

COS 330: Great Ideas in Theoretical Computer Science

Fall 2026

Precept 5

Having a Ball

PROBLEM-SOLVING SESSION

Today is a **problem-solving session**. There is a single problem, which finishes the analysis of fully random hashing from lecture: we will show that the maximum load is $O(\ln q / \ln \ln q)$, improving on the $O(\ln q)$ bound we proved in class. The problem is split into small guided steps; work through them on your own or in a small group, and ask your TA when you get stuck. Solutions will be shared after precept.

Problem 1: Balls and Bins

We use the setting and notation of lecture 8. Let U be the universe of possible pages, fix a set $S \subseteq U$ of $|S| = m$ pages, and let $h : U \rightarrow [q] = \{0, 1, \dots, q-1\}$ be a *fully random* hash function: each value $h(x)$ is chosen independently and uniformly in $[q]$. The *load* of a location $i \in [q]$ is

$$X_i = |\{x \in S : h(x) = i\}|,$$

the number of pages of S that hash to i . As in lecture, assume that $m \leq q$, and also that $q \geq 3$. Throughout, $\ln$ denotes the natural logarithm.

In lecture we used a Chernoff bound and a union bound to show that, with probability at least $1 - e^{-100}$, every location has load at most $1 + 2 \ln q + 200$. We also claimed that the right behavior is actually $\Theta(\ln q / \ln \ln q)$. In this problem you will prove the upper bound: with probability at least $1 - 1/q$, every location has load less than

$$\left\lceil \frac{4 \ln q}{\ln \ln q} \right\rceil.$$

(This is a classic problem called the *balls and bins* problem: throw m balls into q bins independently and uniformly at random, and ask how full the fullest bin is.)

Why Do We Need a New Argument?

The Chernoff bound we used in lecture says that the probability that a location has load at least $1 + 2L$ is at most about e^{-L} . In other words, if k is the load, then the probability decays like $e^{-k/2}$. For a union bound over q locations we need each location to fail with probability much less than $1/q$, and $e^{-k/2} \leq 1/q$ forces $k \geq 2 \ln q$. So no matter how carefully we choose the

constants, that bound can never give anything better than $\Theta(\ln q)$. To do better, we need a tail bound that decays *faster* than exponentially in k – and we will get one by counting directly.

We are going to start by considering some arbitrary integer $k \geq 1$. This is a parameter that we will use in the analysis and will eventually set to a specific value, after part (c), to get the final bound.

(a) Fix a location $i \in [q]$ and a set $T \subseteq S$ of exactly k pages. What is the probability that every page in T is mapped to location i ?

Recall that h is fully random, so the values $h(x)$ for $x \in T$ are independent and uniform in $[q]$.

(b) Show that for every location $i \in [q]$,

$$\Pr[X_i \geq k] \leq \binom{m}{k} q^{-k},$$

and conclude, using $m \leq q$, that $\Pr[X_i \geq k] \leq 1/k!$.

Proof Guide

1. Think about each group of k pages separately: part (a) tells you how likely it is that one particular group all lands at location i .
2. Argue that whenever location i has load at least k , some group of k pages must all have landed there.
3. Use this to relate the event you care about to the events from part (a), and then apply the union bound.
4. Count how many groups of k pages there are.
5. For the second part, use that $\binom{m}{k} \leq \frac{m^k}{k!}$, and then use the assumption relating m and q .

(c) Show that $\Pr[X_i \geq k] \leq (e/k)^k$. How does this compare to the tail bound we got from Chernoff in lecture, which decays like $e^{-k/2}$?

Proof Guide

1. Start from the bound you proved in part (b), and use that $k! \geq (k/e)^k$ for every integer $k \geq 1$ (you may use this without proof).
2. For the comparison, rewrite your new bound as an exponential, and compare how fast its exponent grows with the exponent in the Chernoff bound.

Now it's time to choose the value of k . Set

$$k = \left\lceil \frac{4 \ln q}{\ln \ln q} \right\rceil.$$

For this choice of k , one can show that $(e/k)^k \leq q^{-2}$. Proving this only takes some calculations and simplifications of logarithms, so we will skip it: just assume that it is true. Combined with part (c), it gives

$$\Pr[X_i \geq k] \leq q^{-2} \quad \text{for every location } i \in [q].$$

(d) Conclude that with probability at least $1 - 1/q$, every location has load less than k , where $k = \lceil 4 \ln q / \ln \ln q \rceil$ as above. In particular, the maximum load is $O(\ln q / \ln \ln q)$ with probability at least $1 - 1/q$.

(Note that you can solve this with exactly the same argument we used in lecture, but now with the improved bound above.)
