CS 4850 Lecture Notes
April 17, 2009
Taken by Eric Frackleton

Data Streams:
We have a sequence of integers $a_1, a_2, \ldots, a_n$ all in $\{1, \ldots, m\}$ where both m and n are large. We would like to determine the number of distinct elements in the sequence. If we want the exact answer, any algorithm would require m bits of space because the set of distinct elements could be any of the 2^m subsets of $\{1, \ldots, m\}$ and knowing only the count at any given point in time is insufficient because you have to be able to deduce if the next one has been seen before.

Low Memory Approximation Algorithm: We want to determine if the number of distinct elements is above t with high probability using only $logt$ space.
It should be the case that as the number of distinct elements increases, so does the probability the algorithm returns yes. This probability should be low when there are $< \frac{t}{2}$ distinct elements and high when there are $> 2t$ distinct elements.
Let $h : \{1, \ldots, m\} \to \{1, \ldots, t\}$ be a hash function that takes each number from 1 to m to some number between 1 and t with the same probability. Such a function may have the form $h(x) = ax+b \mod t$ for some a and b and will require $logt$ space to store. The algorithm returns yes if $h(a_i) = 1$ for some i.
We will show that if there are at least $2t$ distinct elements, the algorithm says yes with probability at least 0.865 and if there are at most $\frac{t}{2}$ distinct elements, it says yes with probability at most 0.4. For each symbol a in the sequence, $Pr[h(a) = 1] = \frac{1}{t}$. Therefore $Pr[h(a) \neq 1$ for d distinct elements$] = (1 - \frac{1}{t})^d$. This function decreases as d increases. Hence if $d < \frac{t}{2}$, we have that $Pr[h(a) \neq 1$ for d distinct elements$] > (1 - \frac{1}{t})^{\frac{t}{2}} \approx \frac{1}{\sqrt{e}} \approx 0.6$. Therefore, the probability it returns yes is at most $1 - \frac{1}{\sqrt{e}} \approx 0.4$. Similarly, if $d > 2t$, we have that $Pr[h(a) \neq 1$ for d distinct elements$] < (1-\frac{1}{t})^{2t} \approx \frac{1}{e^2} \approx 0.135$. Therefore, the probability it returns yes is at least $1-\frac{1}{e^2} \approx 0.865$.

Another approximation algorithm: Let S be the set of distinct elements in the stream and min be the minimal element of S. If we divide the range 1 to m up into $|S| + 1$ regions, we expect the border between the first and second region to be a reasonable approximation of min. Therefore, we can approximate $|S|$ from min as $min \approx \frac{m}{|S|+1}$ so $|S| \approx \frac{m}{min} - 1$.
Problem: The sequence is likely not going to be random. For example, it could be a group of small elements. We will account for this next time using hash functions, but for now assume the elements of the sequence are random.
Lemma: If a_i are selected uniformly at random from 1 to m then with probability at least $\frac{2}{3}$, we have that $\frac{d}{6} \leq \frac{m}{min} \leq 6d$ where $d = |S|$.
Proof: First we want to show that $Pr[\frac{m}{min} > 6d] < \frac{1}{6}$: We have $Pr[\frac{m}{min} > 6d] = Pr[min < \frac{m}{6d}] = Pr[\exists k \text{ s.t. } b_k < \frac{m}{6d}]$ where b_k is the k^{th} element of S. Let $z_k = 1$ if $b_k < \frac{m}{6d}$ and 0 otherwise. Also let $z = \sum_{k=1}^{d} z_k$.
What is the probability that $z_k = 1$? It is $\frac{1}{6d}$ because there are m choices of b_k. Therefore, $E[z_k] = \frac{1}{6d}$ so $E[z] = \frac{1}{6}$.
What is probability that z differs from its expected value by enough to have value 1?
What is probability that value of z is at least 6 times its expected value? Use Markov inequality: $Pr[X \geq aE[X]] \leq \frac{1}{a}$. Then $Pr[z \geq 6E[z]] \leq \frac{1}{6}$ so $Pr[z \geq 1] \leq \frac{1}{6}$ and therefore $Pr[\frac{m}{min} > 6d] \leq \frac{1}{6}$ because $z \geq 1$ iff $min < \frac{m}{6d}$.
It just remains to show something similar for $\frac{m}{min} < \frac{d}{6}$, which is left for next time.