Advanced Algorithms 12/05, 2017

Lecture 1

Lecturer: Mohsen Ghaffari *Scribe: Roko Zaja*

This draft has not been checked by the lecturer yet.

1 Streaming & Sketching Algorithms 2 — Moment Estimation and Space Complexity Lower Bounds

1.1 Alon-Matias-Szegedy's 2nd Moment Estimator

We consider a stream of numbers $a_1, a_2, ..., a_m \in \{1, \ldots, n\}$. Let f_j denote the number of times that j appears in the stream. We want to estimate the 2nd moment:

$$F^2 = \sum_{j=1}^{n} f_j^2.$$

The AMS Algorithm:

- Pick a random number $r_j \in \{+1, -1\}$ for each $j \in \{1, \ldots, n\}$.
- Maintain $Z = \sum\limits_{j=1}^{n} r_j f_j$ throughout the stream.
- Output Z^2.

Claim 1. *Z^2 is an unbiased estimator of F^2.*

Proof.

$$E[(\sum_{j=1}^{n} r_j f_j)] = E[\sum_{j=1}^{n} rj^2 fj^2 + 2\sum_{j_1<j_2} r_{j_1} r_{j_2} f_{j_1} f_{j_2}] = \sum_{j=1}^{n} f_j^2$$

□

$Var(Z^2) = E[Z^4] - E[Z^2]^2 = \sum\limits_{j=1}^{n} f_j^4 + 6\sum\limits_{j=1}^{n} f_{j_1}^2 f_{j_2}^2 - (\sum\limits_{j=1}^{n} f_j^2)^2 = 4\sum\limits_{j=1}^{n} f_{j_1}^2 f_{j_2}^2 \le 2(\sum\limits_{j=1}^{n} f_j^2)^2$

$P[|Z^2 - F^2| \ge \epsilon F^2] \le \frac{2}{\epsilon^2}$ (using Chebyshev's inequality)

We can repeat this k times and take the average. Variance is now $\le \frac{2}{k\epsilon^2}$. By setting $k = \frac{C}{\epsilon^2}$ we get a constant bound on the probability. We also notice that 4-wise independence between $r_j s$ is sufficient in our calculation. For $k$wise independence, $k \log n$ bits suffice. In our case, it follows that $O(\log n)$ bits suffice.
The basic idea in our algorithm works in the similar fashion when trying to estimate F_k [AMS96].

Lemma 2. *(Johnson-Lindenstrauss) For any set of n points in R^D there exist a mapping $f : R^n \to R^k$ ($k = O(\frac{\log n}{\epsilon^2})$ such that for every pair $u, v \in S$ we have:*

$$(1-\epsilon)||v-u|| \le ||f(v) - f(u)|| \le (1+\epsilon)||v-u||.$$

General Proof Outline: Construct a random projection over k dimensional subspaces. prove that the expected value of the Euclidian distance of the random projection is equal to the Euclidian distance of the original subspace. prove that the variance of the Euclidian distance is greater than the specified error factor only with a probability $\frac{2}{n^2}$ such that the union bound of this probability across all pairs of points is less than $1 - \frac{1}{n}$.

1.2 Distinct elements

Theorem 3. *Any deterministic streaming algorithm that computes a $\frac{9}{8}-$approximation of DE needs $\Omega(n)$ bits of memory.*

The notion of communication complexity was introduced by Yao in 1979,[1] who investigated the following problem involving two separated parties (Alice and Bob). Alice receives an n-bit string x and Bob another n-bit string y, and the goal is for one of them (say Bob) to compute a certain function $f(x, y)$ with the least amount of communication between them. For sending X from Alice to Bob we need $\log_2 |X|$ bits.

Lemma 4. *There exists a collection X of subsets of $\{1, \dots, n\}$ such that:*
(1)$|X| = 2^{\Omega(n)}$;
(2)$\forall S_1, S_2 \in X$ we have $|S_1 \setminus S_2| \geq \frac{n}{8}$.

Construction:
(1) greedily
(2) random: Pick each set S_i by picking each j with probability $\frac{1}{2}$. From Chernoff we get: $P[S_i \setminus S_j] \leq \frac{n}{8}$ is exponentially small e^{-cn}.

1.3 Graph streaming and sketching algorithms

Over the last decade, there has been considerable interest in designing algorithms for processing massive graphs in the data stream model. The original motivation was two-fold: a) in many applications, the dynamic graphs that arise are too large to be stored in the main memory of a single machine and b) considering graph problems yields new insights into the complexity of stream computation. However, the techniques developed in this area are now finding applications in other areas including data structures for dynamic graphs, approximation algorithms, and distributed and parallel computation.

Our problem is to identify the connected components.

Definition 5. ***(Connected components)*** *Given an undirected graph $G = (V, E)$ with n nodes and m edges, connected components (CC) is the problem of determining a function $c : V \to \{0, \dots, n-1\}$ such that $\forall u, v \in V, c(u) = c(v)$, if and only if u and v are connected by a path in G.*

Outgoing edge queries:
Given a set S, is there an edge going out of S?

(Promise: There is either a single edge going out of S or no edge.)

(u, v) identified by concatenating the IDs of its two endpoints($O(\log n)$ bit identifier).

References

[AMS96] Noga Alon, Yossi Matias, and Mario Szegedy. The space complexity of approximating the frequency moments. In *Proceedings of the twenty-eighth annual ACM symposium on Theory of computing*, pages 20–29. ACM, 1996.