A. Probability tools

1. Let \( f : (0, +\infty) \to \mathbb{R}_+ \) be a function admitting an inverse \( f^{-1} \) and let \( X \) be a random variable. Show that if for any \( t > 0 \), \( \Pr[X > t] \leq f(t) \), then, for any \( \delta > 0 \), with probability at least \( 1 - \delta \), \( X \leq f^{-1}(\delta) \).

   \textit{Solution:} For any \( \delta > 0 \), let \( t = f^{-1}(\delta) \). Plugging this in \( \Pr[X > t] \leq f(t) \) yields \( \Pr[X > f^{-1}(\delta)] \leq \delta \), that is \( \Pr[X \leq f^{-1}(\delta)] \geq 1 - \delta \). \( \square \)

2. Let \( X \) be a discrete random variable taking non-negative integer values. Show that \( \mathbb{E}[X] = \sum_{n \geq 1} \Pr[X \geq n] \) (\textit{hint:} rewrite \( \Pr[X = n] \) as \( \Pr[X \geq n] - \Pr[X \geq n + 1] \)).

   \textit{Solution:} We assume that \( X \) is a bounded random variable to avoid any convergence issues (although the statement is still true in the general case).

   By definition of expectation and using the hint, we can write
   \[
   \mathbb{E}[X] = \sum_{n \geq 0} n \Pr[X = n] = \sum_{n \geq 1} n(\Pr[X \geq n] - \Pr[X \geq n + 1]).
   \]

   Note that in this sum, for \( n \geq 1 \), \( \Pr[X \geq n] \) is added \( n \) times and subtracted \( n - 1 \) times, thus \( \mathbb{E}[X] = \sum_{n \geq 1} \Pr[X \geq n] \).

   More generally, by definition of the Lebesgue integral, for any non-negative random variable \( X \), the following identity holds:
   \[
   \mathbb{E}[X] = \int_0^{+\infty} \Pr[X \geq t] \, dt.
   \]

   \( \square \)

B. Label bias

1. Let \( D \) be a distribution over \( \mathcal{X} \) and let \( f : \mathcal{X} \to \{-1, +1\} \) be a labeling function. Suppose we wish to find a good approximation of the label bias of
the distribution $D$, that is of $p_+$ defined by:

$$p_+ = \Pr_{x \sim D}[f(x) = +1].$$

Let $S$ be a finite labeled sample of size $m$ drawn i.i.d. according to $D$. Use $S$ to derive an estimate $\hat{p}_+$ of $p_+$. Show that for any $\delta > 0$, with probability at least $1 - \delta$, $|p_+ - \hat{p}_+| \leq \sqrt{\frac{\log(2/\delta)}{2m}}$ (carefully justify all steps).

**Solution:** Let $\hat{p}_+$ be the fraction of positively labeled points in $S = (x_1, \ldots, x_m)$:

$$\hat{p}_+ = \frac{1}{m} \sum_{i=1}^{m} 1_{f(x_i) = +1}$$

Since the points are drawn i.i.d.,

$$E[\hat{p}_+] = \frac{1}{m} \sum_{i=1}^{m} E_{S \sim D^m}[1_{f(x) = +1}] = \frac{1}{m} E_{S \sim D^m}[1_{f(x) = +1}] = \frac{1}{m} E_{x \sim D}[1_{f(x) = +1}] = p_+.$$ 

Thus, by Hoeffding’s inequality, for any $\epsilon > 0$,

$$\Pr[|p_+ - \hat{p}_+| > \epsilon] \leq 2e^{-2m\epsilon^2}.$$ 

Setting $\delta$ to match the right-hand side yields the result. \(\square\)

C. Learning in the presence of noise

1. In Lecture 2, we showed that the concept class of axis-aligned rectangles is PAC-learnable. Consider now the case where the training points received by the learner are subject to the following noise: points negatively labeled are unaffected by noise but the label of a positive training point is randomly flipped to negative with probability $\eta \in (0, \frac{1}{2})$. The exact value of the noise rate $\eta$ is not known to the learner but an upper bound $\eta' < 1/2$. Show that the algorithm described in class returning the tightest rectangle containing positive points can still PAC-learn axis-aligned rectangles in the presence of this noise. To do so, you can proceed using the following steps:

   (a) Using the notation of the lecture slides, assume that $\Pr[R] > \epsilon$. Suppose that $\text{error}(R') > \epsilon$. Give an upper bound on the probability that $R'$ misses a region $r_j$, $j \in [1, 4]$ in terms of $\epsilon$ and $\eta'$.
Solution: The probability that $R'$ misses region $r_j$ is the product of the probability $p$ for each point $x_i$ of the training sample to either not fall in $r_j$ or be positive and fall in $r_j$ with the label flipped to negative due to noise.

\[
p = \Pr[x \not\in r_j \lor (x \in r_j \land x \text{ positive } \land \text{ label of } x \text{ flipped})]
= \Pr[x \not\in r_j] + \Pr[(x \in r_j \land \text{ label of } x \text{ flipped})]
= (1 - \Pr[x \in r_j]) + \eta \Pr[x \in r_j]
= (1 - \eta)(1 - \Pr[x \not\in r_j]) + \eta
\leq (1 - \eta)(1 - \epsilon/4) + \eta
= (1 - \epsilon/4) + \eta \epsilon/4 \leq 1 - \epsilon(1 - \eta')/4.
\]

(b) Use that to give an upper bound on $\Pr[\text{error}(R') > \epsilon]$ in terms of $\epsilon$ and $\eta'$ and conclude by giving a sample complexity bound.

Solution: The probability that $\Pr[\text{error}(R') > \epsilon]$ is upper bounded by the probability that $R'$ misses at least one region $r_j$. Thus, by the union bound,

\[
\Pr[\text{error}(R') > \epsilon] \leq 4 \left(1 - \epsilon(1 - \eta')/4\right)^m \leq 4e^{-m\epsilon(1-\eta')/4}.
\]

Setting $\delta$ to match the upper bound leads to the following: with probability at least $1 - \delta$, for $m \geq \frac{4}{(1-\eta')^2} \log \frac{4}{\delta}$, $\text{error}(R') \leq \epsilon$.

2. [Bonus question] In this section, we will seek a more general result. We consider a finite hypothesis set $H$, assume that the target concept is in $H$, and adopt the following noise model: the label of a training point received by the learner is randomly changed with probability $\eta \in (0, \frac{1}{2})$. The exact value of the noise rate $\eta$ is not known to the learner but an upper bound $\eta'$ is supplied to him with $\eta \leq \eta' < 1/2$.

(a) For any $h \in H$, let $d(h)$ denote the probability that the label of a training point received by the learner disagrees with the one given by $h$. Let $h^*$ be the target hypothesis, show that $d(h^*) = \eta$.

Solution: The probability that the label of a point be incorrect is $\eta$. A label is incorrect iff it differs from the label given by the target $h^*$. □
(b) More generally, show that for any $h \in H$, $d(h) = \eta + (1 - 2\eta) \text{error}(h)$, where $\text{error}(h)$ denotes the generalization error of $h$.

**Solution:** The label of a point disagrees with the one given by $h$ either because its label is correct (probability $1 - \eta$) and $h$ misclassifies that point (probability $\text{error}(h)$), or because its label is incorrect (probability $\eta$) and $h$ classifies it correctly (probability $1 - \text{error}(h)$). Since these two events are disjoint, the probability of their union is the sum of the probability and

\[
d(h) = (1 - \eta)\text{error}(h) + \eta(1 - \text{error}(h)) \]

\[
= \eta + (1 - 2\eta) \text{error}(h). \]

\[\square\]

(c) Fix $\epsilon > 0$ for this and all the following questions. Use the previous questions to show that if $\text{error}(h) > \epsilon$, then $d(h) - d(h^*) \geq \epsilon'$, where $\epsilon' = \epsilon(1 - 2\eta')$.

**Solution:** In view of the previous question, if $\text{error}(h) > \epsilon$,

\[
d(h) = \eta + (1 - 2\eta) \text{error}(h) \]

\[
\geq \eta + (1 - 2\eta)\epsilon \]

\[
\geq \eta + (1 - 2\eta')\epsilon \]

\[
= d(h^*) + (1 - 2\eta')\epsilon, \]

where we used $d(h^*) = \eta$. \[\square\]

(d) For any hypothesis $h \in H$ and sample $S$ of size $m$, let $\hat{d}(h)$ denote the fraction of the points in $S$ whose labels disagree with those given by $h$. We will consider the algorithm $L$ which, after receiving $S$, returns the hypothesis $h_S$ with the smallest number of disagreements (thus $\hat{d}(h_S)$ is minimal). To show PAC-learning for $L$, we will show that for any $h$, if $\text{error}(h) > \epsilon$, then with high probability $\hat{d}(h) \geq \hat{d}(h^*)$. First, show that for any $\delta > 0$, with probability at least $1 - \delta/2$, for $m \geq \frac{2}{\epsilon^2} \log \frac{2}{\delta}$, the following holds:

\[
\hat{d}(h^*) - d(h^*) \leq \epsilon' / 2 \]

**Solution:** By Hoeffding’s inequality $\Pr[\hat{d}(h^*) - d(h^*) > \epsilon' / 2] \leq e^{-m\epsilon'^2 / 2}$. Setting $\delta / 2$ to match the right-hand side yields the result. \[\square\]
(e) Second, show that for any $\delta > 0$, with probability at least $1 - \delta/2$, for $m \geq \frac{2}{\epsilon^2} (\log |H| + \log \frac{2}{\delta})$, the following holds for all $h \in H$:

$$d(h) - \hat{d}(h) \leq \epsilon'/2$$

Solution: By the union bound and Hoeffding’s inequality $\Pr[\exists h: d(h) - \hat{d}(h) > \epsilon'/2] \leq \frac{|H|e^{-m\epsilon'^2/2}}{\delta}$. Setting $\delta/2$ to match the right-hand side yields the result. $\square$

(f) Finally, show that for any $\delta > 0$, with probability at least $1 - \delta$, for $m \geq \frac{2}{\epsilon'^2} (\log |H| + \log \frac{2}{\delta})$, the following holds for all $h \in H$ with $\text{error}(h) > \epsilon$:

$$\hat{d}(h) - \hat{d}(h^*) \geq 0.$$ 

(hint: use $\hat{d}(h) - \hat{d}(h^*) = [\hat{d}(h) - d(h)] + [d(h) - d(h^*)] + [d(h^*) - \hat{d}(h^*)]$ and use previous questions to lower bound each of these three terms).

Solution: By the union bound, for any $\delta > 0$, with probability at least $1 - \delta$, for $m \geq \frac{2}{\epsilon'^2} (\log |H| + \log \frac{2}{\delta})$, both inequalities of the previous two questions hold, the previous one for all $h \in H$. Thus, using the equality of the hint, with probability at least $1 - \delta$, for $m \geq \frac{2}{\epsilon'^2} (\log |H| + \log \frac{2}{\delta})$, the following holds for all $h \in H$ with $\text{error}(h) > \epsilon$:

$$\hat{d}(h) - \hat{d}(h^*) = [\hat{d}(h) - d(h)] + [d(h) - d(h^*)] + [d(h^*) - \hat{d}(h^*)] \geq -\epsilon'/2 + \epsilon' - \epsilon'/2 = 0,$$

and thus such hypotheses $h$ are not selected by $L$ since they do not admit a minimal $\hat{d}(h)$.

This shows that algorithm $L$ can be used for PAC-learning in the presence of the noise described and in the consistent case where the target concept is in $H$. Nevertheless, the computational complexity of $L$ is in general not polynomial. In general, the problem of finding the hypothesis with minimal $\hat{d}(h)$ is NP-complete. $\square$