1 Two Bins

TODO: This section is going to be super epic, but is not quite true yet. In this section, we consider the problem of hashing n elements into 2 bins. We give some bounds on the maxload, and propose a conjecture of a tight concentration problem for the maxload in the 2 bin setting that, if proven could be helpful in reducing the n bin hashing problem to this simpler problem.

Definition 1. In ***Uniform Two Bin*** LH, denoted UH/2, we hash elements from a universe $[p]$ via hash functions parameterized by $a \in [p] \setminus \{0\}, b \in [p]$ defined as

$$h_{a,b}(x) = \left\lfloor \frac{\mathfrak{m}_p(ax+b)}{p/2} \right\rfloor .$$

In ***Multiplicative Two Bin*** LH, denoted MH/2, we hash elements from a universe $[p]$ via hash functions parameterized by $a \in [p] \setminus \{0\}$, defined by

$$h_a(x) = \left\lfloor \frac{\mathfrak{m}_p(ax)}{p/2} \right\rfloor .$$

We remark that, while in the n bin case the shift factor was not important for our purposes, only corresponding to a factor-of-2 difference in maxload, it is more impactful in the 2 bin setting, because for two bins maxload lies in an interval of multiplicative size 2, namely $[n/2, n]$.

Question 1 (Reduction to Two Bins)**.** One potential idea for proving a bound on the maxload achieved in the n bin setting is to first derive a strong concentration bound on the maxload obtained in the 2 bin setting, and then recursively apply this to subsets of the bins. We propose this as an open question.

In order for the technique in Question 1 to succeed, it would be helpful to establish a result of the following form:

Conjecture 1. For any $k \in \Theta(1)$, probability of maxload larger than $n/2 + k\sqrt{n}$ is at most

$$2^{-\Theta(k^2)} + 1/p.$$

Clearly fully-independently thrown balls will achieve the probability tail bound of the conjecture. However, this conjecture seems quite challenging to prove. We prove two weakenings of this conjecture.

Proposition 1. The expected maxload of UH/2 is at most

$$n/2 + \sqrt{n}/2.$$

Proof. We say that $x, y \in X$ ***collide*** if they hash to the same bin. The bins to which distinct $x, y \in X$ map to under UH/2 are pairwise-independent. Thus, the expected number of collisions is $\frac{1}{2}\binom{n}{2}$. On the other hand, if the maxload is m, the number of collisions is

$$\binom{m}{2} + \binom{n-m}{2}.$$

Let random variable M denote the maxload. Combining our two observations yields:

$$\begin{aligned} \frac{1}{2}\binom{n}{2} &\ge \mathbb{E}\left[\binom{M}{2} + \binom{n-M}{2}\right] && (1)\\ &\ge \binom{\mathbb{E}[M]}{2} + \binom{n-\mathbb{E}[M]}{2}, && (2)\end{aligned}$$

where (2) follows by Jensen's inequality due to the convexity of the right hand side of (1). Let μ denote $\mathbb{E}[M]$. (2) is a quadratic in μ; applying the quadratic formula we find

$$\mu \le n/2 + \sqrt{n}/2.$$

We remark that similar simple analysis can give weak tail-bounds. For instance, if we fix $k > 0$ and let δ denote the probability of having maxload at least $n/2(1 + k/\sqrt{n})$, we find

$$\delta \le \frac{1}{k^2}.$$

TODO: There should be some dependence on p here that got swept under the rug? □

We now establish Theorem 1, the analogue of Proposition 1 for function MH/2. It is interesting that we can obtain a similar bound with the simpler hash function and without pairwise independence. In fact, this prompts the following question:

Question 2 (Triple Collisions)**.** Although MH/2, UH/2 do not exhibit 3-wise independence, it may be possible to bound the degree to which they fail to be 3-wise independent, in a manner similar to how we bound the overlap in the proof of Theorem 1. Analysis of colliding trios, or higher numbers of elements, could potentially help give stronger bounds on maxload. However this seems quite challenging.

Theorem 1. The expected maxload of MH/2 is

$$n/2 + \mathcal{O}(\log^2 n).$$

Proof. As in Proposition 1, our proof is based on counting the expected number of collisions, and comparing this to the required number of collisions to have a certain maxload. The difficulty in this proof however, is that elements no longer exhibit pairwise independence, so computing the expected number of collisions is complicated. In particular, some elements collide with probability much larger than $1/2$. For instance, $1, 3$ collide with probability $2/3$, as shown in Figure 1.

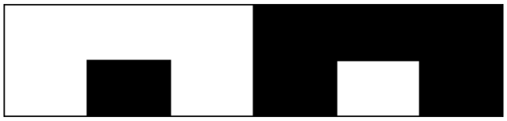

Figure 1: A depiction of the bins that $1, 3$ hash to under values of $a \in [p]$ (the horizontal axis). (Bin 0 is white, Bin 1 is black)

In order to bound the expected number of collisions, we essentially show that for any particular element $x \in [p]$, there are only very few $y \in [p]$ where x, y collide with probability much larger than $1/2$. By symmetry (more precisely, the existence of multiplicative inverses), it does not matter which x we choose to compare with. We will analyze the probability of elements y colliding with 1. Instead of working with probability, we work with the following scaled version of the collisions probability.

Definition 2. The ***overlap*** of $x \in [p]$ is the number of $a \in [p/2]$ such that $1, x$ collide (i.e., hash to the same bin). Equivalently, this is the number of $a \in [p/2]$ such that

$$\mathfrak{m}_p(ax) < p/2.$$

The ***excess overlap*** of x, denoted $\mathfrak{e}(x)$, is the overlap of x minus $p/4$. The ***contribution*** of a set $S \subset [p/2]$ to $\mathfrak{e}(x)$ is the difference between the number of $s \in S$ with $\mathfrak{m}_p(as) < p/2$ and the number of $s \in S$ with $\mathfrak{m}_p(as) > p/2$. Note that we can bound $\mathfrak{e}(x)$ by partitioning $[p/2]$ into disjoint subsets $S_1, S_2 \dots$, and summing the contribution of each S_i to $\mathfrak{e}(x)$.

We proceed to derive bounds on $\mathfrak{e}(x)$ as a function of x, and eventually use these to obtain bounds on $\sum_{x \in X} \mathfrak{e}(x)$. First, we give a simple bound that highlights the general technique required.

Claim 1.

$$\mathfrak{e}(x) \le \mathcal{O}(x + p/x).$$

Proof. We group elements into ***stacks*** consisting of p/x contiguous elements each. A stack is a set of contiguous elements that don't experience overlap. We get ± 1 error per stack, yielding a total of x error, plus p/x error for the final ascent. □

Claim 1 is fairly tight for small x; as hinted by Figure 1 small (odd) x achieve maxload $\Theta(p/x)$. The x in Claim 1 is generally quite weak, but there are some $x > p/n$ with very large maxload. This generally happens if x wraps around to a small number quickly. This suggests that we need a better way of grouping the elements. The following lemma, which is a beautiful elementary fact of number theory, gives such a method.

Lemma 1. For each $x \in [p]$, there exists $m \in [n]$ and $k \in [\lceil p/n \rceil]$ such that

$$xm \equiv k \mod p.$$

Furthermore, this m, k uniquely characterize x.

Proof. One way is you could take $x = p/c$ and look at the wrap-around points. Balance stuff. After w wraparounds you get size like x/w. **TODO:** formalize this

Note that $m^{-1}k$ (where m^{-1} denotes the multiplicative inverse of m modulo p) can take on at most $n \cdot \lceil p/n \rceil$ values, because $m \in [n], k \in [\lceil p/n \rceil]$. However, we associated an m, k pair with all elements of $[p]$ thus they are unique. **TODO:** actually this doesn't quite make sense, for divisibility reasons. I only really care about injective, not surjective. □

Now that we have shown all x are of the form described in Lemma 1, we use this form to give a more accurate bound on $\mathfrak{e}(x)$.

Lemma 2. **TODO:** put a floor/ceiling on all the fractions everywhere! Let $m \le n$ be the smallest m such that $\mathfrak{m}_p(xm) \le p/n$, and let $k = \mathfrak{m}_p(xm) \le p/n$. Then,

$$\mathfrak{e}(x) \le \mathcal{O}\left(\frac{p}{km} + k + m\right).$$

Proof. It may be helpful to refer to Figure 2 for understanding of the terminology used in this lemma.

Define the (i, j)-th ***full line*** as

$$L_{i,j} = i + \frac{mp}{k} \cdot j + m[p/k]$$

for $i \le m, j \le \frac{k}{2m}$.

Claim 2. Each full line contributes $\mathcal{O}(1)$ to $\mathfrak{e}(x)$.

Proof. Fix a full line $\delta + m[p/k]$. The image of $\delta + m[p/k]$ under multiplication by x modulo p is:

$$\mathfrak{m}_p(\delta x), \mathfrak{m}_p((\delta + m)x), \mathfrak{m}_p((\delta + 2m)x), \dots.$$

Let $\delta' = \mathfrak{m}_p(\delta x)$. Because $\mathfrak{m}_p(mx) = k$ we can re-express the image as:

$$\delta', \delta' + k, \delta' + 2k, \dots, \delta' + k \lfloor p/k \rfloor.$$

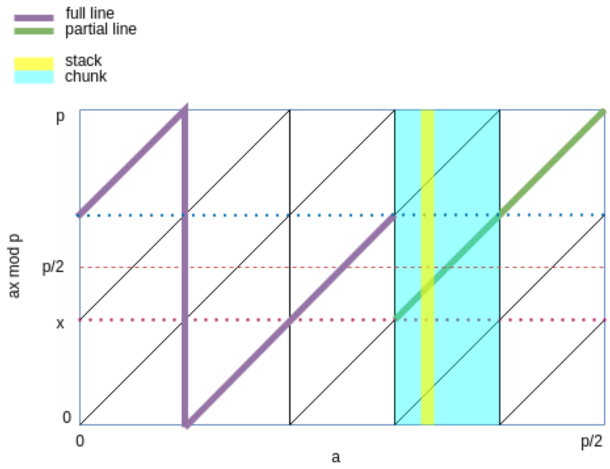

Figure 2: Depiction $\mathfrak{m}_p(ax)$ as a function of a. Partitioned into lines, stacks and chunks.

TODO: I am just making up the floors / ceilings at the moment, plz fix sometime

In other words, the image of the full line consists of $\lfloor p/k \rfloor$ points, which are evenly spaced out, with space k between the points. Thus, we will have $\lfloor p/k \rfloor /2 \pm \mathcal{O}(1)$ above and below the line. In other words, the full line contributes at most $\mathcal{O}(1)$ to $\mathfrak{e}(x)$. □

The full lines take of all but up a suffix of $[p/2]$ of size at most $m \lfloor p/k \rfloor$. Let s be the start of this suffix. We partition this suffix into m ***partial lines***, with the i-th partial line defined as:

$$P_i = (s + i + m[p/k]) \cap [p/2].$$

Unlike full lines, the partial lines need not lie half above and half below $p/2$ (because they may be cut off part-way through). We now further partition the partial lines.

The i-th ***stack*** is

$$S_i = (s + im + [m]) \cap [p/2].$$

We say stack S_i is ***full*** if $|S_i| = m$, and ***partially-full*** if $0 < |S_i| < m$. The i-th ***chunk*** is

$$C_i = \bigcup_{j \in i\frac{p}{km} + [\frac{p}{km}]} S_j.$$

Chunk C_i is a ***full chunk*** if $|C_i| = p/k$. The ***final chunk*** is the final non-empty C_i; this chunk is the only non-empty but not-necessarily-full chunk. A ***line-chunk*** is the restriction of a partial line to a particular chunk.

Claim 3. Each full chunk contributes $\mathcal{O}(1)$ to $\mathfrak{e}(x)$.

Proof. Each stack in a full chunk consists of m contiguous indices. By virtue of the minimality condition on k in the lemma statement, distinct x, y in the same stack are separated by distance at least p/n. In particular, **TODO:** ohh this is tricky **TODO:** hmm,I'm not really sure this is true. it certainly seems difficult to prove...

Case 3.1. m is odd.
TODO: justify all of this By construction the line-chunks within a chunk all reside in disjoint contiguous sub-intervals of $[p]$. Thus, there is exactly one line-chunk which crosses $p/2$. There are $(m-1)/2$ line-chunks above and below $p/2$, so these contribute 0 to $\mathfrak{e}(x)$. The line-chunk which crosses $p/2$ is above and below $p/2$ for half of the time, with ±1 error, so it contributes ±1 to $\mathfrak{e}(x)$.

Case 3.2. m is even.
TODO: justify this: There are $m/2$ line-chunks above and below $p/2$, except for one line-chunk barely touches $p/2$, but overall this only causes a ±1 contribution to $\mathfrak{e}(x)$.

□

Claim 4. The contribution of the final chunk to $\mathfrak{e}(x)$ is at most

$$\mathcal{O}\left(\frac{p}{km} + m\right).$$

Proof. The final chunk consists of at most $\frac{p}{km}$ points per partial line, so in particular each line travels distance at most p/m. This means that only $\mathcal{O}(1)$ partial lines cross $p/2$. We can pair up all the non-crossing partial lines, with one left over if there are an odd number. All of these lines will consist of the same number of full stacks ±1, along with a partial stack. We pay $\mathcal{O}(m)$ for the partial stack, and $\frac{p}{km}$ for the segment of the final un-paired partial line lying in the final chunk. Finally, we pay an additional $\frac{p}{km}$ for the partial line which crosses $p/2$. **TODO:** so this, like the thing above, relies on an unproven, and frankly unlikely to be true, assertion that each full stack has a contribution of $\pm\mathcal{O}(1)$ to the excess overlap. while this seems somewhat reasonable, by virtue of a heuristic that the points in a stack are nearly evenly distributed, it seems very difficult to prove (and again, somewhat likely to be false). small matter! another partition ought to do the trick. I still think that the parameterization is clever, and unearths the bx well. □

There are at most k full lines and at most m full chunks. The full lines, full chunks and final chunk taken together consitute a partition of

$[\lfloor p/2 \rfloor]$. Summing the contributions to $\mathfrak{e}(x)$ from full lines, full chunks, and the final chunk as bounded in Claim 2,Claim 3,Claim 4 gives a bound on the excess overlap. In particular we have:

$$\mathfrak{e}(x) \leq \mathcal{O}\left(\frac{p}{km} + k + m\right).$$

□

Corollary 1.

$$\sum_{x \in X} \mathfrak{e}(x) \leq \mathcal{O}(p \log^2 n).$$

Proof. From Lemma 2 we obtain the bound

$$\sum_{x \in X} \mathfrak{e}(x) \leq \mathcal{O}\left(\sum_{m,k \leq \sqrt{n}} \frac{p}{mk} + \frac{p}{n} + n\right) \leq \mathcal{O}(p \log^2 n).$$

□

As a straightforward corollary of Corollary 1 we obtain a bound on the expected number of collisions.

Corollary 2. The total expected number of collisions is at most

$$n^2/4 + \mathcal{O}(n \log^2 n).$$

Proof. Modular inverses exist, so everything is symmetric.

And this is what you get when you sum stuff up. The $n^2/4$ is important BTW. □

Finally, we complete the proof by comparing the expected number of collisions to the number of collisions induced by the maxload.

Corollary 3. The expected maxload of $\mathsf{MH}/2$ is

$$n/2 + \mathcal{O}(\log^2 n).$$

Proof. Let μ denote the expected maxload. By convexity, Jensen's inequality says

$$n^2/4 + \mathcal{O}(n \log^2 n) \geq \binom{\mu}{2} + \binom{n-\mu}{2}.$$

Doing some algebra we obtain:

$$\mu \leq n/2 + \mathcal{O}(\log^2 n).$$

□

□

2 Two Bins, Try 2 (with circles)

Now we give a bound on $\sum_{x \in X} \mathfrak{e}(x)$. We conjecture that the bound we give is far from tight. In particular we propose Conjecture 2. However, the proof in Theorem 2 does well enough at helping us understand the shape of $\sum_{x \in X} \mathfrak{e}(x)$, in particular showing that it much smaller than $p \cdot n$.

Conjecture 2.

$$\sum_{x \in X} \mathfrak{e}(x) \leq (\lg^2 n)/n.$$

Proof. The argument would be as follows: for each x, we can write $xm \equiv k$ for $m < n, k < p/n$. Then we partition $[p/2]$ into full lines then chunks and the final chunk.

Difficulty: segments of m contiguous things are really not so uniform, even though intuitively they behave like uniform things. And this is kind of hard to think about. But intuitively this should give you like $p/(km)$ behavior bound. This is supported by extensive simulations. If you add it up it gives you like $\log^2 n$. □

Theorem 2.

$$\sum_{x \in X} \mathfrak{e}(x) \leq p \cdot n^{3/4}.$$

Proof. First, we demonstrate the power of our partitioning technique for bounding excess overlap and eliminate a portion of $[p/2]$ which our later methods are less effective against.

Claim 5. Let $x < p/n$. Then

$$\mathfrak{e}(x) \leq \mathcal{O}(x + p/x).$$

Proof. We partition $[p/2]$ into ***stacks*** which are contiguous groups of $m \in \lceil p/x \rceil \pm \mathcal{O}(1)$ elements $a + [m]$ which do not experience overlap, i.e.,

$$\mathfrak{m}_p(ax) < \mathfrak{m}_p((a+1)x) < \ldots < \mathfrak{m}_p((a+m-1)x).$$

Clearly there are $\Theta(x)$ stacks, and each stack, except for the final one, contributes $\pm\Theta(1)$ to $\mathfrak{e}(x)$. The final stack might not have $\Omega(p/x)$ elements. However, it certainly contributes at most $\mathcal{O}(p/x)$ to the $\mathfrak{e}(x)$. Summing the contributions of all stacks gives the desired bound on $\mathfrak{e}(x)$. □

Corollary 4.

$$\sum_{x \in X \cap [\lceil p/n \rceil]} \mathfrak{e}(x) \leq \mathcal{O}(p \log n).$$

Proof. This follows immediately from Claim 5. □

Because of Corollary 4 it now suffices to analyze the case where $x > p/n$. We remark that small odd numbers x result in precisely $\Theta(p/x)$ excess overlap, so our understanding of the excess overlap incurred by small x is relatively tight. However, small odd x are not the only x which incur large excess overlap. If x wraps around to a small value quickly, then x can also have large excess overlap. In what follows we give some (very lose) methods for bounding these messier, more subtly bad x.

TODO: make an alias for $\lfloor p/2 \rfloor$

Definition 3. Fix $L \in \mathbb{N}$ and $x \in [p/2] \setminus \{0\}$. We define values

$$\delta_i = \delta_i(x, L), c_i = c_i(x, L), \ell_i = \ell_i(x, L)$$

as $\delta_0 = x$, $c_0 = 1$, and for $i \in \{1, 2, \dots, L-1\}$

$$\ell_{i-1} = \left(\underset{\ell \in \{\lceil p/\delta_{i-1} \rceil, \lfloor p/\delta_{i-1} \rfloor\}}{\operatorname{argmin}} |\ell_{i-1} \cdot \delta_{i-1}|_p \right) - 1,$$

$$\delta_i = |\ell_{i-1} \cdot \delta_{i-1}|_p,$$

$$c_i = c_{i-1} \cdot \ell_{i-1} = \prod_{j \in [i-1]} \ell_j.$$

While δ_i, c_i, ℓ_i are functions of x, L we will leave this dependence implicit so long as x, L are defined in context. Intuitively, δ_i is obtained by traveling once around a circle of $[p/2]$ points with stride δ_{i-1} and then taking the remainder obtained at the end.

We partition $[p/2]$ into L ***levels*** and then further partition the levels into ***circles***. For each $i \in [L]$, level i is composed of c_i circles. Intuitively, a circle on level i is composed of ℓ_i elements of $[p/2]$. However, if i is too large then all $[p/2]$ elements will have already been used up. We say that a circle on level i is ***full*** if it actually has ℓ_i elements, and that a level is ***full*** if all circles on that level are full. We say that a L is a ***valid*** number of levels if all but the final level is full (and the final level is non-empty). Note that in the final level circles could either have more or less than ℓ_L elements.

Formally, circle j on level i consists of the elements:

$$[p/2] \bigcap \left(c_i \cdot [\ell_i] + j + \sum_{k=0}^{i-1} c_k \ell_k \right).$$

Claim 6. The image of a level i circle under $y \mapsto \mathfrak{m}_p(xy)$ consists of elements spaced out by δ_i.

Proof. This is an immediate consequence of our definition. This property is illustrated in Figure 3. □

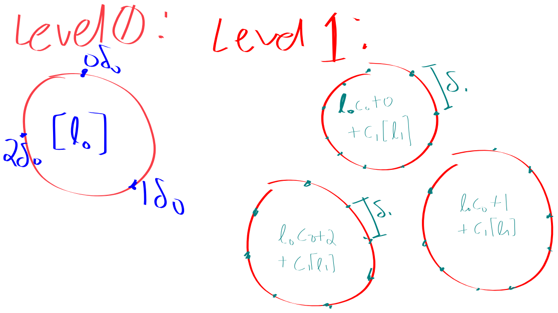

Figure 3: Circles and Levels

The utility of the level/circle method of viewing x is explained in Lemma 3.

Lemma 3. Fix $x > p/n$, and let $L \in \mathbb{N}$ be a valid number of levels for x such that x has m elements on level L. Then

$$\mathfrak{e}(x) \le \mathcal{O}(c_L + p/c_L) + ??.$$

TODO: where ?? is the noise for number of circles before the final level

Proof. We partition the elements on level L into contiguous sets of c_L elements, which we term ***full cycles***. There are $\mathcal{O}(c_L)$ elements at the end which are not included in any full cycle, but constitute a ***partial cycle***

Claim 7. Each full cycle contributes $\mathcal{O}(1)$ to $\mathfrak{e}(x)$.

Proof. We prove Claim 7 by induction on L.

If $L = 1$ it is clear: this is simply saying that evenly spaced points spanning the circle results in $\pm\Theta(1)$ contribution to $\mathfrak{e}(x)$.

ok. a bound of c_{L-1} is obvious; just cancel circles with themselves. (but iirc this is not good enough bound) The induction is somehow supposed to show that the "extra" points in each circles can cancel with each other?.

maybe worth taking a look at lemma 7 to see exactly what we have to work with...

Let $k \equiv mx$. For simplicity let k be positive (I mean less than $p/2$); everything is symmetric if k is negative.

Our partition is as follows: contiguous blocks of m elements. Ok now we need to show that this works.

First a simple case, which is our base case for an induction. We say that x has a single level if $m = \lceil p/x \rceil$. Then, it's clear that m contiguous elements has ± 1 excess overlap contribution.

Now let's ramp it up a level, and do x which is like two levels or whatever. So now we have a circle, and basically think of m circles sprouting out of this original circle. Think of k being supa supa small. So now each of these 'lil circles has one weird "stitch" thing. Ok but basically we can pair most of the circle-y dudes up with themselves. But maybe they each have some extra points. But then we also pair up these extra points so its fine.

And then we can induct that for any number of levels. With the only inductive thing that we need being like each circle that the final circles sprout from has exactly one stitch and beyond that has just uniform distance things.

Anyways once we have the whole $\mathcal{O}(1)$ err per m steps thing we are super chilling for our $m + p/m$ bound.

□

There are $\Theta(p/c_L)$ full cycles, and each contributes at most $\mathcal{O}(1)$ to $\mathfrak{e}(x)$ by Claim 7. There are at most c_L elements in the partial cycle; these elements contribute at most c_L to $\mathfrak{e}(x)$. The circles before the final level contribute at most ?? to $\mathfrak{e}(x)$. Combined, this gives the desired bound on $\mathfrak{e}(x)$. □

For $m < \sqrt{n}$ we say that x is ***unavoidably**-m-*circle if there exists a level ℓ such that $c_\ell = m$, while $c_{\ell+1} > p/\sqrt{n}$. Intuitively, this means that the granularity on level ℓ is very small.

Lemma 4. Fix $m < \sqrt{n}$. There are at most $\mathcal{O}(m\sqrt{n})$ unavoidably-m-circle x.

Proof. Assume that for some ℓ, we have $c_\ell(x) = m$. To be unavoidably-m-circle, x must satisfy

$$\left(\sum_{k=0}^{\ell-1} c_k \ell_k\right) \cdot x \equiv \delta \mod p$$

TODO: I think this is the right number for some very small δ. In particular, $c_{\ell+1}$ must satisfy:

$$c_{\ell+1} \approx mp/\delta > p/\sqrt{n}.$$

TODO: there is some fudging about the fact that $m \neq \sum c\ell$; but intuitively $m \approx \sum c\ell$ so I won't press the issue right now This results in the requirement $\delta < m\sqrt{n}$. However, for each such value of δ there is a unique solution to $mx \equiv \delta \mod p$, and this solution might not even have $c_\ell(x) = m$. Thus, there are at most to at most $m\sqrt{n}$ values of x which are unavoidably-m-circle.

□

Claim 8. Adding up stuff we get average excess overlap $p/n^{1/4}$.

Proof. We combine our two lemmas and add stuff up. Adding stuff up, in the worst case, we obtain:

$$\sqrt{n}p + 2\sqrt{n}p/2 + \cdots + n^{1/4}\sqrt{n}p/n^{1/4} = n^{3/4}p.$$

□

□

3 Try 3 of the insane proof

Theorem 3.

$$\sum_{x\in X} \mathfrak{e}(x) \leq p \cdot n^{11/12}.$$

Proof.

Claim 9. $\mathfrak{e}(x) \leq x + p/x$

Proof. Partition to stacks. □

Corollary 1.

$$\sum_{x\in X\cap[\lceil p/n^{1/12}\rceil]} \mathfrak{e}(x) \leq \mathcal{O}(pn^{11/12}).$$

Proof. Immediate from above claim. □

Lemma 5. Fix x, and let L be a valid number of levels for x. Then,

$$\mathfrak{e}(x) \leq p/\ell_L + c_L\ell_L.$$

TODO: isn't there some noise for non ultimate levels

Proof. Partition $[p/2]$ into groups of c_{L+1} elements. By [claim proving that circles consist of evenly spaced elements] we incur error at most $\mathcal{O}(1)$ per all ℓ_L points in each circle. Thus each group contributes at most c_L to $\mathfrak{e}(x)$, so the total contribution from all groups together is at most $pc_L/(\ell_L c_L) = p/\ell_L$. Finally, there are at most c_{L+1} elements left over which do not fit into a group. These contribute at most c_{L+1} to $\mathfrak{e}(x)$, which together with the bound on the contribution to $\mathfrak{e}(x)$ from groups gives the desired bound on $\mathfrak{e}(x)$. Note that our estimate of the elements which do not fit in a group is quite weak if $L > 1$; However, it is simple and sufficient for our purposes. □

Claim 10.

$$c_L < \ell_L.$$

TODO: is this actually true? I think probably not. do I really need it?

Proof. **Intuition**: ℓ at least doubles each time. The hardest c_L/ℓ_L ratio you can get is probably if we just barely double each time. But in that case it does hold. □

Claim 11. For any $x > p/n^{1/12}$, there exists an $L \in \mathbb{N}$ such that $C_{L+1} < p/n^{1/12}$.

Proof. $p/x \approx n^{1/12} \ll p/n^{1/12}$. □

Definition 4. We say that x is z-bad if there is a valid level $L \in N$ such that $\ell_L(x) = z$, $c_{L+1} < p/n^{1/12}$ but $c_{L+2} > p/n^{1/12}$.

By Claim 11 any $x > p/n^{1/12}$ is z-bad for some $z \in [p]$.

Lemma 6. The number of z-bad integers $x \in [p/n^{1/12}, p]$ is at most $z^3 n^{1/6}$.

Proof. For x to be z-bad we must have $\ell_{L+1}\ell_L c_L > p/n^{1/12}$. Of course $\ell_{L+1} \approx p/\delta_{L+1}$, so this translates into

$$\delta_{L+1} < z^2 n^{1/12}.$$

On the other hand, $\delta_L = c_L \cdot x$, and so

$$\delta_{L+1} = \lfloor p/\delta_L \rfloor \delta_L = (p/(c_L x)c_L)\, x.$$

Of course $p/(c_L x)c_L < n^{1/12} z$, so we obtain the equation

$$\eta x \equiv \delta$$

where $\eta < n^{1/12} z$, $\delta < z^2 n^{1/12}$. Clearly, there are at most $n^{1/6} z^3$ solutions to this equation.

Intuitively, this is expressing the fact that a quick drastic decrease in granularity is fairly rare. □

Corollary 2.

$$\sum_{x \in X \cap [p/n^{1/12}, p]} \mathfrak{e}(x) \le pn^{11/12}.$$

Proof. The worst case is

$$\sum_{k=1}^{n^{1/4}} pn^{1/6} z^3 / z = pn^{11/12}.$$

□

Combining Corollary[A] and Corollary[B] gives the desired bound on $\sum_{x \in X} \mathfrak{e}(x)$.

□

4 Try 4 of the proof

Theorem 4.

$$\sum_{x \in X} \mathfrak{e}(x) \le pn^{3/4}.$$

Proof.

Claim 12. All but at most $n^{1/4}$ of $x \in [p]$ are of the form

$$x \equiv m^{-1} k \mod p$$

for some $m < n^{1/4}, k \in (n^{1/2}, p/n^{1/4})$.

Proof. Clear. □

Lemma 7. Let x be of the form $x \equiv m^{-1}k$ for some $m < n^{1/4}, k \in (n^{1/2}, p/n^{1/4})$. Then,

$$\mathfrak{e}(x) \le mp/k + k.$$

Proof. Think of it as being m circles at granularity k. Every mp/k steps we get $\pm m$ contribution to $\mathfrak{e}(x)$. Add on mp/k because we can't control the end. This gives the desired bound. □

Corollary 3. Thus we get $pn^{3/4}$.

Proof. Add it up. □

□

5 Another remark about circles

Imagine that you have some set of m_i, k_i such that $xm_i \equiv k_i$ for each i. Imagine further that you could form a partition

$$p/2 = \sum_i \zeta_i \lfloor p/k_i \rfloor m_i.$$

Then you can view $p/2$ as being the union over i of $\zeta_i m_i$ circles of granularity k_i.

References

[1] Noga Alon, Martin Dietzfelbinger, Peter Bro Miltersen, Erez Petrank, and Gabor Tardos. Is linear hashing good? In *Conference Proceedings of the Annual ACM Symposium on Theory of Computing*, pages 465–474. Association for Computing Machinery (ACM), January 1997. ISSN: 0734-9025.

[2] G. H. Hardy and E. M. Wright. *An Introduction to the Theory of Numbers*. Oxford University Press, Oxford, fifth edition, 1979.

[3] Mathias Bæk Tejs Knudsen. Linear hashing is awesome, June 2017. arXiv:1706.02783 [cs].

[4] Alan Siegel. On Universal Classes of Extremely Random Constant-Time Hash Functions. *SIAM Journal on Computing*, 33(3):505–543, January 2004. Publisher: Society for Industrial and Applied Mathematics.