CS5330: Randomized Algorithms Spring 2020

Week 3 – Additional Notes

Prof. Arnab Bhattacharyya

1 Variance analysis for Count Sketch

The simplified CountSketch algorithm given in the slides defines:

$$\hat{f}_i = \sigma_i \cdot C[h(i)]$$

As in the slides, if we let $X_{i,a}$ indicate the collision event $h(i) = h(a)$, then $\hat{f}_i = f_i + \sum_{a \neq i} \sigma_i \sigma_a X_{i,a} \cdot f_a$. Using the pairwise independence of the signs, we saw that $\mathbb{E}[\hat{f}_i] = f_i$. We now complete the variance analysis that is omitted in the slides:

$$\begin{aligned}
\mathsf{Var}[\hat{f}_i] &= \mathbb{E}\left[\left(\sum_{a \neq i} \sigma_i \sigma_a X_{i,a} \cdot f_a\right)^2\right] \\
&= \mathbb{E}\left[\sum_{a \neq i} \sigma_i^2 \sigma_a^2 X_{i,a}^2 \cdot f_a^2 + \sum_{\substack{a \neq b \\ a \neq i \\ b \neq i}} \sigma_i^2 \sigma_a \sigma_b X_{i,a} X_{i,b} \cdot f_a f_b\right] \\
&= \mathbb{E}\left[\sum_{a \neq i} X_{i,a} \cdot f_a^2 + \sum_{\substack{a \neq b \\ a \neq i \\ b \neq i}} \sigma_a \sigma_b X_{i,a} X_{i,b} \cdot f_a f_b\right] \\
&= \sum_{a \neq i} \mathbb{E}[X_{i,a}] \cdot f_a^2 + \sum_{\substack{a \neq b \\ a \neq i \\ b \neq i}} \mathbb{E}[\sigma_a \sigma_b] \cdot \mathbb{E}[X_{i,a} X_{i,b}] \cdot f_a f_b
\end{aligned}$$

The last line follows from linearity of expectation as well as from the independence between the choice of the signs and the hash function. Note that $\mathbb{E}[\sigma_a \sigma_b] = 0$ for $a \neq b$, and $\mathbb{E}[X_{i,a}] = \Pr(h(i) = h(a)) \leq 1/w$. Plugging into the above, $\mathsf{Var}[\hat{f}_i] \leq \frac{1}{w} \sum_a f_a^2$.

Using Chebyshev,

$$\Pr\left[|\hat{f}_i - f_i| > \epsilon \sqrt{\sum_a f_a^2}\right] \leq \frac{\mathsf{Var}[\hat{f}_i]}{\epsilon^2 \sum_a f_a^2} \leq \frac{1}{w\epsilon^2} \leq 1/3$$

where the last inequality uses our choice of $w > 3/\epsilon^2$.

2 References

The very readable paper that introduced CountMin Sketch is [CM05]. Count Sketch was given by Charikar, Chen and Farach-Colton in [CCFC04].

The LazyMedian algorithm is due to Motwani and Raghavan (apparently as part of the work on their book, though I am not really sure). Our discussion is based on Section 3.5 of the Mitzenmacher and Upfal book; please go through it for the full analysis. We discussed in class why LazyMedian only requires $3n/2 + o(n)$ comparisons (though this is not mentioned in the book). The reason is that in Step 5 of Algorithm 3.1 in the book, we can compare each element in S first to d and then only if it's more than d, to u. If $\mathcal{E}_1, \mathcal{E}_2, \mathcal{E}_3$ do not hold, then the number of elements in S less than d is at least $\frac{n}{2} - 4n^{3/4}$. For these elements, we don't need to compare them to u. Hence, we can perform $2n - (\frac{n}{2} - 4n^{3/4}) = 3n/2 + o(n)$ comparisons to complete Step 5.

For deterministic median selection, both the best upper bound and the best lower bound for the number of comparsions are due to Dor and Zwick. They showed in [DZ01] that $(2 + 2^{-50})n$ comparisons are required, while in [DZ99], they designed an algorithm making $2.95n$ comparisons.

References

[CCFC04] Moses Charikar, Kevin Chen, and Martin Farach-Colton. Finding frequent items in data streams. *Theoretical Computer Science*, 312(1):3–15, 2004.

[CM05] Graham Cormode and Shan Muthukrishnan. An improved data stream summary: the count-min sketch and its applications. *Journal of Algorithms*, 55(1):58–75, 2005.

[DZ99] Dorit Dor and Uri Zwick. Selecting the median. *SIAM Journal on Computing*, 28(5):1722–1758, 1999.

[DZ01] Dorit Dor and Uri Zwick. Median selection requires $(2+\epsilon)n$ comparisons. *SIAM Journal on Discrete Mathematics*, 14(3):312, 2001.