Algorithms for Big Data (Fall 25)

Lecture 5

Frequency Moments and AMS Sampler/Estimator

Ali Vakilian (vakilian@vt.edu)

Frequency Moments

Input: A data stream $S = (e_1, e_2, e_3, \dots, e_N)$, that are seen one by one, where each $e_i \in [n]$ (for known n or an upper bound on n).

- Let f_i denotee the frequency of item i in the stream
- Consider vector $\boldsymbol{f} = (f_1, \dots, f_n)$

The Goal: Given $k \geq 0$, compute the k-th moment of $\boldsymbol{f}$ denoted as

$$F_k = \sum_{i \in [n]} f_i^k$$

Example: $n = 9$ and stream is $9, 1, 1, 3, 5, 8, 9, 7, 2, 1, 3, 9, 8, 4$

- $F_1 = 14$
- $F_2 = 30$

$$\boldsymbol{f} = (3,1,2,1,1,0,1,2,3)$$

Frequency Moments

Input: A data stream $S = (e_1, e_2, e_3, \ldots, e_N)$, that are seen one by one, where each $e_i \in [n]$ (for known n or an upper bound on n).

- Let f_i denotee the frequency of item i in the stream
- Consider vector $\boldsymbol{f} = (f_1, \ldots, f_n)$

The Goal: Given $k \geq 0$, compute the k-th moment of $\boldsymbol{f}$ denoted as

$$F_k = \sum_{i \in [n]} f_i^k$$

Generalization. Estimate $g(S)$ defined as $\sum_{i \in [n]} g_i(f_i)$ where $g_i: \mathbb{R} \to \mathbb{R}$ and $g_i(0) = 0$.

- F_k: For every i, $g_i(x) = x^k$
- Entropy of the stream is defined as $\sum_{i \in [n]} f_i \log f_i$, i.e., $g_i(x) = x \log x$.

  (assume $0 \log 0 = 0$)

Frequency Moments: Questions

(I) Estimation. Given k, estimate F_k exactly/approximately using small memory in one pass over the stream.

(II) Sampling. Given k, sample an item i proportional to f_i^k / F_k using small memory in one pass over the stream.

(III) Sketching. Given k, create a small size summary (sketch) of the frequency vector providing point query (or other statistics), in one pass over the stream.

F_2 Estimation

(I) Estimation. Estimate F_2 exactly/approximately using small memory in one pass over the stream.

(II) Sampling. Sample an item i proportional to f_i^2/F_2 using small memory in one pass over the stream.

- To compute F_2 exactly, we need to keep track of f_i for all $i \in [n]$.
+ However, we afford to keep track of the frequency of a single (or few) item.

Let's try it …

(Recap) When Variance is Small Enough?

If we want to apply Chebyshev's inequality,

$$\Pr\left[|X - \mathbb{E}[X]| > c\mathbb{E}[X]\right] \leq \frac{\mathrm{Var}[X]}{c^2(\mathbb{E}[X])^2}$$

So, we will get $(\epsilon, O(1))$-relative estimate if

$$\mathrm{Var}[X] = \mathbb{E}[X^2] - \mathbb{E}[X]^2 \leq \epsilon^2 \mathbb{E}[X]^2$$

Which holds when

$$\mathbb{E}[X^2] \leq \epsilon^2 \cdot \mathbb{E}[X]^2$$

by Averaging & Median Trick

We can boost it to (ϵ, δ)-relative estimate, in $O(1/\epsilon^2 \log 1/\delta)$ space

F_k-Estimation (Simple Algorithm)

- Let $Z = n \cdot f_i^k$

Is it an unbiased estimation?

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

Though, the issue is its Variance:

$$\mathrm{Var}[Z] = n \cdot F_{2k} - F_k^2$$

- it can get as large as $O(nF_k^2)$
- The averaging technique will need $O(n)$ repetitions which is not good!

Simple Sampling Approach:

sample $i \in [n]$ uniformly at random
$f_i \leftarrow 0$
while an item e in stream arrives:
    if $e = i$ **then**
        $f_i \leftarrow f_i + 1$
return $n \cdot f_i^k$

F_2-Estimation via Sampling

- It's more natural to sample an item proportional to its frequency (Importance/Weighted Sampling)

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

$$\mathbb{E}[Z^2] = F_1 F_{2k-1}$$

Exercise. $F_1 F_{2k-1} \leq n^{1-\frac{1}{k}} \cdot (F_k)^2$

But, how to perform this sampling?

Importance Sampling Approach:

sample $i \in [n]$ at random $\propto \frac{f_i}{F_1}$
$f_i \leftarrow 0$
while an item e in stream arrives:
  if $e = i$ **then**
    $f_i \leftarrow f_i + 1$
return $F_1 \cdot f_i^{k-1}$

$O\left(\epsilon^{-2} n^{1/k}\right)$ samples suffices to get $\mathbb{E}[Z_{\text{avg}}^2] \leq \epsilon^2 \mathbb{E}[Z_{\text{avg}}]^2$

Reservoir Sampling?

- We only get a *random sample* by the end of the stream, not at the beginning
- Still, it has some nice properties, useful for us:
  - sample is u.a.r among the stream seen so far

Sampling technique known as
AMS Sampling

ReservoirSample (stream):

sample $\leftarrow \emptyset$, $t \leftarrow 0$

foreach item x in stream:

$t \leftarrow t + 1$

// Replace with probability $\frac{1}{t}$

if RandomUniform$(0{,}1) < \frac{1}{t}$:

sample $\leftarrow x$

return sample

AMS Sampling

- M is the length of the stream
- $\boldsymbol{e}$ is the value of the sample by Reservoir Sampling, so far.
- R_t is the index of the sample by Reservoir Sampling, so far.
- C is the number of times item $\boldsymbol{e}$ is seen in the stream after index R_t

AMS-Sample (stream):

$M \leftarrow 0, C \leftarrow 0, e \leftarrow \perp$

foreach item e_t in the stream:

$M \leftarrow M + 1$

Maintain R_t via Reservoir Sampling

if R_t is kept the same as R_{t-1} **then**

if $e_t = \boldsymbol{e}$ **then** $C \leftarrow C + 1$

else

$\boldsymbol{e} \leftarrow e_t, R_t \leftarrow t$ and $C \leftarrow 1$

return $M(C^k - (C-1)^k)$

AMS Sampling

- M is the length of the stream
- $\boldsymbol{e}$ is the value of the sample by Reservoir Sampling, so far.
- R_t is the index of the sample by Reservoir Sampling, so far.
- C is the number of times item $\boldsymbol{e}$ is seen in the stream after index R_t

	e_1	e_2	e_3	e_4	e_5	e_6
	1	2	1	3	2	2
$\boldsymbol{e}$						
R_t						
C						

AMS-Sample (stream):

$M \leftarrow 0, C \leftarrow 0, e \leftarrow \perp$

foreach item e_t in the stream:

$M \leftarrow M + 1$

Maintain R_t via Reservoir Sampling

if R_t is kept the same as R_{t-1} **then**

if $e_t = \boldsymbol{e}$ **then** $C \leftarrow C + 1$

else

$\boldsymbol{e} \leftarrow e_t, R_t \leftarrow t$ and $C \leftarrow 1$

return $M(C^k - (C-1)^k)$

AMS Sampling

- M is the length of the stream
- $\boldsymbol{e}$ is the value of the sample by Reservoir Sampling, so far.
- R_t is the index of the sample by Reservoir Sampling, so far.
- C is the number of times item $\boldsymbol{e}$ is seen in the stream after index R_t

$\boldsymbol{T = 1}$

	e_1 ★	e_2	e_3	e_4	e_5	e_6
	1	2	1	3	2	2
$\boldsymbol{e}$	1					
R_t	1					
C	1					

AMS-Sample (stream):

$M \leftarrow 0, C \leftarrow 0, e \leftarrow \perp$

foreach item e_t in the stream:

$M \leftarrow M + 1$

Maintain R_t via Reservoir Sampling

if R_t is kept the same as R_{t-1} **then**

if $e_t = \boldsymbol{e}$ **then** $C \leftarrow C + 1$

else

$\boldsymbol{e} \leftarrow e_t, R_t \leftarrow t$ and $C \leftarrow 1$

return $M(C^k - (C-1)^k)$

AMS Sampling

- M is the length of the stream
- $\boldsymbol{e}$ is the value of the sample by Reservoir Sampling, so far.
- R_t is the index of the sample by Reservoir Sampling, so far.
- C is the number of times item $\boldsymbol{e}$ is seen in the stream after index R_t

$\boldsymbol{T = 2}$

	e_1	e_2 ★	e_3	e_4	e_5	e_6
	1	2	1	3	2	2
$\boldsymbol{e}$	1	2				
R_t	1	2				
C	1	1				

AMS-Sample (stream):

$M \leftarrow 0, C \leftarrow 0, e \leftarrow \perp$

foreach item e_t in the stream:

$M \leftarrow M + 1$

Maintain R_t via Reservoir Sampling

if R_t is kept the same as R_{t-1} **then**

if $e_t = \boldsymbol{e}$ **then** $C \leftarrow C + 1$

else

$\boldsymbol{e} \leftarrow e_t, R_t \leftarrow t$ and $C \leftarrow 1$

return $M(C^k - (C-1)^k)$

AMS Sampling

- M is the length of the stream
- $\boldsymbol{e}$ is the value of the sample by Reservoir Sampling, so far.
- R_t is the index of the sample by Reservoir Sampling, so far.
- C is the number of times item $\boldsymbol{e}$ is seen in the stream after index R_t

$\boldsymbol{T = 3}$

	e_1	e_2 ★	e_3	e_4	e_5	e_6
	1	2	1	3	2	2
$\boldsymbol{e}$	1	2	2			
R_t	1	2	2			
C	1	1	1			

AMS-Sample (stream):

$M \leftarrow 0, C \leftarrow 0, e \leftarrow \perp$

foreach item e_t in the stream:

  $M \leftarrow M + 1$

  Maintain R_t via Reservoir Sampling

  if R_t is kept the same as R_{t-1} **then**

    if $e_t = \boldsymbol{e}$ **then** $C \leftarrow C + 1$

  else

    $\boldsymbol{e} \leftarrow e_t, R_t \leftarrow t$ and $C \leftarrow 1$

return $M(C^k - (C-1)^k)$

AMS Sampling

- M is the length of the stream
- $\boldsymbol{e}$ is the value of the sample by Reservoir Sampling, so far.
- R_t is the index of the sample by Reservoir Sampling, so far.
- C is the number of times item $\boldsymbol{e}$ is seen in the stream after index R_t

$\boldsymbol{T = 4}$

	e_1	e_2	e_3	e_4	e_5	e_6
	1	2	1	3	2	2
$\boldsymbol{e}$	1	2	2	2		
R_t	1	2	2	2		
C	1	1	1	1		

AMS-Sample (stream):

$M \leftarrow 0, C \leftarrow 0, e \leftarrow \perp$

foreach item e_t in the stream:

$M \leftarrow M + 1$

Maintain R_t via Reservoir Sampling

if R_t is kept the same as R_{t-1} **then**

if $e_t = \boldsymbol{e}$ **then** $C \leftarrow C + 1$

else

$\boldsymbol{e} \leftarrow e_t, R_t \leftarrow t$ and $C \leftarrow 1$

return $M(C^k - (C-1)^k)$

AMS Sampling

- M is the length of the stream
- $\boldsymbol{e}$ is the value of the sample by Reservoir Sampling, so far.
- R_t is the index of the sample by Reservoir Sampling, so far.
- C is the number of times item $\boldsymbol{e}$ is seen in the stream after index R_t

$\boldsymbol{T = 5}$

	e_1	e_2 ★	e_3	e_4	e_5	e_6
	1	2	1	3	2	
$\boldsymbol{e}$	1	2	2	2	2	
R_t	1	2	2	2	2	
C	1	1	1	1	2	

AMS-Sample (stream):

$M \leftarrow 0, C \leftarrow 0, e \leftarrow \perp$

foreach item e_t in the stream:

$M \leftarrow M + 1$

Maintain R_t via Reservoir Sampling

if R_t is kept the same as R_{t-1} **then**

if $e_t = \boldsymbol{e}$ **then** $C \leftarrow C + 1$

else

$\boldsymbol{e} \leftarrow e_t, R_t \leftarrow t$ and $C \leftarrow 1$

return $M(C^k - (C-1)^k)$

AMS Sampling

- M is the length of the stream
- $\boldsymbol{e}$ is the value of the sample by Reservoir Sampling, so far.
- R_t is the index of the sample by Reservoir Sampling, so far.
- C is the number of times item $\boldsymbol{e}$ is seen in the stream after index R_t

$\boldsymbol{T = 6}$

	e_1	e_2 ★	e_3	e_4	e_5	e_6
	1	2	1	3	2	2
$\boldsymbol{e}$	1	2	2	2	2	2
R_t	1	2	2	2	2	2
C	1	1	1	1	2	3

AMS-Sample (stream):

$M \leftarrow 0, C \leftarrow 0, e \leftarrow \perp$

foreach item e_t in the stream:

$M \leftarrow M + 1$

Maintain R_t via Reservoir Sampling

if R_t is kept the same as R_{t-1} **then**

if $e_t = \boldsymbol{e}$ **then** $C \leftarrow C + 1$

else

$\boldsymbol{e} \leftarrow e_t, R_t \leftarrow t$ and $C \leftarrow 1$

return $M(C^k - (C-1)^k)$

AMS Sampling

- M is the length of the stream
- $\boldsymbol{e}$ is the value of the sample by Reservoir Sampling, so far.
- R_t is the index of the sample by Reservoir Sampling, so far.
- C is the number of times item $\boldsymbol{e}$ is seen in the stream after index R_t

AMS-Sample (stream):

$M \leftarrow 0, C \leftarrow 0, e \leftarrow \perp$

foreach item e_t in the stream:

$M \leftarrow M + 1$

Maintain R_t via Reservoir Sampling

if R_t is kept the same as R_{t-1} **then**

if $e_t = \boldsymbol{e}$ **then** $C \leftarrow C + 1$

else

$\boldsymbol{e} \leftarrow e_t, R_t \leftarrow t$ and $C \leftarrow 1$

return $M(C^k - (C-1)^k)$

Another run

	e_1	e_2	e_3	e_4	e_5	e_6
	1	2	1	3	2	2
$\boldsymbol{e}$						
R_t						
C						

AMS Sampling

- M is the length of the stream
- $\boldsymbol{e}$ is the value of the sample by Reservoir Sampling, so far.
- R_t is the index of the sample by Reservoir Sampling, so far.
- C is the number of times item $\boldsymbol{e}$ is seen in the stream after index R_t

$\boldsymbol{T'=1}$

	e_1	e_2	e_3	e_4	e_5	e_6
	1	2	1	3	2	2
$\boldsymbol{e}$	1					
R_t	1					
C	1					

AMS-Sample (stream):

$M \leftarrow 0, C \leftarrow 0, e \leftarrow \perp$

foreach item e_t in the stream:

  $M \leftarrow M + 1$

  Maintain R_t via Reservoir Sampling

  if R_t is kept the same as R_{t-1} **then**

    if $e_t = \boldsymbol{e}$ **then** $C \leftarrow C + 1$

  else

    $\boldsymbol{e} \leftarrow e_t, R_t \leftarrow t$ and $C \leftarrow 1$

return $M(C^k - (C-1)^k)$

AMS Sampling

- M is the length of the stream
- $\boldsymbol{e}$ is the value of the sample by Reservoir Sampling, so far.
- R_t is the index of the sample by Reservoir Sampling, so far.
- C is the number of times item $\boldsymbol{e}$ is seen in the stream after index R_t

$\boldsymbol{T' = 2}$

	e_1 ★	e_2	e_3	e_4	e_5	e_6
	1	2	1	3	2	2
$\boldsymbol{e}$	1	1				
R_t	1	1				
C	1	1				

AMS-Sample (stream):

$M \leftarrow 0, C \leftarrow 0, e \leftarrow \perp$

foreach item e_t in the stream:

$M \leftarrow M + 1$

Maintain R_t via Reservoir Sampling

if R_t is kept the same as R_{t-1} **then**

if $e_t = \boldsymbol{e}$ **then** $C \leftarrow C + 1$

else

$\boldsymbol{e} \leftarrow e_t, R_t \leftarrow t$ and $C \leftarrow 1$

return $M(C^k - (C-1)^k)$

AMS Sampling

- M is the length of the stream
- $\boldsymbol{e}$ is the value of the sample by Reservoir Sampling, so far.
- R_t is the index of the sample by Reservoir Sampling, so far.
- C is the number of times item $\boldsymbol{e}$ is seen in the stream after index R_t

$\boldsymbol{T'} = \boldsymbol{3}$

	e_1	e_2	e_3 ★	e_4	e_5	e_6
	1	2	1	3	2	2
$\boldsymbol{e}$	1	1	1			
R_t	1	1	3			
C	1	1	1			

AMS-Sample (stream):

$M \leftarrow 0, C \leftarrow 0, e \leftarrow \perp$

foreach item e_t in the stream:

$M \leftarrow M + 1$

Maintain R_t via Reservoir Sampling

if R_t is kept the same as R_{t-1} **then**

if $e_t = \boldsymbol{e}$ **then** $C \leftarrow C + 1$

else

$\boldsymbol{e} \leftarrow e_t, R_t \leftarrow t$ and $C \leftarrow 1$

return $M(C^k - (C-1)^k)$

AMS Sampling

- M is the length of the stream
- $\boldsymbol{e}$ is the value of the sample by Reservoir Sampling, so far.
- R_t is the index of the sample by Reservoir Sampling, so far.
- C is the number of times item $\boldsymbol{e}$ is seen in the stream after index R_t

$\boldsymbol{T' = 4}$

	e_1	e_2	e_3	e_4	e_5	e_6
	1	2	1	3	2	2
$\boldsymbol{e}$	1	1	1	1		
R_t	1	1	3	3		
C	1	1	1	1		

AMS-Sample (stream):

$M \leftarrow 0, C \leftarrow 0, e \leftarrow \perp$

foreach item e_t in the stream:

$M \leftarrow M + 1$

Maintain R_t via Reservoir Sampling

if R_t is kept the same as R_{t-1} **then**

if $e_t = \boldsymbol{e}$ **then** $C \leftarrow C + 1$

else

$\boldsymbol{e} \leftarrow e_t, R_t \leftarrow t$ and $C \leftarrow 1$

return $M(C^k - (C-1)^k)$

AMS Sampling

Theorem. The estimate Z returned by **AMS-Sample** is an unbiased estimate of F_k

Theorem. $\mathrm{Var}[Z] \le k \cdot n^{1-\frac{1}{k}} \cdot (F_k)^2$

Observations.

- $M =$
- $\forall i \in [n], \Pr[R_M = i] =$

AMS-Sample (stream):

$M \leftarrow 0, C \leftarrow 0, e \leftarrow \perp$

foreach item e_t in the stream:

$M \leftarrow M + 1$

Maintain R_t via Reservoir Sampling

if R_t is kept the same as R_{t-1} **then**

if $e_t = \boldsymbol{e}$ **then** $C \leftarrow C + 1$

else

$\boldsymbol{e} \leftarrow e_t, R_t \leftarrow t$ and $C \leftarrow 1$

return $M(C^k - (C-1)^k)$

AMS Sampling: Expectation Analysis

Let t be the last time (i.e., $R_M = t$) the reservoir sampling gets updated:

- i.e., $R_M = t$, and at the end of the stream, $\boldsymbol{e} = e_t$
- conditioned on $\boldsymbol{e}$ be the value at the end of the stream, the index t (i.e., the value of R_M) is uniformly distributed among the f_e choices of $\boldsymbol{e}$
  - **why?**
  - by the property of reservoir sampling (any index is sampled w.p. $1/M$)
  - if $R_M = \boldsymbol{e}$, the value of C is equally likely to be any of $\{1, \dots, f_e\}$

Theorem. The estimate Z returned by **AMS-Sample** is an unbiased estimate of F_k

AMS Sampling: Expectation Analysis

$$\mathbb{E}[Z] = \sum_{i=1}^{n} \Pr[R_M = i] \cdot \sum_{t=1}^{f_i} \Pr[C = t] \cdot M \cdot \left(t^k - (t-1)^k\right)$$

$$= \sum_{i=1}^{n} \frac{f_i}{F_1} \cdot \sum_{t=1}^{f_i} \frac{1}{f_i} \cdot F_1 \cdot \left(t^k - (t-1)^k\right)$$

$$= \sum_{i=1}^{n} \sum_{t=1}^{f_i} \left(t^k - (t-1)^k\right) = \boldsymbol{F_k}$$

telescopes to f_i^k

Theorem. The estimate Z returned by **AMS-Sample** is an unbiased estimate of F_k

Theorem. $\mathrm{Var}[Z] \leq k \cdot n^{1-\frac{1}{k}} \cdot (F_k)^2$

AMS Sampling: Expectation Analysis

$$\mathbb{E}[Z^2] = \sum_{i=1}^{n} \Pr[R_M = i] \cdot \sum_{t=1}^{f_i} \Pr[C = t] \cdot M^2 \cdot \left(t^k - (t-1)^k\right)^2$$

$$= \sum_{i=1}^{n} \frac{f_i}{F_1} \cdot \sum_{t=1}^{f_i} \frac{1}{f_i} \cdot F_1^2 \cdot \left(t^k - t^{k-1}\right)^2$$

$$= F_1 \cdot \sum_{i=1}^{n} \sum_{t=1}^{f_i} \left(t^k - (t-1)^k\right)^2$$

$$\leq F_1 \cdot k F_{2k-1}$$

Theorem. $\mathrm{Var}[Z] \leq k \cdot n^{1-\frac{1}{k}} \cdot (F_k)^2$

AMS Sampling: Expectation Analysis

$$\sum_{i=1}^{n}\sum_{t=1}^{f_i}\left(t^k-(t-1)^k\right)^2 \le \sum_{i=1}^{n}\sum_{t=1}^{f_i}\left(t^k-(t-1)^k\right)\cdot\left(kt^{k-1}\right)$$

Mean Value Theorem. $\left(t^k-(t-1)^k\right) \le kt^{k-1}$

$$a^k - b^k = (a-b)\sum_{i=0}^{k-1} a^{k-1-i} b^i$$

$$\left(t^k-(t-1)^k\right) = \sum_{i=0}^{k-1} t^{k-i-1}\cdot(t-1)^i \le \sum_{i=0}^{k-1} t^{k-1} = kt^{k-1}$$

Theorem. $\mathrm{Var}[Z] \le k\cdot n^{1-\frac{1}{k}}\cdot(F_k)^2$

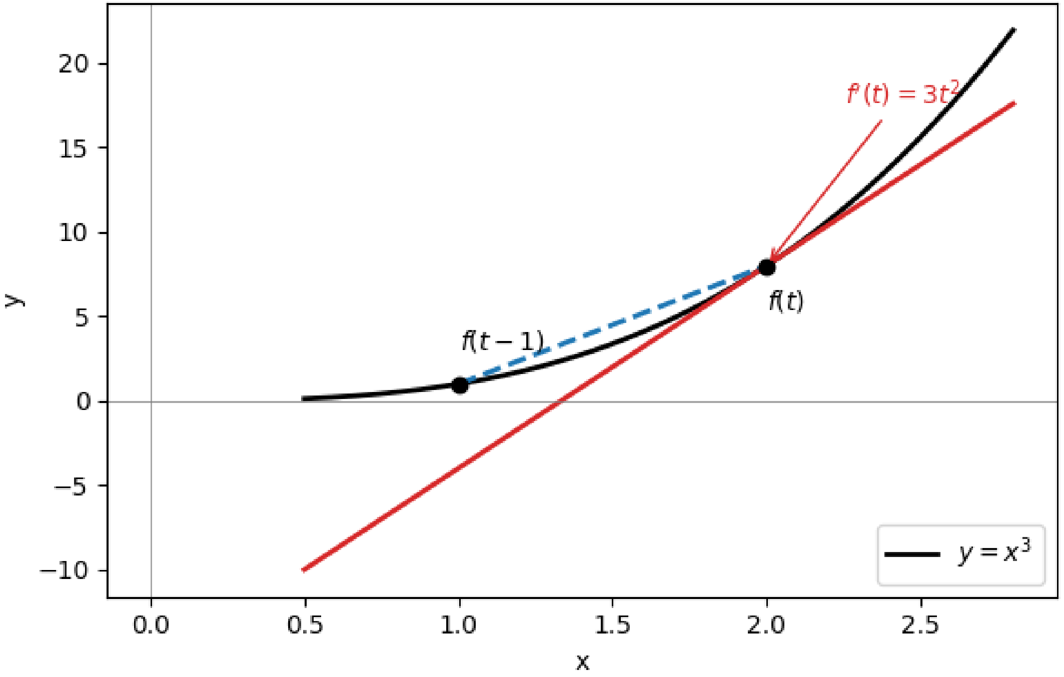
AMS Sampling: Expectation Analysis

$$\sum_{i=1}^{n} \sum_{t=1}^{f_i} \left(t^k - (t-1)^k\right)^2 \leq \sum_{i=1}^{n} \sum_{t=1}^{f_i} \left(t^k - (t-1)^k\right) \cdot \left(kt^{k-1}\right)$$

Mean Value Theorem. $\left(t^k - (t-1)^k\right) \leq kt^{k-1}$

$$)^k\right) = \sum_{i=0}^{k-1} t^{k-i-1} \cdot (t-1)^i \leq \sum_{i=0}^{k-1} t^{k-1} = kt^{k-1}$$

$$f(x) = x^k \Longrightarrow f(t) - f(t-1) \leq f'(t)$$



Theorem. $\mathrm{Var}[Z] \leq k \cdot n^{1-\frac{1}{k}} \cdot (F_k)^2$

AMS Sampling: Expectation Analysis

$$\sum_{i=1}^{n} \sum_{t=1}^{f_i} \left(t^k - (t-1)^k\right)^2 \leq \sum_{i=1}^{n} \sum_{t=1}^{f_i} \left(t^k - (t-1)^k\right) \cdot (k t^{k-1})$$

$$\leq k \sum_{i=1}^{n} f_i^{k-1} \sum_{t=1}^{f_i} \left(t^k - (t-1)^k\right)$$

telescopes to f_i^k

Theorem. $\mathrm{Var}[Z] \leq k \cdot n^{1-\frac{1}{k}} \cdot (F_k)^2$

AMS Sampling: Expectation Analysis

$$\sum_{i=1}^{n}\sum_{t=1}^{f_i}\left(t^k-(t-1)^k\right)^2 \leq \sum_{i=1}^{n}\sum_{t=1}^{f_i}\left(t^k-(t-1)^k\right)\cdot\left(kt^{k-1}\right)$$

$$\leq k\sum_{i=1}^{n} f_i^{k-1}\sum_{t=1}^{f_i}\left(t^k-(t-1)^k\right)$$

$$\leq k\sum_{i=1}^{n} f_i^{k-1} f_i^k$$

$$= \boldsymbol{kF_{2k-1}}$$

Theorem. $\mathrm{Var}[Z] \leq k\cdot n^{1-\frac{1}{k}}\cdot (F_k)^2$

AMS Sampling: Expectation Analysis

$$\mathbb{E}[Z^2] = \sum_{i=1}^{n} \Pr[R_M = i] \cdot \sum_{t=1}^{f_i} \Pr[C = t] \cdot M^2 \cdot \left(t^k - (t-1)^k\right)^2$$

$$= \sum_{i=1}^{n} \frac{f_i}{F_1} \cdot \sum_{t=1}^{f_i} \frac{1}{f_i} \cdot F_1^2 \cdot \left(t^k - t^{k-1}\right)^2$$

$$= F_1 \cdot \sum_{i=1}^{n} \sum_{t=1}^{f_i} \left(t^k - (t-1)^k\right)^2$$

$$\leq F_1 \cdot k F_{2k-1}$$

$$\leq k \cdot n^{1-\frac{1}{k}} \cdot (F_k)^2$$

Theorem. $\mathrm{Var}[Z] \leq k \cdot n^{1-\frac{1}{k}} \cdot (F_k)^2$

AMS Sampling

Theorem. The estimate Z returned by **AMS-Sample** is an unbiased estimate of F_k

Theorem. $\mathrm{Var}[Z] \leq k \cdot n^{1-\frac{1}{k}} \cdot (F_k)^2$

AMS-Sample (stream):

$M \leftarrow 0, C \leftarrow 0, \boldsymbol{e} \leftarrow \perp$

foreach item e_t in the stream:

  $M \leftarrow M + 1$

  Maintain R_t via Reservoir Sampling

  if R_t is kept the same as R_{t-1} **then**

    if $e_t = \boldsymbol{e}$ **then** $C \leftarrow C + 1$

  else

    $\boldsymbol{e} \leftarrow e_t, R_t \leftarrow t$ and $C \leftarrow 1$

return $M(C^k - (C-1)^k)$

By averaging $\Omega(\frac{1}{\epsilon^2} kn^{1-1/k})$ estimators, and applying Chebyshev's inequality:

we get $(1 \pm \epsilon)$ estimate to F_k with constant probability.

AMS-Estimator Wrap-Up

- AMS-Estimator gives a $(1 \pm \epsilon)$-estimation of F_k in $O\left(\frac{1}{\epsilon^2} \cdot n^{1-\frac{1}{k}}\right)$ space.

Is it tight? Can we do better?

- For $k > 2$, it is known to be tight.
- What about F_2?

Estimating F_2

Input: A data stream $S = (e_1, e_2, e_3, \dots, e_N)$, that are seen one by one, where each $e_i \in [n]$ (for known n or an upper bound on n).

- Let f_i denotee the frequency of item i in the stream
- Consider vector $\boldsymbol{f} = (f_1, \dots, f_n)$

The Goal: Compute F_2

The generic AMS estimator gives $(1 \pm \epsilon)$-estimation in $O(\frac{1}{\epsilon^2}\sqrt{n})$ space.

Can we do better?