FTML practical session 13

13 juin 2025

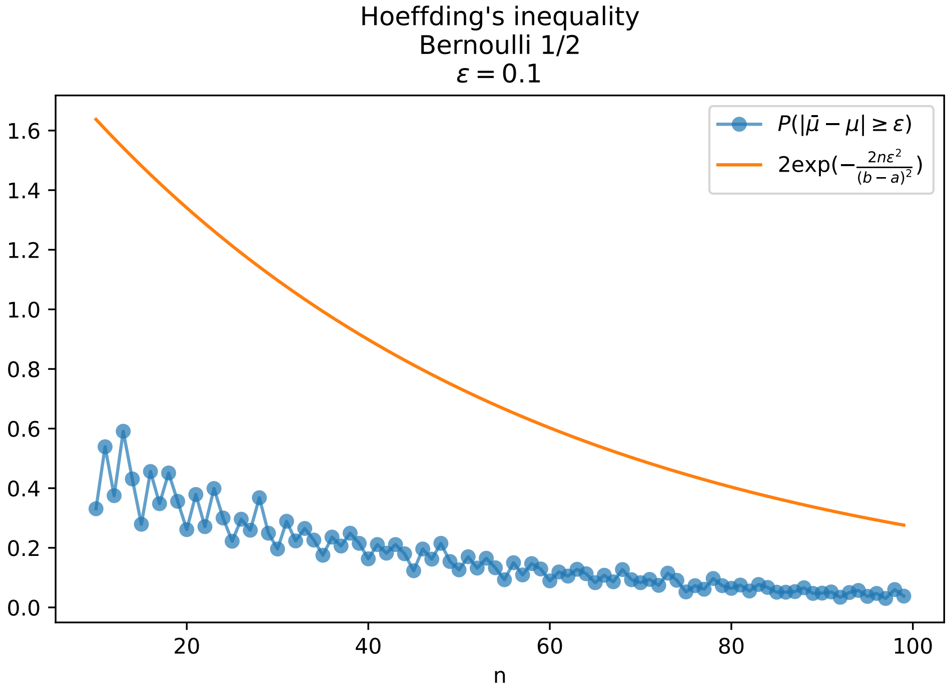

FIGURE 1 – Simulation of Hoeffding's inequality

TABLE DES MATIÈRES

1 HOEFFDING'S INEQUALITY

The following result will be useful in order to proove the bound on the estimation error in section 2

Theorème 1. *Hoeffding's inequality*
Let $(X_i)_{1\leqslant i\leqslant n}$ *be* n *i.i.d real random variables such that* $\forall i \in [1, n]$, $X_i \in [a, b]$ *and* $E(X_i) = \mu \in \mathbb{R}$. *Let* $\bar{\mu} = \frac{1}{n}\sum_{i=1}^{n} X_i$.
Then $\forall \epsilon > 0$,

$$P\Big(|\bar{\mu} - \mu| \geqslant \epsilon\Big) \leqslant 2\exp\left(-\frac{2n\epsilon^2}{(b-a)^2}\right) \tag{1}$$

Run a simulation that allows to visualize Hoeffding's inequality, with a random variable of your choice, like in figures 2 and 3, where a Bernoulli variable of parameter $p = 1/2$ is used.

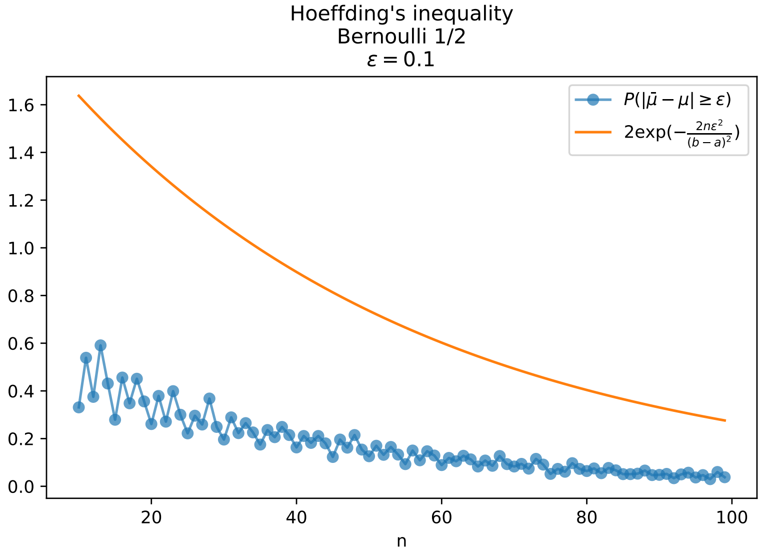

Figure 2 – Hoeffding's inequality with a Bernoulli variable of parameter $p = 1/2$ and $\epsilon = 0.1$

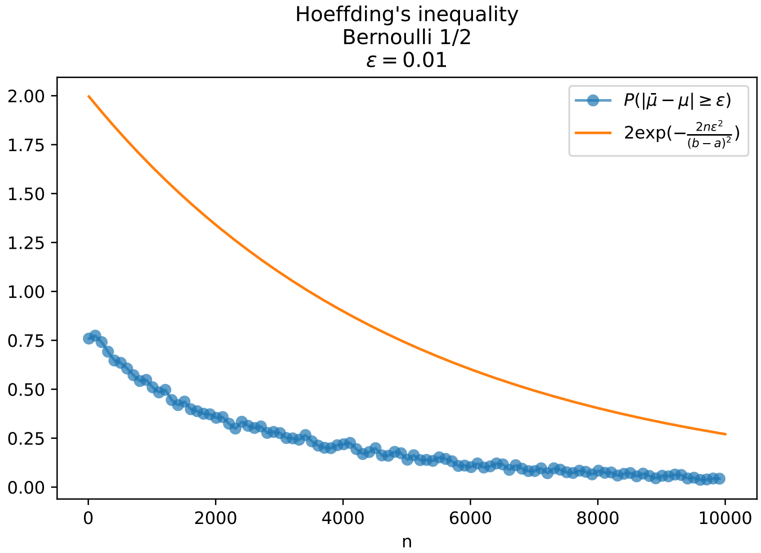

Figure 3 – Hoeffding's inequality with a Bernoulli variable of parameter $p = 1/2$ and $\epsilon = 0.01$

2 BOUND ON THE ESTIMATION ERROR

We consider a usual supervised learning setting, and the a space of functions F in which we choose our estimators. The dataset contains n samples, the empircial risk is noted R_n and the real risk R. If f_n is the empircial risk minimizer, and f_a the optimal estimator in F, we have seen this result during the lectures :

$$0 \leqslant R(f_n) - R(f_a) \leqslant 2 \sup_{h \in F} |R(h) - R_n(h)| \tag{2}$$

Our objective is to bound equation 2. We make the following additional hypotheses :

— F is finite, with $|F|$ elements.

— The loss l is uniformly bounded : $l(\hat{y}, y) \in [a, b]$ with a and b real numbers.

We will also use the following result :

Proposition 2. *Boole's inequality*
Let $A_1, A_2, \dots$, be acountable set of events of a probability space $\{\Omega, \mathcal{F}, P\}$.
Then,

$$P\Big(\cup_{i \geqslant 1} A_i \Big) \leqslant \sum_{i \geqslant 1} P(A_i) \tag{3}$$

Step 1] Using 2, we have that :

$$P\Big(R(f_n) - R(f_a) \geqslant t\Big) \leqslant P\Big(2 \sup_{h \in F} |R(h) - R_n(h)| \geqslant t\Big) \tag{4}$$

Step 2]
Show that

$$P\Big(2 \sup_{h \in F} |R(h) - R_n(h)| \geqslant t\Big) \leqslant \sum_{h \in F} P\big(2|R(h) - R_n(h)| \geqslant t\big) \tag{5}$$

Step 3]
Show that

$$P\Big(R(f_n) - R(f_a) \geqslant t\Big) \leqslant 2|F| \exp\Big(- \frac{nt^2}{2(b-a)^2} \Big) \tag{6}$$

Step 4]
We write

$$\delta = 2|F| \exp\Big(- \frac{nt^2}{2(b-a)^2} \Big) \tag{7}$$

Show that with probability larger than $1 - \delta$,

$$R(f_n) \leqslant R(f_a) + 2\sqrt{\frac{2(b-a)^2 \Big(\log(\frac{2}{\delta}) + \log(|F|) \Big)}{n}} \tag{8}$$

In which situations do we have for instance that $a = 0$ and $b = 1$?

Using the code of **tp_02_ols** or **tp_01/**, observe the $\frac{1}{\sqrt{n}}$ behavior of the test error as a function of the number of train samples.