

Optimal Bounds for Johnson-Lindenstrauss Transforms and Streaming Problems with Subconstant Error

T. S. JAYRAM and DAVID P. WOODRUFF, IBM Almaden

The Johnson-Lindenstrauss transform is a dimensionality reduction technique with a wide range of applications to theoretical computer science. It is specified by a distribution over projection matrices from $\mathbb{R}^n \rightarrow \mathbb{R}^k$ where $k \ll n$ and states that $k = O(\varepsilon^{-2} \log 1/\delta)$ dimensions suffice to approximate the norm of any fixed vector in \mathbb{R}^n to within a factor of $1 \pm \varepsilon$ with probability at least $1 - \delta$. In this article, we show that this bound on k is optimal up to a constant factor, improving upon a previous $\Omega((\varepsilon^{-2} \log 1/\delta) / \log(1/\varepsilon))$ dimension bound of Alon. Our techniques are based on lower bounding the information cost of a novel one-way communication game and yield the first space lower bounds in a data stream model that depend on the error probability δ .

For many streaming problems, the most naïve way of achieving error probability δ is to first achieve constant probability, then take the median of $O(\log 1/\delta)$ independent repetitions. Our techniques show that for a wide range of problems, this is in fact optimal! As an example, we show that estimating the ℓ_p -distance for any $p \in [0, 2]$ requires $\Omega(\varepsilon^{-2} \log n \log 1/\delta)$ space, even for vectors in $\{0, 1\}^n$. This is optimal in all parameters and closes a long line of work on this problem. We also show the number of distinct elements requires $\Omega(\varepsilon^{-2} \log 1/\delta + \log n)$ space, which is optimal if $\varepsilon^{-2} = \Omega(\log n)$. We also improve previous lower bounds for entropy in the strict turnstile and general turnstile models by a multiplicative factor of $\Omega(\log 1/\delta)$. Finally, we give an application to one-way communication complexity under product distributions, showing that, unlike the case of constant δ , the VC-dimension does not characterize the complexity when $\delta = o(1)$.

Categories and Subject Descriptors: F.2.3 [Analysis of Algorithms and Problem Complexity]: Tradeoffs between Complexity Measures

General Terms: Algorithms, Theory

Additional Key Words and Phrases: Communication complexity, data streams, distinct elements, entropy, frequency moments

ACM Reference Format:

Jayram, T. S. and Woodruff, D. P. 2013. Optimal bounds for Johnson-Lindenstrauss transforms and streaming problems with subconstant error. *ACM Trans. Algorithms* 9, 3, Article 26 (June 2013), 17 pages.
DOI: <http://dx.doi.org/10.1145/2483699.2483706>

1. INTRODUCTION

The Johnson-Lindenstrauss transform is a fundamental dimensionality reduction technique with applications to many areas such as nearest-neighbor search [Ailon and Chazelle 2009; Indyk and Motwani 1998], compressed sensing [Candès and Tao 2006], computational geometry [Clarkson 2008], data streams [Alon et al. 1999; Indyk 2006], graph sparsification [Spielman and Srivastava 2008], machine learning [Langford et al. 2007; Shi et al. 2009; Weinberger et al. 2009], and numerical linear algebra [Clarkson and Woodruff 2009; Drineas et al. 2007; Rokhlin et al. 2009; Sarlós 2006]. It is given by a projection matrix that maps vectors in \mathbb{R}^n to \mathbb{R}^k , where $k \ll n$, while seeking to approximately preserve their norm. The classical result states that a

Authors' address: T. S. Jayram and D. P. Woodruff, IBM Almaden, 650 Harry Road, San Jose, CA 95120; email: {jayram, dpwoodru}@us.ibm.com.

Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies show this notice on the first page or initial screen of a display along with the full citation. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, to republish, to post on servers, to redistribute to lists, or to use any component of this work in other works requires prior specific permission and/or a fee. Permissions may be requested from Publications Dept., ACM, Inc., 2 Penn Plaza, Suite 701, New York, NY 10121-0701 USA, fax +1 (212) 869-0481, or permissions@acm.org.

© 2013 ACM 1549-6325/2013/06-ART26 \$15.00

DOI: <http://dx.doi.org/10.1145/2483699.2483706>

random projection with $k = O(\frac{1}{\varepsilon^2} \log 1/\delta)$ dimensions suffices to approximate the norm of any fixed vector in \mathbb{R}^n to within a factor of $1 \pm \varepsilon$ with probability at least $1 - \delta$. This is a remarkable result because the target dimension is *independent* of n . Because the transform is linear, it also preserves the pairwise distances of the vectors in this set, which is what is needed for most applications. The projection matrix is itself produced by a random process that is oblivious to the input vectors. Since the original work of Johnson and Lindenstrauss, it has been shown [Achlioptas 2003; Arriaga and Vempala 1999; Dasgupta and Gupta 2003; Indyk and Motwani 1998] that the projection matrix could be constructed element-wise using the standard Gaussian distribution or even uniform ± 1 variables [Achlioptas 2003]. By setting the size of the target dimension $k = O(\frac{1}{\varepsilon^2} \log 1/\delta)$, the resulting matrix, suitably scaled, is guaranteed to approximate the norm of any single vector with failure probability δ .

Due to its algorithmic importance, there has been a flurry of research aiming to improve upon these constructions that address both the time needed to generate a suitable projection matrix as well as to produce the transform of the input vectors [Ailon and Chazelle 2009, 2010; Ailon and Liberty 2009, 2010; Liberty et al. 2008]. In the area of data streams, the Johnson-Lindenstrauss transform has been used in the seminal work of Alon et al. [1999] as a building block to produce *sketches* of the input that can be used to estimate norms. For a stream with $\text{poly}(n)$ increments/decrements to a vector in \mathbb{R}^n , the size of the sketch can be made to be $O(\frac{1}{\varepsilon^2} \log n \log 1/\delta)$. To achieve even better update times, Thorup and Zhang [2004], building upon the COUNT SKETCH data structure of Charikar et al. [2002], use an ultra-sparse transform to estimate the norm, but then have to take a median of several estimators in order to reduce the failure probability. This is inherently nonlinear but suggests the power of such schemes in addressing sparsity as a goal; in contrast, a single transform with constant sparsity per column fails to be an (ε, δ) -JL transform [Dasgupta et al. 2010; Matousek 2008].

In this article, we consider the central lower bound question of Johnson-Lindenstrauss transforms: how good is the upper bound on the target dimension of $k = O(\frac{1}{\varepsilon^2} \log 1/\delta)$ on the target dimension to approximate the norm of a fixed vector in \mathbb{R}^n ? Alon [2003] gave a near-tight lower bound of $\Omega(\frac{1}{\varepsilon^2} (\log 1/\delta) / \log(1/\varepsilon))$, leaving an asymptotic gap of $\log(1/\varepsilon)$ between the upper and lower bounds. In this article, we close the gap and resolve the optimality of Johnson-Lindenstrauss transforms by giving a lower bound of $k = \Omega(\frac{1}{\varepsilon^2} \log 1/\delta)$ dimensions. More generally, we show that any sketching algorithm for estimating the norm (whether linear or not) of vectors in \mathbb{R}^n must use space at least $\Omega(\frac{1}{\varepsilon^2} \log n \log 1/\delta)$ to approximate the norm within a $1 \pm \varepsilon$ factor with a failure probability of at most δ . By a simple reduction, we show that this result implies the aforementioned lower bound on Johnson-Lindenstrauss transforms.

Our results come from lower-bounding the information cost of a novel one-way communication complexity problem. One can view our results as a strengthening of the augmented-indexing problem [Ba et al. 2010; Bar-Yossef et al. 2004; Clarkson and Woodruff 2009; Kane et al. 2010a; Miltersen et al. 1998] to very large domains. Our technique is far-reaching, implying the first lower bounds for the space complexity of streaming algorithms that depends on the error probability δ . The connection to streaming follows via a standard reduction from a two-player communication problem to a streaming problem. In this reduction, the first player runs the streaming algorithm on her input, and passes the state to the second player, who continues running the streaming algorithm on his input. If the output of the streaming algorithm can be used to solve the communication problem, then the size of its state must be at least the communication required of the communication problem.

In many cases, our results are tight. For instance, for estimating the ℓ_p -norm for any $p \geq 0$ in the turnstile model,¹ we prove an $\Omega(\varepsilon^{-2} \log n \log 1/\delta)$ space lower bound for streams with $\text{poly}(n)$ increments/decrements. This resolves a long sequence of work on this problem [Indyk and Woodruff 2003; Kane et al. 2010a; Woodruff 2004] and is simultaneously optimal in ε, n , and δ . For $p \in (0, 2]$, this matches the upper bound of Kane et al. [2010a]. Indeed, in Kane et al. [2010a], it was shown how to achieve $O(\varepsilon^{-2} \log n)$ space and constant probability of error. To reduce this to error probability δ , run the algorithm $O(\log 1/\delta)$ times in parallel and take the median. Surprisingly, this is optimal! For estimating the number of distinct elements in a data stream, we prove an $\Omega(\varepsilon^{-2} \log 1/\delta + \log n)$ space lower bound, improving upon the previous $\Omega(\log n)$ bound of Alon et al. [1999] and $\Omega(\varepsilon^{-2})$ bound of Indyk and Woodruff [2003] and Woodruff [2004]. In Kane et al. [2010a, 2010b], an $O(\varepsilon^{-2} + \log n)$ -space algorithm is given with constant probability of success. We show that if $\varepsilon^{-2} = \Omega(\log n)$, then running their algorithm in parallel $O(\log 1/\delta)$ times and taking the median of the results is optimal. Similarly, we improve the known $\Omega(\varepsilon^{-2} \log n)$ bound for estimating the entropy in the turnstile model to $\Omega(\varepsilon^{-2} \log n \log 1/\delta)$, and we improve the previous $\Omega(\varepsilon^{-2} \log n / \log 1/\varepsilon)$ bound [Kane et al. 2010a] for estimating the entropy in the strict turnstile model to $\Omega(\varepsilon^{-2} \log n \log 1/\delta / \log 1/\varepsilon)$. Entropy has become an important tool in databases as a way of understanding database design, enabling data integration, and performing data anonymization [Srivastava and Venkatasubramanian 2010]. Estimating this quantity in an efficient manner over large sets is a crucial ingredient in performing this analysis (see the recent tutorial in Srivastava and Venkatasubramanian [2010] and the references therein).

Kremer et al. [1999] showed the surprising theorem that for constant error probability δ , the one-way communication complexity of a function under product distributions coincides with the VC-dimension of the communication matrix for the function. We show that for sub-constant δ , such a nice characterization is not possible. Namely, we exhibit two functions with the same VC-dimension whose communication complexities differ by a multiplicative $\log 1/\delta$ factor.

Organization. In Section 2, we give preliminaries on communication and information complexity. In Section 3, we give our lower bound for augmented-indexing over larger domains. In Section 4, we give the improved lower bound for Johnson-Lindenstrauss transforms and the streaming and communication applications previously mentioned. In Section 5, we discuss open problems.

2. PRELIMINARIES

Let $[a, b]$ denote the set of integers $\{i \mid a \leq i \leq b\}$, and let $[n] = [1, n]$. Random variables will be denoted by upper case Roman or Greek letters, and the values they take by (typically corresponding) lowercase letters. Probability distributions will be denoted by lowercase Greek letters. A random variable X with distribution μ is denoted by $X \sim \mu$. If μ is the uniform distribution over a set \mathcal{U} , then this is also denoted as $X \in_R \mathcal{U}$.

2.1. One-Way Communication Complexity

Let \mathcal{D} denote the input domain and \mathcal{O} the set of outputs. Consider the two-party communication model, where Alice holds an input $x \in \mathcal{D}$ and Bob holds an input $y \in \mathcal{D}$. Their goal is to solve some relation problem $\mathcal{Q} \subseteq \mathcal{D} \times \mathcal{D} \times \mathcal{O}$. For each $(x, y) \in \mathcal{D}^2$, the set $\mathcal{Q}_{xy} = \{z \mid (x, y, z) \in \mathcal{Q}\}$ represents the set of possible answers on input (x, y) . Let $\mathcal{L} \subseteq \mathcal{D}^2$ be the set of legal or *promise* inputs, that is, pairs (x, y) such that $\mathcal{Q}_{xy} \neq \mathcal{O}$.

¹Technically, for $p < 1$, ℓ_p is not a norm, but it is still a well-defined quantity.

is a (partial) function on \mathcal{D}^2 if for every (x, y) , Q_{xy} has size 1 or $Q_{xy} = \mathcal{O}$. In a *one-way communication protocol* \mathcal{P} , Alice sends a single message to Bob, following which Bob outputs an answer in \mathcal{O} . The maximum length of Alice's message (in bits) over all inputs is the *communication cost* of the protocol \mathcal{P} . The protocol is allowed to be randomized in which the players have *private* access to an unlimited supply of random coins. The protocol solves the communication problem Q if the answer on any input $(x, y) \in \mathcal{L}$ belongs to Q_{xy} with failure probability at most δ . Note that the protocol is legally defined for all inputs, however, no restriction is placed on the answer of the protocol for non-promise inputs. The *one-way communication complexity* of Q , denoted by $R_\delta^\rightarrow(Q)$, is the minimum communication cost of a protocol for Q with failure probability at most δ . A related complexity measure is distributional complexity $D_{\mu, \delta}^\rightarrow(Q)$ with respect to a distribution μ over \mathcal{L} . This is the cost of the best *deterministic* protocol for Q that has error probability at most δ when the inputs are drawn from distribution μ . By Yao's lemma, $R_\delta^\rightarrow(Q) = \max_\mu D_{\mu, \delta}^\rightarrow(Q)$. Define $R_\delta^{\rightarrow, \parallel}(Q) = \max_{\text{product } \mu} D_{\mu, \delta}^\rightarrow(Q)$, where now the maximum is taken only over product distributions μ on \mathcal{L} (if no such distribution exists then $R_\delta^{\rightarrow, \parallel}(Q) = 0$). Here, by product distribution, we mean that Alice and Bob's inputs are chosen independently.

We note that the public-coin one-way communication complexity, that is, the one-way communication complexity in which the parties additionally share an infinitely long random string and denoted $R_\delta^{\rightarrow, \text{pub}}$, is at least $R_\delta^\rightarrow - O(\log I)$, where I is the sum of input lengths to the two parties [Kremer et al. 1999].

Another restricted model of communication is *simultaneous* or *sketch*-based communication, where Alice and Bob each send a message (sketch) depending only on her/his own input (as well as private coins) to a referee. The referee then outputs the answer based on the two sketches. The communication cost is the maximum sketch sizes (in bits) of the two players.

Note. When δ is fixed (say 1/4), we will usually suppress it in the terms involving δ .

2.2. Information Complexity

We summarize basic properties of entropy and mutual information (for proofs, see Chapter 2 of Cover and Thomas [1991]).

PROPOSITION 2.1.

- (1) *Entropy Span.* If X takes on at most s values, then $0 \leq H(X) \leq \log s$.
- (2) $I(X : Y) \stackrel{\text{def}}{=} H(X) - H(X|Y) \geq 0$, that is, $H(X | Y) \leq H(X)$.
- (3) *Chain rule.* $I(X_1, X_2, \dots, X_n : Y | Z) = \sum_{i=1}^n I(X_i : Y | X_1, X_2, \dots, X_{i-1}, Z)$
- (4) *Subadditivity.* $H(X, Y | Z) \leq H(X | Z) + H(Y | Z)$, and equality holds if and only if X and Y are independent conditioned on Z .
- (5) *Fano's inequality:* Let A be a "predictor" of X , that is, there is a function g such that $\Pr[g(A) = X] \geq 1 - \delta$ for some $\delta < 1/2$. Let \mathcal{U} denote the support of X , where $|\mathcal{U}| \geq 2$. Then, $H(X | A) \leq \delta \log(|\mathcal{U}| - 1) + h_2(\delta)$, where $h_2(\delta) \stackrel{\text{def}}{=} \delta \log \frac{1}{\delta} + (1 - \delta) \log \frac{1}{1-\delta}$ is the binary entropy function.

Recently, the *information complexity* paradigm, in which the information about the inputs revealed by the message(s) of a protocol is studied, has played a key role in resolving important communication complexity problems [Barak et al. 2010; Bar-Yossef et al. 2002; Chakrabarti et al. 2001; Harsha et al. 2007; Jain et al. 2008]. We do not need the full power of these techniques in this article. There are several possible definitions of information complexity that have been considered depending on the

Problem: $\text{IND}_{\mathcal{U}}^a$

Promise Inputs:

Alice gets $\mathbf{x} = (x_1, x_2, \dots, x_N) \in \mathcal{U}^N$.

Bob gets $\mathbf{y} = (y_1, y_2, \dots, y_N) \in (\mathcal{U} \cup \{\perp\})^N$ such that for some (unique) i :

- (1) $y_i \in \mathcal{U}$,
- (2) $y_k = x_k$ for all $k < i$,
- (3) $y_{i+1} = y_{i+2} = \dots = y_N = \perp$

Output:

Does $x_i = y_i$ (yes/no)?

Fig. 1. Communication problem $\text{IND}_{\mathcal{U}}^a$.

application. Our definition is tuned specifically for one-way protocols, similar in spirit to Bar-Yossef et al. [2002] (see also Barak et al. [2010]).

Definition 2.2. Let \mathcal{P} be a one-way protocol. Suppose μ is a distribution over its input domain \mathcal{D} . Let Alice's input X be chosen according to μ . Let A be the random variable denoting Alice's message on input $X \sim \mu$; A is a function of X and Alice's private coins. The *information cost* of \mathcal{P} under μ is defined to be $I(X : A)$.

The *one-way information complexity* of a problem Q with respect to μ and δ , denoted by $\text{IC}_{\mu, \delta}^{\rightarrow}(Q)$, is defined to be the minimum information cost of a one-way protocol under μ that solves Q with failure probability at most δ .

By the entropy span bound (Proposition 2.1),

$$I(X : A) = H(A) - H(A | X) \leq H(A) \leq |A|,$$

where $|A|$ denotes the length of Alice's message.

PROPOSITION 2.3. *For every probability distribution μ on inputs,*

$$R_{\delta}^{\rightarrow}(Q) \geq \text{IC}_{\mu, \delta}^{\rightarrow}(Q).$$

2.3. JL Transforms

Definition 2.4. A random family \mathcal{F} of $k \times n$ matrices A , together with a distribution μ on \mathcal{F} , forms a Johnson-Lindenstrauss transform with parameters ε, δ , or (ε, δ) -JLT for short, if for any fixed vector $x \in \mathbb{R}^n$,

$$\Pr_{A \sim \mu} [(1 - \varepsilon)\|x\|_2^2 \leq \|Ax\|_2^2 \leq (1 + \varepsilon)\|x\|_2^2] \geq 1 - \delta.$$

We say that k is the dimension of the transform.

3. AUGMENTED INDEXING ON LARGE DOMAINS

For a sufficiently large universe \mathcal{U} and an element $\perp \notin \mathcal{U}$, fix $\mathcal{U} \cup \{\perp\}$ to be the input domain. Consider the decision problem known as *augmented indexing* with respect to \mathcal{U} ($\text{IND}_{\mathcal{U}}^a$) as shown in Figure 1.

Let μ be the uniform distribution on \mathcal{U} and let μ^N denote the product distribution on \mathcal{U}^N .

THEOREM 3.1. *Suppose the failure probability $\delta \leq \frac{1}{4|\mathcal{U}|}$. Then,*

$$\text{IC}_{\mu^N, \delta}^{\rightarrow}(\text{IND}_{\mathcal{U}}^a) \geq N \log(|\mathcal{U}|)/2$$

PROOF. The proof uses some of the machinery developed for direct sum theorems in information complexity.

Let $\mathbf{X} = (X_1, X_2, \dots, X_N) \sim \mu^N$, and let A denote Alice's message on input \mathbf{X} in a protocol for $\text{IND}_{\mathcal{U}}^a$ with failure probability δ . By the chain rule for mutual information (Proposition 2.1),

$$\begin{aligned} \text{I}(\mathbf{X} : A) &= \sum_{i=1}^N \text{I}(X_i : A \mid X_1, X_2, \dots, X_{i-1}) \\ &= \sum_{i=1}^N \text{H}(X_i \mid X_1, X_2, \dots, X_{i-1}) \\ &\quad - \text{H}(X_i \mid A, X_1, X_2, \dots, X_{i-1}). \end{aligned} \quad (1)$$

Fix a coordinate i within the sum in this equation. By independence, the first expression: $\text{H}(X_i \mid X_1, X_2, \dots, X_{i-1}) = \text{H}(X_i) = \log|\mathcal{U}|$. For the second expression, fix an element $a \in \mathcal{U}$ and let \mathbf{Y}_a denote $(X_1, X_2, \dots, X_{i-1}, a, \perp, \dots, \perp)$. Note that when Alice's input is \mathbf{X} , the input that Bob is holding is *exactly* \mathbf{Y}_a for some i and a . Let $B(A, \mathbf{Y}_a)$ denote Bob's output on Alice's message A . Then,

$$\Pr[B(A, \mathbf{Y}_a) = 1 \mid X_i = a] \geq 1 - \delta$$

and for every $a' \neq a$,

$$\Pr[B(A, \mathbf{Y}_{a'}) = 0 \mid X_i = a] \geq 1 - \delta$$

Therefore, by the union bound,

$$\Pr\left[B(A, \mathbf{Y}_a) = 1 \wedge \bigwedge_{a' \neq a} B(A, \mathbf{Y}_{a'}) = 0 \mid X_i = a\right]$$

is at least $1 - \delta|\mathcal{U}| \geq \frac{3}{4}$. Thus, there is a predictor for X_i using X_1, X_2, \dots, X_{i-1} and A with failure probability at most $1/4$. By Fano's inequality,

$$\begin{aligned} \text{H}(X_i \mid A, X_1, X_2, \dots, X_{i-1}) &\leq \frac{1}{4} \log(|\mathcal{U}| - 1) + h_2\left(\frac{1}{4}\right) \\ &\leq \frac{1}{2} \log(|\mathcal{U}|), \end{aligned}$$

since $|\mathcal{U}|$ is sufficiently large. Substituting in (1), we conclude

$$\text{I}(\mathbf{X} : A) \geq \frac{N \log(|\mathcal{U}|)}{2}. \quad \square$$

COROLLARY 3.2. *Let $|\mathcal{U}| = 1/4\delta$. Then $R_{\delta}^{\rightarrow}(\text{IND}_{\mathcal{U}}^a) = \Omega(N \log 1/\delta)$.*

Remark 3.3. Consider a variant of $\text{IND}_{\mathcal{U}}^a$ where for the index i of interest, Bob does not get to see all of the prefix x_1, x_2, \dots, x_{i-1} of \mathbf{x} . Instead, for every such i , there is a subset $J_i \subseteq [i-1]$ depending on i such that he gets to see only x_k for $k \in J_i$. In this case, he has even less information than what he had for $\text{IND}_{\mathcal{U}}^a$ so every protocol for this problem is also a protocol for $\text{IND}_{\mathcal{U}}^a$. Therefore, the one-way communication lower bound of Corollary 3.2 holds for this variant.

Remark 3.4. Now, consider the standard indexing problem IND where Bob gets an index i and a single element y , and the goal is to determine whether $x_i = y$. This is equivalent to the setting of the previous remark where $J_i = \emptyset$ for every i . The proof of Theorem 3.1 can be adapted to show that $R_\delta^{\rightarrow, \parallel}(\text{IND}) = \Omega(N \log 1/\delta)$ for $|\mathcal{U}| = \frac{1}{8\delta}$. Let μ be the distribution where Alice gets \mathbf{X} uniformly chosen in \mathcal{U}^N and Bob's input (I, Y) is uniformly chosen in $[N] \times \mathcal{U}$. As in the proof of the theorem, let A be the message sent by Alice on input \mathbf{X} . Let δ_i denote the expected error of the protocol conditioned on $I = i$. By an averaging argument, for at least half the indices i , $\delta_i \leq 2\delta$. Fix such an i . Look at the last expression bounding the information cost in (1). Using $H(X_i | A, X_1, X_2, \dots, X_{i-1}) \leq H(X_i | A)$ and then proceeding as before, there exists an estimator β_i such that

$$\Pr[\beta_i(A) \neq X_i | I = i] \leq |\mathcal{U}|\delta_i \leq 2|\mathcal{U}|\delta \leq \frac{1}{4},$$

implying that $I(X_i : A) \geq (1/2) \log(|\mathcal{U}|)$. The lower bound follows since there are at least $N/2$ such indices.

3.1. An Encoding Scheme

Let $\Delta(\mathbf{x}, \mathbf{y}) \stackrel{\text{def}}{=} |\{i \mid x_i \neq y_i\}|$ denote the Hamming distance between two vectors \mathbf{x}, \mathbf{y} over some domain. We present an encoding scheme that transforms the inputs of $\text{IND}_{\mathcal{U}}^{\text{a}}$ into well-crafted gap instances of the Hamming distance problem. This will be used in the applications to follow.

LEMMA 3.5. *Consider the problem $\text{IND}_{\mathcal{U}}^{\text{a}}$ on length $N = bm$, where $m = \frac{1}{4\epsilon^2}$ is odd and b is some parameter. Let $\alpha \geq 2$ be an integer. Let (\mathbf{x}, \mathbf{y}) be a promise input to the problem to $\text{IND}_{\mathcal{U}}^{\text{a}}$. Then there exist encoding functions $\mathbf{x} \rightsquigarrow \mathbf{u} \in \{0, 1\}^n$ and $\mathbf{y} \rightsquigarrow \mathbf{v} \in \{0, 1\}^n$, where $n = O(\alpha^b \cdot \frac{1}{\epsilon^2} \cdot \log 1/\delta)$, depending on a shared random string \mathbf{s} that satisfy the following: suppose the index i (which is determined by \mathbf{y}) for which the players need to determine whether $x_i = y_i$ belongs to $[(p-1)m+1, pm]$, for some p . Then, \mathbf{u} can be written as $(\mathbf{u}_1, \mathbf{u}_2, \mathbf{u}_3) \in \{0, 1\}^{n_1} \times \{0, 1\}^{n_2} \times \{0, 1\}^{n_3}$ and \mathbf{v} as $(\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3) \in \{0, 1\}^{n_1} \times \{0, 1\}^{n_2} \times \{0, 1\}^{n_3}$ such that:*

- (1) $n_2 = n \cdot \alpha^{-p}(\alpha - 1)$ and $n_3 = n \cdot \alpha^{-p}$;
- (2) each of the \mathbf{u}_i 's and \mathbf{v}_i 's have exactly half of their coordinates set to 1;
- (3) $\Delta(\mathbf{u}_1, \mathbf{v}_1) = 0$ and $\Delta(\mathbf{u}_3, \mathbf{v}_3) = n_3/2$;
- (4) if (\mathbf{x}, \mathbf{y}) is a no instance, then with probability at least $1 - \delta$,

$$\Delta(\mathbf{u}_2, \mathbf{v}_2) \geq n_2(\frac{1}{2} - \frac{\epsilon}{3});$$

- (5) if (\mathbf{x}, \mathbf{y}) is a yes instance, then with probability at least $1 - \delta$,

$$\Delta(\mathbf{u}_2, \mathbf{v}_2) \leq n_2(\frac{1}{2} - \frac{2\epsilon}{3}).$$

PROOF. We first define and analyze a basic encoding scheme. Let $\mathbf{w} \in \mathcal{U}^m$. Let $s : \mathcal{U}^m \rightarrow \{-1, +1\}^m$ be a random hash function defined by picking m random hash functions $s^1, \dots, s^m : \mathcal{U} \rightarrow \{-1, 1\}$ and then setting $s = (s^1, \dots, s^m)$. We define $\text{enc}_1(\mathbf{w}, s)$ to be the majority of the ± 1 values in the m components of $s(\mathbf{w})$. This is well defined since m is odd. We contrast this with another encoding defined with an additional parameter $j \in [m]$. Define $\text{enc}_2(\mathbf{w}, j, s)$ to be just the j th component of $s(\mathbf{w})$.

To analyze this scheme, fix two vectors $\mathbf{w}, \mathbf{z} \in \mathcal{U}^m$ and an index j . If $w_j \neq z_j$, then

$$\Pr[\text{enc}_1(\mathbf{w}, s) \neq \text{enc}_2(\mathbf{z}, j, s)] = \frac{1}{2}.$$

On the other hand, suppose $w_j = z_j$. Then, by a standard argument involving the binomial coefficients,

$$\Pr[\text{enc}_1(\mathbf{w}, \mathbf{s}) \neq \text{enc}_2(\mathbf{z}, j, \mathbf{s})] \leq \frac{1}{2} \left(1 - \frac{1}{2\sqrt{m}}\right) = \frac{1}{2} - \varepsilon.$$

We repeat this scheme to amplify the gap between the two cases. Let $\mathbf{s} = (s_1, s_2, \dots, s_k)$ be a collection of $k = \frac{10}{\varepsilon^2} \cdot \log 1/\delta$ independent and identically distributed random hash functions each mapping \mathcal{U}^m to $\{-1, +1\}^m$. Define

$$\text{enc}_1(\mathbf{w}, \mathbf{s}) = (\text{enc}_1(\mathbf{w}, s_1), \text{enc}_1(\mathbf{w}, s_2), \dots, \text{enc}_1(\mathbf{w}, s_k)),$$

and

$$\text{enc}_2(\mathbf{z}, j, \mathbf{s}) = (\text{enc}_1(\mathbf{z}, j, s_1), \dots, \text{enc}_2(\mathbf{z}, j, s_k)),$$

For ease of notation, let $\mathbf{w}' = \text{enc}_1(\mathbf{w}, \mathbf{s})$ and $\mathbf{z}' = \text{enc}_2(\mathbf{z}, j, \mathbf{s})$.

FACT 3.6. *Let X_1, X_2, \dots, X_k be a collection of independent and identically distributed 0-1 Bernoulli random variables each with probability of equaling 1 equal to p . Set $\bar{X} = \sum_i X_i/k$. Then,*

$$\begin{aligned} \Pr[\bar{X} < p - h] &< \exp(-2h^2k), \text{ and} \\ \Pr[\bar{X} > p + h] &< \exp(-2h^2k). \end{aligned}$$

In this fact, with $k = 10\varepsilon^{-2} \log 1/\delta$ and $h = \varepsilon/3$, we obtain that the tail probability is at most δ . In the case $w_j \neq z_j$, we have $p = \frac{1}{2}$, so

$$\Pr[\Delta(\mathbf{w}', \mathbf{z}') < k(\frac{1}{2} - \frac{\varepsilon}{3})] \leq \delta. \quad (2)$$

In the second case, $p = \frac{1}{2} - \varepsilon$,

$$\Pr[\Delta(\mathbf{w}', \mathbf{z}') > k(\frac{1}{2} - \frac{2\varepsilon}{3})] \leq \delta. \quad (3)$$

The two cases differ by a factor of at least $1 + \varepsilon/3$ for ε less than a small enough constant.

Divide $[N]$ into b blocks where the q th block equals $[(q-1)m+1, qm]$ for every $q \in [b]$. We use this to define an encoding for promise inputs (\mathbf{x}, \mathbf{y}) to the problem $\text{IND}_{\mathcal{U}}^a$, where the goal is to decide for an index i belonging to block p whether $x_i = y_i$. Let $j = i - (p-1)m$ denote the offset of i within block p . We also think of \mathbf{x} and \mathbf{y} as being analogously divided into b blocks $\mathbf{x}_{[1]}, \mathbf{x}_{[2]}, \dots, \mathbf{x}_{[b]}$ and $\mathbf{y}_{[1]}, \mathbf{y}_{[2]}, \dots, \mathbf{y}_{[b]}$ respectively. Thus, the goal is to decide whether the j th components of $\mathbf{x}_{[p]}$ and $\mathbf{y}_{[p]}$ are equal.

Fix a block index q . Let $\mathbf{s}_{[q]}$ denote a vector of k independent and identically distributed random hash functions corresponding to block q . Compute $\text{enc}_1(\mathbf{x}_{[q]}, \mathbf{s}_{[q]})$ and then repeat each coordinate of this vector α^{b-q} times. Call the resulting vector $\mathbf{x}'_{[q]}$. For $\mathbf{y}_{[q]}$, the encoding $\mathbf{y}'_{[q]}$ depends on the relationship of q to p and additionally on j (both p and j are determined by \mathbf{y}). If $q < p$, we use the same encoding function as that for $\mathbf{x}_{[q]}$, that is, $\text{enc}_1(\mathbf{y}_{[q]}, \mathbf{s}_{[q]})$ repeated α^{b-q} times. If $q > p$, the encoding is a 0 vector of length $\alpha^{b-q} \cdot k$. If $q = p$, the encoding equals $\text{enc}_2(\mathbf{y}_{[p]}, j, \mathbf{s}_{[p]})$ using the second encoding function, again repeated α^{b-p} times. For each q , the lengths of both $\mathbf{x}'_{[q]}$ and $\mathbf{y}'_{[q]}$ equal $\alpha^{b-q} \cdot k$. Finally, define a dummy vector $\mathbf{x}_{[b+1]}$ of length $k/(\alpha-1)$ all of whose components equal 1, and another dummy vector $\mathbf{y}_{[b+1]}$ of the same length all of whose

components equal 0. We define the encoding $\mathbf{x} \rightsquigarrow \mathbf{u}$ to be the concatenation of all $\mathbf{x}'_{[q]}$ for all $1 \leq q \leq b + 1$. Similarly for $\mathbf{y} \rightsquigarrow \mathbf{v}$. The encodings have length

$$\begin{aligned} n &= k/(\alpha - 1) + \sum_{1 \leq q \leq b} \alpha^{b-q} \cdot k \\ &= \alpha^b \cdot k/(\alpha - 1) \\ &= O(\alpha^b \cdot \frac{1}{\epsilon^2} \cdot \log 1/\delta). \end{aligned}$$

Moreover, the values are in $\{-1, 0, +1\}$ but a simple fix to be described at the end will transform this into a 0-1 vector.

We now define the split $\mathbf{u} = (\mathbf{u}_1, \mathbf{u}_2, \mathbf{u}_3)$ and $\mathbf{v} = (\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3)$. Define \mathbf{u}_1 (respectively, $\mathbf{u}_2, \mathbf{u}_3$) to be the concatenation of all $\mathbf{x}'_{[q]}$ for $q < p$ (respectively, $q = p, q > p$). Define \mathbf{v}_c for $c = 1, 2, 3$ analogously.

First, note that $\mathbf{u}_1 = \mathbf{v}_1$ because $\mathbf{x}'_{[q]} = \mathbf{y}'_{[q]}$ for $q < p$. Next, the lengths of \mathbf{u}_3 and \mathbf{v}_3 equal

$$\begin{aligned} k/(\alpha - 1) + \sum_{p+1 \leq q \leq b} \alpha^{b-q} \cdot k &= \alpha^{b-p} \cdot k/(\alpha - 1) \\ &= n \cdot \alpha^{-p} = n_3. \end{aligned}$$

Since \mathbf{u}_3 is a ± 1 vector while \mathbf{v}_3 is a 0 vector, $\|\mathbf{u}_3 - \mathbf{v}_3\|_1 = n_3$. Last, we look at \mathbf{u}_2 and \mathbf{v}_2 . Their lengths equal $\alpha^{b-p} \cdot k = n \cdot \alpha^{-p}(\alpha - 1) = n_2$. We now analyze $\|\mathbf{u}_2 - \mathbf{v}_2\|_1 = 2\Delta(\mathbf{x}'_{[p]}, \mathbf{y}'_{[p]})$. We distinguish between the *yes* and *no* instances via (2) and (3). For a *no* instance, $x_i \neq y_i$, so by (2), with probability at least $1 - \delta$,

$$\|\mathbf{u}_2 - \mathbf{v}_2\|_1 \geq 2n_2(\frac{1}{2} - \frac{\epsilon}{3}).$$

For a *yes* instance, a similar calculation using (3) shows that with probability at least $1 - \delta$,

$$\|\mathbf{u}_2 - \mathbf{v}_2\|_1 \leq 2n_2(\frac{1}{2} - \frac{2\epsilon}{3}).$$

To obtain the required 0-1 vectors, apply a simple transformation of $\{-1 \rightarrow 0101, 0 \rightarrow 0011, +1 \rightarrow 1010\}$ to \mathbf{u} and \mathbf{v} . This produces 0-1 inputs having a relative Hamming weight of exactly half in each of the \mathbf{u}_i 's and \mathbf{v}_i 's. The length quadruples while a norm distance of d translates to a Hamming distance of $2d$, which translates to the bounds stated in the lemma. \square

4. APPLICATIONS

Throughout we assume that $n^{1-\gamma} \geq \frac{1}{\epsilon^2} \log 1/\delta$ for an arbitrarily small constant $\gamma > 0$. For several of the applications in this section, the bounds will be stated in terms of communication complexity that can be translated naturally to memory lower bounds for analogous streaming problems.

4.1. Approximating the Hamming Distance

Consider the problem HAM where Alice gets $\mathbf{x} \in \{0, 1\}^n$, Bob gets $\mathbf{y} \in \{0, 1\}^n$, and their goal is to produce a $1 \pm \epsilon$ -approximation of $\Delta(x, y)$.

THEOREM 4.1. $R_\delta^{\rightarrow}(\text{HAM}) = \Omega(\frac{1}{\epsilon^2} \cdot \log n \cdot \log 1/\delta)$

PROOF. We reduce IND_U^a to HAM using the encoding given in Lemma 3.5 with $\alpha = 2$ so that $n_2 = n_3 = n \cdot 2^{-p}$. With probability at least $1 - \delta$, the *yes* instances are encoded

to have Hamming distance at most $n \cdot 2^{-p}(1 - \frac{2\varepsilon}{3})$ while the *no* instances have distance at least $n \cdot 2^{-p}(1 - \frac{\varepsilon}{3})$. Their ratio is at least $1 + \varepsilon/3$. Using a protocol for HAM with approximation factor $1 + \varepsilon/3$ and failure probability δ , we can distinguish the two cases with probability at least $1 - 2\delta$.

Since we assume that $\frac{1}{\varepsilon^2} \cdot \log 1/\delta < n^{1-\gamma}$ for a constant $\gamma > 0$, we can indeed set $b = \Omega(\log n)$, as needed here to fit the vectors into n coordinates. Now, apply Corollary 3.2 to finish the proof. \square

4.2. Estimating ℓ_p -Distances

Since $\Delta(\mathbf{x}, \mathbf{y}) = \|\mathbf{x} - \mathbf{y}\|_p^p$, for $\mathbf{x}, \mathbf{y} \in \{0, 1\}^n$, Theorem 4.1 immediately yields the following for any constant p .

THEOREM 4.2. *The one-way communication complexity of the problem of approximating the $\|\cdot\|_p$ difference of two vectors of length n to within a factor $1 + \varepsilon$ with failure probability at most δ is $\Omega(\frac{1}{\varepsilon^2} \cdot \log n \cdot \log 1/\delta)$.*

4.3. JL Transforms

Recall that n is the dimension of the vectors we seek to significantly reduce using the JL transform. We have the following lower bound.

THEOREM 4.3. *Any (ε, δ) -JLT (F, μ) has dimension $\Omega(\frac{1}{\varepsilon^2} \log 1/\delta)$.*

PROOF. The public-coin one-way communication complexity, that is, the one-way communication complexity in which the parties additionally share an infinitely long random string and denoted $R_\delta^{\rightarrow, \text{pub}}$, is at least $R_\delta^{\rightarrow} - O(\log I)$, where I is the sum of input lengths to the two parties [Kremer et al. 1999]. By Theorem 4.2,

$$\begin{aligned} R_\delta^{\rightarrow, \text{pub}}(\ell_2) &= \Omega\left(\frac{1}{\varepsilon^2} \log n \log 1/\delta\right) - O(\log n) \\ &= \Omega\left(\frac{1}{\varepsilon^2} \log n \log 1/\delta\right). \end{aligned}$$

Consider the following public-coin protocol for ℓ_2 . The parties use the public-coin to agree upon a $k \times n$ matrix A sampled from \mathcal{F} according to μ . Alice computes $A\mathbf{x}$, rounds each entry to the nearest additive multiple of $\varepsilon/(2\sqrt{k})$, and send the rounded vector $\tilde{A}\mathbf{x}$ to Bob. Bob then computes $A\mathbf{y}$, and outputs $\|\tilde{A}\mathbf{x} - A\mathbf{y}\|$. By the triangle inequality,

$$\begin{aligned} \|A\mathbf{y} - A\mathbf{x}\| - \|\tilde{A}\mathbf{x} - A\mathbf{x}\| &\leq \|\tilde{A}\mathbf{x} - A\mathbf{y}\| \\ &\leq \|A\mathbf{y} - A\mathbf{x}\| + \|\tilde{A}\mathbf{x} - A\mathbf{x}\|, \end{aligned}$$

or using the definition of $\tilde{A}\mathbf{x}$,

$$\|A\mathbf{y} - A\mathbf{x}\| - \frac{\varepsilon}{2} \leq \|\tilde{A}\mathbf{x} - A\mathbf{y}\| \leq \|A\mathbf{y} - A\mathbf{x}\| + \frac{\varepsilon}{2}.$$

With probability $\geq 1 - \delta$, we have $\|A(\mathbf{y} - \mathbf{x})\|^2 = (1 \pm \varepsilon)\|\mathbf{y} - \mathbf{x}\|^2$, or $\|A\mathbf{y} - A\mathbf{x}\| = (1 \pm \varepsilon/2)\|\mathbf{y} - \mathbf{x}\|$. Using that $\|\mathbf{y} - \mathbf{x}\| \geq 1$ in Theorem 4.2 if $\|\mathbf{y} - \mathbf{x}\| \neq 0$, we have $\|\tilde{A}\mathbf{x} - A\mathbf{y}\| = (1 \pm \varepsilon)\|\mathbf{x} - \mathbf{y}\|$. Hence,

$$kB = \Omega\left(\frac{1}{\varepsilon^2} \log n \log 1/\delta\right),$$

where B is the maximum number of bits needed to describe an entry of $\tilde{A}\mathbf{x}$. With probability at least $1 - \delta$, $\|A\mathbf{x}\|^2 = (1 \pm \varepsilon)\|\mathbf{x}\|^2$, and so using that $\mathbf{x} \in \{0, 1\}^n$, no entry of $A\mathbf{x}$ can be larger than $2n$. By rescaling δ by a constant, this event also occurs, and so $B = O(\log n + \log 1/\varepsilon + \log k)$. Since we assume that $n \geq \frac{1}{\varepsilon^2} \log n \log 1/\delta$, we have $B = O(\log n)$, and so $k = \Omega\left(\frac{1}{\varepsilon^2} \log 1/\delta\right)$, finishing the proof. \square

4.4. Estimating Distinct Elements and Related Problems

We improve the lower bound for estimating the number F_0 of distinct elements in an insertion-only data stream up to a $(1 \pm \varepsilon)$ -factor with probability at least $1 - \delta$. We let n be the universe size, that is, the total possible number of distinct elements.

THEOREM 4.4. *Any 1-pass streaming algorithm that outputs a $(1 \pm \varepsilon)$ -approximation to F_0 in an insertion-only stream with probability at least $1 - \delta$ must use $\Omega(\varepsilon^{-2} \log 1/\delta + \log n)$ bits of space.*

Remark 4.5. This improves the previous $\Omega(\varepsilon^{-2} + \log n)$ lower bound of Alon et al. [1999], Indyk and Woodruff [2003], and Woodruff [2004].

PROOF. It is enough to show an $\Omega\left(\frac{1}{\varepsilon^2} \cdot \log 1/\delta\right)$ bound since the $\Omega(\log n)$ bound is in Alon et al. [1999]. We reduce $\text{IND}_{\mathcal{U}}^{\alpha}$ to approximating F_0 in a stream. Apply Lemma 3.5 with $\alpha = 2$ and $b = 1$ to obtain \mathbf{u} and \mathbf{v} of length $k = O\left(\frac{1}{\varepsilon^2} \cdot \log 1/\delta\right)$. With $b = p = 1$, with probability at least $1 - \delta$, the Hamming distance for *no* instances is at least $\frac{k}{2}\left(1 - \frac{\varepsilon}{3}\right)$ while for the *yes* instances it is at most $\frac{k}{2}\left(1 - \frac{2\varepsilon}{3}\right)$.

Alice inserts a token i corresponding to each i such that $u_i = 1$. Bob does the same with respect to v_i . Since the Hamming weights of \mathbf{u} and \mathbf{v} are exactly half, by a simple calculation, $2F_0 = \Delta(\mathbf{u}, \mathbf{v}) + k$. Thus, there is a multiplicative gap of at least $1 + \Theta(\varepsilon)$. \square

We note that distinct elements is used as a subroutine in geometric problems in a data stream, such as approximating the Klee's measure. In this problem there is a stream of axis-aligned boxes given on the d -dimensional grid and one wants to approximate the number of points in the union of the boxes. It was recently shown that for constant d , this can be $(1 + \varepsilon)$ -approximated with probability $1 - \delta$ in $\text{poly}(\varepsilon^{-1} \log(n\Delta)) \log 1/\delta$ bits of space and using $\text{poly}(\varepsilon^{-1} \log(n\Delta)) \log 1/\delta$ time to process each arriving box, where $[\Delta]^d$ is the input grid [Tirthapura and Woodruff 2011]. Notice that if time is not taken into consideration, this is just an F_0 -approximation, where for each input box, one inserts the points one by one into a streaming algorithm for approximating F_0 . Conversely, estimating the Klee's measure is at least as hard as estimating F_0 , since F_0 is a special case of the problem when $d = 1$. By Theorem 4.4, we obtain an $\Omega(\varepsilon^{-2} \log 1/\delta)$ lower bound for estimating the Klee's measure, improving the previous $\Omega(\varepsilon^{-2})$ lower bound that came from previous lower bounds for estimating F_0 .

We now turn to two well-studied specializations of the distinct elements problem, the max-dominance norm and the distinct summation.

The max-dominance norm is a useful measure in finance and IP traffic monitoring [Cormode and Muthukrishnan 2003]. Alice has $\mathbf{x} \in \{0, 1, \dots, \text{poly}(n)\}^n$, Bob has $\mathbf{y} \in \{0, 1, \dots, \text{poly}(n)\}^n$, and the max-dominance norm is defined to be $\sum_{j=1}^n \max(\mathbf{x}_j, \mathbf{y}_j)$. The problem of estimating this quantity is investigated in Cormode and Muthukrishnan [2003], Pavan and Tirthapura [2007], Stoev et al. [2007], Stoev and Taqqu [2010], Sun and Poon [2009], and Woodruff [2011].

COROLLARY 4.6. *The one-way communication complexity of the problem of approximating the max-dominance norm is $\Omega(\frac{1}{\epsilon^2} \cdot \log 1/\delta + \log n)$.*

PROOF. The proof of the $\Omega(\frac{1}{\epsilon^2} \cdot \log 1/\delta)$ bound is the same as that in Theorem 4.4, noting that for bit vectors \mathbf{x} and \mathbf{y} , $\sum_{j=1}^n \max(\mathbf{x}_j, \mathbf{y}_j) = F_0$, where, here, F_0 refers to the number of distinct elements in the stream in which Alice inserts a token j corresponding to each j for which $\mathbf{x}_j = 1$, while Bob inserts a token j corresponding to each j for which $\mathbf{y}_j = 1$.

To show the $\Omega(\log n)$ lower bound, we use the well-known fact that the randomized two-party communication complexity of equality on binary strings of length n is $\Omega(\log n)$ [Kushilevitz and Nisan 1997]. Moreover, this continues to hold under the promise that either $\mathbf{x} = \mathbf{y}$ or $\Delta(\mathbf{x}, \mathbf{y}) = n/2$, where \mathbf{x} and \mathbf{y} are Alice and Bob's input, respectively [Buhrman et al. 1998]. We refer to the version of equality with this problem as Gap-Equality, which is equivalent to testing if $\Delta(\mathbf{x}, \mathbf{y}) = 0$ or $\Delta(\mathbf{x}, \mathbf{y}) = n/2$.

Observe that for binary vectors \mathbf{x} and \mathbf{y} ,

$$\sum_{j=1}^n \max(\mathbf{x}_j, \mathbf{y}_j) = \frac{wt(\mathbf{x}) + wt(\mathbf{y}) + \Delta(\mathbf{x}, \mathbf{y})}{2}, \quad (4)$$

since if a coordinate is 1 in either x or y , it contributes one to both sides of the equality, whereas if it occurs zero times it contributes zero to both sides of the equality. Equivalently,

$$\Delta(\mathbf{x}, \mathbf{y}) = 2 \sum_{j=1}^n \max(\mathbf{x}_j, \mathbf{y}_j) - wt(\mathbf{x}) - wt(\mathbf{y}).$$

In the protocol for Gap-Equality, Alice sends a $(1 + 1/15)$ -approximation $\tilde{wt}(\mathbf{x})$ to $wt(\mathbf{x})$ to Bob, which can be done using $C \log \log n$ bits for a constant $C > 0$. Bob computes $wt(\mathbf{y})$ exactly. Alice also sends her message in the protocol for $(1 + 1/15)$ -approximating the max-dominance norm to Bob. From this, Bob computes his estimate Φ of the max-dominance norm.

Since $\tilde{wt}(\mathbf{x})$ and Φ are $(1 + 1/15)$ -approximations to $wt(\mathbf{x})$ and the max-dominance norm, respectively, each of which is at most n , we have that

$$\tilde{\Delta}(\mathbf{x}, \mathbf{y}) = 2\Phi - \tilde{wt}(\mathbf{x}) - wt(\mathbf{y}) = \Delta(\mathbf{x}, \mathbf{y}) \pm \frac{n}{5}.$$

It follows that $\tilde{\Delta}(\mathbf{x}, \mathbf{y})$ can be used to distinguish the case $\Delta(\mathbf{x}, \mathbf{y}) = 0$ from the case $\Delta(\mathbf{x}, \mathbf{y}) = n/2$. Hence, the length of Alice's message must be $\Omega(\log n) - C \log \log n = \Omega(\log n)$. This proves the $\Omega(\log n)$ bound. \square

We also get an optimal lower bound for the distinct summation problem [Pavan and Tirthapura 2007; Woodruff 2011]. Here, for each $j \in [n]$ there is an integer $v_j \in \{1, \dots, \text{poly}(n)\}$. Alice's input is an n -dimensional vector \mathbf{x} with $\mathbf{x}_j \in \{0, v_j\}$, while Bob's input is an n -dimensional vector \mathbf{y} with $\mathbf{y}_j \in \{0, v_j\}$. The problem is to compute the expression

$$\sum_j v_j \cdot \delta(\text{either } \mathbf{x}_j = v_j \text{ or } \mathbf{y}_j = v_j),$$

where δ is the indicator function, in this case indicating that either $\mathbf{x}_j = v_j$ or $\mathbf{y}_j = v_j$ (or both).

COROLLARY 4.7. *The one-way communication complexity of the problem of approximating distinct summation is $\Omega(\frac{1}{\epsilon^2} \cdot \log 1/\delta + \log n)$.*

PROOF. We consider the very simple case in which $v_j = 1$ for all j . In this case the value of distinct summation equals the max-dominance norm between binary vectors \mathbf{x} and \mathbf{y} , where $\mathbf{x}_j = 1$ (respectively, $\mathbf{y}_j = 1$) if Alice (respectively, Bob) has $(j, 1)$ and $\mathbf{x}_j = 0$ (respectively, $\mathbf{y}_j = 0$) if Alice (respectively, Bob) has $(j, 0)$. Conversely, any binary vectors \mathbf{x} and \mathbf{y} give rise to the equivalent distinct summation problem. The corollary now follows by Theorem 4.6, since the proof of that theorem holds even for binary vectors \mathbf{x} and \mathbf{y} . \square

Both the max-dominance norm and the distinct summation problem are special cases of estimating F_0 in a data stream. Indeed, for the max-dominance norm, we can define n disjoint intervals I_j of $\text{poly}(n)$ numbers. We insert the first \mathbf{x}_j items of I_j and first \mathbf{y}_j items of I_j into a stream, so that $\max(\mathbf{x}_j, \mathbf{y}_j)$ is equal to the number of distinct items inserted from I_j . Similarly, for distinct summation, we create disjoint intervals I_j of $\text{poly}(n)$ numbers for each $j \in [n]$. We either insert the first v_j items from I_j into the stream, or 0 items, depending on whether the pair (j, v_j) or $(j, 0)$ occurs in the stream. It follows by using the F_0 algorithm of Kane et al. [2010b], we can achieve $O(\varepsilon^{-2} \log 1/\delta + \log n \log 1/\delta)$ bits of space.

4.5. Estimating Entropy

Our technique improves the lower bound for additively estimating the entropy of a stream. To capture this, the entropy difference of two n -dimensional vectors x and y is the problem of computing

$$H(\mathbf{x}, \mathbf{y}) = \sum_{i=1}^n \frac{|x_i - y_i|}{\|\mathbf{x} - \mathbf{y}\|_1} \log_2 \frac{\|\mathbf{x} - \mathbf{y}\|_1}{|x_i - y_i|}.$$

As usual with entropy, if $x_i - y_i = 0$, the corresponding term in this sum is 0.

THEOREM 4.8. *The one-way communication complexity of the problem of estimating the entropy difference up to an additive ε with probability at least $1 - \delta$ is $\Omega(\varepsilon^{-2} \log n \log 1/\delta)$.*

Remark 4.9. This improves the previous $\Omega(\varepsilon^{-2} \log n)$ lower bound implied by the work of Kane et al. [2010a].

PROOF. We reduce from HAM. Since the input vectors \mathbf{x}, \mathbf{y} to HAM are in $\{0, 1\}^n$, $\|\mathbf{x} - \mathbf{y}\|_1 = \Delta(x, y)$. Also, if $x_i = y_i$, then the contribution to the entropy is 0. Otherwise, the contribution is $\frac{\log_2 \Delta(x, y)}{\Delta(x, y)}$. Hence, $H(\mathbf{x}, \mathbf{y}) = \log_2 \Delta(x, y)$, or $\Delta(x, y) = 2^{H(\mathbf{x}, \mathbf{y})}$. Given an approximation $\tilde{H}(\mathbf{x}, \mathbf{y})$ such that $|\tilde{H}(\mathbf{x}, \mathbf{y}) - H(\mathbf{x}, \mathbf{y})| \leq \varepsilon$ with probability at least $1 - \delta$,

$$\begin{aligned} (1 - \Theta(\varepsilon))\Delta(x, y) &\leq 2^{-\varepsilon} \Delta(x, y) \\ &\leq 2^{\tilde{H}(\mathbf{x}, \mathbf{y})} \\ &\leq 2^\varepsilon \Delta(x, y) \\ &\leq (1 + \Theta(\varepsilon))\Delta(x, y). \end{aligned}$$

so one obtains a $(1 \pm \Theta(\varepsilon))$ -approximation to HAM with the same probability. The lower bound now follows from Theorem 4.1. \square

Entropy estimation has also been studied in the strict turnstile model of streaming, in which one has a stream of tokens that can be inserted or deleted, and the number of tokens of a given type at any point in the stream is nonnegative. We can show an $\Omega(\varepsilon^{-2} \log n \log 1/\delta / \log 1/\varepsilon)$ lower bound as follows.

We apply Lemma 3.5 with $\alpha = \varepsilon^{-2}$, $b = O(\log n / \log(1/\varepsilon))$ to obtain \mathbf{u} and \mathbf{v} . For each coordinate i in \mathbf{u} , Alice inserts a token i if the value at the coordinate equals 0 and a token of $n + i$ if the value equals 1. Let $\mathbf{u} = (\mathbf{u}_1, \mathbf{u}_2, \mathbf{u}_3)$ and $\mathbf{v} = (\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3)$. Bob can compute the split of \mathbf{v} because he can compute n_1, n_2 , and n_3 based on p (which itself depends on i). Bob deletes all the tokens corresponding to coordinates in \mathbf{u}_1 , which is possible because $\mathbf{v}_1 = \mathbf{u}_1$. For coordinates in \mathbf{v}_2 he mimics Alice's procedure, that is, a token i for 0 and a token $n + i$ for 1. For \mathbf{v}_3 , he does nothing. The number of tokens equal $n_3 + 2n_2 = n \cdot \varepsilon^{-2p}(2\varepsilon^{-2} - 1)$. The tokens corresponding to \mathbf{u}_3 appear exactly once. For every coordinate where \mathbf{u}_2 and \mathbf{v}_2 differ, the stream consists of 2 distinct tokens, whereas for each of the remaining coordinates the stream consists of a token appearing twice. Therefore, number of tokens appearing exactly once equals $n_3 + 2\Delta(\mathbf{u}_2, \mathbf{v}_2) = n\varepsilon^{-2p} + 2\Delta(\mathbf{u}_2, \mathbf{v}_2)$. The number of tokens appearing twice equals $n_2 - \Delta(\mathbf{u}_2, \mathbf{v}_2) = n \cdot \varepsilon^{-2p}(\varepsilon^{-2} - 1) - \Delta(\mathbf{u}_2, \mathbf{v}_2)$. In the setting of Theorem A.3 of Kane et al. [2010a], if $\Delta = \Delta(\mathbf{u}_2, \mathbf{v}_2)$, then the entropy H satisfies

$$\Delta = \frac{HR}{2B} + C - C \log R - \frac{A}{2B} \log R,$$

where

$$\begin{aligned} A &= n\varepsilon^{-2p}, \\ B &= n\varepsilon^{-2p+2}(\varepsilon^{-2} - 1), \\ C &= \varepsilon^{-2}, \\ R &= A + 2BC. \end{aligned}$$

Notice that A, B, C , and R are known to Bob. Thus, to decide whether Δ is small or large, it suffices to have a $\frac{1}{\varepsilon}$ -additive approximation to $HR/(2B)$, or since $B/R = \Theta(\varepsilon^2)$, it suffices to have an additive $\Theta(\varepsilon)$ -approximation to H with probability at least $1 - \delta$. The theorem follows by applying Corollary 3.2.

THEOREM 4.10. *Any 1-pass streaming algorithm that outputs an additive ε -approximation to the entropy in the strict turnstile model with probability at least $1 - \delta$ must use $\Omega(\varepsilon^{-2} \log n \log 1/\delta / \log(1/\varepsilon))$ bits of space.*

Remark 4.11. This improves the previous $\Omega(\varepsilon^{-2} \log n / \log 1/\varepsilon)$ lower bound of Kane et al. [2010a].

4.6. VC-Dimension and Subconstant Error

Recall that the VC-dimension $VC(f)$ of the communication matrix for a binary function f is the maximum number ℓ of columns for which all 2^ℓ possible bit patterns occur in the rows of the matrix restricted to those columns. Kremer et al. [1999], show the surprising result that the VC-dimension $VC(f)$ of the communication matrix for f exactly characterizes $R_{1/3}^{\rightarrow, \parallel}(f)$.

THEOREM 4.12. [KREMER ET AL. 1999]. $R_{1/3}^{\rightarrow, \parallel}(f) = \Theta(VC(f))$.

We show that for subconstant error probabilities δ , the VC-dimension does not capture $R_\delta^{\rightarrow, \parallel}(f)$.

THEOREM 4.13. *There exist problems f, g for which $VC(f) = VC(g) = N$, yet*

$$\begin{aligned} - R_\delta^{\rightarrow, \parallel}(f) &= \Theta(N). \\ - R_\delta^{\rightarrow, \parallel}(g) &= \Theta(N \log 1/\delta). \end{aligned}$$

PROOF. For f , we take the Indexing function. Namely, Alice is given $x \in \{0, 1\}^N$, Bob is given $i \in [N]$, and $f(x, i) = x_i$. It is easy to see that $VC(f) = N$, and it is well known [Kremer et al. 1999; Kushilevitz and Nisan 1997] that $R_{\delta}^{\rightarrow, \parallel}(f) = \Theta(N)$ in this case, if, say, $\delta < 1/3$.

For g , we take the problem IND with $|\mathcal{U}| = \frac{1}{8\delta}$. By Remark 3.4 following Corollary 3.2 (and a trivial upper bound), $R_{\delta}^{\rightarrow, \parallel}(\text{IND}) = \Theta(N \log 1/\delta)$. On the other hand, $VC(g) \leq N$ since for each row of the communication matrix, there are at most N ones. Also, $VC(g) = N$ since the matrix for f occurs as a submatrix for g . This completes the proof. \square

The separation in Theorem 4.13 is best possible, since the success probability can always be amplified to $1 - \delta$ with $O(\log 1/\delta)$ independent repetitions.

5. CONCLUSION AND OPEN QUESTIONS

We present the first general technique for achieving lower bounds for data stream problems in terms of the error probability δ . In many cases, we show that, surprisingly, the naïve way of amplifying an algorithm’s success probability by independent repetition is space-optimal.

There are still problems for which we do not understand the optimal dependence on δ . One of these is ℓ_p -norm estimation for $p > 2$. For constant probability of success and constant error ε , known bounds are within a $\log n$ factor of optimal [Andoni et al. 2010; Braverman and Ostrovsky 2010; Ganguly 2011]. It would be interesting to prove lower bounds in terms of δ .

Another interesting direction is to understand if there are any time/space trade-offs in terms of δ . For instance, while the space complexity of ℓ_p -norm estimation for $p \leq 2$ gets multiplied by $O(\log 1/\delta)$, it is not clear that the time to process each stream item also needs to be multiplied by $O(\log 1/\delta)$.

Another example of a space/time trade-off is for finding all items x_i in a data stream for which $x_i^2 \geq \frac{1}{100} \|x\|_2^2$, that is, the ℓ_2 “heavy hitters”. It is known that the space required is $\Theta(\log^2 n)$ [Charikar et al. 2002; Jowhari et al. 2011] for constant probability of success. This is actually achieved by an algorithm that succeeds with probability of success $1 - 1/n$. However, the time complexity of recovering the items x_i from the sketch is n^γ for a constant $\gamma > 0$. More recently, it was shown how to achieve $O(\log^2 n)$ bits of space with a much faster time complexity of $\log^{O(1)}(n)$ [Gilbert et al. 2010], but with only a constant probability of success. Is there a lower bound showing that when $\delta = 1/n$, either the space must be $\omega(\log^2 n)$ or the time must be larger than polylogarithmic?

REFERENCES

- Dimitris Achlioptas. 2003. Database-friendly random projections: Johnson-lindenstrauss with binary coins. *J. Comput. Syst. Sci.* 66, 4, 671–687.
- Nir Ailon and Bernard Chazelle. 2009. The fast Johnson–Lindenstrauss transform and approximate nearest neighbors. *SIAM J. Comput.* 39, 1, 302–322.
- Nir Ailon and Bernard Chazelle. 2010. Faster dimension reduction. *Commun. ACM* 53, 2, 97–104.
- Nir Ailon and Edo Liberty. 2009. Fast dimension reduction using rademacher series on dual BCH codes. *Disc. Computat. Geom.* 42, 4, 615–630.
- Nir Ailon and Edo Liberty. 2010. Almost optimal unrestricted fast Johnson-Lindenstrauss transform. CoRR abs/1005.5513.
- Noga Alon. 2003. Problems and results in extremal combinatorics–I. *Disc. Math.* 273, 1–3, 31–53.
- Noga Alon, Yossi Matias, and Mario Szegedy. 1999. The space complexity of approximating the frequency moments. *J. Comput. Syst. Sci.* 58, 1, 137–147.

- Alexandr Andoni, Robert Krauthgamer, and Krzysztof Onak. 2010. Streaming algorithms from precision sampling. CoRR abs/1011.1263.
- Rosa I. Arriaga and Santosh Vempala. 1999. An algorithmic theory of learning: Robust concepts and random projection. In *Proceedings of FOCS*. 616–623.
- Khanh Do Ba, Piotr Indyk, Eric C. Price, and David P. Woodruff. 2010. Lower bounds for sparse recovery. In *Proceedings of the 21st Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*. 1190–1197.
- Ziv Bar-Yossef, T. S. Jayram, Ravi Kumar, and D. Sivakumar. 2002. Information theory methods in communication complexity. In *Proceedings of the 17th Annual IEEE Conference on Computational Complexity (CCC)*. 93–102.
- Ziv Bar-Yossef, T. S. Jayram, Robert Krauthgamer, and Ravi Kumar. 2004. The sketching complexity of pattern matching. In *Proceedings of the 8th International Workshop on Randomization and Computation (RANDOM)*. 261–272.
- Boaz Barak, Mark Braverman, Xi Chen, and Anup Rao. 2010. How to compress interactive communication. In *Proceedings of STOC*. 67–76.
- Vladimir Braverman and Rafail Ostrovsky. 2010. Recursive sketching for frequency moments. CoRR abs/1011.2571 (2010).
- Harry Buhrman, Richard Cleve, and Avi Wigderson. 1998. Quantum vs. classical communication and computation. In *Proceedings of STOC*. 63–68.
- Emmanuel J. Candès and Terence Tao. 2006. Near-optimal signal recovery from random projections: Universal encoding strategies? *IEEE Trans. Inf. Theory* 52, 12, 5406–5425.
- Amit Chakrabarti, Yaoyun Shi, Anthony Wirth, and Andrew Chi-Chih Yao. 2001. Informational complexity and the direct sum problem for simultaneous message complexity. In *Proceedings of FOCS*. 270–278.
- Moses Charikar, Kevin Chen, and Martin Farach-Colton. 2002. Finding frequent items in data streams. In *Proceedings of ICALP*. 693–703.
- Kenneth L. Clarkson. 2008. Tighter bounds for random projections of manifolds. In *Proceedings of the Symposium on Computational Geometry*. 39–48.
- Kenneth L. Clarkson and David P. Woodruff. 2009. Numerical linear algebra in the streaming model. In *Proceedings of the 41st Annual ACM Symposium on Theory of Computing (STOC)*. 205–214.
- Graham Cormode and S. Muthukrishnan. 2003. Estimating dominance norms of multiple data streams. In *Proceedings of ESA*. 148–160.
- Thomas Cover and Joy Thomas. 1991. *Elements of Information Theory*. Wiley Interscience.
- Anirban Dasgupta, Ravi Kumar, and Tamás Sarlós. 2010. A sparse Johnson-Lindenstrauss transform. In *Proceedings of STOC*. 341–350.
- Sanjoy Dasgupta and Anupam Gupta. 2003. An elementary proof of a theorem of Johnson and Lindenstrauss. *Random Struct. Algor.* 22, 1, 60–65.
- Petros Drineas, Michael W. Mahoney, S. Muthukrishnan, and Tamás Sarlós. 2007. Faster least squares approximation. CoRR abs/0710.1435 (2007).
- Sumit Ganguly. 2011. Polynomial estimators for high frequency moments. CoRR abs/1104.4552 (2011).
- Anna C. Gilbert, Yi Li, Ely Porat, and Martin J. Strauss. 2010. Approximate sparse recovery: Optimizing time and measurements. In *Proceedings of STOC*. 475–484.
- Prahladh Harsha, Rahul Jain, David A. McAllester, and Jaikumar Radhakrishnan. 2007. The communication complexity of correlation. In *Proceedings of the IEEE Conference on Computational Complexity*. 10–23.
- Piotr Indyk. 2006. Stable distributions, pseudorandom generators, embeddings, and data stream computation. *J. ACM* 53, 3, 307–323.
- Piotr Indyk and Rajeev Motwani. 1998. Approximate nearest neighbors: Towards removing the curse of dimensionality. In *Proceedings of STOC*. 604–613.
- Piotr Indyk and David P. Woodruff. 2003. Tight lower bounds for the distinct elements problem. In *Proceedings of the 44th Annual IEEE Symposium on Foundations of Computer Science (FOCS)*. 283–288.
- Rahul Jain, Pranab Sen, and Jaikumar Radhakrishnan. 2008. Optimal direct sum and privacy trade-off results for quantum and classical communication complexity. CoRR abs/0807.1267 (2008).
- Hossein Jowhari, Mert Saglam, and Gábor Tardos. 2011. Tight bounds for LP samplers, finding duplicates in streams, and related problems. In *Proceedings of PODS*. 49–58.
- Daniel M. Kane, Jelani Nelson, and David P. Woodruff. 2010a. On the exact space complexity of sketching and streaming small norms. In *Proceedings of SODA*.
- Daniel M. Kane, Jelani Nelson, and David P. Woodruff. 2010b. An optimal algorithm for the distinct elements problem. In *Proceedings of PODS*. 41–52.

- Ilan Kremer, Noam Nisan, and Dana Ron. 1999. On randomized one-round communication complexity. *Computat. Complex.* 8, 1, 21–49.
- Eyal Kushilevitz and Noam Nisan. 1997. *Communication Complexity*. Cambridge University Press.
- J. Langford, L. Li, and A. Strehl. 2007. Vowpal wabbit online learning. Tech. rep. <http://hunch.net/~p=309>.
- Edo Liberty, Nir Ailon, and Amit Singer. 2008. Dense fast random projections and lean walsh transforms. In *Proceedings of APPROX-RANDOM*. 512–522.
- Jirí Matousek. 2008. On variants of the Johnson-Lindenstrauss lemma. *Random Struct. Algor.* 33, 2, 142–156.
- Peter Bro Miltersen, Noam Nisan, Shmuel Safra, and Avi Wigderson. 1998. On data structures and asymmetric communication complexity. *J. Comput. Syst. Sci.* 57, 1, 37–49.
- A. Pavan and Srikanta Tirhapura. 2007. Range-efficient counting of distinct elements in a massive data stream. *SIAM J. Comput.* 37, 2, 359–379.
- V. Rokhlin, A. Szlam, and M. Tygert. 2009. A randomized algorithm for principal component analysis. *SIAM J. Matrix Anal. Appl.* 31, 3, 1100, 1124.
- Tamás Sarlós. 2006. Improved approximation algorithms for large matrices via random projections. In *Proceedings of FOCS*. 143–152.
- Q. Shi, J. Petterson, G. Dror, J. Langford, A. J. Smola, A. Strehl, and V. Vishwanathan. 2009. Hash kernels. In *Proceedings of AISTATS 12*.
- Daniel A. Spielman and Nikhil Srivastava. 2008. Graph sparsification by effective resistances. In *Proceedings of STOC*. 563–568.
- Divesh Srivastava and Suresh Venkatasubramanian. 2010. Information theory for data management. In *Proceedings of the International Conference on Management of Data (SIGMOD'10)*. ACM, New York, 1255–1256. DOI:<http://dx.doi.org/10.1145/1807167.1807337>.
- Stilian Stoev and Murad S. Taqqu. 2010. Max-stable sketches: Estimation of Lp-norms, dominance norms and point queries for non-negative signals. CoRR abs/1005.4344 (2010).
- Stilian Stoev, Marios Hadjieleftheriou, George Kollios, and Murad S. Taqqu. 2007. Norm, point, and distance estimation over multiple signals using max-stable distributions. In *Proceedings of ICDE*. 1006–1015.
- He Sun and Chung Keung Poon. 2009. Two improved range-efficient algorithms for F_0 estimation. *Theor. Comput. Sci.* 410, 11, 1073–1080.
- Mikkel Thorup and Yin Zhang. 2004. Tabulation based 4-universal hashing with applications to second moment estimation. In *Proceedings of the 15th Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*. 615–624.
- S. Tirhapura and D. Woodruff. 2011. Approximating the Klee’s measure problem on a stream of rectangles. Manuscript.
- Kilian Q. Weinberger, Anirban Dasgupta, Josh Attenberg, John Langford, and Alex J. Smola. 2009. Feature hashing for large scale multitask learning. CoRR abs/0902.2206 (2009).
- David P. Woodruff. 2004. Optimal space lower bounds for all frequency moments. In *Proceedings of the 15th Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*. 167–175.
- David P. Woodruff. 2011. Near-optimal private approximation protocols via a black box transformation. In *Proceedings of STOC*. 735–744.

Received August 2011; revised December 2012; accepted December 2012