Streaming Algorithms

Meng-Tsung Tsai

03/13/2018

References

- "Pairwise Independence and Derandomization," Luby and Wigderson (2005)
- "Sketch Techniques for Approximate Query Processing," Cormode

Pairwise Independence

Family of functions

Let H be a family of functions, e.g. $H = \{h_{a,b}(x) : a, b \in \mathbf{Z}_p\}$ where

$$h_{a,b}(x) = ax+b \bmod p.$$

Let h be a function sampled uniformly at random from H. We say H is **pairwise independent** if for each $i \neq j \in \mathbf{Z}_p$

$$\Pr[h(i) = a \wedge h(j) = b] = 1/p^2.$$

Family of functions

Let H be a family of functions, e.g. $H = \{h_{a,b}(x) : a, b \in \mathbf{Z}_p\}$ where

$$h_{a,b}(x) = ax+b \bmod p.$$

Let h be a function sampled uniformly at random from H. We say H is **pairwise independent** if for each $z_1 \neq z_2 \in \mathbf{Z}_p$, for each $y_1, y_2 \in \mathbf{Z}_p$

$$\Pr[h(z_1) = y_1 \wedge h(z_2) = y_2] = 1/p^2.$$

<u>Theorem 1</u>. $H = \{h_{a,b}(x) : a, b \in \mathbf{Z}_p\}$ is pairwise independent.

Illustration of Theorem 1

h	**0**	**1**	**...**
$h_{0,0}$	0	0	
$h_{0,1}$	1	1	
$h_{0,2}$	2	2	
$h_{1,0}$	0	1	
$h_{1,1}$	1	2	
$h_{1,2}$	2	0	
$h_{2,0}$	0	2	
$h_{2,1}$	1	0	
$h_{2,2}$	2	1	

Proof of Theorem 1

Recall that $H = \{h_{a,b}(x) : a, b \in \mathbf{Z}_p\}$ where $h_{a,b}(x) = ax+b \bmod p$.

Let $z_1 \neq z_2$ in $\mathbf{Z}_p$. If $h = h_{a,b}(x)$ for $a = 0$, then $h(z_1) = b = h(z_2)$.

$\Rightarrow \Pr[h(z_1) = b \wedge h(z_2) = b] = \Pr[h = h_{0,b}] = 1/p^2$.

If $h = h_{a,b}(x)$ for $a \neq 0$, then for each (y_1, y_2) where $y_1 \neq y_2$ there exists a unique $h_{a,b}$ so that $h_{a,b}(z_1) = y_1$ and $h_{a,b}(z_2) = y_2$. (Why?)

For each $a \neq 0$, b in $\mathbf{Z}_p$,
$h_{a,b}(z_1)$ and $h_{a,b}(z_2)$ maps (z_1, z_2) into (y_1, y_2) for some $y_1 \neq y_2$.

For each $y_1 \neq y_2$,
some $h_{a,b}$ for $a \neq 0$ maps (y_1, y_2) into (z_1, z_2).

Implications of Theorem 1

For each p, there exists a pairwise independent family H of functions so that each function in H can be represented in O(log p) space. (How?)

Of course there are many pairwise independent families, but few of which can use logarithmic space to represent each function in them.

Count-Min Sketch

Problem Defintion

Input: a sequence of n elements $e_1, e_2, \ldots, e_n$ where each e_i in $[U] = \{1, \ldots, U\}$. Let |U| be a prime w.l.o.g.

Output: for each $k \in [U]$, output the frequency $f(k) = \sum_{i \in [n]} \mathbf{1}[e_i = k]$. In words, f(k) is the number of e_i in the sequence that has value k.

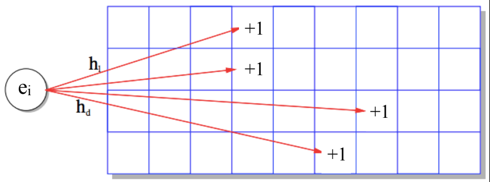
Goal: using o(U log n) bits to get an approximate $\hat{f}(k)$ for each f(k).

Algorithm

Sample functions $h_1, h_2, \ldots, h_d$ independently, uniformly at random from $H_w = \{h_{a,b}(x)\%w : a, b \in \mathbf{Z}_p\}$ where $w \ll p = |U|$.

$T[d][w] \leftarrow \{0, 0, \ldots, 0\}$.

foreach e_i {
  foreach h_j {
    $T[j][h_j(e_i)]$ ++;
  }
}



let $\hat{f}(k) = \min_{j \in [d]} \{T[j][h_j(k)]\}$;

For each hash function h_j

Observe that $T[j][a] = \sum_{i \in [n]} \mathbf{1}[h_j(e_i) = a]$ and therefore

$$T[j][h_j(k)] \geq f(k).$$

Because h_j is sampled from a pairwise independent H, we have

$$\Pr[h_j(\ell) = h_j(k)] = 1/w \text{ for every } \ell \neq k.$$

Hence, the expected noise $E[\mathscr{E}_j] = \sum_{\ell \neq k,\, h_j(\ell) = h_j(k)} E[T[j][h_j(\ell)]] = n/w$.

Let $w = 2/\varepsilon$. Then $E[\mathscr{E}_j] = \varepsilon n/2$. By Markov inequality,

$$\Pr[\mathscr{E}_j \geq \varepsilon n] \leq 1/2.$$

For all hash functions $h_1, h_2, ..., h_d$

$\Pr[\min_{j\in[d]} \mathscr{E}_j \geq \varepsilon n]$

$= \prod_{j\in[d]} \Pr[\mathscr{E}_j \geq \varepsilon n]$ (Why?)

$\leq 1/2^d$

Pick d = log nU. Then for a certain k in [U], the estimate $\hat{f}(k)$ has the additive error bounded to within εn with probability at least 1-1/(nU). Formally,

$$0 \leq \hat{f}(k) - f(k) \leq \varepsilon n$$

By the union bound, $\hat{f}(k)$ for all k in [U] have the additive error bounded to within εn with probability at least 1-1/n.

Result

By the Count-Min Sketch, one can over-estimate each f(k) to within the additive error εn with probability at least $1-1/n^{\Omega(1)}$ using O((1/ε) log nU) space and O(n log nU) time.

By CM sketch, can we output a set S so that all $k \in S$ have $f(k) \geq (n)^{1/2}$ with high probability?

Result

By the Count-Min Sketch, one can over-estimate each f(k) to within the additive error εn with probability at least $1-1/n^{\Omega(1)}$ using O((1/ε) log nU) space and O(n log nU) time.

By CM Sketch, can we output a set S so that every $k \in U$ that has $f(k) \geq (n)^{1/2}$ is contained in S with high probability?