Lecture 5: Pairwise Independence and streaming

January 22, 2019

Lecturer: Eric Vigoda *Scribes: Aditi Laddha, He Jia*

Disclaimer: *These notes have not been subjected to the usual scrutiny reserved for formal publications.*

5.1 Pairwise Independent

Suppose $X_1, \ldots, X_n$ are n random variables on Ω.

Definition 5.1 *$X_1, \ldots, X_n$ are mutually independent if for all $\alpha_2, \ldots, \alpha_n \in \Omega$*

$$\Pr(X_1 = \alpha_1, \ldots, X_n = \alpha_n) = \prod_{i=1}^{n} \Pr(X_i = \alpha_i).$$

Definition 5.2 *$X_1, \ldots, X_n$ are pairwise independent if for all $i, j \in \{1, \ldots, n\}$, $\alpha, \beta \in \Omega$*

$$\Pr(X_i = \alpha_1, X_j = \beta) = \Pr(X_i = \alpha) \Pr(X_j = \beta).$$

Next we give two examples of pairwise independent variables.

5.1.1 Simple Construction

Suppose $X_1, \ldots, X_m \in \{0, 1\}$ are m mutually independent random bits with Bernoulli distribution of $p = 1/2$. We will make $2^m - 1$ pairwise independent random variables $Y_1, \ldots, Y_{2^m-1} \in \{0, 1\}$. Enumerate all non-empty subsets of $\{1, \ldots, m\}$ as $S_1, \ldots, S_{2^m-1}$. Let

$$Y_i = \bigoplus_{j \in S_i} X_j = \sum_{j \in S_i} X_j \mod 2.$$

Lemma 5.3 *$Y_1, \ldots, Y_{2^m-1}$ are pairwise independent.*

Proof: (Uniform) First show that $\Pr(Y_i = 0) = \Pr(Y_i = 1) = 1/2$.
Suppose $S_i = \{t_1, \ldots, t_\ell\} \subset \{1, \ldots, m\}$. Then

$$Y_i = \sum_{j=1}^{\ell} X_{t_j} \mod 2 = (\sum_{j=1}^{\ell-1} X_{t_j} \mod 2 + X_{t_\ell}) \mod 2.$$

Reveal $X_{t_1}, \ldots, X_{t_{\ell-1}}$. We can see

$$\Pr(Y_i = 1) = \Pr(X_{t_\ell} = 0 \cap \sum_{j=1}^{\ell-1} X_{t_j} \mod 2 = 1) + \Pr(X_{t_\ell} = 1 \cap \sum_{j=1}^{\ell-1} X_{t_j} \mod 2 = 0) = \frac{1}{2}.$$

(Pairwise independent) For any $i \neq j$, without loss of generality, we may assume $S_i \setminus S_j \neq \emptyset$. Then

$$\Pr(Y_i = \alpha, Y_j = \beta) = \Pr(Y_i = \alpha \mid Y_j = \beta) \Pr(Y_j = \beta) = \Pr(Y_i = \alpha \mid Y_j = \beta) \times \frac{1}{2}.$$

Take $t \in S_i \setminus S_j$. Reveal $\{X_1, \ldots, X_m\} \setminus \{X_t\}$. Then with probability 1/2, $X_t = 1$ and Y_j flips; with probability 1/2, $X_t = 0$ and Y_j is the same. Therefore,

$$\Pr(Y_i = \alpha \mid Y_j = \beta)) = \Pr(Y_i = \alpha \mid X_1, \ldots, X_m \setminus X_t) = \frac{1}{2}.$$

Thus

$$\Pr(Y_i = \alpha, Y_j = \beta) = \frac{1}{4}.$$

■

5.1.2 Hashing

For prime p, given a, b which are independent and uniform over $\{0, \ldots, p-1\}$. We construct $Y_0, \ldots, Y_{p-1}$ which are pairwise independent and uniform over $\{0, \ldots, p-1\}$. Namely, let

$$Y_i = a + ib \mod p.$$

Lemma 5.4 *$Y_0, \ldots, Y_{p-1}$ are pairwise independent.*

Proof: (Uniform) First show that $\Pr(Y_i = \alpha) = 1/p$.
For any $b, i, \alpha \in \{0, \ldots, p-1\}$

$$\Pr(Y_i = \alpha) = \Pr(a + ib \equiv \alpha \mod p) = \Pr(a \equiv \alpha - ib \mod p) = \frac{1}{p}$$

since there is a unique such $a \in \{0, \ldots, p-1\}$.
(Pairwise independent) Consider $i, j \in \{0, \ldots, p-1\}, i \neq j$ and $\alpha, \beta \in \{0, \ldots, p-1\}$, we will show

$$\Pr(Y_i = \alpha, Y_j = \beta) = \frac{1}{p^2}.$$

$$Y_i = \alpha \iff a + ib \equiv \alpha \mod p$$
$$Y_j = \beta \iff a + jb \equiv \beta \mod p$$

Thus

$$\begin{aligned} \alpha - \beta &\equiv (a + ib) - (a + jb) \mod p \\ \alpha - \beta &\equiv b(i - j) \mod p \\ b &\equiv \frac{\alpha - \beta}{i - j} \\ \text{and} \qquad a &\equiv \alpha - ib \mod p \end{aligned}$$

So there is a unique (a, b) pair so that $Y_i = \alpha, Y_j = \beta$. Therefore,

$$\Pr(Y_i = \alpha, Y_j = \beta) = \Pr(b \equiv \frac{\alpha - \beta}{i - j}, a \equiv \alpha - ib \mod p) = \frac{1}{p^2}.$$

■

5.2 Application: Streaming

Given a stream $S = \{s_1, s_2, \dots, s_m\}$ where $\forall i, s_i \in \{1, \dots, n\}$ and m is a very large number. The elements of the sequence are given one by one and cannot be stored. Define $f_i = |\{s_j \in S : s_j = i\}|$. For an integer $k \geq 1$, the kth frequency moment is defined as

$$F_k = \sum_{i=1}^{n} f_i^k$$

F_0 is the number of distinct elements in $S = |\{i : f_i > 0\}|$

Definition 5.5 *For integer k, $zeros(k) = \#$ trailing zeros in binary representation of $k = \max_{l \geq 0}\{l : 2^l$ divides $k\}$*

5.2.1 The AMS algorithm

Find a prime p such that $n \leq p < 2n$. Pad f such that $f_i = 0, \forall i \in \{n+1, \dots, p\}$.

Algorithm 1: AMS Algorithm for estimating F_0

input : $S = \{s_1, s_2, \dots, s_m\}$ where $s_i \in \{1, \dots, n\}$.
output: $\hat{d}$, a (3,0.96) approximation of F_0.

Choose a, b randomly from $\{0, 1, \dots, p-1\}$ and define $h(k) = a + kb \mod p$;
z = 0;
for $i \leftarrow 1$ **to** m **do**
    compute $zeros(h(s_i))$;
    if $zeros(h(s_i)) > z$ **then**
        $z = zeros(h(s_i))$
Output $2^{z+1/2}$;

5.3 Analysis of Algorithm

5.3.1 Space Complexity

$a, b \leq p \leq 2n$, so space needed to store a, b is $O(\log(n))$ and $z \leq \log(n)$, so space needed to store z is $O(\log\log(n))$.

Overall space needed is $O(\log(n))$.

5.3.2 Failure Probability

Let $F_0 = d = |\{i : f_i > 0\}|$ and let the output of the Algorithm 1 be $\hat{d}$.

Lemma 5.6

$$\Pr(\hat{d} \geq 3d \ \textit{or} \ d \leq \frac{\hat{d}}{3}) \leq 0.96.$$

For $k \in \{1, \dots, p\}$ and integer $l \geq 0$,

$$\Pr(zeros(h(k)) \geq l) = \Pr(\text{last } l \text{ bits of } h(k) \text{ are all } 0) = \frac{1}{2^l}$$

because $h(k)$ is a uniformly random bit string.

We cannot use Chernoff bounds on $h(k)$ as they are not mutually independent only pairwise independent.

For $k \in \{1, \ldots, n\}$ and integer $l \geq 0$, define a random variable

$$X_{k,l} = \begin{cases} 1 & \text{if } zeros(h(k)) \geq l \\ 0 & \text{otherwise} \end{cases}$$

Let

$$Y_l = \sum_{k: f_k > 0} X_{k,l}$$

If $2^{z+1/2}$ is the output of algorithm 1, then

$$z \geq l \Leftrightarrow Y_l > 0$$
$$z \leq l - 1 \Leftrightarrow Y_l = 0$$

For a fixed l, $X_{1,l}, \ldots, X_{n,l}$ are pairwise independent due to the construction of h from last section.

$$\mathbb{E}(X_{k,l}) = \Pr[zeros(h(k)) \geq l] = \frac{1}{2^l}$$

and

$$var(X_{k,l}) = \mathbb{E}(X_{k,l}^2) - \mathbb{E}(X_{k,l})^2 \leq \mathbb{E}(X_{k,l}^2) = \frac{1}{2^l}$$

$$\mathbb{E}(Y_l) = \mathbb{E}(\sum_{k: f_k > 0} X_{k,l}) = \sum_{k: f_k > 0} \mathbb{E}(X_{k,l}) = \frac{d}{2^l}$$

$$\text{Var}(Y_l) = \text{Var}(\sum_{k: f_k > 0} X_{k,l}) = \sum_{k: f_k > 0} \text{Var}(X_{k,l})$$

Linearity of variance holds over pairwise independent variables too. So, $var(Y_l) \leq \frac{d}{2^l}$.

By Markov's inequality,

$$\begin{aligned} \Pr(Y_l > 0) &= \Pr(Y_l \geq 1) \\ &\leq \frac{\mathbb{E}(Y_l)}{1} \\ &\leq d2^{-l} \end{aligned}$$

Using Chebyshev's inequality,

$$\begin{aligned} \Pr(Y_l = 0) &\leq \Pr(|Y_l - \mathbb{E}[Y_l]| \geq d2^{-l}) \\ &\leq \frac{\text{Var}(Y_l)}{(d2^{-l})^2} \\ &= \frac{2^l}{d} \end{aligned}$$

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

$$\begin{aligned}
\Pr(\hat{d} \geq 3d) &= \Pr[z \geq a] \\
&= \Pr(Y_a > 0) \\
&\leq \frac{d}{2^a} \\
&\leq \frac{2^{a+1/2}}{3.2^a} \\
&= \frac{\sqrt{2}}{3} \\
&< 0.48
\end{aligned}$$

and

$$\begin{aligned}
\Pr(\hat{d} \leq d/3) &= \Pr[z \leq b] \\
&= \Pr[Y_{b+1} = 0] \\
&\leq \frac{2^{b+1}}{d} \\
&\leq \frac{2^{b+1}}{3.2^{b+1/2}} \\
&= \frac{\sqrt{2}}{3} \\
&< 0.48
\end{aligned}$$

This gives a 3-approximation algorithm with error probability 0.96.

To get a 3-approximation algorithm with error probability $\leq \delta$, we can boost the algorithm by running it $r = O(\log(\frac{1}{\delta}))$. Let the outputs be $\hat{d}_1, \hat{d}_2, \ldots, \hat{d}_r$. Then with probability $\geq 1-\delta$, the median of $\hat{d}_1, \hat{d}_2, \ldots, \hat{d}_r$ is a 3-approximation of d. This takes $O(\log(\frac{1}{\delta}) \log(n))$ bits. For every run of the AMS algorithm, we select a new hash function making each run independent of the others.

References

[1] Alon, N. & Matias, Y. & Szegedy, M. The space complexity of approximating the frequency moments. *Journal of Computer and system sciences*, pages 137–147, 1999