2.1 Maximal Inequalities for Sub-Gaussian Random Variables

Let $X_1, \ldots, X_n$ be a collection of sub-Gaussian random variables (not necessarily independent) with mean $\mu$ and variance proxy $\sigma^2$. Prove that
\[
E \max_{i \in [n]} (X_i - \mu) \leq \sigma \sqrt{2 \log n},
\]
\[
P \left( \max_{i \in [n]} (X_i - \mu) > \varepsilon \right) \leq ne^{-\varepsilon^2/(2\sigma^2)}.
\]

2.2 Maximal Inequalities for Linear Predictors

Let $X_1, \ldots, X_d$ be a collection of independent sub-Gaussian random variables with mean $\mu$ and variance proxy $\sigma^2$, and denote $X = (X_1, \ldots, X_d) \in \mathbb{R}^d$. Let $B := \{w \in \mathbb{R}^d : \|w\|_2 \leq 1\}$. Using covering numbers, prove that
\[
E \sup_{w \in B} w^\top (X - \mu) \leq 4\sigma \sqrt{d},
\]
\[
P \left( \sup_{w \in B} w^\top (X - \mu) > \varepsilon \right) \leq 6d e^{-\varepsilon^2/(8\sigma^2)}.
\]

Hint: Let $C \subseteq B$ be a 1/2-cover of the unit ball $B$ with respect to the $\ell_2$ norm. Relate $\sup_{w \in B} w^\top (X - \mu)$ with $\max_{c \in C} c^\top (X - \mu)$ and use that for any $w \in B$, $w^\top (X - \mu)$ is sub-Gaussian with variance proxy $\sigma^2$.

2.3 Linear Predictors $\ell_1/\ell_\infty$ Constraints

Give a proof of Proposition 3.4 in the Lecture Notes that does not use Hölder’s inequality.

2.4 Proof of Sauer-Shelah’s Lemma 4.11

In the following, let $d = \mathcal{VC}(A)$.

1. For any $n \geq 1$, prove that
\[
\tau_A(n) \leq \sum_{k=0}^{d} \binom{n}{k}
\]
using that $|A \cap \{x_1, \ldots, x_n\}| \leq |y \subseteq \{x_1, \ldots, x_n\} : A \text{ shatters } y|$. The last statement is known as Pajor’s theorem, whose proof is a non-trivial exercise in combinatorics.
2. For any $d \leq n$, prove that
\[
\sum_{k=0}^{d} \binom{n}{k} \leq \left( \frac{en}{d} \right)^d.
\]

Hint: Write $\sum_{k=0}^{d} \binom{n}{k}$ in terms of the probability mass function of the binomial distribution.

### 2.5 VC Dimension

Compute the VC dimension of the following families of classifiers.

1. Signed intervals over the real line: $A = \{ x \in \mathbb{R} \rightarrow u(21_{w^- \leq x \leq w^+} - 1) : w^- \leq w^+, u \in \{-1, 1\} \}$.

2. Axis-aligned rectangles in $\mathbb{R}^2$: $A = \{ (x_1, x_2) \in \mathbb{R}^2 \rightarrow 21_{w^-_1 \leq x_1 \leq w^+_1, w^-_2 \leq x_2 \leq w^+_2} - 1 : w^-_1 \leq w^+_1, w^-_2 \leq w^+_2 \}$.

3. Axis-aligned rectangles in $\mathbb{R}^d$: $A = \{ x \in \mathbb{R}^d \rightarrow 21_{w^-_i \leq x_i \leq w^+_i} \forall i \in [d] - 1 : w^-_i \leq w^+_i \forall i \in [d] \}$.

Let $A_1, \ldots, A_\ell$ be given hypothesis classes over the same domain $\mathcal{X}$. Let $d = \max_{i \in [\ell]} \mathsf{VC}(A_i)$. Prove that
\[
\mathsf{VC}\left( \bigcup_{i \in [\ell]} A_i \right) \leq 6d \log(3d) + 3d + 3 \log \ell.
\]

Hint: Take a set of $k$ points in $\mathcal{X}$ and assume that they are shattered by the union class $A$ so that $\tau_A(k) = 2^k$. At the same time, use Sauer-Shelah’s Lemma, Lemma 4.11, to show that $\tau_A(k) \leq \ell(ek)^d$. Use that if $x < a \log x + b$ for given $a \geq 1, b > 0$, then $x < 4a \log(2a) + 2b$ (see Lemma A.2 in the book *Understanding Machine Learning: From Theory to Algorithms* Textbook by Shai Ben-David and Shai Shalev-Shwartz)

### 2.6 Monotonicity of Packing and Covering Numbers

Prove Proposition 5.1 in the Lecture Notes.

### 2.7 Concentration Inequalities: Confidence Intervals and Sensitivity to Variance

A biased coin with heads probability $p$ is tossed $n$ times. Compute a lower bound for the probability that the number of heads obtained is between $pn - \sqrt{n}$ and $pn + \sqrt{n}$ using, respectively,

1. Markov’s inequality;
2. Chebyshev’s inequality;
3. Hoeffding’s inequality;
4. Bernstein’s inequality.

Evaluate the lower bounds in the case $p = 1/2$ (fair coin) and $p = 0.99$. Comment your findings.
2.8 Bounds in probability with empirical Rademacher complexity

Assume that the loss function \( \ell \) is bounded in the interval \([0, c]\). Prove that with probability at least \( 1 - \delta \) we have

\[
r(A^*) - r(a^*) < 4 \text{Rad}(\mathcal{L} \circ \{Z_1, \ldots, Z_n\}) + 5c\sqrt{\frac{\log(2/\delta)}{n}}.
\]

This statement is analogous (modulo constants) to the statement in Theorem 6.13, when the (deterministic) quantity \( \mathbb{E}\text{Rad}(\mathcal{L} \circ \{Z_1, \ldots, Z_n\}) \) is replaced with the (random) quantity \( \text{Rad}(\mathcal{L} \circ \{Z_1, \ldots, Z_n\}) \), a.k.a. the empirical Rademacher complexity. See also Remark 2.12. This bound is data-dependent as it depends on the training set \( Z_1, \ldots, Z_n \). The bound in Theorem 6.13 only depends on the distribution of the data (via the expected value), not on the data itself.

2.9 From Bounds in Probability to Bounds in Expectation

Let \( X \) be a random variable such that for any \( \varepsilon \in \mathbb{R} \) we have

\[
\mathbb{P}(|X - \mathbb{E}X| > \varepsilon) \leq 2e^{-\varepsilon^2/(2\sigma^2)}.
\]

Prove the following.

1. For any \( k \geq 1 \) we have

\[
\mathbb{E}[|X - \mathbb{E}X|^k] \leq (2\sigma^2)^{k/2}k\Gamma(k/2),
\]

where the Gamma function is defined as \( \Gamma(t) = \int_0^\infty x^{t-1}e^{-x}dx \) for any \( t > 0 \).

Hint: Use that for any non-negative random variable \( X \) we have \( \mathbb{E}X = \int_0^\infty \mathbb{P}(X > \varepsilon)d\varepsilon \).

2. For any \( \lambda \in \mathbb{R} \) we have

\[
\mathbb{E}e^{\lambda(X - \mathbb{E}X)} \leq e^{c\sigma^2\lambda^2},
\]

for a constant \( c > 0 \).

Hint: Use the series expansion for the exponential: \( e^x = \sum_{k=0}^\infty \frac{x^k}{k!} \), that the Gamma function is non-decreasing with \( \Gamma(k) = (k - 1)! \) and \( 2(k!)^2 \leq (2k)! \) for any integer \( k \geq 1 \).