1 Basic probability tools

1.1 Markov


\[ E[X] \geq \Pr[X \geq a] \cdot a + \Pr[X \leq a] \cdot 0. \]

1.2 Chebyshev

\[
\Pr[|X - E[X]| \geq a] = \Pr[(x - E[x])^2 \geq a^2] \\
\leq \frac{E[(x - E[x])^2]}{a^2} = \frac{\text{Var}[X]}{a^2}
\]

1.3 Example: Coupon Collector

\(X_i\): # boxes bought when you have exactly \(i - 1\) different coupons, and after buying these boxes, you have \(i\) different coupons.

Let \(X = \sum_{i \in [n]} X_i\): # boxes one bought until at least one of every type of coupon is obtained.

Let \(p_i\) be the probability of obtaining a new coupon when we have exactly \(i - 1\) different coupons.

\(p_i = 1 - \frac{i-1}{n}\). Then \(E[X_i] = \frac{1}{p_i} = \frac{n}{n-i+1}\). Therefore \(E[X] = nH_n \approx n \ln n\).

By Markov, \(\Pr[X \geq 2n \ln n] \leq 1/2\).

Get a better bound using Chebyshev. First, note that \(X_i\)’s are independent, hence \(\text{Var}[X] = \sum_{i \in [n]} \text{Var}[X_i]\).

Let \(Y\) be a geometric random variable with parameters \(p\), then

\[ E[Y] = \sum_{i=1}^{p} \Pr[X \geq i] = \sum_{i=1}^{\infty} (1 - p)^{i-1} = 1/p, \]

and

\[ E[Y^2] = \sum_{i=1}^{\infty} p(1 - p)^{i-1} i^2 \]

\[ = \frac{2 - p}{p^2} < 1/p^2. \]
Therefore,
\[
\text{Var}[X] = \sum_{i \in [n]} \text{Var}[X_i] \leq \sum_{i \in [n]} \left( \frac{n}{n - i + 1} \right)^2 = n^2 \sum_{i \in [n]} \frac{1}{i^2} \leq n^2 \cdot \frac{\pi^2}{6}.
\]

And
\[
\Pr[|X - n \ln n| \geq n \ln n] \leq \frac{n^2 \pi^2/6}{(n \ln n)^2} = \Theta\left( \frac{1}{\ln^2 n} \right).
\]

Get an even better bound using the union bound. The probability that the \(i\)-th coupon is NOT obtained after \(n \ln n + cn\) steps is
\[
\left( 1 - \frac{1}{n} \right)^{n(\ln n + c)} < e^{-(\ln n + c)} = \frac{1}{e^c \cdot n}.
\]
Setting \(c = \log n\), the probability is at most \(1/n\). Then use the union bound.

1.4 Chernoff, estimating parameters

We want to show that \(\Pr[p \in [\bar{p} - \delta, \bar{p} + \delta]] \geq 1 - \gamma\).

\(X = n\bar{p}\) has a binomial distribution w/ parameter \(n, p\). So \(\mathbb{E}[X] = np\). If \(p \not\in [\bar{p} - \delta, \bar{p} + \delta]\), then we have one of the followings.

1. \(X > \mathbb{E}[X] + \delta/p \cdot \mathbb{E}[X]\).
2. \(X < \mathbb{E}[X] - \delta/p \cdot \mathbb{E}[X]\).

Therefore,
\[
\Pr[X \notin [\mathbb{E}[X] - \delta/p \cdot \mathbb{E}[X], \mathbb{E}[X] + \delta/p \cdot \mathbb{E}[X]]] < 2e^{-\frac{2(\delta/p)^2}{n}} = 2e^{-2n\delta^2}.
\]

2 FM sketch

Denote \(d\) the number of distinct elements.

Basic intuition is that we expect 1 out of the \(d\) distinct elements to hit zeros(\(h(j)\)) > \(\log d\), and we don’t expect any elements to hit zeros(\(h(j)\)) \(\gg \log d\).

For each \(j \in [n]\), each \(r \geq 0\), let \(X_{r,j}\) be the indicator variable for the event that \(\text{zeros}(h(j)) \geq r\). Let \(X_r = \sum_{j: f_j > 0} X_{r,j}\). At the end of the process,

\[
X_r > 0 \iff z \geq r.
\]
\[
X_r = 0 \iff z \leq r - 1.
\]

Since \(h(j)\) is uniformly distributed over the \((\log n)\)-bit strings, we have \(\mathbb{E}[X_{r,j}] = 1/2^r\). Therefore
1. $\mathbb{E}[X_r] = \sum_{j:f_j > 0} \mathbb{E}[X_{r,j}] = d/2^{r}$. 

2. $\text{Var}[X_r] = \sum_{j:f_j > 0} \text{Var}[X_{r,j}] \leq \sum_{j:f_j > 0} \mathbb{E}[X_{r,j}^2] = \sum_{j:f_j > 0} \mathbb{E}[X_{r,j}] = d/2^{r}$. 

Thus, using Markov and Chebyshev, we have

1. $\Pr[X_r > 0] = \Pr[X_r \geq 1] \leq \frac{\mathbb{E}[Y_r]}{d} = \frac{d}{2^{r}}$, and

2. $\Pr[X_r = 0] = \Pr[|X_r - \mathbb{E}[X_r]| \geq d/2^{r}] \leq \frac{\text{Var}[X_r]}{(d/2^{r})^2} \leq 2^r/d$. 

Let $\hat{d} = 2^{z+1/2}$. Let $a$ be the smallest integer such that $2^a + 1/2 \geq 3d$, and $b$ be the largest integer such that $2^{b+1} \leq d/3$. Then

1. $\Pr[\hat{d} \geq 3d] = \Pr[z \geq a] = \mathbb{E}[X_a > 0] = \frac{d}{2^{r}} \leq \frac{\sqrt{2}}{3} \approx 47\%$, and

2. $\Pr[\hat{d} \leq d/3] = \Pr[z \leq b] = \mathbb{E}[X_{b+1} = 0] = \frac{2^{b+1} - 1}{d} \leq \frac{\sqrt{2}}{3} \approx 47\%$. 

Success ratio about 14%.

**Probability amplification.** Can we boost the success probability to $1 - \delta$? The idea is to run $k$ copies of this algorithm in parallel, using mutually independent random hash functions, and output the median of the $k$ answers.

1. If this median exceeds $3d$, then at least $k/2$ copies having answer exceeds $3d$, whereas we only expect $\sqrt{2}/3 \cdot k < k/2$ of them to exceed $3d$. By a standard Chernoff bound, the probability of this event is just $2^{-\Omega(k)}$, which is $\delta$ if we choose $k = \Theta(\log(1/\delta))$.

2. The arguments for the other side is similar.

If $\delta$ is a constant, we only needs $\Theta(1)$ hash functions. Store and compute a suitable hash function needs $O(\log n)$ bits, and store $z$ needs $O(\log \log n)$ bits. Therefore the total space used is $O(\log n)$ bits.

### 3 BJKST sketch

Same as FM sketch, for each $j \in [n]$, each $r \geq 0$, let $X_{r,j}$ be the indicator variable for the event that $\text{zeros}(h(j)) \geq r$. Let $X_r = \sum_{j:f_j > 0} X_{r,j}$. At the end of the process, 

$X_r > 0 \iff z \geq r$. 

$X_r = 0 \iff z < r - 1$. 

Since $h(j)$ is uniformly distributed over the $(\log n)$-bit strings, we have $\mathbb{E}[X_{r,j}] = 1/2^r$. Therefore

1. $\mathbb{E}[X_r] = \sum_{j:f_j > 0} \mathbb{E}[X_{r,j}] = d/2^{r}$. 

2. $\text{Var}[X_r] = \sum_{j:f_j > 0} \text{Var}[X_{r,j}] \leq \sum_{j:f_j > 0} \mathbb{E}[X_{r,j}^2] = \sum_{j:f_j > 0} \mathbb{E}[X_{r,j}] = d/2^{r}$. 

3
When we have a failure? $|X_z 2^z - d| \geq \epsilon d$. Which is equivalent to

$$\log n \bigcup_{r=0}^{2^z} (|X_r - d/2^r| \geq \epsilon d/2^r) \cap (r = z).$$

The idea is to show:

1. For small values of $r$, $|X_r - d/2^r| \geq \epsilon d/2^r$ is unlikely.
2. For large values of $r$, $r = z$ is unlikely.

Now we prove these two claims. Let $s$ be an integer such that $12/\epsilon^2 \leq d/2^s < 24/\epsilon^2$. Any $r \geq s$ will be considered large, and any $r < s$ will be considered small. We start with $r = 1$ since when $r = 0$ there is no failure. Observe that

$$\log n \bigcup_{r=0}^{2^z} (|X_r - d/2^r| \geq \epsilon d/2^r) \cap (r = z) \subseteq \bigcup_{r=1}^{s-1} (|X_r - d/2^r| \geq \epsilon d/2^r) \cup \bigcup_{r=s}^{\log n} (r = z)$$

Using Chebyshev, we have for $r < s$,

$$\Pr[|X_r - \epsilon d/2^r| \geq \epsilon d/2^r] = \Pr[|X_r - \mathbb{E}[X_r]| \geq \epsilon d/2^r] \leq \frac{\text{Var}[X_r]}{(\epsilon d/2^r)^2} = \frac{2^s/(\epsilon^2 d)}{\epsilon^2}$$

Therefore,

$$\sum_{r<s} \Pr[|X_r - \epsilon d/2^r| \geq \epsilon d/2^r] \leq \frac{2^s/d \cdot \epsilon^2}{1/12} \leq 1/12.$$ 

Using Markov, we have

$$\Pr[r \geq s] = \Pr[X_{s-1} > c/\epsilon^2] \quad \text{(by the algorithm)}$$

$$\leq \mathbb{E}[X_{s-1}] \cdot \epsilon^2/c$$

$$\leq 48/c = 1/6 \quad \text{(by setting } c = 288)$$

The final result follows by a union bound.

Space: 2 hash functions + $O(1/\epsilon^2)$ elements in the universe $O(\log n/\epsilon^2)$.

**Comment 1** The current best algorithm by Kane et al. (PODS 2010) achieves space $O(1/\epsilon^2 + \log n)$ bits, which is also tight.

### 4 Linear Sketch for Distinct Element

Why linear sketch is ideal for streaming algorithms? – It can handle both insertion and deletion.

Why our algorithm is a linear sketch? – Each sum is a linear combination of entries $x_1, \ldots, x_n$ (write out the matrix $M$).
4.1 Proof of the first lemma

\[ P = (1 - 1/T)^D \approx e^{-D/T}. \]
Thus if \( D \geq (1 + \epsilon)T \) then \( P \leq e^{-(1+\epsilon)} < 1/e - \epsilon/3 \) (use basic calculus).

The other direction is similar.

The second lemma is just a Chernoff.

5 Space-saving

Very similar to Misra-Gries.

6 Count-min

For each \( k \in [d] \), for each \( i \in [n] \), let \( j = h_k(i) \). W.l.o.g, let’s consider \( i = 1 \). We now try to analysis the noise/excess in the counter \( Z_{h_k(1)}^k \), denoted by \( Y_k \). \( Y \) can be expressed as \( Y_{k,2} + Y_{k,3} + \ldots + Y_{k,n} \), where \( Y_{k,i} \) is the contribution that \( j \) makes if it collides with item 1 under hash function \( h_k \), that is, \( h_k(j) = h_k(1) \). It is easy to see that

\[
Y_{i,j} = \begin{cases} 
x_j, & \text{w. pr. } 1/k, \\
0, & \text{w. pr. } 1 - 1/k.
\end{cases}
\]

Obviously, \( \mathbb{E}[Y_{i,j}] = x_j/k \) for all \( j \), and

\[
\mathbb{E}[Y_i] = \mathbb{E}[Y_{k,2}] + \mathbb{E}[Y_{k,3}] + \ldots + \mathbb{E}[Y_{k,n}] = \frac{\|x\|_1 - x_1}{k} \leq \frac{\|x\|_1}{k} = \frac{\epsilon}{2} \|x\|_1,
\]

by linearity of expectation and our choice of \( k \).

By Markov, \( \Pr[Y_i \geq \epsilon \|x\|_1] \leq 1/2. \)
Take the \( \min \) among \( d = \log(1/\delta) \) independent counters will boost this probability to \( 1/2^d = 1 - \delta \).

7 Count-sketch

Fix a \( k \in [d] \). For each \( i \in [n] \), let \( j = h_k(i) \). W.l.o.g, let’s consider \( i = 1 \). We now try to analysis the noise/excess in the counter \( Z_{h_k(1)}^k \), denoted by \( Y \). \( Y \) can be expressed as \( Y_2 + Y_3 + \ldots + Y_n \), where \( Y_i \) is the contribution that \( j \) makes if it collides with item 1 under hash function \( h_k \), that is, \( h_k(j) = h_k(1) \). It is easy to see that

\[
Y_j = \begin{cases} 
x_j, & \text{w. pr. } 1/(2w), \\
-x_j, & \text{w. pr. } 1/(2w), \\
0, & \text{w. pr. } 1 - 1/w.
\end{cases}
\]

Obviously, \( \mathbb{E}[Y_j] = 0 \) for all \( j \), and \( \mathbb{E}[Y] = 0 \) by linearity of expectation.

Note that \( Y_j \)'s are pairwise independent as the hash functions chosen are 2-universal, therefore we have \( \text{Var}[Y] = \text{Var}[Y_2] + \ldots + \text{Var}[Y_n] \).

It is easy to calculate the variance of \( Y_j \): \( \text{Var}[Y_j] = \mathbb{E}[Y_j^2] - 0 = x_j^2/w. \) Therefore

\[
\text{Var}[Y] = \frac{x_2^2 + \ldots + x_n^2}{w} \leq \frac{\|x\|_2^2}{w}.
\]

The rest is a standard Chebyshev + Chernoff.
8 GKMS $L_2$ point query

By JL-lemma, $s = \|Rx\|_2 = (1 \pm \epsilon) \|x\|_2$ w.h.p.

Denote $\tilde{b} = x/s$, $\tilde{c} = e_i$, and $\tilde{a} = \tilde{b} - \tilde{c}$. Then w.h.p., $\|\tilde{b}\|_2 = 1 \pm \epsilon$. Let’s assume (for simplicity) first $\|\tilde{b}\|_2 = 1$, and we will put $\epsilon$ back later.

Since $\|\tilde{a}\|_2^2 = \|\tilde{b} - \tilde{c}\|_2^2 = (\tilde{b} - \tilde{c})^2 = \|b\|_2^2 + \|c\|_2^2 - 2\tilde{b}\tilde{c}$, we have

$$1 - \|\tilde{a}\|_2^2 / 2 = \left(\|b\|_2^2 + \|c\|_2^2 - \|\tilde{a}\|_2^2\right) / 2 = \tilde{b}\tilde{c}.$$

Going back to $x, s, e_i$, we have

$$1 - \|x/s - e_i\|_2^2 / 2 = xe_i / s \iff x_i = \left(1 - \|x/s - e_i\|_2^2 / 2\right) s.$$

Now we deal with “epsilon-acrobatics”.

$$\|Rx/s - Re_i\|_2^2 / 2 = \|R(x/s - e_i)\|_2^2 / 2$$

$$= (1 \pm \epsilon) \|x/s - e_i\|_2^2 / 2$$

$$= (1 \pm \epsilon) \left(\frac{x}{(1 \pm \epsilon) \|x\|_2} - e_i\right)^2 / 2$$

$$= (1 \pm \epsilon) \left(\frac{1}{1 \pm \epsilon} + 1 - 2xe_i / \|x\|_2 (1 \pm \epsilon)\right) / 2$$

$$= (1 \pm c\epsilon) \left(1 - xe_i / \|x\|_2\right),$$

for some absolute constant $c > 1$.

We next have

$$(1 - \|Rx/s - Re_i\|_2^2 / 2)s = \left(1 - (1 \pm c\epsilon) \left(1 - xe_i / \|x\|_2\right)\right) \cdot (1 \pm \epsilon) \|x\|_2$$

$$= \ldots$$

$$= \pm c'\epsilon \|x\|_2 + xe_i,$$

for some absolute constant $c' > 1$. 