Algorithms for Massive Data Lecture Notes

Darshan Thaker

January 24, 2019

Contents

Lecture 1

Approximate Counting

We consider counting up to $n \in \mathbb{N}$ in the streaming model. To motivate this, consider a router that receives many requests and wants to count the number of requests. Naively, to store a counter for n, this would require $O(\log(n))$ bits of space (the big-O arises from the chosen base of the logarithm, which can be absorbed into a constant). Can we do better? The answer is a resounding no, *unless* we allow for approximation and randomness. If we allow approximation, the surprising answer is that we can use only $O(\log\log(n))$ bits of storage.

Algorithm **??**, known as Morris's algorithm, is such an approximate algorithm.

Algorithm 1.1 Approximate Counting Algorithm [Morris 78]

Given: a stream of data.

Initialize variable $X = 0$.
Upon each new element in the stream, increment X with probability $\frac{1}{2^X}$, else do nothing.
Return estimator $E = 2^X - 1$.

We now analyze the correctness and performance of the Morris algorithm. First, let us start with the correctness of the algorithm. Let X_n be the value of X after n increments, and let $E_n = 2^{X_n} - 1$.

Theorem 1.0.1. *The estimator outputted by the Morris algorithm is a unbiased estimator, i.e.*

$$\mathbb{E}_{X_1,\ldots,X_n}[E_n] = n$$

Proof. The proof proceeds by induction over n. Note that the expectation is over the n values of X (since these are all random variables). First, note the base case of $n = 0$. Clearly, $E_0 = 2^{X_0} - 1 = 2^0 - 1 = 0$, thus the base case is satisfied. Now, assume that the hypothesis holds for E_{n-1}, and we will show that it holds also for E_n. First, we will compute $\mathbb{E}_{X_1,\ldots,X_n}[2^{X_n}]$. We can split this expectation into two expectations:

$$\mathbb{E}_{X_1,\ldots,X_n}[2^{X_n}] = \mathbb{E}_{X_1,\ldots,X_{n-1}}\left[\mathbb{E}_{X_n}\left[2^{X_n}\right]\right]$$

This trick is an important trick. It amounts to fixing the first $n-1$ draws of $X_1, \ldots, X_{n-1}$ and then finding the expectation with respect to the last draw X_n. We can always do this by the rules of probability.

$$\begin{aligned}
&= \mathbb{E}_{X_1,\ldots,X_{n-1}}\left[2^{X_{n-1}+1} \cdot \Pr\left[X_{n-1} \text{ incremented}\right] + 2^{X_{n-1}} \cdot \Pr\left[X_{n-1} \text{ not incremented}\right]\right] \\
&= \mathbb{E}_{X_1,\ldots,X_{n-1}}\left[2^{X_{n-1}+1} \cdot \frac{1}{2^{X_{n-1}}} + 2^{X_{n-1}} \cdot \left(1 - \frac{1}{2^{X_{n-1}}}\right)\right] \\
&= \mathbb{E}_{X_1,\ldots,X_{n-1}}\left[2 + 2^{X_{n-1}} - 1\right] \\
&= \mathbb{E}_{X_1,\ldots,X_{n-1}}\left[2^{X_{n-1}}\right] + 1 \\
&= \mathbb{E}_{X_1,\ldots,X_{n-1}}\left[E_{n-1} + 1\right] + 1 \\
&= \mathbb{E}_{X_1,\ldots,X_{n-1}}\left[n - 1 + 1\right] + 1 && \text{(Inductive hypothesis)} \\
&= n + 1
\end{aligned}$$

Thus, by linearity of expectation (which we used above as well), we have that $\mathbb{E}_{X_1,\ldots,X_n}[E_n] = \mathbb{E}_{X_1,\ldots,X_n}[2^{X_n}] - 1 = n$, completing the proof. □

We are not merely interested in the expectation of the estimator, but also in the concentration of measure of the estimator, which we now prove.

Theorem 1.0.2. *The estimator outputted by the Morris algorithm has the following property .*

$$\mathrm{Var}[E_n] \leq \frac{3n(n-1)}{2} + 1 = O(n^2)$$

Proof. First, we note that instead of proving results about $\mathrm{Var}[E_n]$, we can prove results about $\mathbb{E}[2^{2X_n}]$. To see this, note the following (where all expectations are taken over the random $X_1, \ldots, X_n$:

$$\begin{aligned}
\mathrm{Var}\left[E_n\right] &= \mathrm{Var}\left[2^{X_n} - 1\right] \\
&= \mathrm{Var}\left[2^{X_n}\right] \\
&= \mathbb{E}\left[\left(2^{X_n}\right)^2\right] - \left(\mathbb{E}\left[2^{X_n}\right]\right)^2 \\
&\leq \mathbb{E}\left[2^{2X_n}\right]
\end{aligned}$$

We now proceed via induction on n to show that $\mathbb{E}\left[2^{2X_n}\right] \leq \frac{3n(n-1)}{2} + 1$. Consider the base case for $n = 0$. We know that $\mathbb{E}\left[2^{2X_0}\right] = 1$ based on the initialization of the algorithm, which is less than or equal to $\frac{3 \cdot 0 \cdot (0-1)}{2} + 1$. Now, assume that the hypothesis holds for $\mathbb{E}_{X_1,\ldots,X_{n-1}}\left[2^{2X_{n-1}}\right]$. Following the proof almost exactly as Theorem **??**, we can conclude that:

$$
\begin{aligned}
\mathbb{E}_{X_1,\ldots,X_n}[2^{2X_n}] &= \mathbb{E}_{X_1,\ldots,X_{n-1}}\left[\mathbb{E}_{X_n}\left[2^{2X_n}\right]\right] \\
&= \mathbb{E}_{X_1,\ldots,X_{n-1}}\left[2^{2(X_{n-1}+1)}\cdot\frac{1}{2^{X_{n-1}}} + 2^{2X_{n-1}}\cdot\left(1-\frac{1}{2^{X_{n-1}}}\right)\right] \\
&= \mathbb{E}_{X_1,\ldots,X_{n-1}}\left[2^{X_{n-1}+2} + 2^{2X_{n-1}} - 2^{X_{n-1}}\right] \\
&= \mathbb{E}_{X_1,\ldots,X_{n-1}}\left[4\cdot 2^{X_{n-1}} + 2^{2X_{n-1}} - 2^{X_{n-1}}\right] \\
&= \mathbb{E}_{X_1,\ldots,X_{n-1}}\left[3\cdot 2^{X_{n-1}} + 2^{2X_{n-1}}\right] \\
&= 3\cdot\mathbb{E}_{X_1,\ldots,X_{n-1}}\left[2^{X_{n-1}}\right] + \mathbb{E}_{X_1,\ldots,X_{n-1}}\left[2^{2X_{n-1}}\right] \\
&= 3\cdot\mathbb{E}_{X_1,\ldots,X_{n-1}}\left[E_{n-1}+1\right] + \mathbb{E}_{X_1,\ldots,X_{n-1}}\left[2^{2X_{n-1}}\right] \\
&= 3n + \mathbb{E}_{X_1,\ldots,X_{n-1}}\left[2^{2X_{n-1}}\right] && \text{(Theorem ??)} \\
&\leq 3n + \frac{3(n-1)(n-2)}{2} && \text{(Inductive hypothesis)} \\
&= \frac{3n(n-1)}{2} + 1 \\
&= O(n^2)
\end{aligned}
$$

By the principle of mathematical induction, we can conclude the claim holds for all natural numbers n. □

Combining the above two theorems above, we can now state a probabilistic result about the approximation error of the estimator outputted by Morris's algorithm. Before stating this result, recall Chebyshev's inequality, which will come in handy many times. Chebyshev's inequality states that for any random variable X and for all $\lambda > 0$, we have the following bound:

$$\Pr\left[|X - \mathbb{E}\left[X\right]| \geq \lambda\right] \leq \frac{\mathrm{Var}[X]}{\lambda^2}$$

Theorem 1.0.3. *The estimator outputted by Morris's algorithm E_n is in the range $n \pm O(n)$ with probability at least 90%.*

Proof. We use Chebyshev's inequality on the random variable E_n with the value of $\lambda^2 = 10\mathrm{Var}[E_n]$. By Theorem **??** and Theorem **??**, we know that $\mathbb{E}[E_n] = n$ and $\mathrm{Var}[E_n] = O(n^2)$. Thus, by the Chebyshev bound, we have that E_n falls in the range $\mathbb{E}[E_n] \pm \lambda$ with probability at least 90%, or that E_n falls in the range $n \pm O(n)$ with probability at least 90%. □

Finally, we observe that the Morris algorithm uses $O(\log\log(n))$ bits of storage with high probability, since the algorithm simply stores X and returns $E = 2^X - 1$. Since with high probability, E is "close" to n by the above theorem, we conclude that X requires $\log\log(n)$ bits of storage with high probability (at least 90%).

1.0.1 Variance Improvement by Averaging

We can improve the variance of the estimator outputted by the Morris algorithm using a generic technique. This leads to the so-called Morris+ algorithm, described in Algorithm **??**, which simply outputs the average of k independent runs of the Morris algorithm.

Algorithm 1.2 Approximate Counting Algorithm with Better Variance [Morris 78]

Given: a stream of data, $k \in \mathbb{N}$

Run k independent copies of the Morris algorithm (Algorithm **??**) to get k estimators, $E_1, \ldots E_k$.
Return estimator $E = \frac{1}{k}\sum_{i=1}^{k} E_i$.

We now analyze the correctness of this algorithm (using the same notation as the previous section).

Theorem 1.0.4. *The estimator outputted by the Morris+ algorithm is a unbiased estimator, i.e.*

$$\mathbb{E}_{E_1,\ldots,E_n}[E] = n$$

Proof. This follows easily from the linearity of expectation and the application of Theorem **??**.

$$\begin{aligned}\mathbb{E}_{E_1,\ldots,E_n}[E] &= \frac{1}{k}\sum_{i=1}^{k}\mathbb{E}_{E_i}[E_i] \\ &= \frac{1}{k}\cdot kn \\ &= n\end{aligned}$$

□

Similarly, we can analyze the variance of the estimator.

Theorem 1.0.5. *The estimator outputted by the Morris+ algorithm has the following property (for $n \geq 10$).*

$$\mathrm{Var}[E] \leq \frac{3}{2}\cdot\frac{n^2}{k} = O\left(\frac{n^2}{k}\right)$$

Proof. By properties of the variance and the application of Theorem **??**, we know the following:

$$\begin{aligned}\mathrm{Var}\left[\frac{1}{k}\sum_{i=1}^{k}E_i\right] &= \frac{1}{k^2}\sum_{i=1}^{k}\mathrm{Var}[E_i] \\ &\leq \frac{1}{k^2}\cdot k\cdot\frac{3}{2}n^2 && \text{(If } n \geq 10\text{)} \\ &= \frac{3}{2}\cdot\frac{n^2}{k} \\ &= O\left(\frac{n^2}{k}\right)\end{aligned}$$

□

Notice that the above theorem introduces an inverse dependence of the variance on k (which is a parameter we can pick). Thus, we can pick k to reduce the variance of the estimator as much as we desire. In particular, suppose we wish to achieve a $1+\epsilon$ approximation to the desired output with high probability, i.e. the outputted estimator should be in the range $[n,(1+\epsilon)n]$ with high probability. We can now state a result about this approximation using the Morris+ algorithm.

Theorem 1.0.6. *Using $k = O\left(\frac{1}{\epsilon^2}\right)$, the estimator outputted by the Morris+ algorithm E is in the range $n \pm \epsilon\cdot O(n)$ with probability at least 90%.*

Proof. We use Chebyshev's inequality on the random variable E with the value of $\lambda^2 = 10\mathrm{Var}[E]$. By Theorem **??** and Theorem **??**, we know that $\mathbb{E}[E] = n$ and $\mathrm{Var}[E_n] = O\left(\frac{n^2}{k}\right)$. Thus, by the Chebyshev bound, we have that E falls in the range $\mathbb{E}[E] \pm \lambda$ with probability at least 90%, or that E falls in the range $n \pm \frac{1}{\sqrt{k}}O(n)$ with probability at least 90%. If $k = O\left(\frac{1}{\epsilon^2}\right)$, we have that E falls in the range $n \pm \epsilon\cdot O(n)$ with probability at least 90%. □