Homework 3
Due: Solutions

Problem 1

a. Let \( f \) be a function \( f : \{0, 1\}^n \to \{0, 1\} \). The error of this function is:
\[
\Pr(f(X_1, \ldots, X_n) \neq Y) = \sum_{\vec{x} \in \{0, 1\}^n} \Pr(f(\vec{x}) \neq Y | \vec{X} = \vec{x}) \Pr(\vec{X} = \vec{x}) = \sum_{\vec{x} \in \{0, 1\}^n} \Pr(\vec{X} = \vec{x} | f(\vec{x}) \neq Y) \Pr(f(\vec{x}) \neq Y)
\]
Note that since the function \( f \) always returns the same value given \( \vec{x} \), we have that \( \Pr(f(\vec{x}) \neq Y) = \frac{1}{2} \). Also, \( Y \neq f(\vec{x}) \) if and only if \( Y = 1 - f(\vec{x}) \). Therefore, we have that:
\[
\Pr(f(X_1, \ldots, X_n) \neq Y) = \frac{1}{2} \sum_{\vec{x} \in \{0, 1\}^n} \Pr(\vec{X} = \vec{x} | Y = 1 - f(\vec{x}))
= \frac{1}{2} \sum_{\vec{x} \in \{0, 1\}^n} g(\vec{x}; 1 - f(\vec{x}))
\]
Observe that the error is minimized if and only if for any \( \vec{x} \), we choose \( f(\vec{x}) = \arg \max_{y \in \{0, 1\}} g(\vec{x}, y) \).

b. We have that \( f(\vec{x}, \vec{w}^*) \) is equal to 1 if and only if
\[
f(\vec{x}, \vec{w}^*) = 1 \iff \sum_{i=1}^n X_i \ln \left( \frac{p_i}{1-p_i} \right) \geq \sum_{i=1}^n (1-X_i) \ln \left( \frac{p_i}{1-p_i} \right)
\]
\[
\iff \prod_{i: X_i = 1} \frac{p_i}{1-p_i} \geq \prod_{i: X_i = 0} \frac{p_i}{1-p_i}
\]
\[
\iff \prod_{i: X_i = 1} p_i \prod_{i: X_i = 0} (1-p_i) \geq \prod_{i: X_i = 0} p_i \prod_{i: X_i = 1} (1-p_i)
\]
\[
\iff g(\vec{x}, 1) \geq g(\vec{x}, 0)
\]
The first equivalence is due to the definition of \( f \), and the second equivalence is obtained by applying the exponential function to both sides of the inequality. In the last equivalence, we used the definition of \( g \). Optimality follows by point a.
c. For \( i = 1, \ldots, n \), let \( Z_i \) be a binary random variable that denotes the event \( \{ X_i = Y \} \), i.e. \( Z_i = 1 \) if and only if \( X_i = Y \), else \( Z_i = 0 \). We have that \( Z_i \) is distributed as a Bernoulli of parameter \( p_i \), i.e. \( Z_i \sim B(p_i) \), and for \( j \neq i \), \( Z_j \) is independent with \( Z_i \) by problem assumptions.

We have that \( f^* \) makes an error if and only if \( \sum_{i=1}^{n} Z_i w_i \leq W/2 \), where \( W = \sum_{i=1}^{n} w_i \) (the sum of the weights of the correct votes is less than the sum of the incorrect votes; in case of a tie we make the worst case assumption that we are wrong).

Let \( \tilde{Z}_i = Z_i w_i \). Observe that \( \tilde{Z}_i = w_i \) with probability \( p_i \), and \( \tilde{Z}_i = 0 \) with probability \( 1 - p_i \). It is clear that \( \tilde{Z}_i \in [0, w_i] \) if \( p_i > 1/2 \), and \( \tilde{Z}_i \in [-w_i, 0] \) if \( p_i < 1/2 \), and \( E\tilde{Z}_i = p_i w_i \).

We have that:

\[
Pr(f^*(X_1, \ldots, X_n) \neq Y) \leq Pr \left( \sum_{i=1}^{n} \tilde{Z}_i \leq W/2 \right)
= Pr \left( \sum_{i=1}^{n} \tilde{Z}_i - \sum_{i=1}^{n} p_i w_i \leq \sum_{i=1}^{n} w_i \left( p_i - \frac{1}{2} \right) \right)
\]

Note that \( w_i (p_i - 1/2) \leq 0 \) for each \( i = 1, \ldots, n \). Hence, we can apply Hoeffding’s bound and obtain that:

\[
Pr(f^*(X_1, \ldots, X_n) \neq Y) \leq \exp \left( -\frac{2 \left[ \sum_{i=1}^{n} w_i (p_i - 1/2) \right]^2}{\sum_{i=1}^{n} w_i^2} \right)
\]

d. If \( p_i = 0.9 \) for \( i = 1, \ldots, n \), we would have that:

\[
Pr(f^*(X_1, \ldots, X_n) \neq Y) \leq \exp \left( -\frac{2 \left[ \sum_{i=1}^{n} w_i (p_i - 1/2) \right]^2}{\sum_{i=1}^{n} w_i^2} \right) \leq \exp (-2n(0.9 - 0.5))
\]

and we can see that this bound is small for large \( n \) and goes to 0 as \( n \to \infty \).

If we have that \( p_1 \to 1 \), then \( w_1 \to \infty \), and by taking the limit we have that:

\[
Pr(f^*(X_1, \ldots, X_n) \neq Y) \leq \exp(-1/2) \approx 0.6
\]

That is, the bound becomes vacuous. This point shows a limitation of the applicability of Hoeffding’s bound in the case of very heterogenous sums.
Problem 2

a. Suppose that we have $m$ samples for each segment $i$, with $i = 1, \ldots, 5$. Let $\tilde{T}_i$ be the average of those samples for segment $i$. For a given $i$, by using Hoeffding’s bound, we know that:

$$\Pr(\left| \mathbb{E}T_i - \tilde{T}_i \right| > 1) \leq 2 \exp(-2m/100)$$

where we used the fact that $T_i \in [5,15]$. By union bound, we have that the probability that there exists $i$ such that $|\mathbb{E}T_i - \tilde{T}_i| > 1$ is upper bounded by

$$5 \cdot 2 \exp(-2m/100) \leq \delta .$$

By solving the above inequality with respect to $m$, we have that:

$$m \geq 50 \ln \frac{10}{\delta}$$

Therefore, if we have $m \geq 50 \ln \frac{10}{\delta}$ samples for each segment, then with probability at least $1 - \delta$ we have that for any $i = 1, \ldots, 5$, it holds that $|\mathbb{E}T_i - \tilde{T}_i| < 1$. Let $\tilde{T} = \sum_{i=1}^{5} \tilde{T}_i$. Observe that the statement above implies that with probability at least $1 - \delta$, it holds that

$$\left| \mathbb{E}T - \tilde{T} \right| = \left| \sum_{i=1}^{5} \mathbb{E}T_i - \tilde{T}_i \right| \leq \sum_{i=1}^{5} |\mathbb{E}T_i - \tilde{T}_i| \leq 5$$

The first equality is obtained by using linearity of expectation, the second inequality is due to triangle inequality.

b. Let $T = \sum_{i=1}^{5} T_i$. Suppose that we have $m$ samples of the time $T$. We know that $T \in [25,75]$. Let $\tilde{T}$ be the average of the $m$ samples. By using Hoeffding’s inequality, we have that:

$$\Pr(\left| \mathbb{E}T - \tilde{T} \right| > 5) \leq 2 \exp\left( -2 \cdot \frac{25m}{50^2} \right) = 2 \exp\left( -\frac{m}{50} \right)$$

By constraining

$$2 \exp(-m/50) \leq \delta \iff m \geq 50 \ln \frac{2}{\delta}$$

we have that if we compute the mean of $m \geq 50 \ln(2/\delta)$ samples, then with probability at least $1 - \delta$ this value is within 5 minute of the expectation of $T$. 

3
Problem 3

(Additive Bound). Let \( \tilde{\rho} \) be the fraction of edges found out of \( s \) random queries. It holds that \( \mathbb{E}\tilde{\rho} = \rho \). By Hoeffding’s bound, we have that:

\[
\Pr (|\tilde{\rho} - \rho| > \epsilon) \leq 2 \exp \left( -2s\epsilon^2 \right) \leq \delta
\]

Therefore, by taking \( s \geq \frac{1}{2\epsilon^2} \ln \frac{2}{\delta} \) samples, we have that with probability at least \( 1 - \delta \), it holds that \( |\tilde{\rho} - \rho| \leq \epsilon \).

(Multiplicative Bound). Using the same strategy above, we have that:

\[
\Pr (|\tilde{\rho} - \rho| > \rho\epsilon) \leq 2 \exp \left( -2s\epsilon^2 \rho^2 \right) \leq 2 \exp \left( -2s\epsilon^2 d^2 \right)
\]

In the second inequality, we use the fact that \( \rho \geq d \). Again, we set

\[
2 \exp \left( -2s\epsilon^2 d^2 \right) \leq \delta.
\]

Therefore, by taking \( s \geq \frac{1}{2\epsilon^2 d^2} \ln \frac{2}{\delta} \) samples, with probability at least \( 1 - \delta \), we have that \( |\tilde{\rho} - \rho| \leq \rho\epsilon \).

Observe that if \( m = O(n) \), then \( \rho = o(1/n) \), and we need at least \( \Omega\left( \frac{n^2}{\epsilon^2} \ln \frac{1}{\delta} \right) \) samples to obtain an estimate \( |\tilde{\rho} - \rho| \leq \rho\epsilon \) with probability at least \( 1 - \delta \). In comparison, we only need \( \binom{n}{2} \) queries to compute exactly the value \( \rho \).

Therefore, we observe that a sampling strategy is not effective to obtain a multiplicative bound if \( m = O(n) \), i.e. the graph is sparse. A sampling strategy is asymptotically convenient only if \( m = \Omega(n^{1+c}) \) for \( c > 0 \)

Assume now that we do not know a lower bound \( d \) to \( \rho \). We design an algorithm that is able to compute multiplicative bound by iteratively guessing the value of this lower bound.

Let \( d_0 = \frac{1}{4} \), and let \( d_i = \frac{1}{2}d_{i-1} \) for \( i = 1, 2, \ldots \).

At iteration \( i \), we do \( s_i \) random queries and compute the fraction of edges found \( \tilde{\rho}_i \). Fixed \( \epsilon \) and \( \delta_i \) (see next equation), the number of samples \( s_i \) is chosen to satisfy:

\[
\Pr (|\tilde{\rho}_i - \rho| \geq \epsilon d_i) \leq 2 \exp(-2s_i\epsilon^2 d_i^2) \leq \delta_i \tag{1}
\]

Suppose that the event \( |\tilde{\rho}_i - \rho| \leq \epsilon d_i \) is true for every iteration (we will further discuss this later). There are two situations: (i) \( \tilde{\rho}_i - \epsilon d_i \geq d_i \) and (ii) \( \tilde{\rho}_i - \epsilon d_i \leq d_i \).
In case (i), we have that \( \rho \geq \tilde{\rho}_i - \epsilon d_i \geq d_i \) (first inequality due to event, second inequality due to the fact that we are in case (i)), therefore we are guaranteed that our lower bound is correct and we can return \( \tilde{\rho}_i \). In case (ii), we cannot have this guarantee, therefore we iterate again.

Observe that if \( d_i \leq \rho/(1 + 2\epsilon) \), then we have the guarantee that situation (i) will occur (if event \( |\tilde{\rho}_i - \rho| \leq \epsilon d_i \) is true). In fact, we have that:

\[
|\tilde{\rho}_i - \rho| \leq \epsilon d_i \implies \tilde{\rho}_i - \epsilon d_i \geq \rho - 2\epsilon d_i \geq d_i
\]

In the last inequality, we used the fact that \( d_i \leq \rho/(1 + 2\epsilon) \). Therefore, if \( d_i \) is small enough, and \( |\tilde{\rho}_i - \rho| \leq \epsilon d_i \) holds, then we have that situation (i) will occur and the algorithm terminates.

Therefore, the total number of iteration is

\[
\leq -\log_2(\rho) + \log(1 + 2\epsilon) = O\left(\log_2\left(\frac{1}{\rho}\right)\right).
\]

Observe that after \( \log_2 n \) iteration, we can say that \( \rho \in O(1/n) \) (as the algorithm did not terminate earlier) and a sampling strategy is not more effective (see discussion above). Therefore, we run at maximum \( \log_2 n \) iterations of the algorithm, and if the algorithm did not terminate, we query all the edges in the graph.

This algorithm always returns a correct estimate if for \( i = 1, \ldots, \log_2 n \), the events \( |\tilde{\rho}_i - \rho| \leq \epsilon d_i \) hold (we are doing a worst case analysis, if one of this single event is not true we say that the algorithm fails).

We set \( \delta_i = \delta/\log_2 n \). By union bound, the algorithm returns a correct estimate with probability \( \geq 1 - \delta \). The number of random queries required at iteration \( i \) are (solving (1))

\[
s_i \geq \frac{1}{2\epsilon^2 d_i^2} \ln\left(\frac{\log_2 n}{\delta}\right).
\]