Lecture 3: Probabilistic Counting and Morris Counter

09-02-2025 Lecturer: Ali Vakilian | Scribe: Arezoo Sarshartehrani | Editor: Ali Vakilian

1 Estimating the Median in a Stream

While the exact mean of a stream can be computed in minimal space, the median, being a rank-based statistic, is far more challenging. Exact computation would require storing the entire stream, which violates the memory constraints of the streaming model. We therefore turn to randomized approximation approaches.

The standard approach is to maintain a uniform random sample of the stream using Reservoir Sampling. After the stream has passed, the median of this sample is returned as an estimate of the true median. The central question is how large a sample is needed to ensure the estimate is accurate.

Accuracy Guarantee. To formalize accuracy, we define an ϵ**-approximate median** as any value whose rank is between $(\frac{1}{2} - \epsilon)N$ and $(\frac{1}{2} + \epsilon)N$ in the sorted stream. The following theorem guarantees that our sampling-based approach can find such a value with high probability.

Theorem 1.1. *By choosing a sample of size $k = O(\frac{1}{\epsilon^2} \log \frac{1}{\delta})$, the median of the sample is an ϵ-approximate median of the full stream with probability at least $1 - \delta$.*

Proof. Let S be the random sample of size k. Let M_L be the set of the smallest $(\frac{1}{2} - \epsilon)N$ elements in the stream and M_U be the set of the largest $(\frac{1}{2} - \epsilon)N$ elements. The algorithm fails if the median of S lies in either M_L or M_U.

A sufficient condition for the median of S to lie in M_L is if more than half of the sample, $|S|/2$, is drawn from M_L. Let's analyze the probability of this event. Let $|S_L| = |S \cap M_L|$. Since S is a uniform sample, $|S_L|$ is a sum of k independent Bernoulli trials, each with success probability $p = (\frac{1}{2} - \epsilon)$. The expected size is $\mathbb{E}[|S_L|] = kp = k(\frac{1}{2} - \epsilon)$.

The failure condition $|S_L| > k/2$ represents a deviation from the mean of at least $k/2 - k(\frac{1}{2} - \epsilon) = k\epsilon$. By the Chernoff-Hoeffding inequality:

$$\Pr[|S_L| > k/2] = \Pr[|S_L| - \mathbb{E}[|S_L|] > k\epsilon] \leq \exp(-2(k\epsilon)^2/k) = \exp(-2k\epsilon^2)$$

To bound the total failure probability by δ, we use the union bound for the two symmetric failure events (too many samples from M_L or too many from M_U). We set the probability of each tail event to be at most $\delta/2$:

$$\exp(-2k\epsilon^2) \leq \frac{\delta}{2} \implies -2k\epsilon^2 \leq \ln(\delta/2) \implies k \geq \frac{\ln(2/\delta)}{2\epsilon^2}$$

This shows that a sample size of $k = O(\frac{1}{\epsilon^2} \log \frac{1}{\delta})$ is sufficient. □

2 Probabilistic Counting with Morris's Counter

A fundamental problem in data streaming is to estimate the total number of items, N, in a stream. While a trivial deterministic counter requires $O(\log N)$ bits, this can still be too large. Information-theoretic lower bounds show that any deterministic algorithm for this task requires $\Omega(\log N)$ bits, so to achieve a more space-efficient solution, we must turn to randomization and approximation.

Applications. This task of approximate counting is crucial in many large-scale systems. For instance, network routers count the total number of data packets to monitor traffic load and detect denial-of-service attacks. Web services track the total number of error logs to detect system failures in real-time. Financial systems monitor the total volume of transactions to gauge system health and throughput.

The Morris Counter. A classic probabilistic algorithm for this problem is the Morris Counter (1978). Instead of incrementing a counter for every item, it maintains a value X and increments it stochastically.

Algorithm 1 Morris Counter

1: Initialize $X \leftarrow 0$.
2: **for** each item in the stream **do**
3: With probability $1/2^X$, set $X \leftarrow X + 1$.
4: **return** $2^X - 1$.

The core idea is that as the counter's value X increases, the probability of an increment, $1/2^X$, decreases exponentially. This creates a logarithmic relationship between the counter and the true count, leading to a highly compressed representation.

2.1 Analysis of the Morris Counter

We analyze the Morris Counter by studying the random variable $Y_n = 2^{X_n}$, where X_n is the value of the counter after n items.

Expectation. The estimator for the count n is $\hat{n} = Y_n - 1$. We can show this estimator is unbiased.

Theorem 2.1. *For the Morris Counter,* $\mathbb{E}[Y_n] = n + 1$.

Proof. We establish a recurrence for $\mathbb{E}[Y_n]$ using the law of total expectation. Given the value of the counter at step $n-1$, X_{n-1}, we compute the conditional expectation of Y_n:

$$\begin{aligned}\mathbb{E}[Y_n|X_{n-1}] &= \left(2^{X_{n-1}+1}\right) \cdot \frac{1}{2^{X_{n-1}}} + \left(2^{X_{n-1}}\right) \cdot \left(1 - \frac{1}{2^{X_{n-1}}}\right) \\ &= 2 + 2^{X_{n-1}} - 1 = 2^{X_{n-1}} + 1 = Y_{n-1} + 1\end{aligned}$$

Taking the expectation over all values of X_{n-1}, we get $\mathbb{E}[Y_n] = \mathbb{E}[Y_{n-1}] + 1$. With the base case $\mathbb{E}[Y_0] = 2^0 = 1$, an induction analysis using this recurrence solves to $\mathbb{E}[Y_n] = n + 1$. Thus, the estimator $\hat{n} = Y_n - 1$ is unbiased, as $\mathbb{E}[\hat{n}] = (n+1) - 1 = n$. □

Space Complexity. The space used by the algorithm is the number of bits needed to store X_n. We can bound the expected value of X_n using Jensen's Inequality. Since $f(x) = 2^x$ is a convex function, $f(\mathbb{E}[Z]) \leq \mathbb{E}[f(Z)]$. Applying this with $Z = X_n$:

$$2^{\mathbb{E}[X_n]} \leq \mathbb{E}[2^{X_n}] = n + 1 \implies \mathbb{E}[X_n] \leq \log_2(n+1)$$

The expected number of bits needed to store X_n is approximately $\mathbb{E}[\log_2 X_n]$. Since $g(x) = \log_2 x$ is a concave function, Jensen's inequality is reversed: $\mathbb{E}[g(Z)] \leq g(\mathbb{E}[Z])$.

$$\mathbb{E}[\log_2 X_n] \leq \log_2(\mathbb{E}[X_n]) \leq \log_2(\log_2(n+1))$$

This shows that the expected space complexity is $O(\log \log n)$, an exponential improvement over the naive counter.

Variance and Concentration. To understand the estimator's accuracy, we compute its variance.

$$\mathrm{Var}(\hat{n}) = \mathrm{Var}(Y_n - 1) = \mathrm{Var}(Y_n) = \mathbb{E}[Y_n^2] - (\mathbb{E}[Y_n])^2$$

A similar inductive proof shows that $\mathbb{E}[Y_n^2] = \frac{3}{2}n^2 + \frac{3}{2}n + 1$. Combining these results yields:

$$\mathrm{Var}(\hat{n}) = \left(\frac{3}{2}n^2 + \frac{3}{2}n + 1\right) - (n+1)^2 = \frac{n(n-1)}{2} = O(n^2)$$

The variance is quadratic in n. Applying Chebyshev's inequality, $\Pr[|\hat{n} - n| \geq \epsilon n] \leq \frac{\mathrm{Var}(\hat{n})}{(\epsilon n)^2} \approx \frac{n^2/2}{\epsilon^2 n^2} = \frac{1}{2\epsilon^2}$. This bound is quite weak; it only guarantees a constant-factor approximation with constant probability.

2.2 Improving the Estimator: Averaging and the Median Trick

To achieve a high-precision $(1 \pm \epsilon)$-approximation with probability at least $1 - \delta$, we must reduce the high variance of the basic Morris Counter. This is done with a standard and powerful two-level technique that first reduces variance and then amplifies the success probability.

Step 1: Variance Reduction via Averaging. The first step is to create a "base estimator" that is reasonably accurate. While a single Morris counter is unbiased, its $O(n^2)$ variance is too high. We can dramatically reduce this variance by averaging.

Method. We run k independent copies of the Morris counter in parallel. Let their final estimates be $\hat{n}_1, \ldots, \hat{n}_k$. Our new base estimator is their average:

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

Analysis. By linearity of expectation, this new estimator is still unbiased ($\mathbb{E}[\hat{n}_{\text{avg}}] = n$). Because the counters are independent, the variance of the average is the average of the variances:

$$\mathrm{Var}(\hat{n}_{\text{avg}}) = \frac{1}{k^2}\sum_{i=1}^{k} \mathrm{Var}(\hat{n}_i) \approx \frac{1}{k^2} \cdot k \cdot \frac{n^2}{2} = \frac{n^2}{2k}$$

The variance is now reduced by a factor of k. We can apply Chebyshev's inequality to see how accurate this makes our estimate. The probability that our base estimator's error exceeds ϵn is:

$$\Pr\left[|\hat{n}_{\text{avg}} - n| > \epsilon n\right] \leq \frac{\mathrm{Var}(\hat{n}_{\text{avg}})}{(\epsilon n)^2} \approx \frac{n^2/(2k)}{\epsilon^2 n^2} = \frac{1}{2k\epsilon^2}$$

The median trick (our next step) works well if each base estimator has a constant success probability, like $3/4$. To achieve this, we set the failure probability above to be at most $1/4$:

$$\frac{1}{2k\epsilon^2} \leq \frac{1}{4} \implies k \geq \frac{2}{\epsilon^2}$$

By choosing $k = O(1/\epsilon^2)$, we now have a base estimator that is a $(1 \pm \epsilon)$-approximation with constant probability.

Remark. Note that if we wanted to achieve the final failure probability δ using *only* the averaging technique, we would set the Chebyshev bound directly to δ. This would require choosing $k \geq \frac{1}{2\epsilon^2\delta}$, meaning the number of counters would be $O(\frac{1}{\epsilon^2\delta})$. This leads to a natural question: can we achieve the desired confidence with a better dependence on $1/\delta$?

Step 2: Probability Amplification via the Median Trick. Our base estimator is now accurate with a constant probability (e.g., 75%), but we need it to be accurate with a very high probability ($1-\delta$). We achieve this by running multiple independent base estimators and taking their median.

Method. We create ℓ independent copies of our averaged estimator, $E_1, \ldots, E_\ell$. (This means we are now running a total of $k \times \ell$ Morris counters). Our final estimate is:

$$\hat{n}_{\text{final}} = \text{median}(E_1, \ldots, E_\ell)$$

Analysis. The key insight is that for the final median to be wrong, a majority of the base estimators must be wrong. Let's define an indicator variable Z_i to be 1 if the estimator E_i fails (i.e., its error is greater than ϵn). From Step 1, we know $\Pr[Z_i = 1] \leq 1/4$.

The median fails only if the total number of failures, $Z = \sum Z_i$, is at least $\ell/2$. The expected number of failures is low: $\mu = \mathbb{E}[Z] \leq \ell/4$. We are asking for the probability of a large deviation where the number of failures is at least twice its expectation. We can bound this using a Chernoff bound:

$$\Pr[\text{median fails}] = \Pr[Z \geq \ell/2] \leq \Pr[Z \geq 2\mu] \leq e^{-\mu/3} \approx e^{-\ell/12}$$

The probability of the median failing decreases exponentially with ℓ. To meet our final goal, we set this failure probability to be at most δ:

$$e^{-\ell/12} \leq \delta \implies \ell \geq 12\ln(1/\delta)$$

Therefore, by choosing $\ell = O(\log(1/\delta))$, we can make the success probability arbitrarily close to 1.

Takeaway 2.1

The combination of **averaging to reduce variance** followed by the **median trick to amplify probability** is a fundamental and widely used technique in the design of randomized algorithms. It provides a standard recipe for turning a weak estimator (one that is correct on average but unreliable) into one with arbitrarily strong (ϵ, δ) guarantees. For the Morris Counter, this method yields a final space complexity of $O\left(\frac{1}{\epsilon^2}\log\frac{1}{\delta}\log\log n\right)$.