CS 598CSC: Algorithms for Big Data	**Lecture date: Sept 2, 2014**
Instructor: Chandra Chekuri	**Scribe: Chandra Chekuri**

1 AMS Sampling

We have seen reservoir sampling and the related weighted sampling technique to obtain independent samples from a stream without the algorithm knowing the length of the stream. We now discuss a technique to sample from a stream $\sigma = a_1, a_2, \ldots, a_m$ where the tokens a_j are integers from $[n]$ and we wish to estimate a function

$$g(\sigma) := \sum_{i \in [n]} g(f_i)$$

where f_i is the frequency of i and g is a real-valued function such that $g(0) = 0$. A natural example is to estimate frequency moments $F_k = \sum_{i \in [n]} f_i^k$; here we have $g(x) = x^k$, a convex function for $k \geq 1$. Another example is the empirical entropy of σ defined as $\sum_{i \in [n]} p_i \log p_i$ where $p_i = \frac{f_i}{m}$ is the empirical probability of i; here $g(x) = x \log x$.[1]

AMS sampling from the famous paper [**?**] gives an unbiased estimator for $g(\sigma)$. The estimator is based on a random variable Y defined as follows. Let J be a uniformly random sample from $[m]$. Let $R = |\{j \mid a_j = a_J, J \leq j \leq m\}|$. That is, R is the count of the number of tokens after J that are for the same coordinate. Then, let Y the estimate defined as:

$$Y = m(g(R) - g(R-1)).$$

The lemma below shows that Y is an unbiased estimator of $g(\sigma)$.

Lemma 1

$$\mathbf{E}[Y] = g(\sigma) = \sum_{i \in [n]} g(f_i).$$

Proof: The probability that $a_J = i$ is exactly f_i/m since J is a uniform sample. Moreover if $a_J = i$ then R is distributed as a uniform random variable over $[f_i]$.

$$\begin{aligned}
\mathbf{E}[Y] &= \sum_{i \in [n]} \Pr[a_J = i] \mathbf{E}[Y | a_J = i] \\
&= \sum_{i \in [n]} \frac{f_i}{m} \mathbf{E}[Y | a_J = i] \\
&= \sum_{i \in [n]} \frac{f_i}{m} \sum_{\ell=1}^{f_i} m \frac{1}{f_i} \left(g(\ell) - g(\ell - 1)\right) \\
&= \sum_{i \in [n]} g(f_i).
\end{aligned}$$

□

One can estimate Y using small space in the streaming setting via the reservoir sampling idea for generating a uniform sample. The algorithm is given below; the count R gets reset whenever a new sample is picked.

[1] In the context of entropy, by convention, $x \log x = 0$ for $x = 0$.

```
AMSEstimate:
s ← null
m ← 0
R ← 0
While (stream is not done)
   m ← m + 1
   a_m is current item
   Toss a biased coin that is heads with probability 1/m
   If (coin turns up heads)
         s ← a_m
         R ← 1
   Else If (a_m == s)
         R ← R + 1
endWhile
Output m(g(R) − g(R − 1))
```

To obtain a (ϵ, δ)-approximation via the estimator Y we need to estimate $\mathbf{Var}[Y]$ and apply standard tools. We do this for frequency moments now.

1.1 Application to estimating frequency moments

We now apply the AMS sampling to estimate F_k the k'th frequency moment for $k \geq 1$. We have already seen that Y is an exact statistication estimator for F_k when we set $g(x) = x^k$. We now estimate the variance of Y in this setting.

Lemma 2 *When $g(x) = x^k$ and $k \geq 1$,*

$$\mathbf{Var}[Y] \leq kF_1F_{2k-1} \leq kn^{1-\frac{1}{k}}F_k^2.$$

Proof:

$$\begin{aligned}
\mathbf{Var}[Y] &\leq \mathbf{E}[Y^2] \\
&\leq \sum_{i \in [n]} \Pr[a_J = i] \sum_{\ell=1}^{f_i} \frac{m^2}{f_i}\left(\ell^k - (\ell-1)^k\right)^2 \\
&\leq \sum_{i \in [n]} \frac{f_i}{m} \sum_{\ell=1}^{f_i} \frac{m^2}{f_i}(\ell^k - (\ell-1)^k)(\ell^k - (\ell-1)^k) \\
&\leq m \sum_{i \in [n]} \sum_{\ell=1}^{f_i} k\ell^{k-1}(\ell^k - (\ell-1)^k) \qquad \text{(using } (x^k - (x-1)^k) \leq kx^{k-1}\text{)} \\
&\leq km \sum_{i \in [n]} f_i^{k-1} f_i^k \\
&\leq kmF_{2k-1} = kF_1F_{2k-1}.
\end{aligned}$$

We now use convexity of the function x^k for $k \geq 1$ to prove the second part. Note that $\max_i f_i = F_\infty$.

$$F_1F_{2k-1} = (\sum_i f_i)(\sum_i f_i^{2k-1}) \leq (\sum_i f_i)F_\infty^{k-1}(\sum_i f_i^k) \leq (\sum_i f_i)(\sum_i f_i^k)^{\frac{k-1}{k}}(\sum_i f_i^k).$$

Using the preceding inequality, and the inequality $(\sum_{i=1}^n x_i)/n \leq ((\sum_{i=1}^n x_i^k)/n)^{\frac{1}{k}}$ for all $k \geq 1$ (due to the convexity of the function $g(x) = x^k$), we obtain that

$$F_1 F_{2k-1} \leq (\sum_i f_i)(\sum_i f_i^k)^{\frac{k-1}{k}}(\sum_i f_i^k) \leq n^{1-1/k}(\sum_i f_i^k)^{\frac{1}{k}}(\sum_i f_i^k)^{\frac{k-1}{k}}(\sum_i f_i^k) \leq n^{1-1/k}(\sum_i f_i^k)^2.$$

□

Thus we have $\mathbf{E}[Y] = F_k$ and $\mathbf{Var}[Y] \leq kn^{1-1/k}F_k^2$. We now apply the trick of reducing the variance and then the median trick to obtain a high-probability bound. If we take h independent estimators for Y and take their average the variance goes down by a factor of h. We let $h = \frac{c}{\epsilon^2}kn^{1-1/k}$ for some fixed constant c. Let Y' be the resulting averaged estimator. We have $\mathbf{E}[Y'] = F_k$ and $\mathbf{Var}[Y'] \leq \mathbf{Var}[Y]/h \leq \frac{\epsilon^2}{c}F_k^2$. Now, using Chebyshev, we have

$$\Pr[|Y' - \mathbf{E}[Y']| \geq \epsilon \mathbf{E}[Y']] \leq \mathbf{Var}[Y']/(\epsilon^2 \mathbf{E}[Y']^2) \leq \frac{1}{c}.$$

We can choose $c = 3$ to obtain a $(\epsilon, 1/3)$-approximation. By using the median trick with $\Theta(\log \frac{1}{\delta})$ independent estimators we can obtain a (ϵ, δ)-approximation. The overall number of estimators we run independently is $O(\log \frac{1}{\delta} \cdot \frac{1}{\epsilon^2} \cdot n^{1-1/k})$. Each estimator requires $O(\log n + \log m)$ space since we keep track of one index from $[m]$, one count from $[m]$, and one item from $[n]$. Thus the space usage to obtain a (ϵ, δ)-approximation is $O(\log \frac{1}{\delta} \cdot \frac{1}{\epsilon^2} \cdot n^{1-1/k} \cdot (\log m + \log n))$. The time to process each stream element is also the same.

The space complexity of $\tilde{O}(n^{1-1/k})$ is not optimal for estimating F_k. One can achieve $\tilde{O}(n^{1-2/k})$ which is optimal for $k > 2$ and one can in fact achieve poly-logarithmic space for $1 \leq k \leq 2$. We will see these results later in the course.

Bibliographic Notes: See Chapter 1 of the draft book by McGregor and Muthukrishnan; see the application of AMS sampling for estimating the entropy. See Chapter 5 of Amit Chakrabarti for the special case of frequency moments explained in detail. In particular he states a clean lemma that bundles the variance reduction technique and the median trick.