To appear in STACS 2010 - Symposium on Theoretical Aspects of Computer Science
www.stacs-conf.org

AMS WITHOUT 4-WISE INDEPENDENCE ON PRODUCT DOMAINS

VLADIMIR BRAVERMAN [1] AND KAI-MIN CHUNG [2] AND ZHENMING LIU [3] AND MICHAEL MITZENMACHER [4] AND RAFAIL OSTROVSKY [5]

[1] University of California Los Angeles. Supported in part by NSF grants 0716835, 0716389, 0830803, 0916574 and Lockheed Martin Corporation.
E-mail address: vova@cs.ucla.edu
URL: http://www.cs.ucla.edu/~vova

[2] Harvard School of Engineering and Applied Sciences. Supported by US-Israel BSF grant 2006060 and NSF grant CNS-0831289.
E-mail address: kmchung@fas.harvard.edu
URL: http://people.seas.harvard.edu/~kmchung/

[3] Harvard School of Engineering and Applied Sciences. Supported in part by NSF grant CNS-0721491. The work was finished during an internship in Microsoft Research Asia.
E-mail address: zliu@fas.harvard.edu
URL: http://people.seas.harvard.edu/~zliu/

[4] Harvard School of Engineering and Applied Sciences. Supported in part by NSF grant CNS-0721491 and research grants from Yahoo!, Google, and Cisco.
E-mail address: michaelm@eecs.harvard.edu
URL: http://www.eecs.harvard.edu/~michaelm/

[5] University of California Los Angeles. Supported in part by IBM Faculty Award, Lockheed-Martin Corporation Research Award, Xerox Innovation Group Award, the Okawa Foundation Award, Intel, Teradata, NSF grants 0716835, 0716389, 0830803, 0916574 and U.C. MICRO grant.
E-mail address: rafail@cs.ucla.edu
URL: http://www.cs.ucla.edu/~rafail

ABSTRACT. In their seminal work, Alon, Matias, and Szegedy introduced several sketching techniques, including showing that 4-wise independence is sufficient to obtain good approximations of the second frequency moment. In this work, we show that their sketching technique can be extended to product domains $[n]^k$ by using the product of 4-wise independent functions on $[n]$. Our work extends that of Indyk and McGregor, who showed the result for $k = 2$. Their primary motivation was the problem of identifying correlations in data streams. In their model, a stream of pairs $(i, j) \in [n]^2$ arrive, giving a joint distribution (X, Y), and they find approximation algorithms for how close the joint distribution is to the product of the marginal distributions under various metrics, which naturally corresponds to how close X and Y are to being independent. By using our technique, we obtain a new result for the problem of approximating the ℓ_2 distance between the joint distribution and the product of the marginal distributions for k-ary vectors, instead of just pairs, in a single pass. Our analysis gives a randomized algorithm that is a $(1 \pm \epsilon)$ approximation (with probability $1 - \delta$) that requires space logarithmic in n and m and proportional to 3^k.

1998 ACM Subject Classification: F.2.1, G.3 .
Key words and phrases: Data Streams, Randomized Algorithms, Streaming Algorithms, Independence, L_2 Norm, Sketching.
THIS PAPER IS A MERGE FROM THE WORK OF [7, 9, 11]

SYMPOSIUM ON THEORETICAL ASPECTS OF COMPUTER SCIENCE

© V. Braverman, K. Chung, Z. Liu, M. Mitzenmacher, and R. Ostrovsky
To appear in STACS 2010

1. Introduction

In their seminal work, Alon, Matias and Szegedy [4] presented celebrated sketching techniques and showed that 4-wise independence is sufficient to obtain good approximations of the second frequency moment. Indyk and McGregor [14] make use of this technique in their work introduce the problem of measuring independence in the streaming model. There they give efficient algorithms for approximating pairwise independence for the ℓ_1 and ℓ_2 norms. In their model, a stream of pairs $(i, j) \in [n]^2$ arrive, giving a joint distribution (X, Y), and the notion of approximating pairwise independence corresponds to approximating the distance between the joint distribution and the product of the marginal distributions for the pairs. Indyk and McGregor state, as an explicit open question in their paper, the problem of whether one can estimate k-wise independence on k-tuples for any $k > 2$. In particular, Indyk and McGregor show that, for the ℓ_2 norm, they can make use of the product of 4-wise independent functions on $[n]$ in the sketching method of Alon, Matias, and Szegedy. We extend their approach to show that on the product domain $[n]^k$, the sketching method of Alon, Matias, and Szegedy works when using the product of k copies of 4-wise independent functions on $[n]$. The cost is that the memory requirements of our approach grow exponentially with k, proportionally to 3^k.

Measuring independence and k-wise independence is a fundamental problem with many applications (see e.g., Lehmann [16]). Recently, this problem was also addressed in other models by, among others, Alon, Andoni, Kaufman, Matulef, Rubinfeld and Xie [1]; Batu, Fortnow, Fischer, Kumar, Rubinfeld and White [5]; Goldreich and Ron [12]; Batu, Kumar and Rubinfeld [6]; Alon, Goldreich and Mansour [3]; and Rubinfeld and Servedio [19]. As a specific example, identifying correlations between columns of a table is a fundamental problem in relational databases (see, e.g., Ilyas, Markl, Haas, Brown and Aboulnaga [13]; Haas and Brown [10]; and Poosala and Ioannidis [18]). For queries with predicates over multiple attributes, estimating selectivity is an important part of constructing effective query plans. If there is no prior knowledge, then the "statistical independence assumption" is typical. Under this assumption, the selectivity is estimated by a product of columns' cardinalities, i.e., assuming that columns are statistically independent. However, when this assumption is not correct, it may result in suboptimal query plans and decrease performance significantly (see, e.g., Poosala and Ioannidis [18]). For data warehouse and OLAP applications, finding correlated columns may optimize schema and thus improve performance significantly (see e.g., Kimball and Caserta [15]).

Traditional non-parametric methods of testing independence over empirical data usually require space complexity that is polynomial to either the support size or input size. The scale of contemporary data sets often prohibits such space complexity. It is therefore natural to ask whether we will be able to design algorithms to test for independence in streaming model. Interestingly, this specific problem appears not to have been introduced until the work of Indyk and McGregor. While arguably results for the ℓ_1 norm would be stronger than for the ℓ_2 norm in this setting, the problem for ℓ_2 norms is interesting in its own right. The problem for the ℓ_1 norm has been recently resolved by Braverman and Ostrovsky in [8]. They gave an $(1 \pm \epsilon, \delta)$-approximation algorithm that makes a single pass over a data stream and uses polylogarithmic memory.

1.1. Our Results

In this paper we generalize the "sketching of sketches" result of Indyk and McGregor. Our specific theoretical contributions can be summarized as follows:

Main Theorem.

Let $\vec{v} \in \mathrm{R}^{(n^k)}$ be a vector with entries $\vec{v}_{\mathbf{p}} \in R$ for $\mathbf{p} \in [n]^k$. Let $h_1, \ldots, h_k : [n] \to \{-1, 1\}$ be independent copies of 4-wise independent hash functions; that is, $h_i(1), \ldots, h_i(n) \in \{-1, 1\}$ are 4-wise independent hash functions for each $i \in [k]$, and $h_1(\cdot), \ldots, h_k(\cdot)$ are mutually independent. Define $H(\mathbf{p}) = \prod_{i=1}^{k} h_j(p_j)$, and the sketch $Y = \sum_{p \in [n]^k} \vec{v}_{\mathbf{p}} H(p)$.

We prove that the sketch Y can be used to give an efficient approximation for $\|\vec{v}\|^2$; our result is stated formally in Theorem 4.2. Note that H is not 4-wise independent.

As a corollary, the main application of our main theorem is to extend the result of Indyk and McGregor [14] to detect the dependency of k random variables in streaming model.

Corollary 1.1. *For every $\epsilon > 0$ and $\delta > 0$, there exists a randomized algorithm that computes, given a sequence $a_1, \ldots, a_m$ of k-tuples, in one pass and using $O(3^k \epsilon^{-2} \log \frac{1}{\delta}(\log m + \log n))$ memory bits, a number Y so that the probability Y deviates from the ℓ_2 distance between product and joint distribution by more than a factor of $(1 + \epsilon)$ is at most δ.*

1.2. Techniques and a Historical Remark

This paper is a merge from [7, 11, 9], where the same result was obtained with different proofs. The proof of [11] generalizes the geometric approach of Indyk and McGregor [14] with new geometric observations. The proofs of [7, 9] are more combinatorial in nature. These papers offer new insights, but due to the space limitation, we focus on the proof from [9] in this paper. Original papers are available on line and are recommended to the interested reader.

2. The Model

We provide the general underlying model. Here we mostly follow the notation of [7, 14].

Let S be a stream of size m with elements $a_1, \ldots, a_m$, where $a_i \equiv (a_i^1, \ldots, a_i^k) \in [n]^k$. (When we have a sequence of elements that are themselves vectors, we denote the sequence number by a subscript and the vector entry by a superscript when both are needed.) The stream S defines an *empirical* distribution over $[n]^k$ as follows: the frequency $f(\omega)$ of an element $\omega \in [n]^k$ is defined as the number of times it appears in S, and the empirical distribution is

$$\Pr[\omega] = \frac{f(\omega)}{m} \quad \text{for any } \omega \in [n]^k.$$

Since $\omega = (\omega_1, \ldots, \omega_k)$ is a vector of size k, we may also view the streaming data as defining a joint distribution over the random variables $X_1, \ldots, X_k$ corresponding to the values in each dimension. (In the case of $k = 2$, we write the random variables as X and Y rather than X_1 and X_2.) There is a natural way of defining marginal distribution for the random variable X_i: for $\omega_i \in [n]$, let $f_i(\omega_i)$ be the number of times ω_i appears in the ith coordinate of an element of S, or

$$f_i(\omega_i) = \left|\{a_j \in S : a_j^i = \omega_i\}\right|.$$

The empirical marginal distribution $\Pr_i[\cdot]$ for the ith coordinate is defined as

$$\Pr_i[\omega_i] = \frac{f_i(\omega_i)}{m} \quad \text{for any } \omega_i \in [n].$$

Next let $\vec{v}$ be the vector in $\mathrm{R}^{[n]^k}$ with $\vec{v}_\omega = \Pr[\omega] - \prod_{1 \le i \le k} \Pr_i[\omega_i]$ for all $\omega \in [n]^k$. Our goal is to approximate the value

$$\|\vec{v}\| \equiv \left(\sum_{\omega \in [n]^k} \left| \Pr[\omega] - \prod_{1 \le i \le k} \Pr_i[\omega_i] \right|^2 \right)^{\frac{1}{2}}. \tag{2.1}$$

This represent the ℓ_2 norm between the tensor of the marginal distributions and the joint distribution, which we would expect to be close to zero in the case where the X_i were truly independent.

Finally, our algorithms will assume the availability of 4-wise independent hash functions. For more on 4-wise independence, including efficient implementations, see [2, 20]. For the purposes of this paper, the following simple definition will suffice.

Definition 2.1. *(4-wise independence)* A family of hash functions $\mathcal{H}$ with domain $[n]$ and range $\{-1, 1\}$ is *4-wise independent* if for any distinct values $i_1, i_2, i_3, i_4 \in [n]$ and any $b_1, b_2, b_3, b_4 \in \{-1, 1\}$, the following equality holds,

$$\Pr_{h \leftarrow \mathcal{H}}[h(i_1) = b_1, h(i_2) = b_2, h(i_3) = b_3, h(i_4) = b_4] = 1/16.$$

Remark 2.2. In [14], the family of 4-wise independent hash functions $\mathcal{H}$ is called 4-wise independent random vectors. For consistencies within our paper, we will always view the object $\mathcal{H}$ as a hash function family.

3. The Algorithm and its Analysis for $k = 2$

We begin by reviewing the approximation algorithm and associated proof for the ℓ_2 norm given in [14]. Reviewing this result will allow us to provide the necessary notation and frame the setting for our extension to general k. Moreover, in our proof, we find that a constant in Lemma 3.1 from [14] that we subsequently generalize appears incorrect. (Because of this, our proof is slightly different and more detailed than the original.) Although the error is minor in the context of their paper (it only affects the constant factor in the order notation), it becomes more important when considering the proper generalization to larger k, and hence it is useful to correct here.

In the case $k = 2$, we assume that the sequence $(a_1^1, a_1^2), (a_2^1, a_2^2), \ldots, (a_m^1, a_m^2)$ arrives an item by an item. Each (a_i^1, a_i^2) (for $1 \le i \le m$) is an element in $[n]^2$. The random variables X and Y over $[n]$ can be expressed as follows:

$$\begin{cases} \Pr[i,j] = \Pr[X = i, Y = j] = |\{\ell : (a_\ell^1, a_\ell^2) = (i,j)\}|/m \\ \Pr_1[i] = \Pr[X = i] = |\{\ell : (a_\ell^1, a_\ell^2) = (i,\cdot)\}|/m \\ \Pr_2[j] = \Pr[Y = j] = |\{\ell : (a_\ell^1, a_\ell^2) = (\cdot,j)\}|/m. \end{cases}$$

We simplify the notation and use $p_i \equiv \Pr[X = i]$, $q_j \equiv \Pr[Y = j]$, $r_{i,j} = \Pr[X = i, Y = j]$. and $s_{i,j} = \Pr[X = i] \Pr[Y = j]$.

Indyk and McGregor's algorithm proceeds in a similar fashion to the streaming algorithm presented in [4]. Specifically let $s_1 = 72\epsilon^{-2}$ and $s_2 = 2\log(1/\delta)$. The algorithm computes s_2 random variables $Y_1, Y_2, \ldots, Y_{s_2}$ and outputs their median. The output is the algorithm's estimate on the norm of v defined in Equation 2.1. Each Y_i is the average of s_1 random variables Y_{ij}: $1 \le j \le s_1$, where Y_{ij} are independent, identically distributed random variables. Each of the variables $D = D_{ij}$ can be computed from the algorithmic routine shown in Figure 1.

2-D APPROXIMATION $\left((a_1^1, a_1^2), \ldots, (a_m^1, a_m^2)\right)$

1 Independently generate 4-wise independent random functions h_1, h_2 from $[n]$ to $\{-1, 1\}$.
2 **for** $c \leftarrow 1$ **to** m
3 **do** Let the cth item $(a_c^1, a_c^2) = (i, j)$
4 $t_1 \leftarrow t_1 + h_1(i)h_2(j), t_2 \leftarrow t_2 + h_1(i), t_3 \leftarrow t_3 + h_2(j)$.
5 Return $Y = (t_1/m - t_2t_3/m^2)^2$.

Figure 1: The procedure for generating random variable Y for $k = 2$.

Notice that by the end of the process 2-D APPROXIMATION, we have $t_1/m = \sum_{i,j\in[n]} h_1(i)h_2(j)r_{i,j}$, $t_2/m = \sum_{i\in[n]} h_1(i)p_i$, and $t_3/m = \sum_{i\in[n]} h_2(i)q_i$. Also, when a vector is in $\mathrm{R}^{(n^2)}$, its indices can be represented by $(i_1, i_2) \in [n]^2$. In what follows, we will use a bold letter to represent the index of a high dimensional vector, e.g., $v_{\mathbf{i}} \equiv v_{i_1,i_2}$. The following Lemma shows that the expectation of Y is $\|v\|^2$ and the variance of Y is at most $8(\mathrm{E}[Y])^2$ because $\mathrm{E}[Y^2] \le 9\mathrm{E}[Y]^2$.

Lemma 3.1. *([14]) Let h_1, h_2 be two independent instances of 4-wise independent hash functions from $[n]$ to $\{-1, 1\}$. Let $v \in \mathrm{R}^{n^2}$ and $H(\mathbf{i})(\equiv H\big((i_1, i_2)\big) = h_1(i_i) \cdot h_2(i_2)$. Let us define $Y = \Big(\sum_{\mathbf{i}\in[n]^2} H(\mathbf{i})v_{\mathbf{i}}\Big)^2$. Then $\mathrm{E}[Y] = \sum_{\mathbf{i}\in[n]^2} \vec{v}_{\mathbf{i}}^2$ and $\mathrm{E}[Y^2] \le 9(\mathrm{E}[Y])^2$, which implies $\mathrm{Var}[Y] \le 8E^2[Y]$.*

Proof. We have $\mathrm{E}[Y] = \mathrm{E}[(\sum_{\mathbf{i}} H(\mathbf{i})\vec{v}_{\mathbf{i}})^2] = \sum_{\mathbf{i}} \vec{v}_{\mathbf{i}}^2\mathrm{E}[H^2(\mathbf{i})] + \sum_{\mathbf{i}\neq\mathbf{j}} \vec{v}_{\mathbf{i}}\vec{v}_{\mathbf{j}}\mathrm{E}[H(\mathbf{i})H(\mathbf{j})]$. For all $\mathbf{i} \in [n]^2$, we know $h^2(\mathbf{i}) = 1$. On the other hand, $H(\mathbf{i})H(\mathbf{j}) \in \{-1, 1\}$. The probability that $H(\mathbf{i})H(\mathbf{j}) = 1$ is $\Pr[H(\mathbf{i})H(\mathbf{j}) = 1] = \Pr[h_1(i_1)h_1(j_1)h_2(i_2)h_2(j_2) = 1] = 1/16 + \binom{4}{2}1/16 + 1/16 = 1/2$. The last equality holds is because $h_1(i_1)h_1(j_1)h_2(i_2)h_2(j_2) = 1$ is equivalent to saying either all these variables are 1, or exactly two of these variables are -1, or all these variables are -1. Therefore, $\mathrm{E}[h(\mathbf{i})h(\mathbf{j})] = 0$. Consequently, $\mathrm{E}[Y] = \sum_{\mathbf{i}\in[n]^2}(\vec{v}_{\mathbf{i}})^2$.

Now we bound the variance. Recall that $\mathrm{Var}[Y] = \mathrm{E}[Y^2] - \mathrm{E}[Y]^2$, we bound

$$\mathrm{E}[Y^2] = \sum_{\mathbf{i},\mathbf{j},\mathbf{k},\mathbf{l}\in[n]^2} \mathrm{E}[H(\mathbf{i})H(\mathbf{j})H(\mathbf{k})h(\mathbf{l})]\vec{v}_{\mathbf{i}}\vec{v}_{\mathbf{j}}\vec{v}_{\mathbf{k}}\vec{v}_{\mathbf{l}} \le \sum_{\mathbf{i},\mathbf{j},\mathbf{k},\mathbf{l}\in[n]^2} |\mathrm{E}[H(\mathbf{i})H(\mathbf{j})H(\mathbf{k})H(\mathbf{l})]| \cdot |\vec{v}_{\mathbf{i}}\vec{v}_{\mathbf{j}}\vec{v}_{\mathbf{k}}\vec{v}_{\mathbf{l}}|.$$

Also $|\mathrm{E}[H(\mathbf{i})H(\mathbf{j})H(\mathbf{k})H(\mathbf{l})]| \in \{0, 1\}$. The quantity $\mathrm{E}[H(\mathbf{i})H(\mathbf{j})H(\mathbf{k})H(\mathbf{l})] \neq 0$ if and only if the following relation holds,

$$\forall s \in [2] : ((i_s = j_s) \wedge (k_s = l_s)) \vee ((i_s = k_s) \wedge (j_s = l_s)) \vee ((i_s = l_s) \wedge (k_s = j_s)). \tag{3.1}$$

Denote the set of 4-tuples $(\mathbf{i}, \mathbf{j}, \mathbf{k}, \mathbf{l})$ that satisfy the above relation by $\mathcal{D}$. We may also view each 4-tuple as an ordered set that consists of 4 points in $[n]^2$. Consider the unique smallest axes-parallel rectangle in $[n]^2$ that contains a given 4-tuple in $\mathcal{D}$ (i.e. contains the four ordered points). Note this could either be a (degenerate) line segment or a (non-degenerate) rectangle, as we discuss below. Let $M : \mathcal{D} \to \{A, B, C, D\}$ be the function that maps an element $\sigma \in \mathcal{D}$ to the smallest rectangle $ABCD$ defined by σ. Since a rectangle can be uniquely determined by its diagonals, we may write $M : \mathcal{D} \to (\chi_1, \chi_2, \varphi_1, \varphi_2)$, where $\chi_1 \le \chi_2 \in [n]$, $\varphi_1 \le \varphi_2 \in [n]$ and the corresponding rectangle is understood to be the one with diagonal $\{(\chi_1, \varphi_1), (\chi_2, \varphi_2)\}$. Also, the inverse function $M^{-1}(\chi_1, \chi_2, \varphi_1, \varphi_2)$ represents the pre-images of $(\chi_1, \chi_2, \varphi_1, \varphi_2)$ in $\mathcal{D}$. $(\chi_1, \chi_2, \varphi_1, \varphi_2)$ is degenerate if either $\chi_1 = \chi_2$ or $\varphi_1 = \varphi_2$, in which case the rectangle (and its diagonals) correspond to the segment itself, or $\chi_1 = \chi_2$ and $\varphi_1 = \varphi_2$, and the rectangle is just a single point.

Example 3.2. Let $\mathbf{i} = (1, 2)$, $\mathbf{j} = (3, 2)$, $\mathbf{k} = (1, 5)$, and $\mathbf{l} = (3, 5)$. The tuple is in $\mathcal{D}$ and its corresponding bounding rectangle is a non-degenerate rectangle. The function $M(\mathbf{i}, \mathbf{j}, \mathbf{k}, \mathbf{l}) = (1, 3, 2, 5)$.

Example 3.3. Let $\mathbf{i} = \mathbf{j} = (1, 4)$ and $\mathbf{k} = \mathbf{l} = (3, 7)$. The tuple is also in $\mathcal{D}$ and minimal bounding rectangle formed by these points is an interval $\{(1, 4), (3, 7)\}$. The function $M(\mathbf{i}, \mathbf{j}, \mathbf{k}, \mathbf{l}) = (1, 3, 4, 7)$.

To start we consider the non-degenerate cases. Fix any $(\chi_1, \chi_2, \varphi_1, \varphi_2)$ with $\chi_1 < \chi_2$ and $\phi_1 < \phi_2$. There are in total $\binom{4}{2}^2 = 36$ tuples $(\mathbf{i}, \mathbf{j}, \mathbf{k}, \mathbf{l})$ in $\mathcal{D}$ with $M(\mathbf{i}, \mathbf{j}, \mathbf{k}, \mathbf{l}) = (\chi_1, \chi_2, \varphi_1, \varphi_2)$. Twenty-four of these tuples correspond to the setting where none of $\mathbf{i}, \mathbf{j}, \mathbf{k}, \mathbf{l}$ are equal, as there are twenty-four permutations of the assignment of the labels $\mathbf{i}, \mathbf{j}, \mathbf{k}, \mathbf{l}$ to the four points. (This corresponds to the first example). In this case the four points form a rectangle, and we have $|\vec{v}_{\mathbf{i}}\vec{v}_{\mathbf{j}}\vec{v}_{\mathbf{k}}\vec{v}_{\mathbf{l}}| \le \frac{1}{2}((\vec{v}_{\chi_1,\varphi_1}\vec{v}_{\chi_2,\varphi_2})^2 + (\vec{v}_{\chi_1,\varphi_2}\vec{v}_{\chi_2,\varphi_1})^2)$. Intuitively, in these cases, we assign the "weight" of the tuple to the diagonals.

The remaining twelve tuples in $M^{-1}(\chi_1, \chi_2, \varphi_1, \varphi_2)$ correspond to intervals. (This corresponds to the second example.) In this case two of $\mathbf{i}, \mathbf{j}, \mathbf{k}, \mathbf{l}$ correspond to one endpoint of the interval, and the other two

labels correspond to the other endpoint. Hence we have either $|\vec{v}_{\mathbf{i}}\vec{v}_{\mathbf{j}}\vec{v}_{\mathbf{k}}\vec{v}_{\mathbf{l}}| = (\vec{v}_{\chi_1,\varphi_1}\vec{v}_{\chi_2,\varphi_2})^2$ or $|\vec{v}_{\mathbf{i}}\vec{v}_{\mathbf{j}}\vec{v}_{\mathbf{k}}\vec{v}_{\mathbf{l}}| = (\vec{v}_{\chi_1,\varphi_2}\vec{v}_{\chi_2,\varphi_1})^2$, and there are six tuples for each case.

Therefore for any $\chi_1 < \chi_2 \in [n]$ and $\varphi_1 < \varphi_2 \in [n]$ we have:

$$\sum_{\substack{(\mathbf{i},\mathbf{j},\mathbf{k},\mathbf{l})\in \\ M^{-1}(\chi_1,\chi_2,\varphi_1,\varphi_2)}} |v_{\mathbf{i}}v_{\mathbf{j}}v_{\mathbf{k}}v_{\mathbf{l}}| \le 18((v_{\chi_1,\varphi_1}v_{\chi_2,\varphi_2})^2 + (v_{\chi_1,\varphi_2}, v_{\chi_2,\varphi_1})^2).$$

The analysis is similar for the degenerate cases, where the constant 18 in the bound above is now quite loose. When exactly one of $\chi_1 = \chi_2$ or $\varphi_1 = \varphi_2$ holds, the size of $M^{-1}(\chi_1, \chi_2, \varphi_1, \varphi_2)$ is $\binom{4}{2} = 6$, and the resulting intervals correspond to vertical or horizontal lines. When both $\chi_1 = \chi_2$ and $\varphi_1 = \varphi_2$, then $|M^{-1}(\chi_1, \chi_2, \varphi_1, \varphi_2)| = 1$. In sum, we have

$$\begin{aligned}
\sum_{\mathbf{i},\mathbf{j},\mathbf{k},\mathbf{l}\in\mathcal{D}} |\vec{v}_{\mathbf{i}}\vec{v}_{\mathbf{j}}\vec{v}_{\mathbf{k}}\vec{v}_{\mathbf{l}}| &= \sum_{\substack{\chi_1\le\chi_2\\ \varphi_1\le\varphi_2}} \sum_{\substack{(\mathbf{i},\mathbf{j},\mathbf{k},\mathbf{l})\in \\ M^{-1}(\chi_1,\chi_2,\varphi_1,\varphi_2)}} |\vec{v}_{\mathbf{i}}\vec{v}_{\mathbf{j}}\vec{v}_{\mathbf{k}}\vec{v}_{\mathbf{l}}| \\
&\le \sum_{\substack{\chi_1<\chi_2\\ \varphi_1<\varphi_2}} 18((\vec{v}_{\chi_1,\varphi_1}\vec{v}_{\chi_2,\varphi_2})^2 + (\vec{v}_{\chi_1,\varphi_2}\vec{v}_{\chi_2,\varphi_1})^2) + \sum_{\substack{\chi_1=\chi_2\\ \varphi_1<\varphi_2}} 6((\vec{v}_{\chi_1,\varphi_1}\vec{v}_{\chi_2,\varphi_2})^2 + (\vec{v}_{\chi_1,\varphi_2}\vec{v}_{\chi_2,\varphi_1})^2) \\
&\quad + \sum_{\substack{\chi_1<\chi_2\\ \varphi_1=\varphi_2}} 6((\vec{v}_{\chi_1,\varphi_1}\vec{v}_{\chi_2,\varphi_2})^2 + (\vec{v}_{\chi_1,\varphi_2}\vec{v}_{\chi_2,\varphi_1})^2) + \sum_{\substack{\chi_1=\chi_2\\ \varphi_1=\varphi_2}} (\vec{v}_{\chi_1,\varphi_1}\vec{v}_{\chi_2,\varphi_2})^2 \\
&\le 9 \sum_{\substack{\mathbf{i}\in[n]^2\\ \mathbf{j}\in[n]^2}} (\vec{v}_{\mathbf{i}}\vec{v}_{\mathbf{j}})^2 = 9\mathrm{E}^2[Y].
\end{aligned}$$

Finally, we have $\sum_{\mathbf{i},\mathbf{j},\mathbf{k},\mathbf{l}\in[n]^2} |\mathrm{E}[H(\mathbf{i})H(\mathbf{j})H(\mathbf{k})H(\mathbf{l})]| \cdot |\vec{v}_{\mathbf{i}}\vec{v}_{\mathbf{j}}\vec{v}_{\mathbf{k}}\vec{v}_{\mathbf{l}}| \le \sum_{\mathbf{i},\mathbf{j},\mathbf{k},\mathbf{l}\in\mathcal{D}} |\vec{v}_{\mathbf{i}}\vec{v}_{\mathbf{j}}\vec{v}_{\mathbf{k}}\vec{v}_{\mathbf{l}}| \le 9\mathrm{E}^2[Y]$ and $\mathrm{Var}[Y] \le 8\mathrm{E}[Y]^2$. ∎

We emphasize the geometric interpretation of the above proof as follows. The goal is to bound the variance by a constant times $\mathrm{E}^2[Y] = \sum_{\mathbf{i},\mathbf{j}\in[n]^2}(\vec{v}_{\mathbf{i}}v_{\mathbf{j}})^2$, where the index set is the set of all possible lines in plane $[n]^2$ (each line appears twice). We first show that $\mathrm{Var}[Y] \le \sum_{\mathbf{i},\mathbf{j},\mathbf{k},\mathbf{l}\in\mathcal{D}} |\vec{v}_{\mathbf{i}}\vec{v}_{\mathbf{j}}\vec{v}_{\mathbf{k}}\vec{v}_{\mathbf{l}}|$, where the 4-tuple index set corresponds to a set of rectangles in a natural way. The main idea of Indyk and McGregor is to use inequalities of the form $|\vec{v}_{\mathbf{i}}\vec{v}_{\mathbf{j}}\vec{v}_{\mathbf{k}}\vec{v}_{\mathbf{l}}| \le \frac{1}{2}((\vec{v}_{\chi_1,\varphi_1}\vec{v}_{\chi_2,\varphi_2})^2 + (\vec{v}_{\chi_1,\varphi_2}\vec{v}_{\chi_2,\varphi_1})^2)$ to assign the "weight" of each 4-tuple to the diagonals of the corresponding rectangle. The above analysis shows that 18 copies of all lines are sufficient to accommodate all 4-tuples. While similar inequalities could also assign the weight of a 4-tuple to the vertical or horizontal edges of the corresponding rectangle, using vertical or horizontal edges is problematic. The reason is that there are $\Omega(n^4)$ 4-tuples but only $O(n^3)$ vertical or horizontal edges, so some lines would receive $\Omega(n)$ weight, requiring $\Omega(n)$ copies. This problem is already noted in [7].

Our bound here is $\mathrm{E}[Y^2] \le 9\mathrm{E}^2[Y]$, while in [14] the bound obtained is $\mathrm{E}[Y^2] \le 3\mathrm{E}^2[Y]$. There appears to have been an error in the derivation in [14]; some intuition comes from the following example. We note that $|\mathcal{D}|$ is at least $\binom{4}{2}^2 \cdot \binom{n}{2}^2 = 9n^4 - 9n^2$. (This counts the number of non-degenerate 4-tuples.) Now if we set $v_i = 1$ for all $1 \le i \le n^2$, we have $\mathrm{E}[Y^2] \ge |\mathcal{D}| = 9n^4 - 9n^2 \sim 9\mathrm{E}^2(D)$, which suggests $\mathrm{Var}[D] > 3\mathrm{E}^2[D]$. Again, we emphasize this discrepancy is of little importance to [14]; the point there is that the variance is bounded by a constant factor times the square of the expectation. It is here, where we are generalizing to $k \ge 3$, that the exact constant factor is of some importance.

Given the bounds on the expectation and variance for the $D_{i,j}$, standard techniques yield a bound on the performance of our algorithm.

Theorem 3.4. *For every $\epsilon > 0$ and $\delta > 0$, there exists a randomized algorithm that computes, given a sequence $(a_1^1, a_1^2), \ldots, (a_m^1, a_m^2)$, in one pass and using $O(\epsilon^{-2}\log\frac{1}{\delta}(\log m + \log n))$ memory bits, a number* Med *so that the probability* Med *deviates from $\|v\|^2$ by more than ϵ is at most δ.*

Proof. Recall the algorithm described in the beginning of Section 3: let $s_1 = 72\epsilon^{-2}$ and $s_2 = 2\log\delta$. We first computes s_2 random variables $Y_1, Y_2, \ldots, Y_{s_2}$ and outputs their median Med, where each Y_i is the average of s_1 random variables Y_{ij}: $1 \le j \le s_1$ and Y_{ij} are independent, identically distributed random variables computed by Figure 1. By Chebyshev's inequality, we know that for any fixed i,

$$\Pr\left(\left|Y_i - \|\vec{v}\|\right| \right) \ge \epsilon\|\vec{v}\|] \le \frac{\mathrm{Var}(Y_i)}{\epsilon^2\|\vec{v}\|^2} = \frac{(1/s_1)\mathrm{Var}[Y]}{\epsilon^2\|\vec{v}\|^2} = \frac{(9\epsilon^2/72)\|\vec{v}\|^2}{\epsilon^2\|\vec{v}\|^2} = \frac{1}{8}.$$

Finally, by standard Chernoff bound arguments (see for example Chapter 4 of [17]), the probability that more than $s_2/2$ of the variables Y_i deviate by more than $\epsilon\|\vec{v}\|$ from $\|\vec{v}\|$ is at most δ. In case this does not happen, the median Med supplies a good estimate to the required quantity $\|\vec{v}\|$ as needed. ∎

4. The General Case $k \ge 3$

Now let us move to the general case where $k \ge 3$. Recall that $\vec{v}$ is a vector in R^{n^k} that maintains certain statistics of a data stream, and we are interested in estimating its ℓ_2 norm $\|\vec{v}\|$. There is a natural generalization for Indyk and McGregor's method for $k = 2$ to construct an estimator for $\|\vec{v}\|$: let $h_1, \ldots, h_k : [n] \to \{-1, 1\}$ be independent copies of 4-wise independent hash functions (namely, $h_i(1), \ldots, h_i(n) \in \{-1, 1\}$ are 4-wise independent hash functions for each $i \in [k]$, and $h_1(\cdot), \ldots, h_k(\cdot)$ are mutually independent.). Let $H(\mathbf{p}) = \prod_{i=1}^{k} h_j(p_j)$. The estimator Y is defined as $Y \equiv \left(\sum_{\mathbf{p}\in[n]^k} \vec{v}_{\mathbf{p}} H(\mathbf{p})\right)^2$.

Our goal is to show that $\mathrm{E}[Y] = \|\vec{v}\|^2$ and $\mathrm{Var}[Y]$ is reasonably small so that a streaming algorithm maintaining multiple independent instances of estimator Y will be able to output an approximately correct estimation of $\|\vec{v}\|$ with high probability. Notice that when $\|\vec{v}\|$ represents the ℓ_2 distance between the joint distribution and the tensors of the marginal distributions, the estimator can be computed efficiently in a streaming model similarly to as in Figure 1. We stress that our result is applicable to a broader class of ℓ_2-norm estimation problems, as long as the vector $\vec{v}$ to be estimated has a corresponding efficiently computable estimator Y in an appropriate streaming model. Formally, we shall prove the following main lemma in the next subsection.

Lemma 4.1. *Let $\vec{v}$ be a vector in R^{n^k}, and $h_1, \ldots, h_k : [n] \to \{-1, 1\}$ be independent copies of 4-wise independent hash functions. Define $H(\mathbf{p}) = \prod_{i=1}^{k} h_j(p_j)$, and $Y \equiv \left(\sum_{\mathbf{p}\in[n]^k} \vec{v}_{\mathbf{p}} H(\mathbf{p})\right)^2$. We have $\mathrm{E}[Y] = ||\vec{v}||$ and $\mathrm{Var}[Y] \le 3^k \mathrm{E}[Y]^2$.*

We remark that the bound on the variance in the above lemma is tight. One can verify that when the vector $\vec{v}$ is a uniform vector (i.e., all entries of $\vec{v}$ are the same), the variance of Y is $\Omega(3^k E[Y]^2)$. With the above lemma, the following main theorem mentioned in the introduction immediately follows by a standard argument presented in the proof of Theorem 3.4 in the previous section.

Theorem 4.2. *Let $\vec{v}$ be a vector in $\mathrm{R}^{[n]^k}$ that maintains an arbitrary statistics in a data stream of size m, in which every item is from $[n]^k$. Let $\epsilon, \delta \in (0, 1)$ be real numbers. If there exists an algorithm that maintains an instance of Y using $O(\mu(n, m, k, \epsilon, \delta))$ memory bits, then there exists an algorithm Λ such that:*

(1) With probability $\ge 1 - \delta$ the algorithm Λ outputs a value between $[(1-\epsilon)\|\vec{v}\|^2, (1+\epsilon)\|\vec{v}|^2]$ and
(2) the space complexity of Λ is $O(3^k \frac{1}{\epsilon^2} \log \frac{1}{\delta} \mu(n, m, k, \epsilon, \delta))$.

As discussed above, an immediate corollary is the existence of a one-pass space efficient streaming algorithm to detect the dependency of k random variables in ℓ_2-norm:

Corollary 4.3. *For every $\epsilon > 0$ and $\delta > 0$, there exists a randomized algorithm that computes, given a sequence $a_1, \ldots, a_m$ of k-tuples, in one pass and using $O(3^k \epsilon^{-2} \log \frac{1}{\delta}(\log m + \log n))$ memory bits, a number Y so that the probability Y deviates from the square of the ℓ_2 distance between product and joint distribution by more than a factor of $(1 + \epsilon)$ is at most δ.*

4.1. Analysis of the Sketch Y

This section is devoted to prove Lemma 4.1, where the main challenge is to bound the variance of Y. The geometric approach of Indyk and McGregor [14] presented in Section 3 for the case of $k = 2$ can be extended to analyze the general case. However, we remark that the generalization requires new ideas. In particular, instead of performing "local analysis" that maps each rectangle to its diagonals, a more complex "global analysis" is needed in higher dimensions to achieve the desired bounds. The alternative proof we present here utilizes similar ideas, but relies on a more combinatorial rather than geometric approach.

For the expectation of Y, we have

$$\begin{aligned}\mathrm{E}[Y] &= \mathrm{E}\left[\sum_{\mathbf{p},\mathbf{q}\in[n]^k} \vec{v}_{\mathbf{p}} \cdot \vec{v}_{\mathbf{q}} \cdot H(\mathbf{p}) \cdot H(\mathbf{q})\right] \\ &= \sum_{\mathbf{p}\in[n]^k} \vec{v}_{\mathbf{p}}^2 \cdot \mathrm{E}\left[H(\mathbf{p})^2\right] + \sum_{\mathbf{p}\neq\mathbf{q}\in[n]^k} \vec{v}_{\mathbf{p}} \cdot \vec{v}_{\mathbf{q}} \cdot \mathrm{E}\left[H(\mathbf{p})H(\mathbf{q})\right] \\ &= \sum_{\mathbf{p}\in[n]^k} \vec{v}_{\mathbf{p}}^2 = ||\vec{v}||^2,\end{aligned}$$

where the last equality follows by $H(\mathbf{p})^2 = 1$, and $\mathrm{E}\left[H(\mathbf{p})H(\mathbf{q})\right] = 0$ for $\mathbf{p} \neq \mathbf{q}$.

Now, let us start to prove $\mathrm{Var}[Y] \leq 3^k \mathrm{E}[Y]^2$. By definition, $\mathrm{Var}[Y] = \mathrm{E}[(Y - \mathrm{E}[Y])^2]$, so we need to understand the following random variable:

$$Err \equiv Y - \mathrm{E}[Y] = \sum_{\mathbf{p}\neq\mathbf{q}\in[n]^k} H(\mathbf{p})H(\mathbf{q})\vec{v}_{\mathbf{p}}\vec{v}_{\mathbf{q}}. \tag{4.1}$$

The random variable Err is a sum of terms indexed by pairs $(\mathbf{p}, \mathbf{q}) \in [n]^k \times [n]^k$ with $\mathbf{p} \neq \mathbf{q}$. At a very high level, our analysis consists of two steps. In the first step, we group the terms in Err properly and simplify the summation in each group. In the second step, we expand the square of the sum in $\mathrm{Var}[Y] = \mathrm{E}[Err^2]$ according to the groups and apply Cauchy-Schwartz inequality three times to bound the variance.

We shall now gradually introduce the necessary notation for grouping the terms in Err and simplifying the summation. We remind the reader that vectors over the reals (e.g., $\vec{v} \in R^{n^k}$) are denoted by $\vec{v}, \vec{w}, \vec{r}$, and vectors over $[n]$ are denoted by $\mathbf{p}, \mathbf{q}, \mathbf{a}, \mathbf{b}, \mathbf{c}, \mathbf{d}$ and referred as *index vectors*. We use $S \subseteq [k]$ to denote a subset of $[k]$, and let $\bar{S} = [k] \backslash S$. We use $\mathrm{Ham}(\mathbf{p}, \mathbf{q})$ to denote the *Hamming distance* of index vectors $\mathbf{p}, \mathbf{q} \in [n]^k$, i.e., the number of coordinates where $\mathbf{p}$ and $\mathbf{q}$ are different.

Definition 4.4. *(Projection and inverse projection)* Let $\mathbf{c} \in [n]^k$ be an index vector and $S \subseteq [k]$ a subset. We define the *projection of* $\mathbf{c}$ *to* S, denoted by $\Phi_S(\mathbf{c}) \in [n]^{|S|}$, to be the vector $\mathbf{c}$ restricted to the coordinates in S. Also, let $\mathbf{a} \in [n]^{|S|}$ and $\mathbf{b} \in [n]^{k-|S|}$ be index vectors. We define the *inverse projection of* $\mathbf{a}$ *and* $\mathbf{b}$ *with respect to* S, denoted by $\Phi_S^{-1}(\mathbf{a}, \mathbf{b}) \in [n]^k$, as the index vector $\mathbf{c} \in [n]^k$ such that $\Phi_S(\mathbf{c}) = \mathbf{a}$ and $\Phi_{\bar{S}}(\mathbf{c}) = \mathbf{b}$.

We next define *pair groups* and use the definition to group the terms in Err.

Definition 4.5. *(Pair Group)* Let $S \subseteq [k]$ be a subset of size $|S| = t$. Let $\mathbf{c}, \mathbf{d} \in [n]^t$ be a pair of index vectors with $\mathrm{Ham}(\mathbf{c}, \mathbf{d}) = t$ (i.e., all coordinates of $\mathbf{c}$ and $\mathbf{d}$ are distinct.). The *pair group* $\sigma_S(\mathbf{c}, \mathbf{d})$ is the set of pairs $(\mathbf{p}, \mathbf{q}) \in [n]^k \times [n]^k$ such that (i) on coordinate S, $\Phi_S(\mathbf{p}) = \mathbf{c}$ and $\Phi_S(\mathbf{q}) = \mathbf{d}$, and (ii) on coordinate $\bar{S}$, $\mathbf{p}$ and $\mathbf{q}$ are the same, i.e., $\Phi_{\bar{S}}(\mathbf{p}) = \Phi_{\bar{S}}(\mathbf{q})$. Namely,

$$\sigma_S(\mathbf{c}, \mathbf{d}) = \left\{(\mathbf{p}, \mathbf{q}) \in [n]^k \times [n]^k : \Big(\mathbf{c} = \Phi_S(\mathbf{p})\Big) \wedge \Big(\mathbf{d} = \Phi_S(\mathbf{q})\Big) \wedge \Big(\Phi_{\bar{S}}(\mathbf{p}) = \Phi_{\bar{S}}(\mathbf{q})\Big)\right\}. \tag{4.2}$$

To give some intuition for the above definitions, we note that for every $\mathbf{a} \in [n]^{|\bar{S}|}$, there is a unique pair $(\mathbf{p}, \mathbf{q}) \in \sigma_S(\mathbf{c}, \mathbf{d})$ with $\mathbf{a} = \Phi_{\bar{S}}(\mathbf{p}) = \Phi_{\bar{S}}(\mathbf{q})$, and so $|\sigma_S(\mathbf{c}, \mathbf{d})| = n^{|\bar{S}|}$. On the other hand, for every pair

$(\mathbf{p}, \mathbf{q}) \in [n]^k \times [n]^k$ with $\mathbf{p} \neq \mathbf{q}$, there is a unique non-emtpy $S \subseteq [k]$ such that $\mathbf{p}$ and $\mathbf{q}$ are distinct on exactly coordinates in S. Therefore, $(\mathbf{p}, \mathbf{q})$ belongs to exactly one pair group $\sigma_S(\mathbf{c}, \mathbf{d})$. It follows that we can partition the summation in Err according to the pair groups:

$$Err = \sum_{\substack{S \subseteq [k] \\ S \neq \emptyset}} \sum_{\substack{\mathbf{c}, \mathbf{d} \in [n]^{|S|}, \\ \mathrm{Ham}(\mathbf{c}, \mathbf{d}) = |S|}} \sum_{\substack{(\mathbf{p}, \mathbf{q}) \in \\ \sigma_S(\mathbf{c}, \mathbf{d})}} H(\mathbf{p}) H(\mathbf{q}) \vec{v}_{\mathbf{p}} \vec{v}_{\mathbf{q}}. \tag{4.3}$$

We next observe that for any pair $(\mathbf{p}, \mathbf{q}) \in \sigma_S(\mathbf{c}, \mathbf{d})$, since $\mathbf{p}$ and $\mathbf{q}$ agree on coordinates in $\bar{S}$, the value of the product $H(\mathbf{p})H(\mathbf{q})$ depends only on S, $\mathbf{c}$ and $\mathbf{d}$. More precisely,

$$H(\mathbf{p})H(\mathbf{q}) = \prod_{i \in [k]} h_i(p_i) h_i(q_i) = \left(\prod_{i \in S} h_i(p_i) h_i(q_i) \right) \cdot \left(\prod_{i \in \bar{S}} h_i(p_i)^2 \right) = \prod_{i \in S} h_i(p_i) h_i(q_i),$$

which depends only on S, $\mathbf{c}$ and $\mathbf{d}$ since $\Phi_S(\mathbf{p}) = \mathbf{c}$ and $\Phi_S(\mathbf{q}) = \mathbf{d}$. This motivates the definition of *projected hashing*.

Definition 4.6. *(Projected hashing)* Let $S = \{s_1, s_2, \ldots, s_t\}$ be a subset of $[k]$, where $s_1 < s_2 < \cdots < s_j$. Let $\mathbf{c} \in [n]^t$. We define the *projected hashing* $H_S(\mathbf{c}) = \prod_{i \leq t} h_{s_i}(c_i)$.

We can now translate the random variable Err as follows:

$$Err = \sum_{\substack{S \subseteq [k] \\ S \neq \emptyset}} \sum_{\substack{\mathbf{c}, \mathbf{d} \in [n]^{|S|}, \\ \mathrm{Ham}(\mathbf{c}, \mathbf{d}) = |S|}} \left(H_S(\mathbf{c}) H_S(\mathbf{d}) \sum_{\substack{(\mathbf{p}, \mathbf{q}) \in \\ \sigma_S(\mathbf{c}, \mathbf{d})}} \vec{v}_{\mathbf{p}} \vec{v}_{\mathbf{q}} \right). \tag{4.4}$$

Fix a pair group $\sigma_S(\mathbf{c}, \mathbf{d})$, we next consider the sum $\sum_{(\mathbf{p}, \mathbf{q}) \in \sigma_S(\mathbf{c}, \mathbf{d})} \vec{v}_{\mathbf{p}} \vec{v}_{\mathbf{q}}$. Recall that for every $\mathbf{a} \in [n]^{|\bar{S}|}$, there is a unique pair $(\mathbf{p}, \mathbf{q}) \in \sigma_S(\mathbf{c}, \mathbf{d})$ with $\mathbf{a} = \Phi_{\bar{S}}(\mathbf{p}) = \Phi_{\bar{S}}(\mathbf{q})$. The sum can be viewed as the inner product of two vectors of dimension $n^{|\bar{S}|}$ with entries indexed by $\mathbf{a} \in [n]^{|\bar{S}|}$. To formalize this observation, we introduce the definition of *hyper-projection* as follows.

Definition 4.7. *(Hyper-projection)* Let $\vec{v} \in R^{n^k}$, $S \subseteq [k]$, and $\mathbf{c} \in [n]^{|S|}$. The *hyper-projection* $\Upsilon_{S,\mathbf{c}}(\vec{v})$ of $\vec{v}$ (with respect to S and $\mathbf{c}$) is a vector $\vec{w} = \Upsilon_{S,\mathbf{c}}(\vec{v})$ in $\mathrm{R}^{[n]^{k-|S|}}$ such that $\vec{w}_{\mathbf{d}} = \vec{v}_{\Phi_{\bar{S}}^{-1}(\mathbf{c}, \mathbf{d})}$ for all $\mathbf{d} \in [n]^{k-|S|}$.

Using the above definition, we continue to rewrite the Err as

$$Err = \sum_{\substack{S \subseteq [k] \\ S \neq \emptyset}} \sum_{\substack{\mathbf{c}, \mathbf{d} \in [n]^{|S|}, \\ \mathrm{Ham}(\mathbf{c}, \mathbf{d}) = |S|}} H_S(\mathbf{c}) H_S(\mathbf{d}) \cdot \langle \Upsilon_{S,\mathbf{c}}(\vec{v}), \Upsilon_{S,\mathbf{d}}(\vec{v}) \rangle. \tag{4.5}$$

Finally, we consider the product $H_S(\mathbf{c})H_S(\mathbf{d})$ again and introduce the following definition to further simplify the Err.

Definition 4.8. *(Similarity and dominance)* Let t be a positive integer.

- Two pairs of index vectors $(\mathbf{c}, \mathbf{d}) \in [n]^t \times [n]^t$ and $(\mathbf{a}, \mathbf{b}) \in [n]^t \times [n]^t$ are *similar* if for all $i \in [t]$, the two sets $\{c_i, d_i\}$ and $\{a_i, b_i\}$ are equal. We denote this as $(\mathbf{a}, \mathbf{b}) \sim (\mathbf{c}, \mathbf{d})$.
- Let $\mathbf{c}$ and $\mathbf{d} \in [n]^t$ be two index vectors. We say $\mathbf{c}$ *is dominated by* $\mathbf{d}$ if $c_i < d_i$ for all $i \in [t]$. We denote this as $\mathbf{c} \prec \mathbf{d}$. Note that $\mathbf{c} \prec \mathbf{d} \Rightarrow \mathrm{Ham}(\mathbf{c}, \mathbf{d}) = t$.

Now, note that if $(\mathbf{a}, \mathbf{b}) \sim (\mathbf{c}, \mathbf{d})$, then $H_S(\mathbf{a})H_S(\mathbf{b}) = H_S(\mathbf{c})H_S(\mathbf{d})$ since the value of the product $H_S(\mathbf{c})H_S(\mathbf{d})$ depends on the values $\{c_i, d_i\}$ only as a set. It is also not hard to see that $\sim$ is an equivalence

relation, and for every equivalent class $[(\mathbf{a},\mathbf{b})]$, there is a unique $(\mathbf{c},\mathbf{d}) \in [(\mathbf{a},\mathbf{b})]$ with $\mathbf{c} \prec \mathbf{d}$. Therefore, we can further rewrite the Err as

$$Err = \sum_{\substack{S\subseteq[k]\\ S\neq\emptyset}} \sum_{\mathbf{c}\prec\mathbf{d}\in[n]^{|S|}} H_S(\mathbf{c})H_S(\mathbf{d}) \cdot \left(\sum_{(\mathbf{a},\mathbf{b})\sim(\mathbf{c},\mathbf{d})} \langle \Upsilon_{S,\mathbf{a}}(\vec{v}), \Upsilon_{S,\mathbf{b}}(\vec{v})\rangle \right). \tag{4.6}$$

We are ready to bound the term $\mathrm{E}[Err^2]$ by expanding the square of the sum according to Equation (4.6). We first show in Lemma 4.9 below that all the cross terms in the following expansion vanish.

$$\mathrm{Var}[Y] = \sum_{\substack{S,S'\subseteq[k]\\ S,S'\neq\emptyset}} \sum_{\substack{\mathbf{c}\prec\mathbf{d}\in[n]^{|S|}\\ \mathbf{c}'\prec\mathbf{d}'\in[n]^{|S|'}}} \mathrm{E}[H_S(\mathbf{c})H_S(\mathbf{d})H_{S'}(\mathbf{c}')H_{S'}(\mathbf{d}')]\cdot$$
$$\left[\left(\sum_{(\mathbf{a},\mathbf{b})\sim(\mathbf{c},\mathbf{d})} \langle \Upsilon_{S,\mathbf{a}}(\vec{v}), \Upsilon_{S,\mathbf{b}}(\vec{v})\rangle \right) \left(\sum_{(\mathbf{a}',\mathbf{b}')\sim(\mathbf{c}',\mathbf{d}')} \langle \Upsilon_{S',\mathbf{a}'}(\vec{v}), \Upsilon_{S',\mathbf{b}'}(\vec{v})\rangle \right) \right]. \tag{4.7}$$

Lemma 4.9. *Let S and S' be subsets of $[k]$, and $\mathbf{c} \prec \mathbf{d} \in [n]^{|S|}$ and $\mathbf{c}' \prec \mathbf{d}' \in [n]^{|S'|}$ index vectors. We have $E[H_S(\mathbf{c})H_S(\mathbf{d})H_{S'}(\mathbf{c}')H_{S'}(\mathbf{d}')] \in \{0,1\}$. Furthermore, we have $\mathrm{E}[H_S(\mathbf{c})H_S(\mathbf{d})H_{S'}(\mathbf{c}')H_{S'}(\mathbf{d}')] = 1$ iff $(S = S') \wedge (\mathbf{c} = \mathbf{c}') \wedge (\mathbf{d} = \mathbf{d}')$.*

Proof. Recall that $h_1, \ldots, h_k$ are independent copies of 4-wise independent uniform random variables over $\{-1, 1\}$. Namely, for every $i \in [k]$, $h_i(1), \ldots, h_i(n)$ are 4-wise independent, and $h_1(\cdot), \ldots, h_k(\cdot)$ are mutually independent. Observe that for every $i \in [k]$, there are at most 4 terms out of $h_i(1), \ldots, h_i(n)$ appearing in the product $H_S(\mathbf{c})H_S(\mathbf{d})H_{S'}(\mathbf{c}')H_{S'}(\mathbf{d}')$. It follows that all distinct terms appearing in $H_S(\mathbf{c})H_S(\mathbf{d})H_{S'}(\mathbf{c}')H_{S'}(\mathbf{d}')$ are mutually independent uniform random variable over $\{-1, 1\}$. Therefore, the expectation is either 0, if there is some $h_i(j)$ that appears an odd number of times, or 1, if all $h_i(j)$ appear an even number of times. By inspection, the latter case happens if and only if $(S = S') \wedge (\mathbf{c} = \mathbf{c}') \wedge (\mathbf{d} = \mathbf{d}')$. ∎

By the above lemma, Equation (4.7) is simplified to

$$\mathrm{Var}[Y] = \sum_{\substack{S\subseteq[k]\\ S\neq\emptyset}} \sum_{\mathbf{c}\prec\mathbf{d}\in[n]^{|S|}} \left(\sum_{(\mathbf{a},\mathbf{b})\sim(\mathbf{c},\mathbf{d})} \langle \Upsilon_{S,\mathbf{a}}(\vec{v}), \Upsilon_{S,\mathbf{b}}(\vec{v})\rangle \right)^2. \tag{4.8}$$

We next apply the Cauchy-Schwartz inequality three times to bound the above formula. Consider a subset $S \subseteq [k]$ and a pair $\mathbf{c} \prec \mathbf{d} \in [n]^{|S|}$. Note that there are precisely $2^{|S|}$ pairs $(\mathbf{a}, \mathbf{b})$ such that $(\mathbf{a}, \mathbf{b}) \sim (\mathbf{c}, \mathbf{d})$. Thus, by the Cauchy-Schwartz inequality:

$$\begin{aligned}
\left(\sum_{\substack{(\mathbf{a},\mathbf{b})\in[n]^{|S|}\\ (\mathbf{a},\mathbf{b})\sim(\mathbf{c},\mathbf{d})}} \langle \Upsilon_{S,\mathbf{a}}(\vec{v}), \Upsilon_{S,\mathbf{b}}(\vec{v})\rangle \right)^2 &\leq 2^{|S|} \sum_{\substack{(\mathbf{a},\mathbf{b})\in[n]^{|S|}\\ (\mathbf{a},\mathbf{b})\sim(\mathbf{c},\mathbf{d})}} (\langle \Upsilon_{S,\mathbf{a}}, \Upsilon_{S,\mathbf{b}}\rangle)^2 \\
&\leq 2^{|S|} \sum_{\substack{(\mathbf{a},\mathbf{b})\in[n]^{|S|}\\ (\mathbf{a},\mathbf{b})\sim(\mathbf{c},\mathbf{d})}} \langle \Upsilon_{S,\mathbf{a}}(\vec{v}), \Upsilon_{S,\mathbf{a}}(\vec{v})\rangle \cdot \langle \Upsilon_{S,\mathbf{b}}, \Upsilon_{S,\mathbf{b}}(\vec{v})\rangle.
\end{aligned}$$

Notice that in the second inequality, we applied Cauchy-Schwartz in a component-wise manner. Next, for a subset $S \subseteq [k]$, we can apply the Cauchy-Schwartz inequality a third time (from the third line to the fourth

line) as follows:

$$\begin{aligned}
&\sum_{\mathbf{c}\prec\mathbf{d}\in[n]^{|S|}} \left(\sum_{\substack{(\mathbf{a},\mathbf{b})\in[n]^{|S|}\\(\mathbf{a},\mathbf{b})\sim(\mathbf{c},\mathbf{d})}} \langle \Upsilon_{S,\mathbf{a}}(\vec{v}), \Upsilon_{S,\mathbf{b}}(\vec{v})\rangle \right)^2 \\
\leq\ & 2^{|S|} \sum_{\mathbf{c}\prec\mathbf{d}\in[n]^{|S|}} \sum_{\substack{(\mathbf{a},\mathbf{b})\in[n]^{|S|}\\(\mathbf{a},\mathbf{b})\sim(\mathbf{c},\mathbf{d})}} \langle \Upsilon_{S,\mathbf{a}}(\vec{v}), \Upsilon_{S,\mathbf{a}}(\vec{v})\rangle \cdot \langle \Upsilon_{S,\mathbf{b}}(\vec{v}), \Upsilon_{S,\mathbf{b}}(\vec{v})\rangle \\
=\ & 2^{|S|} \sum_{\substack{\mathbf{c},\mathbf{d}\in[n]^{|S|}\\ \mathrm{Ham}(\mathbf{c},\mathbf{d})=|S|}} \langle \Upsilon_{S,\mathbf{c}}(\vec{v}), \Upsilon_{S,\mathbf{c}}(\vec{v})\rangle \cdot \langle \Upsilon_{S,\mathbf{d}}(\vec{v}), \Upsilon_{S,\mathbf{d}}(\vec{v})\rangle \\
\leq\ & 2^{|S|} \sum_{\mathbf{c},\mathbf{d}\in[n]^{|S|}} \langle \Upsilon_{S,\mathbf{c}}(\vec{v}), \Upsilon_{S,\mathbf{c}}(\vec{v})\rangle \cdot \langle \Upsilon_{S,\mathbf{d}}(\vec{v}), \Upsilon_{S,\mathbf{d}}(\vec{v})\rangle \\
=\ & 2^{|S|} \left(\sum_{\mathbf{c}\in[n]^{|S|}} \langle \Upsilon_{S,\mathbf{c}}(\vec{v}), \Upsilon_{S,\mathbf{c}}(\vec{v})\rangle \right)^2 .
\end{aligned}$$

Finally, we note that by definition, we have $\sum_{\mathbf{c}\in[n]^{|S|}} \langle \Upsilon_{S,\mathbf{c}}(\vec{v}), \Upsilon_{S,\mathbf{c}}(\vec{v})\rangle = ||\vec{v}||^2$, which equals to $\mathrm{E}[Y]$. It follows that the variance in Equation (4.8) can be bounded by

$$\mathrm{Var}[Y] \leq \sum_{S\subseteq[k],S\neq\emptyset} 2^{|S|}\cdot \mathrm{E}[Y]^2 = \mathrm{E}[Y]^2 \sum_{i=1}^{k} \binom{k}{i} 2^i = (3^k-1)\mathrm{E}[Y]^2,$$

which finishes the proof of Lemma 4.1.

5. Conclusion

There remain several open questions left in this space. Lower bounds, particularly bounds that depend non-trivially on the dimension k, would be useful. There may still be room for better algorithms for testing k-wise independence in this manner using the ℓ_2 norm. A natural generalization would be to find a particularly efficient algorithm for testing k-out-of-s-wise independence (other than handling each set of k variable separately). More generally, a question given in [14], to identify random variables whose correlation exceeds some threshold according to some measure, remains widely open.

References

[1] Noga Alon, Alexandr Andoni, Tali Kaufman, Kevin Matulef, Ronitt Rubinfeld, and Ning Xie. Testing k-wise and almost k-wise independence. In *STOC '07: Proceedings of the thirty-ninth annual ACM symposium on Theory of computing*, pages 496–505, New York, NY, USA, 2007. ACM.

[2] Noga Alon, László Babai, and Alon Itai. A fast and simple randomized parallel algorithm for the maximal independent set problem. *J. Algorithms*, 7(4):567–583, 1986.

[3] Noga Alon, Oded Goldreich, and Yishay Mansour. Almost k-wise independence versus k-wise independence. *Inf. Process. Lett.*, 88(3):107–110, 2003.

[4] Noga Alon, Yossi Matias, and Mario Szegedy. The space complexity of approximating the frequency moments. *J. Comput. Syst. Sci.*, 58(1):137–147, 1999.

[5] T. Batu, L. Fortnow, E. Fischer, R. Kumar, R. Rubinfeld, and P. White. Testing random variables for independence and identity. In *FOCS '01: Proceedings of the 42nd IEEE symposium on Foundations of Computer Science*, page 442, Washington, DC, USA, 2001. IEEE Computer Society.

[6] Tugkan Batu, Ravi Kumar, and Ronitt Rubinfeld. Sublinear algorithms for testing monotone and unimodal distributions. In *STOC '04: Proceedings of the thirty-sixth annual ACM symposium on Theory of computing*, pages 381–390, New York, NY, USA, 2004. ACM.

[7] Vladimir Braverman and Rafail Ostrovsky. Measuring k-wise independence of streaming data. *CoRR*, abs/0806.4790v1, 2008.

[8] Vladimir Braverman and Rafail Ostrovsky. Measuring independence of datasets. *CoRR*, abs/0903.0034, 2009.

[9] Vladimir Braverman and Rafail Ostrovsky. AMS without $\mathbf{4}$-wise independence on product domains. *CoRR*, abs/0806.4790, 2009.

[10] Paul G. Brown and Peter J. Hass. Bhunt: automatic discovery of fuzzy algebraic constraints in relational data. In *VLDB '2003: Proceedings of the 29th international conference on Very large data bases*, pages 668–679. VLDB Endowment, 2003.

[11] Kai-Min Chung, Zhenming Liu, and Michael Mitzenmacher. Testing k-wise independence over streaming data. Unpublished manuscript, available at `http://www.eecs.harvard.edu/~michaelm/postscripts/sketchexttemp.pdf`, 2009.

[12] Oded Goldreich and Dana Ron. On testing expansion in bounded-degree graphs. Technical report, Electronic Colloquium on Computational Complexity, 2000.

[13] Ihab F. Ilyas, Volker Markl, Peter Haas, Paul Brown, and Ashraf Aboulnaga. Cords: automatic discovery of correlations and soft functional dependencies. In *SIGMOD '04: Proceedings of the 2004 ACM SIGMOD international conference on Management of data*, pages 647–658, New York, NY, USA, 2004. ACM.

[14] Piotr Indyk and Andrew McGregor. Declaring independence via the sketching of sketches. In *SODA '08: Proceedings of the nineteenth annual ACM-SIAM symposium on Discrete algorithms*, pages 737–745, Philadelphia, PA, USA, 2008. Society for Industrial and Applied Mathematics.

[15] Ralph Kimball and Joe Caserta. *The Data Warehouse ETL Toolkit: Practical Techniques for Extracting, Cleanin*. Wiley, 1 edition, September 2004.

[16] E. L. Lehmann and Springer. *Testing Statistical Hypotheses (Springer Texts in Statistics)*. Springer, January 1997.

[17] Michael Mitzenmacher and Eli Upfal. *Probability and Computing: Randomized Algorithms and Probabilistic Analysis*. Cambridge University Press, New York, NY, USA, 2005.

[18] Viswanath Poosala and Yannis E. Ioannidis. Selectivity estimation without the attribute value independence assumption. In *VLDB '97: Proceedings of the 23rd International Conference on Very Large Data Bases*, pages 486–495, San Francisco, CA, USA, 1997. Morgan Kaufmann Publishers Inc.

[19] Ronitt Rubinfeld and Rocco A. Servedio. Testing monotone high-dimensional distributions. In *STOC '05: Proceedings of the thirty-seventh annual ACM symposium on Theory of computing*, pages 147–156, New York, NY, USA, 2005. ACM.

[20] Mikkel Thorup and Yin Zhang. Tabulation based 4-universal hashing with applications to second moment estimation. In *SODA '04: Proceedings of the fifteenth annual ACM-SIAM symposium on Discrete algorithms*, pages 615–624, Philadelphia, PA, USA, 2004. Society for Industrial and Applied Mathematics.

This work will be licensed under the Creative Commons Attribution-NoDerivs License. To view a copy of this license, visit http://creativecommons.org/licenses/by-nd/3.0/.