

Introduction to Statistical Learning

Exercise set #2

Exercise 1 -

1. (Hoeffding's Lemma) Consider Z a random variable such that : $\mathbb{E}(Z) = 0$ and $\mathbb{P}(Z \in [a, b]) = 1$ almost surely. Prove the following upper bound : for any $s > 0$,

$$\mathbb{E}(e^{sZ}) \leq \exp\left(\frac{s^2(b-a)^2}{8}\right)$$

2. (Hoeffding's inequality) Consider $Z_1, \dots, Z_n$ IID over $[0, 1]$ and $\bar{Z}_n = \frac{1}{n} \sum_{i=1}^n Z_i$. Show that we have, for any $t > 0$,

$$\mathbb{P}(\bar{Z}_n - \mathbb{E}(Z_1) > t) \leq \exp(-2nt^2),$$

and

$$\mathbb{P}(\bar{Z}_n - \mathbb{E}(Z_1) < -t) \leq \exp(-2nt^2),$$

Correction 1 -

1. Using the convexity of $\exp$,

$$\begin{aligned} \mathbb{E}(e^{sZ}) &= \int_{-\infty}^{\infty} e^{sz} d\mathbb{P}_Z(z) \\ &= \int_a^b e^{sz} d\mathbb{P}_Z(z) \\ &\leq \int_a^b \left(\frac{z-a}{b-a} e^{sb} + \frac{b-z}{b-a} e^{sa} \right) d\mathbb{P}_Z(z) \\ &\leq -\frac{a}{b-a} e^{sb} + \frac{b}{b-a} e^{sa} + \underbrace{\frac{\mathbb{E}(z)}{b-a} e^{sb}}_{=0} - \underbrace{\frac{\mathbb{E}(z)}{b-a} e^{sa}}_{=0} \\ &\leq e^{sa} \frac{b - ae^{s(b-a)}}{b-a} \end{aligned}$$

Let $p := \frac{-a}{b-a}$, $u := s(b-a)$. Let $\phi_p(u) := -up + \ln(1-p+pe^u)$, so that

$$e^{\phi_p(u)} = e^{-up + \ln(1-p+pe^u)} = e^{-up} (1-p+pe^u) = e^{-sa} \left(\underbrace{1 + \frac{a}{b-a}}_{=\frac{b-a+a}{b-a}=\frac{b}{b-a}} - \frac{a}{b-a} e^{s(b-a)} \right).$$

Hence,

$$\mathbb{E}(e^{sZ}) \leq e^{\phi_p(u)}.$$

Using Taylor-Lagrange inequality and because $\phi_p(0) = \phi'_p(0) = 0$, this yields

$$\mathbb{E}(e^{sZ}) \leq e^{\frac{u^2}{2} \|\phi''_p\|_{\infty}}.$$

Conclusion. One may show that $\|\phi''_p\|_{\infty} \leq \frac{1}{4}$, so that

$$\mathbb{E}(e^{sZ}) \leq e^{\frac{u^2}{8}} = e^{\frac{s^2(b-a)^2}{8}}.$$

2. Using Markov's inequality, we have for any $s > 0$,

$$\begin{aligned}\mathbb{P}(\bar{Z}_n - \mathbb{E}(\bar{Z}_n) > t) &= \mathbb{P}(e^{s(\bar{Z}_n - \mathbb{E}(\bar{Z}_n))} > e^{st}) \\ &\leq \frac{\mathbb{E}(e^{s(\bar{Z}_n - \mathbb{E}(\bar{Z}_n))})}{e^{st}} \\ &\leq e^{-st} \prod_{i=1}^n \mathbb{E}(e^{\frac{s(Z_i - \mathbb{E}(Z_i))}{n}})\end{aligned}$$

We may apply Hoeffding's lemma on $Z_i - \mathbb{E}(Z_i) \in [-\mathbb{E}(Z_i), 1 - \mathbb{E}(Z_i)]$: for any $s > 0$,

$$\begin{aligned}\mathbb{P}(\bar{Z}_n - \mathbb{E}(\bar{Z}_n) > t) &\leq e^{-st} \prod_{i=1}^n \exp \frac{s^2}{8n^2} \\ &\leq e^{\frac{s^2}{8n} - st}\end{aligned}$$

In particular, this inequality holds for $s^* = \arg \min e^{\frac{s^2}{8n} - st}$. One can show that $s^* = 4nt$ and $e^{\frac{(4nt)^2}{8n} - (4nt)t} = e^{-2nt^2}$

Conclusion. Because $\mathbb{P}(\bar{Z}_n - \mathbb{E}(\bar{Z}_n) > t)$ does not depend on s , we can conclude that

$$\mathbb{P}(\bar{Z}_n - \mathbb{E}(\bar{Z}_n) > t) \leq e^{-2nt^2}.$$

The same reasoning can be done on $\mathbb{E}(Z_i) - Z_i$ to prove that

$$\mathbb{P}(\bar{Z}_n - \mathbb{E}(\bar{Z}_n) < -t) \leq e^{-2nt^2}.$$

Exercise 2 -

1. (Azuma's inequality) Consider $V = (V_1, \dots, V_n, \dots)$ and $Z = (Z_1, \dots, Z_n, \dots)$ two sequences of random variables. We assume the following : for any $n \geq 1$,

- V_n is a function of $Z_1, \dots, Z_n$,
- $\mathbb{E}(V_{n+1} \mid Z_1, \dots, Z_n) = 0$,
- there exists U_n a measurable function of $Z_1, \dots, Z_{n-1}$ and $c_n \geq 0$ such that : $U_n \leq V_n \leq U_n + c_n$.

Prove that, for any $t > 0$,

$$\mathbb{P}\left(\sum_{i=1}^n V_i > t\right) \leq \exp\left(-\frac{2t^2}{\sum_{i=1}^n c_i^2}\right)$$

and

$$\mathbb{P}\left(\sum_{i=1}^n V_i < -t\right) \leq \exp\left(-\frac{2t^2}{\sum_{i=1}^n c_i^2}\right)$$

2. (McDiarmid's Inequality) Use the previous question to prove McDiarmid's inequality.

Correction 2 -

1. First show an intermediate lemma.

Lemma 1. Let V, Z such that $\mathbb{E}(V \mid Z) = 0$ and assume there exists f a measurable function and $c \geq 0$ such that $f(Z) \leq V \leq f(Z) + C$. Then, we have that

$$\forall s > 0, \quad \mathbb{E}(e^{sV} \mid Z) \leq e^{s^2 c^2 / 8}.$$

Proof. For any $z \in \mathbb{R}$, let $a = f(z), b = f(z) + C$. Then, as in the proof of Hoeffding's lemma, we can use the convexity of exp:

$$\forall s > 0, \mathbb{E}(e^{sV} \mid Z = z) \leq \frac{b - \mathbb{E}(V \mid Z = z)}{b - a} e^{sa} + \frac{\mathbb{E}(V \mid Z = z) - a}{b - a} e^{sb}.$$

The rest of the proof is essentially the same, yielding the result. $\square$

Using Markov's inequality, we have, for any $s > 0$,

$$\begin{aligned} \mathbb{P}\left(\sum_{i=1}^n V_i > t\right) &= \mathbb{P}\left(\exp\left(s \sum_{i=1}^n V_i\right) > e^{st}\right) \\ &\leq \frac{\mathbb{E}(\exp(\sum_{i=1}^n V_i))}{\exp(st)} \\ &\leq \exp(-st) \mathbb{E}\left[\mathbb{E}\left(\exp\left(sV_n + s \sum_{i=1}^{n-1} V_i\right) \mid Z_1, \dots, Z_{n-1}\right)\right] \\ &\leq \exp(-st) \mathbb{E}\left[e^{s^2 c_n^2 / 8} \mathbb{E}\left(\exp\left(s \sum_{i=1}^{n-1} V_i\right) \mid Z_1, \dots, Z_{n-1}\right)\right] \\ &\leq \exp(-st + s^2 c_n^2 / 8) \mathbb{E}\left[\exp\left(s \sum_{i=1}^{n-1} V_i\right)\right] \end{aligned}$$

where we used the lemma on $V_n \mid (Z_1, \dots, Z_{n-1})$. Then, a recurrence yields

$$\mathbb{P}\left(\sum_{i=1}^n V_i > t\right) \leq \exp\left(-st + \sum_{i=1}^n s^2 c_i^2 / 8\right).$$

Conclusion. The result is true in particular for $s := \frac{4t}{\sum_{i=1}^n c_i^2}$, yielding

$$\mathbb{P}\left(\sum_{i=1}^n V_i > t\right) \leq \exp\left(-\frac{2t^2}{\sum_{i=1}^n c_i^2}\right).$$

The same reasoning also gives

$$\mathbb{P}\left(\sum_{i=1}^n V_i < -t\right) \leq \exp\left(-\frac{2t^2}{\sum_{i=1}^n c_i^2}\right).$$

2. Recall McDiarmid's inequality.

Bounded difference assumption. We assume that for some $c > 0$,

$$\sup_{z_1, \dots, z_n, z'_i} |f(z_1, \dots, z_n) - f(z_1, \dots, z'_i, \dots, z_n)| \leq \frac{c}{n}.$$

Theorem 1 (McDiarmid's Inequality). Consider $Z_1, \dots, Z_n$ IID. Under the bounded difference assumption, we have for any $t > 0$,

$$\begin{aligned} \mathbb{P}(f(Z_1, \dots, Z_n) - \mathbb{E}(f(Z_1, \dots, Z_n)) > t) &\leq \exp\left(\frac{-2nt^2}{c^2}\right) \\ \mathbb{P}(f(Z_1, \dots, Z_n) - \mathbb{E}(f(Z_1, \dots, Z_n)) < -t) &\leq \exp\left(\frac{-2nt^2}{c^2}\right) \end{aligned}$$

Proof. Let $V_i := [\mathbb{E}(f(Z_1, \dots, Z_n) \mid Z_1, \dots, Z_i) - \mathbb{E}(f(Z_1, \dots, Z_n) \mid Z_1, \dots, Z_{i-1})]$ (function of $Z_1, \dots, Z_i$ — first assumption of Azuma). Then,

$$\begin{aligned} \sum_{i=1}^n V_i &= \mathbb{E}(f(Z_1, \dots, Z_n) \mid Z_1, \dots, Z_n) - \mathbb{E}(f(Z_1, \dots, Z_n)) \\ &= f(Z_1, \dots, Z_n) - \mathbb{E}(f(Z_1, \dots, Z_n)) \end{aligned}$$

Moreover,

$$\begin{aligned} \mathbb{E}(V_i \mid Z_1, \dots, Z_{i-1}) &= \mathbb{E}(\mathbb{E}(f(Z_1, \dots, Z_n) \mid Z_1, \dots, Z_i) \\ &\quad - \mathbb{E}(f(Z_1, \dots, Z_n) \mid Z_1, \dots, Z_{i-1}) \mid Z_1, \dots, Z_{i-1}) \\ &= \mathbb{E}(\mathbb{E}(f(Z_1, \dots, Z_n) \mid Z_1, \dots, Z_i) \mid Z_1, \dots, Z_{i-1}) \\ &\quad - \mathbb{E}(f(Z_1, \dots, Z_n) \mid Z_1, \dots, Z_{i-1})) \\ &= \mathbb{E}(f(Z_1, \dots, Z_n) \mid Z_1, \dots, Z_{i-1}) - \mathbb{E}(f(Z_1, \dots, Z_n) \mid Z_1, \dots, Z_{i-1}) \\ &= 0, \end{aligned}$$

which ensures the second assumption of Azuma. Now, for any $i = 1, \dots, n$, let

$$\begin{aligned} W_i &:= \sup_z \mathbb{E}(f(Z_1, \dots, Z_n) \mid Z_1, \dots, Z_{i-1}, Z_i = z') - \mathbb{E}(f(Z_1, \dots, Z_n) \mid Z_1, \dots, Z_{i-1}), \\ X_i &:= \inf_{z'} \mathbb{E}(f(Z_1, \dots, Z_n) \mid Z_1, \dots, Z_{i-1}, Z_i = z') - \mathbb{E}(f(Z_1, \dots, Z_n) \mid Z_1, \dots, Z_{i-1}), \end{aligned}$$

so that $X_i \leq V_i \leq W_i$, and

$$\begin{aligned} W_i - X_i &\leq \sup_{z, z'} (\mathbb{E}(f(Z_1, \dots, Z_n) \mid Z_1, \dots, Z_{i-1}, Z_i = z) \\ &\quad - \mathbb{E}(f(Z_1, \dots, Z_n) \mid Z_1, \dots, Z_{i-1}, Z_i = z')) \\ &\leq \frac{c}{n}, \end{aligned}$$

by bounded difference assumption. This proves the third assumption of Azuma: $X_i \leq V_i \leq X_i + \frac{c}{n}$. All the assumptions of Azuma's inequality for V_i are now verified.

Conclusion. Applying Azuma on V_i yields

$$\begin{aligned} \mathbb{P}(f(Z_1, \dots, Z_n) - \mathbb{E}(f(Z_1, \dots, Z_n)) > t) &\leq \exp\left(\frac{-2nt^2}{c^2}\right), \\ \mathbb{P}(f(Z_1, \dots, Z_n) - \mathbb{E}(f(Z_1, \dots, Z_n)) < -t) &\leq \exp\left(\frac{-2nt^2}{c^2}\right). \end{aligned}$$

□

Exercise 3 - (Application of McDiarmid's concentration inequality) Let $\mathcal{F}$ be a class of $[0, 1]$ -valued functions. Show that, with probability at least $1 - \delta$,

$$\mathcal{R}_n(\mathcal{F}) \leq \hat{\mathcal{R}}_n(\mathcal{F}) + \sqrt{\frac{\log(1/\delta)}{2n}},$$

and also that

$$\sup_{f \in \mathcal{F}} \left(\mathbb{E}(f(Z_1)) - \frac{1}{n} \sum_{i=1}^n f(Z_i) \right) \leq 2\hat{\mathcal{R}}_n(\mathcal{F}) + 3\sqrt{\frac{\log(2/\delta)}{2n}}.$$

Correction 3 - Let $D_n := (Z_1, \dots, Z_n)$. Let

$$h(Z_1, \dots, Z_n) := \hat{\mathcal{R}}_n(\mathcal{F}) = \mathbb{E}_\varepsilon \left(\sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^n \varepsilon_i f(Z_i) \mid D_n \right),$$

where the expectation is taken over ε only. Let us verify the assumptions of McDiarmid's inequality:

- $Z_1, \dots, Z_n$ are IID.
- h verifies the bounded difference assumption: let $j \in \{1, \dots, n\}$, $z_1, \dots, z_n, z, z'$, and $d_n = (z_1, \dots, z_j = z, \dots, z_n)$, $d'_n = (z_1, \dots, z_j = z', \dots, z_n)$. Fix $\varepsilon \in \{-1, 1\}^n$ and define

$$S(f) := \sum_{i \neq j} \varepsilon_i f(z_i), \quad A(z) := \sup_{f \in \mathcal{F}} \frac{1}{n} [\varepsilon_j f(z) + S(f)],$$

so that $h(d_n) = \mathbb{E}_\varepsilon(A(z))$ and $h(d'_n) = \mathbb{E}_\varepsilon(A(z'))$. For any $f \in \mathcal{F}$, since f is $[0, 1]$ -valued, $|f(z) - f(z')| \leq 1$, hence

$$\frac{1}{n} [\varepsilon_j f(z) + S(f)] = \frac{1}{n} [\varepsilon_j f(z') + S(f)] + \frac{\varepsilon_j}{n} (f(z) - f(z')) \leq A(z') + \frac{1}{n}.$$

Taking the supremum over $f \in \mathcal{F}$ gives $A(z) \leq A(z') + \frac{1}{n}$, and by symmetry $|A(z) - A(z')| \leq \frac{1}{n}$. Taking the expectation over ε ,

$$|h(d_n) - h(d'_n)| \leq \mathbb{E}_\varepsilon (|A(z) - A(z')|) \leq \frac{1}{n}.$$

Hence h verifies the bounded difference assumption with $c = 1$.

We can now apply McDiarmid's inequality: for any $t > 0$,

$$\begin{aligned} \mathbb{P}(h(Z_1, \dots, Z_n) - \mathbb{E}(h(Z_1, \dots, Z_n)) > t) &\leq \exp(-2nt^2) \\ \mathbb{P}(h(Z_1, \dots, Z_n) - \mathbb{E}(h(Z_1, \dots, Z_n)) < -t) &\leq \exp(-2nt^2) \end{aligned}$$

Since $h(Z_1, \dots, Z_n) = \hat{\mathcal{R}}_n(\mathcal{F})$ and $\mathbb{E}(h(Z_1, \dots, Z_n)) = \mathcal{R}_n(\mathcal{F})$, this gives with $\delta := e^{-2nt^2} \iff t = \sqrt{\frac{\ln 1/\delta}{2n}}$,

$$\mathbb{P} \left(\hat{\mathcal{R}}_n(\mathcal{F}) - \mathcal{R}_n(\mathcal{F}) < -\sqrt{\frac{\ln 1/\delta}{2n}} \right) \leq \delta$$

Conclusion.

$$\mathcal{R}_n(\mathcal{F}) \leq \hat{\mathcal{R}}_n(\mathcal{F}) + \sqrt{\frac{\ln 1/\delta}{2n}} \quad \text{with probability } 1 - \delta.$$

Let $\phi(Z_1, \dots, Z_n) := \sup_{f \in \mathcal{F}} \left(\mathbb{E}(f(Z_1)) - \frac{1}{n} \sum_{i=1}^n f(Z_i) \right)$. ϕ verifies the bounded difference assumption with $c = 1$. Thanks to McDiarmid, we have

$$\mathbb{P} \left(\phi(Z_1, \dots, Z_n) > \mathbb{E}(\phi(Z_1, \dots, Z_n)) + \sqrt{\frac{\ln 2/\delta}{2n}} \right) \leq \frac{\delta}{2},$$

which ensures $\phi(Z_1, \dots, Z_n) \leq \mathbb{E}(\phi(Z_1, \dots, Z_n)) + \sqrt{\frac{\ln 2/\delta}{2n}}$ with probability $1 - \frac{\delta}{2}$. We now have to

bound $\mathbb{E}(\phi(Z_1, \dots, Z_n))$ to conclude.

$$\begin{aligned}
\mathbb{E}(\phi(Z_1, \dots, Z_n)) &= \mathbb{E}_Z \left(\sup_{f \in \mathcal{F}} \left(\mathbb{E}(f(Z_1)) - \frac{1}{n} \sum_{i=1}^n f(Z_i) \right) \right) \\
&= \mathbb{E}_Z \left(\sup_{f \in \mathcal{F}} \left(\mathbb{E}_{Z'} \left(\frac{1}{n} \sum_{i=1}^n f(Z'_i) \right) - \frac{1}{n} \sum_{i=1}^n f(Z_i) \right) \right) \\
&\leq \mathbb{E}_{Z, Z'} \left(\sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^n (f(Z'_i) - f(Z_i)) \right) \\
&\leq \mathbb{E}_{Z, Z', \varepsilon} \left(\sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^n \varepsilon_i (f(Z'_i) - f(Z_i)) \right) \\
&\leq \mathbb{E}_{Z, \varepsilon} \left(\sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^n \varepsilon_i f(Z_i) \right) + \mathbb{E}_{Z', \varepsilon} \left(\sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^n -\varepsilon_i f(Z'_i) \right) \\
&\leq 2\mathcal{R}_n(\mathcal{F}).
\end{aligned}$$

The fourth line holds because $\mathcal{L}(f(Z'_i) - f(Z_i)) = \mathcal{L}(\varepsilon_i(f(Z'_i) - f(Z_i)))$, and the fifth one because $\sup(x + y) \leq \sup x + \sup y$. Now, we can conclude by combining the two following inequalities:

$$\begin{aligned}
\sup_{f \in \mathcal{F}} \left(\mathbb{E}(f(Z_1)) - \frac{1}{n} \sum_{i=1}^n f(Z_i) \right) &\leq 2\mathcal{R}_n(\mathcal{F}) + \sqrt{\frac{\ln 2/\delta}{2n}} \quad \text{with probability } 1 - \frac{\delta}{2}, \\
\mathcal{R}_n(\mathcal{F}) &\leq \hat{\mathcal{R}}_n(\mathcal{F}) + \sqrt{\frac{\ln 2/\delta}{2n}} \quad \text{with probability } 1 - \frac{\delta}{2}.
\end{aligned}$$

Conclusion. With probability $1 - \delta$,

$$\sup_{f \in \mathcal{F}} \left(\mathbb{E}(f(Z_1)) - \frac{1}{n} \sum_{i=1}^n f(Z_i) \right) \leq 2\hat{\mathcal{R}}_n(\mathcal{F}) + 3\sqrt{\frac{\ln 2/\delta}{2n}}.$$

Exercise 4 - (Sauer's lemma) Consider $\mathcal{A}$ a class of subsets of $\mathbb{R}^d$ with VC dimension $V < \infty$ and growth $\gamma(\mathcal{A}, n)$, $\forall n \geq 1$. Show that :

$$\forall n \geq 1, \gamma(\mathcal{A}, n) \leq \sum_{i=0}^V \binom{n}{i}.$$

Correction 4 - Recall that the growth $\gamma(\mathcal{A}, n)$ is defined by

$$\gamma(\mathcal{A}, n) := \max_{x_1, \dots, x_n \in \mathbb{R}^d} \left| \{A \cap \{x_1, \dots, x_n\} : A \in \mathcal{A}\} \right|,$$

i.e., the maximal number of distinct subsets of $\{x_1, \dots, x_n\}$ that can be obtained by intersecting it with the sets of $\mathcal{A}$. First, because the number of subsets of any set of n points is bounded by

$$\sum_{k=0}^n \binom{n}{k} = 2^n, \text{ we have}$$

$$\gamma(\mathcal{A}, n) \leq 2^n.$$

The VC dimension of $\mathcal{A}$ is defined by

$$V := \max\{n \in \mathbb{N} : \gamma(\mathcal{A}, n) = 2^n\}.$$

Shattering. Let S be a finite set and $\mathcal{C}$ a family of subsets of S . We say that $T \subset S$ is *shattered* by $\mathcal{C}$ if every subset of T can be obtained by intersecting T with an element of $\mathcal{C}$:

$$\forall U \subset T, \exists B \in \mathcal{C}, \quad B \cap T = U.$$

We denote by $\text{sh}(\mathcal{C})$ the set of all the subsets of S shattered by $\mathcal{C}$. Note that $\emptyset \in \text{sh}(\mathcal{C})$ as soon as $\mathcal{C}$ is not empty.

Intuition. If you think of $\mathcal{C}$ as the $\{-1, +1\}$ -valued functions, saying that T is shattered by $\mathcal{C}$ means that for any combination of values $\{-1, +1\}$ taken by the elements $x_i \in T$, there exists a classifier in $\mathcal{C}$ that achieves zero-loss on T .

The proof of the exercise relies on the following lemma.

Lemma 2. *Let S be a finite set and $\mathcal{C}$ a family of subsets of S . Then $|\mathcal{C}| \leq |\text{sh}(\mathcal{C})|$.*

Proof. By induction on $m := |S|$.

Case $m = 0$. Then $S = \emptyset$, and either $\mathcal{C} = \emptyset$ and $|\mathcal{C}| = 0$, or $\mathcal{C} = \{\emptyset\}$ and $\text{sh}(\mathcal{C}) = \{\emptyset\}$, so that $|\mathcal{C}| = 1 = |\text{sh}(\mathcal{C})|$.

Induction step. Assume the result for sets of size $m - 1$, and let $|S| = m \geq 1$. Fix $x \in S$, let $S' := S \setminus \{x\}$, and split $\mathcal{C}$ according to whether its elements contain x or not:

$$\mathcal{C}_0 := \{B \in \mathcal{C} : x \notin B\}, \quad \mathcal{C}_1 := \{B \setminus \{x\} : B \in \mathcal{C}, x \in B\}.$$

Both are families of subsets of S' , and $|\mathcal{C}| = |\mathcal{C}_0| + |\mathcal{C}_1|$. The induction hypothesis applied on S' gives

$$|\mathcal{C}| = |\mathcal{C}_0| + |\mathcal{C}_1| \leq |\text{sh}(\mathcal{C}_0)| + |\text{sh}(\mathcal{C}_1)| = |\text{sh}(\mathcal{C}_0) \cup \text{sh}(\mathcal{C}_1)| + |\text{sh}(\mathcal{C}_0) \cap \text{sh}(\mathcal{C}_1)|,$$

because $|A| + |B| = |A \cup B| + |A \cap B|$. We now exhibit as many distinct sets shattered by $\mathcal{C}$.

- Let $T \in \text{sh}(\mathcal{C}_0) \cup \text{sh}(\mathcal{C}_1)$ and $U \subset T$. There exists $B \in \mathcal{C}$ such that $B \cap T = U$ or $(B \setminus \{x\}) \cap T = U$. Because $x \notin T$, both cases give $B \cap T = U$. Hence $T \in \text{sh}(\mathcal{C})$, and $x \notin T$.
- Let $T \in \text{sh}(\mathcal{C}_0) \cap \text{sh}(\mathcal{C}_1)$ and $U \subset T \cup \{x\}$. If $x \notin U$, as $T \in \text{sh}(\mathcal{C}_0)$ there exists $B \in \mathcal{C}$ with $x \notin B$ and $B \cap T = U$, so that $B \cap (T \cup \{x\}) = U$. If $x \in U$, as $T \in \text{sh}(\mathcal{C}_1)$ there exists $B \in \mathcal{C}$ with $x \in B$ and $(B \setminus \{x\}) \cap T = U \setminus \{x\}$, so that $B \cap (T \cup \{x\}) = U$. Hence $T \cup \{x\} \in \text{sh}(\mathcal{C})$, and $x \in T \cup \{x\}$.

The sets of the first kind do not contain x , those of the second kind do. Therefore all these shattered sets are distinct, and

$$|\text{sh}(\mathcal{C})| \geq |\text{sh}(\mathcal{C}_0) \cup \text{sh}(\mathcal{C}_1)| + |\text{sh}(\mathcal{C}_0) \cap \text{sh}(\mathcal{C}_1)| \geq |\mathcal{C}|. \quad \square$$

Let $n \geq 1$, $x_1, \dots, x_n \in \mathbb{R}^d$, $S := \{x_1, \dots, x_n\}$ and $m := |S| \leq n$. Consider $\mathcal{C}$ defined by

$$\mathcal{C} := \{A \cap S : A \in \mathcal{A}\}.$$

Let $T \in \text{sh}(\mathcal{C})$ and $k := |T|$. Since $T \subset S$, we have $(A \cap S) \cap T = A \cap T$, so that $\{A \cap T : A \in \mathcal{A}\}$ contains the 2^k subsets of T , because T is shattered by $\mathcal{C}$. Thus $\gamma(\mathcal{A}, k) = 2^k$, and $k \leq V$ by definition of V . In other words, $\text{sh}(\mathcal{C})$ only contains subsets of S of size at most V , and there are $\sum_{i=0}^V \binom{m}{i}$ of them. The lemma then yields

$$\left| \{A \cap \{x_1, \dots, x_n\} : A \in \mathcal{A}\} \right| = |\mathcal{C}| \leq |\text{sh}(\mathcal{C})| \leq \sum_{i=0}^V \binom{m}{i} \leq \sum_{i=0}^V \binom{n}{i}.$$

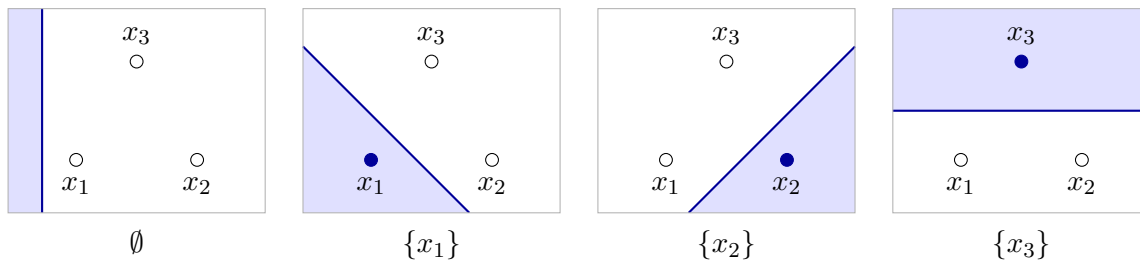
Conclusion. The bound holds for any $x_1, \dots, x_n \in \mathbb{R}^d$, so taking the maximum gives

$$\forall n \geq 1, \quad \gamma(\mathcal{A}, n) \leq \sum_{i=0}^V \binom{n}{i}.$$

Exercise 5 - (VC dimension of half-spaces) Consider $\mathcal{A}$ the class of half-spaces and show that its VC dimension is $V(\mathcal{A}) = d + 1$. *Hint: first prove the upper bound on Radon's theorem, and then build a separating hyperplane for any arbitrary labeling for some set of $d + 1$ points.*

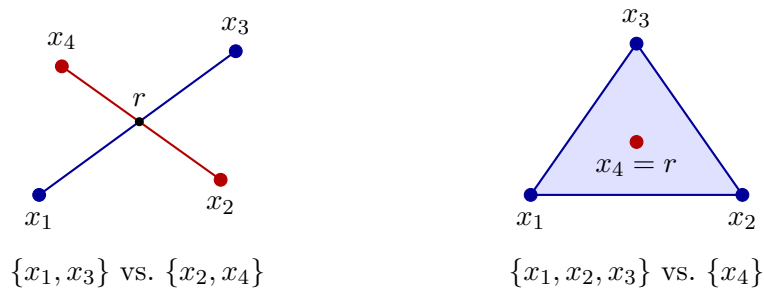
Correction 5 -

Lower bound ($d = 2$). Three non-aligned points are shattered by half-planes.



The four remaining subsets are obtained with the opposite half-planes.

Upper bound ($d = 2$). Four points always split into two groups whose convex hulls intersect: no line separates them.



A half-plane and its complement are both convex: if they separated the two groups, the common point r would lie in both.

Conclusion. Graphically, in the case $d = 2$, we have $V(\mathcal{A}) = 3$.

Upper bound using Radon's theorem.

Theorem 2 (Radon's theorem). *Any set of $d + 2$ points in $\mathbb{R}^d$ can be partitioned into two sets whose convex hull intersect, i.e.,*

$$\forall X = \{x_1, \dots, x_{d+2}\} \subset \mathbb{R}^d, \exists X_1, X_2 \subset \mathbb{R}^d, X_1 \cup X_2 = X, X_1 \cap X_2 = \emptyset, \text{Conv}(X_1) \cap \text{Conv}(X_2) \neq \emptyset.$$

Note that if two sets can be separated by a hyperplane, their convex hull also must be separated by this same hyperplane. For the sets X_1 and X_2 from Radon's theorem, this would mean that $\text{Conv}(X_1)$ and $\text{Conv}(X_2)$ are separated by a hyperplane, while having a non-empty intersection, which is absurd. Hence, $V(\mathcal{A}) < d + 2$.

Lower bound. The sets of half-spaces of $\mathbb{R}^d$ are described by

$$\mathcal{A} := \{x = (x^1, \dots, x^d) \in \mathbb{R}^d, \exists c \in \mathbb{R}, \alpha \in \mathbb{R}^d, \alpha^\top x + c \geq 0\}.$$

Let the $d + 1$ points be the origin and the canonical base $e = (0_{\mathbb{R}^d}, e_1, \dots, e_d)$. Then, the half-space defined by $Y^\top Px + \frac{y_0}{2} \geq 0$ verifies

$$\forall i \in 0, \dots, d, Y^\top Px_i + \frac{y_0}{2} = y_i + \frac{y_0}{2} \begin{cases} < 0 & \text{if } y_i = -1, \\ > 0 & \text{if } y_i = +1. \end{cases}$$

Hence, $d + 1 \leq V(\mathcal{A})$.

Conclusion. In the general case, $V(\mathcal{A}) = d + 1$.

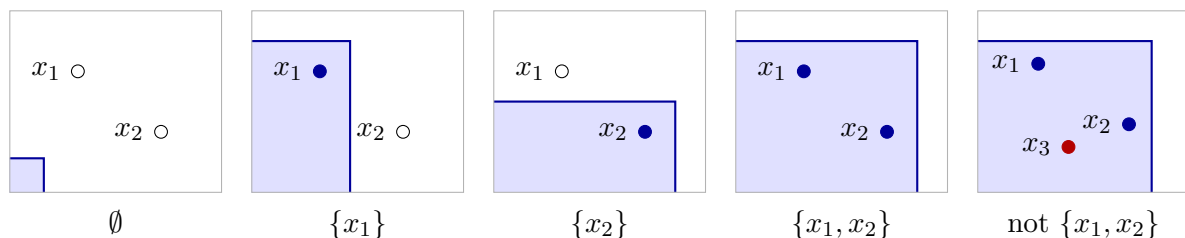
Exercise 6 - Compute the VC dimension $V(\mathcal{A})$ in the following cases:

1. $\mathcal{A} = \{[-\infty, x_1] \times \cdots \times [-\infty, x_d], x_1, \dots, x_d \in \mathbb{R}\}$,
2. $\mathcal{A}$ is the class of all rectangles of $\mathbb{R}^2$ with axis-orthogonal edges.
3. $\mathcal{A}$ is the class of all rectangles of $\mathbb{R}^2$.
4. $\mathcal{A}$ is the class of all triangles of $\mathbb{R}^2$.

Correction 6 -

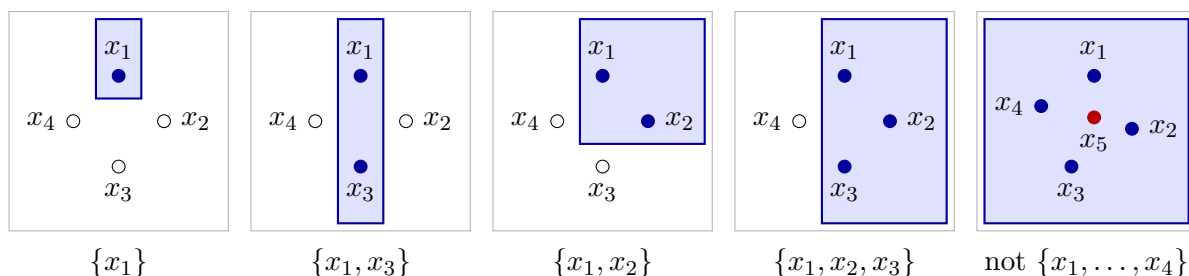
1. $\mathcal{A} = \{[-\infty, x_1] \times \cdots \times [-\infty, x_d], x_1, \dots, x_d \in \mathbb{R}\}$. In general, $V(\mathcal{A}) = d$.
 - *Lower bound.* The canonical basis $\{e_1, \dots, e_d\}$ is shattered: for $S \subset \{1, \dots, d\}$, the quadrant $\prod_{k=1}^d [-\infty, a_k]$ with $a_k = 1$ if $k \in S$ and $a_k = \frac{1}{2}$ otherwise contains e_k if and only if $k \in S$.
 - *Upper bound.* Among $d + 1$ points, select for each coordinate k a point maximizing this coordinate. This gives at most d points, and any quadrant containing them contains all the $d + 1$ points: this selection cannot be obtained without the remaining points.

For $d = 2$, this gives $V(\mathcal{A}) = 2$:



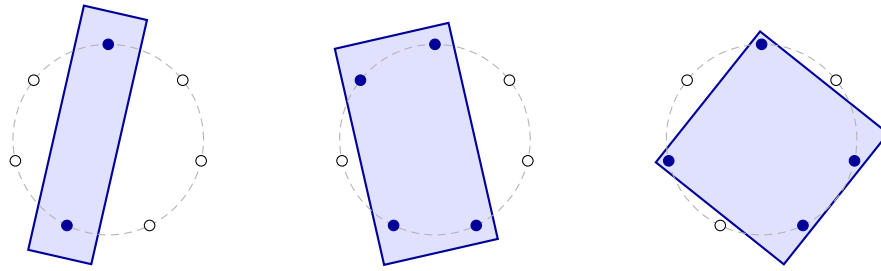
Two points are shattered (first four panels). With three points, a quadrant containing the topmost and the rightmost points always contains the third one (last panel).

2. $\mathcal{A}$ is the class of all rectangles of $\mathbb{R}^2$ with axis-orthogonal edges. $V(\mathcal{A}) = 4$.

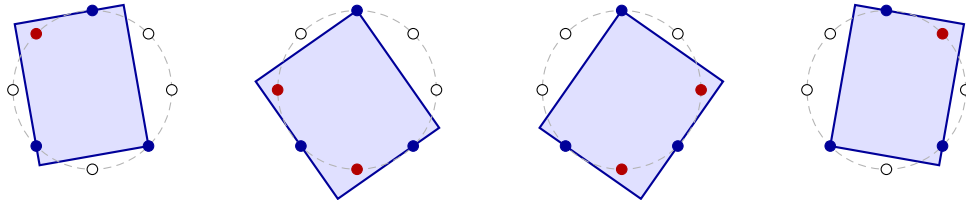


Four points in diamond position are shattered (first four panels; the other subsets are obtained similarly). With five points, a rectangle containing the topmost, rightmost, bottommost and leftmost points always contains the fifth one (last panel).

3. $\mathcal{A}$ is the class of all rectangles of $\mathbb{R}^2$. $V(\mathcal{A}) = 7$.

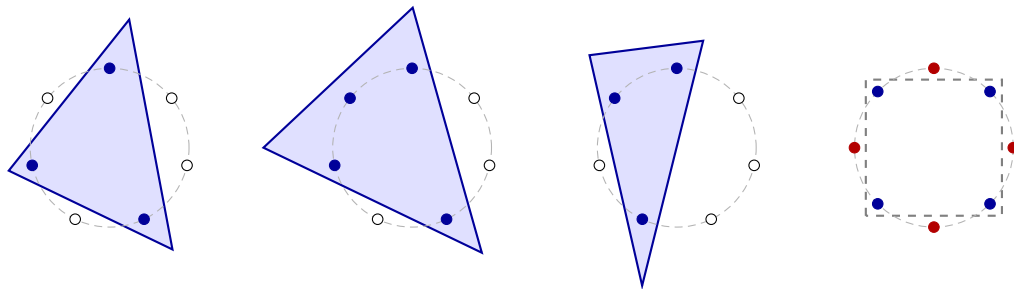


Seven points regularly spaced on a circle are shattered: three examples, with the selected points in blue.



Upper bound on an example: with eight points regularly spaced on a circle, the three blue points cannot be selected. Whatever its orientation, the smallest rectangle containing them contains another point, in red (four orientations shown).

4. $\mathcal{A}$ is the class of all triangles of $\mathbb{R}^2$. $V(\mathcal{A}) = 7$.



Seven points on a circle are shattered: each side of the triangle cuts off one group of consecutive points, and there are at most three groups to cut off (first three panels). With eight points on a circle, selecting every other point leaves four groups to cut off, which requires four sides (last panel).

Exercise 7 - (Relation between Rademacher average and combinatorial complexities)

Consider a class $\mathcal{G}$ of binary valued functions with shattering coefficient function denoted by $\gamma(\mathcal{G}, n)$. Show that :

$$\mathbb{E}(\hat{\mathcal{R}}_n(\mathcal{G})) \leq \sqrt{\frac{2 \ln(\gamma(\mathcal{G}, n))}{n}}$$

and if $V(\mathcal{G}) < \infty$ then show that, for some constant $C > 0$, we have :

$$\mathbb{E}(\hat{\mathcal{R}}_n(\mathcal{G})) \leq C \sqrt{\frac{V(\mathcal{G}) \log(n)}{n}}$$

Correction 7 - Let $D_n := (Z_1, \dots, Z_n)$ be the sample. By definition, the empirical Rademacher average is a random variable depending on D_n :

$$\hat{\mathcal{R}}_n(\mathcal{G}) = \mathbb{E}_\varepsilon \left(\sup_{g \in \mathcal{G}} \frac{1}{n} \sum_{i=1}^n \varepsilon_i g(Z_i) \mid D_n \right),$$

where the expectation is taken over ε only, the sample being fixed. Let $\mathcal{G}_{D_n}$ be the ensemble of $(g(z_1), \dots, g(z_n))$ for any $g \in \mathcal{G}$:

$$\mathcal{G}_{D_n} := \{(g(z_i))_{1 \leq i \leq n}, g \in \mathcal{G}\},$$

so that a change a variables yield

$$\hat{\mathcal{R}}_n(\mathcal{G}) = \mathbb{E}_\varepsilon \left(\sup_{z \in \mathcal{G}_{D_n}} \frac{1}{n} \sum_{i=1}^n \varepsilon_i z_i \mid D_n \right).$$

But because g is $\{-1, +1\}$ -valued, we have $\|z\| \leq \sqrt{n}$. Applying Massart's lemma, this gives

$$\hat{\mathcal{R}}_n(\mathcal{G}) \leq \frac{\sqrt{2n \ln |\mathcal{G}_{D_n}|}}{n}.$$

But we also have the bound

$$|\mathcal{G}_{D_n}| \leq \gamma(\mathcal{G}, n) := \max_{(z_i)_{1 \leq i \leq n}} |\{(g(z_i))_{1 \leq i \leq n}, g \in \mathcal{G}\}|.$$

Hence, for any sample D_n ,

$$\hat{\mathcal{R}}_n(\mathcal{G}) \leq \sqrt{\frac{2 \ln \gamma(\mathcal{G}, n)}{n}},$$

and taking the expectation over D_n gives $\mathbb{E}(\hat{\mathcal{R}}_n(\mathcal{G})) \leq \sqrt{\frac{2 \ln \gamma(\mathcal{G}, n)}{n}}$. Now, assume that $V(\mathcal{G}) < +\infty$. If $n \leq V(\mathcal{G})$, it means that $\gamma(\mathcal{G}, n) = 2^n$, so that

$$\hat{\mathcal{R}}_n(\mathcal{G}) \leq \sqrt{2}.$$

If $n > V(\mathcal{G}) =: V$, $\gamma(\mathcal{G}, n) \leq \left(\frac{en}{V}\right)^V$, and

$$\begin{aligned} \hat{\mathcal{R}}_n(\mathcal{G}) &\leq \sqrt{\frac{2 \ln \left(\frac{en}{V}\right)^V}{n}} \\ &\leq \sqrt{\frac{2V \left(\ln \left(\frac{n}{V}\right) + 1\right)}{n}} \\ &\leq \sqrt{2(\ln V - 1) \frac{V \ln n}{n}} \\ &\leq C \sqrt{\frac{V \ln n}{n}}. \end{aligned}$$

Conclusion. In both cases, for any sample D_n , $\hat{\mathcal{R}}_n(\mathcal{G}) \leq C \sqrt{\frac{V \ln n}{n}}$. Taking the expectation,

$$\mathbb{E}(\hat{\mathcal{R}}_n(\mathcal{G})) \leq C \sqrt{\frac{V \ln n}{n}}.$$