Algorithms for Large/Real-time Data Set

Lecturer: Sanjeev Arora, COS 521, Fall 2005

Scribe Notes: Chee Wei Tan

1 Computing Frequency Moments

We begin with an example: Suppose we have a router that wants to compute the frequency moments of Internet Protocol (IP) destination addresses from a sequence of incoming data packets in real-time. Each data item has a label $i \in \{1, 2, \ldots, n\}$. In our example, label i corresponds to an IP destination address. Let m_i be the number of items with label i. Define

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

Our goal is to compute $F_k,\ k = 1, 2, \ldots$.

Computing F_1 is trivial as we only need to keep a counter for each data item. Space requirement is $O(\log A)$ where A is the length of the sequence of data. Then, how do we compute F_2? A trivial approach is to maintain a counter for each m_i, but the space complexity becomes $O(n \log A)$.

We describe below a $(1+\epsilon)$ approximation algorithm to compute F_2 [1] by maintaining a single counter variable. It is shown in [1] that the space complexity is $O(\frac{1}{\epsilon \log n \log})$ for $k = 2$, and $O(n^{1-1/k})$ for $n \geq 3$, and these bounds are tight.

Input: A random hash function h that requires a random seed of $O(\log n)$ bits and the label i as input, and ouput a random variable ϵ_i.

Output: An unbiased estimate of F_2.

Step 1: Initialize the counter variable, `counter` to 0.

Step 2: For each received item with label i, $\texttt{counter} \leftarrow \texttt{counter} + \epsilon_{\texttt{i}}$.

At the end of the input sequence, we have the counter value, $\texttt{counter} = \sum_{i=1}^{n} m_i \epsilon_i$.

Before we provide an analysis of the above algorithm, we first introduce the notion of 4-wise independence of a sequence of random variables.

Definition 1. A sequence of random variables taking values in $\{-1, 1\}$, i.e., $\epsilon_1, \epsilon_2, \ldots, \epsilon_n \in \{-1, 1\}$ is 4-wise independent if any 4-tuple of random variables $\epsilon_i, \epsilon_j, \epsilon_k, \epsilon_l$ is jointly independent.

It can be shown that a 4-wise independence sequence of random variables is also k-wise independent for k less than 4.

Assuming that the sequence of random variables ϵ_i satisfies 4-wise independence, we show that the above algorithm is an unbiased estimator of the expected value of F_2.

$$\begin{aligned} \text{Let} \quad X &= (\texttt{counter})^2 \\ &= \left(\sum_{i=1}^{n} m_i \epsilon_i\right)^2. \end{aligned}$$

Taking expectation over our choice of the random seed,

$$\begin{aligned} \text{Let} \quad E[X] &= E\Big[(\sum_{i=1}^{n} m_i \epsilon_i)^2\Big] \\ &= E\Big[\sum_{i=1}^{n}\sum_{j=1}^{n} m_i m_j \epsilon_i \epsilon_j\Big] \\ &= \sum_{i=1}^{n}\sum_{j=1}^{n} m_i m_j E[\epsilon_i \epsilon_j] \\ &= \sum_{i=1}^{n} m_i^2, \end{aligned} \tag{1}$$

where the last equality follows from the fact that ϵ_i and ϵ_j are pairwise independent for $i \neq j$ and each of them has zero mean, i.e., $E[\epsilon_i \epsilon_j] = E[\epsilon_i^2] = 1$ if $i = j$, and $E[\epsilon_i \epsilon_j] = E[\epsilon_i]E[\epsilon_j] = 0$ if $i \neq j$. Hence, X is an unbiased estimator of F_2.

To access how good the algorithm gives an estimate of F_2, we compute the variance of the estimation.

$$\begin{aligned} \text{Let} \quad Var[X] &= E[X^2] - (E[X])^2 \\ &= E\Big[(\sum_{i=1}^{n} m_i \epsilon_i)^4\Big] - \left(\sum_{i=1}^{n} m_i^2\right)^2 \\ &= E\Big[\sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{k=1}^{n}\sum_{l=1}^{n} m_i m_j m_k m_l \epsilon_i \epsilon_j \epsilon_k \epsilon_l\Big] - \left(\sum_{i=1}^{n} m_i^4 - 2\sum_{i=1}^{n}\sum_{j=1, j\neq i}^{n} m_i^2 m_j^2\right) \end{aligned}$$

But,

$$\begin{aligned} E\Big[\sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{k=1}^{n}\sum_{l=1}^{n} m_i m_j m_k m_l \epsilon_i \epsilon_j \epsilon_k \epsilon_l\Big] &= E\Big[\sum_{i=1}^{n} m_i^4 \epsilon_i^4\Big] + 6E\Big[\sum_{i=1}^{n}\sum_{j=1, j\neq i}^{n} m_i^2 m_j^2 \epsilon_i^2 \epsilon_j^2\Big] \\ &= \sum_{i=1}^{n} m_i^4 + 6\sum_{i=1}^{n}\sum_{j=1, j\neq i}^{n} m_i^2 m_j^2 \end{aligned} \tag{2}$$

where we make use of the fact that $E[\epsilon_i \epsilon_j \epsilon_k \epsilon_l] = 0$ if any index of i, j, k, l occurs an odd number of times.

Hence, we have

$$\begin{aligned} Var[X] &= \sum_{i=1}^{n} m_i^4 + 6 \sum_{i=1}^{n} \sum_{j=1, j\neq i}^{n} m_i^2 m_j^2 - (\sum_{i=1}^{n} m_i^4 - 2 \sum_{i=1}^{n} \sum_{j=1, j\neq i}^{n} m_i^2 m_j^2) \\ &= 4 \sum_{i=1}^{n} \sum_{j=1, j\neq i}^{n} m_i^2 m_j^2 \leq 2(E[X])^2 = 2F_2^2, \end{aligned} \quad (3)$$

where the last inequality follows from the fact that

$$\left(\sum_{i=1}^{n} m_i^2 \right)^2 - 2 \sum_{i=1}^{n} \sum_{j=1, j\neq i}^{n} m_i^2 m_j^2 = \sum_{i=1}^{n} m_i^4 \geq 0.$$

It is clear from above that the estimation error can be large. However, we can further reduce the variance of the estimation error by *repeated sampling*, i.e., take k independent copies of X and take the average of $X_1, \ldots, X_k$. Let $Y = \sum_{i=1}^{k} X_i$. Then we have

$$E[Y] = \sum_{i=1}^{k} E[X_i] = kF_2 \quad \text{and} \quad Var[Y] = \sum_{i=1}^{k} Var[X_i] \leq 2kF_2^2,$$

where we make use of the fact that X_i's are independent and the last inequality follows from (3). Hence, by Chebyshev's inequality, the average sampled value of F_2 is

$$\frac{Y}{k} \sim O\left(E[X] \pm \frac{\sqrt{2k} F_2}{k} \right)$$

and we can get a good approximation of $(1 \pm \epsilon) F_2$ by selecting k such that $k \geq 2/\epsilon^2$. The space requirement for repeated sampling is $O(\frac{1}{\epsilon} \log n)$.

For more detailed analysis of computing the frequency moment, please refer to [1].

2 Dimension Reduction

This section finds practical application in image processing where we want to store large vectors with each entry containing a pixel value. Suppose we are given vectors $u_1, u_2, \ldots, u_m \in \mathbf{R}^n$ where n is very large. We desire a more compact representation of these vectors, i.e., $u_1', u_2', \ldots, u_m' \in \mathbf{R}^d$ where $d \ll n$ such that

$$\|u_i' - u_j'\|_2 \in (1 \pm \epsilon) \|u_i - u_j\|_2,$$

where $\|.\|_2$ denotes the Euclidean norm.

It can be shown that this is possible if $d \sim O(\frac{\log m + \log n}{\epsilon^2})$. We next illustrate an algorithm to compute $u_i^{'}, \forall i$ that gives a good approximation to the above criteria.

Step 1: Pick d random vectors of dimension n, e.g.,

$$[\epsilon_{11} \quad \epsilon_{12} \quad \ldots \quad \epsilon_{1n}]^T, [\epsilon_{21} \quad \epsilon_{22} \quad \ldots \quad \epsilon_{2n}]^T, \ldots, [\epsilon_{n1} \quad \epsilon_{n2} \quad \ldots \quad \epsilon_{nn}]^T,$$

where $\epsilon_{ij} \in \{-1, 1\}$ are independent random variables for all i and j, and $[.]^T$ denotes the transpose operator.

Step 2: For each vector u, use the random linear map $\mathbf{R}^n \to \mathbf{R}^d$ in Step 1 to get a column vector

$$u^{'} = [[\epsilon_{11} \quad \epsilon_{12} \quad \ldots \quad \epsilon_{1n}]^T u \ \ldots \ [\epsilon_{n1} \quad \epsilon_{n2} \quad \ldots \quad \epsilon_{nn}]^T u]^T.$$

For every two vector $u^{'}$ and $v^{'}$ obtained above, let

$$\|u^{'} - v^{'}\|_2^2 = \sum_{l=1}^{d} \left(\sum_{i=1}^{n} \epsilon_{il}(u_i - v_i) \right)^2.$$

Taking expectation,

$$\begin{aligned} E[\|u^{'} - v^{'}\|_2^2] &= E\left[\sum_{l=1}^{d} \left(\sum_{i=1}^{n} \epsilon_{il}(u_i - v_i) \right)^2 \right] \\ &= d \sum_{i=1}^{n} (u_i - v_i)^2 \\ &= d\|u - v\|_2^2. \end{aligned} \tag{4}$$

But we need $\binom{m}{2}$ different $\|u - v\|_2$ such that $\|u_i^{'} - v_j^{'}\|_2 \leq (1 \pm \epsilon)\|u_i - v_j\|_2$. Lastly, we can use Chernoff bound to show that

$$Pr\{\text{any of the} \binom{m}{2} \text{different} \ \|u - v\|_2 \ \text{deviates by more than} \ 1 \pm \epsilon\} < \frac{1}{m^3}.$$

For more details, please refer to Sanjeev's online note on dimension reduction.

References

[1] Alan, Matias and Szegedy "The space complexity of approximating the frequency moments", STOC, 1996