1 Probabilistic inequalities

In this question you will be asked to derive the three most used probabilistic inequalities for a specific random variable. Let \( x_1, \ldots, x_n \) be independent \( \{-1, 1\} \) valued random variables. Each \( x_i \) takes the value 1 with probability \( 1/2 \) and -1 else. Let \( X = \sum_{i=1}^{n} x_i \).

1. Let the random variable \( Y \) be defined as \( Y = |X| \). Prove that Markov’s inequality holds for \( Y \). Hint: note that \( Y \) takes integer values. Also, there is no need to compute \( \Pr[Y = i] \).

2. Prove Chebyshev’s inequality for the above random variable \( X \). You can use the fact that Markov’s inequality holds for any positive variable regardless of your success (or lack of if) in the previous question. Hint: \( \text{Var}[X] = E[(X - E[X])^2] \).

3. Argue that
\[
\Pr[X > a] = \Pr[\prod_{i=1}^{n} e^{\lambda x_i} > e^{\lambda a}] \leq \frac{E[\prod_{i=1}^{n} e^{\lambda x_i}]}{e^{\lambda a}}
\]
for any \( \lambda \in [0, 1] \). Explain each transition.

4. Argue that:
\[
\frac{E[\prod_{i=1}^{n} e^{\lambda x_i}]}{e^{\lambda a}} = \frac{\prod_{i=1}^{n} E[e^{\lambda x_i}]}{e^{\lambda a}} = \frac{(E[e^{\lambda x_1}])^n}{e^{\lambda a}}
\]
What property of the random variables \( x_i \) did we use in each transition?

5. Conclude that \( \Pr[X > a] \leq e^{-\frac{a^2}{2n}} \) by showing that:
\[
\exists \lambda \in [0, 1] \text{ s.t. } \frac{(E[e^{\lambda x_1}])^n}{e^{\lambda a}} \leq e^{-\frac{a^2}{2n}}
\]
Hint: For the hyperbolic cosine function we have \( \cosh(x) = \frac{1}{2}(e^x + e^{-x}) \leq e^{x^2/2} \) for \( x \in [0, 1] \ \lambda \in [0, 1] \).
The number of unique elements in an array

setup

In this question we will approximate the number of unique elements in a array $L$ of known length $n$ into which we have random access. The array contains $m$ unique elements $a_1, \ldots, a_m$ each of which appears $n(a_i)$ times, i.e., $\sum_{i=1}^{m} n(a_i) = n$. We define the following sampling procedure:

1. Pick $j$ uniformly at random from $[1, \ldots, n]$
2. $x \leftarrow L[j]$
3. return $x$

questions

1. Define $p(x)$ as the probability that the sampling procedure above returns element $x$. Compute $p(x)$ as a function of $n$ and $n(x)$
2. Let $f(x) = \frac{n}{n(x)}$. Compute:
   $$E_{x \sim \text{smp}}[f(x)]$$
   where $x \sim \text{smp}$ denoted that $x$ is chosen according the sampling procedure above.
3. A list is said to be $k$-frequency-bounded if no item in it appears more than $k$ times. In Other words, $\max_{i \in [1,\ldots,m]} n(a_i) \leq k$. Show that for a $k$-frequency-bounded list $L$ we have that:
   $$\text{Var}_{x \sim \text{smp}}[f(x)] \leq km^2$$
4. Let $Y = \frac{1}{s} \sum_{\ell=1}^{s} f(x_{\ell})$ where $x_{\ell}$ are chosen independently from the list according to the sampling procedure. Compute $E[Y]$ and show that $\text{Var}[Y] \leq km^2/s$.
5. Use Chebyshev’s inequality to find a value for $s$ such that for any $k$-frequency-bounded list and any two constants $\varepsilon \in [0,1]$ and $\delta \in [0,1]$: 
   $$\Pr[|Y - m| > \varepsilon m] < \delta.$$ 
   $s$ should be a function of $k$, $\varepsilon$ and $\delta$. 


2 Approximate pie-charts

setup

A list $A$ of length $n$ contains $m$ distinct items. Each of which appears $n_i$ times, i.e., $\sum_{i=1}^{m} n_i = n$. We define the frequency $f_i$ of item $i$ as $n_i/n$. A circle divided into sections relative to $f_i$ is called a pie-chart and we would like to produce one. Alas, the list $A$ is very long and we would rather perform $o(n)$ operations to produce it. Our strategy is to sample $s$ items from the list uniformly at random with replacement and output the histogram of $s$. More formally, let $s_i$ denote the number of times item $i$ appeared in the sample and $g_i = s_i/s$. We would want to have that for each item:

$$f_i - \tau \leq g_i \leq f_i + \tau.$$ 

The value of $\tau$ is the prescribed precision, for example, 1%. Note that it is an additive error and not a multiplicative one.

questions

1. Compute $E[g_i]$.

2. Bound from above the probability of a large deviation. In other words, bound $\Pr[|g_i - f_i| > \tau]$.

3. Find a value for $s$ such that with probability at least $1 - \delta$ for all $i$ we have $|g_i - f_i| \leq \tau$.

4. Bonus question: show that the condition of 3 hold also for:

$$s \geq \frac{4 \log(2m/\delta)}{m \tau^2}.$$
3 Bloom-like filter

setup

This question will deal with a data structure for holding a set of objects in a space efficient manner such that membership queries can be performed quickly and reliably. For lack of a better name we will call this data-structure a bloom-like filter. Bloom-like filters consist of \( k \) bit arrays \( B_1, \ldots, B_k \) each of length \( n \) (all bits initially set to \( \text{False} \)). They are also associated with \( k \) hash functions \( h_1, \ldots, h_k \). Each hash function \( h_i : x \rightarrow [1, \ldots, n] \) is chosen independently at random from a family \( H \) such that for any object, \( x \), in the universe \( \Pr_{h \sim H}[h(x) = i] = 1/n \). We define the following two operations on bloom-like filters.

1. \( \text{insert}(x) \)
2. for \( i \) in \([1, \ldots, k]\)
3. \( B_i[h_i(x)] = \text{True} \)

1. \( \text{query}(x) \)
2. for \( i \) in \([1, \ldots, k]\)
3. if \( B_i[h_i(x)] == \text{False} \)
4. return \( \text{False} \)
5. return \( \text{True} \)

questions

1. Argue that for any element \( x \) which was inserted into the bloom-like filter (\( \text{insert}(x) \) was performed) the output of \( \text{query}(x) \) is \( \text{True} \).

2. Assume we have inserted exactly \( n \) different items into the bloom-like filter. What is the probability that \( \text{query}(x^{\text{new}}) \) return \( \text{True} \) for \( x^{\text{new}} \) which was not inserted. Provide a bound for this probability which does not depend on \( n \) (you can assume \( n \) is larger than 2)

3. We now query the bloom-like filter with \( m \) different new objects \( x_1^{\text{new}}, \ldots, x_m^{\text{new}} \). Provide a value for \( k \) such that \( \text{query}(x_i^{\text{new}}) \) returns \( \text{False} \) for all the \( m \) new objects with probability at least \( 1 - \delta \). Note, the randomness is only the choice of the hash functions.
4 Useful facts

1. For any vector \( x \in \mathbb{R}^d \) we define the \( p \)-norm of \( x \) as follows:

\[
||x||_p = \left( \sum_{i=1}^{d} (x(i))^p \right)^{1/p}
\]

2. **Markov’s inequality:** For any non-negative random variable \( X \):

\[
\Pr[X > t] \leq \frac{E[X]}{t}.
\]

3. **Chebyshev’s inequality:** For any random variable \( X \):

\[
\Pr[|X - E[X]| > t] \leq \frac{\text{Var}[X]}{t^2}.
\]

4. **Chernoff’s inequality:** Let \( x_1, \ldots, x_n \) be independent \{0, 1\} valued random variables. Each \( x_i \) takes the value 1 with probability \( p_i \) and 0 else. Let \( X = \sum_{i=1}^{n} x_i \) and let \( \mu = E[X] = \sum_{i=1}^{n} p_i \). Then:

\[
\Pr[X > (1 + \epsilon)\mu] \leq e^{-\mu \epsilon^2 / 4}
\]
\[
\Pr[X < (1 - \epsilon)\mu] \leq e^{-\mu \epsilon^2 / 2}
\]

Or in a another convenient form:

\[
\Pr[|X - \mu| > \epsilon\mu] \leq 2e^{-\mu \epsilon^2 / 4}.
\]

5. **Hoeffding’s inequality:** Let \( x_1, \ldots, x_n \) be independent random variables taking values in \{+1, -1\} each with probability 1/2, then:

\[
\Pr[\left| \sum_{i=1}^{n} x_i a_i \right| > t] \leq 2e^{-\frac{t^2}{2\sum a_i^2}}.
\]

6. For any \( x \geq 2 \) we have:

\[
e^{-1} \geq (1 - \frac{1}{x})^{-x} \geq \frac{2}{3} e^{-1}
\]

7. For convenience:

\[
\frac{3}{5} \leq 1 - e^{-1} \approx 0.632 \leq \frac{2}{3} \quad \text{and} \quad \frac{3}{4} \leq 1 - \frac{2}{3} e^{-1} \approx 0.754 \leq \frac{4}{5}
\]