Warm-up

Problem 1. Generalise Eq. (22) of the lecture notes to $m > n$ bins, to compute directly

$$\mathbb{E}[\text{empty bins after } m \text{ balls}]$$

and solve for m to get this expectation to be at most $1/2$. Show you retrieve the $\Theta(n \log n)$ bound.

Problem 2. Use Chebyshev's inequality to bound the probability that $m(n)$, the number of balls needed to hit every bin at least once, is greater than $\alpha n \ln n$ (for $\alpha > 1$).

Problem solving

Problem 3. Let $c > 0$ some constant to be determined later. We want to show that, when throwing $m = cn \ln n$ balls into n bins (uniformly and independently at random), with high probability *every* bin has $\Theta(\ln n)$ balls. That is, with probability at least $1 - o(1)$ we have both that minimum load at least $c_1 \ln n$ and the maximum load at most $c_2 \ln n$, for some constants $0 < c_1 < c_2$.

a) Let L_i the load of bin i, for a fixed $1 \leq i \leq n$. Compute $\mathbb{E}[L_i]$ and $\mathrm{Var}[L_i]$.

b) Use Chebyshev to bound

$$\Pr\left[L_i \notin \left[\frac{1}{2}c \ln n, \frac{3}{2}c \ln n\right] \right]$$

Is it enough to conclude?

c) Show, using a Chernoff bound, that

$$\Pr\left[L_i \notin \left[\frac{1}{2}c \ln n, \frac{3}{2}c \ln n\right] \right] \leq \frac{2}{n^{c/12}}$$

(What does Hoeffding's give?)

d) Pick a suitable value of $c > 0$ to conclude that

$$\Pr\left[\forall i, L_i \in \left[\frac{1}{2}c \ln n, \frac{3}{2}c \ln n\right] \right] \geq 1 - \frac{2}{n}$$

Problem 4. Suppose that instead of throwing m balls into n bins where each bin has the same probability $1/n$, now bin i has probability p_i, where $\sum_{i=1}^{n} p_i = 1$. We will see this vector of probabilities as a vector $p \in [0,1]^n$.

a) As a function of p, what is the probability to get a collision when $m = 2$?

b) What is the expected number of collisions, $\mathbb{E}[c(m,n)]$ when throwing $m \geq 2$ balls with replacement?

(If you want to go further, try to compute or bound the variance as a function of $\|p\|_2, \|p\|_3, m$. It is not easy.)

Problem 5. (Guided tutorial) Consider the "best of two choices" strategy: when throwing ball t, we select *two* bins independently and uniformly at random, and put the ball in the least full of the two (breaking ties arbitrarily). We will (not) prove the following result stated in the lecture:

> *(The Power of Two Choices)* The expected maximum load $\hat{L}(n)$ when throwing independently n balls into n bins using the "best of two choices" strategy satisfies
>
> $$\hat{L}(n) \leq \log\log n + O(1)$$

but at least give a sketch of proof that should provide some intuition. (Compare it to the $O(\frac{\log n}{\log\log n})$ bound when using only "one choice"! This is a very useful and surprising result: see `https://www.eecs.harvard.edu/~michaelm/postscripts/handbook2001.pdf` for a survey and applications.)

a) Denote by B_i, for $1 \leq i \leq n$, the number of bins that have at least i balls after throwing n balls according to the best-of-two-choices strategy. Explain why $B_2 \leq \frac{n}{2}$.

b) Let B'_i (for $i \geq 3$) be the number of balls which, *at the time they were thrown and then added to a bin*, were the i-th or more in their chosen bin. Argue that $B_i \leq B'_i$.

c) Explain why, at any step $1 \leq t \leq n$ (when we threw the t-th ball), there were at most B_i bins with at least i balls. Deduce that the probability that ball t chooses a bin containing already at least $i \geq 2$ balls is at most $(B_i/n)^2$.

d) Show that $\mathbb{E}\left[B'_{i+1}\right] \leq \frac{B_i^2}{n}$.

e) Ignoring all dependencies for now (dependence between events, things are equal to their expectation, etc.), explain how this hints at a recurrence relation of the form

$$B_{i+1} \leq \frac{B_i^2}{n} \qquad \text{("Wishful thinking")}$$

Solve this recurrence relation: what upper bound for B_i ($i \geq 2$) would this give?

$$B_i \leq \frac{n}{2^{2^{i-2}}}$$

f) Conclude by given the maximum i (according to this "wishful thinking bound") for which $B_i \geq 1$. Explain how that would imply the result.

This *would* conclude the proof assuming everything behaves exactly as expected, to get the above recurrence relation. To make this formal, we would need to argue that each B_i concentrates tightly around its expectation (and keep track of the small deviations around them), and to do that we would need a bit more than Chernofff/Hoeffding since $B_1, \ldots, B_n$ are very much dependent. There *are* ways to handle these dependencies, but they are beyond the scope here.

a) To conclude: *why stop at two choices?* Going above the same outline as above, sketch why, we $d \geq 2$ choices instead, we would get an expected max load of

$$\log_d \log n + O(1) = \frac{\log \log n}{\log d} + O(1)$$

that is, not a breathtaking improvement.

Problem 6. Let's get back to throwing n balls into n bins independently and uniformly at random. Show that, for large enough n, the expected number of empty bins approaches n/e, where $e \approx 2.718$ is the base of the natural logarithm.

Problem 7. You have been playing the Australian 1st Division lottery, which requires you to guess correctly 6 numbers out of 45 to win. You have consistently lost, and are suspecting the lottery is rigged.

a) If the lottery was fair, what is the probability that your ticket (a single ticket) wins? Call this probability p.

b) Assuming the total prize is $\$30,000,000$ and a ticket is \$0.60, what is the expected reward if you play one ticket? 100 (different) tickets?

c) You suspect that half of the possible outcomes actually never show up, due to an issue in the lottery design or some foul play. Of course, you don't have much to back this up, and have no idea *which* half of the outcomes would still show up. As a function of p (in big-Oh notation), how many tickets would you need to play before having any statistical evidence to prove or disprove your suspicion?

Advanced

Problem 8. *(Poissonization.)* In the setting of Problem 4, suppose that instead of throwing m balls, we first draw the value $M \sim \text{Poi}(m)$, and then throw M independent balls into the n bins. Let $N_1, \ldots, N_n$ the number of balls falling into bins $1, 2, \ldots n$ respectively.

a) Show that $N_1, \ldots, N_n$ are independent.

b) Rewrite the number of collisions $\tilde{c}(m, n)$ as a function of $N_1, \ldots, N_n$.

c) Compute $\mathbb{E}[\tilde{c}(m, n)]$.

d) Compute $\text{Var}[\tilde{c}(m, n)]$.

e) Conclude by giving a bound on the number m sufficient to approximate $\|p\|_2$ to within a factor 2 with probability at least $9/10$.

Problem 9. Go over the MGF-based proof that $L(n) \leq \frac{2\ln n}{\ln\ln(en)}$ from the lecture notes. Using the same approach, show that if $X_1, \ldots, X_n$ are (not necessarily independent) Gaussian random variables with mean zero and variance σ^2, then

$$\mathbb{E}\left[\max_{1\leq i\leq n} X_i\right] \leq \sqrt{2\sigma^2 \ln n}\,.$$

As a corollary, show that

$$\mathbb{E}\left[\max_{1\leq i\leq n} |X_i|\right] \leq \sqrt{2\sigma^2 \ln(2n)}\,.$$