Tight Lower Bounds for the Distinct Elements Problem

Piotr Indyk*
MIT
indyk@theory.lcs.mit.edu

David Woodruff†
MIT
dpwood@mit.edu

Abstract

We prove strong lower bounds for the space complexity of (ϵ, δ)-approximating the number of distinct elements F_0 in a data stream. Let m be the size of the universe from which the stream elements are drawn. We show that any one-pass streaming algorithm for (ϵ, δ)-approximating F_0 must use $\Omega\left(\frac{1}{\epsilon^2}\right)$ space when $\epsilon = \Omega\left(m^{-\frac{1}{9+k}}\right)$, for any $k > 0$, improving upon the known lower bound of $\Omega\left(\frac{1}{\epsilon}\right)$ for this range of ϵ. This lower bound is tight up to a factor of $(\log\log m + \log(1/\epsilon))$. Our lower bound is derived from a reduction from the one-way communication complexity of approximating a boolean function in Euclidean space. The reduction makes use of a low-distortion embedding from an l_2 to an l_1 norm.

1 Introduction

Let **a**$= a_1, \ldots, a_n$ be a sequence of elements, which we will refer to as a *stream*, from a universe of size m, which we denote by $[m] = \{0, \ldots, m-1\}$. In this paper we examine the space complexity of algorithms that count the number of distinct elements $F_0 = F_0(\mathbf{a})$ in **a**. All algorithms will be given one-pass over the elements of **a**, which are arranged in adversarial order. An algorithm A is said to (ϵ, δ) approximate F_0 on stream **a** if A outputs a number $\tilde{F_0}$ such that $\mathbf{Pr}[|\tilde{F_0} - F_0| > \epsilon F_0] < \delta$. Since there is a provable deterministic space lower bound of $\Omega(m)$ for computing or even approximating F_0 within a multiplicative factor of $(1 \pm \epsilon)$ [1], there has been considerable effort to devise randomized approximation algorithms.

There are several practical motivations for designing space-efficient algorithms for approximating F_0. Computing the number of distinct elements is very valuable to the database community. Query optimizers can use F_0 to find the number of unique values in a database with a certain attribute without having to perform an expensive sort on the values. With commercial databases approaching the size of 100 terabytes, it is infeasible to make multiple passes over the data since the sheer amount of time spent looking at the data is prohibitive.

Other applications include networking: internet routers that only get one quick view at incoming packets can gather the number of distinct destination addresses passing through them with only limited memory. For an application of F_0 estimation algorithms to detecting Denial of Service attacks, see [2].

There have been a multitude (e.g., [7, 1, 4]) of algorithms proposed for computing F_0 in a data stream, beginning with the work of Flajolet and Martin [7]. The best known algorithm was presented in [4] and (ϵ, δ) approximates F_0 in space $O(((\frac{1}{\epsilon})^2 \log\log m + \log m \log \frac{1}{\epsilon}) \log \frac{1}{\delta})$. In this paper we will take δ to be constant.

For reasonable values of m and ϵ (e.g., $m = 2^{32}$, $\epsilon = 10\%$), the storage bound of the algorithms is dominated by the $1/\epsilon^2$ term. If one wants even better approximation quality (e.g., $\epsilon = 1\%$), the quadratic dependence on $1/\epsilon$ constitutes a severe drawback of the existing algorithms. This raises the question (posed in [4]) if it is possible to reduce the dependence on $1/\epsilon$. Till now, the best known lower bound for the problem was $\Omega(\log m + 1/\epsilon)$ [1, 3].

In this paper we answer this question in the negative. In particular, we show that any algorithm for approximating F_0 up to a factor of $(1 + \epsilon)$ requires $\Omega(1/\epsilon^2)$ storage, as long as $1/\epsilon = o(m^\alpha)$ for certain $\alpha > 0$. This matches (up to a factor of $\log\log m$ for small ϵ, and $\log(1/\epsilon)$ for large ϵ) the upper bound of [4].

1.1 Overview of techniques and technical results

One way of establishing a lower bound is to lower bound the communication complexity of computing certain boolean functions and reduce them to that of computing F_0 [3]. In this model there are two parties, Alice and Bob, who have inputs x and y respectively and wish to compute the value of a boolean function $f(x, y)$ with probability at

*This work was supported in part by NSF ITR grant CCR-0220280 and Sloan Research Fellowship.
†Supported by Akamai Presidential Fellowship for the duration of this work.

least $1-\delta$. The idea is that Alice can run a distinct elements algorithm A on x and transmit the state S of A to Bob. Then Bob can plug S into his copy of A and continue the computation on y. This computation will return a number $\tilde{F_0}$ which is a $(1 \pm \epsilon)$ approximation to $F_0(x \circ y)$ with probability at least $1-\delta$. If $\tilde{F_0}$ can be used to determine $f(x,y)$ with error probability less than δ, then the communication cost is just the space used by A, which must be at least the one-way communication complexity of computing $f(x,y)$. In [1] the authors reduce the communication complexity of computing equality $EQ(x,y)$ to computing $F_0(x \circ y)$. Since the randomized communication complexity of $EQ(x,y)$ is $\Theta(\log m)$, this established an $\Omega(\log m)$ lower bound for computing F_0.

We would like to obtain lower bounds in terms of the approximation error ϵ. In [3] the one-way communication complexity of the ϵ-set disjointness problem is used to derive an $\Omega(\frac{1}{\epsilon})$ space lower bound for approximating F_0. Here Alice is given a set $x \subseteq [m]$ with $|x| = \frac{m}{2}$ and Bob is given a set $y \subseteq [m]$ with $|y| = \epsilon * m$. Furthermore, both parties are given the promise that either $y \subseteq x$ or $y \cap x = \emptyset$. In case of the former, $F_0(x \circ y) = \frac{m}{2}$ and in the latter $F_0(x \circ y) = \frac{m}{2} + \epsilon * m$. Hence, an algorithm which (ϵ, δ) approximates F_0 can distinguish these two cases. However, this reduction is particularly weak since Alice can replace the F_0 approximation algorithm with an algorithm which simply samples $O(\frac{1}{\epsilon})$ of the elements s_i of $[m]$, checks if $s_i \in x$, and sends the result to Bob. This motivates the search for promise problems which use the full power of a distinct elements algorithm as a subroutine.

Given a stream $\mathbf{a}$, its characteristic vector $v_{\mathbf{a}}$ is the m-dimensional vector with ith coordinate 1 iff element i of $[m]$ appears in $\mathbf{a}$. Note that $F_0(\mathbf{a})$ is just the Hamming weight of $v_{\mathbf{a}}$. One natural boolean function to consider is the following: Alice and Bob are given $x, y \in \{0,1\}^m$ with the promise that either $\Delta(x,y) \leq \frac{m}{2} - \epsilon m$, in which case $f(x,y) = 0$, or $\Delta(x,y) \geq \frac{m}{2}$, in which case $f(x,y) = 1$. Here $\Delta(x,y)$ denotes the Hamming distance between x and y, that is, the number of bit positions in which x and y differ. Alice and Bob view their inputs x and y as characteristic vectors of certain streams $\mathbf{a}_x$ and $\mathbf{a}_y$. The value $F_0(\mathbf{a}_x \circ \mathbf{a}_y)$ is then just the Hamming weight of $x \vee y$, the bitwise OR of x and y. By constraining the weights of the inputs x and y to be close to each other and less than $\frac{m}{2} - \epsilon m$, one can use an (ϵ, δ) F_0-approximation algorithm to determine the value of $f(x,y)$ with error probability at most δ. Unfortunately, it is rather difficult to lower bound the one-way communication complexity of f directly.

Fix ϵ and set $t = t(\epsilon) = \frac{1}{\epsilon^2}$, where we assume t is a power of 2 for convenience. In this paper we consider the one-way communication complexity of computing the following related promise problem $\Pi_{l_2}(t)$. Alice has a vector $x \in [0,1]^t$ with small rational coordinates and with $||x||_2 = 1$. Bob has a basis vector y drawn from the standard basis $\{e_1, \ldots, e_t\}$ of R^t. Both parties are given the promise that either $\langle x, y \rangle = 0$ or $\langle x, y \rangle = \frac{1}{\sqrt{t}}$, where $\langle , \rangle$ denotes Euclidean inner product. We show the one-way communication complexity of deciding $\Pi_{l_2}(t)$ with error probability at most δ when x, y are drawn from a uniform distribution is $\Omega(t)$. We do this by using tools from information theory developed in [5] which generalize the notion of VC-dimension [10] to shatter coefficients.

We then reduce $\Pi_{l_2}(t)$ to that of approximating F_0 of a certain stream. In the reduction we use a deterministic $(1+\gamma)$-distorting embedding ϕ between an l_2^t and an l_1^d norm (defined below) developed in [6], where $d = O\left(t \frac{\log(\frac{1}{\gamma})}{\gamma^2}\right)$. Alice computes $\phi(x)$ and Bob computes $\phi(y)$. Depending on whether $\langle x, y \rangle = 0$ or $\langle x, y \rangle = \frac{1}{\sqrt{t}}$, $||x-y||_2^2 = ||x||_2^2 + ||y||_2^2 - 2\langle x, y \rangle$ will be 2 or $2(1 - \frac{1}{\sqrt{t}})$. By choosing γ appropriately small, $|\phi(x) - \phi(y)|_1 \approx ||x-y||_2$. We then rationally approximate the coordinates of $\phi(x)$ and $\phi(y)$ and scale all coordinates by a common denominator.

We convert the integer coordinates of the scaled $\phi(x)$ and $\phi(y)$ into their unary equivalents, obtaining bit strings x' and y' of length m, where m is determined by parameters chosen in the reduction (specifically, m is a function of ϵ and γ). We show how a $(\frac{1}{\sqrt{t}} = \epsilon, \delta)$ approximation algorithm computing $F_0(a_{x'} \circ a_{y'})$ can decide $\Pi_{l_2}(t)$ with probability at least $1-\delta$. Since the communication complexity of $\Pi_{l_2}(t)$ is $t = \frac{1}{\epsilon^2}$, the space complexity of (ϵ, δ) approximating F_0 in a universe of dimension m is $\Omega\left(\frac{1}{\epsilon^2}\right)$. The goal is then to find the smallest m for a fixed ϵ so that an (ϵ, δ) F_0-approximation algorithm can decide $\Pi_{l_2}(t)$. We determine m to be $\omega\left(\frac{1}{\epsilon^9} \log\left(\frac{1}{\epsilon}\right)\right)$, which shows an $\Omega\left(\frac{1}{\epsilon^2}\right)$ space lower bound for $\epsilon = \Omega(m^{-\frac{1}{9+k}})$, for any $k > 0$, and hence an $m^{\frac{2}{9}}$ lower bound for all smaller ϵ.

2 Preliminaries

2.1 Communication Complexity

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 $f(x,y)$. In the protocols we consider, 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. Note that only one message is sent, and it can only be from Alice to Bob.

Definition 1 *For each randomized protocol Π as described above for computing f, the communication cost of Π is the expected length of the longest message sent from Alice to Bob over all inputs. The δ-**error randomized communication complexity** of f, $R_\delta(f)$, is the communication cost of* *the optimal protocol computing f with error probability δ (that is,* $\Pr[\Pi(x,y) \neq f(x,y)] \leq \delta$*).*

For deterministic protocols with input distribution μ, define $D_{\mu,\delta}(f)$, the **δ-error μ-distributional communication complexity** of f, to be the communication cost of an optimal such protocol. Using the Yao Minimax Principle, $R_\delta(f)$ is bounded from below by $D_{\mu,\delta}$ for any μ [12].

2.2 VC dimension

Let $\mathcal{F} = \{f : \mathcal{X} \to \{0,1\}\}$ be a family of Boolean functions on a domain $\mathcal{X}$. Each $f \in \mathcal{F}$ can be viewed as a $|\mathcal{X}|$-bit string $f_1 \dots f_{|\mathcal{X}|}$.

Definition 2 *For a subset $\mathcal{S} \subseteq \mathcal{X}$, the **shatter coefficient** $SC(f_S)$ of $\mathcal{S}$ is given by $|\{f|_S\}_{f \in \mathcal{F}}|$, the number of distinct bit strings obtained by restricting $\mathcal{F}$ to $\mathcal{S}$. The l-th shatter coefficient $SC(\mathcal{F}, l)$ of $\mathcal{F}$ is the largest number of different bit patterns one can obtain by considering all possible $f|_S$, where S ranges over all subsets of size l. If the shatter coefficient of $\mathcal{S}$ is $2^{|\mathcal{S}|}$, then $\mathcal{S}$ is **shattered** by $\mathcal{F}$. The **VC dimension** of $\mathcal{F}$, VCD($\mathcal{F}$), is the size of the largest subset $\mathcal{S} \subseteq \mathcal{X}$ shattered by $\mathcal{F}$.*

The connection between VC dimension and randomized one-way communication complexity was first explored in [9]. The following theorem lower bounds the (one-way) communcation complexity of f in terms of information theory.

Theorem 3 *For every $f : \mathcal{X} \times \mathcal{Y} \to \{0,1\}$ and every $0 < \delta < 1$, there exists a distribution μ on $\mathcal{X} \times \mathcal{Y}$ such that*

$$D_{\mu,\delta}(f) \geq \ell(1 - H_2(\delta)),$$

where $\ell = VCD(f_{\mathcal{X}})$.

The following generalization of this theorem [5] is useful when computing $VCD(f_{\mathcal{X}})$ is difficult.

Theorem 4 *For every function $f : \mathcal{X} \times \mathcal{Y} \to \{0,1\}$, every $l \geq VCD(f_{\mathcal{X}})$, and every $\delta > 0$, there exists a distribution μ on $\mathcal{X} \times \mathcal{Y}$ such that:*

$$D_{\mu,\delta}(f) \geq \log(SC(f_{\mathcal{X}}, l)) - l \cdot H_2(\delta)$$

2.3 Embeddings

For a survey on low-distortion embeddings the reader is referred to [8]. The l_p^t norm in $\mathcal{R}^t$ of a vector x is defined to be $||x||_p = \left(\sum_{i=1}^t x_i^p\right)^{\frac{1}{p}}$. A $(1+\gamma)$-distortion embedding $\phi : l_r^t \to l_s^d$ is a mapping such that: for any $p, q \in l_r^t$,

$$\frac{1}{1+\gamma}||p-q||_r \leq ||\phi(p) - \phi(q)||_s \leq ||p-q||_r$$

We will need the following theorem [6] in our reduction:

Theorem 5 *For every γ, there exists a $(1+\gamma)$-distortion embedding $\phi : l_2^t \to l_1^d$ with $d = O\left(t\frac{\log(\frac{1}{\gamma})}{\gamma^2}\right)$.*

We will use the notation $|x|$ to mean the l_1 norm of x and $||x||$ to mean the l_2 norm of x.

3 Reduction

We proceed as outlined in section 1.1. Recall that ϵ is fixed, and we set $t = t(\epsilon) = \Theta\left(\frac{1}{\epsilon^2}\right)$, which we assume to be a power of 2 for convenience. We first define and lower bound the communication complexity of $\Pi_{l_2}(t)$.

3.1 Complexity of $\Pi_{l_2}(t)$

Let $E = \{e_1, \dots, e_t\}$ be the standard basis of t unit vectors in R^t. We define the promise problem $\Pi'_{l_2}(t)$ as follows:

$$\begin{aligned}\Pi'_{l_2}(t) \quad &= \quad \{(x,y) \in [0,1]^t \times E \mid ||x|| = 1 \\ &\quad\quad \text{and either } \langle x,y \rangle = 0 \text{ or } \langle x,y\rangle = \frac{2}{\sqrt{t}}\}\end{aligned}$$

We will only be interested in tuples $(x,y) \in \Pi_{l_2}(t)$ in which the descriptions of the coordinates of x and y are finite, so we impose the constraint that these coordinates be rational. We will also need to assign probability distributions on $\mathcal{X} \times \mathcal{Y}$ to use Theorem 4, so it will be convenient to assume these rational numbers have size bounded by $B = \lceil 2\log t \rceil$, where the size of a rational number $\frac{p}{q} = \lceil \log_2(p) + \log_2(q) + 1 \rceil$. We define the promise problem $\Pi_{l_2}(t)$:

$$\begin{aligned}\Pi_{l_2}(t) \quad &= \quad \{(x,y) | (x,y) \in \Pi'_{l_2}(t) \text{ and } \forall i, \\ &\quad\quad x_i \text{ and } y_i \text{ are rational with size } \leq B\}\end{aligned}$$

We let $\mathcal{X}$ denote the set of all x for which there exists a y such that $(x,y) \in \Pi_{l_2}(t)$, and we define $\mathcal{Y}$ similarly. For $(x,y) \in \Pi_{l_2}(t)$, we define $f(x,y) = 0$ if $\langle x,y \rangle = 0$ and $f(x,y) = 1$ if $\langle x,y \rangle = \frac{2}{\sqrt{t}}$. As stated in the preliminaries, we can view $f(x,y)$ as a family of functions $\mathcal{F} = \{f_x(y) : \mathcal{Y} \to \{0,1\} \,|\, x \in \mathcal{X}\}$, where $f_x(y) = f(x,y)$.

Theorem 6 *The $\frac{t}{4}$th shatter coefficient of $\mathcal{F}$ is $2^{H_2(\frac{1}{4})t}$.*

Proof For any subset $T = \{e_{i_1}, \dots e_{i_{\frac{t}{4}}}\} \subseteq \mathcal{Y}$ of $\frac{t}{4}$ vectors, we define x_T to be the normalized average $\frac{2}{\sqrt{t}} \sum_{e \in T} e$. From our assumptions on t, the coordinates of x_T are rational with size bounded above by B. We define the set $\mathcal{X}_1 \subseteq \mathcal{X}$ as $\mathcal{X}_1 = \{x_T \mid T \subseteq \mathcal{Y}\}$. The claim is that every length-t bit string with exactly $\frac{t}{4}$ 1s will occur in the truth table of $f_{\mathcal{X}_1}$. Consider any such string with 1s in positions $i_1, \dots, i_{\frac{t}{4}}$, and let $T = \{e_{i_1}, \dots e_{i_{\frac{t}{4}}}\}$. Then for all $e \in T$, $\langle x_T, e \rangle = \langle \frac{2}{\sqrt{t}} \sum_{e \in T} e, e \rangle = \frac{2}{\sqrt{t}}$ so that $f_{x_T}(e) = 1$, and for all $e \notin T$, $\langle x_T, e\rangle = 0$ since x_T is in an orthogonal subspace to e, and hence $f_{x_T}(e) = 0$. Since there are $\binom{t}{\frac{t}{4}} \approx 2^{H_2(\frac{1}{4})t}$ such strings, the theorem follows. ■

Corollary 7 *For all $\delta < \frac{1}{4}$, the one-way communication complexity $R_\delta(f)$ is $\Omega(t)$.*

Proof We can apply Theorem 5 so long as $t \geq VCD(f_{\mathcal{X}_1})$, but this is clear since $|\mathcal{Y}| = t$ so that for all subsets $\mathcal{X}_1 \subseteq \mathcal{X}$, $VCD(f_{\mathcal{X}_1}) \leq t$. We deduce that there exists an input distribution μ such that $D_{\mu,\delta}(f)$ is at least $\log(SC(f_{\mathcal{X}_1}, t)) - tH_2(\delta) \approx t(H_2(\frac{1}{4}) - H_2(\delta))) = \Omega(t)$. By the Yao minimax principle [12] the corollary follows. ■

We will need the following connection between l_2^t distances and dot products: Let $(x, y) \in \Pi_{l_2}(t)$. Using the relation $||y - x||^2 = ||y||^2 + ||x||^2 - 2\langle x, y\rangle$, the property that $||y|| = ||x|| = 1$, and the inequality $\sqrt{1-\epsilon} < 1 - \frac{\epsilon}{2}$ for $0 < \epsilon < 1$, we see:

$$\begin{aligned} f(x,y) = 0 \quad &\Rightarrow \quad ||y-x||^2 = 2 \\ &\Rightarrow \quad ||y-x|| = \sqrt{2} \\ f(x,y) = 1 \quad &\Rightarrow \quad ||y-x||^2 = 2 - \frac{4}{\sqrt{t}} \\ &\Rightarrow \quad ||y-x|| = \sqrt{2}\sqrt{1-\frac{2}{\sqrt{t}}} \\ &\qquad\qquad < \sqrt{2}(1-\frac{1}{\sqrt{t}}) \end{aligned}$$

3.2 Embedding l_2^t into $l_1^{poly(t)}$

Let $\gamma = \gamma(t)$ be a function to be specified later. Let ϕ be a $(1+\gamma)$-distortion embedding $\phi : l_2^t \to l_1^d$, with $d = O\left(t\frac{\log(\frac{1}{\gamma})}{\gamma^2}\right)$, as per Theorem 6. Alice and Bob can construct the same embedding ϕ locally without any communication overhead. Let (x, y) be an instance of $\Pi_{l_2}(t)$ and let Alice possess x and Bob possess y. Alice computes $\phi(x)$, Bob computes $\phi(y)$. By the distance-distorting properties of ϕ, we have:

$$\frac{1}{1+\gamma} \leq \quad |\phi(x)|,\ |\phi(y)| \quad \leq 1 \qquad (1)$$

$$\frac{||y-x||}{1+\gamma} \leq \quad |\phi(y) - \phi(x)| \quad \leq ||y-x|| \qquad (2)$$

3.3 Rational Approximation

We will need the coordinates of $\phi(x)$ and $\phi(y)$ to be rational. This will change $|\phi(x)|$, $|\phi(y)|$, and $|\phi(x) - \phi(y))|$, but not by much if we choose a good rational approximation. We fix a function $z = z(t) : \mathcal{N} \to \mathcal{N}$ to be specified later. Let $[z]$ denote the set of nonnegative integers less than z. Let $\{r\}$ denote the fractional part of a real number r, so that $r - \{r\}$ is an integer. Then it easily follows that for any real number r, there exists a unique $s = s(r) \in [z]$ with $0 \leq \{r\} - \frac{s}{z} \leq \frac{1}{z}$. For this value of s we define the rational approximation $\psi(r)$ of r to be $\psi(r) = (r - \{r\}) + \frac{s}{z}$. For a d-dimensional real vector $v = (r_1, \ldots, r_d)$, we define $\psi(v) = (\psi(r_1), \ldots, \psi(r_d))$.

By our choice of rational approximation, we have:

$$|\phi(x)| - \frac{d}{z} \leq |\psi(\phi(x))| \leq |\phi(x)| + \frac{d}{z} \qquad (3)$$

$$|\phi(y)| - \frac{d}{z} \leq |\psi(\phi(y))| \leq |\phi(y)| + \frac{d}{z} \qquad (4)$$

$$|\phi(y) - \phi(x)| - \frac{d}{z} \leq |\psi(\phi(y)) - \psi(\phi(x))| \qquad (5)$$

$$|\psi(\phi(y)) - \psi(\phi(x))| \leq |\phi(y) - \phi(x)| + \frac{d}{z} \qquad (6)$$

3.4 Reduction to Distinct Elements

Alice and Bob now convert their transformed inputs to integer vectors. To do this, they scale each coordinate by z, scaling the norm of their vectors by z. For d-dimensional vectors $v = (v_1, \ldots, v_d)$, let $s(v)$ denote the vector $(z * v_1, \ldots, z * v_d)$, so that Alice and Bob now have $s(\psi(\phi(x)))$ and $s(\psi(\phi(y)))$ respectively.

The idea is to convert each of these integer coordinates to their unary representation so that we can reduce this problem to that of computing Hamming distances between bit strings. Since $|\psi(\phi(x))| \leq |\phi(x)| + \frac{d}{z} \leq 2$, each coordinate of $\psi(\phi(x))$ is rational with absolute value less than or equal to 2. Hence, each coordinate of $s(\psi(\phi(x)))$ is an integer with absolute value less than or equal to $2z$. For each integer i, $-2z \leq i \leq 2z$, we define its unary equivalent $u(i)$ to be a bit string of length $4z$ with first $2z + i$ bit positions to be 1s, and remaining bit positions to be 0s, namely, $u(i) = 1^{i+2z}0^{2z-i}$. For a d-dimensional integer vector $v = (v_1, \ldots, v_d)$ with coordinates in the range $[-2z, 2z]$, we define $u(v)$ to be $(u(v_1), \ldots, u(v_d))$. For any two such vectors v_1, v_2 with coordinates in $[-2z, 2z]$, it is easy to see that $|v_1 - v_2| = \Delta(u(v_1), u(v_2))$, where Δ refers to Hamming distance. Let $x' = u(s(\psi(\phi(x))))$ and $y' = u(s(\psi(\phi(y))))$. Note that both x' and y' are bit strings of length $(4z + 1)d$.

Let $wt(x)$ denote the number of 1s in bit string x. Since $|s(\psi(\phi(x)))| = wt(x')$ and $|s(\psi(\phi(y)))| = wt(y')$, combining (1), (3) and (4), we see that:

$$z\left(\frac{1}{1+\gamma} - \frac{d}{z}\right) \leq wt(x') \leq z\left(1 + \frac{d}{z}\right)$$

$$z\left(\frac{1}{1+\gamma} - \frac{d}{z}\right) \leq wt(y') \leq z\left(1 + \frac{d}{z}\right)$$

Also, since $\Delta(x',y') = |s(\psi(\phi(x))) - s(\psi(\phi(y)))| = z|\psi(\phi(x)) - \psi(\phi(y))|$, combining (2), (5), and (6), we observe:

$$\frac{z||y-x||}{1+\gamma} - d \leq \Delta(x',y') \leq z||y-x|| + d \tag{7}$$

We now transform these observations on Hamming distances into observations on the number of distinct elements of streams. Alice and Bob can pretend their bit strings x' and y' are the characteristic vectors of certain streams $\mathbf{a}_{x'}$ and $\mathbf{a}_{y'}$ in a universe of size $(4z+1)d$. There are an unbounded number of streams that Alice and Bob can choose from which have characteristic vectors x' and y'. They choose two such streams arbitrarily, say $\mathbf{a}_{x'}$ and $\mathbf{a}_{y'}$, which will be fed into local copies of an F_0 approximation algorithm. Let $\mathbf{a}_{x'} \circ \mathbf{a}_{y'}$ be the stream which is the concatenation of streams $\mathbf{a}_{x'}$ and $\mathbf{a}_{y'}$, so that its characteristic vector is just the bitwise OR $x' \vee y'$ of x' and y'.

Claim 8 *Let $x', y', \boldsymbol{a}_{x'}, \boldsymbol{a}_{y'}$ be as above. Suppose $\beta_1(z) \leq wt(x'), wt(y') \leq \beta_2(z)$ for some $\beta_1(z) \leq \beta_2(z)$ functions of z. Then $\beta_1(z) + \frac{\Delta(x',y')}{2} \leq F_0(\boldsymbol{a}_{x'} \circ \boldsymbol{a}_{y'}) \leq \beta_2(z) + \frac{\Delta(x',y')}{2}$.*

Proof As noted above, $F_0(\mathbf{a}_{x'} \circ \mathbf{a}_{y'})$ is just $wt(x' \vee y')$. Let c be the number of positions which are 1 in y' but 0 in x'. Then $\Delta(x',y') = (wt(x') - (wt(y') - c)) + c = wt(x') - wt(y') + 2c$, so that $c = \frac{1}{2}(\Delta(x',y') + wt(y') - wt(x'))$ and $wt(x' \vee y') = wt(x') + c = \frac{1}{2}(\Delta(x',y') + wt(x') + wt(y'))$. The claim follows from the bounds on $wt(x'), wt(y')$. ■

Now suppose $f(x,y) = 0$. Then $||x-y|| = \sqrt{2}$, so that we have $\frac{z\sqrt{2}}{1+\gamma} - d \leq \Delta(x',y')$ by (7). By Claim 8 we conclude,

$$\begin{aligned} F_0(\mathbf{a}_{x'} \circ \mathbf{a}_{y'}) &\geq -\frac{3d}{2} + \frac{z}{1+\gamma} + \frac{z\sqrt{2}}{2+2\gamma} \\ &= -\frac{3d}{2} + \frac{z\left(1+\frac{1}{\sqrt{2}}\right)}{1+\gamma} \end{aligned}$$

On the other hand, if $f(x,y) = 1$, we have

$$||x-y|| < \sqrt{2}\left(1 - \frac{1}{\sqrt{t}}\right)$$

so that

$$\Delta(x',y') < z\left(\sqrt{2}\left(1-\frac{1}{\sqrt{t}}\right) + \frac{d}{z}\right)$$

by (7), and hence by Claim 8 we have

$$\begin{aligned} F_0(\mathbf{a}_{x'} \circ \mathbf{a}_{y'}) &\leq z + d + \frac{z}{\sqrt{2}} - \frac{z}{\sqrt{2t}} + \frac{d}{2} \\ &= \frac{3d}{2} + z\left(1+\frac{1}{\sqrt{2}}\right) - \frac{z}{\sqrt{2t}} \end{aligned}$$

It remains to choose z and γ so that these two cases can be distinguished by an $\left(\epsilon = \frac{1}{\sqrt{t}}, \delta\right)$ approximation algorithm for distinct elements. Let $\epsilon' = c * \epsilon$ for some small constant c to be determined shortly. Suppose we have an (ϵ', δ) F_0-approximation algorithm which can distinguish these two cases. Back substituting ϵ for $\frac{1}{\sqrt{t}}$, we want:

$$\begin{aligned} &(1+\epsilon')\left(\frac{3d}{2} + z\left(1+\frac{1}{\sqrt{2}}\right) - \frac{\epsilon z}{\sqrt{2}}\right) \\ &< (1-\epsilon')\left(-\frac{3d}{2} + \frac{z}{1+\gamma}\left(1+\frac{1}{\sqrt{2}}\right)\right) \end{aligned} \tag{8}$$

If $d = o(z\epsilon)$, there exists a constant ϵ_0 such that for all $\epsilon < \epsilon_0$, (8) is equivalent to:

$$\begin{aligned} &(1+\epsilon')\left(z\left(1+\frac{1}{\sqrt{2}}\right) - \frac{\epsilon z}{\sqrt{2}}\right) \\ &< (1-\epsilon')\left(\frac{z}{1+\gamma}\left(1+\frac{1}{\sqrt{2}}\right)\right) \end{aligned}$$

Setting $c = \frac{1}{3(\sqrt{2}+1)}$, one finds after some algebra that for sufficiently small, but constant ϵ, one can set $\gamma = \Theta(\epsilon)$.

Our computations mean that $d = O\left(\frac{1}{\epsilon^4}\log\left(\frac{1}{\epsilon}\right)\right)$ so that $d = o(z\epsilon)$ so long as we set $z = \omega\left(\frac{1}{\epsilon^5}\log\left(\frac{1}{\epsilon}\right)\right)$ and hence in dimension $\Theta(zd) = \omega\left(\frac{1}{\epsilon^9}\log\left(\frac{1}{\epsilon}\right)\right)$ an $(\Theta(\epsilon), \delta)$ F_0 approximator can distinguish the above two cases with error probability at most δ, which means by the reduction that it must take $\Omega\left(\frac{1}{\epsilon^2}\right)$ space.

Let $m = \Theta\left(\frac{1}{\epsilon^9}\log\left(\frac{1}{\epsilon}\right)\right)$. Then we see that for all $\epsilon = \Omega\left(m^{-\frac{1}{9+k}}\right)$, for any $k > 0$, there is a space lower bound of $\Omega\left(\frac{1}{\epsilon^2}\right)$ on (ϵ, δ) approximating the number of distinct elements in a universe of size m.

4 Conclusions

We have shown a tight space lower bound of $\Omega\left(\frac{1}{\epsilon^2}\right)$ on (ϵ, δ) approximating the number of distinct elements in a universe of size m when $\epsilon = \Omega\left(m^{-\frac{1}{9+k}}\right)$, for any $k > 0$.

The upper bound of $o(m^{\frac{1}{9}})$ on $1/\epsilon$ can be somewhat relaxed by strengthening the analysis presented in this paper. For example, one could use a *randomized* embedding of the relevant l_2^t vectors into l_1^d; this would give $d = O(\log t/\gamma^2)$ and lead to a somewhat higher upper bound on $1/\epsilon$. Instead of following along these lines, we mention that, very recently, the second author managed to improve the upper bound on $1/\epsilon$ to $m^{1/2}$ [11], which is optimal. The approach of that paper is as follows. Observe that one can reformulate the reductions given in Section 3 as a method for constructing a set of vectors in $\{0,1\}^m$ that results in a large bound for the shatter coefficient of the function $f(x, y)$. The set is constructed indirectly: first, the vectors are chosen in l_2^t, then they are mapped into l_1^d, and then finally into the Hamming space. This indirect route blows up the dimension of the vectors by a polynomial factor. Instead, in [11], the vectors are constructed directly in the Hamming space via a (fairly involved) probabilistic argument.

5 Acknowledgments

For helpful discussions, we thank: Johnny Chen, Ron Rivest, Naveen Sunkavally, Hoeteck Wee. Also many thanks to Hoeteck for reading and commenting on previous drafts of this paper.

References

[1] N. Alon, Y. Matias, and M. Szegedy. The space complexity of approximating the frequency moments. In *Proceedings of the 28th Annual ACM Symposum on the Theory of Computing*, p. 20-29, 1996. Also; J. Comp. Sys. Sci. 58 (1999), 137-147.

[2] A. Akella, A. R. Bharambe, M. Reiter and S. Seshan. Detecting DDoS attacks on ISP networks. In *Proceedings of the Workshop on Management and Processing of Data Streams*, 2003.

[3] Z. Bar Yossef. The complexity of massive data set computations. Ph.D. Thesis, U.C. Berkeley, 2002.

[4] Z. Bar Yossef, T.S. Jayram, R. Kumar, D. Sivakumar, and Luca Trevisan. Counting distinct elements in a data stream. *RANDOM 2002, 6th. International Workshop on Randomization and Approximation Techniques in Computer Science*, p. 1-10, 2002.

[5] Z. Bar Yossef, T.S. Jayram, R. Kumar, and D. Sivakumar. Information Theory Methods in Communication Complexity. *17th IEEE Annual Conference on Computational Complexity*, p. 93-102, 2002.

[6] T. Figiel, J. Lindenstrauss, and V. D. Milman. The Dimension of Almost Spherical Sections of Convex Bodies. *Acta Mathematica* (139) 53-94, 1977.

[7] P. Flajolet and G.N. Martin. Probabilistic counting algorithms for data base applications. *Journal of Computer and System Sciences*, 18(2) 143-154, 1979.

[8] P. Indyk. Algorithmic applications of low-distortion geometric embeddings. In *Proceedings of the 42nd Annual IEEE Symposium on Foundations of Computer Science*, invited talk, Las Vegas, Nevada, 14-17 October, 2001.

[9] I. Kremer, N. Nisan, and D. Ron. On randomized one-round communication complexity. *Computational Complexity*, 8(1):21-49, 1999.

[10] V.N. Vapnik and A.Y. Chervonenkis. On the uniform converges of relative frequences of events to their probabilities. *Theory of Probability and its Applications*, XVI(2):264-280, 1971.

[11] D. P. Woodruff. Optimal space lower bounds for all frequency moments. Available: http://web.mit.edu/dpwood/www

[12] A. C-C. Yao. Lower bounds by probabilistic arguments. In *Proceedings of the 24th Annual IEEE Symposium on Foundations of Computer Science*, p. 420-428, 1983.