Lecture 12 CountMin Sketch, CountSketch

Now we describe CountMin Sketch for solving (k, ℓ_1)-point query problem, and CountSketch for solving (k, ℓ_2)-point query problem (i.e. upon a query i, it returns $\tilde{x}_i$ such that $\tilde{x}_i = x_i \pm \|x\|_2/\sqrt{k}$)

I CountMin Sketch

Let w, d be some parameters.

① Choose 2-wise independent hash functions $h_1, \dots, h_d : [n] \longrightarrow [w]$.

Initialize $C[\ell, s] = 0$ for each $1 \le \ell \le d$, $1 \le s \le w$.

② for each item $e_t = (i_t, \Delta_t)$ in the stream

for each $\ell = 1, \dots, d$, update

$$C[\ell, h_\ell(i_t)] \leftarrow C[\ell, h_\ell(i_t)] + \Delta_t$$

③ for each $i \in [n]$, set $\tilde{x}_i = \min_{\ell=1,\dots,d} C[\ell, h_\ell(i)]$

④ upon a query(i), output $\tilde{x}_i$.

Lemma 1: Consider strict turnstile model (i.e. $x \geq 0$ at any time). Let $d = \Omega(\log(1/\delta))$ and $w > 2k$. Then for any fixed $i \in [n]$, $x_i \le \tilde{x}_i$, and

$$\Pr[\tilde{x}_i \geq x_i + \|x\|_1/k] \leq \delta.$$

Proof: Fix ℓ and $i \in [n]$; $h_\ell(i)$ is the bucket that h_ℓ hashes i to.

$Z_\ell = C[\ell, h_\ell(i)]$ is the counter value that i is hashed to.

Note $$E[Z_\ell] = x_i + \sum_{i' \neq i} \Pr[h_\ell(i') = h_\ell(i)]\, x_{i'}$$

$$= x_i + \sum_{i' \neq i} x_{i'}/w$$ → as h_ℓ is choose from a 2-wise independent hash family

$$< x_i + \frac{1}{2k}\|x\|_1$$

Since we are considering strict turnstile, $Z_\ell - x_i$ is nonnegative. By Markov's inequality,

$$\Pr[Z_\ell - x_i > \frac{\|x\|_1}{k}] < \frac{1}{2}$$

Since the d hash functions are independent,

$$\Pr[\min_\ell Z_\ell \geq x_i + \varepsilon\|x\|_1] < \left(\frac{1}{2}\right)^d < \delta. \quad \square$$

Note space complexity: $d \cdot w$ counters

If we set $d = \Omega(\ln n)$ and $w = 3k$, then with probability $1 - \frac{1}{n}$ for all $i \in [n]$, $\tilde{x}_i \leq x_i + \frac{\|x\|_1}{k}$.

The corresponding space is $O(k \ln n)$ counters
(If the ~~total~~ counter is at most m, then $O(k \ln n \cdot \ln m)$ bits of space)

II Count Sketch

Let w, d be some parameters.

① Choose 2-wise independent hash functions $h_1, \cdots h_d : [n] \longrightarrow [w]$;
Choose 2-wise independent hash functions $g_1, \cdots g_d : [n] \rightarrow \{-1, 1\}$;

Initialize $C[\ell, s] = 0$ for each $1 \le \ell \le d$, $1 \le s \le w$.

② for each item $e_t = (i_t, \Delta_t)$ in the stream
for each $\ell = 1, \cdots, d$, update

$$C[\ell, h_\ell(i_t)] \leftarrow C[\ell, h_\ell(i_t)] + g(i_t) \cdot \Delta_t$$

③ for each $i \in [n]$,
set $\tilde{x}_i = \text{median}\{g_1(i) C[1, h_1(i)], \cdots, g(i) \cdot C[\ell, h(i)]\}$

④ upon a query(i), output $\tilde{x}_i$

Intuition:

① Each hash function h_ℓ spreads the elements across w buckets

② The hash function g_ℓ induces cancellations

③ Answer may be negative even if $x \geq 0$; we take the median

> Lemma 2. Let $d \geq 4 \log \frac{1}{\delta}$ and $w > 3k^2$. Then for any fixed $i \in [n]$, $E[\tilde{x}_i] = x_i$ and
> $$\Pr\left[|\tilde{x}_i - x_i| \geq \frac{\|x\|_2}{k}\right] \leq \delta$$

Comparison to Count Min

1. Error guarantee is with respect to $\|x\|_2$ instead of $\|x\|_1$. For $x \geq 0$, $\|x\|_2 \leq \|x\|_1$ and in some cases $\|x\|_2 \ll \|x\|_1$.
2. Space increases to $O(k^2 \log n)$ counters from $O(k \log n)$ counters. (corresponding to error prob $\leq \frac{1}{n}$)

Proof of Lemma 2:

Fix an $i \in [n]$ and $\ell \in [d]$. Let $Z_\ell = g_\ell(i)\, C[\ell, h_\ell(i)]$.

For $i' \in [n]$ let $Y_{i'}$ be the indicator random variable that is 1 if $h_\ell(i) = h_\ell(i')$: that is i and i' collide in h_ℓ.

Note $E[Y_{i'}] = E[Y_{i'}^2] = 1/w$, as h_ℓ 2-wise independent

Note $Z_\ell = g_\ell(i) C[\ell, h_\ell(i)]$

$$= g_\ell(i) g_\ell(i) x_i + \sum_{i' \neq i} g_\ell(i) \cdot g_\ell(i') x_{i'} Y_{i'}$$

Thus, $E[Z_\ell] = x_i + \sum_{i' \neq i} E[g_\ell(i) g_\ell(i') Y_{i'}] x_{i'}$

Note $Y_{i'}$ is independent of $g_\ell(i)$ and $g_\ell(i')$,
so $E[g_\ell(i) g_\ell(i') \cdot Y_{i'}] = E[g_\ell(i) g_\ell(i')] \cdot E[Y_{i'}]$

Now since g_ℓ is 2-wise independent, so
$\Pr[g_\ell(i) = g_\ell(i')] = \frac{1}{2}$, and $\Pr[g_\ell(i) = -g_\ell(i')] = \frac{1}{2}$

Thus $E[g_\ell(i) g_\ell(i')] = 1 \times \frac{1}{2} + (-1) \times \frac{1}{2} = 0$

Thus $E[Z_\ell] = x_i + 0 = x_i$

Now we analyze $Var[Z_\ell]$.

$$\mathrm{Var}[Z_\ell] = E[(Z_\ell - x_i)^2]$$

$$= E\left[\left(\sum_{i' \neq i} g_\ell(i) \cdot g_\ell(i') Y_{i'} \cdot x_{i'}\right)^2\right]$$

$$= E\left[\sum_{i' \neq i} Y_{i'}^2 x_{i'}^2 + \sum_{\substack{i' \neq i \\ i'' \neq i \\ i' \neq i''}} x_{i'} x_{i''} g_\ell(i') g_\ell(i'') Y_{i'} Y_{i''}\right]$$

$$= \sum_{i' \neq i} x_{i'}^2 \cdot E[Y_{i'}^2] + 0 \leftarrow$$ the same reason as before

$$< \|x\|_2^2 / w$$

Therefore $E[Z_\ell] = x_i$ and $\mathrm{Var}[Z_\ell] \leq \|x\|_2^2 / w$.

Using Chebyshev

$$\Pr\left[|Z_\ell - x_i| \geq \frac{\|x\|_2}{k}\right] \leq \frac{\mathrm{Var}[Z_\ell]}{\frac{\|x\|_2^2}{k^2}} \leq \frac{k^2}{w} < \frac{1}{3}$$

Via the Chernoff bound:

$$\Pr\left[\left|\mathrm{median}\{Z_1, \ldots, Z_d\} - x_i\right| \geq \frac{\|x\|_2}{k}\right] \leq e^{-c \cdot d}$$

$$\leq \delta. \quad \square$$

Thus with $d = \Theta(\ln n)$, $w = 4k^2$, with prob. $1 - \frac{1}{n}$, for all $i \in [n]$:

$$|\hat{x}_i - x_i| \leq \frac{\|x\|_2}{k}$$

Total space $O(k^2 \cdot \log n)$ counters. and $O(k^2 \ln n \cdot \ln m)$ bits (if total counter is $\leq m$)