Lecture 5: Frequency Moments and AMS Sampler

09-09-2025   Lecturer: Ali Vakilian | Scribe: Caleb McIrvin | Editor: Ali Vakilian

1 Frequency Moment Generalization

While our main focus is on estimating frequency moments (e.g., $F_k = \sum_i f_i^k$), it is important to recognize that many of the underlying sampling techniques can be extended to estimate more general functions of a stream's frequency vector, f. This is particularly true for functions that can be expressed as a sum over the items in the universe, where each term in the sum depends only on the frequency of a single item. Such functions are known as *separable sum functions* and have the general form: $g(f) = \sum_{i=1}^{U} \phi(f_i)$, where U is the size of the universe and ϕ is some function applied to the frequency of each item, where $\phi(0) = 0$. The k-th frequency moment is a classic example of a separable sum function, where $\phi(z) = z^k$.

2 F_2 estimation

This lecture introduces *sampling-based* techniques for estimating frequency moments, $F_k = \sum_{i=1}^{n} f_i^k$, for $k \geq 2$. These moments are fundamental statistics that capture the shape of a data distribution and are a core component in many machine learning applications. For example, the second moment, F_2, is central to computing Euclidean distances and related error measures like MSE. Exact computation is costly because it requires maintaining the full frequency vector $(f_i)_{i \in [n]}$. We will see that by sampling a few items in a carefully manner and tracking only a small subset of frequencies, we can obtain accurate approximations to F_k with sublinear space and per-update time.

2.1 Warm-up: Simple Algorithm via Uniform Sampling

Intuitively, a simple estimator of F_k can be obtained by storing the frequency of a single randomly sampled element and using the result to estimate the k-th frequency moment. While this estimator is unbiased, it suffers from high variance, as we will see.

Algorithm 1 Uniform Sampling Approach

```
1: sample i ∈ [n] uniformly at random
2: f_i ← 0
3: while an item e arrives in stream do
4:     if e = i then
5:         f_i ← f_i + 1
6: return n · f_k^i.
```

The resulting estimator can be formulated as $Z = nf_i^k$. As mentioned previously, this estimator is unbiased. To see this, we take the expectation $\mathbb{E}[Z]$ and note that it is equivalent to the frequency moment F_k, as follows: $\mathbb{E}[Z] = \frac{1}{n} \sum_{i \in [n]} nf_i^k = \sum_{i \in [n]} f_i^k = F_k$.

As only f_i is stored, this algorithm uses $O(\log n)$ bits of space, which is efficient. However, the variance of this estimator is quite large.

Lemma 2.1. $\mathrm{Var}[Z] = nF_{2k} - F_k^2$.

Proof. Compute the second moment: $\mathbb{E}[Z^2] = \mathbb{E}\left[n^2 f_i^{2k}\right] = n^2 \cdot \frac{1}{n}\sum_{i=1}^{n} f_i^{2k} = nF_{2k}$. Therefore $\mathrm{Var}[Z] = \mathbb{E}[Z^2] - \big(\mathbb{E}[Z]\big)^2 = nF_{2k} - F_k^2$. □

Implication for averaging (why this is not useful). Let $\bar{Z} = \frac{1}{t}\sum_{\ell=1}^{t} Z^{(\ell)}$ be the average of t independent copies (using independent sampled indices). Then $\mathrm{Var}[\bar{Z}] = \frac{1}{t}\mathrm{Var}[Z] = \frac{1}{t}\big(nF_{2k} - F_k^2\big)$. In the worst case (e.g., when all mass is on a single coordinate), we have $F_k = |f_{i^\star}|^k$ and $F_{2k} = |f_{i^\star}|^{2k}$, hence

$$\frac{\mathrm{Var}[Z]}{F_k^2} = \frac{nF_{2k} - F_k^2}{F_k^2} = n - 1.$$

By Chebyshev's inequality, to get a constant-probability constant-factor approximation (e.g., relative error $\leq 1/2$ with probability $\geq 2/3$), one needs $t \geq \Theta\left(\frac{\mathrm{Var}[Z]}{F_k^2}\right) = \Theta(n)$. Thus, naive averaging requires $\Theta(n)$ independent repetitions, which defeats the purpose: it is comparable to tracking the entire frequency vector. Consequently, this simple uniform-coordinate sampling approach has too large a variance to be useful for $(1 \pm \varepsilon)$-approximation, motivating more sophisticated sampling approaches that achieve small variance with *sublinear* space.

2.2 Importance Sampling Algorithm

How can we reduce the estimator's variance without increasing the space? Previously we sampled an index uniformly from $[n]$, so the chance of selecting item i did not reflect how often it appears in the stream. This is a poor strategy for estimating F_k (for $k \geq 2$), which is dominated by high-frequency coordinates. A natural fix is *weighted* sampling: choose item i with probability proportional to its frequency f_i. In this part, we show a streaming implementation that uses small sketches and achieves much smaller variance at essentially the same space cost. Algorithmically, this becomes

Algorithm 2 Importance Sampling Approach

sample $i \in [n] \propto \frac{f_i}{F_i}$
$f_i \leftarrow 0$
while an item e arrives in stream **do**
  if $e = i$ **then**
    $f_i \leftarrow f_i + 1$
return $F_1 \cdot f_k^i$.

Calculating the expectation under this sampling method, we see that this estimator is also unbiased.

$$\mathbb{E}[Z] = \sum_{i \in [n]} \frac{f_i}{F_1}(F_1 f_i^{k-1}) = \sum_{i \in [n]} f_i^k = F_k$$

To see that the variance is well-bounded, we calculate

Lemma 2.2. *For $k \geq 2$, $\mathrm{Var}[Z] \leq n^{1-\frac{1}{k}} F_k^2$.*

Proof. As $\mathrm{Var}[Z] \leq \mathbb{E}[Z^2]$, it is sufficient to prove the stronger inequality $\mathbb{E}[Z^2] \leq n^{1-\frac{1}{k}} F_k^2$.

$$\mathbb{E}[Z^2] = \sum_{i=1}^{n} \left(F_1 f_i^{k-1}\right)^2 \cdot \Pr[\text{sample is } i] = \sum_{i=1}^{n} \left(F_1^2 f_i^{2k-2}\right) \cdot \frac{f_i}{F_1} = F_1 \sum_{i=1}^{n} f_i^{2k-1} = F_1 F_{2k-1}$$

Our task is to prove that $F_1 F_{2k-1} \leq n^{1-\frac{1}{k}} F_k^2$.

Claim 2.3. For any value of $k \geq 1$, $F_1 F_{2k-1} \leq n^{1-\frac{1}{k}} F_k^2$.

Proof. We use three standard inequalities that relate different frequency moments. For any frequency vector f:

(i) For any $p \geq q \geq 1$, it holds that $F_p \leq F_q \cdot (\max_j f_j)^{p-q}$.

(ii) The maximum frequency is bounded by the k-th moment: $\max_j f_j \leq \left(\sum_i f_i^k\right)^{1/k} = F_k^{1/k}$.

(iii) By Hölder's inequality, the L_1 and L_k norms are related: $F_1 \leq n^{1-1/k} F_k^{1/k}$.

We can now bound the term $\mathbb{E}[Z^2]$ by applying these inequalities in sequence.

$$\begin{aligned}
\mathbb{E}[Z^2] &= F_1 F_{2k-1} \\
&\leq F_1 \cdot \left(F_k \cdot (\max_j f_j)^{(2k-1)-k} \right) && \text{by Inequality (i)} \\
&= F_1 F_k (\max_j f_j)^{k-1} \\
&\leq F_1 F_k \left(F_k^{1/k} \right)^{k-1} && \text{by Inequality (ii)} \\
&= F_1 F_k F_k^{(k-1)/k} = F_1 F_k^{(2k-1)/k} \\
&\leq \left(n^{1-1/k} F_k^{1/k} \right) \cdot F_k^{(2k-1)/k} && \text{by Inequality (iii)} \\
&= n^{1-1/k} F_k^{(1+2k-1)/k} = n^{1-1/k} F_k^2
\end{aligned}$$

□

We have shown that $\mathbb{E}[Z^2] \leq n^{1-\frac{1}{k}} F_k^2$. Since $\mathrm{Var}(Z) < \mathbb{E}[Z^2]$, the lemma holds. □

This variance bound can be used to achieve a $(1 \pm \varepsilon)$-relative estimate with constant success probability. By averaging $m = O(\varepsilon^{-2} n^{1-1/k})$ independent copies of the base estimator Z, the variance of the resulting average estimator, Z_{avg}, is reduced. An application of Chebyshev's inequality shows this is sufficient for a constant probability guarantee:

$$\Pr\left[|Z_{\text{avg}} - F_k| > \varepsilon F_k \right] \leq \frac{\mathrm{Var}[Z_{\text{avg}}]}{(\varepsilon F_k)^2} \leq \frac{n^{1-1/k} F_k^2 / m}{\varepsilon^2 F_k^2} = O(1)$$

Since we are tracking $O(\varepsilon^{-2} n^{1-1/k})$ estimators, each requiring polylogarithmic space, the overall space complexity becomes $\widetilde{O}(\varepsilon^{-2} n^{1-1/k})$.

While this importance sampling estimator has low variance, it introduces a significant challenge: it requires sampling an item i with a probability, f_i/F_1, that depends on the final frequencies, which are unknown at the start of the stream. This creates a classic "chicken-and-egg" problem, as the algorithm needs a sample at the beginning based on information that is only available at the end. A standard method like *Weighted Reservoir Sampling* might seem like a solution, as it can produce a sample with the desired weighted probabilities. However, it can only guarantee this property for the sample available *after* the entire stream has been processed.

2.3 AMS Sampling

The uniform random sampling nature of reservoir sampling enables it to be used as a subroutine in another sampling method, *AMS sampling* [Alon et al., 1996]. We will see that AMS sampling results in an unbiased, sublinear variance estimator of the frequency moment given a single pass over the stream.

Algorithm 3 AMS-Sample (Stream)

$M \leftarrow 0, C \leftarrow 0, e \leftarrow \perp$
for each item e_t in the stream **do**
  $M \leftarrow M + 1$
  Maintain R_t via reservoir sampling
  if R_t is kept the same as R_{t-1} **then**
    if $e_t = e$ **then**
      $C \leftarrow C + 1$
  else
    $e \leftarrow e_t$
    $C \leftarrow 1$
return $M(C^k - (C-1)^k)$

The algorithm uses three variables: e stores the value of the sampled item, R_t records the stream index where it was sampled, and C counts all subsequent occurrences of e after that index.

Lemma 2.4. *The estimate Z returned by AMS-Sample is unbiased.*

Proof. First note that by the guarantee of the Reservoir sampling, for every $i \in [n]$, $\Pr[e = i] = f_i/F_1$. Let t be the last time the reservoir sampling gets updated, i.e. $e = e_t$ and $R_M = t$. Consider an item $i \in [n]$. If we know that the item sampled is i (i.e., $e = i$), then R_M is uniformly distributed with probability $\frac{1}{f_i}$ among all possible occurrences of i in the stream. Again, we are using the fact that Reservoir sampling, pick any index in the stream uniformly at random; i.e., with probability $1/M$. As a result, the value of C is uniformly sampled from $\{1, \ldots, f_e\}$.

$$
\begin{aligned}
\mathbb{E}[Z] &= \sum_{i=1}^{n} \Pr\left[e = i\right] \sum_{t=1}^{f_i} \Pr\left[C = t\right] \left(M(t^k - (t-1)^k)\right) \\
&= \sum_{i=1}^{n} \frac{f_i}{F_1} \sum_{t=1}^{f_i} \frac{1}{f_i} \left(F_1(t^k - (t-1)^k)\right) = \sum_{i=1}^{n} \sum_{t=1}^{f_i} (t^k - (t-1)^k) \\
&= \sum_{i=1}^{n} f_i^k && \triangleright \text{ The inner sum telescopes to } f_i^k \\
&= F_k
\end{aligned}
$$

□

Next, we bound the variance of the estimate Z.

Theorem 2.5. $\mathrm{Var}\left[Z\right] \leq kn^{1-\frac{1}{k}}(F_k)^2$.

Proof. We provide a stronger upperbound by showing an upperbound for $\mathbb{E}[Z^2]$.

$$\begin{aligned}
\mathbb{E}[Z^2] &= \sum_{i=1}^{n} \Pr[e=i] \sum_{t=1}^{f_i} \Pr[C=t]\, M^2 \left(t^k - (t-1)^k\right)^2 \\
&= \sum_{i=1}^{n} \frac{f_i}{F_i} \sum_{t=1}^{f_i} \frac{1}{f_i} F_1^2 (t^k - t^{k-1})^2 = F_1 \sum_{i=1}^{n} \sum_{t=1}^{f_i} (t^k - (t-1)^k)^2 \\
&\le F_1 \sum_{i=1}^{n} \sum_{t=1}^{f_i} (t^k - (t-1)^k)(kt^{k-1}) && \triangleright \text{ Mean Value Theorem} \\
&\le kF_1 \sum_{i=1}^{n} f_i^{k-1} \sum_{t=1}^{f_i} (t^k - (t-1)^k) \\
&\le kF_1 \sum_{i=1}^{n} f_i^{k-1} f_i^k && \triangleright \text{ The inner sum telescopes to } f_i^k \\
&\le kF_1 F_{2k-1} \\
&\le k \cdot n^{1-\frac{1}{k}} \cdot (F_k)^2 && \triangleright \text{ by Claim 2.3}
\end{aligned}$$

□

By averaging $O(\varepsilon^{-2} n^{1-1/k})$ independent estimators, Chebyshev's inequality guarantees a $(1 \pm \epsilon)$-relative estimate for F_k with constant probability. The space complexity of this approach, $\widetilde{O}(n^{1-1/k})$, is known to be essentially almost optimal for any $k > 2$ [Bar-Yossef et al., 2004, Chakrabarti et al., 2003]. This highlights a key distinction for the second moment, as we will see in a future lecture where we describe a significantly improved, polylogarithmic-space estimator for F_2.

References

Noga Alon, Yossi Matias, and Mario Szegedy. The space complexity of approximating the frequency moments. In *Proceedings of the twenty-eighth annual ACM symposium on Theory of computing*, pages 20–29, 1996.

Ziv Bar-Yossef, T. S. Jayram, Ravi Kumar, and D. Sivakumar. An information statistics approach to data stream and communication complexity. *Journal of Computer and System Sciences*, 68(4):813–844, 2004. doi: 10.1016/j.jcss.2003.11.006.

Amit Chakrabarti, Subhash Khot, and Xiaodong Sun. Near-optimal lower bounds on the multi-pass space complexity of approximating frequency moments. In *Proceedings of the 44th Annual IEEE Symposium on Foundations of Computer Science (FOCS)*, pages 367–376, 2003. doi: 10.1109/SFCS.2003.1238221.