Problem 1 [10 + 10 + 10 = 30 pts]: Consider an experiment in which you toss a fair coin (i.e., the probability of turning up heads is 1/2 for each toss) \( n \) times. All the tosses are independent. Answer the following questions:

(I) Let \( X_i \) be a 0–1 random variable which is 1 if the total number of heads you saw in tosses 1, 2, \ldots, \( i \) is even. The random variable \( X_i \) is defined for each index \( i = 1, 2, \ldots, n \). Let \( X = X_1 + X_2 + \ldots + X_n \). What is the expected value \( E[X] \)?

(II) Give the best bound you can, in terms of \( n, \delta \) for the following tail probability \( \Pr[[X - E[X]] \geq \delta E[X]] \) where \( 0 < \delta \leq 1 \).

(III) Let random variables \( Y_1, \ldots, Y_n \) be defined as follows: \( Y_i = 1 \) if the number of heads you have seen till the \( i \)th toss is divisible by 3. Are \( Y_1, \ldots, Y_n \) independent?

Solution to Problem 1:

(I) Consider any sequence of tosses upto the \((i - 1)\)th toss. For each such sequence, there is precisely one result of the \( i \)th toss, out of the two possible choices, that makes the total number of heads even upto the \( i \)th toss. Therefore, it is easily seen that \( E[X_i] = \Pr[X_i = 1] = 1/2 \). Thus, by linearity of expectation, \( E[X] = \sum_{i=1}^n E[X_i] = n/2 \).

(II) We first argue that \( X_1, \ldots, X_n \) are independent random variables, and then we use a Chernoff bound to estimate the desired tail probability. Let \( i_1 < i_2 < \ldots < i_k \) be distinct indices between 1 and \( n \). We show that,

\[
\Pr[X_{i_1} = 1 | X_{i_2} = a_2 \land \ldots \land X_{i_{k-1}} = a_{k-1}] = \Pr[X_{i_k} = 1],
\]

where the \( a_i \) are 0 or 1. Then, by an easy induction argument one can show that,

\[
\Pr[X_{i_1} = a_1 \land X_{i_2} = a_2 \ldots \land X_{i_k} = a_k] = \prod_{j=1}^k \Pr[X_{i_j} = a_j](= 1/2^k),
\]

thereby proving the independence of the \( X_i \). To prove the claim notice that the event \( X_{i_1} = a_1 \land \ldots \land X_{i_{k-1}} = a_{k-1} \) leads to a set of allowable sequences upto the \( i_k \)-th toss. (Actually the same inductive proof also shows that this set of allowable sequences is non-empty, though we have not stated that explicitly in the induction statement.) For each such allowable sequence, and each remainder sequence from \( i_{k-1} \)-th upto the \( i_k \)-th sequence, there will be exactly 1 choice of toss which makes the total number of heads up to the \( i_k \)-th toss even. As such conditioned on the event we still have that \( X_{i_k} = 1 \) with probability 1/2.

Having established independence we can use Chernoff bound to get \( \Pr[[X - E[X]] \geq \delta E[X]] \leq 2e^{-\frac{n\delta^2}{6}} \).

(III) No, the \( Y_i \) are not independent. For example \( \Pr[Y_1 = 1 \land Y_2 = 0 \land Y_3 = 1] = 0 \neq \Pr[Y_1 = 1] \cdot \Pr[Y_2 = 0] \cdot \Pr[Y_3 = 1] \) since if \( Y_1 = 1 \) the first toss must have been a ‘T’, and since \( Y_2 = 0 \) the second toss must have been a ‘H’, but then for any possible choice of toss we cannot have \( Y_3 = 1 \). On the other hand, the sequence of tosses ‘TTT’ achieves \( Y_1 = 1 \), and \( Y_3 = 1 \) thereby proving \( \Pr[Y_1 = 1] > 0, \Pr[Y_3 = 1] > 0 \) and the sequence of tosses ‘THT’ shows that \( \Pr[Y_2 = 0] > 0 \).

Problem 2 [20 pts]: Let \( X = \sum_{i=1}^n X_i \) where the \( X_i \) are independent 0–1 random variables, and let \( \mu = E[X] \). Suppose \( \mu_L, \mu_H \) are chosen such that \( \mu_L \leq \mu \leq \mu_H \). Then for any \( \delta > 0 \), show that

\[
\Pr[X \geq (1 + \delta) \mu_H] \leq \left( \frac{e^\delta}{(1 + \delta)^{1+\delta}} \right)^{\mu_H}.
\]

Similarly, for \( 0 < \delta < 1 \),

\[
\Pr[X \leq (1 - \delta) \mu_L] \leq \left( \frac{e^{-\delta}}{(1 - \delta)^{1-\delta}} \right)^{\mu_L}.
\]

Solution 2: We only show one part of this - the argument for the second is entirely similar. To show

\[
\Pr[X \geq (1 + \delta) \mu_H] \leq \left( \frac{e^\delta}{(1 + \delta)^{1+\delta}} \right)^{\mu_H},
\]
we proceed as usual through the proof steps for the Chernoff bound, i.e., starting with

$$\Pr[X \geq a] = \Pr[e^{tX} \geq e^{ta}] \leq \frac{E[e^{tx}]}{e^{ta}},$$

where $t > 0$ and $a = (1 + \delta)\mu_H$. Now we use the independence of the $X_i$ to write $E[e^{tX}] = \prod_{i=1}^n E[e^{tX_i}]$ and then observing that $E[e^{tX_i}] = (1 - p_i) + p_i e^t = 1 + p_i (e^t - 1) \leq e^{p_i (e^t - 1)}$ where $p_i = \Pr[X_i = 1]$. When we take the product we can write $\prod_{i=1}^n E[e^{tX_i}] \leq e^{(e^t - 1)\mu}$ and here we observe that we can bound this by $e^{(e^t - 1)\mu_H}$ as $\mu \leq \mu_H$. For the rest of the proof we can pretend that we are instead analyzing Poisson random variables whose mean is $\mu_H$ instead of $\mu$ and the rest of the algebra stays the same. In the end we get bounds with $\mu_H$ in place of $\mu$.

**Problem 3 [30 pts]:** Analyze the following randomized binary search procedure for the expected running time. Give an exact expression for the expected running time (number of comparisons) if you can, for the cases $x$ in the array (may also need to fix the rank) and $x$ not in the array. To use this function we will call $\text{RandomizedBinarySearch}(A, 1, n, x)$.

Assume that the function $\text{RAND}(a, b)$ called below for integers $a < b$, generates an integer uniformly at random in the range $[a, b]$.

```
Data: Array $A[1 \ldots n]$ of distinct numbers (sorted), indexes lo, hi, and number $x$
Result: The index $i \in [lo, hi]$ such that $A[i] = x$ or $-1$

$n \leftarrow hi - lo + 1$;
if $n \leq 0$ then
    return $-1$;
if $n = 1$ and $A[lo] = x$ then
    return lo;
else if $n = 1$ then
    return $-1$;
x \leftarrow RAND(lo, hi);
if $A[x] = x$ then
    return $x$;
else if $A[x] < x$ then
    return $\text{RandomizedBinarySearch}(A, r + 1, hi, x)$;
else
    return $\text{RandomizedBinarySearch}(A, lo, r - 1, x)$;

Algorithm 1: RandomizedBinarySearch($A, lo, hi, x$)
```

**Solution 3:** There are essentially 3 different cases to analyze 2 of which are similar. The cases are as follows:

(A) $x > A[n]$,
(B) $x < A[1]$,
(C) $x = A[\ell]$, for some $1 \leq \ell \leq n$.

We get the same expression for comparisons in cases (A), (B) above but something different in (C) where it now depends on $\ell$ although they are all asymptotically the same. A basic observation which directly leads to the desired expressions is the following. We illustrate it with respect to case (A) but it can be easily generalized to cases (B) and (C). From that, we can derive all the expectations.

The basic question we want to ask is: what is the probability that $x$ is compared to $A[i]$ where $1 \leq i \leq n$ is a given index. Here we are assuming $x > A[n]$. Similar to the analysis of randomized QuickSort one can see that this event happens only when the first random choice $r$ within the index range $[i \ldots n]$ is $i$ itself. If this doesn’t happen and the first random index in that range is $j > i$ then the search will recurse on the subarray $A[j + 1 \ldots n]$ and $A[i]$ will be never be compared to $x$. So, the probability of that happening is precisely $1/(n - i + 1)$. Thus in this case, as well as case (B) where the corresponding expression is $1/i$ the answer is just $1 + 1/2 + \ldots + 1/n = H_n$. For case (C) the expression is precisely $1/(i - \ell) + 1$, and so the expression in this case is, as can be easily verified is $H_{\ell} - 1 + H_{n - \ell + 1}$.

**Problem 4 [20 pts]:** Prove the following inequality by an appropriate application of the Chernoff bound. Here $m \geq 1$
is an integer.

\[ \sum_{i=2m}^{3m} \binom{3m}{i} \frac{1}{2^i} \leq \left( \frac{27e}{32} \right)^m. \]

[Hint: Consider applying the Chernoff bound to a sum of 3m Bernoulli random variables whose mean is m.]

**Solution 4:** As suggested, we consider Bernoulli random variables \( X_1, \ldots, X_{3m} \) where for each \( 1 \leq j \leq 3m \) we have \( \Pr[X_j = 1] = 1/3 \). Then the mean of \( X = \sum_{j=1}^{3m} X_j \) is \( m \). Now, we consider the probability \( \Pr[X \geq 2m] \). Since \( X \) has the Binomial distribution \( B(3m, 1/3) \) the expression for the probability is

\[ \sum_{i=2m}^{3m} \binom{3m}{i} \left( \frac{1}{3} \right)^i \left( \frac{2}{3} \right)^{3m-i} = \frac{2^{3m} \sum_{i=2m}^{3m} \binom{3m}{i}}{3^{2m}} \frac{1}{2^i}. \]

Now, we use the Chernoff bound to estimate \( \Pr[X \geq (1 + 1)E[X]] \) to get that this is at most \( \left( \frac{e}{(1+1)(1+1)} \right)^m = \left( \frac{4}{5} \right)^m \). By bringing the factor outside the summation on the LHS we get the desired inequality.

**Problem 5 [Extra credit]:** Consider an experiment which proceeds in rounds. For the first round, we have \( n \) balls, which are thrown independently and uniformly at random into \( n \) bins. After round \( i \), for \( i \geq 1 \), we discard every ball that fell into a bin by itself in round \( i \). The remaining balls are retained for round \( i + 1 \), in which they are thrown independently and uniformly at random into the \( n \) bins. We stop if no ball remains. Show that with some constant probability the number of rounds is at most \( O(\log \log n) \). [This is an easier version of a problem from the Randomized Algorithms book by Motwani and Raghavan.]

**Solution 5:** We first show that with constant probability, after a constant number of rounds the number of balls falls down to some fraction of \( n \). We call this phase (1). Then, in phase (2), with constant probability in \( O(\log \log n) \) rounds the experiment ends. We will assume that \( n \geq 2 \) otherwise the assertion is “obvious” (except that \( \log \log n \) is undefined!).

Phase (1): Let \( Z_1, Z_2, \ldots \) denote the number of balls in round 1, 2, \ldots. Remember that these are random variables. In some round, say \( l \), let there be \( m \) balls (initially \( m = n \)), i.e., \( Z_l = m \). Let \( X_i \) for \( 1 \leq i \leq n \) be a Bernoulli random variable that is 1 if bin \( i \) receives exactly 1 ball and is 0 otherwise. One can easily see that \( \Pr[X_i = 1] = m \cdot 1/n \cdot (1-1/n)^{m-1} \). Let \( X = X_1 + \ldots + X_n \). One sees easily that the number of balls going to the next round is \( m - X \) (this is a random variable).

Now \( E[X] = m(1-1/n)^{m-1} \geq m(1-1/n)^n \geq m/e^2 \) as one easily checks that \( 1-x \geq e^{-2x} \) for \( 0 \leq x \leq 1/2 \) and notice that \( 1/n \leq 1/2 \). Therefore, the expected number of balls going to next round is at most \( m(1-1/e^2) \leq m(1-1/e) = 7m/8 \). So, \( E[Z_{l+1} | Z_l] \leq 7Z_l/8 \). As such, \( \Pr[Z_{l+1} \geq \frac{15}{16} Z_l | Z_l] \leq \frac{E[Z_{l+1} | Z_l]}{15/16 Z_l} \leq 14/15 \). Thus all the events \( Z_2 \leq 15/16 Z_1, Z_3 \leq 15/16 Z_2, \ldots \) happen with at least a constant probability which is at least \( 1/15 \). After a constant number of rounds \( t \) we have that with some constant probability (which is at least \( (1/15)^t \)), \( Z_{t+1} \leq Z_t/128 \). Now, phase (2) starts.

Phase (2): For this phase we start the rounds again from 1, i.e., our numbering starts from 1 again for convenience of writing. Thus we denote the number of balls in first round of phase (2) as \( Z_1 \), those in second round as \( Z_2 \) etc. We recall that \( Z_1 \leq n/128 \). First, we derive another estimate for \( E[Z_{l+1} | Z_l] \). Using our earlier calculation, we have that:

\[ E[Z_{l+1} | Z_l] = Z_l \left( 1 - \left( 1 - \frac{1}{n} \right)^{Z_l} \right) \leq Z_l \left( 1 - \left( 1 - \frac{1}{n} \right)^{Z_l} \right) \leq Z_l \left( 1 - \left( 1 - Z_l/n \right) \right) \leq Z_l^2/n, \]

where we have used the inequality \( 1 - ax \leq (1-x)^a \) for \( 0 \leq x \leq 1 \). Consider the following sequence defined recursively:

\[ b_1 = 128, b_2 = b_1^2/2^2, b_3 = b_2^2/2^3, \ldots, b_i = b_{i-1}^2/2^i, \ldots \]

By writing \( b_i = 2^{x_i} \) we can see that \( x_1, x_2, \ldots \), satisfy the recurrence:

\[ x_1 = 7, x_2 = 2x_1 - 2, \ldots, x_i = 2x_{i-1} - i, \ldots \]

One easily shows that \( x_i \) increase exponentially, for example one can show by induction that \( x_i \geq 4 \left( \frac{3}{4} \right)^i \) for all \( i \geq 1 \). Let \( E_i \) denote the event that \( Z_i \leq n/b_i \).

Now, from our analysis above \( E[Z_{l+1} | Z_l] \leq Z_l^2/n \). Suppose \( a_1, \ldots, a_i \) are any integers with \( a_i \leq n/b_i \). Then \( E[Z_{l+1} | Z_l = a_1 \land Z_2 = a_2 \land \ldots \land Z_i = a_i] \leq \frac{n}{b_i} \). Since this holds for each possible combination of integers we have that:

\[ E[Z_{l+1} | E_1 \land E_2 \land \ldots \land E_l] \leq \frac{n}{b_i^l}. \]
Therefore,
\[
\Pr[E_{i+1} | E_1 \land E_2 \land \ldots \land E_i] = 1 - \Pr[(Z_{i+1} > n/b_{i+1}) | E_1 \land E_2 \land \ldots \land E_i] \\
\geq 1 - \frac{\mathbb{E}[Z_{i+1} | E_1 \land E_2 \land \ldots \land E_i]}{\frac{n}{b_{i+1}}} \geq 1 - \frac{n}{b_{i+1}^2} \geq 1 - \frac{1}{2^{i+1}}.
\]

Now, we are almost done. Let \(E_0\) denote the event that after \(t\) rounds we will have that the number of balls is at most \(n/128\) (\(t\) was a constant that appeared earlier). We have seen that \(\Pr[E_0] = \Omega(1)\). Now let \(E_1, E_2, \ldots\), denote the events as defined above for phase (2). Clearly, \(\Pr[E_1 | E_0] = 1\) and as we showed, \(\Pr[E_2 | E_1 \land E_0] \geq (1 - \frac{1}{2^2})\). In general we will have, \(\Pr[E_i | E_{i-1} \land \ldots \land E_1 \land E_0] \geq (1 - 1/2^i)\). Therefore the probability, \(\Pr[E_0 \land E_1 \ldots \land E_r] = \Pr[E_0] \Pr[E_1 | E_0] \Pr[E_2 | E_1 \land E_0] \ldots \Pr[E_r | E_{r-1} \land \ldots \land E_0] \geq \Omega(1)(1 - \frac{1}{2^2})(1 - \frac{1}{2^3}) \ldots \geq 1 - \frac{1}{2^2} - \frac{1}{2^3} \ldots \geq 1/2\). However, if all the events \(E_0, E_1, E_2, \ldots, E_r\) happen and \(r \geq O(\log \log n)\) the experiment must have ended (recall that \(b_i \geq 2^{1/(3/2)^i}\)). In total thus we took a constant plus an additional number of rounds bounded by \(O(\log \log n)\).