Declaring Independence via the Sketching of Sketches

Piotr Indyk
indyk@mit.edu

Andrew McGregor
andrewm@ucsd.edu

Abstract

We consider the problem of identifying correlations in data streams. Surprisingly, our work seems to be the first to consider this natural problem. In the centralized model, we consider a stream of pairs $(i, j) \in [n]^2$ whose frequencies define a joint distribution (X, Y). In the distributed model, each coordinate of the pair may appear separately in the stream. We present a range of algorithms for approximating to what extent X and Y are independent, i.e., how close the joint distribution is to the product of the marginals. We consider various measures of closeness including ℓ_1, ℓ_2, and the mutual information between X and Y. Our algorithms are based on "sketching sketches", i.e., composing small-space linear synopses of the distributions. Perhaps ironically, the biggest technical challenges that arise relate to ensuring that different components of our estimates are sufficiently independent.

1 Introduction

The data-stream model has enjoyed considerable attention for more than ten years and a wide range of problems have been tackled including estimating quantiles (e.g., [19, 14, 22]), frequency moments and finding frequent items (e.g., [2, 25, 9, 7]), estimating the difference between the underlying distribution of two streams (e.g., [24, 13, 20]), histograms and clustering (e.g., [21, 11, 12]), graph problems (e.g., [5, 16, 17]), among others. For a more comprehensive overview of the area the reader is directed to excellent surveys [28, 4]. Surprisingly, the very natural problem of identifying correlations in data-streams has, to date, not been considered. Note that two random variables can have similar distributions and yet be entirely independent. Conversely, one can be a function of the other and yet have distributions that are far apart. In this paper, we consider the problem of approximating the degree of correlation between two random variables.

Identifying correlations is a fundamental problem in a variety of settings including network monitoring, sensor networks, and communication applications. For example, correlation between the traffic observed at two different routers could provide an early warning of the onset of a coordinated denial of service attack or the existence of zombie machines under some central control. In a sensor network, correlations between concurrent measurements at different sensors may help determine the geometry of the area in which the sensors have been deployed. In many communication problems, multiple signals need to be transmitted simultaneously. If these signals are correlated that it makes sense to jointly encode the signals [15].

Another application would be to determine the utility of various k-gram models of text [27]. Models for language become more accurate as they take into account more context of a given work or character, e.g., the previous k words or characters. However, the computational overhead of using the model increases as one increases k. Determining the strength of the correlation between characters k apart in the string would help determine the optimum value of k.

Empirical Distributions and Partial Independence: Consider two random variables X and Y on $[n]$. X and Y are independent if $\Pr[X = i, Y = j] = \Pr[X = i] \Pr[Y = j]$ for all $i, j \in [n]$. In this paper we consider X, Y and the joint random variable (X, Y) being defined empirically: the stream codifes pairs from $[n]^2$ and $\Pr[X = i, Y = j]$ is defined as the fraction of pairs equal to (i, j), $\Pr[X = i]$ will the fraction of pairs of the form $(i, \cdot)$ etc.[1] This being the case, even if the pairs originate from two independent sources, it is unlikely that the distributions defined empirically will be perfectly independent. However, for sufficiently long streams, if the pairs are defined by two independent sources then $\Pr[X = i, Y = i]$

[1] Considering empirical distributions rather than the distribution of a supposed "source" is standard in the literature on the data-stream model. The problem of combining the restrictions of the data-stream model with the notion of learning something about a source has only recently been considered [10, 23].

should approach $\Pr[X=i]\Pr[Y=i]$ for all i,j. When processing finite length streams it would therefore be useful to approximate this distance between the joint distribution is to the product distribution. There are numerous ways of doing this. In the paper we consider approximating the ℓ_1 and ℓ_2 difference between the joint distribution and the product distribution. We also consider approximating the *mutual information* between X and Y:

$$I(X;Y) = H(X) - H(X|Y)$$

where $H(X) = -\sum_i \Pr[X=i]\lg\Pr[X=i]$ is the entropy of the distribution X and

$$H(X|Y) = \sum_{i,j} \Pr[X=i,Y=j]\lg\frac{\Pr[Y=j]}{\Pr[X=i,Y=j]}$$

is the conditional entropy. The mutual information should be zero if X and Y are independent. There is a natural relationship between mutual information and the ℓ_1 distance between the joint distribution and the product distribution. If $p_i = \Pr[X=i]$ and $p_i^j = \Pr[X=i|Y=j]$ then the ℓ_1 distance equals $E_{i\sim Y}[\ell_1(p,p^j)]$ while $I(X;Y) = E_{i\sim Y}[D_{\mathrm{KL}}(p,p^j)]$ where D_{KL} is the Kullback-Liebler divergence, $D_{\mathrm{KL}}(p,q) = \sum_i p_i \lg(q_i/p_i)$.

There are numerous ways in which the stream can codify pairs from $[n]^2$. In the *centralized model* the elements of the stream are the pairs themselves. For example these pairs could correspond to the source IP's of two packets being forwarded at the same time at a network router. In the *distributed model* there are two streams that are being observed at two different locations. Each stream will define a marginal. We assume the stream are synchronized in the sense that we consider the i-th element of the stream defining X occurring at the same time as the i-th element of the stream defining Y. These two elements define the pair from $[n]^2$ and therefore empirically define a joint distribution (X,Y) as described above.

Our Results and Techniques: Most of our algorithms for estimating independence are based on sketching sketches, i.e., composing small-space synopses of data. Perhaps ironically, the main difficulty in this approach are ensuring that there is sufficient independence between components of our estimators. For example, standard results that show 4-wise independence is sufficient for various types are sketch, are not enough when composing these sketches. Similarly, subtle issues arise in the application of pseudo-random-generators that need to be addressed.

Our algorithms and lower-bounds include:

1. A 1-pass, $\tilde{O}(\epsilon^{-2}\log\delta^{-1})$-space, $(1+\epsilon)$-factor approximation (w.p. $1-\delta$) of the ℓ_2 difference between the joint and product distributions in the centralized model.

2. For the ℓ_1 difference between the joint and product distributions in the centralized model we present:

   (a) A 1-pass, $\tilde{O}(\log\delta^{-1})$-space, $(1+\epsilon)$-factor approximation w.p. $1-\delta$.

   (b) A 1-pass, $\tilde{O}(\epsilon^{-2}n\log\delta^{-1})$-space, $(1+\epsilon)$-factor approximation w.p. $1-\delta$.

   (c) A 2-pass, $\tilde{O}(\epsilon^{-4}\log^2\delta^{-1})$-space, ϵ-additive approximation w.p. $1-\delta$.

3. A 1-pass, $\tilde{O}(\epsilon^{-2}\log\delta^{-1})$-space, ϵ-additive approximation (w.p. $1-\delta$) of $I(X;Y)$ in the centralized model. We show that any 5/4-factor approximation (w.p. 3/4) requires $\Omega(n)$ space.

4. Finally we present mainly negative results in the distributed model in the distributed model.

Related Work: The problems of testing independence and k-wise independence were considered by Batu et al. [6] and Alon et al. [1] in the model in which independent samples from the joint distribution were available. Here the goal was primarily to minimize the number of samples required to determine whether the relevant random variables were independent or far from independent. In contrast, we consider the space complexity of estimating how close the variables are to being independent and consider the distributions to be empirically defined by the data available.

Chakribarti et al. [8] considered the problem of trying to determine the k-th order entropy of a string given the constraints of the data stream model. The k-th order entropy captures how "surprising" a character is in the string given one knows the preceding k characters. This is related to the correlation between characters that are k apart in the string since if two such characters are independent, then the entropy of a character given the previous k characters should be equal to the

entropy of a character given the previous $k-1$ characters.

Finally, we note that our problem is related, but different, to the problem of trying to estimate correlated aggregates [18, 3]. In this model, the stream consists of a series of pairs (x, y). The goal is to estimate functions such as the median of all x for which the corresponding y is less than the average value of y.

2 Preliminaries

Notation: Let $[n] := \{1, \ldots, n\}$ and write $x \in_R S$ to mean that x is randomly chosen from the multi-set S. For random variables X and Y on the same base set, we say $X \sim Y$ if they have the same distribution. We say $\hat{Q}$ is an (t, δ)-approx. for Q if for some $1/t \leq c \leq 1$, $\Pr[c \leq \hat{Q}/Q \leq tc] \geq 1 - \delta$. Similarly $\hat{Q}$ is an (ϵ, δ)-additive-approx. for Q if $\Pr[|\hat{Q} - Q| \leq \epsilon] \geq 1 - \delta$. $\tilde{O}$ omits all factors of $\text{polylog}(m, n)$. Finally we denote the ℓ_2 norm of x as $\|x\|$ and the ℓ_1 norm as $|x|$.

Lower-Bounds: The proofs of our lower bounds all use the following standard technique of reduction from a communication complexity problem. Rather than repeating details in each proof we review the general proof template here. Consider a 2-party communication problem in which Alice has input x and Bob has input y and together they wish to compute $f(x, y)$. We suppose that there exists a streaming algorithm $\mathcal{A}$ that takes P passes over a stream and uses W working memory to approximate some quantity. If there exists sets $S_A(x)$ and $S_B(y)$ such that the value returned by $\mathcal{A}$ on the stream formed by any ordering of $S_A(x) \cup S_B(y)$ determines $f(x, y)$, then there exists a $(2P-1)$-round protocol that requires $O(PW)$ bits: Alice runs $\mathcal{A}$ on $S_A(x)$, communicates the memory state of $\mathcal{A}$, Bob runs $\mathcal{A}$ initiated with this memory state on $S_B(x)$ and communicates the memory state of $\mathcal{A}$ to Alice and so on. Hence, a lower bound for the communication problem yields a lower-bound for P and/or W.

3 Approximating Partial Independence

Consider a stream $\langle a_1, \ldots, a_m \rangle$ where $a_k \in [n]^2$ and define random variables X and Y on $[n]$ by

$$\begin{aligned} \Pr[X = i, Y = j] &= |\{k : a_k = (i, j)\}|/m \\ \Pr[X = i] &= |\{k : a_k = (i, \cdot)\}|/m \\ \Pr[Y = j] &= |\{k : a_k = (\cdot, j)\}|/m. \end{aligned}$$

Let $r_{ij} = \Pr[X = i, Y = j]$, $p_i = \Pr[X = i]$, $q_j = \Pr[Y = j]$ and $s_{ij} = \Pr[X = i] \Pr[Y = j]$. I.e., r is the joint distribution, p and q are the marginal distributions of r, and s is the product distribution of the marginals. In the next three sections we present algorithms and lower-bounds for approximating the degree of independence between X and Y.

3.1 Approximating ℓ_2-independence: In the classic result of Alon, Matias, and Szegedy [2] it was shown that numerous 4-wise independent vectors $z \in \{-1, 1\}^{n^2}$ could be used to estimate the ℓ_2 difference between two distributions on $[n]^2$. For our application, the elements of z will be the elements of the outer product of two vectors $x, y \in \{-1, 1\}^n$ which are 4-wise independent. As such, they can be shown to 3-wise independent but not 4-wise independent, e.g.,

$$z_{1,1} z_{2,2} = (x_1^1 x_1^2)(x_2^1 x_2^2) = (x_1^1 x_2^2)(x_2^1 x_1^2) = z_{1,2} z_{2,1} \ .$$

However, by exploiting the geometry of the dependencies, the next lemma establishes that the elements of z are still sufficiently independent.

LEMMA 3.1. *Consider $x^1, x^2 \in \{-1, 1\}^n$ where each vector is 4-wise independent. Let $v \in \mathbb{R}^{n^2}$ and $z_{\mathbf{i}} = x_{i_1}^1 x_{i_2}^2$. Define $\Upsilon = (\sum_{\mathbf{i} \in [n]^2} z_{\mathbf{i}} v_{\mathbf{i}})^2$. Then $E[\Upsilon] = \sum_{\mathbf{i} \in [n]^2} v_{\mathbf{i}}^2$ and $\text{Var}[\Upsilon] \leq 3 (E[\Upsilon])^2$.*

Proof. While z is 2-wise independent, as noted above z is not 4-wise independent. However, z is "almost" 4-wise independent in the sense that $z_{\mathbf{i}}, z_{\mathbf{j}}, z_{\mathbf{k}}$, and $z_{\mathbf{l}}$ only fail to be independent if

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

and in which case $z_{\mathbf{i}} z_{\mathbf{j}} z_{\mathbf{k}} z_{\mathbf{l}} = 1$. Let D be the set of $(\mathbf{i}, \mathbf{j}, \mathbf{k}, \mathbf{l})$ that satisfy Eq. 3.1.

The expectation inequality follows in the standard way because z is 2-wise independent:

$$E[\Upsilon] = E\left[\left(\sum_{\mathbf{i} \in [n]^2} z_{\mathbf{i}} v_{\mathbf{i}}\right)^2\right] = \sum_{\mathbf{i} \in [n]^2} v_{\mathbf{i}}^2 \ .$$

Algorithm: $\|r - s\|$ **Approximation**

1. Compute $O(\epsilon^{-2} \log \delta^{-1})$ independent one-dimensional sketches:
   (a) Let $x, y \in_R \{-1, 1\}^n$ where the each vector is 4-wise independent
   (b) Let $t_1 = t_2 = t_3 = 0$
   (c) On seeing the stream element $a_k = (i, j)$:

$$t_1 \leftarrow t_1 + x_i y_j, \quad t_2 \leftarrow t_2 + x_i, \quad t_3 \leftarrow t_3 + y_j \ .$$

   Note that by the end of the stream:

$$\frac{t_1}{m} = \sum_{i,j \in [n]} x_i y_j r_{i,j}, \quad \frac{t_2}{m} = \sum_{i \in [n]} x_i p_i, \quad \frac{t_3}{m} = \sum_{j \in [n]} y_j q_j \ .$$

   (d) Let $\Upsilon = (t_1/m - t_2 t_3/m^2)^2$
2. Group into $O(\log \delta^{-1})$ groups of $O(\epsilon^{-2})$. Return the median of the mean of each group.

Figure 1: Single-Pass Approximation of $\|r - s\|$

We can rewrite the second moment as follows:

$$\begin{aligned} E\left[\Upsilon^2\right] &= E\left[\left(\sum_{\mathbf{i} \in [n]^2} z_{\mathbf{i}} v_{\mathbf{i}}\right)^4\right] \\ &= E\left[\sum_{\mathbf{i},\mathbf{j},\mathbf{k},\mathbf{l} \in [n]^2} z_{\mathbf{i}} z_{\mathbf{j}} z_{\mathbf{k}} z_{\mathbf{l}} v_{\mathbf{i}} v_{\mathbf{j}} v_{\mathbf{k}} v_{\mathbf{l}}\right] \\ &= \sum_{(\mathbf{i},\mathbf{j},\mathbf{k},\mathbf{l}) \in D} v_{\mathbf{i}} v_{\mathbf{j}} v_{\mathbf{k}} v_{\mathbf{l}} \end{aligned}$$

We now note that,

$$\begin{aligned} D = \{(\mathbf{i},\mathbf{j},\mathbf{k},\mathbf{l}) : \exists a, b, c, d \in [n] : \\ \{\mathbf{i},\mathbf{j},\mathbf{k},\mathbf{l}\} = \{(a,b),(a,d),(b,c),(b,d))\}\} \ , \end{aligned}$$

i.e., the quadruples in D are (possibly degenerate) rectangles. Because of this, we may associate a two pairs to each element of D corresponding to the both pairs of opposite corners. Furthermore,

$$\begin{aligned} 2 v_{\mathbf{i}} v_{\mathbf{j}} v_{\mathbf{k}} v_{\mathbf{l}} \leq \min\{&(v_{\mathbf{i}} v_{\mathbf{j}})^2 + (v_{\mathbf{k}} v_{\mathbf{l}})^2, \\ &(v_{\mathbf{i}} v_{\mathbf{k}})^2 + (v_{\mathbf{j}} v_{\mathbf{l}})^2, (v_{\mathbf{i}} v_{\mathbf{l}})^2 + (v_{\mathbf{j}} v_{\mathbf{k}})^2\} \ . \end{aligned}$$

Hence we may charge each quadruple in D to two diagonal pairs such that each diagonal pair is charged a total of 3 times its contribution. The result follows since $\mathrm{Var}\left[\Upsilon\right] \leq E\left[\Upsilon^2\right]$.

Constructing $\sum_{i,j \in [n]} x_i y_j r_{i,j}$ is simple since the pairs (i, j) arrive together. It turns out the constructing $\sum_{i,j \in [n]} x_i y_j p_i q_j$ is also simple because a sketch of a product of distribution is the product of sketches of the distributions.

LEMMA 3.2. *Consider $x^1, x^2 \in \{-1, 1\}^n$ and let $v^1, v^2 \in \mathbb{R}^n$. Then*

$$(x^1.v^1)(x^2.v^2) = \sum_{\mathbf{i} \in [n]^2} x^1_{i_1} x^2_{i_2} v^1_{i_1} v^2_{i_2} \ .$$

The algorithm is presented in Fig. 1 and the proof of correctness is given in the next theorem.

THEOREM 3.1. *There exists a single-pass, $\tilde{O}(\epsilon^{-2} \log \delta^{-1})$-space $(1 + \epsilon, \delta)$ approx. for $\|r - s\|$.*

Proof. By appealing to Lemma 3.1 and Lemma 3.2, $E\left[\Upsilon\right] = \sum_{i,j \in [n]} (r_{i,j} - p_i q_j)^2$. By Lemma 3.1 and the Chebychev bound, averaging $O(\epsilon^{-2})$ independent Υ returns a $(1 + \epsilon, 1/4)$-approximation. Taking the median of $O(\log \delta^{-1})$ averages returns an $(1 + \epsilon, \delta)$-approximation as desired.

It remains to be argued that the space requirement is as stated. This follows because there are only $O(\epsilon^{-2} \log \delta^{-1})$ independent estimators and each only requires $O(\log m + \log n)$ space. The 4-wise independent vectors x and y can be constructed using standard techniques, e.g., via using the parity check matrix of BCH codes [2].

3.2 Approximating ℓ_1-independence: In this section we use ideas from Indyk [24] to design a small-space $O(\log n)$ multiplicative approximation for $|r-s|$. Our approach is similar to that used to estimate $\|r-s\|$ in the previous section but we face two new obstacles in the composition of the appropriate sketches. First, 4-wise-independent vectors will not be sufficient and we will need to resort to the machinery of pseudo-random-generators (PGR). This will introduce a subtle issue concerning the order in which random bits are accessed. Secondly, since there is no sketch of ℓ_1 into small dimension ℓ_1, the sketches do not compose as easily as they did with ℓ_2. In particular, since the median operation is not linear we will not be able to achieve a $(1+\epsilon,\delta)$-approx. in small space. Rather we will be forced to use T-truncated-Cauchy distributed vectors for the inner sketch and the usual Cauchy distributed vectors for the outer sketch.

Definition 3.1. (T-truncated-Cauchy) *The Cauchy distribution has density function $\pi^{-1}/(1+x^2)$. Let $T>0$, $X \sim$ Cauchy, and define*

$$Y = -TI_{[X\le -T]} + XI_{[-T<X<T]} + TI_{[X\ge T]}$$

where $I_{[\cdot]}$ is the indicator function. We say $Y \sim$ T-truncated-Cauchy.

We now prove the necessary properties of the T-truncated-Cauchy distribution.

Lemma 3.3. *Let $T = 100n$ and $y \in \mathbb{R}^n$ where each coordinate $y_i \sim$ T-truncated-Cauchy. Let $v^1,\ldots,v^n \in \mathbb{R}^n$. Then for sufficiently large n,*

$$\Pr\left[1/100 \le \frac{\sum_j |y.v^j|}{\sum_j |v^j|} \le 20\ln n\right] \ge 9/10$$

Proof. Consider an element y_i of y. It can be shown [24, Lemma 5] that,

$$E\left[|y_i|\right] \le \ln(T^2+1)/\pi + O(1) \ ,$$

and hence, for sufficiently large n,

$$E\left[\sum_{j\in[n]} |y.v^j|\right] \le \sum_{i,j\in[n]} E\left[|y_i v_i^j|\right] \le \sum_{j\in[n]} |v^j| \ln n$$

Therefore, by Markov's inequality,

$$\Pr\left[\sum_j |y.v^j| \ge 20\ln n \sum_j |v^j|\right] \le 1/20 \ .$$

We now consider a lower bound for $\sum_{j\in[n]} |y.v^j|$. If T was infinite then because the Cauchy distribution is 1-stable, for any $j\in[n]$,

$$\Pr\left[|y.v^j| \le \frac{|v^j|}{100}\right] = \frac{1}{\pi}\int_{-1/100}^{1/100} \frac{1}{1+x^2} \le \frac{1}{100} \ .$$

However, the probability that there exists an i such that $|y_i| = T$ is at most

$$\frac{2n}{\pi}\int_{-\infty}^{-T} \frac{1}{1+x^2} \le \frac{2n}{T\pi} \le 1/100$$

for $T = 100n$. Therefore

$$\Pr\left[|y.v^j| \le |v^j|/100\right] \le 1/100 + 1/100 = 1/50 \ .$$

Then, by Markov's inequality,

$$\Pr\left[\sum_{j\in[n]} |y.v^j| \le \frac{1}{100}\sum_{j\in[n]} |v^j|\right] \le 1/20 \ .$$

The algorithm is presented in Fig. 2. The algorithm is an "ideal" algorithm in the sense that we assume access to random oracle and that computation can done with infinite precision. The precision issues can be addressed as in [24]. However, the argument used in [24] to show that pseudo-random-generators can be used rather than a fully random oracle needs to be slightly massaged for our purposes. In [24], the argument first considered a sorted stream such that the necessary random bits could be generated on the fly. Therefore the algorithm used sufficiently little space that it could be argued that only a few truly random bits were required. Unfortunately, in general it is impossible to order (i,j) pairs such that all pairs $(i,\cdot)$ are consecutive and all pairs $(\cdot,j)$ are consecutive.

To argue that we can still use a PRG for both x and y we proceed in two steps. First we consider the pairs ordered by grouping together on the first argument i. For a fixed i, we may use a truly random x_i and a pseudo-random y. Therefore, we can perform the whole computation using bit-by-bit access to truly random x, and sufficiently small space including the PRG seed. Now we repeat the argument on x to construct a small space algorithm with no oracle assumptions.

Theorem 3.2. *There exists a single-pass, $\tilde{O}(\ln \delta^{-1})$-space, $(O(\ln n),\delta)$-approx. for $|r-s|$.*

Approximation of $|r-s|$:

1. Repeat $O(\log \delta^{-1})$ times:
   (a) Let $s = O(1)$ and $T = 100n$.
   (b) Let $x^1, \ldots, x^s \in \mathbb{R}^n$ where each $x_i^j \sim$ Cauchy and is independent.
   (c) Let $y \in \mathbb{R}^n$ where each $y_i \sim T$-truncated-Cauchy and is independent.
   (d) Let $t_1^r = t_2^r = 0$ for $r \in [s]$ and $t_3 = 0$
   (e) On seeing the stream element $a_k = (i, j)$:

$$t_1^r \leftarrow t_1^r + x_i^r y_j, \quad t_2^r \leftarrow t_2^r + x_i^r, \quad t_3 \leftarrow t_3 + y_j \ .$$

   Note that by the end of the stream:

$$\frac{t_1^r}{m} = \sum_{i,j \in [n]} x_i^r y_j r_{i,j}, \quad \frac{t_2^r}{m} = \sum_{i \in [n]} x_i^r p_i, \quad \frac{t_3}{m} = \sum_{j \in [n]} y_j q_j \ .$$

   (f) Let $\Upsilon = \text{median}(|(t_1^1/m - t_2^1 t_3/m^2)^2|, \ldots, |(t_1^n/m - t_2^n t_3/m^2)^2|)$
2. Return the median of the esimators.

Figure 2: Single-Pass Approximation of $|r-s|$

Proof. Let $u_i = \sum_{j \in [n]} y_j (r_{i,j} - p_i q_j)$. By Lemma 3.3, $\Pr\left[1/100 \leq \frac{|u|}{|r-s|} \leq 20 \ln n\right] \geq 9/10$. Since the elements of x^r are p-stable,

$$|t_1^r/m - t_2^r t_3/m| = \left| \sum_{i,j \in [n]} x_i^r y_j (r_{i,j} - p_i q_j) \right| = \left| \sum_{i \in [n]} x_i^r u_i \right| \sim |u||X| \ ,$$

where $X \sim$ Cauchy. Hence, taking the median of $O(1)$ estimators yields a $(O(\log n), 1/5)$-approximation of $|u|$. Repeating the process $O(\log \delta^{-1})$ times and taking the median yields a $(O(\log n), \delta)$-approximation of $|r-s|$ as required.

Other Algorithms: In the remainder of this section we present two other approximation algorithms with different accuracy guarantees and resource use. If a $(1+\epsilon)$ multiplicative approximation is necessary then this is possible if we permit significant increase in space. Specifically, with $\tilde{O}(\epsilon^{-2} n \ln \delta^{-1})$ space we can compute p and use the normal ℓ_1-sketch algorithm to $(1+\epsilon, \delta/n)$-approx $|q - q^i|$ where $q^i = (r_{i,1}/p_i, \ldots, r_{i,n}/p_i)$. Call this approximation Υ_i. Then $\sum_i \Upsilon_i$ is an $(1+\epsilon, \delta)$-approx. for $|r-s|$.

THEOREM 3.3. *There exists a single-pass, $\tilde{O}(\epsilon^{-2} n \ln \delta^{-1})$-space $(1+\epsilon, \delta)$-approx. for $|r-s|$.*

Alternatively, there exists a two-pass algorithm that returns an (ϵ, δ)-additive-approx while only using $\tilde{O}(\epsilon^{-4} \ln \delta^{-1})$-space. In the first pass we take a set S of $O(\epsilon^{-2} \log \delta^{-1})$ samples from q.In the second pass, for each sample $i \in S$ we $(1+\epsilon, \delta/|S|)$-approx $|q - q^i|$. By an application of the Chernoff-bound, the mean of these estimates yields a $(1+\epsilon, \delta)$-additive-approx. for $|r-s|$ because $|r-s| = E_{i \sim p}[|q-q^i|]$ and $|q-q^i| \in [0,2]$.

THEOREM 3.4. *There exists a 2-pass, $\tilde{O}(\epsilon^{-4} \ln \delta^{-1})$-space (ϵ, δ)-additive-approx. for $|r-s|$.*

3.3 Mutual Information Approximation: Recall that the mutual information between two random variables is defined as $I(X;Y) = H(X) - H(X|Y)$ where $H(X) = -\sum_i p_i \lg p_i$ is the entropy and $H(X|Y) = -\sum_{i,j} r_{i,j} \lg(r_{i,j}/q_j)$ is the conditional entropy. In this section we show that arbitrary precision multiplicative approximation of the mutual information requires $\Omega(m)$ space. However, arbitrarily small additive approximation is possible in small space.

LEMMA 3.4. *There exists a one-pass*

$\tilde{O}(\epsilon^{-2}\log\delta^{-1})$-*space* (ϵ,δ)-*additive approx of the mutual information* $I(X;Y)$. *Any single pass,* $(5/4,1/4)$-*approx of* $I(X;Y)$ *requires* $\Omega(n)$ *space.*

Proof. For the upper-bound we write

$$I(X;Y) = H(X) + H(Y) - H(X,Y)$$

and approximate each term up to $\pm\epsilon/3$ using [8]. For the lower-bound let $(\sigma, j) \in \mathbb{F}_2^{n-1} \times [n]$ be an instance of `Index` where $\sum_i \sigma_i = (n-1)/2$. Let $S_A = \{(1,i),(2,i) : \sigma_i = 1, i \in [n-1]\}$ and $S_B = \{(1,j),(2,n)\}$. Note that $H(X) = 1$ and

$$\begin{aligned} H(X|Y) &= (1-q_j-q_n)H(1/2,1/2) + \\ &\qquad \sigma_j q_j H(2/3,1/3) \\ &= \frac{n-1-2\sigma_j+3\sigma_j H(2/3,1/3)}{n+1} \ . \end{aligned}$$

Therefore,

$$I(X;Y) = \frac{1}{n+1}\begin{cases} 2 & \text{if } \sigma_j = 0 \\ 4-3H(2/3,1/3) & \text{if } \sigma_j = 1 \end{cases}$$

and hence, approximating $I(X;Y)$ by a factor at most $\sqrt{2/(4-3H(2/3,1/3))} \geq 1.25$ determines the value of σ_i.

A similar result holds for approximating the conditional entropy.

Lemma 3.5. *There exists a one-pass* $\tilde{O}(\epsilon^{-2}\log\delta^{-1})$-*space* (ϵ,δ)-*additive approx of* $H(X|Y)$. *Any constant pass,* $(\alpha, 1/4)$-*approx of* $H(X|Y)$ *requires* $\Omega(n)$ *space for any constant* $\alpha > 1$.

Proof. For the upper-bound we write

$$H(X|Y) = H(X,Y) - H(Y)$$

and approximate each term upto $\pm\epsilon/2$ using [8]. For the lower bound, let (σ,ρ) be an instance of disjointness with weight of σ and ρ being $n/4$. $S_A = \{(1,i) : \sigma_i = 1\}$ and $S_B = \{(2,i) : \rho_i = 1\}$. Then $H(X|Y) = 0$ is $\sigma.\rho = 0$ and $H(X|Y) = 4/n$ if $\sigma.\rho = 1$.

4 Distributed Correlation

Consider a stream $S = \langle a_1, \ldots, a_{2m}\rangle$ where $a_k \in [n]\times[m]\times\{1,2\}$ and, for each $i \in [m]$ there exists exactly one item of the form $(\cdot, i, 1)$ and exactly one item of the form $(\cdot, i, 2)$. Define $b_i = (j,k)$ if $(j,i,1),(k,i,2) \in S$. Define the distributions r, p, q, r analogous to the definition in the previous section.

It turns out that estimating correlation in this model is hard even in the binary case (i.e., $n = 2$.) Note that there is natural connection to computing Hamming distances [29, 26, 13]: Consider $p = (1/2,1/2)$, $q = (1/2,1/2)$ and let $d = |\{k : b_k = (i,j), i \neq j\}|$. Then,

$$\begin{aligned} |r-s| &= \frac{1}{m}\left(\left|\frac{d}{2}-\frac{m}{4}\right| + 2\left|\frac{d}{2}-\frac{m}{4}\right| + \left|\frac{d}{2}-\frac{m}{4}\right|\right) \\ &= |2d/m - 1| \end{aligned}$$

Lemma 4.1. *For* $n = 2$*: any constant pass algorithm determining if* X *and* Y *are independent requires* $\Omega(m)$ *space but an* (ϵ,δ)*-additive-approx. to* $|r-s|$ *is possible in* $O(\epsilon^{-2}\log\delta^{-1})$ *space. More generally, an* (ϵ,δ)*-additive-approx. is possible with* $O(n^2\epsilon^{-2}\log n\delta^{-1})$ *space.*

Proof. For the lower-bound let $(\sigma,\rho) \in \mathbb{F}_2^n \times \mathbb{F}_2^n$ be an instance of `Disjointness`.

$$\begin{aligned} S_A &= \{(0,4i+1,1),(\sigma_i,4i+2,1), \\ &\qquad (\bar{\sigma}_i,4i+3,1),(1,4i+4,1) : i \in [m]\} \\ S_B &= \{(\bar{\rho}_i,4i+1,2),(\rho_i,4i+2,2), \\ &\qquad (0,4i+3,2),(1,4i+4,2) : i \in [m]\} \end{aligned}$$

Note that $p = (1/2,1/2)$ and $q = (1/2,1/2)$. Therefore, $|r-s| = |d/(2m)-1|$. Since $d = 2m$ iff $\rho.\sigma = 0$ the lower-bound follows.

For the upper-bound, let S be a set of $O(n^2\epsilon^{-2}\log n\delta^{-1})$ values from $[m]$ chosen at random (with replacement). As the stream arrives compute p_i and q_i for all $i,j \in [n]$ and

$$\tilde{r}_{i,j} = |\{a_k = (i,j) : k \in |S|\}|/|S| \ .$$

Then, by an application of the Chernoff-bound and the union bound, $|\tilde{r}-r| \leq \epsilon$ with probability at least $1-\delta$. Hence $|s-\tilde{r}| = |s-r| \pm \epsilon$.

5 Conclusions and Open Problems

We presented a range of algorithms for approximating the degree of correlation between two random variables defined empirically by a stream.

An obvious open question is to improve the approximation of the ℓ_1 distance between the joint distribution and the product distribution. Is a one-pass $(1+\epsilon,\delta)$-approximation possible in poly-logarithmic space? More generally, extending the algorithms to test for k-wise independence would

be interesting. Another related problem that deserves attention is the problem of identifying pairs of random variables whose "correlation" (according to some measure) exceeds some threshold.

Acknowledgments: The second author would like to thank Graham Cormode for helpful conversations.

References

[1] N. Alon, A. Andoni, T. Kaufman, K. Matulef, R. Rubinfeld, and N. Xie, *Testing k-wise and almost k-wise independence.*, in STOC, 2007, pp. 496–505.

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

[3] R. Ananthakrishna, A. Das, J. Gehrke, F. Korn, S. Muthukrishnan, and D. Srivastava, *Efficient approximation of correlated sums on data streams.*, IEEE Trans. Knowl. Data Eng., 15 (2003), pp. 569–572.

[4] B. Babcock, S. Babu, M. Datar, R. Motwani, and J. Widom, *Models and issues in data stream systems*, ACM Symposium on Principles of Database Systems, (2002), pp. 1–16.

[5] Z. Bar-Yossef, R. Kumar, and D. Sivakumar, *Reductions in streaming algorithms, with an application to counting triangles in graphs*, in ACM-SIAM Symposium on Discrete Algorithms, 2002, pp. 623–632.

[6] T. Batu, L. Fortnow, E. Fischer, R. Kumar, R. Rubinfeld, and P. White, *Testing random variables for independence and identity.*, in FOCS, 2001, pp. 442–451.

[7] L. Bhuvanagiri, S. Ganguly, D. Kesh, and C. Saha, *Simpler algorithm for estimating frequency moments of data streams.*, in ACM-SIAM Symposium on Discrete Algorithms, 2006, pp. 708–713.

[8] A. Chakrabarti, G. Cormode, and A. McGregor, *A near-optimal algorithm for computing the entropy of a stream*, in ACM-SIAM Symposium on Discrete Algorithms, 2007, pp. 328–335.

[9] A. Chakrabarti, S. Khot, and X. Sun, *Near-optimal lower bounds on the multi-party communication complexity of set disjointness.*, in IEEE Conference on Computational Complexity, 2003, pp. 107–117.

[10] K. L. Chang and R. Kannan, *The space complexity of pass-efficient algorithms for clustering.*, in ACM-SIAM Symposium on Discrete Algorithms, 2006, pp. 1157–1166.

[11] M. Charikar, C. Chekuri, T. Feder, and R. Motwani, *Incremental clustering and dynamic information retrieval.*, SIAM J. Comput., 33 (2004), pp. 1417–1440.

[12] M. Charikar, L. O'Callaghan, and R. Panigrahy, *Better streaming algorithms for clustering problems.*, in ACM Symposium on Theory of Computing, 2003, pp. 30–39.

[13] G. Cormode, M. Datar, P. Indyk, and S. Muthukrishnan, *Comparing data streams using hamming norms (How to zero in).*, IEEE Trans. Knowl. Data Eng., 15 (2003), pp. 529–540.

[14] G. Cormode, F. Korn, S. Muthukrishnan, and D. Srivastava, *Space- and time-efficient deterministic algorithms for biased quantiles over data streams.*, in ACM Symposium on Principles of Database Systems, 2006, pp. 263–272.

[15] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, John Wiley & Sons, 1991.

[16] J. Feigenbaum, S. Kannan, A. McGregor, S. Suri, and J. Zhang, *Graph distances in the streaming model: the value of space.*, in ACM-SIAM Symposium on Discrete Algorithms, 2005, pp. 745–754.

[17] ———, *On graph problems in a semi-streaming model*, Theoretical Computer Science, 348 (2005), pp. 207–216.

[18] J. Gehrke, F. Korn, and D. Srivastava, *On computing correlated aggregates over continual data streams.*, in SIGMOD Conference, 2001, pp. 13–24.

[19] M. Greenwald and S. Khanna, *Efficient online computation of quantile summaries.*, in ACM International Conference on Management of Data, 2001, pp. 58–66.

[20] S. Guha, P. Indyk, and A. McGregor, *Sketching information divergences*, in Conference on Learning Theory, 2007, pp. 424–438.

[21] S. Guha, N. Koudas, and K. Shim, *Approximation and streaming algorithms for histogram construction problems*, ACM Trans. Database Syst., 31 (2006), pp. 396–438.

[22] S. Guha and A. McGregor, *Lower bounds for quantile estimation in random-order and multi-pass streaming*, in International Colloquium on Automata, Languages and Programming, 2007, pp. 704–715.

[23] ———, *Space-efficient sampling*, in AISTATS, 2007, pp. 169–176.

[24] P. Indyk, *Stable distributions, pseudorandom generators, embeddings, and data stream computation.*, J. ACM, 53 (2006), pp. 307–323.

[25] P. Indyk and D. P. Woodruff, *Optimal approximations of the frequency moments of data streams.*, in ACM Symposium on Theory of Computing, 2005, pp. 202–208.

[26] T. S. Jayram, R. Kumar, and D. Sivakumar, *Simple lower bound on one-way Gap-Hamming.*, in `www.cse.iitk.ac.in/users/sganguly/slides/ravikumar.pdf`, 2007.

[27] C. D. Manning and H. Schtze, *Foundations of Statistical Natural Language Processing*, MIT Press, 1999.

[28] S. Muthukrishnan, *Data streams: Algorithms and applications*, Now Publishers, (2006).

[29] D. P. Woodruff, *Optimal space lower bounds for all frequency moments.*, in ACM-SIAM Symposium on Discrete Algorithms, 2004, pp. 167–175.