Approximate Counting

Given stream $e_1, e_2, \ldots, e_N$

Goal: output approximation of N (total # elements) trivial: $O(\log N)$ space

using $\ll \log N$ space $\Omega(\log n)$ for deterministic counting

(related, but more difficult: count # distinct elements)

Morris Counter

Initialize: $x \leftarrow 0$

Process token e_i:

With prob. $\frac{1}{2^x}$:

$x \leftarrow x + 1$

Output: $\hat{n} = 2^x - 1$

Lemma Let C_n be the r.v. equal to 2^x after processing token # n (with $C_0 = 2^0 = 1$)

Then $E[C_n] = n + 1$ for all $n \geq 0$.

$\rightarrow$ In particular: $E[C_N] = N + 1$ and $E[\hat{n}] = E[C_N - 1] = n$

i.e., $\hat{n}$ is an unbiased estimator of n

Proof: By induction on n

Base case: $n = 0$

$C_0 = 1 = 0 + 1$ ✓

Inductive step ($n \to n+1$)

Consider r.v. $Z_{n+1} = \begin{cases} 1 & \text{if } x \text{ increased in iteration } i+1 \\ 0 & \text{otherwise} \end{cases}$

$\Pr[Z_{n+1} = 1] = \frac{1}{C_n}$

↑ current state of 2^x

$$C_{n+1} = \begin{cases} 2 \cdot C_n & \text{if } Z_{n+1} = 1 \\ C_n & \text{if } Z_{n+1} = 0 \end{cases}$$

$$= C_n + Z_{n+1} \cdot C_n = (1 + Z_{n+1}) C_n$$

$$E[C_{n+1}] = E[(1 + Z_{n+1}) C_n]$$

intuition: "$(1 + \frac{1}{C_n}) \cdot C_n = 1 + C_n$"

Law of total expectation →

$$= \sum_x E[(1 + Z_{n+1}) \cdot C_n \mid C_n = x] \cdot \Pr[C_n = x]$$

$$= \sum_x E[(1 + Z_{n+1}) \cdot x \mid C_n = x] \cdot \Pr[C_n = x]$$

$$= \sum_x (1 + \underbrace{E[Z_{n+1} \mid C_n = x]}) \cdot x \cdot \Pr[C_n = x]$$

$$= \sum_x \left(1 + \frac{1}{x}\right) \cdot x \cdot \Pr[C_n = x]$$

$$= \sum_x (x+1) \cdot \Pr[C_n = x]$$

$$= \sum_x E[C_n + 1 \mid C_n = x] \cdot \Pr[C_n = x] = E[C_n + 1]$$

$$= E[C_n] + 1 \overset{IH}{=} (n+1) + 1 \quad \checkmark$$

Lemma: For all $n \geq 0$, $\mathrm{Var}[C_n] = \frac{n \cdot (n-1)}{2}$

Proof Recall: $\mathrm{Var}[X] = E[X^2] - (E[X])^2$

We will prove by induction that $E[C_n^2] = 1 + \frac{3(n+1)n}{2}$

(Once we have this: $\mathrm{Var}[C_n] = E[C_n^2] - \underbrace{(E[C_n])^2}_{\text{II prev. Lemma}}$

$$= 1 + \frac{3n \cdot (n+1)}{2} - (n+1)^2 = \ldots = \frac{n(n-1)}{2}$$

Base case: $E[C_0^2] = E[1^2] = 1$ ✓

Inductive step (very similar to before, use same notation)

$n \to n+1$

$$E[C_{n+1}^2] = \sum_x E[C_{n+1}^2 \mid C_n = x] \cdot \Pr[C_n = x]$$ Recall: $C_{n+1} = (1 + Z_{n+1}) \cdot C_n$

$$= \sum_x E[(1 + 2Z_{n+1} + \underbrace{Z_{n+1}^2}_{= Z_{n+1}}) \cdot C_n^2 \mid C_n = x] \cdot \Pr[C_n = x]$$

$$= \sum_x E[(1 + 3Z_{n+1}) \cdot x^2 \mid C_n = x] \cdot \Pr[C_n = x]$$

$$= \sum_x (1 + 3E[Z_{n+1} \mid C_n = x]) \cdot x^2 \cdot \Pr[C_n = x]$$

$$= \sum_x (1 + 3 \cdot \tfrac{1}{x}) \cdot x^2 \cdot \Pr[C_n = x]$$

$$= \sum_x (x^2 + 3x) \cdot \Pr[C_n = x] = \sum_x E[C_n^2 + 3C_n \mid C_n = x] \cdot \Pr[C_n = x]$$

$$= E[C_n^2 + 3C_n]$$

IH + prev. Lemma

$$\leq 1 + \frac{3(n+1)n}{2} + 3(n+1)$$

$$= \ldots = 1 + \frac{3(n+2)(n+1)}{2}$$ ✓

Remark: Let X be r.v. for the final value of X
(so: $\log_2 X$ is the number of bits required by the counter)
Would like to have a bound for $E[X]$

The following holds: $2^{E[X]} \leq E[2^X]$ by Jensen's Inequality, 2^x convex

$= N + 1$

$\Rightarrow \mathbb{E}[X] \leq \log(N+1)$

$\Rightarrow$ expected # bits $\leq \log\log(N+1)$

$\rightarrow$ Can run several "in parallel" and control the probability that at least one does not overflow beyond $O(\log\log N)$ bits

Variance Reduction $\rightarrow$ Works for any unbiased estimator

Recall $\hat{n}$ was an unbiased estimator of N ($\mathbb{E}[\hat{n}] = N$) with

$$\mathrm{Var}(\hat{n}) = \frac{n\cdot(n-1)}{2} \leq \frac{n^2}{2} \qquad \hat{\sigma} := \sqrt{\mathrm{Var}(\hat{n})} \leq \frac{n}{\sqrt{2}}$$

Bound error probability using Chebysher's inequality

$$\Pr[\underbrace{|\hat{n} - N| \geq \varepsilon\cdot n}_{\text{fail to approximate}}] = \Pr[|\hat{n} - N| \geq \varepsilon\cdot\sqrt{2}\cdot\frac{n}{\sqrt{2}}] \overset{?}{\leq} \Pr[|\hat{n} - N| \geq \varepsilon\sqrt{2}\cdot\hat{\sigma}] \leq \frac{1}{2\varepsilon^2}$$

Good approx: $(1-\varepsilon)N \leq \hat{n} \leq (1+\varepsilon)N$

Not good enough. Goal reduce variance to decrease failure probability

Method: Run k independent instances of Morris counter in parallel with final estimates $\hat{n}_1, \hat{n}_2, \ldots, \hat{n}_k$

Return $\hat{n}_{avg} = \frac{1}{k} \sum_{i=1}^{k} \hat{n}_i$

Analysis:

$$E[\hat{n}_{avg}] = \frac{1}{k} \sum_{i=1}^{k} E[\hat{n}_i] = \frac{1}{k} \sum_{i=1}^{k} n = n$$

→ average is still an unbiased estimator

$$Var[\hat{n}_{avg}] = \frac{1}{k^2} \cdot \sum_{i=1}^{k} Var(\hat{n}_i) \leq \frac{1}{k^2} \sum_{i=1}^{k} \frac{n^2}{2} = \frac{n^2}{2k}$$

↑ independent instances

$$\hat{\sigma}_{avg} = \sqrt{Var[\hat{n}_{avg}]} \leq \frac{n}{\sqrt{2k}}$$

Chebyshev's Inequality:

$$Pr[|\hat{n}_{avg} - n| > \varepsilon n] = Pr[|\hat{n}_{avg} - n| > \varepsilon \cdot \sqrt{2k} \cdot \frac{n}{\sqrt{2k}}]$$

$$\leq Pr[|\hat{n}_{avg} - n| > \varepsilon \sqrt{2k} \cdot \hat{\sigma}_{avg}]$$

$$\leq \frac{1}{2k\varepsilon^2} \leq \frac{1}{4}$$

Goal: $\Pr[\text{"failure"}] \leq \frac{1}{4}$

↳ set k such that $\frac{1}{2k\varepsilon^2} \leq 4 \Leftrightarrow k \geq \frac{2}{\varepsilon^2}$

⇒ set $k = \lceil \frac{2}{\varepsilon^2} \rceil$

Summary: by running $\Theta(\frac{1}{\varepsilon^2})$ instances of Morris counter and averaging the results, we return an estimate $\hat{n}_{avg}$ such that $(1-\varepsilon)N \leq \hat{n}_{avg} \leq (1+\varepsilon)N$ with probability $\geq \frac{3}{4}$

→ multiplicative overhead in space compared to single instance

Question: How to reduce failure probability (now $\leq \frac{1}{4}$) to $\leq \delta$ for a tunable parameter δ