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.

Solution 1. By linearity of expectation,

$$\mathbb{E}\left[\text{empty bins after } n \text{ balls}\right] = \sum_{i=1}^{n} \Pr[\, i\text{-th bin empty}\,] = n\left(1 - \frac{1}{n}\right)^{m} \leq n \cdot e^{-m/n}.$$

Solving

$$n \cdot e^{-m/n} \leq \frac{1}{2}$$

for m gives, taking logarithms, $\ln(2n) \leq \frac{m}{n}$, that is $m = \Omega(n \log n)$.

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 $\alpha > 1$).

Solution 2. Denote T_i as the r.v. that counts how many balls to throw to hit the i-th bin after the $(i-1)$ bins are filled.

$$m(n) = T_1 + T_2 + \cdots + T_n.$$

To use Chebyshev's inequality, we need to know the expectation and variance. T_i follows a geometric distribution, whose mean and variance are $\frac{1}{p_i}$ and $\frac{1-p_i}{p_i^2}$ (resp.), and here $p_i = \frac{n-i+1}{n}$. Following the lecture notes, we have

$$\mathbb{E}[m(n)] = \sum_{i=1}^{n} \mathbb{E}[T_i] = \sum_{i=1}^{n} \frac{n}{n-i+1} = nH_n = n \log n + O(n) \leq 2n \log n.$$

$$\sigma^2 = \text{Var}[m(n)] \leq \frac{\pi^2}{6} n^2.$$

By Chebyshev's inequality, let $t = (\alpha - 2) n \log n$,

$$\Pr[m \geqslant \alpha \cdot n \log n] \leq \Pr[|m - \mathbb{E}[m]| \geqslant t] \leq \frac{\text{Var}[m]}{t^2},$$

and thus

$$\Pr[m \geqslant \alpha n \log n] \leq \frac{\frac{\pi^2}{6} n^2}{((\alpha - 2) \cdot n \log n)^2} = O\left(\frac{1}{\log^2 n}\right)$$

Problem 3. Show that the expected number of balls $M_2(n)$ we need to throw before each of the n bins contains at least *two* balls is $\Theta(n \log n)$.

Solution 3. First, $M_2(n) \geq M(n)$, since to get at least two balls in each bin, we need to first have at least one ball in each bin. This shows that $M_2(n) = \Omega(n \log n)$.

For the upper bound, we will show that $M_2(n) \leq 2 \cdot M(n)$: this will prove $M_2(n) = O(n \log n)$. To see why this inequality holds, consider the following: wait until each bin has at least one ball (this takes expected time $M(n)$), and then "forget" the number of balls in each bin, and wait until each bin receives at least one new ball (i.e., "restart from scratch"). This second stage also take $M(n)$ balls in expectation. At the end of the two stages, clearly, each bin contains at least $1 + 1 = 2$ balls: this may have happens much before, but at the very least it is the case after waiting until stage 2 is completed.

Problem solving

Problem 4. 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 $\operatorname{Var}[L_i]$.

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

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

Solution 4.

a) As $L_i \sim \operatorname{Bin}\left(m, \frac{1}{n}\right)$, we get $\mathbb{E}[L_i] = \frac{m}{n} = c \ln n$ and $\operatorname{Var}[L_i] = \frac{m}{n}\left(1 - \frac{1}{n}\right) \leq \frac{m}{n} = c \ln n$.

b) Since

$$\Pr\left[L_i \notin \left[\frac{1}{2}c\ln n, \frac{3}{2}c\ln n\right] \right] = \Pr\left[|L_i - \mathbb{E}[L_i]| \geqslant \frac{1}{2}c\ln n \right],$$

by Chebyshev, using the bound on $\mathrm{Var}[L_i]$ above have

$$\Pr\left[L_i \notin \left[\frac{1}{2}c\ln n, \frac{3}{2}c\ln n\right] \right] \leq \frac{4c\ln n}{c^2\ln^2 n} = \frac{4}{c\ln n}$$

This is small (for large enough n), but not small enough for our purposes: as we want to bound that probability that *any* of the L_i's is large, we need to take a union bound over all n of them. That would lead to a bound of $n \cdot \frac{4}{c\ln n}$, which is vacuous (completely useless): this is greater than 1!

c) Apply the Chernoff bound (L_i is a sum of independent Bernoulli draws)

$$\Pr\left[|L_i - c\log n| \geqslant \frac{1}{2}c\log n\right] \leq 2\exp\left(-\frac{c\log n}{12}\right)$$

and so

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

d) We note that

$$\Pr\left[\forall i, L_i \in \left[\frac{1}{2}c\log n, \frac{3}{2}c\log n\right]\right] = 1 - \Pr\left[\exists i, L_i \notin \left[\frac{1}{2}c\log n, \frac{3}{2}c\log n\right]\right].$$

By the union bound,

$$\begin{aligned}\Pr\left[\exists i, L_i \notin \left[\frac{1}{2}c\log n, \frac{3}{2}c\log n\right]\right] &\leq \sum_{i=1}^{n}\Pr\left[L_i \notin \left[\frac{1}{2}c\log n, \frac{3}{2}c\log n\right]\right] \\ &\leq n \cdot \frac{2}{n^{c/12}} = \frac{2}{n^{\frac{c}{12}-1}}.\end{aligned}$$

Choosing $c = 24$ suffices.

Problem 5. 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.)

Solution 5.

a) Denote X_i as the indicator r.v. for a collision at i-th bin. Using independence between draws:

$$\Pr[X_i = 1] = \Pr\left[\text{first ball hits } i\text{-th bin}\right] \cdot \Pr\left[\text{second ball hits } i\text{-th bin}\right] = p_i^2.$$

$$\Pr\left[\text{collision}\right] = \sum_{i=1}^{n} \Pr[X_i = 1] = \sum_{i=1}^{n} p_i^2 = \|p\|_2^2,$$

where $\|\cdot\|_2$ denotes the 2-norm of a vector.

b) Write Y_j for the indicator of a collision between draws $1 \le c_j < d_j \le m$: there are $\binom{m}{2}$ such indicators. Using linearity of expectation,

$$\mathbb{E}[c(m,n)] = \sum_{j=1}^{\binom{m}{2}} \mathbb{E}[Y_j] = \sum_{j=1}^{\binom{m}{2}} \Pr[Y_j] = \frac{m(m-1)}{2}\|p\|_2^2$$

which behaves like $\Theta(m^2\|p\|_2^2)$. As a sanity check, when p is the uniform distribution $(1/n, 1/n, \ldots, 1/n)$, we have $\|p\|_2^2 = \sum_{i=1}^{n} \frac{1}{n} = \frac{1}{n}$, and we retrieve the result seen in class.

Problem 6. (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
>
> $$\hat{L}(n) \le \log\log n + O(1)$$

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

b) Let B_i' (for $i \ge 3$) 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 $B_i \le B_i'$.

c) Explain why, at any step $1 \le t \le n$ (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 $i \ge 2$ balls is at most $(B_i/n)^2$.

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

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.

Solution 6.

a) $B_2 > \frac{n}{2}$ by definition would mean that there are more than $n/2$ bins with more than 2 balls. So strictly more than $(n/2) \cdot 2 = n$ balls are in the bins, which exceeds n the number of balls actually thrown.

b) For every bin that has at least i balls after all n balls are thrown, we look at how its last ball was thrown. Its last ball will be the i-th ball into the bin, which gets counted into B'_i. So every bin counted by B_i has a ball that is a member to be counted by B'_i and thus $B_i \leq B'_i$.

c) Since B_i is computed at the end (after throwing all n balls), and the process does not remove balls from bins – at every step the balls in each bin only increase. So the number of bins with at least i balls is bounded by B_i.
Denote $B_{i,t}$ the number of bins with at least i balls at time t and $B_{i,t} \leq B_i$.

$$\Pr\left[t \text{ ball chooses bin with at least } i \text{ balls}\right] = \left(\frac{B_{i,t}}{n}\right)^2 \leq \left(\frac{B_i}{n}\right)^2.$$

d) Fix some $i \geqslant 3$ and B_i (conditioning on B_i being some number). Denote $X_1, \ldots, X_n$ the indicator for the j-th ball when thrown, were the $(i+1)$-th or more in their chosen bin.

$$\mathbb{E}[B'_{i+1}] = \mathbb{E}\left[\sum_{j=1}^{n} X_j\right] = \sum_{j=1}^{n} \Pr[X_j] \leq n \cdot \left(\frac{B_i}{n}\right)^2 = \frac{B_i^2}{n}.$$

e) In expectation, we see the following,

$$\mathbb{E}[B_{i+1}] \leq \mathbb{E}[\mathbb{E}[B'_{i+1} \mid B_i]] \leq \mathbb{E}\left[\frac{B_i^2}{n}\right].$$

Now, handwaving (this is *not* a full, valid proof), we assume that "things behave exactly like their expectation:" $B_i \approx \mathbb{E}[B_i]$. We proceed by induction with $B_2 \leq \frac{n}{2}$ and $B_{i+1} \leq \frac{B_i^2}{n}$.
For $k = 2$, we have $B_2 \leq \frac{n}{2}$.
Now suppose $B_k \leq \frac{n}{2^{2^{k-2}}}$, then

$$B_{k+1} \leq \frac{B_k^2}{n} \leq \frac{\left(\frac{n}{2^{2^{k-2}}}\right)^2}{n} \leq \frac{\frac{n^2}{2^{2^{k-1}}}}{n} = \frac{n}{2^{2^{k-1}}}.$$

f) Solving the inequality:

$$1 \leq \frac{n}{2^{2^{i-2}}} \Rightarrow 2^{2^{i-2}} \leq n \Rightarrow 2^{i-2} \leq \log_2 n \Rightarrow i \leq 2 + \log_2 \log_2 n.$$

This tells us that the largest number i for which there is at least *one* bin with at least i balls is not more $\log_2 \log_2 n + 2$. That's just a contrived way to say that the maximum load is at most $\log_2 \log_2 n + 2$, since the maximum load is the maximum number i of balls that can be found in at least one bin.

g) For $d \geq 2$ choices, one can go through the same steps as above to see that the recurrence becomes

$$B_{i+1} \leq \frac{B_i^d}{n^{d-1}}.$$

Solving it gives the claimed bound.

Problem 7. 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.

Solution 7. See Solution to Problem 1. It only remains to show (if you do not want to take it for granted) that

$$\lim_{n \to \infty} (1 - 1/n)^n = e^{-1}$$

or, equivalently (since $(1-1/n)^n = e^{\frac{\ln(1-1/n)}{(1/n)}}$, that

$$\lim_{n\to\infty} \frac{\ln(1-1/n)}{(1/n)} = -1$$

Setting $f(x) = \ln(1-x)$ (such that $f(0) = 0$, this would be showing that $f'(0) = \lim_{x\to 0} \frac{f(x)-f(0)}{x} = -1$, which could can easily check by differentiating f.

Problem 8. 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 $\$1,000,000$ 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?

Solution 8.

a) All possible guesses $\binom{45}{6}$. Suppose the draw is uniformly at random,

$$\Pr[\text{win}] = \frac{1}{\binom{45}{6}} = p.$$

b) Denote X as the random variable for how much money one would win/lose from one ticket.

$$\Pr[X = 1{,}000{,}000 - 0.6] = \frac{1}{\binom{45}{6}} = p \text{ and } \Pr[X = -0.6] = 1 - \frac{1}{\binom{45}{6}} = 1 - p.$$

$$\mathbb{E}[X] = 1 \times 10^6 p - 0.6 = 1 \times 10^6 \times \frac{1}{\binom{45}{6}} - 0.6 \approx -0.48.$$

Playing 100 games, by linearity of expectation, one loses ≈ 48 dollars in expectation. *Note: this is **very** realistic. In real life, the expected gain is negative!*

c) If you do not observe a collision, all you see is a sequence of unique numbers: which reveals absolutely nothing about what the underlying distribution of outcomes is. Put differently, *conditioned on not seeing a collision*, what you observe has exactly the same probability under (1) the uniform distribution over the $\binom{45}{6}$ outcomes, and (2) a distribution only uniform over an (unknown) subset of *half* these outcomes. So to have a chance to conclude *anything*, you need to make enough observations to have a decent chance to observe a collision (in at least one of the two cases) – which, by the birthday paradox seen in class, will be after playing $\Omega(\sqrt{n})$ tickets, where here $n = 1/p = \frac{1}{2}\binom{45}{6}$.

Advanced

Problem 9. *(Poissonization. ⋆⋆)* In the setting of Problem 5, 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 $N_1, \dots, N_n$ the number of balls falling into bins $1, 2, \dots n$ respectively.

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

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

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 $9/10$.

Solution 9.

a) We will show (a stronger statement) that $N_j \sim \mathrm{Poi}(mp_j)$ independently. Indeed, we show that for every possible $k_1, \ldots, k_j$ satisfying the following,

$$\sum_{j=1}^{n} k_j = \sum_{j=1}^{n} N_j = M = k,$$

its probability mass function can be written as follows:

$$\begin{aligned}
\sum_{k_1,\ldots,k_n:\sum k_j = k} \prod_{j=1}^{n} \Pr[N_j = k_j] &= \sum_{k_1,\ldots,k_n:\sum k_j = k} \prod_{j=1}^{n} \frac{(mp_j)^{k_j} e^{-mp_j}}{k_j!} \\
&= \sum_{k_1,\ldots,k_n:\sum k_j = k} m^{\sum_{j=1}^{n} k_j} e^{-\sum_{j=1}^{n} mp_j} \prod_{j=1}^{n} \frac{p_j^{k_j}}{k_j!} \\
&= m^k e^{-m} \sum_{k_1,\ldots,k_n:\sum k_j = k} \prod_{j=1}^{n} \frac{p_j^{k_j}}{k_j!} \\
&= \frac{m^k e^{-m}}{k!}, \\
&= \Pr[M = k],
\end{aligned}$$

where the second last equality is obtained via the multinomial theorem (note: this may not be the most elegant proof)
Some other proof references: see, e.g.,

- *https://people.csail.mit.edu/ronitt/COURSE/F20/Handouts/scribe14.pdf*
- *https://math.stackexchange.com/a/1355399/75808.*

b) The number of collision in the i-th bin is simply $\binom{N_i}{2}$. We can then rewrite

$$\tilde{c}(m,n) = \sum_{i=1}^{n} \binom{N_i}{2} = \sum_{i=1}^{n} \frac{N_i^2 - N_i}{2}.$$

c) Using the expression for the moments of a Poisson random variable (to be computed, or available in a textbook or on Wikipedia!), we then get

$$\mathbb{E}[\tilde{c}(m,n)] = \sum_{i=1}^{n} \mathbb{E}\left[\frac{N_i^2 - N_i}{2}\right] = \sum_{i=1}^{n} \frac{(mp_i)^2}{2} = \frac{m^2}{2}\|p\|_2^2.$$

d) Similarly, for the variance, thanks to independence we have

$$\operatorname{Var}[\tilde{c}(m,n)] = \sum_{i=1}^{n} \operatorname{Var}\left[\binom{N_i}{2}\right] = \frac{1}{4}\sum_{i=1}^{n} \operatorname{Var}\left[N_i^2 - N_i\right]$$

and so this boils down to computing

$$\operatorname{Var}\left[N_i^2 - N_i\right] = \mathbb{E}\left[N_i^4 - 2N_i^3 + N_i^2\right] - m^2 p_i^2$$

for $N_i \sim \text{Poisson}(mp_i)$. which can be done by expanding the square and a sequence of (cumbersome) series manipulations. This will give

$$\operatorname{Var}\left[N_i^2 - N_i\right] = m^4 p_i^4 + 4m^3 p_i^3 + m^2 p_i^2$$

and so, summing over i,

$$\operatorname{Var}[\tilde{c}(m,n)] = \frac{1}{4}m^4\|p\|_4^4 + m^3\|p\|_3^3 + \frac{1}{4}m^2\|p\|_2^2$$

c) By Chebyshev's inequality, setting $X = \sqrt{\frac{2}{m^2}\tilde{c}(m,n)}$ as our estimator (which satisfies $\mathbb{E}\left[X^2\right] = \|p\|_2^2$ based on what we did above)

$$\begin{aligned}
\Pr[\, X \notin [\|p\|_2/2, 2\|p\|_2]\,] &= \Pr\Big[\, X^2 \notin [\|p\|_2^2/4, 4\|p\|_2^2]\,\Big] \\
&\leq \Pr\left[\, |X^2 - \mathbb{E}[X^2]| \geq \frac{3}{4}\|p\|_2^2 \right] \\
&\leq \frac{\operatorname{Var}[X^2]}{(3/4)^2\|p\|_2^4} = \frac{(2/m^2)^2 \operatorname{Var}[\tilde{c}(m,n)]}{(3/4)^2\|p\|_2^4} \\
&= \frac{64}{9m^4\|p\|_2^4}\left(\frac{1}{4}m^4\|p\|_4^4 + m^3\|p\|_3^3 + \frac{1}{4}m^2\|p\|_2^2\right) \\
&= \frac{16}{9}\frac{\|p\|_4^4}{\|p\|_2^4} + \frac{64}{9m}\frac{\|p\|_3^3}{\|p\|_2^4} + \frac{1}{4m^2\|p\|_2^2}
\end{aligned}$$

Now, for the whole thing to be at most $1/10$, it's enough to choose m such that each of the three terms is at most $1/30$. For that, the last term implies we should make sure $m \geq \frac{1}{2\sqrt{30}\|p\|_2}$, the second will require $m \geq \frac{640\|p\|_3^3}{3\|p\|_2^4}$, and the first... is annoying, as we have no control over it! It does not depend on m.... so what can we do? Looks like we are in bad shape...

First, let's simplify our task. By monotonicity of ℓ_q norms (for vector norms), we have $\|p\|_4 \leq \|p\|_2$ and $\|p\|_3 \leq \|p\|_2$, and so we can bound our variance as

$$\operatorname{Var}[\tilde{c}(m,n)] \leq \frac{1}{4}m^4\|p\|_2^4 + m^3\|p\|_2^3 + \frac{1}{4}m^2\|p\|_2^2$$

at least we got rid of the annoying 3- and 4-norms... if we apply Chebyshev with this (weaker, but simpler) variance bound, we get

$$\Pr[\, X \notin [\|p\|_2/2, 2\|p\|_2]\,] \leq \frac{16}{9}\frac{\|p\|_2^4}{\|p\|_2^4} + \frac{64}{9m}\frac{\|p\|_3^2}{\|p\|_2^4} + \frac{1}{4m^2\|p\|_2^2} = \frac{16}{9} + \frac{(64/9)}{m\|p\|_2} + \frac{1}{4m^2\|p\|_2^2}\,.$$

The first term is still very bad, because it does not depend on m (and is definitely bigger than 1). But here's a simple trick: instead of using X^2 as our estimate, take $T = 10$ (for instance) independent copies $X_1^2, \ldots, X_T^2$ of X^2, and use their average $Y = \frac{1}{T}(X_1^2, \ldots, X_T^2)$ as our estimate. The expectation doesn't change (we just took an average), but the variance decreases by $T^2 = 100$! That will take care of the first term, and only cost us $T = 100$ times as many samples... Now we get

$$\Pr\left[\,\sqrt{Y} \notin [\|p\|_2/2, 2\|p\|_2]\,\right] \leq \frac{1}{100}\left(\frac{16}{9} + \frac{(64/9)}{m\|p\|_2^2} + \frac{1}{4m^2\|p\|_2^2}\right)$$

The first term is now always good: $16/900 < 1/30$. The second term will be good (less than $1/30$) for $m = \Theta(1/\|p\|_2)$. The third term will also be good (less than $1/30$) for $m = \Theta(1/\|p\|_2)$. So all together, it suffices to take $m = \Theta(1/\|p\|_2)$ samples to succeed with probability at least $9/10$.

Last detail: but we don't know $\|p\|_2$, that's the whole point! How do we choose m? Well, one can show that $\|p\|_2 \geq 1/\sqrt{n}$ always (try it), so it's always enough to take $\boxed{m = O(\sqrt{n})}$ samples...

Problem 10. 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

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

Solution 10. Let $X_1, \ldots, X_n$ be $\mathcal{N}(0, \sigma^2)$ and they do not have to be independent.

$$\begin{aligned}
\mathbb{E}[\max_{1\le i\le n} X_i] &= \frac{1}{t}\mathbb{E}[\max_{1\le i\le n} tX_i] \\
&= \frac{1}{t}\mathbb{E}[\max_{1\le i\le n} \ln(\exp(tX_i))] \\
&= \frac{1}{t}\mathbb{E}[\log(\max_{1\le i\le n} \exp(tX_i))] && \text{(monotonicity of } \log(\cdot)) \\
&\le \frac{1}{t}\mathbb{E}\left[\log\left(\sum_{i=1}^{n} \exp(tX_i)\right)\right] \\
&\le \frac{1}{t}\ln\left(\mathbb{E}\left[\sum_{i=1}^{n} \exp(tX_i)\right]\right) && \text{(Jensen)} \\
&= \frac{1}{t}\ln(n\mathbb{E}[\exp(tX_1)]) && \text{(linearity)} \\
&= \frac{1}{t}\ln\left(n\exp\left(\frac{1}{2}\sigma^2 t^2\right)\right) = \frac{\ln n}{t} + \frac{1}{2}\sigma^2 t \le \sqrt{2\sigma^2 \ln n}.
\end{aligned}$$

Note that,

$$\max_{1\le i\le n} |X_i| = \max_{1\le i\le n} \max(X_i, -X_i).$$

Of course, X_i and $-X_i$ are not independent, but to apply the previous result they do not need to be! So this problem can be reduced to $Y_1, \ldots, Y_{2n} \sim \mathcal{N}(0, \sigma^2)$ and a max over them. Applying the previous bound to this new problem with size of $2n$, we conclude the proof.