1 Hoeffding’s inequality

We prove Hoeffding’s lemma and leave Hoeffding’s inequality as an exercise.

**Definition 2.1.** Let $X$ be an arbitrary domain. A function $f : X \to \mathbb{R}$ is called convex if:

\[
(\forall a, b \in X, s \in [0, 1]) \ f((1 - s)a + sb) \leq (1 - s)f(a) + sf(b)
\]

**Lemma 2.1** (Hoeffding’s lemma). Assume that the random variable $x \in [0, 1]$ has mean $\mathbb{E}[x] = \mu$. We have that:

\[
\mathbb{E}[e^{t(x-\mu)}] \leq e^{\frac{1}{8}t^2}
\]

for all $t \in \mathbb{R}$.

**Proof.** Invoke Definition 2.1 with $f(x) = e^{t(x-\mu)}$, $a = 0$, $b = 1$:

\[
(\forall s \in [0, 1]) \ f(s) \leq (1 - s)f(0) + sf(1)
\]

\[
\Rightarrow (\forall x \in [0, 1]) \ f(x) \leq (1 - x)f(0) + xf(1)
\]

\[
\Rightarrow (\forall x \in [0, 1]) \ e^{t(x-\mu)} \leq (1 - x)e^{-t\mu} + xe^{t(1-\mu)}
\]

By computing expectations on both sides, we get:

\[
\mathbb{E}[e^{t(x-\mu)}] \leq (1 - \mathbb{E}[x])e^{-t\mu} + \mathbb{E}[x]e^{t(1-\mu)}
\]

\[
= (1 - \mu)e^{-t\mu} + \mu e^{t(1-\mu)}
\]

\[
= e^{-t\mu}(1 - \mu + \mu e^t)
\]

\[
= e^{g(t)}
\]

where:

\[
g(t) = -t\mu + \log(1 - \mu + \mu e^t)
\]

It is easy to note that $g(0) = 0$ and that:

\[
\frac{\partial g}{\partial t}(t) = -\mu + \frac{\mu e^t}{1 - \mu + \mu e^t} \Rightarrow \frac{\partial g}{\partial t}(0) = 0
\]
Let \( w = \frac{\mu e^t}{1 - \mu + \mu e^t} \), then:

\[
\frac{\partial^2 g}{\partial t^2} (t) = \frac{\mu e^t (1 - \mu + \mu e^t) - \mu e^t \mu e^t}{(1 - \mu + \mu e^t)^2} = w(1 - w) \leq 1/4
\]

By Taylor’s theorem, for every real \( t \) there exists a \( v \in [0, t] \) such that:

\[
g(t) = g(0) + t \frac{\partial g}{\partial t} (0) + \frac{1}{2} t^2 \frac{\partial^2 g}{\partial t^2} (v) \leq \frac{1}{2} t^2 \frac{1}{4} = t^2 / 8
\]

which proves our claim.

\[\Box\]

## 2 Exercises

a) Prove the following (look at the proofs of Corollaries 1.2 and 1.3, and use Hoeffding’s lemma 2.1):

**Corollary 2.1** (Hoeffding’s inequality). Assume that \( x_1 \ldots x_n \) are \( n \) independent random variables with support on \([0, 1]\) and mean \( \mu \). Fix \( \varepsilon > 0 \). We have that:

\[
P \left[ \left| \frac{1}{n} \sum_{i=1}^{n} x_i - \mu \right| \geq \varepsilon \right] \leq 2e^{-2n\varepsilon^2}
\]

b) Prove the following (look at the proofs of Corollaries 1.2 and 1.3):

**Corollary 2.2** (Hoeffding’s inequality). Assume that \( x_1 \ldots x_n \) are \( n \) independent random variables, where each \( x_i \in [a_i, b_i] \). Fix \( \varepsilon > 0 \). We have that:

\[
P \left[ \left| \frac{1}{n} \sum_{i=1}^{n} x_i - \mathbb{E} \left[ \frac{1}{n} \sum_{i=1}^{n} x_i \right] \right| \geq \varepsilon \right] \leq \frac{4e^{-2\varepsilon^2}}{\sum_{i=1}^{n} (b_i - a_i)^2}
\]

## 3 Application: Empirical Risk Minimization with a Finite Hypothesis Class

One of the main goals of machine learning is to minimize a risk with respect to a data distribution. Unfortunately, we never observe the data distribution directly, but a finite set of samples drawn from it. Assume an algorithm “learns” by minimizing an empirical risk, i.e., a risk that depends on a training set. Here we prove a generalization result of this learning procedure.
Theorem 2.1. Assume that $x \in X$ and $y \in Y$ where $X$ and $Y$ are arbitrary domains. Assume that the pair $(x, y)$ follows an arbitrary distribution $D$. Assume that $(x_1, y_1) \ldots (x_n, y_n)$ are $n$ i.i.d. samples drawn from the distribution $D$. Assume that $F$ is a finite set of functions, i.e., $F = \{f_1 \ldots f_k\}$ where $k < +\infty$ and $(\forall j) f_j : X \to Y$. The expected risk and its minimizer are defined as:

$$R(f) = \mathbb{E}_{(x, y) \sim D}[1[f(x) \neq y]]$$

$$\hat{f} = \arg\min_{f \in F} R(f)$$

The empirical risk and its minimizer are defined as:

$$\hat{R}(f) = \frac{1}{n} \sum_{i=1}^{n} 1[f(x_i) \neq y_i]$$

$$\hat{f} = \arg\min_{f \in F} \hat{R}(f)$$

Fix $\delta \in (0, 1)$. We have that:

$$\mathbb{P} \left[ |\hat{R}(f) - R(f)| < \sqrt{\frac{2(\log k + \log (2/\delta))}{n}} \right] \geq 1 - \delta$$

or equivalently, if $n \geq \frac{2(\log k + \log (2/\delta))}{\varepsilon^2}$ then:

$$\mathbb{P} \left[ |\hat{R}(f) - R(f)| < \varepsilon \right] \geq 1 - \delta$$

Proof. Fix a function $f \in F$. Define the random variable $z = 1[f(x) \neq y] \in [0, 1]$. Note that the expected and empirical risks are:

$$\overline{R}(f) = \mathbb{E}_{(x, y) \sim D}[z]$$

$$\hat{R}(f) = \frac{1}{n} \sum_{i=1}^{n} z_i$$

and moreover $\mathbb{E}[z_i] = \overline{R}(f)$, thus:

$$\mathbb{E}[\hat{R}(f)] = \mathbb{E}\left[\frac{1}{n} \sum_{i=1}^{n} z_i\right]$$

$$= \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}[z_i]$$

$$= \overline{R}(f)$$

By the Hoeffding’s inequality (Corollary 2.1) for a single hypothesis $f \in F$, we have:

$$\mathbb{P} \left[ |\hat{R}(f) - R(f)| \geq \varepsilon \right] \leq 2e^{-2n\varepsilon^2}$$
By applying the union bound for all $k$ functions in $\mathcal{F}$ and by Hoeffding’s inequality (Corollary 2.1), we have:

\[
P \left[ (\exists f \in \mathcal{F}) \left| \hat{R}(f) - \mathcal{R}(f) \right| \geq \varepsilon \right] = P \left[ \bigcup_{f \in \mathcal{F}} \left[ \left| \hat{R}(f) - \mathcal{R}(f) \right| \geq \varepsilon \right] \right] \\
\leq \sum_{f \in \mathcal{F}} P \left[ \left| \hat{R}(f) - \mathcal{R}(f) \right| \geq \varepsilon \right] \\
\leq 2ke^{-2n\varepsilon^2}
\]

Equivalently:

\[
P \left[ (\forall f \in \mathcal{F}) \left| \hat{R}(f) - \mathcal{R}(f) \right| < \varepsilon \right] = P \left[ \bigcap_{f \in \mathcal{F}} \left[ \left| \hat{R}(f) - \mathcal{R}(f) \right| < \varepsilon \right] \right] \\
= 1 - P \left[ (\exists f \in \mathcal{F}) \left| \hat{R}(f) - \mathcal{R}(f) \right| \geq \varepsilon \right] \\
\geq 1 - 2ke^{-2n\varepsilon^2} \tag{1}
\]

Let $\delta = 2ke^{-2n\varepsilon^2}$, then $\varepsilon = \sqrt{\frac{\log k + \log (2/\delta)}{2n}}$. Finally since $\hat{f}$ minimizes $\hat{R}$ we know that $\hat{R}(\hat{f}) \leq \mathcal{R}(\mathcal{F})$. From eq.(1) and the above, we have:

\[
\mathcal{R}(\hat{f}) - \mathcal{R}(\mathcal{F}) < \hat{R}(\hat{f}) + \varepsilon - \hat{R}(\mathcal{F}) + \varepsilon \\
\leq 2\varepsilon
\]

which proves our claim. $\square$

Expressions of the form of eq.(1) are called uniform convergence.

4 Exercises

a) Assume that $\mathcal{X} = \mathbb{R}^p$ for some number of features $p$. As in binary classification, assume that $\mathcal{Y} = \{-1, +1\}$. First, assume that $\mathcal{F}$ is the set of linear classifier functions of the form:

\[
f(x) = \begin{cases} 
+1 & \text{if } \langle w, x \rangle \geq 0 \\
-1 & \text{if } \langle w, x \rangle < 0
\end{cases}
\]

for some $w \in \{-1, 0, +1\}^p$. How many vectors $w$ are in the set $\{-1, 0, +1\}^p$? In other words, what is $k$ in Theorem 2.1? Now, assume that $\mathcal{F}$ is the set of linear classifier functions where $w \in \{-1, 0, +1\}^p$ and where $w$ has at most $s$ non-zero elements, for some fixed value $s$. What is $k$ in Theorem 2.1?

b) Assume that $\mathcal{A}$ is an event that depends on a random variable $x$. Fix $a$, $b$ and $\delta \in (0, 1)$. Assume that $P[\mathcal{A}(a)] \leq \delta$ and $P[\mathcal{A}(b)] \leq \delta$. Furthermore, assume that if not $\mathcal{A}(a)$ and not $\mathcal{A}(b)$ then $(\forall x \in [a, b])$ not $\mathcal{A}(x)$. Find $c$ in the expression $P[(\forall x \in [a, b])$ not $\mathcal{A}(x)] \geq c$. 

4