

Frequency Estimation over a Stream

Notes 28 · Mathematical summary

Sariel Har-Peled

CS 574 · University of Illinois, Urbana-Champaign · Fall 2026

Model and moments

$$\sigma = (s_1, \dots, s_m) \in [n]^m, \quad f_i = |\{t : s_t = i\}|, \quad F_k = \sum_{i=1}^n f_i^k.$$

$$F_0 = |\{i : f_i > 0\}|, \quad F_1 = m, \quad F_\infty = \max_i f_i.$$

Fixed integer $k \geq 1$; $0 < \varepsilon, \delta < 1$; one insertion-only pass.

$$\Pr[|\hat{F}_k - F_k| > \varepsilon F_k] \leq \delta.$$

Word model: $O(\log(n + 1) + \log(m + 1))$ bits; empty stream $\mapsto 0$.

Averaging lemma

Independent X_j with $\mathbb{E}X_j = \mu > 0$, $\text{Var}(X_j) = \nu < \infty$.

$$\alpha = \max\{1, \lceil 4\nu/(\varepsilon^2\mu^2) \rceil\}, \quad A = \alpha^{-1} \sum_{j=1}^{\alpha} X_j.$$

$$\mathbb{E}A = \mu, \quad \text{Var}(A) = \nu/\alpha.$$

$$\Pr[|A - \mu| \geq \varepsilon\mu] \leq \frac{\nu}{\alpha\varepsilon^2\mu^2} \leq \frac{1}{4}.$$

Median amplification

Independent blocks; $\beta = 2\lceil 12 \ln(1/\delta) \rceil + 1$.

$$Z = \text{median}(A_1, \dots, A_\beta), \quad G = \sum_i \mathbf{1}[|A_i - \mu| \leq \varepsilon\mu].$$

$$\mathbb{E}G \geq 3\beta/4, \quad \Pr[|Z - \mu| > \varepsilon\mu] \leq \Pr[G < \beta/2] \leq e^{-\beta/24} \leq \delta.$$

$$M = \alpha\beta = O\left(\left(1 + \frac{\nu}{\varepsilon^2\mu^2}\right) \log \frac{2}{\delta}\right), \quad \text{aggregation time } O(M).$$

Reservoir and suffix rank

At position t : with probability $1/t$, set $(v, r) \leftarrow (s_t, 1)$; otherwise increment r if $s_t = v$.

$$\Pr[P = p] = \frac{1}{p} \prod_{t=p+1}^m \left(1 - \frac{1}{t}\right) = \frac{1}{m}.$$

$$r = |\{t \geq P : s_t = s_P\}|, \quad X = m(r^k - (r-1)^k).$$

Replacement always resets r , including replacement by the same value.

Unbiasedness by telescoping

$$\Delta_k(r) = r^k - (r-1)^k.$$
$$\mathbb{E}X = \frac{1}{m} \sum_{i=1}^n \sum_{r=1}^{f_i} m \Delta_k(r) = \sum_i f_i^k = F_k.$$

$$k=1 : X = m; \quad k=2 : X = m(2r-1).$$

More generally, $g(0) = 0$ implies

$$\mathbb{E}[m(g(r) - g(r-1))] = \sum_i g(f_i).$$

Squared-increment bound

For $1 \leq r \leq f$,

$$0 \leq \Delta_k(r) \leq kr^{k-1} \leq kf^{k-1}.$$

$$\sum_{r=1}^f \Delta_k(r)^2 \leq kf^{k-1} \sum_{r=1}^f \Delta_k(r) = kf^{2k-1}.$$

$$\mathbb{E}X^2 = m \sum_i \sum_{r=1}^{f_i} \Delta_k(r)^2 \leq kmF_{2k-1}.$$

Compare frequency moments

Convexity:

$$\left(\frac{1}{n} \sum_i f_i \right)^k \leq \frac{1}{n} \sum_i f_i^k \implies m \leq n^{1-1/k} F_k^{1/k}.$$

With $L = \max_i f_i$, $L^k \leq F_k$:

$$F_{2k-1} \leq L^{k-1} F_k \leq F_k^{(2k-1)/k}.$$

$$\boxed{m F_{2k-1} \leq n^{1-1/k} F_k^2.}$$

General-moment theorem

$$\text{Var}(X) \leq \mathbb{E}X^2 \leq kmF_{2k-1} \leq kn^{1-1/k}F_k^2.$$

Choose $\alpha = \lceil 4kn^{1-1/k}/\varepsilon^2 \rceil$ and $\beta = 2\lceil 12\ln(1/\delta) \rceil + 1$ independent blocks/copies.

$$\Pr[|\hat{F}_k - F_k| > \varepsilon F_k] \leq \delta,$$

space and direct time/item: $O\left(kn^{1-1/k}\varepsilon^{-2}\log\frac{2}{\delta}\right).$

Fixed k ; $O(1)$ words per reservoir sketch.

A signed sketch for F_2

Unbiased four-wise independent signs $b(i) \in \{-1, +1\}$:

$$T \leftarrow T + b(s_t), \quad T = \sum_{t=1}^m b(s_t) = \sum_{i=1}^n f_i b(i), \quad X = T^2.$$

$$\mathbb{E}X = \sum_i f_i^2 + 2 \sum_{i < j} f_i f_j \mathbb{E}[b(i)b(j)] = F_2.$$

Fourth-moment calculation

Quadruple pattern	Expected sign product
i, i, i, i	1
At least one singleton index	0
Two distinct pairs	1

$$\mathbb{E}T^4 = F_4 + 6 \sum_{i < j} f_i^2 f_j^2.$$

$$\text{Var}(T^2) = 4 \sum_{i < j} f_i^2 f_j^2 \leq 2F_2^2.$$

Polynomial seed

Odd prime $n < p = O(\max\{n, 2\})$; independent uniform $c_0, \dots, c_5 \in \mathbb{F}_p$.

$$g(x) = \sum_{j=0}^5 c_j x^j \pmod{p}.$$

For $r \leq 6$ distinct inputs, the Vandermonde system has rank r :

$$\Pr[g(i_1) = y_1, \dots, g(i_r) = y_r] = p^{-r}.$$

Six coefficients: $O(\log(n + 1))$ bits. Fixed-degree evaluation: $O(1)$ word operations.

Retain zeros and normalize

$$b(i) = \begin{cases} 0, & g(i) = p - 1, \\ +1, & g(i) \text{ odd}, \\ -1, & g(i) \text{ even}, g(i) \neq p - 1. \end{cases} \quad q = \frac{p-1}{p} \geq \frac{2}{3}.$$
$$\mathbb{E}b(i) = 0, \quad \mathbb{E}b(i)^2 = \mathbb{E}b(i)^4 = q.$$

$$T = \sum_i f_i b(i), \quad X = T^2/q, \quad \mathbb{E}X = F_2.$$

No conditioning on nonzero outputs.

Ternary variance

$$\mathbb{E}T^4 = qF_4 + 6q^2 \sum_{i < j} f_i^2 f_j^2.$$

$$\begin{aligned}\text{Var}(T^2/q) &= (1/q - 1)F_4 + 4 \sum_{i < j} f_i^2 f_j^2 \\ &= 2F_2^2 + (1/q - 3)F_4 \leq 2F_2^2.\end{aligned}$$

Average the F_2 sketches

$$\alpha = \lceil 16/\varepsilon^2 \rceil, \quad A_i = \frac{1}{\alpha} \sum_{j=1}^{\alpha} X_{i,j}.$$

All coefficient vectors are independent.

$$\text{Var}(A_i) \leq \frac{2F_2^2}{\alpha}, \quad \Pr[|A_i - F_2| > \varepsilon F_2] \leq \frac{1}{8}.$$

Amplify the F_2 estimate

$$U = \sum_i \mathbf{1}[|A_i - F_2| > \varepsilon F_2].$$

$$\Pr[U \geq \beta/2] \leq 7^{-\beta/2} \mathbb{E}[7^U] \leq \left(\frac{\sqrt{7}}{4}\right)^\beta \leq e^{-\beta/4}.$$

$$\beta = 2 \lceil 2 \ln(1/\delta) \rceil + 1, \quad \hat{F}_2 = \text{median}(A_1, \dots, A_\beta).$$

$$\Pr[|\hat{F}_2 - F_2| > \varepsilon F_2] \leq \delta.$$

F_2 theorem and resource bounds

For any fixed insertion-only stream and $0 < \varepsilon, \delta < 1$,

$$\Pr[(1 - \varepsilon)F_2 \leq \hat{F}_2 \leq (1 + \varepsilon)F_2] \geq 1 - \delta.$$

$$\text{space and time/item: } O\left(\varepsilon^{-2} \log \frac{2}{\delta}\right).$$

Per sketch: six field coefficients and $T \in [-m, m]$; constant-word arithmetic. Prime selection is preprocessing.
Empty stream: return 0.

Bounds at a glance

Target	Basic estimator	Relative variance bound
F_k	$m(r^k - (r - 1)^k)$	$kn^{1-1/k}$
F_2	$(\sum_i f_i b(i))^2 / q$	2

$$\text{copies} = O\left((1 + R\epsilon^{-2}) \log(2/\delta)\right), \quad R \geq \text{Var}(X) / (\mathbb{E}X)^2.$$

Sources

Notes 28, *Frequency Estimation over a Stream*; results attributed to (Alon, Matias, & Szegedy, 1999). Background: (Motwani & Raghavan, 1995).

Alon, N., Matias, Y., & Szegedy, M. (1999). The space complexity of approximating the frequency moments. *J. Comput. Syst. Sci.*, 58(1), 137–147. <https://doi.org/10.1006/jcss.1997.1545>

Motwani, R., & Raghavan, P. (1995). *Randomized algorithms*. Cambridge University Press. <https://doi.org/10.1017/cbo9780511814075>