The Average-Case Complexity of Counting Distinct Elements

David P. Woodruff
IBM Almaden
650 Harry Road San Jose, CA 95120
dpwoodru@us.ibm.com

ABSTRACT

We continue the study of approximating the number of distinct elements in a data stream of length n to within a $(1\pm\epsilon)$ factor. It is known that if the stream may consist of arbitrary data arriving in an arbitrary order, then any 1-pass algorithm requires $\Omega(1/\epsilon^2)$ bits of space to perform this task. To try to bypass this lower bound, the problem was recently studied in a model in which the stream may consist of arbitrary data, but it arrives to the algorithm in a random order. However, even in this model an $\Omega(1/\epsilon^2)$ lower bound was established. This is because the adversary can still choose the data arbitrarily. This leaves open the possibility that the problem is only hard under a pathological choice of data, which would be of little practical relevance.

We study the average-case complexity of this problem under certain distributions. Namely, we study the case when each successive stream item is drawn independently and uniformly at random from an unknown subset of d items for an unknown value of d. This captures the notion of *random uncorrelated data*. For a wide range of values of d and n, we design a 1-pass algorithm that bypasses the $\Omega(1/\varepsilon^2)$ lower bound that holds in the adversarial and random-order models, thereby showing that this model admits more space-efficient algorithms. Moreover, the update time of our algorithm is optimal. Despite these positive results, for a certain range of values of d and n we show that estimating the number of distinct elements requires $\Omega(1/\varepsilon^2)$ bits of space even in this model. Our lower bound subsumes previous bounds, showing that even for natural choices of data the problem is hard.

Categories and Subject Descriptors

F.2 [**Analysis of Algorithms**]: Theory

General Terms

Algorithms, Performance

ICDT 2009, March 23–25, 2009, Saint Petersburg, Russia.

Keywords

Data Streams, Distinct Elements

1. INTRODUCTION

In recent years the amount of available data has been tremendous. Transaction data, the web and web access logs, as well as network traffic are in abundance. Because of the sheer size of the data, classical algorithms for computing over such data are no longer deemed practical. Processors do not have the capability to store the entire input and making multiple passes over it is prohibitive. Even the most basic statistics of the data cannot be computed exactly by such algorithms in these settings, and so algorithms must be both approximate and probabilistic. This model for algorithm design is known as the streaming model and has become increasingly popular, dating back to the work of Flajolet and Martin [16], and resurging with the seminal work of Alon, Matias, and Szegedy [2]. For a survey, see the book by Muthukrishnan [30].

One of the most well-studied problems in the theory of data streams is the computation or approximation of the number of distinct elements. In the database community, query optimizers can use this statistic to find the number of unique attribute values without having to perform an expensive sort. It is also useful for planning how to execute SQL queries and joins while avoiding redundant computation. Other applications include internet routing, for which internet routers can gather the number of distinct sources and destinations passing through them with only limited memory. This is especially useful for detecting Denial of Service attacks, see [1].

Let $\mathbf{a} = a_1, a_2, \ldots, a_n$ be a sequence of n items, which we refer to as a stream, drawn from a universe of items $[m] = \{1, 2, \ldots, m\}$. We let $F_0 = F_0(\mathbf{a})$ denote the number of distinct elements in the stream. An algorithm A is said to ϵ-approximate F_0 if it outputs an estimate $\tilde{F}_0$ for which $\Pr[|\tilde{F}_0 - F_0| < \epsilon F_0] > 2/3$, where the probability is over the coin tosses of A. The $2/3$ probability can be amplified by taking the median of several independent estimates. We consider algorithms that make a constant (usually one) number of passes over $\mathbf{a}$, and we are concerned with the memory required of such algorithms.

The first streaming algorithm for estimating F_0 in a data stream is due to Flajolet and Martin [16], who assumed the existence of hash functions with certain randomness properties that were unknown to exist. The best known algorithms [5] ϵ-approximate F_0 in $\tilde{O}\left(1/\epsilon^2\right)$ space with one pass, where the notation $\tilde{O}(f)$ means $f \cdot \text{polylog}(nm/\epsilon)$. These algo-

rithms work for an arbitrary stream **a**, i.e., they are data-independent and order-independent.

For a practical setting of m and ϵ (say, $m = 2^{32}$ and $\epsilon = 10\%$), the space complexity is dominated by the $1/\epsilon^2$ term. As the quality of approximation improves, say to $\epsilon = 1\%$, the quadratic dependence on $1/\epsilon$ is a major shortcoming of existing algorithms, and a natural question is if this depedence is optimal.

Using a reduction from the communication complexity of equality, Alon, Matias, and Szegedy [2] showed that for adversarially chosen streams **a**, any one-pass algorithm A which ϵ-approximates $F_0(\mathbf{a})$ must use $\Omega(\log m)$ space. Bar-Yossef [4] showed an $\Omega(1/\epsilon)$ bound via a reduction from the set-disjointness problem. Indyk and Woodruff introduced a new problem, the gap-Hamming problem, to improve this to $\Omega(1/\epsilon^2)$ bits of space, provided $\epsilon = \Omega(m^{-1/(9+k)})$, for any $k > 0$. The analysis was improved by Woodruff [33] who showed an optimal[1] $\Omega(1/\epsilon^2)$ bound for any $\epsilon = \Omega(1/\sqrt{m})$. The proofs were simplified by Jayram, Kumar, and Sivakumar [22].

Despite these negative results, practitioners may need to ϵ-approximate F_0 even when, say, $\epsilon = 1\%$. The question that naturally arises is whether the data stream model, as defined, naturally captures what is happening in practice. Guha and McGregor ([19], [20]) showed that in many real-world applications, the assumption that the data arrives in an adversarial order is too strong. Indeed, consider a setting for which the semantics of the data imply the data is randomly ordered. For example, if employees in a database are sorted by surname and there is no correlation between the ordering of surnames and salaries, then salaries can be thought of as ordered randomly. Several query optimizers already make these assumptions ([19], [20]). Other instances include the "backing sample" architecture proposed by Gibbons, Matias, and Poosala ([17], [18]) for studying aggregate properties in a database, for which the ordering of the data is random by design. Guha and McGregor refer to this model, in which the data is adversarially-chosen but the order in which it appears to a streaming algorithm is random over all possible permutations, as the *random-order model*. This model has become quite popular; see the works of ([9], [10], [20], [19], [21]).

A natural question is whether this model admits algorithms which can ϵ-approximate F_0 in less space. The answer to this turns out to be negative, as shown by Chakrabarti *et al* [9], who show that any algorithm for ϵ-approximating F_0 in the random-order model needs $\Omega(1/\epsilon^2)$ space. The result follows by arguing that random permutations of an adversarially-chosen stream requiring $\Omega(1/\epsilon^2)$ space in the standard model, will, most of the time, still require this amount of space. Thus, the same motivation that led to the creation of the random-order model still exists. Is it possible to suitably adjust the random-order model to better reflect what is happening in practice, so that there is some hope of designing more space-efficient algorithms?

1.1 Our Contributions

We propose the *random-data* model, for which each item of the data stream is drawn from an unknown distribution $\mathcal{D}$. Distibution $\mathcal{D}$ is defined by probabilities $p_1, \ldots, p_m$, and item x occurs as the next item in the stream with probability p_x. Since $\mathcal{D}$ is the same for each successive item, the *random-data* model is contained in the *random-order* model, as all permutations of the data are equally likely. The models are quite different though, since the distribution on items in the random-data model is a product distribution (namely, $\mathcal{D}^n$), whereas in the random-order model any symmetric distribution on any choice of data is possible. Thus, the random-data model better captures the situation when the data is uncorrelated.

The random-data model has been implicitly studied before. Guha and McGregor [21] study the setting in which each element of a data stream is a sample drawn independently from some unknown distribution. In this model they estimate the density function of the unknown distribution, which has applications to learning theory. This is useful for separating the sample and space complexity of learning algorithms. The random-data model is also referred to as the generation oracle model in property testing of distributions [6]. Moreover, the random-data model was assumed by Motwani and Vassilvitskii [29], who studied sampling-based F_0-estimators under the assumption that the distribution of data is Zipfian. Such an assumption turns out to be useful for estimating statistics of the Web Graph and word frequencies in many languages. Many statistical algorithms used in practice for estimating the number of distinct elements based on sampling techniques already impose such an assumption (see the first paragraph in Section 1 of [31], and the many references in [7] at `www.stat.cornell.edu/~bunge/`), without which their performance is known to be poor [11]. One important model is the Generalized Inverse Gaussian Poisson (GIGP) model [8], which allows the data to come from uniform, Zipfian, and other distributions. Finally, the idea of studying streaming algorithms through an instance-specific lens was posed by Kumar and Panigrahy [25], who studied the problem of finding frequent items in terms of the distribution of item frequencies.

We prove bounds on the average-case complexity of estimating the number of distinct elements when $\mathcal{D}$ is uniform over an unknown subset of d items, chosen from the universe $[m]$, for some unknown value of d. That is, exactly d of the item probabilities p_x are non-zero, and they are equal to $\frac{1}{d}$. These distributions capture the natural situation when there are d distinct universe items, for some unknown value of d, and you see a stream of n uncorrelated samples from this universe, i.e., you sample from a set of unknown size with replacement.

Our choice of distribution is fairly robust in the sense that other natural distributions can be reduced to a d-uniform distribution. For example, a distribution with a few items which occur with much larger probability than the remaining items, which are approximately uniformly distributed (i.e., have the same probability up to a small relative error), can be reduced to a d-uniform distribution. Indeed, algorithmically, a heavy-hitters algorithm, such as CountMin [13] or CountSketch [12], can be used to first find and count the items that occur with large probability. These items can be filtered from the stream, and an algorithm for d-uniform distributions, such as the one described in the next paragraph, can provide a good estimate to F_0 on the filtered stream (even though the probabilities are only approximately $\frac{1}{d}$). Similarly, our lower bounds also apply to such distributions since the few heavy-hitters have a negligible contribution

[1]This is optimal since there is an $O(m)$-space algorithm which just maintains the characteristic vector of the underlying set of the data stream.

to F_0. Our algorithm also applies to pseudorandom distributions with support size d, i.e., distributions on d items that cannot be distinguished from uniform with a polynomial number of samples.

The interesting properties of these distributions are that (1) they are fairly natural, (2) for a certain range of d, we show that one can *beat* the space lower bound that holds for adversarially-chosen data, and (3) for another range of d, we show that the lower bound for adversarially-chosen data carries over to these distributions.

More precisely, when $d = \Omega(1/\varepsilon^2)$ and $d \leq n$, we design a 1-pass algorithm that uses an expected $O(d(\log 1/\varepsilon)/(n\varepsilon^2) + \log m)$ bits of space. Moreover, its worst-case update time (i.e., the worst-case processing time per stream item) is $O(1)$ on words of size $O(\log m)$. Notice that if $n = \omega(d \log 1/\varepsilon)$, the algorithm breaks the $\Omega(1/\varepsilon^2)$-space lower bound that holds in the adversarial and random-order models[2]. Even for $d \leq n = O(d \log 1/\varepsilon)$, our algorithm outperforms the best known algorithms in the adversarial and random-order models [5]. The first algorithm in [5] has time $O(\log m \log 1/\varepsilon)$ and space $O(1/\varepsilon^2 \log m)$. Thus, our time and space are always better. The second algorithm in [5] has time $\Omega(1/\varepsilon^2)$ and space $O(1/\varepsilon^2 \log\log m)$, so our time is better, while our space is better for $n = \Omega(d(\log 1/\varepsilon)/\log\log m)$. The third algorithm in [5] has time $\Omega(1/\varepsilon^2)$ and space $O(1/\varepsilon^2(\log\log m + \log 1/\varepsilon))$, so our time and space are always better. This last algorithm has amortized time $O(\log m + \log 1/\varepsilon)$, which is worse than our worst-case time.

Despite these positive results, our main technical contribution is to show that if $n, d = \Theta(1/\varepsilon^2)$, then estimating F_0 requires $\Omega(1/\varepsilon^2)$ bits of space even in the random data model. The lower bound holds even if $\mathcal{D}$ is known to the algorithm. This subsumes all lower bounds in other data stream models for estimating F_0, showing that even for a natural choice of data the problem is hard. Unlike the adversarially-chosen distributions implicitly used in previous lower bounds, which have trivial $O(\log m)$-space 2-pass algorithms, our hard distribution for 1-pass algorithms is the first candidate for proving an $\Omega(1/\varepsilon^2)$-space bound for any constant number of passes, which would resolve a conjecture of Kumar [24]. To support this claim, in the related decision-tree model of computation, this distribution was shown to require depth $\Omega(1/\varepsilon^2)$; see Section 4.5 of [34].

Techniques: Our algorithm for $d = \Omega(1/\varepsilon^2)$ and $d \leq n$ is based on the observation that the frequency of an item in the stream should be about n/d. If n/d were larger than $1/\varepsilon^2$, we could obtain a $(1 \pm \varepsilon)$-approximation to d with constant probability simply from the frequency of the first item in the data stream. Using a balls-and-bins occupancy bound of Kamath *et al* [23] and the fact that $d \leq n$, we can show that a good estimate of d implies a good estimate of F_0. However, suppose $1 \leq n/d < 1/\varepsilon^2$. Then we can instead store the first $O(1/\varepsilon^2)$ items, treat these as a set, and count the number of times some item in the remainder of the stream occurs in this set. This is correct, but unnecessary if d is much less than n. We instead look at the frequency of the first $O(1)$ stream items in the first half of the stream, and use these to obtain a constant factor approximation to d. On the remaining half of the stream we create a set from the first $O(\tilde{d}/(n\varepsilon^2))$ items, where $\tilde{d}$ is our $O(1)$-approximation to d, and count the number of times some item in the remainder of the stream occurs in this set. This makes the space sensitive to the ratio d/n, as the space is now $O(\tilde{d}(\log n)/(n\varepsilon^2) + \log n)$, since we have $O(\tilde{d}/(n\varepsilon^2))$ $\log n$-bit numbers, and we store a single counter. To reduce the logarithmic factor, we sub-sample the universe so that our d-uniform distribution becomes a $\Theta(1/\varepsilon^2)$-uniform distribution over a smaller universe. Now we can store items using $O(\log 1/\varepsilon)$ bits, and we put the items in a perfect hash table to support $O(1)$ update-time. We spread the construction of the perfect hash table over multiple stream updates, so that our worst-case update time is always $O(1)$. We show that estimating the distribution's support size in the sub-sampled stream can be used to estimate d well, and thus can be used to approximate F_0 of the original stream. Finally, we show that we can assume $n \leq m^4$, so $\log n = O(\log m)$, giving the claimed overall space complexity.

To obtain our 1-pass $\Omega(1/\varepsilon^2)$ lower bound in the random data model for $n, d = \Theta(1/\varepsilon^2)$, we look at the 1-round distributional complexity of a two-party communication problem. It is essential that the distribution depend on $\mathcal{D}$, and so we cannot look at the more powerful notion of randomized communication complexity used in previous work. We also consider the distributional complexity of a function rather than that of a promise problem in previous work, and give a combinatorial proof that rectangles in the communication matrix have low discrepancy.

[2]The $\Omega(1/\varepsilon^2)$ bound holds in both models for any $n = \Omega(1/\varepsilon^2)$, since we may take the known hard instance with stream length $\Theta(1/\varepsilon^2)$ and insert $n - \Theta(1/\varepsilon^2)$ copies of a new item x. This only changes F_0 by 1, and the sub-stream restricted to items other than x is the same as before (i.e., the random-order property is preserved).

2. PRELIMINARIES

2.1 The Random-Data Model

DEFINITION 1. *In the* **random-data model** *there is a distribution $\mathcal{D}$ over items $[m]$, and a stream of n independently-drawn samples $D_1, \ldots, D_n$ from $\mathcal{D}$ is seen by a streaming algorithm A in that order. We say an algorithm A $(1 \pm \varepsilon)$-approximates a function $f(D_1, \ldots, D_n)$ if it outputs an estimate $\tilde{f}$ for which $\Pr[|f - \tilde{f}| > \epsilon f] < 1/10$, where the probability is now over both the coin tosses of A and the random variables $D_1, \ldots, D_n$.*

We will restrict our attention to distributions $\mathcal{D}$ on $[m]$ that are uniform over a subset of $[m]$, though as noted earlier, other natural distributions can be reduced to these. Letting d be the size of the subset, we call such a distribution d-uniform.

We can assume that $m \leq \text{poly}(n)$. Indeed, otherwise we may hash the universe down to a set of size n^3, for which the probability of a collision is negligible.

2.2 Communication Complexity

We will need tools from communication complexity for our lower bounds. We assume the reader is familiar with a few standard notions in communication complexity. The interested reader may consult the book by Kushilevitz and Nisan [26] for more detail. Let $f : \mathcal{X} \times \mathcal{Y} \to \{0, 1\}$ be a Boolean function. We will consider two parties, Alice and Bob, receiving x and y respectively, who try to compute $f(x, y)$. We look at protocols which can err on a certain fraction of inputs.

DEFINITION 2. *Let μ be a distribution on $X \times Y$. The (μ, δ)-distributional communication complexity of f, $D_{\mu,\delta}(f)$, is the cost of the best deterministic protocol that gives the correct answer for f on at least a $1-\delta$ fraction of all inputs in $X \times Y$, weighted according to μ.*

In the one-way model, Alice computes some function $A(x)$ of x and sends the result to Bob. Bob then attempts to compute $f(x, y)$ from $A(x)$ and y. Only one message is sent, and it is from Alice to Bob. We define $D^{1-way}_{\mu,\delta}(f)$ similarly to $D_{\mu,\delta}(f)$, but for one-way protocols. We consider a problem tailored to our streaming application. It may seem artificial, but it turns out to be what we need.

DEFINITION 3. *For $x, y \in \{0,1\}^d$, let $\tau = wt(x)+wt(y)-2wt(x)wt(y)/d$, where $wt(x)$ is the Hamming weight of x, i.e., the number of 1s in x. In this problem, Alice is given the inputs $wt(y)$ and x, while Bob is given the inputs $wt(x)$ and y. Let $HAM_d(x, y) = 1$ if $\Delta(x, y) > \tau$, and otherwise let $HAM_d(x, y) = 0$. Here $\Delta(x, y)$ denotes the Hamming distance between x and y, that is, the number of positions that differ.*

3. BREAKING THE $\Omega(1/\varepsilon^2)$ BARRIER

We give an algorithm which illustrates that for a wide range of values of d and n, one can estimate F_0 with less memory and update time in the random data model than required in the adversarial and random-order models. We start by showing that for $d = \Omega(1/\varepsilon^2)$, estimating F_0 reduces to estimating d. For ease of presentation, some proofs are deferred to the appendix. The next lemma can be proven using a strong tail bound [23] for balls-and-bins distributions.

LEMMA 4. *Let $W > 0$ be any positive constant, and suppose $\nu/\varepsilon^2 \le d \le W \cdot n$ for a sufficiently large constant $\nu > 0$. Let d' be such that $d \le d' \le (1+\varepsilon')d$ for a sufficiently small $\varepsilon' = \Theta(\varepsilon)$. Then with probability at least 99/100, over the random data stream, the quantity, $\tilde{F}_0 = d'\left[1-\left(1-\frac{1}{d'}\right)^n\right]$, is a $(1 \pm \varepsilon)$-approximation to F_0.*

We use this lemma to design an algorithm when $d = \Omega(1/\varepsilon^2)$ and $d \le n$. Notice that if $d = o(1/\varepsilon^2)$, we can simply store the hashes of all distinct items in $O(d \log 1/\varepsilon + \log m) = o((\log 1/\varepsilon)/\varepsilon^2 + \log m)$ bits of space. Let ε' be as in Lemma 4, which is any sufficiently small value of the order $\Theta(\varepsilon)$.

We assume, as is typical for streaming algorithms, that n is known in advance, though any $O(1)$-approximation would work in the following algorithm with minor modifications. Moreover, even if one only has an upper bound n' on n, where $n' = \text{poly}(n)$, one can adapt the algorithm with minor modifications. Indeed, one can "guess" $n = 2^i$ for $i = 0, \ldots, \log_2 n'$, and run our algorithm in parallel for each guessed value. At the end of the stream, n is known, and one can extract the information from the state of the algorithm corresponding to the guess which is within a factor of 2. The space only blows up by a $\log n' = O(\log n)$ factor.

THEOREM 5. *If $\nu/\varepsilon^2 \le d \le n$ for a sufficiently large constant $\nu > 0$, then F_0-estimator outputs a $(1\pm\varepsilon)$-approximation to F_0 with probability at least 9/10. The algorithm is 1-pass and uses an expected $O(d(\log 1/\varepsilon)/(n\varepsilon^2) + \log m)$ bits of space. The worst-case update time is $O(1)$ on words of size $O(\log m)$.*

We can assume that $n \le m^4$. If n were any larger, we could restrict our attention to the first m^4 positions in the stream. The above algorithm would have expected space $O(d(\log d)/(m^4\varepsilon^2)+\log m)$ bits. This is $O(1/(m^2\varepsilon^2)+\log m)$ bits because $d \le m$. Moreover, $\varepsilon \ge 1/(m+1)$ since $\varepsilon = 1/(m+1)$ corresponds to computing F_0 exactly. So by replacing n with m^4, the space is still $O(\log m)$ bits. Hence, $\log n = \Theta(\log m)$ (since we also have $m \le n^3$).

The 1/10 error probability can be reduced to probability δ by assuming that $n = \Omega(d \log 1/\delta)$. The algorithm breaks the stream into $\iota = \Theta(\log 1/\delta)$ contiguous substreams each of length $n/\iota \ge d$. It runs F_0-estimator on each substream and outputs the median of its estimates. The output will be a $(1 \pm \varepsilon)$-approximation with probability $\ge 1-\delta$, while by a Markov bound the space will be $O(d(\log 1/\varepsilon)(\log 1/\delta)/(n\varepsilon^2)+\log m \log(1/\delta))$ with arbitrarily large constant probability. We will need the following inequality, proven in the appendix.

CLAIM 6. *Let $0 \le x < 1$ and $y \ge 1$, $x, y \in \mathbf{R}$. Then, $xy - \frac{(xy)^2}{2} \le 1-(1-x)^y \le xy$.*

Proof of Theorem 5: We define several natural probabilistic events.

DistinctStart: We have that $d \ge \nu/\varepsilon^2$ for a sufficiently large constant ν, and so we can assume that

$$\Pr[a_1, \ldots, a_5 \text{ are distinct }] \ge 1 - 25/d \ge 299/300.$$

We call the event that $a_1, \ldots, a_5$ are distinct DistinctStart, which we condition on.

IGood: Notice that $I-5$ is a geometric random variable with expectation $d/5$. By Claim 6 and DistinctStart, we have that $\Pr[I-5 \le d/100]$ is at most

$$1-(1-5/d)^{d/100} \le 5d/(100d) = 1/20.$$

Moreover,

$$\Pr[I-5 > d] = (1-5/d)^d \le e^{-5} \le e^{-\ln 100} \le 1/100.$$

We condition on event IGood: $I-5 \in [d/100, d]$. Hence,

$$A \in \left[\left\lfloor \frac{(c')^2(\varepsilon'')^2 d}{10^4} \right\rfloor, \left\lfloor \frac{(c')^2(\varepsilon'')^2 d}{10^2} \right\rfloor\right],$$

$$B \in \left[\left\lceil \frac{d}{40(c')^2(\varepsilon'')^2 n} \right\rceil, \left\lceil \frac{5d}{2(c')^2(\varepsilon'')^2 n} \right\rceil\right].$$

Using that $d \ge \nu/\varepsilon^2$ for a sufficiently large $\nu > 0$, we can assume $A \ge 1$. Also, by definition $B \ge 1$.

GoodSubstream: Let D be the set of items x in the support of distribution $\mathcal{D}$ for which $g(x) = 1$. Then $\mathbf{E}[|D|] = d/A$. Since g is pairwise-independent, $\mathbf{Var}[|D|] \le d/A$. By Chebyshev's inequality,

$$\Pr\left[\left||D| - \frac{d}{A}\right| > c'\varepsilon''\frac{d}{A}\right] \le \frac{A}{(c')^2(\varepsilon'')^2 d} \le \frac{1}{100},$$

where we conditioned on the event IGood. By the definition of ε'', if this event that we call GoodSubstream occurs, then we have

$$||D| - d/A| \le c'\varepsilon'' d/A = (\varepsilon'/3)(d/A).$$

F_0-estimator:

1. If any of the first $5 = \lceil \ln 100 \rceil$ items are duplicates, output fail. Otherwise, store the set $S = \{a_1, \ldots, a_5\}$ of the first 5 items.
2. Let $I > 5$ be the first position of a duplicate item in S in the first $n/2$ positions of the stream. If no such I exists, output fail.
3. Set the parameters:
   $c' = 100(1152/(9717 \ln 200))^{1/2}$,
   $\varepsilon'' = \varepsilon'/(3c')$,
   $A = \lfloor (c')^2(\varepsilon'')^2(I-5)/100 \rfloor$,
   $B = \lceil 5(I-5)/(2(c')^2(\varepsilon'')^2 n) \rceil$.
   If $A = 0$, then output fail.
4. Let $g : [m] \to [A]$ and $h : [m] \to [\lceil 1/(\varepsilon'')^5 \rceil]$ be 2-wise independent functions.
5. Let $\mathbf{b} = b_1, \ldots, b_r$ be the sub-stream of $a_{n/2+1}, \ldots, a_n$ of items a_j for which $g(a_j) = 1$. That is, b_k is the k-th item among $a_{n/2+1}, \ldots, a_n$ which hashes to 1.
6. Store the set T of the at most B distinct values $h(b_1), \ldots, h(b_B)$ in a perfect hash table.
7. Let C be the number of items in stream $\mathbf{b}$ after position B which hash to a value in T.
8. Set $d' = \frac{A(r-B)|T|}{(1-2\varepsilon'/5)(1-2\varepsilon'')C}$.
9. Output the nearest integer to $d'\left[1 - \left(1 - \frac{1}{d'}\right)^n\right]$.

We condition on this event in the remainder of the proof. Thus,

$$|D| \in [(1-\varepsilon'/3)d/A, (1+\varepsilon'/3)d/A].$$

NoCollisions: Conditioned on GoodSubstream, it follows from the definition of A that $|D| = O(1/\varepsilon^2)$, and since the range of h has size $\lceil 1/(\varepsilon'')^5 \rceil$, the probability there are no collisions is at least $1 - O(\varepsilon)$, which can be assumed to be at least $299/300$ for ε less than a small enough positive constant. We condition on this event, which we call NoCollisions, in the remainder of the proof.

LongSubstream: Conditioned on GoodSubstream, each item in stream $\mathbf{a}$ in positions $n/2+1, \ldots, n$ occurs independently in $\mathbf{b}$ with probability $|D|/d \geq (1-\varepsilon'/3)/A$ (notice that the randomness defining g is independent of the randomness in the data stream, so whether two different items in $\mathbf{a}$ occur in D are indeed independent events). So,

$$\mathbf{E}[r] \geq (n/2)(1-\varepsilon'/3)/A \geq n/(3A)$$

for small enough ε'. Using independence, by a Chernoff bound,

$$\begin{aligned} \Pr[|\mathbf{E}[r] - r| \geq (1/2)\mathbf{E}[r]] &\leq 2e^{-\mathbf{E}[r]/12} \\ &\leq 2e^{-n/(36A)} \\ &= 2e^{-\Theta(n/(d\varepsilon^2))} \\ &\leq \frac{1}{300}, \end{aligned}$$

where the last equality follows from the fact that $A = \Theta(d\varepsilon^2)$, and the last inequality follows from the fact that $n \geq d$ and ε can be made sufficiently small. Call this event that $r \geq n/(6A)$ LongSubstream. We condition on it. Then,

$$r - B \geq \frac{n}{6A} - B \geq \frac{100n}{6(c')^2(\varepsilon'')^2(I-5)} - \frac{5(I-5)}{2(c')^2(\varepsilon'')^2 n} - 1.$$

Notice that, since ε'' can be made arbitrarily small,

$$\begin{aligned} \frac{100}{6(c')^2(\varepsilon'')^2} - \frac{5}{2(c')^2(\varepsilon'')^2} - 1 &\geq \frac{50}{3(c')^2(\varepsilon'')^2} - \frac{3}{(c')^2(\varepsilon'')^2} \\ &= \frac{41}{9(c')^2(\varepsilon'')^2}. \end{aligned}$$

Since $n \geq (I-5)$, we thus have

$$r - B \geq \frac{n}{I-5} \cdot \frac{41}{9(c')^2(\varepsilon'')^2}.$$

GoodEstimation: We now show that the output is a good approximation to F_0. We calculate the expected size of T. Observe that $b_{n/2+1}, \ldots, b_{n/2+B}$ are random and independent elements of D. Now, by event LongSubstream, $r - B \geq 1$, so

$$\mathbf{E}[C] = \frac{(r-B)|T|}{|D|} \geq \frac{n}{I-5} \cdot \frac{41}{9(c')^2(\varepsilon'')^2} \cdot \frac{|T|}{|D|},$$

where the expectation is taken over the randomness in the data stream in the last $r-B$ positions of $\mathbf{b}$, for given values of $r, B, |T|, |D|$, and $I-5$. Using GoodSubstream, $|D| \leq 2d/A$, and so,

$$\mathbf{E}[C] \geq \frac{n}{I-5} \cdot \frac{41}{9(c')^2(\varepsilon'')^2} \cdot \frac{A|T|}{2d} = \frac{41nA|T|}{18(I-5)d(c')^2(\varepsilon'')^2}.$$

Since $A \geq 1$ and $A = \lfloor (c')^2(\varepsilon'')^2(I-5)/100 \rfloor$, we have

$$A \geq (c')^2(\varepsilon'')^2(I-5)/200.$$

$$\begin{aligned}\mathbf{E}[C] &\geq \frac{41nA|T|}{18(I-5)d(c')^2(\varepsilon'')^2}\\ &\geq \frac{41n|T|(c')^2(\varepsilon'')^2(I-5)}{3600(I-5)d(c')^2(\varepsilon'')^2}\\ &= \frac{41n|T|}{3600d}.\end{aligned}$$

Using NoCollisions, we have $\mathbf{E}[|T|] = |D|(1-(1-1/|D|)^B)$, where the expectation is over the first B items of $\mathbf{b}$, for a given B and $|D|$. Now,

$$|D| \geq 2d/A \geq 200d/((c')^2(\varepsilon'')^2(I-5))$$

using GoodSubstream. Since $I-5 \leq d$ and ε'' can be made sufficiently small, $|D| \geq 40$, so if $B=1$, then $|D| \geq 40B$. Otherwise, $B>1$, and since

$$B = \lceil 5(I-5)/(2(c')^2(\varepsilon'')^2 n)\rceil,$$

it follows that

$$B \leq 5(I-5)/((c')^2(\varepsilon'')^2 n).$$

Now,

$$I-5 \leq d \leq n,$$

and so

$$|D| \geq 200/((c')^2(\varepsilon'')^2),$$

while

$$B \leq 5/((c')^2(\varepsilon'')^2),$$

so again $|D| \geq 40B$. Using Claim 6,

$$\begin{aligned}\mathbf{E}[|T|] &= |D|(1-(1-1/|D|)^B)\\ &\geq |D|(B/|D| - B^2/(2|D|^2))\\ &\geq B - B^2/(2|D|)\\ &\geq B - B^2/(80B)\\ &\geq 79B/80.\end{aligned}$$

Taking the expectation over g and the second half of stream $\mathbf{a}$,

$$\begin{aligned}\mathbf{E}[C] &\geq \frac{41n\mathbf{E}|\mathbf{T}|}{3600d}\\ &\geq \frac{41\cdot 79nB}{3600\cdot 80d}\\ &\geq \frac{41\cdot 79nd}{40\cdot(c')^2\cdot(\varepsilon'')^2\cdot n\cdot 3600\cdot 80\cdot d}\\ &= \frac{c''}{(c')^2\cdot(\varepsilon'')^2},\end{aligned}$$

where

$$c'' = 41\cdot 79/(40\cdot 3600\cdot 80).$$

Now, c' was chosen so that we have

$$\mathbf{E}[C] \geq 3\cdot\ln(200)/(\varepsilon'')^2.$$

Since the C_k are independent Bernoulli random variables, by a Chernoff bound,

$$\begin{aligned}\Pr[|C-\mathbf{E}[C]| &\geq \varepsilon''\mathbf{E}[C]]\\ &\leq 2e^{-(\varepsilon'')^2\mathbf{E}[C]/3}\\ &\leq 2e^{-\ln 200}\\ &\leq 1/100.\end{aligned}$$

If $|C-\mathbf{E}[C]| < \varepsilon''\mathbf{E}[C]$, then we say that event GoodEstimation has occurred. By a union bound, we have that the events DistinctStart, IGood, GoodSubstream, NoCollisions, LongSubstream, and GoodEstimation simultaneously occur with probability at least

$$1-\frac{1}{300}-\frac{1}{20}-\frac{1}{100}-\frac{1}{100}-\frac{1}{300}-\frac{1}{300}-\frac{1}{100} = \frac{19}{20}-\frac{4}{100}.$$

Since GoodEstimation occurs,

$$|C-\mathbf{E}[C]| \leq \varepsilon''\mathbf{E}[C].$$

Recalling that

$$\mathbf{E}[C] = (r-B)|T|/|D|,$$

we have $(r-B)|T|/C$ is a $(1\pm 2\varepsilon'')$-approximation to $|D|$ for sufficiently small ε''. It follows that $A(r-B)|T|/C$ is a $(1\pm 2\varepsilon'/5)(1\pm 2\varepsilon'')$-approximation to d (for small enough ε'). It follows by scaling by $(1-2\varepsilon'/5)(1-2\varepsilon'')$, we have

$$d \leq d' \leq (1+2\varepsilon'/5)(1+2\varepsilon'')d/((1-2\varepsilon'/5)(1-2\varepsilon'')) \leq (1+\varepsilon')d$$

(for small enough ε'). Thus, we may apply Lemma 4, which shows that with probability at least 99/100, the output is a $(1\pm\varepsilon)$-approximation to F_0 (when we take the nearest integer, we can have an additional additive 1 error, which since $d, n = \Omega(1/\varepsilon^2)$, is negligible). The overall correctness probability is at least $19/20 - 4/100 - 1/100 = 9/10$.

Steps 1-5 require $O(\log m)$ bits of space. Steps 6-9 require $O(B\log 1/\varepsilon + \log m)$ bits. Note that in expectation the hash table will have size $O(B\log 1/\varepsilon)$ bits [28]. The expected space complexity is

$$\begin{aligned}&O(\mathbf{E}[B]\log 1/\varepsilon + \log m)\\ =\ &O(\mathbf{E}[I-5](\log 1/\varepsilon)/(\varepsilon^2 n) + \log m)\\ =\ &O(d(\log 1/\varepsilon)/(\varepsilon^2 n) + \log m),\end{aligned}$$

as desired. Our update time is $O(1)$ for the first and second halves of the stream. Since we use a perfect hash table, checking membership can be done in $O(1)$ time. The only issue is how to quickly build the table. With high probability, the table can be built in $O(B)$ time [28] (if not, we output fail). This $O(B)$ work can be spread out over the $\Theta(B)$ udpates following b_B. Note that LongSubstream guarantees that $r-B = \Omega(B)$, so this is possible. Thus, the worst-case update time is $O(1)$. □

We remark that there are some streams for which F_0-Estimator always outputs the wrong answer. This is unavoidable if one wants to maintain small space complexity, since some of these streams do in fact correspond to the "hard instances" used to establish the $\Omega(1/\varepsilon^2)$ space lower bound in the case of adversarially-chosen data, though they occur with very low probability.

4. DISTINCT ELEMENTS IS HARD EVEN FOR RANDOM STREAMS

We define the distribution μ_d on $\{0,1\}^d$ for $d = 1/\varepsilon^2$ to be the distribution of characteristic vectors induced by the product distribution $\mathcal{D}^{n/2}$, where $n = \Theta(d)$ is such that $1-(1-1/d)^{n/2} \in [1/3, 2/3]$. We may assume, by adjusting d, n by constant factors, that they are integers. Notice that if $X \sim \mu_d$, then $X_i = 1$ if the "i-th bin" contains one of the "$n/2$ balls". Notice that

$$\mathbf{E}[X_i] = 1-(1-1/d)^{n/2}.$$

It is well-known (see, e.g., chapter 5 of [27]) that $wt(X)$ is tightly concentrated around its expectation since it corresponds to the number of non-empty bins when throwing $n/2$ balls into d bins. Moreover, it is known that

$$\Pr[d/4 \leq wt(X) \leq 3d/4] = 1 - o(1).$$

By symmetry, for $0 \leq k \leq d$ the distribution $\mu_d \mid wt(X) = k$ is uniform on strings of weight k. Put $\mu = \mu_d \times \mu_d$.

LEMMA 7. *Let $k, \ell \in [d/4, 3d/4]$ be arbitrary, and let $X \sim \mu_d \mid wt(X) = k$ and $Y \sim \mu_d \mid wt(Y) = \ell$ be independent random variables. Then for any constant $\delta > 0$, for a sufficiently small choice of the constant $\alpha > 0$,*

$$\Pr[|\Delta(X,Y) - \mathbf{E}[\Delta(X,Y)]| \geq \alpha\sqrt{d}] > 1 - \delta.$$

Also,

$$\Pr[\Delta(X,Y) < \mathbf{E}[\Delta(X,Y)]] \in [1/2 - o(1), 1/2 + o(1)].$$

PROOF. By symmetry, $\mu_d \mid wt(X) = k$ (respectively, $\mu_d \mid wt(Y) = \ell$) is uniform over all strings containing exactly k (respectively ℓ) ones. Suppose X has ones in the set $A \subseteq [d]$ of size k, and Y has ones in the set $B \subseteq [d]$ of size ℓ. Then $\Delta(X,Y) = k + \ell - 2|A \cap B|$. The random variable $|A \cap B|$ is identically distributed to $|A \cap B| \mid (A = a)$ for any fixed set a of size k, so we may assume that

$$A = [k] = \{1, 2, \ldots, k\}.$$

Thus, for any

$$i \in \{0, 1, \ldots, \min(k, \ell)\},$$

we have

$$\Pr[|[k] \cap B| = i] = \frac{\binom{k}{i}\binom{d-k}{\ell-i}}{\binom{d}{\ell}}.$$

Thus, $|[k] \cap B|$ follows a hypergeometric distribution. We need the following normal approximation to the hypergeometric distribution[3] [15].

THEOREM 8. *For $0 < p < 1$, $q = 1 - p$, and any $0 \leq r \leq M$, if $N \to \infty$, $M \to \infty$ so that $M/N \to \nu \in (0,1)$, and $(r - Mp)/\sqrt{Mpq} \to x$, then for $a = 1/(1-\nu)$,*

$$\frac{\binom{Np}{r}\binom{Nq}{M-r}}{\binom{N}{M}} = (1 - o(1))\frac{e^{-ax^2/2}}{\sqrt{2\pi Mpq(1-\nu)}}.$$

Setting $N = d$, $p = k/d$, $r = i$, $q = 1 - k/d$, and $M = \ell$, we have $p, q \in [1/4, 3/4]$, $M/N = \nu \in [1/4, 3/4]$ and $(r - Mp)/\sqrt{Mpq} = x = (i - k\ell/d)/\Theta(\sqrt{d})$. For $i = k\ell/d + \alpha\sqrt{d}$ for any $\alpha \in \mathbf{R}$, we thus have,

$$\Pr[|[k] \cap B| = i] = \binom{k}{i}\binom{d-k}{\ell-i}/\binom{d}{\ell} = \Theta(e^{-\Theta(\alpha^2)}/\sqrt{d}),$$

where the constants in the $\Theta(\cdot)$ notation are absolute. Thus,

$$\Pr[||[k] \cap B| - k\ell/d| \leq \alpha\sqrt{d}/2] \leq \alpha\sqrt{d}/\Theta(\sqrt{d}).$$

For small enough α, this is $\leq \delta$. Now,

$$\mathbf{E}[\Delta(X,Y)] = k + \ell - 2\mathbf{E}[A \cap B],$$

where

$$\mathbf{E}[A \cap B] = k\ell/d,$$

and so

$$\Pr[|\Delta(X,Y) - \mathbf{E}[\Delta(X,Y)]| \geq \alpha\sqrt{d}]$$

$$= \Pr[\mathbf{E}[||[k] \cap B| - k\ell/d| \geq \alpha\sqrt{d}/2] \geq 1 - \delta.$$

This proves the first part of the lemma. The second part follows from the symmetry of Theorem 8. □

Let $S_{\mathcal{D},\epsilon,\ell}(F_0)$ be the minimum space complexity, over all algorithms A which ϵ-approximate F_0 with ℓ-passes in the random-data model with item distribution $\mathcal{D}$ with probability at least $1 - \delta/2$ for a constant $\delta > 0$.

THEOREM 9. *For any $\epsilon > 0$ and constants $\delta, \ell > 0$, we have that*

$$S_{\mathcal{D},\epsilon',\ell}(F_0) \geq D_{\mu,\delta}(HAM_d)$$

and

$$S_{\mathcal{D},\epsilon',1}(F_0) \geq D^{1-way}_{\mu,\delta}(HAM_d),$$

where $\epsilon' = \Theta(\epsilon)$.

PROOF. Let M be an ℓ-pass ϵ'-approximation algorithm for F_0 in the random-data model with distribution $\mathcal{D}$, which succeeds with probability at least $1 - \delta/2$, and let $n/2$ and μ_d be as above. Alice is given $X \sim \mu_d$, and Bob is given $Y \sim \mu_d$, where X and Y are independent. Moreover, Alice is also given $wt(Y)$ and Bob is also given $wt(X)$. Conditioned on X, Alice chooses a random stream $\mathbf{a}_X$ of length $n/2$ with characteristic vector X. Conditioned on Y, Bob chooses a random stream $\mathbf{a}_Y$ of length $n/2$ with characteristic vector y. Alice runs algorithm M on $\mathbf{a}_X$ and transmits the state of M to Bob, who continues running M on $\mathbf{a}_y$.

Observe that $\mathbf{a}_X \circ \mathbf{a}_Y$ is a uniformly random stream of length n, with items independently distributed according to $\mathcal{D}$. This follows from the fact that $\mathbf{a}_X$ and $\mathbf{a}_Y$ are independent, and they were drawn uniformly at random. Indeed, X (resp. Y) is a random characteristic vector, and $\mathbf{a}_X$ (resp. $\mathbf{a}_Y$) is random conditioned on having characteristic vector X (resp. Y).

By Lemma 7 and that

$$\Pr[d/4 \leq wt(X), wt(Y) \leq 3d/4] = 1 - o(1),$$

there is a constant $\alpha > 0$ for which with probability at least $1 - \delta/2$, we have

$$|\Delta(x,y) - \mathbf{E}[\Delta(x,y)]| \geq \alpha\sqrt{d}.$$

We condition on this event, denoted $\mathcal{E}$. Put $\epsilon' = \alpha\epsilon/2$.

Let $\tilde{F}_0$ be the output of M. The claim is that $\tilde{F}_0$ along with $wt(X)$ and $wt(Y)$ can be used to decide HAM_d. We first decompose F_0. We have,

$$F_0(\mathbf{a}_X \circ \mathbf{a}_Y)$$

$$= wt(X \wedge Y) + \Delta(X,Y) = \frac{1}{2}\left(wt(X) + wt(Y) + \Delta(X,Y)\right),$$

and so

$$\Delta(X,Y) = 2F_0(\mathbf{a}_X \circ \mathbf{a}_Y) - (wt(X) + wt(Y)).$$

Define the quantity E to be

$$E = 2M(\mathbf{a}_X \circ \mathbf{a}_Y) - (wt(X) + wt(Y)).$$

[3] See also `http://www.dartmouth.edu/~chance/teaching_aids/books_articles/probability_book/pinsky-hypergeometric.pdf`

Let

$$\tau = wt(X) + wt(Y) - wt(X)wt(Y)/d,$$

which Bob can compute. If $E > \tau$, Bob outputs 1, otherwise Bob outputs 0.

For correctness, suppose M outputs a $(1 \pm \epsilon')$ approximation to $F_0(\mathbf{a}_X \circ \mathbf{a}_Y)$. Conditioned on $\mathcal{E}$, and using the fact that $\mathbf{E}[\Delta(X,Y)] = \tau$, we have two cases.

Case 1: Suppose

$$\Delta(X,Y) > \tau + \alpha\sqrt{d}.$$

Using that $d = 1/\epsilon^2$, we have

$$\begin{aligned} E &\geq 2(1-\epsilon')F_0 - wt(X) - wt(Y) = \Delta(X,Y) - 2\epsilon' F_0 \\ &\geq \Delta(X,Y) - 2\epsilon' d \\ &\geq \Delta(X,Y) - \alpha\sqrt{d} \\ &> \tau + \alpha\sqrt{d} - \alpha\sqrt{d} \\ &= \tau. \end{aligned}$$

Case 2: Suppose

$$\Delta(X,Y) < \tau - \alpha\sqrt{d}.$$

Then,

$$\begin{aligned} E &\leq 2(1+\epsilon')F_0 - wt(X) - wt(Y) \\ &\leq \Delta(X,Y) + 2\epsilon' F_0 \\ &\leq \Delta(X,Y) + 2\epsilon' d \\ &\leq \Delta(X,Y) + \alpha\sqrt{d} \\ &< \tau - \alpha\sqrt{d} + \alpha\sqrt{d} \\ &= \tau. \end{aligned}$$

Since M outputs a $(1 \pm \epsilon')$ approximation to $F_0(\mathbf{a}_X \circ \mathbf{a}_Y)$ with probability at least $1 - \delta/2$, and

$$|\Delta(X,Y) - \tau| \geq \alpha\sqrt{d}$$

with probability at least $1-\delta/2$, the parties can solve HAM_d with probability at least $1-\delta$ and communication $\ell S_{\mathcal{D},\epsilon',\ell}(F_0)$. On the other hand, the communication must be at least $D_{\mu,\delta}(HAM_d)$. If M is 1-pass, then the protocol is 1-way. This concludes the proof. □

By Theorem 9, to lower bound the space complexity of approximating F_0 in the random-data model in one pass, it suffices to give a lower bound on the distributional complexity $D^{1-way}_{\mu,\delta}(HAM_d)$, where $\mu = \mu_d \times \mu_d$, and μ_d is a distribution on $\{0,1\}^d$ of characteristic vectors of $\mathcal{D}^n$. As mentioned above, it is well-known that $wt(X)$ is tightly concentrated around its expectation since it corresponds to the number of non-empty bins when throwing n balls into d bins. In particular,

$$\Pr[d/4 \leq wt(X) \leq 3d/4] = 1 - o(1).$$

By a union bound,

$$\Pr[d/4 \leq wt(X), wt(Y) \leq 3d/4] = 1 - o(1).$$

To simplify the analysis, it would be nice if we could assume that there exist $k, \ell \in [d/4, 3d/4]$ for which $wt(X) = k$ and $wt(Y) = \ell$. This will make X uniform over strings of weight k, and Y uniform over strings of weight ℓ. Let

$$\rho_{k,\ell} = \mu_t \mid (wt(X) = k) \times \mu_t \mid (wt(Y) = \ell).$$

The next lemma uses that Alice is given $wt(Y)$ and Bob is given $wt(X)$. Its proof appears in the appendix.

LEMMA 10. *There are $k, \ell \in [d/4, 3d/4]$ for which*

$$D_{\mu,\delta}(HAM_d) \geq D_{\rho_{k,\ell},\delta/2}(HAM_d).$$

This also holds for 1-way protocols.

In the remainder we fix $wt(X) = k$ and $wt(Y) = \ell$ for k and ℓ satisfying the premise of Lemma 10. To lower bound $D^{1-way}_{\mu,\delta}(HAM_d)$, it thus suffices to lower bound the value $D^{1-way}_{\rho_{k,\ell},\delta/2}(HAM_d)$. For notational convenience, put $\rho = \rho_{k,\ell}$. Note that for such a distribution, we do not need to give Alice $wt(Y)$ or Bob $wt(X)$, since we can assume the protocol has these values hardwired.

Fix a 1-round protocol Π realizing $D^{1-way}_{\rho,\delta/2}(HAM_d)$. Let $g : \{0,1\}^d \times \{0,1\}^d \to \{0,1\}$ be such that $g(x,y) = 1$ iff $HAM_d(x,y) = 1$. We assume $z \stackrel{\text{def}}{=} D^{1-way}_{\rho,\delta/2}(g) = o(d)$, and derive a contradiction. Let M be the single message sent from Alice to Bob in Π. Let A be the (deterministic) algorithm run by Bob on M and Y. We have

$$\Pr_{(X,Y)\sim\rho}[A(M,Y) = g(X,Y)] \geq 1 - \delta/2.$$

We need Fano's inequality:

FACT 11. *([14]) For $R, S \in \{0,1\}$ and a function h,*

$$H(\Pr[h(R) \neq S]) \geq H(S \mid R),$$

where for $x \in [0,1]$, $H(x) = x \log \frac{1}{x} + (1-x)\log\frac{1}{1-x}$ is the binary entropy function. Here, $H(0) = H(1) = 0$.

Applying this with $h = A$, $R = (M,Y)$, and $S = g(X,Y)$, we have

$$H(g(X,Y) \mid M,Y) \leq H(\delta/2).$$

We now lower bound $H(g(X,Y) \mid M,Y)$ as a positive constant independent of δ, which will show a contradiction for small enough δ.

For any $r \in \{0,1\}^d$, let S_r be the set of $x \in \{0,1\}^d$ for which $M = r$. Then

$$\mathbf{E}[|S_M|] = \binom{d}{k}/2^z.$$

By a Markov argument,

$$\Pr[|S_M| \geq \binom{d}{k}/2^{z+1}] \geq \frac{1}{2}.$$

Let us condition on the event

$$\mathcal{E} : |S_M| \geq \binom{d}{k}/2^{z+1}.$$

By concavity of the entropy,

$$H(g(X,Y) \mid M,Y)$$

$$\geq H(g(X,Y) \mid M,Y,\mathcal{E}) \Pr[\mathcal{E}] \geq H(g(X,Y) \mid M,Y,\mathcal{E})/2.$$

Now let S be any set of weight$-k$ vectors for which

$$|S| \geq \binom{d}{k}/2^{z+1}.$$

The number of $y \in \{0,1\}^d$ of weight ℓ for which $g(x,y) = 1$ for any given x of weight k is independent of the particular x. Let q denote this quantity, where $1 \leq q \leq \binom{d}{\ell}$. By the second part of Lemma 7,

$$q/\binom{d}{\ell} = 1/2 \pm o(1).$$

For y of weight ℓ, let $V_y = \Pr_{x \in S}[g(x,y) = 1]$. By averaging,

$$\mathbf{E}_y[V_y] = q/\binom{d}{\ell}.$$

For $u \in S$, let $C_u = 1$ if $g(u,y) = 1$, and $C_u = 0$ otherwise. Then

$$V_y = \frac{1}{|S|}\sum_{u \in S} C_u.$$

We use the second-moment method (see, e.g., [3] for an introduction to this technique). Consider

$$\mathbf{Var}_y[V_y] = \frac{1}{|S|^2}\left[\sum_{u,v \in S} \mathbf{E}[C_u C_v] - \mathbf{E}[C_u]\mathbf{E}[C_v]\right].$$

Then

$$\mathbf{E}[C_u] = q/\binom{d}{\ell}$$

for all $u \in S$. Moreover,

$$\mathbf{E}[C_u^2] = \mathbf{E}[C_u] = q/\binom{d}{\ell}.$$

Thus,

$$\begin{aligned}
&\mathbf{Var}_y[V_y] \\
= \;& \frac{1}{|S|^2}\left[\frac{q|S|}{\binom{d}{\ell}}\left[1 - \frac{q}{\binom{d}{\ell}}\right] + \sum_{u \neq v}\left[\mathbf{E}[C_u C_v] - \frac{q^2}{\binom{d}{\ell}^2}\right]\right] \\
= \;& o(1) + \frac{1}{|S|^2}\sum_{u \neq v}\left[\mathbf{E}[C_u C_v] - \frac{q^2}{\binom{d}{\ell}^2}\right].
\end{aligned}$$

The difficulty is in bounding $\mathbf{E}[C_u C_v]$. Now we use the fact that $|S|$ is large.

FACT 12. *([32]) Let $0 < c < 1/2$ be a constant. For any $u \in \{0,1\}^d$, the number of $v \in \{0,1\}^d$ for which $\Delta(u,v) < cd$ or $\Delta(u,v) > (1-c)d$ is at most*

$$2 \cdot 2^{H(c)d},$$

where H is the binary entropy function.

FACT 13. *([32]) Let $0 < c < 1/2$ be a constant. Then*

$$\binom{d}{cd} \geq 2^{dH(c) - o(d)}.$$

Now,

$$|S| \geq \binom{d}{k}/2^{z+1},$$

and so using that $k \in [d/4, 3d/4]$ and $z = o(d)$, by Fact 13 we have $|S| \geq 2^{dH(1/4) - o(d)}$. Now using Fact 12, it follows that of the $\binom{|S|}{2}$ pairs $u, v \in S$ with $u \neq v$, all but $2|S|2^{dH(1/5)}$ of them have Hamming distance at least $d/5$ and at most $4d/5$. Thus, at least an $\alpha \geq 1/2$ of the pairs have this property. Using that

$$q/\binom{d}{\ell} = 1/2 \pm o(1)$$

in the final inequality,

$$\begin{aligned}
\mathbf{Var}_y[V_y] &= o(1) + \frac{1}{|S|^2}\sum_{u \neq v}\left[\mathbf{E}[C_u C_v] - \frac{q^2}{\binom{d}{\ell}^2}\right] \\
&= o(1) + \\
&\quad \frac{1}{|S|^2}\sum_{\Delta(u,v) \leq d/5 \text{ or } \Delta(u,v) \geq 4d/5}\left[\mathbf{E}[C_u C_v] - \frac{q^2}{\binom{d}{\ell}^2}\right] \\
&\quad + \frac{1}{|S|^2}\sum_{d/5 < \Delta(u,v) < 4d/5}\left[\mathbf{E}[C_u C_v] - \frac{q^2}{\binom{d}{\ell}^2}\right] \\
&\leq o(1) + \\
&\quad \frac{(1-\alpha)q}{\binom{d}{\ell}}\left[1 - \frac{q}{\binom{d}{\ell}}\right] + \\
&\quad \sum_{d/5 < \Delta(u,v) < 4d/5}\left[\frac{\mathbf{E}[C_u C_v]}{|S|^2} - \frac{q^2}{|S|^2\binom{d}{\ell}^2}\right] \\
&\leq o(1) + \\
&\quad \frac{1-\alpha}{4} + \\
&\quad \alpha \cdot \max_{(u,v) \;|\; d/5 < \Delta(u,v) < 4d/5}\left[\mathbf{E}[C_u C_v] - \frac{1}{4}\right].
\end{aligned}$$

Our goal is to show that $\mathbf{Var}_y[V_y]$ is a constant strictly less than $1/4$. Now,

$$\begin{aligned}
\mathbf{E}[C_u C_v] &= \Pr_y[g(v,y) = 1 \mid g(u,y) = 1]\frac{q}{\binom{d}{\ell}} \\
&= \frac{\Pr_y[g(v,y) = 1 \mid g(u,y) = 1]}{2} \pm o(1).
\end{aligned}$$

By the above expressions, to show that $\mathbf{Var}_y[V_y]$ is at most a constant strictly less than $1/4$ it suffices to show that there exists a constant $\beta > 0$ for which

$$\max_{(u,v) \;|\; d/5 < \Delta(u,v) < 4d/5} \Pr_y[g(v,y) = 1 \mid g(u,y) = 1] < 1 - \beta.$$

Fix any u, v for which

$$d/5 < \Delta(u,v) < 4d/5.$$

By relabeling coordinates, we may assume that $u = 1^k 0^{d-k}$, and that

$$v = 1^{k'} 0^{k-k'} 1^{k-k'} 0^{d-2k+k'}$$

for some k'. Notice that $\Delta(u,v) = 2(k-k')$, so we know that

$$k - k' \in [d/10, 2d/5].$$

Consider a random weight-ℓ vector y for which $g(u,y) = 1$. By definition, this means that

$$\Delta(u,y) > k + \ell - 2k\ell/d.$$

We just need to show that with probability $\Omega(1)$, we have $g(v,y) = 0$, that is,

$$\Delta(v,y) < k + \ell - 2k\ell/d.$$

It will be more convenient to argue in terms of sets. Let S be the set of coordinates in $[k]$ for which y is 1, and let T be the set of coordinates in $\{k+1,\ldots,d\}$ for which y is 1. Then

$$\Delta(u,y) = k - |S| + \ell - |S| = k + \ell - 2|S|,$$

which we know is greater than $k + \ell - 2k\ell/d$. Thus,

$$|S| < k\ell/d.$$

Now, $|S|$ is distributed as a conditional distribution of a hypergeometric distribution. It follows as in the proof of Lemma 7 (using Theorem 8), that with arbitrarily large constant probability, we have

$$|S| = k\ell/d - \Omega(\sqrt{d})$$

(the probability tends to 1 as the constant in the $\Omega(\cdot)$ notation tends to ∞). We condition on this event.

Conditioned on $|S| = i$ for any

$$i \in [k\ell/d - \Omega(\sqrt{d}), k\ell/d),$$

S is a random subset of size i contained in $[k]$. Moreover, T is a random subset of size $\ell - i$ contained in $\{k+1,\ldots,d\}$. Letting

$$C = \{k+1,\ldots,2k-k'\},$$

then

$$\Delta(v,y) = k' - |S\cap[k']| + |S| - |S\cap[k']| + k - k' + (\ell - i) - 2|T\cap C|,$$

which equals

$$k + \ell - 2|[k']\cap S| - 2|T\cap C|.$$

Now, $|[k']\cap S|$ is hypergeometrically distributed with mean ik'/k and $|T\cap C|$ is hypergeometrically distributed with mean $(\ell-i)(k-k')/(d-k)$. Write $|[k']\cap S|$ as $ik'/k + \gamma_1$, and write $|T\cap C|$ as $(\ell-i)(k-k')/(d-k) + \gamma_2$. Put

$$i = k\ell/d - \Omega(\sqrt{d}).$$

By direct substitution,

$$\Delta(v,y) = k + \ell - \frac{2k\ell}{d} + \Theta(\sqrt{d}) - 2\gamma_1 - 2\gamma_2.$$

Thus, to show that

$$\Delta(v,y) < k + \ell - 2k\ell/d$$

with constant probability, it suffices to show that for any constant $c > 0$, we have

$$c\sqrt{d} < \gamma_1 + \gamma_2$$

with constant probability. Now, as in the proof of Lemma 7, we have

$$\Pr[\gamma_1 \geq c\sqrt{d}] = \Omega(1)$$

for any constant $c > 0$. Moreover, $\gamma_2 > 0$ with probability at least $1/2 - o(1)$. Note that conditioned on $|S| = i$ for any $i \in [k\ell/d - \Omega(\sqrt{d}), k\ell/d)$, γ_1 and γ_2 are independent. Thus, for any value of i in this range, with probability $\Omega(1)$, $c\sqrt{d} < \gamma_1 + \gamma_2$. It follows that conditioned on

$$|S| \in [k\ell/d - \Omega(\sqrt{d}), k\ell/d),$$

we have $c\sqrt{d} < \gamma_1 + \gamma_2$ with constant probability. Since $|S|$ is in this range with arbitrarily large constant probability, we have $\Pr[g(v,y) = 0] = \Omega(1)$.

Hence, we have that $\mathbf{Var}_y[V_y] = \zeta$ for a constant ζ strictly less than $1/4$. Define the constant

$$\zeta' = \sqrt{\frac{\zeta}{2} + \frac{1}{8}},$$

and note that $\zeta' < 1/2$. It follows by Chebyshev's inequality that,

$$\Pr_y[|V_y - 1/2| > \zeta']$$

$$\leq \Pr_y\left[|V_y - \mathbf{E}[V_y]| > \zeta' + o(1)\right] < \frac{\zeta}{(\zeta')^2} + o(1).$$

This is a constant less than 1, so for an $\Omega(1)$ fraction of y,

$$|V_y - 1/2| \leq \zeta'.$$

Consider the event

$$\mathcal{F} : |V_Y - 1/2| \leq \zeta'.$$

Since X and Y are independent, the above analysis implies that

$$\Pr_{X,Y}[\mathcal{F} \mid \mathcal{E}] = \Omega(1).$$

Thus,

$$H(g(X,Y) \mid M, Y, \mathcal{E}) = \Omega(H(g(X,Y) \mid M, Y, \mathcal{E}, \mathcal{F}).$$

But, by definition of V_Y, if $\mathcal{E}\cap\mathcal{F}$ occurs, then

$$1/2 - \zeta' \leq \Pr_X[g(X,Y) = 1] \leq 1/2 + \zeta'.$$

Thus,

$$H(g(X,Y) \mid M, Y, \mathcal{E}, \mathcal{F}) = \Omega(1),$$

where the constant is independent of δ. It follows that

$$H(g(X,Y) \mid M, Y) = \Omega(1).$$

But we have shown that

$$H(g(X,Y) \mid M, Y) \leq H(\delta/2).$$

This is a contradiction for small enough constant δ. So our assumption that $z = o(d)$ was false. We conclude,

THEOREM 14. *$D^{1-way}_{\mu,\delta}(HAM_d) = \Omega(d)$. Hence, for a constant $\delta > 0$, when $n, d = \Theta(1/\varepsilon^2)$, the space complexity of any 1-pass algorithm in the random-data model which ϵ-approximates F_0 with probability at least $1-\delta$ is $\Omega(1/\epsilon^2)$.*

5. CONCLUSION

We introduced the random data model, in which each of n successive stream items is drawn independently and uniformly at random from an unknown set of size d, for an unknown value of d. For a wide range of values of d and n we gave a 1-pass time-optimal algorithm that beats the $\Omega(1/\varepsilon^2)$ space lower bound that holds in the adversarial and random-order models. Nevertheless, for certain values of d and n, we showed that an $\Omega(1/\varepsilon^2)$ space lower bound holds even in this model, subsuming previous lower bounds (since our model is strictly contained in existing models), and showing that even for natural choices of data the problem is hard.

In the future, it would be useful to understand whether our 1-pass algorithm is space-optimal, whether there are other real-world distributions not easily reducible to the ones studied here, and whether multiple passes over the data can help in this model.

Acknowledgments: The author thanks Peter Haas, T.S. Jayram, Phokion Kolaitis, Jelani Nelson, and the anonymous reviewers for many helpful comments.

6. REFERENCES

[1] A. Akella, A. Bharambe, M. Reiter, and S. Seshan. Detecting DDoS attacks on ISP networks. In *ACM SIGMOD/PODS Workshop on Management and Processing of Data Streams (MPDS) FCRC*, 2003.

[2] N. Alon, Y. Matias, and M. Szegedy. The space complexity of approximating the frequency moments. *Journal of Computer and Systems Sciences*, 58(1):137–147, 1999.

[3] N. Alon and J. Spencer. *The Probabilistic Method*. John Wiley, 1992.

[4] Z. Bar-Yossef. *The Complexity of Massive Data Set Computations*. PhD thesis, U.C. Berkeley, 2002.

[5] Z. Bar-Yossef, T. S. Jayram, R. Kumar, D. Sivakumar, and L. Trevisan. Counting distinct elements in a data stream. In *RANDOM*, pages 1–10, 2002.

[6] T. Batu, S. Dasgupta, R. Kumar, and R. Rubinfeld. The complexity of approximating the entropy. *SIAM J. Comput.*, 35(1):132–150, 2005.

[7] J. Bunge. Bibliography on estimating the number of classes in a population. *Manuscript*, 2007.

[8] Q. L. Burrell and M. R. Fenton. Yes, the GIGP really does work - and is workable! *JASIS*, 44(2):61–69, 1993.

[9] A. Chakrabarti, G. Cormode, and A. McGregor. Robust lower bounds for communication and stream computation. In *STOC*, pages 641–650, 2008.

[10] A. Chakrabarti, T. S. Jayram, and M. Pătraşcu. Tight lower bounds for selection in randomly ordered streams. In *SODA*, pages 720–729, 2008.

[11] M. Charikar, S. Chaudhuri, R. Motwani, and V. R. Narasayya. Towards estimation error guarantees for distinct values. In *PODS*, pages 268–279, 2000.

[12] M. Charikar, K. Chen, and M. Farach-Colton. Finding frequent items in data streams. In *ICALP*, pages 693–703, 2002.

[13] G. Cormode and S. Muthukrishnan. Summarizing and mining skewed data streams. In *SDM*, 2005.

[14] T. M. Cover and J. A. Thomas. *Elements of information theory*. Wiley-Interscience, New York, NY, USA, 1991.

[15] W. Feller. *An Introduction to Probability Theory and its Applications*, volume 1. John Wiley and Sons, 3 edition, 1968.

[16] P. Flajolet and G. N. Martin. Probabilistic counting algorithms for data base applications. *Journal of Computer and System Sciences*, 31:182–209, 1985.

[17] P. B. Gibbons and Y. Matias. Synopsis data structures for massive data sets. In *SODA*, pages 909–910, 1999.

[18] P. B. Gibbons, Y. Matias, and V. Poosala. Fast incremental maintenance of approximate histograms. *ACM Trans. Database Syst.*, 27(3):261–298, 2002.

[19] S. Guha and A. McGregor. Approximate quantiles and the order of the stream. In *PODS*, pages 273–279, 2006.

[20] S. Guha and A. McGregor. Lower bounds for quantile estimation in random-order and multi-pass streaming. In *ICALP*, pages 704–715, 2007.

[21] S. Guha and A. Mcgregor. Space-efficient sampling. In *AISTATS*, pages 169–176, 2007.

[22] T. S. Jayram, R. Kumar, and D. Sivakumar. The one-way communication complexity of gap hamming distance. *Manuscript*, 2007.

[23] A. Kamath, R. Motwani, K. V. Palem, and P. G. Spirakis. Tail bounds for occupancy and the satisfiability threshold conjecture. *Random Structures and Algorithms*, 7(1):59–80, 1995.

[24] R. Kumar. Story of distinct elements. *IITK Workshop on Algorithms for Data Streams*, 2006.

[25] R. Kumar and R. Panigrahy. On finding frequent elements in a data stream. In *APPROX-RANDOM*, pages 584–595, 2007.

[26] E. Kushilevitz and N. Nisan. *Communication Complexity*. Cambridge University Press, 1997.

[27] M. Mitzenmacher and E. Upfal. *Probability and Computing: Randomized Algorithms and Probabilistic Analysis*. Cambridge University Press, 2005.

[28] R. Motwani and P. Raghavan. *Randomized Algorithms*. Cambridge University Press, 1995.

[29] R. Motwani and S. Vassilvitskii. Distinct value estimators in power law distributions. In *ANALCO*, 2006.

[30] S. Muthukrishnan. Data streams: algorithms and applications. *Foundations and Trends in Theoretical Computer Science*, 1(2), 2003.

[31] S. Raskhodnikova, D. Ron, A. Shpilka, and A. Smith. Strong lower bounds for approximating distribution support size and the distinct elements problem. In *FOCS*, pages 559–569, 2007.

[32] J. H. van Lint. *An Introduction to Coding Theory*. New York: Springer-Verlag, 1992.

[33] D. Woodruff. Optimal space lower bounds for all frequency moments. In *SODA*, pages 167–175, 2004.

[34] D. Woodruff. *Efficient and Private Distance Approximation in the Communication and Streaming Models*. PhD thesis, MIT, 2007.

APPENDIX

A. MISSING PROOFS

Proof of Lemma 4: The probability that an item i occurs in the stream is $1-(1-1/d)^n$, and so the expected value of F_0 is $d(1-(1-1/d)^n)$. Define the *defect* ζ to be $d-F_0$, which has expected value $\kappa = d(1-1/d)^n$. Since $d \leq Wn$, $\kappa \leq de^{-n/d} \leq d/e^{1/W}$. By Theorem 2 of [23], for any $\theta > 0$,

$$\Pr[|\zeta - \kappa| \geq \theta\kappa] \leq 2e^{-\frac{\theta^2\kappa^2(d-1/2)}{d^2-\kappa^2}} \leq e^{-\frac{c\theta^2\kappa^2}{d}},$$

where $c > 0$ is a constant. Choose θ so that $\theta\kappa = \epsilon' d$ for a value $\epsilon' = \Theta(\epsilon)$ that will be determined by the analysis. Then, $\Pr[|\zeta - \kappa| \geq \epsilon' d] \leq e^{-c(\epsilon')^2 d}$. Using that c is a constant, and the assumption that $d \geq \nu/\varepsilon^2$ for a sufficiently large constant ν, this probability can be made to

be at most $1/100$ for any choice of ε'. Conditioned on $|\zeta - \kappa| \leq \theta\varepsilon' d$, we have $|d\,[1-(1-1/d)^n] - F_0| \leq \varepsilon' d$. Now, $\mathbf{E}[F_0] = d - \mathbf{E}[\zeta] = d - \kappa \geq (1 - 1/e^{1/W})d$. So for small enough $\varepsilon' = \Theta(\varepsilon)$, we have $|d\,[1-(1-1/d)^n] - F_0| \leq \varepsilon\mathbf{E}[F_0]/3$. Consider the quantity $\tilde{F}_0 - d[1-(1-1/d)^n]$, which equals $(d'-d) + d(1-1/d)^n - d'(1-1/d')^n$. The function $x(1-1/x)^n$ is increasing in x, and since $d' \geq d$, we have $\tilde{F}_0 - d[1-(1-1/d)^n] \leq \varepsilon' d$. On the other hand, $\tilde{F}_0 - d[1-(1-1/d)^n]$ is at least $d(1-1/d)^n - d'(1-1/d')^n$. Differentiating with respect to n, we find that it is minimized when

$$d(\ln(1-1/d))(1-1/d)^n - d'(\ln(1-1/d'))(1-1/d')^n = 0.$$

Substituting back into the expression, we find that its minimum value is $d(1-1/d)^n\left[1 - \frac{\ln(1-1/d)}{\ln(1-1/d')}\right]$. Using that $d \geq \nu/\varepsilon^2$, and thus $1/d < 1$, a Taylor series expansion gives us $\ln(1-1/d) = -1/d^2 - \Theta(1/d^4)$, as well as $\ln(1-1/d') = -1/(d')^2 - \Theta(1/d^4)$, where we have used that $d' = \Theta(d)$. Substituting these bounds into the above minimum, we obtain that the minimum is $d(1-1/d)^n\Theta(\varepsilon')$. Finally, using the fact that $d \leq W \cdot n$, this is bounded as $-O(\varepsilon' d)$. Thus, for small enough $\varepsilon' = \Theta(\varepsilon)$, using that $\mathbf{E}[F_0] \geq (1 - 1/e^{1/W})d$, we have that $|\tilde{F}_0 - d[1-(1-1/d)^n]| \leq \varepsilon\mathbf{E}[F_0]/3$. It follows by the triangle inequality that $|\tilde{F}_0 - F_0| \leq \varepsilon\mathbf{E}[F_0]$. □

Proof of Claim 6: For the lower bound,

$$\begin{aligned} 1-(1-x)^y &\geq 1 - e^{-xy} \\ &= 1 - \left(1 - xy + \frac{(xy)^2}{2} - \cdots\right) \\ &= xy - \frac{(xy)^2}{2} + \cdots \\ &\geq xy - \frac{(xy)^2}{2}. \end{aligned}$$

The first inequality follows from the fact that $1 + z \leq e^z$ for all $z \in \mathbf{R}$, see, e.g., [28]. The second inequality follows via the Taylor expansion for e^{-xy}. The equality and final inequality are straightforward.

For the upper bound, we first show $(1-x)^y \geq 1 - xy$. By monotonicity of the $\ln(\cdot)$ function, $(1-x)^y \geq 1 - xy$ iff $\ln(1-x)^y \geq \ln(1-xy)$. We use the Taylor expansion for $\ln(1+x)$, that is, for $|x| < 1$ we have the expansion $\ln(1+x) = \sum_{n=0}^{\infty} \frac{(-1)^n}{n+1} x^{n+1}$. Then,

$$\ln(1-x)^y = -y\sum_{i=1}^{\infty} \frac{x^i}{i+1}.$$

Also,

$$\ln(1-xy) = -\sum_{i=1}^{\infty} \frac{(xy)^i}{i+1}.$$

We will have $\ln(1-x)^y \geq \ln(1-xy)$ if for all $i \geq 1$,

$$-y\frac{x^i}{i+1} \geq -\frac{(xy)^i}{i+1}.$$

This holds provided $y^{i-1} \geq 1$, which holds for $y \geq 1$, as given by the premise of the claim. Thus,

$$1-(1-x)^y \leq 1-(1-xy) \leq xy,$$

completing the proof. □

Proof of Lemma 10: Suppose not, and for each $k, \ell \in [d/4, 3d/4]$, let $\Pi_{k,\ell}$ be a protocol realizing $D_{\rho_{k,\ell},\delta/2}(HAM_d)$. Given $X \sim \mu_d$, with probability $1 - o(1)$, $wt(X), wt(Y) \in [d/4, 3d/4]$. Note that Alice is given $wt(Y)$ and can also deduce $wt(X)$. She then runs $\Pi_{wt(X),wt(Y)}$. Moreover, Bob is given $wt(X)$ and can deduce $wt(Y)$, so he also runs the protocol $\Pi_{wt(X),wt(Y)}$. The error of the protocol is at most $\delta/2 + o(1) \leq \delta$, and the communication is upper bounded by $D_{\rho_{wt(X),wt(Y),\delta/2}}(HAM_d)$. This is a contradiction to the communication having to be at least $D_{\mu,\delta}(HAM_d)$. □