

Counting Distinct Elements in a Stream

Notes 29 · Mathematical summary

Sariel Har-Peled

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

Model and guarantee

$$\sigma = (s_1, \dots, s_m) \in [n]^m, \quad S = \{s_t : 1 \leq t \leq m\}, \quad u = |S| = F_0.$$
$$\Pr[|\hat{u} - u| > \varepsilon u] \leq \delta.$$

One insertion-only pass; words store keys, counts, and hash seeds. Empty stream $\mapsto 0$; $n = 1$ handled exactly.

Minimum of independent uniforms

$$X_1, \dots, X_u \stackrel{\text{iid}}{\sim} \text{Unif}[0, 1], \quad Y = \min_i X_i, \quad u \geq 1.$$

For $0 \leq t \leq 1$,

$$\Pr[Y > t] = \prod_i \Pr[X_i > t] = (1 - t)^u.$$

$$F_Y(t) = 1 - (1 - t)^u, \quad f_Y(t) = u(1 - t)^{u-1} \quad (0 < t < 1).$$

Moments of the minimum

$$\mathbb{E}Y = \int_0^1 (1-t)^u dt = \frac{1}{u+1}.$$

$$\mathbb{E}Y^2 = 2 \int_0^1 t(1-t)^u dt = 2 \left(\frac{1}{u+1} - \frac{1}{u+2} \right) = \frac{2}{(u+1)(u+2)}.$$

$$\text{Var}(Y) = \frac{u}{(u+1)^2(u+2)}, \quad \frac{\text{Var}(Y)}{(\mathbb{E}Y)^2} = \frac{u}{u+2} < 1.$$

Ideal minimum sketch and amplification

Fully random $h : [n] \rightarrow [0, 1]$:

$$Y = \min_t h(s_t) = \min_{i \in S} h(i), \quad \mu = \frac{1}{u+1}.$$

Independent copies; $\alpha = \lceil 64/\varepsilon^2 \rceil$ per average:

$$\Pr[|A - \mu| > (\varepsilon/4)\mu] \leq \frac{16}{\alpha\varepsilon^2} \leq \frac{1}{4}.$$

$\beta = 2\lceil 12\ln(1/\delta) \rceil + 1$ blocks; $Z = \text{median}(A_i)$:

$$\Pr[|Z - \mu| > (\varepsilon/4)\mu] \leq e^{-\beta/24} \leq \delta.$$

Reciprocal conversion

On success, $Z = (1 + \theta)/(u + 1)$, $|\theta| \leq \varepsilon/4$.

$$\hat{u} = \frac{1}{Z} - 1, \quad \frac{|\hat{u} - u|}{u} = \frac{u + 1}{u} \frac{|\theta|}{|1 + \theta|} \leq \frac{2\varepsilon}{4 - \varepsilon} < \varepsilon.$$

$$\Pr[|\hat{u} - u| > \varepsilon u] \leq \delta, \quad O\left(\varepsilon^{-2} \log \frac{2}{\delta}\right) \text{ ideal words.}$$

Assumption: each perfectly random real hash and stored value costs $O(1)$ words.

Pairwise sampling of distinct values

Pairwise-independent bits: $\Pr[B_i = 1] = p$.

$$R = \{i \in S : B_i = 1\}, \quad X = |R| = \sum_{i \in S} B_i.$$

Insert a key only when its fixed bit is 1; store it only once.

$$\mathbb{E}X = pu, \quad \text{Var}(X) = \sum_{i \in S} p(1 - p) = pu(1 - p).$$

$$\Pr[|X/p - u| > \varepsilon u] \leq \frac{1 - p}{\varepsilon^2 pu}.$$

Capped estimator at scale N

Promise: $u > N/4$, $0 < \varepsilon < 1/4$.

$$p = \min\{1, 32/(N\varepsilon^2)\}, \quad T = pN.$$

If $|R| > 8T$, stop this copy and return **large** ($u > 2N$). Otherwise return $\rho = |R|/p$ at end.

$$|R| \leq \lfloor 8T \rfloor + 1 = O(\varepsilon^{-2}).$$

For $p = 1$: no overflow $\Rightarrow \rho = u$; overflow $\Rightarrow u > 8N > 2N$.

One-scale concentration

For $p < 1$, $pu > 8/\varepsilon^2$. Let X be the final conceptual sample size.

$$\mathcal{G} = \{|X - pu| \leq \varepsilon pu\}, \quad \Pr[\mathcal{G}^c] \leq \frac{1}{\varepsilon^2 pu} < \frac{1}{8}.$$

On $\mathcal{G}$:

$$\text{no overflow} \implies |\rho - u| \leq \varepsilon u;$$

$$\text{overflow} \implies X > 8pN \implies u > 8N/(1 + \varepsilon) > 2N.$$

Amplification with mixed outputs

Run r independent copies.

At least $r/2$ declarations $\Rightarrow$ return **large**.

Otherwise $\Rightarrow$ median of the numerical outputs.

Let B count incorrect copies; $\Pr[\text{copy bad}] \leq 1/8$.

$$\Pr[B \geq r/4] \leq 2^{-r/4} \mathbb{E} 2^B \leq \left(\frac{9/8}{2^{1/4}} \right)^r \leq e^{-r/24}.$$

Proof of mixed-output aggregation

On $B < r/4$:

Declaration majority: at least $r/2$ declarations, so some declaration is correct.

Numerical majority: $q > r/2$ numbers and fewer than $r/4 < q/2$ bad numbers; their median is accurate.

$$r = \lceil 24 \ln(2/\eta) \rceil \implies \Pr[\text{incorrect output}] \leq \eta.$$

$$\text{space } O\left(\varepsilon^{-2} \log \frac{2}{\eta}\right), \quad u > N/4.$$

Finite pairwise-independent hash

$n \geq 2$; prime $n \leq Q < 2n$; independent uniform $a, b \in \mathbb{F}_Q$.

$$h(i) = a(i - 1) + b \pmod{Q}.$$

For $i \neq j$ and any x, y , the equations $h(i) = x, h(j) = y$ uniquely determine (a, b) .

$$\Pr[h(i) = x, h(j) = y] = Q^{-2}.$$

Seed (a, b) : $O(\log n)$ bits; new independent seed for each repetition.

Actual sampling probability

$$p_0 = \min\{1, 32/(N\varepsilon^2)\}, \quad K = \lceil Qp_0 \rceil, \quad B_i = \mathbf{1}[h(i) < K].$$
$$p' = K/Q \geq p_0, \quad T = p'N, \quad \rho = |R|/p'.$$

If $p_0 < 1$ and $N < 2n$,

$$p'u \geq p_0u > 8/\varepsilon^2, \quad T \leq 32/\varepsilon^2 + N/Q < 32/\varepsilon^2 + 2.$$

If $p' = 1$, sampling is exact. The cap remains $O(\varepsilon^{-2})$.

Dyadic scales

$$L = \lceil \log_2 n \rceil, \quad N_i = 2^i \quad (1 \leq i \leq L).$$

Run the amplified sampler at each level with failure budget $\eta = \delta/L$. Return the first numerical output; if none exists, return 0.

$$\text{total space } O\left(\varepsilon^{-2} L \log \frac{2L}{\delta}\right).$$

Witness level

For $u \geq 1$,

$$i_* = \min\{i : N_i \geq u\}, \quad u \leq N_{i_*} \leq 2u.$$

For every $j \leq i_*$,

$$N_j/4 \leq N_{i_*}/4 \leq u/2 < u.$$

All levels through i_* satisfy the one-scale promise.

First numerical output: correctness

$$\Pr[\exists j \leq i_* : \text{level } j \text{ incorrect}] \leq i_* \delta / L \leq \delta.$$

On simultaneous success:

$u \leq N_{i_*}$ rules out a correct **large** declaration at i_* .

Thus a number occurs by i_* , and every earlier numerical output is accurate.

$$\Pr[|\hat{u} - u| > \varepsilon u] \leq \delta.$$

Finite-seed theorem

For $n \geq 2$, $0 < \varepsilon < 1/4$, $0 < \delta < 1$:

$$\Pr[(1 - \varepsilon)F_0 \leq \hat{u} \leq (1 + \varepsilon)F_0] \geq 1 - \delta.$$

$$\text{one pass; } O\left(\varepsilon^{-2} \log n \log \frac{2^{\lceil \log_2 n \rceil}}{\delta}\right) \text{ words.}$$

Empty stream and $n = 1$: exact. For larger requested ε , use $\min\{\varepsilon, 1/8\}$.

High-probability corollary

Fix $a > 0$; take $\delta = n^{-a}$. Then

$$\log(2L/\delta) = O(\log n), \quad \boxed{\text{space } O(\varepsilon^{-2} \log^2 n)}.$$

$$\Pr[(1 - \varepsilon)F_0 \leq \hat{u} \leq (1 + \varepsilon)F_0] \geq 1 - n^{-a}.$$

Sources

Notes 29, *Approximating the Number of Distinct Elements in a Stream*. Prerequisites: median amplification, Chebyshev, Chernoff, and polynomial hashing. Background cited by the notes: (Motwani & Raghavan, 1995).

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