CS261 Winter 2018 - 2019
Lecture 16: Sketching (Part 2)

Instructor: Ashish Goel
Scribe: Kaidi Yan
Edited: Geoff Ramseyer

Feburary 27, 2019

1 Distinct-Sum Problem

The distinct-sum problem states the following: Given a stream of data $S = a_1, a_2, ..., a_t, ...$, find the sum of all distinct elements up till time t, assuming each a_i is a positive integer.

The distinct-sum problem is reducible to a count-distinct problem. Given S, construct a new stream S' such that for each $a_i \in S$ where $i = 1, ..., t$, we put elements $\langle a_i, 1\rangle, \langle a_i, 2\rangle, ..., \langle a_i, a_i\rangle$ into S'. We notice that each distinct element a_i in S contributes a_i distinct elements in S'. Therefore the distinct-sum problem on S is reduced to the count-distinct problem on S', which we can apply the count-min sketch to solve.

Unfortunately, computing and updating the sketch can take a long time, as the number of elements of S' generated by a_i in S is a_i, which can be large if a_i is large. A method for sampling the minimum value of a_i random variables in one action would be quite useful. For this purpose, consider the following.

$$q(k) = \min\{h(k, 1), h(k, 2), ..., h(k, k)\}$$

where h is a consistent uniform [0,1] hash function, as defined in the last lecture. Note that $h(k, 1), h(k, 2), ..., h(k, k)$ are i.i.d. Hence, the probability that $q(k) > x$ $(0 \leq x \leq 1)$ is equal to $(1 - x)^k$. Let $z = (1 - x)^k$. Then $x = 1 - z^{1/k}$ and thus $\Pr[q(k) > 1 - z^{1/k}] > z$[1]. We can now compute $q(k)$ using the following function:

Thus, instead of evaluating k hash functions, this algorithm now only needs to evaluate $q(k)$ once. The total number of hash functions we need to evaluate now becomes proportional to the number of elements in S, which is much less than the sum of distinct elements.

[1] In general, if F is the cumulative probability distribution of a random variable X, then you can obtain a sample from this distribution by setting $X = F^{-1}(U)$ where U is a uniform random variable such as the one generated by random().

```
function q(k)
    srandom(k)
    z = random()
    return 1 − z^{1/k}
end function
```

2 Frequency Moment Estimation

Suppose we have a stream of ⟨key, value⟩ pairs $S = \langle k_1, v_1\rangle, \langle k_2, v_2\rangle, ..., \langle k_t, v_t\rangle, ...$, where a key can appear multiple times in the stream (so it's different from a key in a hashtable). For example, a key can be the source and destination IP addresses for a particular TCP/IP session, and its associated value is the amount of traffic sent during that session. Now given stream S, a time t and an integer p, we are asked to compute $F_p(t)$, which is defined as:

$$F_p(t) = \sum_{k \in \{k_1, k_2, ..., k_t\}} \left(|\sum_{i \in [1,t]: k_i = k} v_i| \right)^p$$

where $\{k_1, k_2, ..., k_t\}$ is the set containing $k_1, k_2, ..., k_t$ (note that a key can appear multiple times so the set size can be smaller than t). We call $F_p(t)$ the p-th *frequency moment* of S up till time t. (In the literature, this problem is often formulated with each $v_i = 1$).

When $p = 0$ and $v_i = 1$ for all i, this problem is exactly the problem of counting the number of distinct keys which can solved by the count-min sketch. Practical applications (like analyzing databases and internet traffic patterns) are often concerned with the first and second frequency moments of the data. The next section defines a *p-stable distribution* – this will be useful in sketching $F_p(t)$.

2.1 p-stable Distribution

Here's the definition of a p-stable distribution. For these notes, $p \in (0, 2]$.

Definition 2.1 *A probability distribution D is p-stable if for all $a_1, a_2, ...a_k \in \mathbb{R}$, if $Z_1, Z_2, ..., Z_k$ are i.i.d. random variables with distribution D, then $\sum_{i=1}^{k} a_i Z_i$ has the same distribution as $(\sum_{i=1}^{k} |a_i|^p)^{1/p} Z$ where Z has distribution D.*

Note that $\sum_{i=1}^{k} a_i Z_i$ is also a random variable following some probability distribution.

Consider the case where $p = 2$ and $k = 2$. In order for D to be 2-stable, we need to have $a_1 Z_1 + a_2 Z_2 = \sqrt{a_1^2 + a_2^2} Z$ where $Z_1, Z_2, Z \sim D$ and Z_1, Z_2 are i.i.d. If D is $N(0, \sigma^2)$, which is a normal distribution with mean 0 and standard deviation σ, then $a_1 Z_1 + a_2 Z_2$ also follows a normal distribution. Its mean and variance are:

$$\mathrm{E}[a_1 Z_1 + a_2 Z_2] = a_1 \cdot 0 + a_2 \cdot 0 = 0$$
$$\mathrm{Var}[a_1 Z_1 + a_2 Z_2] = a_1^2 \sigma^2 + a_2^2 \sigma^2 = (a_1^2 + a_2^2)\sigma^2$$

Thus $a_1 Z_1 + a_2 Z_2 = \sqrt{a_1^2 + a_2^2} Z$ where $Z \sim N(0, \sigma^2)$. This argument can be extended to any k, and hence $N(0, \sigma^2)$ is 2-stable. In fact, the normal distributions with zero mean are the only 2-stable distributions, although we omit the proof here. For convenience, we let $\sigma^2 = 1$, so the rest of these notes will use $N(0, 1)$ as a 2-stable distribution.

2.2 Using 2-stable Distributions to Estimate $F_2(t)$

In order to use 2-stable distributions to estimate $F_2(t)$, define the following sketch (note that we are back to using σ as a notation for the sketch, not the standard deviation):

$$\sigma(S) = \langle \sigma_1(S), \sigma_2(S), ..., \sigma_M(S) \rangle$$

where M is a constant to be determined later. For $\sigma_j(S)$ $(j = 1, ..., M)$, assume we have a consistent Gaussian hash function $h_j(k)$ over keys, i.e., $h_j(k) \sim N(0, 1)$ and $h_j(k), h_j(k')$ are i.i.d. unless $k = k'$. Let

$$\sigma_j(S) = \sum_{i=1}^{t} h_j(k_i) v_i = \sum_{k \in \{k_1, ..., k_t\}} h_j(k_i) \sum_{i \in [1,t]: k_i = k} v_i \tag{1}$$

Since $h_j(k_i)$ are i.i.d. and drawn from a 2-stable distribution, we get:

$$\sum_{k \in \{k_1, ..., k_t\}} h_j(k_i) \sum_{i \in [1,t]: k_i = k} v_i = \sqrt{\sum_{k \in \{k_1, ..., k_t\}} (\sum_{i \in [1,t]: k_i = k} v_i)^2} Z \tag{2}$$

where $Z \sim N(0, 1)$. Squaring both side of (2) and combined with (1) we get:

$$\sigma_j(S)^2 = \sum_{k \in \{k_1, ..., k_t\}} (\sum_{i \in [1,t]: k_i = k} v_i)^2 Z^2$$

Notice that $\sum_{k \in \{k_1, ..., k_t\}} (\sum_{i \in [1,t]: k_i = k} v_i)^2$ is exactly $F_2(t)$. Thus, if we square $\sigma_j(S)$ for all j and sum them up, we get:

$$\sum_{j=1}^{M} \sigma_j(S)^2 = F_2(t) \sum_{j=1}^{M} Z_j^2$$

where $Z_j \sim N(0, 1)$. We know $\mathrm{E}[Z_j^2] = \mathrm{Var}[Z_j] - \mathrm{E}[Z_j]^2 = 1 - 0^2 = 1$, and hence $\mathrm{E}[\sum_{j=1}^{M} Z_j^2] = M$. We define the following estimator:

$$\widehat{F_2(t)} = \frac{\sum_{j=1}^{M} \sigma_j(S)^2}{M}$$

The guarantee given by the median lemma holds here for the mean — if $M > \frac{c}{\delta^2} \log \frac{1}{\epsilon}$ for some constant c, then $\widehat{F_2(t)} \in [(1 - \delta) F_2(t), (1 + \delta) F_2(t)]$ with probability at least $1 - \epsilon$.