lambda(X-\mathbb{E}X)} \leq e^{\lambda^2(b-a)^2/8}
\]

Proof

- W.l.o.g., take \( \mathbb{E}X = 0 \). Let \( \psi(\lambda) = \log \mathbb{E} e^{\lambda X} \)

\[
\psi'(\lambda) = \frac{\mathbb{E}[X e^{\lambda X}]}{\mathbb{E} e^{\lambda X}} \quad \psi''(\lambda) = \frac{\mathbb{E}[X^2 e^{\lambda X}]}{\mathbb{E} e^{\lambda X}} - \left( \frac{\mathbb{E}[X e^{\lambda X}]}{\mathbb{E} e^{\lambda X}} \right)^2
\]

- \( \psi''(\lambda) \) is the variance of \( X \) under the distribution \( Q(dx) = \frac{e^{\lambda x}}{\mathbb{E} e^{\lambda X}} P(dx) \)

- \( \psi''(\lambda) = \text{Var}_Q \left( X - \frac{a + b}{2} \right) \leq \mathbb{E}_Q \left[ \left( X - \frac{a + b}{2} \right)^2 \right] \leq \frac{(b-a)^2}{4} \)

- Fundamental Thm of Calculus: \( \psi(\lambda) = \int_0^\lambda \int_0^\mu \psi''(\rho) \, d\rho \, d\mu \leq \frac{\lambda^2 (b-a)^2}{8} \)
Let $X_1, \ldots, X_n$ be $n$ centered random variables bounded in the interval $[a, b]$. Then
\[
E \max_{i \in [n]} X_i \leq \frac{b - a}{\sqrt{2}} \sqrt{\log n}
\]

**Proof**

- $X = \max_{i \in [n]} X_i$. Exponentiate. Jensen’s ineq. as $x \to e^{\lambda x}$ ($\lambda > 0$) is convex:
  \[
  E X = \frac{1}{\lambda} \log e^{\lambda EX} \leq \frac{1}{\lambda} \log E e^{\lambda X}
  \]

- Bound maximum of non-negative numbers by the sum:
  \[
  E e^{\lambda X} = E e^{\lambda \max_{i \in [n]} X_i} = E \max_{i \in [n]} e^{\lambda X_i} \leq E \sum_{i=1}^{n} e^{\lambda X_i} = \sum_{i=1}^{n} E e^{\lambda X_i}
  \]

- Put everything together and use Hoeffding’s lemma ($E e^{\lambda X_i} \leq e^{\lambda^2 (b-a)^2 / 8}$):
  \[
  E \max_{i \in [n]} X_i \leq \frac{1}{\lambda} \log \sum_{i=1}^{n} e^{\lambda^2 (b-a)^2 / 8} = \frac{1}{\lambda} \log n + \frac{\lambda (b - a)^2}{8}
  \]

- Optimizing the bound $\alpha/\lambda + \lambda \beta$ over $\lambda > 0$ yields the minimum is at $\lambda = \sqrt{\alpha/\beta}$ and the optimal value $2\sqrt{\alpha \beta} = (b - a) \sqrt{\log n/2}$
Bound in expectation for finitely-many actions

Bound in expectation (Proposition 2.3)

If the loss function \( \ell \) is bounded by \( c \), we have

\[
E \max_{a \in \mathcal{A}} \{ r(a) - R(a) \} \leq c \frac{\sqrt{2 \log |\mathcal{A}|}}{\sqrt{n}}
\]

**Proof:** Same as above, using the independence of the data \( Z_1, \ldots, Z_n \) (note that for each \( a \in \mathcal{A} \), \( r(a) - R(a) \) is a centered random variable as \( ER(a) = r(a) \))

- Recall wish: \( E \sup_{a \in \mathcal{A}} \{ r(a) - R(a) \} \leq f(\text{dimension, complexity of } \mathcal{A}) \frac{1}{n^\alpha} \)

- The **dimension** of the data is superseded by the boundedness assumption

- \( \alpha = 1/2 \), slow rate

- When \( |\mathcal{A}| < \infty \), \( |\mathcal{A}| \) is a valid notion of complexity of the problem

- When \( |\mathcal{A}| = \infty \), upper bound is trivial and we need another notion of complexity
Rademacher complexity

Rademacher complexity (Definition 2.5)

The Rademacher complexity of a set $\mathcal{T} \subseteq \mathbb{R}^n$ is defined as

$$\text{Rad}(\mathcal{T}) := \mathbb{E} \sup_{t \in \mathcal{T}} \frac{1}{n} \sum_{i=1}^{n} \Omega_i t_i$$

where $\Omega_1, \ldots, \Omega_n \in \{-1, 1\}$ are i.i.d. uniform random variables (Rademacher)

- Measures of complexity: describes how well elements in $\mathcal{T}$ can replicate the sign pattern of a uniform random signal in $\mathbb{R}^n$ (see Problem 1.5)

- Useful properties:

  - $\text{Rad}(c\mathcal{T} + v) = |c| \text{Rad}(\mathcal{T})$ (Proposition 2.6)
  - $\text{Rad}(\mathcal{T} + \mathcal{T}') = \text{Rad}(\mathcal{T}) + \text{Rad}(\mathcal{T}')$ (Proposition 2.7)
  - $\text{Rad}(\text{conv}(\mathcal{T})) = \text{Rad}(\mathcal{T})$ (Proposition 2.8)

  with $\text{conv}(\mathcal{T}) = \{\sum_{j=1}^{m} w_j t_j : w \in \Delta_m, t_1, \ldots, t_m \in \mathcal{T}, m \in \mathbb{N}\}$
Rademacher complexity

**Massart’s Lemma (Lemma 2.9)**

Let $\mathcal{T} \subseteq \mathbb{R}^n$ and $\overline{t} := \frac{1}{|\mathcal{T}|} \sum_{t \in \mathcal{T}} t$. We have

$$\text{Rad}(\mathcal{T}) \leq \max_{t \in \mathcal{T}} \|t - \overline{t}\|_2 \sqrt{\frac{2 \log |\mathcal{T}|}{n}}$$

**Proof:** Similar to ones given above. **Problem 1.6**

**Contraction property - Talagrand’s Lemma (Lemma 2.10)**

Let $\mathcal{T} \subseteq \mathbb{R}^n$. For each $i \in \{1, \ldots, n\}$, let $f_i : \mathbb{R} \to \mathbb{R}$ be a $\gamma$-Lipschitz function. Then,

$$\text{Rad}((f_1, \ldots, f_n) \circ \mathcal{T}) \leq \gamma \text{Rad}(\mathcal{T})$$

with $(f_1, \ldots, f_n) \circ \mathcal{T} := \{(f_1(t_1), \ldots, f_n(t_n)) \in \mathbb{R}^n : t \in \mathcal{T}\}$

**Proof:** **Problem 1.7**