Improving *Tug-of-War* sketch using Control-Variates method

Bhisham Dev Verma
jointly with Rameshwar Pratap (IIT Mandi) and Raghav Kulkarni (CMI)
IIT Mandi

Streaming Datasets

- Many data sources that generates large volume of data are best modeled as data stream

  e.g. : streams of network packets, click stream data, traffic data etc.

- Impractical to store and process the entire data
- By taking one pass over data, quickly build a small summary (a.k.a. sketch)
- Perform computation on sketch to get approximate answer

k-*th* moment and Inner product

- Universe = { a, b, c, ……, z} (size of universe is n)
- $\sigma_1 =$ a, b, a, d, c, b, b, d, e, … and $\boldsymbol{f} = (f_1, f_2, \ldots, f_n)$ is corresponding frequency vector.
- $\sigma_2 =$ a, b, a, d, c, e, c, d, e, b … and $\boldsymbol{g} = (g_1, g_2, \ldots, g_n)$ is corresponding frequency vector.

- $\boldsymbol{k}$-th moment of σ_1 and σ_2 is

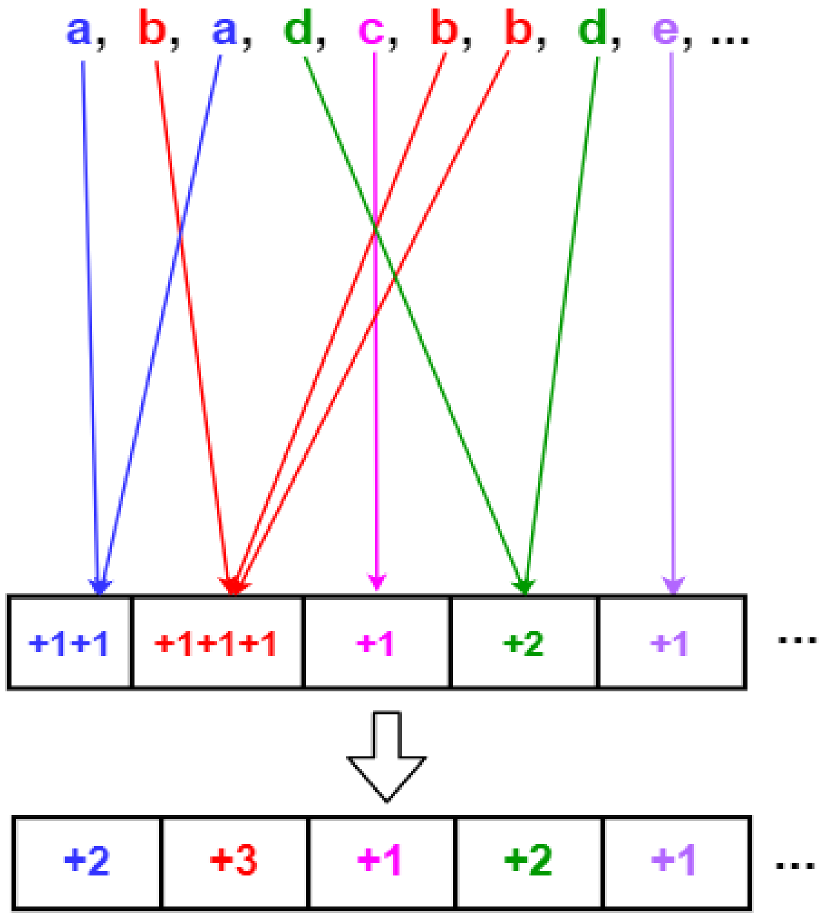
$$\boldsymbol{F_k} = \sum_{i\epsilon[n]} f_i^k \quad \text{and} \quad \boldsymbol{G_k} = \sum_{i\, \epsilon\, n} g_i^k \qquad (1)$$

- Inner product of $\boldsymbol{f}$ and $\boldsymbol{g}$ is

$$\langle \boldsymbol{f}, \boldsymbol{g} \rangle = \sum_{i\epsilon[n]} f_i \cdot g_i \qquad (2)$$

- **Our focus is to find**
  - $\boldsymbol{F_2}$ **moment of the stream**
  - **Inner product**

Naive Method to Compute F_2 moment

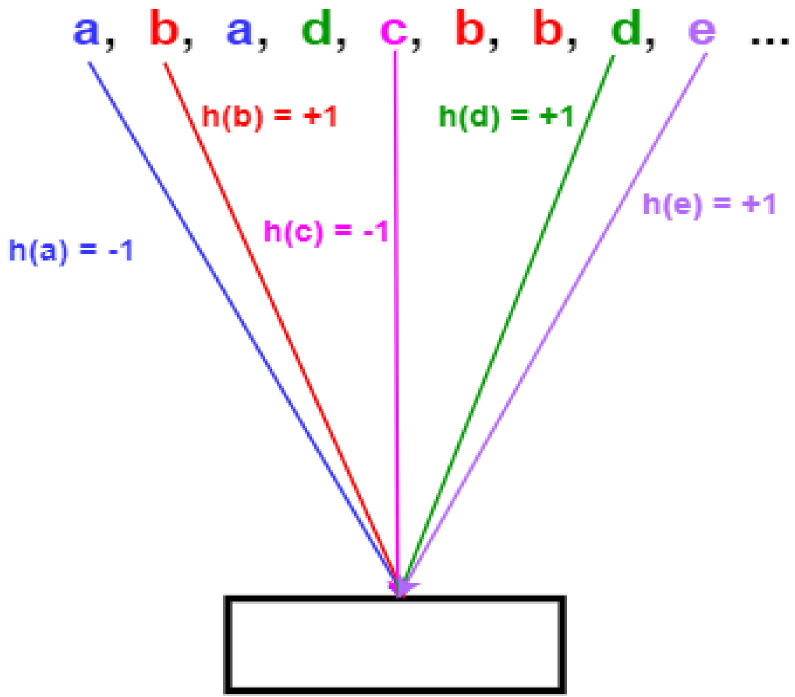


$$F_2 = 2^2 + 3^2 + 1^2 + 2^2 + 1^2 + \cdots$$

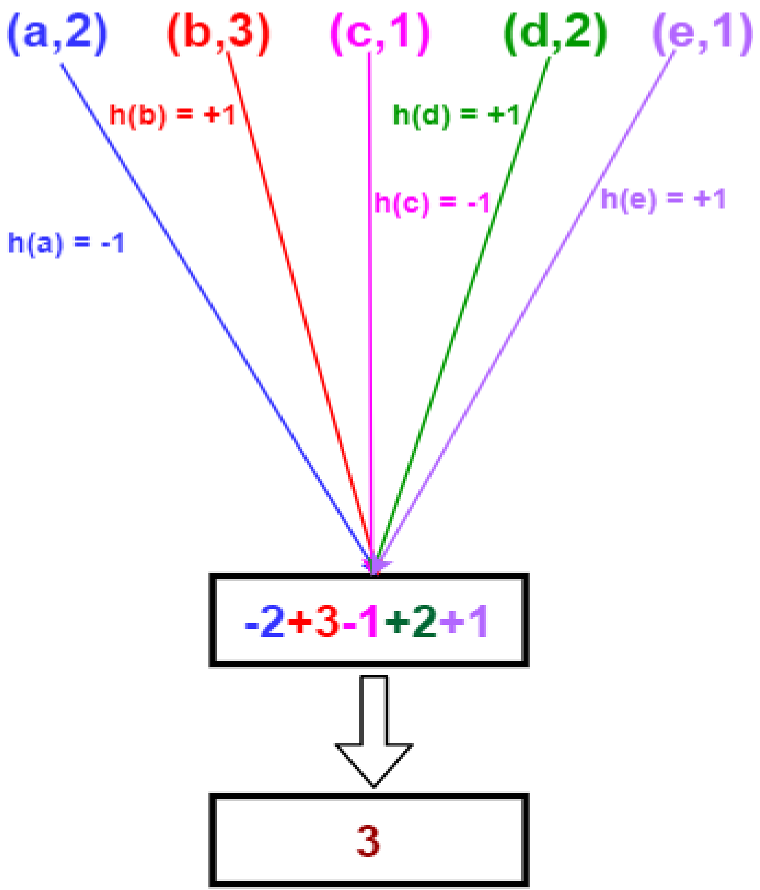
- Data stream of alphabets of length m.
- Universe: = $[n]$ = { a, b, ..., z}
- f_i is frequency of i^{th} item, $i \in [n]$.
- $\boldsymbol{f} = (f_1, f_2, \dots, f_n)$ is a frequency vector.
- Space requirement : $O(n\ logm)$.

Impractical when n and m are very large.

F_2 estimation of a data-stream using *Tug-of-War* sketch



h(.) assign a sign {+1, -1}



Space required : $O(\log m + \log n)$

Estimated $F_2 = 3^2 = 9$

Actual $F_2 = 2^2 + 3^2 + 1^2 + 2^2 + 1^2 = 19$

F_2 estimation of a data-stream using *Tug-of-War* sketch

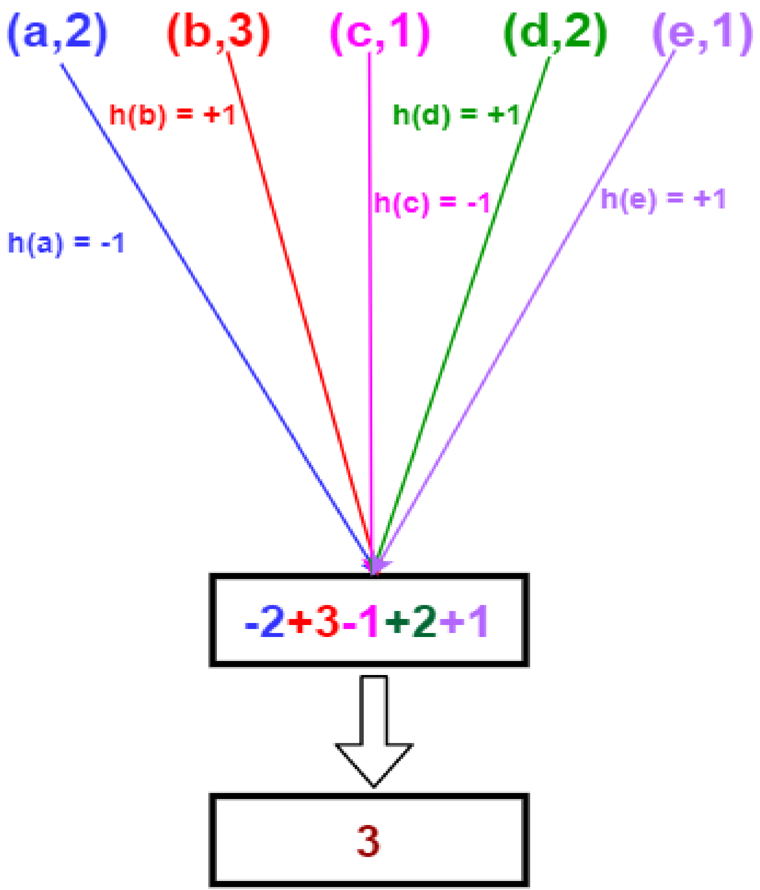
- ❑ $h[n] \rightarrow \{+1, -1\}$
- ❑ f_i frequency of i^{th} item
- ❑ Frequency Vector: $\boldsymbol{f} = (f_1, f_2, \ldots, f_n)$

Estimating F_2:

$$\tilde{X} = \sum_{i \in [n]} f_i h(i)$$

$$X = \tilde{X}^2$$

X is the estimate of F_2



Estimated $F_2 = 3^2 = 9$

Actual $F_2 = 2^2 + 3^2 + 1^2 + 2^2 + 1^2 = 19$

F_2 estimation of a data-stream using *Tug-of-War* sketch

$$\tilde{X} = \sum_{i \in [n]} f_i h(i)$$

$$X = \tilde{X}^2$$

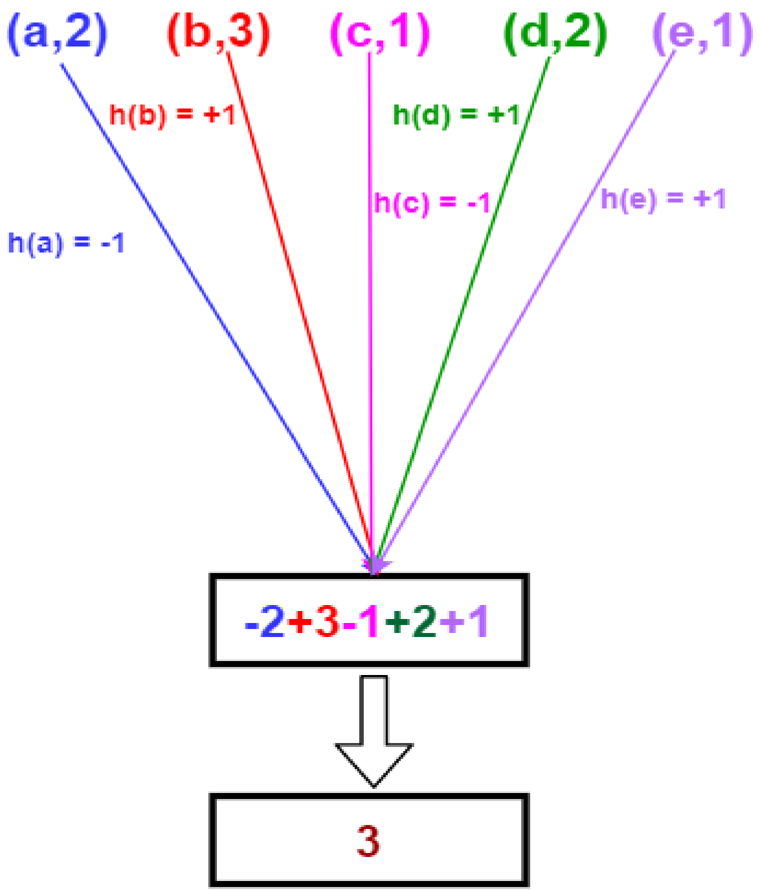
- Statistics of X

$$E[X] = ||\boldsymbol{f}||_2^2 = \boldsymbol{F_2} \qquad (3)$$

and

$$Var[X] = 2(\boldsymbol{F_2^2} - \boldsymbol{F_4}) \qquad (4)$$

Variance is high for large values of f_i



Estimated F_2: $X = 3^2 = 9$

Actual $F_2 = 2^2 + 3^2 + 1^2 + 2^2 + 1^2 = 19$

Variance reduction via Control-Variate (CV)

- ❑ Let X be the r.v. of our estimate
- ❑ Find another r.v. Z s.t. $E[Z]$ is known
- ❑ Our new estimator: $X + c(Z - E[Z])$

$$E[X + c(Z - E[Z])] = E[X]. \tag{5}$$

$$Var[X - c(Z - E[Z])] = Var[X] + c^2 Var[Z] + 2\,Cov[X, Z]. \tag{6}$$

Optimal value of c which minimize equ. (6), say $\hat{c}$ is

$$\hat{c} = -\frac{Cov[X,Z]}{Var[Z]}. \tag{7}$$

Equation (6) and (7), gives

$$Var[X + c(Z - E[Z])] = Var[X] - \frac{Cov[X,Z]^2}{Var[Z]}. \tag{8}$$

Variance reduction via Control-Variate (CV)

Properties of Z:

- ❑ should be easily computable
- ❑ should have low variance
- ❑ should have high covariance with X
- ❑ $E[Z]$ should be known

Improving Tug-of-War using Control-Variate (CV) Method

Tug-of –war estimate: $X = \left(\sum_{i \in [n]} f_i h(i)\right)^2$

We choose CV r.v. $Z = \sum_{i \neq j, i, j \in [n]} h(i) h(j)$

$\Longrightarrow E[Z] = 0$ and $Var[Z] = \boldsymbol{F_0}(\boldsymbol{F_0} - 1)$,

$$Cov[X, Z] = \boldsymbol{F_1^2} - \boldsymbol{F_2}$$

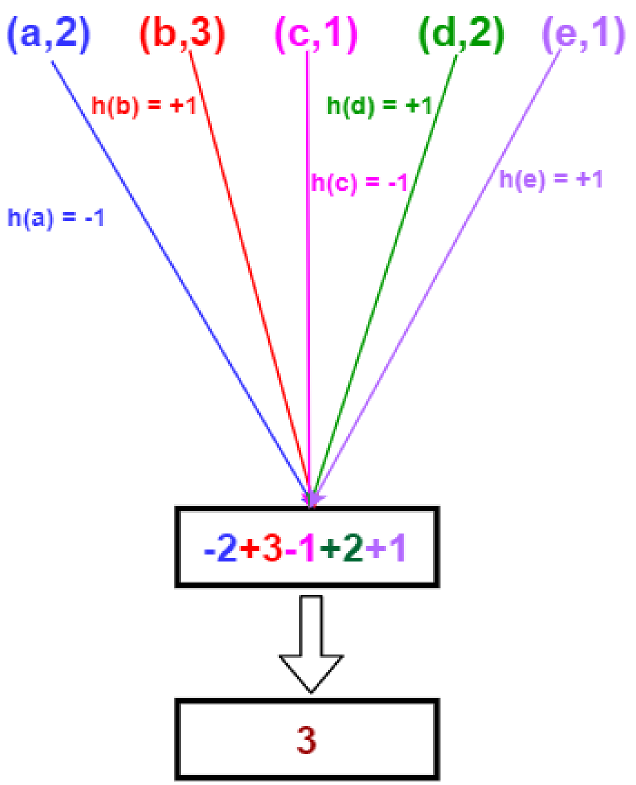
where $\boldsymbol{F_0} := n$ and $\boldsymbol{F_1} := \sum_{i \in [n]} f_i$.

$$\hat{c} = -\frac{Cov[X,Z]}{Var[Z]} = -\frac{\boldsymbol{F_1^2} - \boldsymbol{F_2}}{\boldsymbol{F_0}(\boldsymbol{F_0} - 1)} \qquad (9)$$

$$\text{Variance Reduction} = \frac{Cov[X,Z]^2}{Var[Z]} = \frac{\left(\boldsymbol{F_1^2} - \boldsymbol{F_2}\right)^2}{\boldsymbol{F_0}(\boldsymbol{F_0} - 1)} \qquad (10)$$

Improving Tug-of-War using Control-Variate (CV) Method

- ❑ $X = 9$ (Tug-of-war estimate), $Z = -4,$ and $E[Z] = 0$.
- ❑ Recall $\hat{c} = -\frac{F_1^2 - F_2}{F_0(F_0 - 1)}$.
- ❑ **We compute $\boldsymbol{F_1}$ by maintaining a counter** (in space $\boldsymbol{O(\log m)}$ **).**
- ❑ For $\boldsymbol{F_2}$, we use Tug-of-War estimate as a proxy.



Our CV estimate: $X + \hat{c}(Z - E[Z]) = 9 - \frac{(81 - 9)}{5(5-1)}(-4 - 0)$

$$= 9 - 3.6 \times (-4)$$

$$= \mathbf{23.4}$$

Estimated F_2: X= 3^2 $= 9$

Actual F_2 = $2^2 + 3^2 + 1^2 + 2^2 + 1^2 = 19$

Empirical Evaluation

Datasets

- **Synthetic Datasets**
  - stream of 100000 items
  - frequency of each item is sampled randomly between 1 and 5000.
- **KOS dataset**
  - consist of corpus of document, treat word as an item and number of occurrences in entire corpus as frequency
  - consist of 6906 distinct word and their frequency
- **Transaction datasets**
  - T10I4D100K: consist of 870 distinct items and 1010228 item in total
  - T40I10D100K: consist of 942 distinct items and 3960507 items in total

Empirical Evaluation

Evaluation Metrics:

- ❑ Variance analysis via box-plot
- ❑ Mean absolute error
- ❑ Median of means estimation

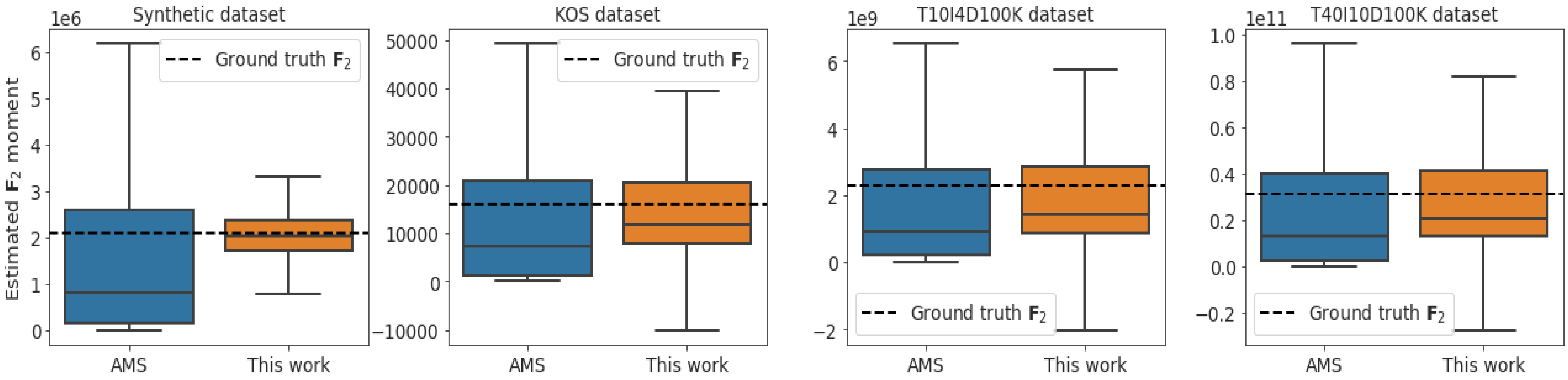


Figure 1

Empirical Evaluation

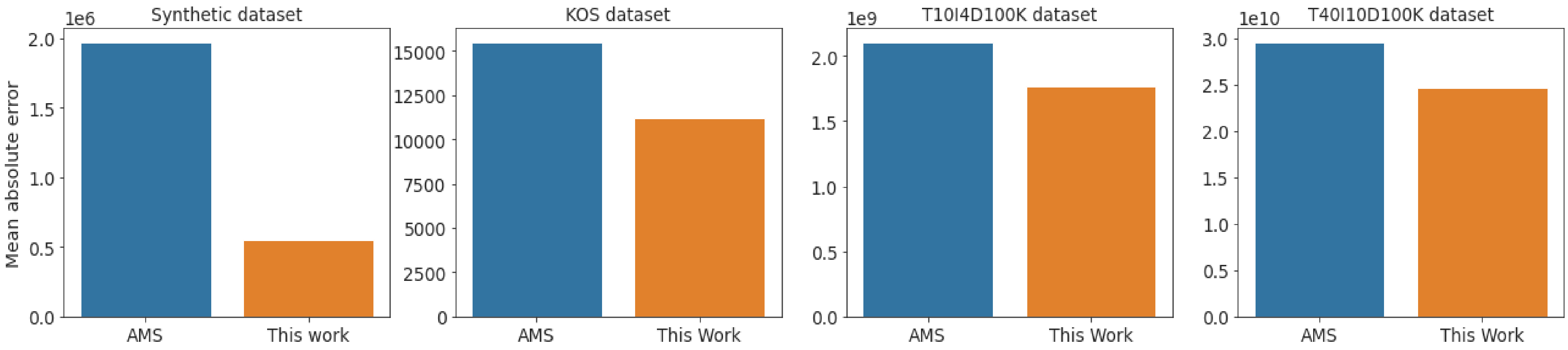


Figure 2

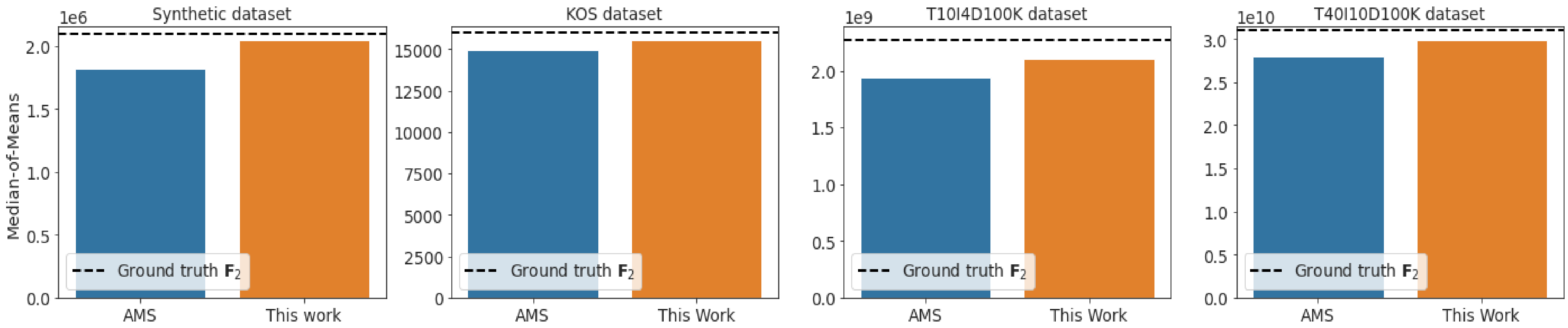


Figure 3

Improving Inner product estimate using CV method

- $\boldsymbol{f} = (f_1, f_2, \dots, f_n)$ is a frequency vector of stream σ_1.
- $\boldsymbol{g} = (g_1, g_2, \dots, g_n)$ is a frequency vector of stream σ_2.
- *Tug-of-War* sketch of streams σ_1 and σ_2 are

$$\tilde{f} = \sum_{i \epsilon [n]} f_i h(i) \quad \text{and} \quad \tilde{g} = \sum_{i \epsilon [n]} g_i h(i)$$

- Inner product estimate of $\boldsymbol{f}$ and $\boldsymbol{g}$ is

$$X^{(2)} = \tilde{f} . \tilde{g}$$

- $E\left[X^{(2)}\right] = \langle \boldsymbol{f}, \boldsymbol{g} \rangle$ (11)
- $Var\left[X^{(2)}\right] = \sum_{i \neq j, i, j \in [n]} f_i^2 g_i^2 + \sum_{i \neq j, i, j \in [n]} f_i g_i f_j g_j$ (12)

Variance is high for large value f_i and g_i

Improving Inner product estimate using CV method

Tug-of-war estimate: $X^{(2)} = \tilde{f}.\tilde{g} = \left(\sum_{i\epsilon[n]} f_i h(i)\right)\left(\sum_{i\epsilon[n]} g_i h(i)\right)$

We choose CV r.v. $Z^{(2)} = \tilde{f}^2 + \tilde{g}^2$

$\Longrightarrow E\left[Z^{(2)}\right] = \boldsymbol{F_2} + \boldsymbol{G_2}$ and $Var\left[Z^{(2)}\right] = 2\left(2\langle \boldsymbol{f}, \boldsymbol{g}\rangle + \boldsymbol{F_2^2} + \boldsymbol{G_2^2}\right)$ (13)

$$Cov\left[X^{(2)}, Z^{(2)}\right] = 2\langle \boldsymbol{f}, \boldsymbol{g}\rangle(\boldsymbol{F_2} + \boldsymbol{G_2}) \quad (14)$$

$$\hat{c} = -\frac{Cov\left[X^{(2)}, Z^{(2)}\right]}{Var\left[Z^{(2)}\right]} = -\frac{\langle \boldsymbol{f}, \boldsymbol{g}\rangle(\boldsymbol{F_2} - \boldsymbol{G_2})}{\left(2\langle \boldsymbol{f}, \boldsymbol{g}\rangle + \boldsymbol{F_2^2} + \boldsymbol{G_2^2}\right)} \quad (15)$$

Variance reduction = $\frac{Cov\left[X^{(2)}, Z^{(2)}\right]^2}{Var\left[Z^{(2)}\right]} = \frac{2\left(\langle \boldsymbol{f}, \boldsymbol{g}\rangle(\boldsymbol{F_2} - \boldsymbol{G_2})\right)^2}{\left(2\langle \boldsymbol{f}, \boldsymbol{g}\rangle + \boldsymbol{F_2^2} + \boldsymbol{G_2^2}\right)}$ (16)

Improving Inner product estimate using CV method

- Our CV estimate of inner product : $X^{(2)} + \hat{c}\left(Z^{(2)} - E[Z^{(2)}]\right)$

Recall:

$$Z^{(2)} = \tilde{f}^2 + \tilde{g}^2 \quad \text{and} \quad E\left[Z^{(2)}\right] = \boldsymbol{F_2} + \boldsymbol{G_2},$$

and

$$\hat{c} = -\frac{\langle \boldsymbol{f}, \boldsymbol{g}\rangle(\boldsymbol{F_2} - \boldsymbol{G_2})}{\left(2\langle \boldsymbol{f}, \boldsymbol{g}\rangle + \boldsymbol{F_2^2} + \boldsymbol{G_2^2}\right)}$$

- For $\langle \boldsymbol{f}, \boldsymbol{g}\rangle$, we use $X^{(2)}$ as a proxy.
- For $\boldsymbol{F_2}$ and $\boldsymbol{G_2}$, we use $\tilde{f}^2$ and $\tilde{g}^2$ obtained by Tug-of-War sketch as a proxy.

Empirical Evaluation

Datasets

- **Synthetic dataset**: We generate a pair of stream using same procedure mentioned for F_2^2 estimation

- **KOS dataset**: We split the corpus into two equal halves consisting of the same number of documents, and we consider each half as a separate data stream.

- **Transaction datasets**: we split the streams in two equal halves and consider each half as a separate data stream

Evaluation Metrics:

- Variance analysis via box-plot
- Mean absolute error
- Median of means estimation

Empirical Evaluation

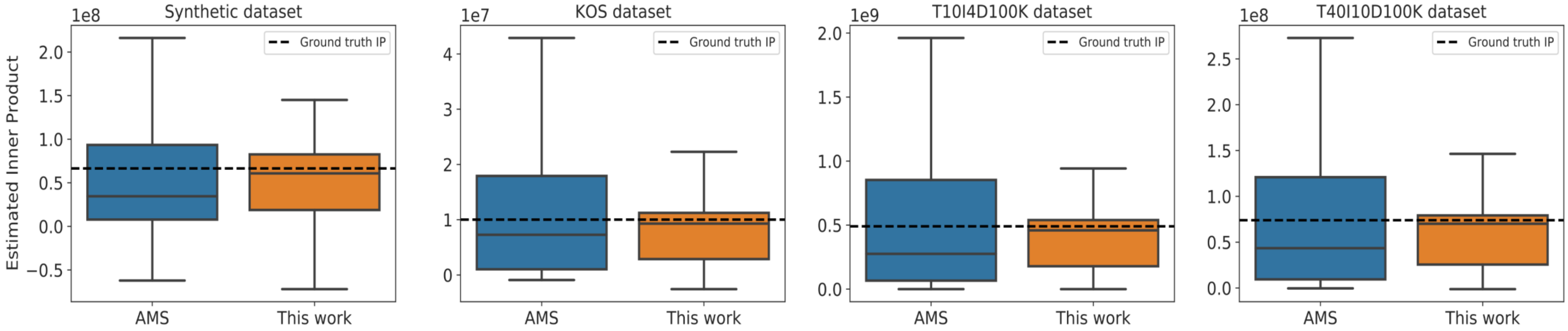


Figure 4

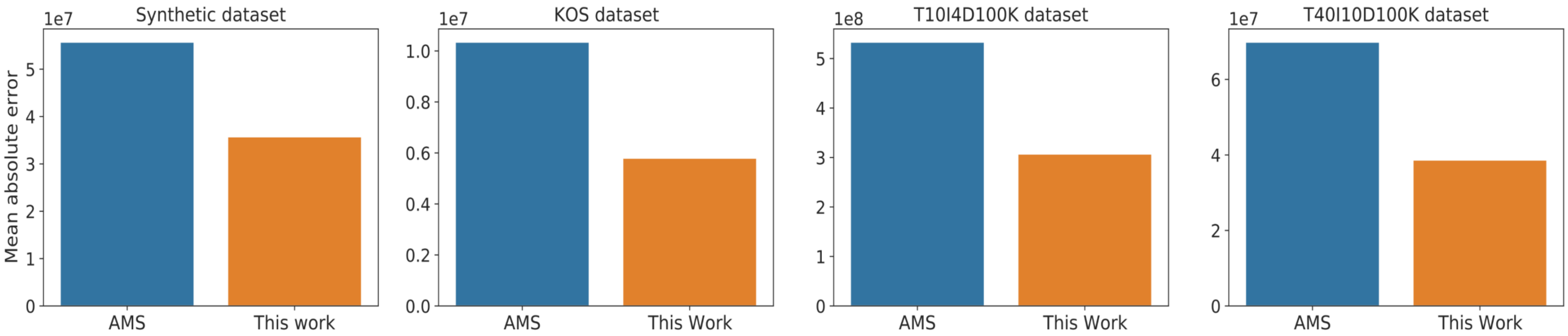


Figure 5

Empirical Evaluation

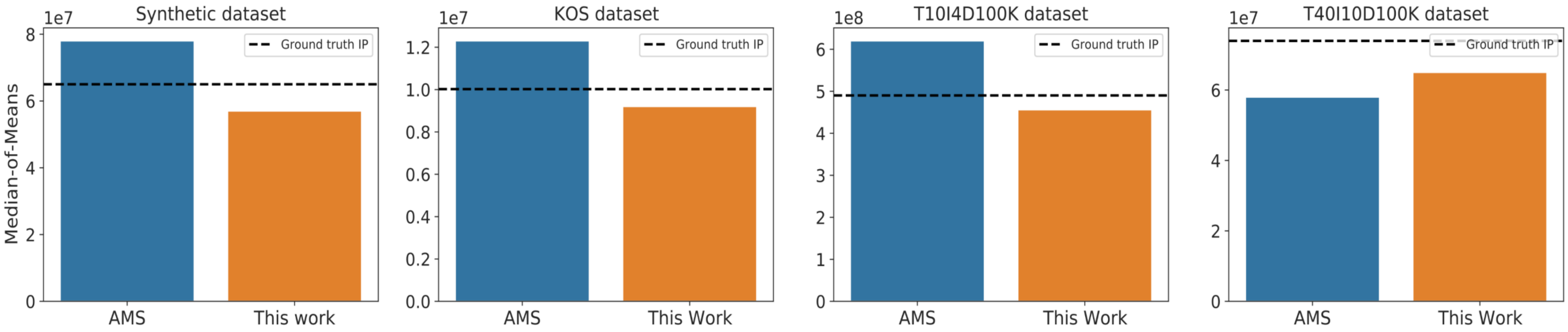


Figure 6

Conclusion and Open Questions

Summary

- ❑ Improving Tug-of-War algorithm for $\boldsymbol{F_2}$ and Inner product estimation using Control-Variate Method.
- ❑ Less overhead and nice empirical performance.

Open Questions

- ❑ Better candidate for Control-variate random variable Z?
- ❑ Possibility of applying in other streaming/randomized algorithms?

Thank You

Questions ?

- bhishamdevverma@gmail.com
- rameshwar.pratap@gmail.com
- kulraghav@gmail.com

References

- Alon, N., Matias, Y., & Szegedy, M. (1999). The space complexity of approximating the frequency moments. *Journal of Computer and system sciences*, *58*(1), 137-147.

- Lavenberg, S. S., & Welch, P. D. (1981). A perspective on the use of control variables to increase the efficiency of Monte Carlo simulations. *Management Science*, *27*(3), 322-335.