General Info

1. Solve 3 out of 4 questions.
2. Each correct answer is worth $3.3$ points.
3. If you have solved more than three questions, please indicate which three you would like to be checked.
4. The exam’s duration is 3 hours. If you need more time please ask the attending professor.
5. Good luck!

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 $X$:
   $$\Pr[X > t] \leq E[X]/t.$$

3. **Chebyshev’s inequality**: For any random variable $X$:
   $$\Pr[|X - E[X]| > t] \leq \text{Var}[X]/t^2.$$

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:
   $$\Pr[X > (1 + \varepsilon)\mu] \leq e^{-\mu \varepsilon^2/4}$$
   $$\Pr[X < (1 - \varepsilon)\mu] \leq e^{-\mu \varepsilon^2/2}$$
   Or in a another convenient form:
   $$\Pr[|X - \mu| > \varepsilon\mu] \leq 2e^{-\mu \varepsilon^2/4}$$

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

6. For any $x \geq 2$ we have:
   $$e^{-1} \geq \left(1 - \frac{1}{x}\right)x = \frac{2}{3} e^{-1}$$

7. For convenience:
   $$\frac{3}{5} \leq 1 - e^{-1} \approx 0.632 \leq \frac{2}{3} \quad \text{and} \quad \frac{3}{4} \leq 1 - \frac{2}{3} e^{-1} \approx 0.754 \leq \frac{4}{5}$$
1 Probabilistic inequalities

setup

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 $X = \sum^n_{i=1} x_i$.

questions

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

$$
\Pr[X > a] = \Pr[\prod^n_{i=1} e^{\lambda x_i} > e^{\lambda a}] \leq \frac{E[\prod^n_{i=1} e^{\lambda x_i}]}{e^{\lambda a}}
$$

for any $\lambda \in [0, 1]$. Explain each transition.

4. Argue that:

$$
\frac{E[\prod^n_{i=1} e^{\lambda x_i}]}{e^{\lambda a}} = \prod^n_{i=1} \frac{E[e^{\lambda x_i}]}{e^{\lambda a}} = \frac{(E[e^{\lambda x_1}])^n}{e^{\lambda a}}
$$

What properties of the random variables $x_i$ did you 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 $x \in [0, 1]$. 

answers

1. 

\[ E[Y] = \sum_{i=0}^{n} Pr[Y = i] \cdot i \]

\[ = \sum_{i=0}^{t} Pr[Y = i] \cdot i + \sum_{i=t+1}^{n} Pr[Y = i] \cdot i \]

\[ \geq \sum_{i=t+1}^{n} Pr[Y = i] \cdot i \]

\[ \geq \sum_{i=t+1}^{n} Pr[Y = i] \cdot t \]

\[ = t \cdot Pr[Y > t] \]

Therefore, \( E[Y] \geq t \cdot Pr[Y > t] \) which is Markov’s inequality.

2. This is identical to the general proof of Chebyshev’s inequality. We define \( Z = (X - E[X])^2 \). Since \( Z \) is positive we can use Markov’s inequality for it and get:

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

Here we used that \( E[Z] = E[(X - E[X])^2] = Var[X] \).

3. First transition:

\[ Pr[X > a] = Pr[\lambda X > \lambda a] = Pr[e^{\lambda X} > e^{\lambda a}] = Pr[e^{\lambda \sum x_i} > e^{\lambda a}] = Pr[\Pi_{i=1}^{n} e^{\lambda x_i} > e^{\lambda a}] \]

These hold due to the monotonicity of multiplication by a positive constant and exponentiation. Now, using Markov’s inequality on the last inequality we get:

\[ Pr[\Pi_{i=1}^{n} e^{\lambda x_i} > e^{\lambda a}] \leq \frac{E[\Pi_{i=1}^{n} e^{\lambda x_i}]}{e^{\lambda a}} \]

4. The first transition is true due to the independence of the variables \( x_i \). This means that \( e^{\lambda x_i} \) are independent. The second transition is due to all expectations of \( e^{\lambda x_i} \) being equal which stems from \( x_i \) being identically distributed.

5. First, we compute the expectation of \( e^{\lambda x_i} \)

\[ E[e^{\lambda x_i}] = \frac{1}{2} e^{\lambda} + \frac{1}{2} e^{-\lambda} = \cosh(\lambda) \leq e^{\lambda^2/2} \]

From the above we have that \( Pr[X > a] \leq e^{n\lambda^2/2} e^{-\lambda a} \). Setting \( \lambda = a/n \) we get \( e^{n\lambda^2/2} e^{-\lambda a} = e^{-\lambda^2} \) which concludes the proof.
2 Approximating the size of a graph

setup

In this question we will try to approximate the size of a graph. A graph \( G(V, E) \) is a set of nodes \( |V| = n \) and a set of edges \( |E| = m \). Each edge \( e \in V \times V \) is a set of two nodes which support it. We assume the graph is simple which means there are no duplicate edges and no self loops (i.e. an edge \( e = (u, u) \)). The degree of a node, \( \text{deg}(u) \), is the number of edges which it supports. More formally \( \text{deg}(u) = |\{ e \in E | u \in e \}| \). The degree of each node in the graph is at least 1. The question refers to the following sampling procedure:

1. \( e = (u, v) \leftarrow \) an edge uniformly at random from \( E \).
2. with probability \( 1/2 \)
3. return \( u \)
4. else
5. return \( v \)

Throughout this question we assume that i) we can sample edges uniformly from the graph ii) that the value of \( m \) is known iii) that given a node \( u \) we can compute \( \text{deg}(u) \). The value of \( n \), however, is unknown.

questions

1. Let \( p(u) \) denote the probability that the sampling procedure returns a specific node, \( u \). Compute \( p(u) \) as a function of \( \text{deg}(u) \) and \( m \). (Note: \( \sum_{u \in V} \text{deg}(u) = 2m \))
2. Let \( f(u) = \frac{2m}{\text{deg}(u)} \). Compute:
   \[ E_{x \sim \text{smpl}}[f(x)] \]
   where \( x \sim \text{smpl} \) denotes that \( x \) is chosen according to the distribution on the nodes generated by the above sampling procedure.
3. We say that a graph is \( d \)-degree-bounded if \( \max_{u \in V} \text{deg}(u) \leq d \). Show that for a \( d \)-degree-bounded graph:
   \[ \text{Var}_{x \sim \text{smpl}}[f(x)] \leq dn^2 \]
4. Let \( Y = \frac{1}{s} \sum_{i=1}^{s} f(x_i) \) where \( x_i \) are nodes chosen independently from the graph according to the above sampling procedure. Compute \( E[Y] \) and show that \( \text{Var}[Y] \leq dn^2 / s \).
5. Use Chebyshev's inequality to find a value for \( s \) such that for any \( d \)-degree-bounded graph and any two constants \( \varepsilon \in [0, 1] \) and \( \delta \in [0, 1] \):
   \[ \Pr[|Y - n| > \varepsilon n] < \delta. \]
   \( s \) should be a function of \( d, \varepsilon \) and \( \delta \).
answers

1. A node is chosen only if an edge it is adjacent to is picked with probability \( \frac{\deg(u)}{m} \) since the edges are chosen uniformly at random. The second event happens with probability \( \frac{1}{2} \) independently of the first event. This gives

\[
p(u) = \frac{\deg(u)}{m} \frac{\deg(2)}{2m} = \frac{\deg(u)}{2m}.
\]

2. By the definition to the expectation:

\[
E_{x \sim \text{smp}}[f(x)] = \sum_{u \in V} p(u)f(u) = \sum_{u \in V} \frac{\deg(u)}{2m} \frac{2m}{\deg(u)} = \sum_{u \in V} 1 = n
\]

3. We say that a graph is \( d \)-degree-bounded if \( \max_{u \in V} \deg(u) \leq d \). Show that for a \( d \)-degree-bounded graph:

\[
\text{Var}_{x \sim \text{smp}}[f(x)] \leq E_{x \sim \text{smp}}[f^2(x)] = \sum_{u \in V} \frac{\deg(u)}{2m} \left( \frac{2m}{\deg(u)} \right)^2 = \sum_{u \in V} \frac{2m}{\deg(u)}
\]

Since \( \deg(u) \geq 1 \) then

\[
\sum_{u \in V} \frac{2m}{\deg(u)} \leq \sum_{u \in V} \frac{2m}{1} = 2mn.
\]

Also, since the graph is \( d \)-degree-bounded \( 2m = \sum_{u \in V} \deg(u) \leq nd \) thus

\[
2mn \leq dn^2.
\]

4. \( Y \) is the average of \( s \) independent copies of \( f(x) \) and therefore, by linearity of the expectation, we have that

\[
\]

Moreover, since the nodes \( x_i \) are chosen independently we have that

\[
\text{Var}[Y] = \frac{1}{s} \sum_{i=1}^{s} \text{Var}[f(x_i)].
\]

Since \( f(x_i) \) distribute identically and substituting \( \text{Var}(x) \leq dn^2 \) we get

\[
\frac{1}{s} \sum_{i=1}^{s} \text{Var}[f(x_i)] \leq \frac{dn^2}{s}.
\]

5. Since \( E[Y] = n \) we get that the above holds if

\[
\Pr[|Y - E[n]| > \varepsilon n] < \frac{\text{Var}[Y]}{\varepsilon^2 n^2} \leq \frac{dn^2}{s} \frac{dn^2}{s} = \frac{d}{s \varepsilon^2}.
\]

The condition that \( \frac{d}{s \varepsilon^2} \leq \delta \) holds for \( s \geq \frac{d}{s \varepsilon^2} \).
3  Approximate median

setup

Given a list $A$ of $n$ numbers $a_1, \ldots, a_n$, we define the rank of an element $r(a_i)$ as the number of elements which are smaller than it. For example, the smallest number has rank zero and the largest has rank $n - 1$. Equal elements are ordered arbitrarily. The median of $A$ is an element $a$ such that $r(a) = n/2$ (rounded either up or down). An $\alpha$-approximate-median is a number $a$ such that:

$$n(1/2 - \alpha) \leq r(a) \leq n(1/2 + \alpha)$$

In this question we sample $k$ elements uniformly at random with replacement from the list $A$. Let the samples be $\{x_1, \ldots, x_k\} = X$. You will be asked to show that the median of $X$ is an $\alpha$-approximate-median of $A$.

questions

1. What is the probability the a randomly chosen element $x$ is such that:

$$r(x) > n(1/2 + \alpha)$$

2. Let us define $X_{>\alpha}$ as the set of samples whose rank is greater than $n(1/2 + \alpha)$. More precisely, $X_{>\alpha} = \{x_i \in X | r(x_i) > n(1/2 + \alpha)\}$. Similarly we define $X_{<\alpha} = \{x_i \in X | r(x_i) < n(1/2 - \alpha)\}$. Prove that if $|X_{>\alpha}| < k/2$ and $|X_{<\alpha}| < k/2$ then the median of $X$ is an $\alpha$-approximate-median of $A$.

3. Let $Z = |X_{>\alpha}|$. Find $t$ for which:

$$\Pr[Z \geq k/2] = \Pr[Z \geq (1 + t)E[Z]]$$

4. Bound from above the probability that $Z \geq k/2$ as tightly as possible. If you do so using a probabilistic inequality, justify your choice.

5. Compute the minimal value for $k$ which will guarantee that $|X_{>\alpha}| < k/2$ and $|X_{<\alpha}| < k/2$ with probability at least $1 - \delta$. 


answers

1. There are \( n(1/2 - \alpha) \) elements for which \( r(x) > n(1/2 + \alpha) \). Since the element is chosen uniformly, the probability of that happening is \( (1/2 - \alpha) \).

2. First we note that the median of \( X \) cannot be either in \( X_{>\alpha} \) or in \( X_{<\alpha} \). This is simply because each of them includes less than half of the elements in \( X \). Moreover, by the definitions of \( X_{>\alpha} \) and \( X_{<\alpha} \) we have:
   \[
   n(1/2 - \alpha) \leq r(\text{median}(X)) \quad \text{and} \quad r(\text{median}(X)) \leq n(1/2 + \alpha)
   \]
   which means that median\((X)\) is an \( \alpha \)-approximate-median of \( A \).

3. Since the probability of a sample being in \( X_{>\alpha} \) is exactly \( 1/2 - \alpha \) and since we have \( k \) independent samples, \( E[Z] = E[|X_{>\alpha}|] = k(1/2 - \alpha) \). Solving for \( t \) we get
   \[
   (1 + t)E[Z] = k/2 \quad \rightarrow \quad (1 + t)(1/2 - \alpha) = 1/2 \quad \rightarrow \quad t = \frac{2\alpha}{1 - 2\alpha}
   \]

4. Since the value of \( Z \) is the sum of independent indicator variables we can apply Chernoff’s inequality. Denoting \( \mu = E[Z] = k(1/2 - \alpha) \) and \( t = \frac{2\alpha}{1 - 2\alpha} \) we have:
   \[
   \Pr[Z \geq k/2] = \Pr[Z \geq (1 + t)\mu] \leq e^{-\mu t^2/4}
   \]

5. Similarly to the the above we can argue that
   \[
   \Pr[|X_{<\alpha}| \geq k/2] \leq e^{-\mu t^2/4}
   \]

From the union bound we have that the probability of the event that \( |X_{<\alpha}| \geq k/2 \) or that \( |X_{>\alpha}| \geq k/2 \) is at most the sum of their probabilities.

\[
\Pr[|X_{<\alpha}| \geq k/2 \cup |X_{>\alpha}| \geq k/2] \leq \Pr[|X_{<\alpha}| \geq k/2] + \Pr[|X_{>\alpha}| \geq k/2] \leq 2e^{-\mu t^2/4}
\]

Demanding that this failure probability is less than \( \delta \) we guarantee success with probability at least \( 1 - \delta \). Substituting \( \mu = k(1/2 - \alpha) \) and \( t = \frac{2\alpha}{1 - 2\alpha} \) this is achieved for
   \[
   2e^{-\mu t^2/4} < \delta \quad \rightarrow \quad k > \frac{4\log(2/\delta)(1/2 - \alpha)}{\alpha^2}
   \]
4 Soulmate search

setup

In this question you will be asked to derive a search algorithm for a unique nearest neighbor given a local sensitive hash function family. We assume a universe of \( n \) objects \( x_1, \ldots, x_n \) and a distance function \( d \). For any pair of points \( 0 \leq d(x_i, x_j) \leq 1 \). Moreover, each point \( x_i \) has exactly one soulmate point \( x_j \) such that \( d(x_i, x_j) \leq r \), \( r \) is known constant. For all other points in the universe \( d(x_i, x_j') > 2r \). You are given a family \( H \) of hash functions such that \( \Pr_{h \sim H}[h(x_i) = h(x_j)] = \frac{1}{1+d(x_i, x_j)} \) for any pair \( (h \sim H \) means that \( h \) is chosen uniformly from \( H \). We also define a bucketing hash function \( g \) which accepts an element \( x \) and returns a list of hash values.

\[
g(x) = [h_1(x), \ldots, h_k(x)]
\]

where each of the hash functions \( h_1, \ldots, h_k \) was chosen uniformly and independently from the family \( H \). We say that \( x_i \) and \( x_j \) are in love if \( g(x_i) = g(x_j) \).

questions

1. What is the probability of two points, whose distance is \( d(x_i, x_j) \), falling in love?

2. Compute a value for \( k \) for which the probability that \( x_i \) and \( x_j \) who are not soulmates \( (d(x_i, x_j) \geq 2r) \) of falling in love is at most \( 1/n \). Or, find \( k \) for which the following holds:

\[
\Pr[g(x_i) = g(x_j) \mid d(x_i, x_j) \geq 2r] \leq 1/n
\]

3. For this value of \( k \), what is the probability that \( x_i \) falls in love with her soulmate? That means \( \Pr[g(x_i) = g(x_j) \mid d(x_i, x_j) \leq r] \). Help: you can use the approximation

\[
\frac{\log(1+r)}{\log(1+2r)} \approx \frac{1}{2}.
\]

4. We now create \( m \) independent copies of \( g \), \( g_1, \ldots, g_m \). We say that \( x_i \) finds \( x_j \) if \( g_\ell(x_i) = g_\ell(x_j) \) for at least one function \( g_\ell \). Give a bound on the value of \( m \) which insures that all \( x_i \) find their soulmates with probability at least \( 1 - \delta \)?

5. Given the above value for \( m \), bound from above the expected number of points \( x_j \) that \( x_i \) fell in love with which were not her soulmates.
answers

1. Since each of the hash functions was chosen independently, we have that for each \( h_\ell(x_i) = h_\ell(x_j) \) for \( x_i \) and \( x_j \) to be in love this must hold for all \( k \) hash functions which happens with probability \( \frac{1}{(1+d(x_i,x_j))^k} \).

2. Using the expression from the previous question for two points for which \( d(x_i,x_j) \geq 2r \) we have:

\[
\frac{1}{(1 + d(x_i,x_j))^k} \leq \frac{1}{(1 + 2r)^k} \leq \frac{1}{n} \quad \Rightarrow \quad k \geq \frac{\log(n)}{\log(1 + 2r)}
\]

3. Substituting \( d(x_i,x_j) \leq r \) and \( k = \frac{\log(n)}{\log(1 + 2r)} \) we get:

\[
\Pr[g(x_i) = g(x_j)] = \frac{1}{(1 + d(x_i,x_j))^k} \geq \frac{1}{(1 + r)^k} = (1 + r)^{-\frac{\log(n)}{\log(1 + 2r)}} = n^{-\frac{\log(1+r)}{\log(1+2r)}} \approx n^{-1/2}
\]

which uses the approximation \( \frac{\log(1+r)}{\log(1+2r)} \approx \frac{1}{2} \).

4. For a point to fail in finding her soulmate, it must fail in falling in love \( m \) consecutive times. The probability of one point failing is therefore \( (1 - n^{-1/2})^m \). By the union bound, the probability of any of the \( n \) points failing is at most \( n(1 - n^{-1/2})^m \), demanding that this is bounded by \( \delta \) yields:

\[
n(1 - n^{-1/2})^m \approx ne^{-m/\sqrt{n}} \leq \delta \quad \Rightarrow \quad m \geq \sqrt{n \log(n/\delta)}
\]

5. We can denote by \( Z_{i,j,\ell} \) the event that point \( x_i \) and \( x_j \) are such that \( d(x_i,x_j) > 2r \) and \( g_\ell(x_i) = g_\ell(x_j) \). The number of points that \( x_i \) falls in love with is bounded by \( \sum_{i=1}^{n} \cup_{\ell=1}^{m} Z_{i,j,\ell} \). Using the linearity of expectation and the fact that \( \Pr[Z_{i,j,\ell} = 1] \leq 1/n \) we have that:

\[
E[\sum_{i=1}^{n} \cup_{\ell=1}^{m} Z_{i,j,\ell}] \leq E[\sum_{i=1}^{n} \sum_{\ell=1}^{m} Z_{i,j,\ell}] = \sum_{i=1}^{n} \sum_{\ell=1}^{m} E[Z_{i,j,\ell}] \leq \sum_{i=1}^{n} \sum_{\ell=1}^{m} 1/n = m
\]