COL202: Discrete Math **Tutorial 8 Solutions**

Instructor: Venkata Koppula

1 General Discrete Probability

1.1. (♦) **k-wise independence:** A set of n events $A_1, A_2, ..., A_n$ is *k-wise independent* if for every set $I \subset [n]$ such that $|I| = k$, the set of events $\{A_i\}_{i \in I}$ are mutually independent.

Consider a source of randomness, where you press a button, and get a uniformly random element in $\mathbb{Z}_q$ (where q is a prime). Each button-pressing costs $Rs.$ 1000. You want a set of n numbers in $\mathbb{Z}_q$ s.t. they're *k-wise independent*. One direct way is to press the button n times to obtain n uniformly random numbers, but do we really need this many calls to the source for *k-wise independence*? Describe a solution that uses only $O(k)$ samples.

More formally, let $t = \Theta(k)$, and for any subset $I \subset [n]$, let $(y_1, y_2, \ldots, y_n)_I$ denote the sequence $(y_i)_{i \in I}$. Describe a deterministic function $F : \mathbb{Z}_q^t \to \mathbb{Z}_q^n$ such that for any subset $I \subset [n]$, $|I| = k$, and any $\theta_1, \theta_2, \ldots, \theta_k \in \mathbb{Z}_q$,

$$\Pr\left[(F(x_1, x_2, \ldots, x_t))_I = (\theta_1, \theta_2, \ldots, \theta_k)\right] = \frac{1}{q^k}$$

where the probability is over the choice of $x_1, x_2, \ldots, x_t$, sampled independently and uniformly at random from $\mathbb{Z}_q$.

Solution: We produce a $k-$wise independent set of n numbers using $t = k$ uniformly random samples from $\mathbb{Z}_q$ (assume $n < q$). Define $F : \mathbb{Z}_q^k \to \mathbb{Z}_q^n$

$$F(x_0, \ldots, x_{k-1}) = (v_1, \ldots, v_n)$$

$$v_i = (x_0 + x_1 i + x_2 i^2 \cdots + x_{k-1} i^{k-1}) \mod q \qquad 1 \le i \le n$$

This can also be written in the form of a linear transformation:

$$\begin{pmatrix} v_1 \\ v_2 \\ \vdots \\ v_n \end{pmatrix} = \begin{pmatrix} 1 & 1 & \ldots & 1 \\ 1 & 2 & \ldots & 2^{k-1} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & n & \ldots & n^{k-1} \end{pmatrix} \begin{pmatrix} x_0 \\ x_1 \\ \vdots \\ x_{k-1} \end{pmatrix}$$

Or in other words

$$\mathbf{v} = \mathbf{F}\mathbf{x}$$

Where all the arithmetic is performed $\mod q$. Now, for a k sized subset $I \subseteq [n]$, what is the probability that $\mathbf{v}_I = (\theta_1, \ldots, \theta_k) = \mathbf{\Theta_I}$? It can be seen that $\mathbf{F}$ has

rank k, so any $k-$rows are linearly independent, that is, the $k \times k$ matrix formed by any $k-$rows is invertible.

Let $I \subseteq [n], |I| = k$ be a set of indices and $\mathbf{F_I}$ be the submatrix of $\mathbf{F}$ limited to the indices I. We can take the inverse of $\mathbf{F_I}$ to uniquely determine the values of $x_0, \ldots, x_{k-1}$:

$$\mathbf{F_I x} = \mathbf{\Theta_I} \implies \mathbf{x} = \mathbf{F_I}^{-1}\mathbf{\Theta_I}$$

Since the x_i are sampled uniformly at random from $\mathbb{Z}_q$, the probability that the sampled value matches the one calculated above is $\frac{1}{q}$ at each index. Hence, the probability of obtaining the correct vector is $\frac{1}{q^k}$

2 Concentration Inequalities

2.1. There are n balls and n bins. Each ball is tossed into one of the n bins, uniformly at random. Let X denote the number of bins with at least two balls. What is $\mathbb{E}[X]$? Give an upper bound on $\Pr[X > 3n/8]$.

Solution: Let X_i be the indicator random variable such that $X_i = 1$ if and only if the number of balls in the i^{th} bin is at least 2. Then $X = \sum_{i=1}^{n} X_i$. Now,

$$\begin{aligned}
\mathbb{E}[X_i] &= \Pr[\text{Bin } i \text{ has at least 2 balls}] \\
&= 1 - \Pr[\text{Bin } i \text{ has no ball}] - \Pr[\text{Bin } i \text{ has 1 ball}] \\
&= 1 - \left(\frac{n-1}{n}\right)^n - \binom{n}{1}\frac{1}{n}\left(\frac{n-1}{n}\right)^{n-1} \\
&= 1 - \left(1 - \frac{1}{n}\right)^n \frac{2n-1}{n-1} \\
&= 1 - \left(1 - \frac{1}{n}\right)^{n-1} \frac{2n-1}{n}
\end{aligned}$$

Thus, by Linearity of Expectation:

$$\mathbb{E}[X] = n - (2n-1)\left(1 - \frac{1}{n}\right)^{n-1}$$

For an upper-bound, we use Markov's Inequality:

$$\Pr\left[X > \frac{3n}{8}\right] \leq \frac{8\mathbb{E}[X]}{3n} = \underbrace{\frac{8}{3} - \frac{8}{3}\left(1 - \frac{1}{n}\right)^{n-1}\frac{2n-1}{n}}_{f(n)}$$

Some analysis shows that $f(n)$ is an increasing function of n. The maximum value is obtained as $n \to \infty$:

$$\lim_{n\to\infty} f(n) = \frac{8}{3}\left(1 - \frac{2}{e}\right) \approx 0.7$$

Which is the required upper bound.

2.2. (★) In class, we saw the exit polls problem. There are n balls in a bag, where each ball is either red or blue colored. You are given the guarantee that either there are at least $2n/3$ blue balls, or at least $2n/3$ red balls. We want to determine whether there are more red balls or more blue balls, by sampling as few balls as possible.

We discussed the following algorithm: let $t = c \log n$ where c is a sufficiently large constant. For $i = 1$ to t:

1. Sample a ball from the bag, uniformly at random. Let z_i denote the color of the ball. After noting the color, put the ball back in the bag.

If the majority of $\{z_1, z_2, \ldots, z_n\}$ are red, then output "More red balls", else output "More blue balls".

We showed that if $t \geq c \log n$ where c is a sufficiently large constant, then the above algorithm gives the correct output with probability at least $1 - 1/n$.

1. Suppose we change the first step of the algorithm as follows: Sample a ball from the bag, uniformly at random. Let z_i denote the color of the ball. After noting the color, **do not** put the ball back in the bag.

Show that there exists a constant c such that the modified algorithm gives the correct output with probability at least $1 - 1/n$.

2.3. We have two coins: one is a fair coin, but the other produces heads with probability $3/4$. One of the two coins is picked, and this coin is tossed n times. Use the Chernoff Bound to determine the smallest n which allows determination of which coin was picked with 95% confidence.

Solution: In this problem, we will be using another form of the Chernoff bound. Using similar techniques as in the proof of your tutorial 7 submission problem, one can derive the following bound for i.i.d. variables X_i:

$$\Pr\left[\frac{1}{n}\sum_i X_i \geq p + \epsilon\right] \leq e^{-2\epsilon^2 n}$$

$$\Pr\left[\frac{1}{n}\sum_i X_i \leq p - \epsilon\right] \leq e^{-2\epsilon^2 n}$$

Try to derive these as an exercise.

Let $X_{i,1}$ be the outcome of the i^{th} toss using the first coin ($p_1 = 1/2$) and $X_{i,2}$ be the outcome of the i^{th} toss using the second coin ($p_2 = 3/4$). The Chernoff bound tells us that if we toss the j^{th} coin $j \in \{1, 2\}$ n times, then the fraction of times we obtain a head would ϵ-close to p_i with high probability. Taking $\epsilon \leq 1/8$ allows us to distinguish between the two coins. Thus

$$\Pr\left[\frac{1}{n}\sum_i X_{i,1} \geq p_1 + \epsilon\right] \leq e^{-2\epsilon^2 n}$$

$$\Pr\left[\frac{1}{n}\sum_i X_{i,2} \le p_2 - \epsilon\right] \le e^{-2\epsilon^2 n}$$

With $\epsilon = 1/8$. Thus the probability is upper bounded by $e^{-\frac{n}{32}}$ which we want to be less than 0.05. This gives us $n \ge 96$.

2.4. We want to store 2 billion records into a hash table that has 1 billion slots. Assuming the records are randomly and independently chosen with uniform probability of being assigned to each slot, two records are expected to be stored in each slot. Of course under a random assignment, some slots may be assigned more than two records.

1. Show that the probability that a given slot gets assigned more than 23 records is less than e^{-36}.
2. Show that the probability that there is a slot that gets assigned more than 23 records is less than e^{-15}.

2.5. Suppose we throw m balls in n bins where each ball is thrown in a bin uniformly at random and independent of other balls. Suppose $m = 2\sqrt{n}$. Use Chernoff bounds plus the union bound to bound the probability that no bin has more than 1 ball.

Solution: To bound the probability that no bin contains more than one ball, we first consider the complementary event: that at least one bin contains two or more balls. We will bound the probability of this complementary event and then use it to infer the desired probability.

Step 1: Define Indicator Variables

For each bin $b \in [n]$ and each ball $i \in [m]$, define an indicator variable:

$$Z_{b,i} = \begin{cases} 1 & \text{if ball } i \text{ falls into bin } b, \\ 0 & \text{otherwise.} \end{cases}$$

These indicator variables are mutually independent for different balls.

Step 2: Number of Balls in a Bin

Let B_b denote the number of balls in bin b:

$$B_b = \sum_{i=1}^{m} Z_{b,i}.$$

The expected number of balls in bin b is:

$$\mathbb{E}[B_b] = \sum_{i=1}^{m} \mathbb{E}[Z_{b,i}] = \sum_{i=1}^{m} \frac{1}{n} = \frac{m}{n} = \frac{2}{\sqrt{n}}$$

substituting $m = 2\sqrt{n}$

Step 3: Applying Chernoff Bounds

We aim to bound the probability that $B_b \geq 2$ for any bin b. We use the following version of Chernoff Bound:

$$\Pr[B_b \geq (1+\delta)\mathbb{E}[B_b]] \leq 2e^{-\delta^2\mathbb{E}[B_b]/3}.$$

Setting $(1+\delta)\mathbb{E}[B_b] = 2$, we solve for δ:

$$1 + \delta = \frac{2}{\mathbb{E}[B_b]} = \sqrt{n} \quad \Rightarrow \quad \delta = \sqrt{n} - 1.$$

Substituting δ back into the Chernoff bound:

$$\Pr[B_b \geq 2] \leq 2e^{-\frac{2(\sqrt{n}-1)^2}{3\sqrt{n}}} \leq 2e^{-\sqrt{n}/2}$$

Step 4: Applying the Union Bound

To bound the probability that *any* bin contains two or more balls, we apply the union bound over all n bins:

$$\Pr[\exists\, b \in [n] \text{ such that } B_b \geq 2] \leq n \cdot 2e^{-\sqrt{n}/2} \leq e^{-n^{1/3}} \text{ (for large enough n)}$$

Thus we get the required probability:

$$\Pr[\forall\, b \in [n] \text{ such that } B_b < 2] = 1 - \Pr[\exists\, b \in [n] \text{ such that } B_b \geq 2] \geq 1 - e^{-n^{1/3}}$$

2.6. Suppose you are given an array A with n distinct integers, and you want to find an approximate median of these n numbers. Any deterministic algorithm will require accessing $\Omega(n)$ elements of A. However, using randomization, we can do this using only $O(\log n)$ accesses. Consider the following algorithm:

1: Choose $k = c \log n$ elements from A, uniformly at random, with replacement.
2: Sort these k elements
3: Return the median x of the k elements

Prove that with probability at least $1 - 1/n$, at least $n/3$ numbers in A are smaller than x, and at least $n/3$ numbers in A are larger than x. You must choose c to be an appropriately large constant.

2.7. (♦) Show that on tossing a fair coin n times, the length of longest contiguous sequence of heads will be $O(\log n)$ with probability at least $1 - 1/n^2$.

Solution:
We show the following :

Let $E_{i,k}$ be the event that there is a sequence of atleast k consecutive heads starting from the i^{th} toss for $1 \le i \le n-k+1$.
Clearly $\mathbf{P}(E_{i,k}) = 1/2^k$
Now, for E_k having the event that there is a sequence of atleast k consecutive heads, we have $E_k = \bigcup_{i=1}^{n-k+1} E_{i,k}$
By union bound, $\mathbf{P}(E_k) \le \sum_{i=1}^{n-k+1} \mathbf{P}(E_{i,k}) = (n-k+1)/2^k$
For $k = 3\log n$, $\mathbf{P}(E_k) \le (n - 3\log n + 1)/n^3 < 1/n^2$
Hence $\mathbf{P}(\bar{E_k}) > 1 - 1/n^2$ for $k = 3\log n$,
where $\bar{E_k}$:= every sequence of consecutive heads has length $< \mathrm{K} = 3\log n$
which shows the required claim.

2.8. In this problem we consider the task of *supervised learning*. We are given a set S of n train samples of the form $(x_0, y_0), \ldots, (x_{n-1}, y_{n-1})$ drawn i.i.d. from some unknown distribution D over pairs (x, y). For simplicity $x_i \in \{0,1\}^m$ and $y_i \in \{0,1\}$. The goal is to find a *classifier* $h : \{0,1\}^m \to \{0,1\}$ that will minimize the test error. One way to find such a classifier is to consider a collection $\mathcal{C}$ of potential classifiers and look at the classifier that does best on the training set S. The test error is defined as $L(h) = Pr_{(x,y)\sim D}[h(x) \neq y]$ and the train error is defined as $\hat{L}_S(h) = \sum_{i\in[n]} |h(x_i) - y_i|/n$, and using the Chernoff bound we can show that as long as the number n of samples is sufficiently larger than logarithm of $|\mathcal{C}|$, the test error will be close to the train error $\forall h \in \mathcal{C}$. So, prove that for every $\epsilon, \delta > 0$, if $n > \log|\mathcal{C}| \log(1/\delta)/\epsilon^2$, then, $\underset{S}{Pr}[\forall_{h\in\mathcal{C}} |L(h) - \hat{L}_S(h)| \le \epsilon] > 1 - \delta$, where the probability is taken over the choice of the set of samples S.

In particular, this tells you that if you have a set of (input, output) pairs in your training data, and you need to choose a hypothesis for future *test* inputs, then the *best* hypothesis is the one that minimizes the error on training data.

Solution:
For any arbitrary $h \in \mathcal{C}$, let $X_{h,s} := n.\vec{L_s}(h) = \sum_{i\in[n]} |h(x_i) - y_i|$
If we let $X_{h,s}^{(i)} = |h(x_i) - y_i| = 1$ if $h(x_i) \neq y_i$ and 0 if $h(x_i) = y_i$
$X_{h,s}$ is a sum of n independent 0,1 random variables.
Moreover $\mathbf{E}X_{h,s} = \sum_{i\in[n]} \mathbf{E}X_{h,s}^{(i)} = \sum_{i\in[n]} \mathbf{P}[h(x_i) \neq y_i] = \sum_{i\in[n]} L(h) = n.L(h)$
Thus, applying the Chernoff bound we get,
$\mathbf{P}[|X_{h,s} - n.L(h)| \ge n.\epsilon] \le 2e^{-n^2\epsilon^2/\ni.L(h)} = 2e^{-n\epsilon^2/\ni.L(h)} \le 2e^{-n\epsilon^2/\ni}$ since L(h) ≤ 1
by the union bound we then have,
$\mathbf{P}[\exists h \in \mathcal{C}, |X_{h,s} - n.L(h)| \ge n.\epsilon] \le \sum_{h\in\mathcal{C}} P[|X_{h,s} - n.L(h)| \ge n.\epsilon] \le 2|\mathcal{C}|.e^{-n\epsilon^2/\ni}$
$\le 2.|\mathcal{C}|e^{-(\log|\mathcal{C}| - log(1/\delta)/3}$ using $n > (log|\mathcal{C}| + log(1/\delta)/\epsilon^2$
$= 2\delta^{1/3}$
$\Rightarrow \mathbf{P}[\exists h \in \mathcal{C}, |X_{h,s} - n.L(h)| < n.\epsilon] \ge 1 - 2\delta^{1/3}$
Finally observe that this is same as
$\mathbf{P}[\exists h \in \mathcal{C}, |X_{h,s} - n.L(h)| \le \epsilon] \ge 1 - 2\delta^{1/3}$

2.9. (†††) **Rumor-spreading:** There are n friends. On day 1, person 1 thinks of a

rumor, and the rumor spreads as follows: every day, if a person knows the rumor, then he/she calls up a uniformly random friend (one of the remaining $n-1$ people, uniformly at random) and shares the rumor. Let X denote the number of days for everyone to know the rumor. Clearly, $X \geq \log n$ since the number of people knowing the rumor can only double everyday.

Prove that there exist constants c_1 and c_2 such that

$$\mathbb{E}[X] \leq c_1 \log n \text{ and } \Pr[X > c_2 \log n] \leq \frac{1}{n}.$$

2.10. (†††) **Balls and Bins, revisited:** There are n balls and n bins. We throw the balls in the bins, one by one, as follows. For $i = 1$ to n, do the following:

1. sample two bins uniformly at random.
2. pick the bin with lower load. If both bins have the same load, then pick the left one. Toss the i^{th} ball in this bin.

Let X denote the max load. What is $\mathbb{E}[X]$? To build intuition for this process, we recommend simulating it for large values of n.

3 Probabilistic Method

3.1. Let $\mathcal{F}$ be a family of subsets of $[2n]$. We say that $\mathcal{F}$ is an *antichain* if for any $A, B \in \mathcal{F}$, A is not a subset of B, and B is not a subset of A. For example, consider the family all subsets of $[2n]$ of size exactly n. Check that these form an antichain. This is the largest possible antichain.

More formally, prove that for any antichain $\mathcal{F}$,

$$|\mathcal{F}| \leq \frac{(2n)!}{n! \cdot n!}.$$

Solution: Consider a random permutation $\pi : [2n] \to [2n]$. We compute the probability of the event that a prefix of this permutation $(\pi(1), \ldots, \pi(k))$ is in $\mathcal{F}$ for some k. Note that this can happen for at most one value of k, since otherwise $\mathcal{F}$ would not be an antichain. For each particular set $A \in \mathcal{F}$,

$$\Pr[A = \{\pi(1), \ldots, \pi(|A|)\}] = \frac{|A|!(2n - |A|)!}{(2n)!}$$

corresponding to all possible orderings of A and $[2n] \setminus A$. By the property of an antichain, these events for different sets $A \in \mathcal{F}$ are disjoint, and hence

$$\begin{aligned}\Pr[\exists A \in \mathcal{F}, A = \{\pi(1), \ldots, \pi(|A|)\}] &= \sum_{A \in \mathcal{F}} \Pr[A = \{\pi(1), \ldots, \pi(|A|)\}] \\ &= \sum_{A \in \mathcal{F}} \frac{|A|!(2n - |A|)!}{(2n)!} = \sum_{A \in \mathcal{F}} \frac{1}{\binom{2n}{|A|}}\end{aligned}$$

Further, use $\Pr[\exists A \in \mathcal{F}, A = \{\pi(1), \ldots, \pi(|A|)\}] \leq 1$ and $\binom{2n}{|A|} \leq \binom{2n}{n}$ $\forall A \subseteq [2n]$ above to obtain

$$1 \geq \sum_{A \in \mathcal{F}} \frac{1}{\binom{2n}{|A|}} \geq \frac{|\mathcal{F}|}{\binom{2n}{n}}$$
$$\Rightarrow |\mathcal{F}| \leq \binom{2n}{n} = \frac{(2n)!}{n! \cdot n!}$$

3.2. (♦) Prove that for every $n \times n$ matrix $\mathbf{A}$ with binary entries, there exists a vector $\mathbf{b} \in \{-1, +1\}^n$ such that the absolute value of the largest entry of $\mathbf{A} \cdot \mathbf{b}$ is at most $4\sqrt{n \ln n}$.

Solution: Consider any vector $\mathbf{a} \in \{0,1\}^n$, and pick a uniformly random $\mathbf{b} \leftarrow \{-1,1\}^n$. We are interested in the probability that $\mathbf{a}^\top \cdot \mathbf{b}$ is greater than $4\sqrt{n \ln n}$. Let $I \subseteq [n]$ be the indices where $\mathbf{a}_i = 1$. Clearly, we only care about $\mathbf{b}_i$ for $i \in I$. Let $|I| = t$, then we are interested in the probability

$$\Pr_b \left[\left| \sum_{i \in I} b_i \right| > 4\sqrt{n \ln n} \right]$$

Let $\tilde{b_i} = \frac{b_i + 1}{2}$ which gives $b_i = 2\tilde{b_i} - 1$. This transformation allows us to convert them to Bernoulli random variables. Applying this substitution to the above equation:

$$\Pr_b \left[\left| 2 \sum_{i \in I} \tilde{b_i} - t \right| > 4\sqrt{n \ln n} \right]$$

$$\implies \Pr_b \left[\left| \sum_{i \in I} \tilde{b_i} - t/2 \right| > 2\sqrt{n \ln n} \right]$$

Now we can apply a Chernoff Bound $\left(\mu = \frac{t}{2}, \delta = \frac{2\sqrt{n \ln n}}{\mu}\right)$:

$$\Pr_b \left[\left| \sum_{i \in I} \tilde{b_i} - t/2 \right| > 2\sqrt{n \ln n} \right] \leq 2e^{-\delta^2 \mu/3} = 2e^{-\frac{4n \ln n}{3t}} \leq 2e^{-\frac{4}{3} \ln n} = \frac{2}{n^{4/3}}$$

Going back to our original problem : consider any n vectors $\mathbf{a_1}, \mathbf{a_2}, \ldots, \mathbf{a_n}$. Sample a uniformly random $\mathbf{b} \leftarrow \{-1,1\}^n$. What is the probability that there exists an $i \in [n]$ such that $\left|\mathbf{a_i}^\top \cdot \mathbf{b}\right| > 4\sqrt{n \ln n}$? This is at most $2/n^{\frac{1}{3}}$ (doing a union bound over all n vectors). Hence, for any matrix $\mathbf{A} \in \{0,1\}^{n \times n}$, if we sample a uniformly random vector $\mathbf{b} \leftarrow \{-1,1\}^n$, then

$$\Pr\left[\text{abs. val. of largest entry of } \mathbf{A} \cdot \mathbf{b} \text{ is greater than } 4\sqrt{n \ln n}\right] \leq 2/n^{1/3} < 1.$$

For $n > 8$. Hence, using the probabilistic method, there exists a vector $\mathbf{b} \in \{-1,1\}^n$

such that all entries of $\mathbf{A} \cdot \mathbf{b}$ are at most $4\sqrt{n \ln n}$.

Note: Here we obtained a solution for $n > 8$. Using a tighter variant of the Chernoff bound, try showing it for all n!