4.5 Streams and Sketches

Thursday, January 30, 2020 12:01 AM

Consider a sequence $a_1, \dots, a_n \in [m]$, with n and m very large.
We want to compute $|\text{uniq}\{a_1, \dots, a_n\}|$.

$O(m)$ solution – bit vector
$O(n \log m)$ solution – store list of items seen
(don't need to organize if we only care about time complexity)

Goal: $O(\log n \cdot \log m)$.

Thm: (lower bound) Any exact deterministic algorithm must use at least m bits of memory on some sequence of length $m+1$.

proof. Suppose we have seen $a_1, \dots, a_m$ already.
Now assume for contradiction that our algorithm uses $< m$ bits of memory on all such sequences.
$|\mathcal{P}([m])| = 2^m$, and $\text{uniq}(\{a_1, \dots, a_m\})$ can be any subset except $\emptyset$,
so $2^m - 1$ possibilities for $\text{uniq}(\{a_1, \dots, a_m\})$.
But we only have 2^{m-1} possible memory states.
So two different subsets $S_1, S_2 \in P([m]) \setminus \emptyset$ must have the same memory state.
$|S_1| = |S_2|$ because otherwise our algorithm must be wrong for one of them.
Now add an $a_{m+1} \in S_1$, $\notin S_2$. Then the new sets would have the same state but different sizes. $\times$

Consider an idealized streaming algorithm (ISA)

1. Pick random hash function $h: [m] \to [0, 1]$.
2. Calculate $z = \min_{i \in \text{stream}} h(i)$.
3. Output $\frac{1}{z} - 1$

Let $\mathcal{S} = \text{unique}\{a_1, \dots, a_n\} = \{j_1, \dots, j_t\}$

$h(j_1), \dots, h(j_t) = X_1, \dots, X_n$ be ind. $\text{Unif}[0, 1]$

$z = \min\{X_i\}_{i=1}^{n}$

Lemma: If $X:\Omega\to[0,+\infty]$ is a nonnegative r.v., then

$$\mathbb{E}X=\int_{[0,+\infty)}\mathbb{P}(X>x)\,dx$$

proof. For every $\omega\in\Omega$, $X(\omega)=\int_{(0,X(\omega))}dx=\int_{[0,+\infty)}\mathbb{1}_{(0,X(\omega))}(x)\,dx$

$$\mathbb{E}X=\int_\Omega X\,d\mathbb{P}=\int_\Omega\int_{[0,+\infty)}\underbrace{\mathbb{1}_{(0,X(\omega))}(x)}_{\text{nonnegative, Tonelli's thm applies for changing integral}}dx\,d\mathbb{P}(\omega)$$

$$=\int_{[0,+\infty)}\int_\Omega\underbrace{\mathbb{1}_{(0,X(\omega))}}_{\text{If } x<X(\omega),\ =1.}(x)\,d\mathbb{P}(\omega)\,dx=\int_{[0,+\infty)}\mathbb{P}(X>x)\,dx$$

Integral of $1\cdot P(\omega)$ for all ω s.t. $x<X(\omega)$.

$\blacksquare$

Claim: $\mathbb{E}Z=\frac{1}{t+1}$.

proof.
$$\mathbb{E}Z=\int_0^\infty\mathbb{P}(Z>\lambda)\,d\lambda=\int_0^1\mathbb{P}(\forall i,\ X_i>\lambda)\,d\lambda$$
$$=\int_0^1\prod_{i=1}^t\mathbb{P}(X_i>\lambda)\,d\lambda=\int_0^1(1-\lambda)^t\,d\lambda=\left[-\frac{(1-\lambda)^{t+1}}{t+1}\right]_0^1=\frac{1}{t+1}.$$

$\blacksquare$

Claim: $\mathbb{E}Z^2=\frac{2}{(t+1)(t+2)}$

$$\boxed{\int u\,dv=uv-\int v\,du}$$

proof.
$$\mathbb{E}Z^2=\int_0^1\mathbb{P}(Z^2>\lambda)\,d\lambda=\int_0^1\mathbb{P}(Z>\sqrt{\lambda})\,d\lambda$$
$$=\int_0^1(1-\sqrt{\lambda})^t\,d\lambda=2\int_1^0u^t(u-1)\,du=2\int_0^1u^t(1-u)\,du$$

Let $u=1-\sqrt{\lambda}$, $\sqrt{\lambda}=1-u$

$du=-\frac{1}{2\sqrt{\lambda}}d\lambda\Rightarrow 2\,du=\frac{1}{u-1}d\lambda$

$d\lambda=2(u-1)du$

$\bar{u}=(1-u)$ $\quad dv=u^t\,dv$

$d\bar{u}=-du$ $\quad v=\frac{1}{t+1}u^{t+1}$

$$=2\left[(1-u)\cdot\frac{1}{t+1}u^{t+1}\Big|_0^1+\int_0^1\frac{1}{t+1}u^{t+1}\,du\right]$$
$$=\frac{2}{(t+1)(t+2)}$$

$\blacksquare$

Thus, $\mathrm{Var}[Z]=\mathbb{E}Z^2-(\mathbb{E}Z)^2=\frac{t}{(t+1)^2(t+2)}<\frac{1}{(t+1)^2}$

Thus, $\mathrm{Var}[Z] = \mathbb{E} Z^2 - (\mathbb{E} Z)^2 = \dfrac{t}{(t+1)^2(t+2)} < \dfrac{1}{(t+1)^2}$

Averaging algorithm

1. Run $q = \dfrac{1}{\varepsilon^2 \eta}$ ISAs in parallel
2. $\bar{Z} = \dfrac{1}{q} \sum_{i=1}^{q} Z_i$
3. output $\dfrac{1}{\bar{Z}} - 1$

Then $\mathbb{E}(\bar{Z}) = \dfrac{1}{t+1}$, $\quad \mathrm{Var}(\bar{Z}) = \dfrac{1}{q} \cdot \dfrac{t}{(t+1)^2(t+2)} < \dfrac{1}{q(t+1)^2}$.

By Chebyshev $\mathbb{P}\left(\left| \bar{Z} - \dfrac{1}{t+1} \right| > \dfrac{\varepsilon}{t+1} \right) < \dfrac{(t+1)^2}{\varepsilon^2} \cdot \dfrac{1}{q(t+1)^2} = \eta$.

Chebyshev

$\mathbb{P}(|X - \mathbb{E}X| \geq a) \leq \dfrac{\mathrm{Var}(X)}{a^2}$

Claim: $\mathbb{P}\left(\left| \left(\tfrac{1}{\bar{Z}} - 1\right) - t \right| > O(\varepsilon)\, t \right) < O(\eta)$

$$\mathbb{P}\left(\left| \bar{Z} - \frac{1}{t+1} \right| > \frac{\varepsilon}{t+1} \right) < \eta$$

$$\mathbb{P}\left(\left| \bar{Z}t + \bar{Z} - 1 \right| > \varepsilon \right) < \eta$$

$$\mathbb{P}\left(\left| \frac{1}{\bar{Z}} - t - 1 \right| > \frac{\varepsilon}{|\bar{Z}|} \right) < \eta$$

w.p. $1-\eta$ $\quad |\bar{Z}| \leq \dfrac{1+\varepsilon}{t+1}$,

$$\Rightarrow \mathbb{P}\left(\left| \frac{1}{\bar{Z}} - t - 1 \right| > \underbrace{\frac{\varepsilon(t+1)}{1+\varepsilon}}_{O(\varepsilon)t} \right) < \underbrace{\eta}_{2\eta}.$$

$\varepsilon(t+1)\left(1 - \varepsilon + \frac{\varepsilon^2}{2} - O(\varepsilon^3)\right)$

$= \varepsilon + t\varepsilon + O(\varepsilon^2)$

$= O(\varepsilon)\, t.$

So, with high prob., our estimator is within a factor $1 + O(\varepsilon)$ of $1+t$.

But we have forgotten to take into account the $\geq n$ bits needed to store our hash function.

Let $\mathcal{H}$ be a set of functions that map $[a] \to [b]$.

Define: $\mathcal{H}$ is a k-wise independent hash family if

$\forall i_1 \neq i_2 \neq \cdots \neq i_k \in [a]$

and $\forall j_1, \ldots, j_k \in [b]$,

$$\mathbb{P}_{h \sim \mathcal{H}}\left(h(i_1) = j_1 \wedge \cdots \wedge h(i_k) = j_k\right) = \frac{1}{b^k}.$$

Ex. The set $\mathcal{H}$ of all functions $[a] \to [b]$ is k-wise ind for every k.

$|\mathcal{H}| = b^a$, so $h \in \mathcal{H}$ is representable in $a \lg b$ bits.

Ex. Let $a = b = q =$ prime power. $\mathcal{H}$ will be the set of all deg $k-1$ polynomials in $\mathbb{F}_q[x]$.

Claim: $\mathcal{H}_{poly\text{-}k}$ is a k-wise family.

proof. Lagrange interpolation. If we know $i_1, \ldots, i_k$ and $j_1, \ldots, j_k$ and that no i's repeat, then

$$p(x) = \sum_{r=1}^{k} \left(\frac{\prod_{y \in [k] \setminus \{r\}} x - i_y}{\prod_{y \in [k] \setminus \{r\}} i_r - i_y} \right) \cdot j_r$$

satisfies $\forall r \;\; p(i_r) = j_r$ and this polynomial is unique because $\mathbb{F}_q$ is a field.

Thus, $|\mathcal{H}_{poly\text{-}k}| = q^k$.

But note that there is precisely one poly of deg $\leq k-1$ s.t. it goes through all (i_r, j_r), so $\mathbb{P}_{h \sim \mathcal{H}_{poly\text{-}k}}\left(h(i_1) = j_1 \wedge \cdots \wedge h(i_k) = j_k\right) = \frac{1}{q^k}$. $\square$

Also, each $h \in \mathcal{H}_{poly\text{-}k}$ is representable using $k \log q$ bits.

How much independence do we need for unique items?

How much independence do we need for unique items?

We can replace independent hash functions with pairwise independence

Careful analysis in book, but basically pairwise independence allows us to sum variances.