Eksamensdisposition - Second Moment

Søren Mulvad, rbn601

17. juni 2019

- **Problem**
- **Kvalitet af Count Sketch estimatet**
  - Forventet værdi
  - Varians
  - Afvigelse fra korrekt resultat
- **Median trick**

Eksamensdisposition - Second Moment

Problem

Vi har en strøm af par $(j_0, \Delta_0), \dots, (j_{s-1}, \Delta_{s-1}) \in [n] \times \mathbb{Z}$ (sættet af alle heltal).
Vi definerer frekvensen f_j for hver eneste $j \in [n]$ som

$$f_j = \sum_{i \in [s], j_i = j} \Delta_i$$

Derudover definerer vi det m'te moment $F_m = \sum_{j \in [n]} f_j^m = ||f||_m^m$. Vi ønsker nu at estimere det 2. moment $F_2 = \sum_{i \in [n]} f_i^2$ ved kun at bruge k tællere.

Count Sketch Algoritme

Algorithm 1: Count Sketch

```
  Init
1 k = ⌈8/ε²⌉
2 C[0,...,k−1] = 0
3 h = 4-universel h : [n] → [k]
4 s = 4-universel s : [n] → {−1,+1}
  Process (j, Δ)
5 C[h(j)] += s(j) · Δ
  Output
6 return X = Σ_{j∈[k]} C[j]²
```

Hvor Process (j, Δ) svarer til vi løbende kører alle par i strømmen igennem i linje 4, hvor vi potentielt kunne stoppe på et vilkårligt tidspunkt.

Kvalitet af Count Sketch estimatet

Forventet værdi

Vi har $C[b] = \sum_{j \in [n]} s(j) f_j [h(j) = b]$, så:

$$\begin{aligned}
X &= \sum_{b \in [k]} \left(\sum_{j \in [n]} s(j) f_j [h_j = b] \right)^2 \\
&= \sum_{b \in [k]} \sum_{i,j \in [n]} s(i)s(j) f_i f_j [h(i) = b = h(j)] \\
&= \sum_{i,j \in [n]} s(i)s(j) f_i f_j [h(i) = h(j)] \\
&= \sum_{i \in [n]} f_i^2 + \sum_{\substack{i,j \in [n] \\ i \neq j}} s(i)s(j) f_i f_j [h(i) = h(j)] \\
&= F_2 + Y
\end{aligned}$$

Hvis vi nu kan vise $\mathbb{E}[Y] = 0$ har vi da vist $\mathbb{E}[X] = F_2$. Pga. vi sagde vores hashing-funktion er 4-universel er $s(i), s(j), h(i)$ og $h(j)$ uafhængige. Derudover har vi $\mathbb{E}[s(i)] = 0$, så alle led i summen bliver 0 hvorved $\mathbb{E}[Y] = 0$.

Varians

Vi ønsker at bestemme variansen af X nu:

$$
\begin{aligned}
\mathrm{Var}\left[X\right] = \mathrm{Var}\left[Y\right] &= \mathbb{E}\left[Y^2\right] - \underbrace{\mathbb{E}\left[Y\right]^2}_{0} \\
&= \mathbb{E}\left[\left(\sum_{\substack{i,j\in[n]\\ i\neq j}} s(i)s(j)f_i f_j[h(i)=h(j)]\right)^2\right] \\
&= \sum_{\substack{i,j,i',j'\in[n]\\ i\neq j,\ i'\neq j'}} \mathbb{E}\left[\Big(s(i)s(j)f_i f_j[h(i)=h(j)]\Big)\Big(s(i')s(j')f_{i'}f_{j'}[h(i')=h(j')]\Big)\right] \qquad (1)
\end{aligned}
$$

Nu tager vi udgangspunkt i ét af leddene i summen i (1). Hvis vi har den situation at en af nøglerne er unik, f.eks. $i \notin \{j, i', j'\}$, så vil i pr. 4-universaliteten være uafhængig af j, i', j' samt hashfunktionen h. hvorved vi ville kunne sætte $\mathbb{E}\left[s(i)\right]$ uden for en parentes i (1) op. Da $\mathbb{E}\left[s(i)\right] = 0$ vil vi derfor få at et led i summen i (1) med en unik nøgle vil give 0.

Vi kan derfor begrænse vores opmærksomhed til kun de led der ikke har unikke nøgler. Siden $i \neq j$ og $i' \neq j'$ må vi enten have $(i,j) = (i',j')$ eller $(i,j) = (j',i')$, altså 2 cases for hvert i og j. Derfor:

$$
\begin{aligned}
\text{Eq. (1)} &= 2\sum_{\substack{i,j\in[n]\\ i\neq j}} \mathbb{E}\left[\Big(s(i)s(j)f_i f_j[h(i)=h(j)]\Big)^2\right] \qquad (2)\\
&= 2\sum_{\substack{i,j\in[n]\\ i\neq j}} \mathbb{E}\left[f_i^2 f_j^2[h(i)=h(j)]\right] \qquad (3)\\
&= 2\sum_{\substack{i,j\in[n]\\ i\neq j}} (f_i^2 f_j^2)\mathbb{P}\left[h(i)=h(j)\right] \qquad (4)\\
&= 2\sum_{\substack{i,j\in[n]\\ i\neq j}} \frac{f_i^2 f_j^2}{k} \\
&= \frac{2}{k}\sum_{\substack{i,j\in[n]\\ i\neq j}} f_i^2 f_j^2 \\
&< \frac{2}{k}\Big(\sum_{i\in[n]} f_i^2\Big)^2 \qquad (5)\\
&= 2F_2^2/k
\end{aligned}
$$

I (2) bruger vi, at der netop var 2 cases hvor nøglerne var lig hinanden, og så har vi bare sat $i' = i$ og $j' = j$ i summen.
I (3) bruger vi, at $s(x)^2 = 1\cdot 1$ eller $(-1)(-1) = 1$ og indikatorvariablen er 0 eller 1, så opløftet i anden er den lig sig selv.
I (4) benytter vi linearity of expectation og igen at $[h(i) = h(j)]$ er en indikatorvariabel.
I (5) må vores udtryk være mindre da vi ser bort fra alle de led hvor $i \neq j$.

Afvigelse ved brug af Chebyshev

Da vi nu har vist

$$
\mathbb{E}\left[X\right] = F_2 \qquad\qquad \mathrm{Var}\left[X\right] < \frac{2F_2^2}{k}
$$

medfører det jf. Chebyshev's ulighed og vores valg af k at:

$$
\mathbb{P}\left[|X - F_2| \geq \epsilon F_2\right] \leq \frac{\mathrm{Var}\left[X\right]}{(\epsilon F_2)^2} \leq \frac{2}{(k\epsilon^2)} = \frac{1}{4}
$$

Median trick

Vi laver t uafhængige estimater $X_0, \ldots, X_{t-1}$ i parallel (ved at bruge forskellige hashfunktioner) og returnerer medianen $X_{(\lceil t/2 \rceil)}$ af de t svar. Vi siger X_i fejler hvis $|X_i - \mathbb{E}[X]| \geq \epsilon F_2$.

Lad $B_i = [X_i \text{ fejler}]$ og lad B være antallet der fejler:

$$B = \sum_{i \in [t]} B_i$$

Da har vi, at hvis $X_{(\lceil t/2 \rceil)}$ fejler, så betyder det at $B \geq t/2$.
Da $\mathbb{P}[B_i = 1] \leq \frac{1}{4}$ må den forventede værdi af B være

$$\mathbb{E}[B] = \mu = \sum_{i \in [t]} \mathbb{E}[B_i] \leq t/4$$

Da kan vi beregne:

$$\mathbb{P}[\text{Median fejler}] = \mathbb{P}[B \geq 2\mu] = \mathbb{P}[B \geq (1+\delta)\mu] \leq e^{-\delta^2 \mu/3} \leq e^{-t/12}$$

Hvor vi bruger Chernoff Bounds til at begrænse sandsynligheden, og herudover i eksponentialet benytter $\delta = 1$.

Herved har vi altså væsentligt begrænset sandsynligheden for at få noget rimelig forkert.