Sketching Information Divergences

Sudipto Guha*[1], Piotr Indyk[2], and Andrew McGregor[3]

[1] University of Pennsylvania
sudipto@cis.upenn.edu
[2] Massachusetts Institute of Technology
indyk@theory.lcs.mit.edu
[3] University of California, San Diego
andrewm@ucsd.edu

Abstract. When comparing discrete probability distributions, natural measures of similarity are not ℓ_p distances but rather are information-divergences such as Kullback-Leibler and Hellinger. This paper considers some of the issues related to constructing small-space *sketches* of distributions, a concept related to dimensionality-reduction, such that these measures can be approximately computed from the sketches. Related problems for ℓ_p distances are reasonably well understood via a series of results including Johnson, Lindenstrauss [27, 18], Alon, Matias, Szegedy [1], Indyk [24], and Brinkman, Charikar [8]. In contrast, almost no analogous results are known to date about constructing sketches for the information-divergences used in statistics and learning theory.

1 Introduction

Which distances can be sketched in sub-linear space? In recent years, streaming algorithms have received significant attention in an attempt to cope with massive datasets [23, 1, 20]. A streaming computation is a sublinear space algorithm that reads the input in sequential order and any item not explicitly remembered is inaccessible. A fundamental problem in the model is the estimation of distances between two objects that are determined by the stream, e.g., the network traffic matrices at two routers. Estimation of distances allows us to construct approximate representations, e.g., histograms, wavelets, Fourier summaries, or equivalently, find models of the input stream, since this problem reduces to finding the "closest" representation in a suitable class. In this paper, the objects of interest are probability distributions defined by a stream of updates as follows.

Definition 1. *Given a data stream $S = \langle a_1, \ldots, a_m \rangle$ with each data item $a_i \in \{p, q\} \times [n]$ we define $S(p) = \langle a^p_1, \ldots, a^p_{m(p)} \rangle$ to be the sub-stream consisting of data items of the form $\langle p, \cdot \rangle$. $S(p)$ defines a distribution $(p_1, \ldots, p_n)$ where $p_i = m(p)_i/m(p)$ and $m(p)_i = |\{j : a^p_j = \langle p, i \rangle\}|$. Similarly define $S(q)$ and q_i.*

* This research was supported by in part by an Alfred P. Sloan Research Fellowship and by NSF Awards CCF-0430376, and CCF-0644119.

One of the cornerstones in the theory of data stream algorithms has been the result of Alon, Matias, and Szegedy [1]. They showed that it is possible to compute a $(1+\epsilon)$-approximation of $\ell_2(p,q)$ using only $\text{poly}(\epsilon^{-1}, \log n)$ space. The algorithm can be viewed as a partial de-randomization of the famous embedding result of Johnson and Lindenstrauss [27, 18]. This result implies that for any two vectors p and q and an $n \times k$ matrix A whose entries are independent $N(0,1)$ random variables, then, with constant probability,

$$(1-\epsilon)\ell_2(p,q) \leq nk^{-1}\ell_2(Ap, Aq) \leq (1+\epsilon)\ell_2(p,q)$$

for some $k = \text{poly}(\epsilon^{-1}, \log n)$. Alon, Matias, and Szegedy demonstrated that an "effective" A can be stored in small space and can be used to maintain a small-space, update-able summary, or *sketch*, of p and q. The ℓ_2 distance between p and q can then be estimated using only the sketches of p and q. While Brinkman and Charikar [8] proved that there was no analogy of the Johnson-Lindenstrauss result for ℓ_1, Indyk [24] demonstrated that $\ell_1(p,q)$ could also be estimated in $\text{poly}(\epsilon^{-1}, \log n)$ space using Cauchy$(0,1)$ random variables rather than $N(0,1)$ random variables. The results extended to all ℓ_p-measures with $0 < p \leq 2$ using stable distributions. Over a sequence of papers [36, 11, 25, 4, 14], ℓ_p and Hamming distances have become well understood. In parallel, several methods of creating summary representations of streams have been proposed for a variety of applications [9, 12, 15]; in terms of distances they can be adapted to compute the Jaccard coefficient (symmetric difference over union) for two sets. One of the principal motivations of this work is to characterize the distances that can be sketched.

The Information Divergences. Applications in pattern matching, image analysis, statistical learning, etc., use distances which are not ℓ_p norms. Several distances[4] such as the Kullback-Leibler and Hellinger divergences are central to estimating the distances between distributions, and have had a long history of study in statistics and information theory literature. We will discuss two broad classes of distance measures (1) f-divergences, which are central to statistical tests and (2) Bregman Divergences which are central to finding optimal models using mathematical programming.

Definition 2 (f-Divergences). *Given two distributions $p = (p_1, \ldots, p_n)$ and $q = (q_1, \ldots, q_n)$ these distances are given by, $\mathcal{D}_f(p,q) = \sum_i p_i f(q_i/p_i)$, for any function f that is convex over $(0,\infty)$ and satisfies $f(1) = 0$. We define $0f(0/0) = 0$, and $0f(a/0) = \lim_{t\to 0} tf(a/t) = a\lim_{u\to\infty} f(u)/u$.*

The quantity q_i/p_i is the "likelihood ratio" and a fundamental aspect of these measures is that these divergences are tied to "ratio tests" in Neyman-Pearson style hypothesis testing [16]. Several of these divergences appear as exponents of error probabilities for optimal classifiers, e.g., in Stein's Lemma.

[4] Several of the "distances" used are not metric, and a more appropriate reference is divergence; we will refer to them as divergences for the rest of the paper.

Results of Csiszár [17], Liese and Vajda [31], and Amari [2, 3] show that f-divergences are the unique class of distances on distributions that arise from a fairly simple set of axioms, e.g., permutation invariance, non-decreasing local projections, and certain direct sum theorems. In many ways these divergences are "natural" to distributions and statistics, in much the same way that ℓ_2 is a natural measure for points in $\mathbb{R}^n$. Given streams $S(p)$ and $S(q)$, it is natural to ask whether these streams are alike or given a prior model of the data, how well does either conform to the prior? These are scenarios where estimation of f-divergences is the most natural problem at hand. Notably, ℓ_1 distance is an f-divergence, $f(u) = |u-1|$, referred to as the Variational distance. However, ℓ_1 distances do not capture the "marginal" utilities of evidence and in innumerable cases Kullback–Leibler ($f(u) = -\log(u)$), Hellinger ($f(u) = (\sqrt{u}-1)^2$), and Jensen–Shannon divergences ($f(u) = -(u+1)\log\frac{1+u}{2} + u\log u$) are preferred. An important "smooth" subclass of the f-divergences are the α-divergences where $f(u) = 1 - u^{(1+\alpha)/2}$.

A major reason for investigating these f-divergences lies in loss functions used in statistical learning. The ℓ_1 distance captures the "hinge loss" and the other divergences are geared towards non-linear losses. To understand the connection better, we need to also discuss the connections between f-divergences and Bregman divergences. The general family of "arcing" [7] and "AnyBoost" [32] family of algorithms fall into a constrained convex programming framework introduced earlier by Bregman [6]. Friedman, Hastie and Tibshirani [26] established the connection between boosting algorithms and logistic loss, and subsequently over a series of papers [30, 29, 28, 13], the study of Bregman divergences and information geometry has become the method of choice for studying exponential loss functions. The connection between loss functions and f-divergences are investigated more recently by Nguyen, Wainright, and Jordan [34].

Definition 3 (Decomposable Bregman Divergences). *Given two distributions $p = (p_1, \ldots, p_n)$ and $q = (q_1, \ldots, q_n)$, the Bregman divergence between p and q is $\mathcal{B}_F(p, q) = \sum_i [F(p_i) - F(q_i) - (p_i - q_i)F'(q_i)]$ for any strictly convex function F.*

Perhaps the most familiar Bregman divergence is ℓ_2^2 with $F(z) = z^2$. The Kullback–Leibler divergence is also a Bregman divergence with $F(z) = z\log z$, and the Itakura–Saito divergence $F(z) = -\log z$. Lafferty et al. [30] suggest $F(z) = -z^\alpha + \alpha z - \alpha + 1$ for $\alpha \in (0, 1)$, $F(z) = z^\alpha - \alpha z + \alpha - 1$ for $\alpha < 0$.

The fundamental use of Bregman divergences is in finding optimal models. Given a distribution q we are interested in finding a p that best matches the data, and this is posed as a convex optimization problem $\min_p \mathcal{B}_F(p, q)$. It is easy to verify that any positive linear combination of Bregman divergences is a Bregman divergence and that the Bregman balls are convex in the first argument but often not in the second. This is the particular appeal of the technique, that the divergence depends on the data naturally and the divergences have come to be known as Information Geometry techniques. Furthermore there is a natural convex duality between the optimum representation p^* under $\mathcal{B}_F$,

and the divergence $\mathcal{B}_F$. This connection to convex optimization is one of the many reasons for the emerging heavy use of Bregman divergences in the learning literature.

Given that we can estimate ℓ_1 and ℓ_2 distances between two streams in small space, it is natural to ask which other f-divergences and Bregman-divergences are sketchable?

Our Contributions: In this paper we take several steps towards a characterization of the distances that can be sketched. Our first results are negative and help us understand why the ℓ_1 and ℓ_2 distances are special among the f and Bregman divergences.

- We prove the *Shift Invariant Theorem* that characterizes a large family of distances that are not estimable in the streaming model. This theorem pertains to decomposable distances, i.e., distances $d : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}^+$ for which there exists a $\phi : \mathbb{R} \times \mathbb{R} \to \mathbb{R}^+$ such that $d(x, y) = \sum_{i \in [n]} \phi(x_i, y_i)$. The theorem suggest that unless $\phi(x_i, y_i)$ is a function of $x_i - y_i$ then the measure d cannot be sketched.
- For all f-divergence where f is twice differentiable and f'' is strictly positive, no polynomial factor approximation of $\mathcal{D}_f(p, q)$ is possible in sub-linear space. Note that for ℓ_1, which can be sketched, the function $f(\zeta) = |\zeta - 1|$ and therefore f'' is not defined at 1.
- For all Bregman divergences $\mathcal{B}_F$ where F is twice differentiable and there exists $\rho, z_0 > 0$ such that,

  $$\forall\, 0 \leq z_2 \leq z_1 \leq z_0, \frac{F''(z_1)}{F''(z_2)} \geq \left(\frac{z_1}{z_2}\right)^{\rho} \text{ or } \forall\, 0 \leq z_2 \leq z_1 \leq z_0, \frac{F''(z_1)}{F''(z_2)} \leq \left(\frac{z_2}{z_1}\right)^{\rho}$$

  then no polynomial factor approximation of $\mathcal{B}_F$ is possible in sub-linear space. This condition effectively states that $F''(z)$ vanishes or diverges monotonically, and polynomially fast, as z approaches zero. Note that for ℓ_2^2, which can be sketched, the function $F(z) = z^2$ and F'' is constant everywhere.

Given the lower bounds, we ask the question of finding additive bounds in sublinear space. We say an algorithm returns an (ϵ, δ)-additive-approximation for a real number Q if it outputs a value $\hat{Q}$ such that $|\hat{Q} - Q| \leq \epsilon$ with probability at least $(1 - \delta)$ over its internal coin tosses. Some results for two pass algorithms were presented in [22]. In this paper we show sharp characterizations about what can be achieved in a single pass.

- Any $(\epsilon, 1/4)$-additive-approximation of an unbounded $\mathcal{D}_f$ requires $\Omega(n)$-space for any constant ϵ. Alternatively if $\mathcal{D}_f$ is bounded then we can (ϵ, δ)-additive-approximate $\mathcal{D}_f$ in $O_\epsilon(\sqrt{n} \log n \log \delta^{-1})$ space[5]. The space bound can be improved for the Jensen-Shannon divergence. Also, for all bounded symmetric f-divergences, we can approximate $\mathcal{D}_f(p, q)$ up to an additive ϵ in $O(\epsilon^{-2} \log \delta^{-1} \log n)$ space if one of p or q is known in advance.

[5] The notation $O_\epsilon(\cdot)$ treats ϵ as if constant

– If $F(0)$ or $F'(0)$ is unbounded, then any $(\epsilon, 1/4)$-additive-approximation of $\mathcal{B}_F$ requires $\Omega(n)$ space for any constant ϵ. Alternatively, if $F(0), F'(0)$ and $F'(1)$ exist, we can approximate $\mathcal{B}_F(p,q)$ in $O_\epsilon(\log n \log \delta^{-1})$ space.

2 Geometry of $\mathcal{D}_f$ and $\mathcal{B}_F$

In this section we first present some simple geometric results that will allow us to make certain useful assumptions about an f or F defining an f–divergence or Bregman divergence.

We start by defining a *conjugate* $f^*(u) = uf(\frac{1}{u})$. We can then write any f-Divergence as,

$$D_f(p,q) = \sum_{i:p_i>q_i} p_i f(q_i/p_i) + \sum_{i:q_i>p_i} q_i f^*(p_i/q_i) \ .$$

The following lemma that demonstrates that we may assume that $f(u) \in [0, f(0)]$ and $f^*(u) \in [0, f^*(0)]$ for $u \in [0,1]$ where both $f(0) = \lim_{u\to 0} f(u)$ and $f^*(0) = \lim_{u\to 0} f^*(u)$ exist if f is bounded.

Lemma 1. *Let f be a real-valued function that is convex on $(0,\infty)$ and satisfies $f(1) = 0$. Then there exists a real-valued function g that is convex on $(0,\infty)$ and satisfies $g(1) = 0$ such that*

1. *$\mathcal{D}_f(p,q) = \mathcal{D}_g(p,q)$ for all distributions p and q.*
2. *g is decreasing in the range $(0,1]$ and increasing in the range $[1,\infty)$. In particular, if f is differentiable at 1 then $g'(1) = 0$.*

Furthermore, if $\mathcal{D}_f$ is bounded then

3. *$g(0) = \lim_{u\to 0} g(u)$ and $g^*(0) = \lim_{u\to 0} g^*(u)$ exists.*

For example, the Hellinger divergence can be realized by either $f(u) = (\sqrt{u}-1)^2$ or $f(u) = 2 - 2\sqrt{u}$. The next lemma will be important when bounding the error terms in our algorithms.

Lemma 2. *For any function f that is positive and convex on $(0,1]$ with $f(1) = 0$, for all $0 < a < b < c \le 1$, $|f(c) - f(b)| \le \frac{c-b}{c-a} f(a)$.*

Similar to Lemma 1, the following lemma demonstrates that, without loss of generality, we may make various assumptions about the F that defines a Bregman divergence.

Lemma 3. *Let F be a differentiable, real valued function that is strictly convex on $(0,1]$ such that $\lim_{u\to 0^+} F(u)$ and $\lim_{u\to 0^+} F'(u)$ exist. Then there exists a differentiable, real valued function G that is strictly convex on $(0,1]$ and,*

1. *$\mathcal{B}_F(p,q) = \mathcal{B}_G(p,q)$ for all distributions p and q.*
2. *$G(z) \ge 0$ for $x \in (0,1]$ and G is increasing in the range $(0,1]$.*
3. *$\lim_{u\to 0^+} G'(u) = 0$ and $\lim_{u\to 0^+} G(u) = 0$.*

3 Techniques

In this section we summarize some of the sketching and sampling techniques that we will use in the algorithms in the subsequent sections. We then review the general approach for proving lower bounds in the data stream model.

AMS-Sketches: A size-k AMS-Sketch of the stream $S(p) = \langle a_1^p, \ldots, a_{m(p)}^p \rangle$ consists of k independent, identically distributed random variables $X_1, \ldots, X_k$. Each X_i is determined by $X_i = |\{j : a_j^p = a_J^p, J \leq j \leq m(p)\}|$ where J is chosen uniformly at random from $[m(p)]$. This sketch is useful for estimating quantities of the form $m(p)^{-1} \sum_{i \in [n]} f(m(p)_i)$ because, if $f(0) = 0$ then

$$E\left[f(X_i) - f(X_i - 1)\right] = m(p)^{-1} \sum_{i \in [n]} f(m(p)_i) \ .$$

It can be constructed by a streaming computation using only $O(k)$ counters [1].

MG-Sketches: A size-k MG-Sketch of the stream $S(p)$ is a deterministic construction that consists of estimates $(\tilde{p}_i)_{i \in [n]}$ for the probability distribution $(p_i)_{i \in [n]}$. These estimates satisfy $p_i - 1/k \leq \tilde{p}_i \leq p_i$ for all $i \in [n]$. Also, at most k values of $\tilde{p}_i$ are non-zero and hence a size k MG-Sketch can be stored with $O(k)$ counters. Furthermore, the sketch can be constructed by a streaming computation using only $O(k)$ counters [33, 5, 19].

Universe-Sampling: A size-k Universe-Sample of $S(p)$ consists of the exact values of p_i for k randomly chosen $i \in [n]$. It can be trivially constructed by a streaming computation using only $O(k)$ counters.

Lower Bounds: A component of the lower bounds we prove in this paper is a reduction from the communication complexity of Set-Disjointness. An instance of this problem consists of two binary strings, $x, y \in \mathbb{F}_2^n$ such that $\sum_i x_i = \sum_i y_i = n/4$. Alice knows the string x and Bob knows the string y. Alice and Bob take turns to send messages to each other with the goal of determining if x and y are disjoint, i.e. $x.y = 0$. Determining if $x.y = 0$ with probability at least $3/4$ requires $\Omega(n)$ bits to be communicated [35].

Our lower bound proofs use the following template. 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. We then show how Alice and Bob can construct a set of stream elements $S_A(x)$ and $S_B(y)$ such that the value returned by $\mathcal{A}$ on the stream containing $S_A(x) \cup S_B(y)$ determines whether $x.y = 0$. Alice and Bob can then emulate $\mathcal{A}$: 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. This protocol transmits $(2P-1)W$ bits and hence if $P = O(1)$, we deduce that $W = \Omega(n)$.

It should be noted that such a style of proof has been used widely. The novelty of our lower bound proofs is in using the geometry of $\mathcal{D}_f$ and $\mathcal{B}_F$ to construct suitable $S_A(x)$ and $S_B(y)$.

4 Multiplicative Approximations

We start with the central theorem of this section, the *Shift Invariance Theorem*. This theorem characterizes a large class of divergences that are not sketchable.

Theorem 1 (Shift Invariance Theorem). *Let $\phi : [0,1] \times [0,1] \to \mathbb{R}^+$ satisfy $\phi(x,x) = 0$ for all $x \in [0,1]$ and there exists $n_0, a, b, c \in \mathbb{N}$ such that for all $n \geq n_0$,*

$$\max\left(\phi\left(\frac{a}{m}, \frac{a+c}{m}\right), \phi\left(\frac{a+c}{m}, \frac{a}{m}\right)\right) > \frac{\alpha^2 n}{4}\left(\phi\left(\frac{b+c}{m}, \frac{b}{m}\right) + \phi\left(\frac{b}{m}, \frac{b+c}{m}\right)\right)$$

where $m = an/4 + bn + cn/2$. Then any algorithm that returns an α approximation of $d(p,q) = \sum_{i \in [5n/4]} \phi(p_i, q_i)$ with probability at least $3/4$ where p and q are defined by a stream of length $O((a+b+c)n)$ over $[5n/4]$ requires $\Omega(n)$ space.

Proof. We refer the reader to the lower bounds template discussed in Section 3. Assume that n is divisible by 4 and $n > n_0$. Let $(x,y) \in \mathbb{F}_2^n \times \mathbb{F}_2^n$ be an instance of Set-Disjointness where $\sum_i x_i = \sum_i y_i = n/4$. Alice and Bob determine the prefix of a stream $S_A(x)$ and the suffix $S_B(y)$ respectively. We first assume that $\phi(a/m, (a+c)/m) \geq \phi((a+c)/m, a/m)$.

$$\begin{aligned} S_A(x) = &\bigcup_{i \in [n]} \{ax_i + b(1-x_i) \text{ copies of } \langle p, i\rangle \text{ and } \langle q, i\rangle\} \\ &\cup \bigcup_{i \in [n/4]} \{b \text{ copies of } \langle p, i+n\rangle \text{ and } \langle q, i+n\rangle\} \\ S_B(y) = &\bigcup_{i \in [n]} \{cy_i \text{ copies of } \langle q, i\rangle\} \cup \bigcup_{i \in [n/4]} \{c \text{ copies of } \langle p, i+n\rangle\} \end{aligned}$$

Observe that $m(p) = m(q) = an/4 + bn + cn/2$ and

$$\mathcal{D}_f(p,q) = (x.y)\phi\left(\frac{a}{m}, \frac{a+c}{m}\right) + (n/4 - x.y)\phi\left(\frac{b}{m}, \frac{b+c}{m}\right) + (n/4)\phi\left(\frac{b+c}{m}, \frac{b}{m}\right) \quad .$$

Therefore,

$$\begin{aligned} x.y = 0 &\Leftrightarrow \mathcal{D}_f(p,q) = (n/4)(\phi(b/m, (b+c)/m) + \phi((b+c)/m, b/m)) \\ x.y = 1 &\Leftrightarrow \mathcal{D}_f(p,q) \geq \alpha^2(n/4)(\phi(b/m, (b+c)/m) + \phi((b+c)/m, b/m)) \end{aligned}$$

Therefore any α-approximation would determine the value of $x.y$ and hence an α-approximation requires $\Omega(n)$ space [35]. If $\phi(a/m, (a+c)/m) \leq \phi((a+c)/m, a/m)$ then the proof follows by reversing the roles of p and q.

The above theorem suggests that unless $\phi(x_i, y_i)$ is some function of $x_i - y_i$ then the distance is not sketchable. The result holds even if the algorithm may take a constant number of passes over the data. We also mention a simpler result

that can be proved using similar ideas to those employed above. This states that if there exist $a, b, c \in \mathbb{N}$ such that

$$\max\left(\frac{\phi(a+c,a)}{\phi(b+c,b)}, \frac{\phi(a,a+c)}{\phi(b,b+c)}\right) > \alpha^2 \ ,$$

then any single-pass α-approximation of $\sum_{i\in[n]} \phi(m(p)_i, m(q)_i)$ requires $\Omega(n)$ space.

We next present two corollaries of Theorem 1. These characterize the f-divergences and Bregman divergences that can be not be sketched. Note that ℓ_1 and ℓ_2^2, which can be sketched, are the only commonly used divergences that do not satisfy the relevant conditions.

Corollary 1 (f-Divergences). *Given an f-divergence $\mathcal{D}_f$, if f is twice differentiable and f'' is strictly positive, then no polynomial factor approximation of $\mathcal{D}_f$ is possible in sub-linear space.*

Proof. We first note that by Lemma 1 we may assume $f(1) = f'(1) = 0$. Let $a = c = 1$ and $b = \alpha^2 n(f''(1)+1)/(8f(2))$ where α is an arbitrary polynomial in n. Note that $f(2) > 0$ because f is strictly convex.

We start by observing that,

$$\phi(b/m, (b+c)/m) = (b/m)f(1+1/b) = (b/m)\left[f(1) + \frac{1}{b}f'(1) + \frac{1}{2!b^2}f''(1+\gamma)\right]$$

for some $\gamma \in [0, 1/b]$ by Taylor's Theorem. Since $f(1) = f'(1) = 0$ and $f''(t)$ is continuous at $t = 1$ this implies that for sufficiently large n, $f''(1+\gamma) \leq f''(1)+1$ and so,

$$\phi(b/m, (b+c)/m) \leq \frac{f''(1)+1}{2mb} = \frac{f''(1)+1}{2f(2)b}m^{-1}f(2) \leq \frac{8}{\alpha^2 n}\phi(a/m, (a+c)/m) \ .$$

Similarly we can show that for sufficiently large n,

$$\phi((b+c)/m, b/m) \leq \frac{8}{\alpha^2 n}\phi(a/m, (a+c)/m) \ .$$

Then appealing to Theorem 1 we get the required result.

Corollary 2 (Bregman Divergences). *Given a Bregman divergences $\mathcal{B}_F$, if F is twice differentiable and there exists $\rho, z_0 > 0$ such that,*

$$\forall\ 0 \leq z_2 \leq z_1 \leq z_0,\ \frac{F''(z_1)}{F''(z_2)} \geq \left(\frac{z_1}{z_2}\right)^{\rho} \ \text{or}\ \forall\ 0 \leq z_2 \leq z_1 \leq z_0,\ \frac{F''(z_1)}{F''(z_2)} \leq \left(\frac{z_2}{z_1}\right)^{\rho}$$

then no polynomial factor approximation of $\mathcal{B}_F$ is possible in sub-linear space.

This condition effectively states that $F''(z)$ vanishes or diverges monotonically, and polynomially fast, as $z \to 0$.

Proof. By the Mean-Value Theorem, for any $m, r \in \mathbb{N}$, there exists $\gamma(r) \in [0,1]$ such that, $\phi(r/m,(r+1)/m) + \phi((r+1)/m, r/m) = m^{-2}F''((r+\gamma(r))/m)$. Therefore, for any $a, b \in \mathbb{N}, c = 1$ and $m = an/4 + bn + n/2$,

$$\frac{\max\left(\phi\left(\frac{a}{m},\frac{a+c}{m}\right),\phi\left(\frac{a+c}{m},\frac{a}{m}\right)\right)}{\phi\left(\frac{b+c}{m},\frac{b}{m}\right)+\phi\left(\frac{b}{m},\frac{b+c}{m}\right)} \geq \frac{1}{2}\frac{F''((a+\gamma(a))/m)}{F''((b+\gamma(b))/m)} \ .$$

If $\forall\ 0 \leq z_2 \leq z_1 \leq z_0,\ F''(z_1)/F''(z_2) \geq (z_1/z_2)^\rho$ then set $a = (\alpha^2 n)^{1/\rho}$ and $b = 1$ where α is an arbitrary polynomial in n. If $\forall\ 0 \leq z_2 \leq z_1 \leq z_0,\ F''(z_1)/F''(z_2) \geq (z_2/z_1)^\rho$ then set $a = 1$ and $b = (\alpha n)^{1/\rho}$. In both cases we deduce that the RHS of Eqn. 1 is greater than $\alpha^2 n/4$. Hence, appealing to Theorem 1, we get the required result.

5 Additive Approximations

In this section we focus on additive approximations. As mentioned earlier, the probability of misclassification using ratio tests is often bounded by $2^{-\mathcal{D}_f}$, for certain $\mathcal{D}_f$. Hence, an additive ϵ approximation translates to a multiplicative 2^ϵ factor for computing the error probability. Our goal is the characterization of divergences that can be approximated additively.

5.1 Lower bound for f-divergences

In this section we show that to additively approximate $\mathcal{D}_f(p,q)$ up to any additive $\epsilon > 0$, $\mathcal{D}_f$ must be bounded.

Theorem 2. *Any $(\epsilon, 1/4)$-additive-approximation of an unbounded $\mathcal{D}_f$ requires $\Omega(n)$ space. This applies even if one of the distributions is known to be uniform.*

Proof. We refer the reader to the template for lower bounds discussed in Section 3. Let $(x,y) \in \mathbb{F}_2^n \times \mathbb{F}_2^n$ be an instance of Set-Disjointness. Then define q be the following stream elements.

$$S_A(x) = \{1 - x_i \text{ copies of } \langle q, i\rangle \text{ for } i \in [n]\}$$
$$S_B(y) = \{1 - y_i \text{ copies of } \langle q, i\rangle \text{ for } i \in [n]\}$$

Let p be the uniform distribution. If $\lim_{u\to 0} f(u)$ is unbounded then $\mathcal{D}_f(p,q)$ is finite iff $x.y = 0$. If $\lim_{u\to\infty} \frac{1}{u} f(u)$ is unbounded then $\mathcal{D}_f(q,p)$ is finite iff $x.y = 0$.

5.2 Upper bounds for f-divergences

In this section we show an additive approximation that complements the lower bound in the previous section. Note that since for any f–divergence, a function $af(\cdot)$ for $a > 0$ gives another f–divergence, the best we can hope for is an approximation which is dependent on $\max\{\lim_{u\to 0} f(u), \lim_{u\to\infty} \frac{1}{u} f(u)\}$. In

The algorithm f-Est(p,q): Let ϵ be a user-specified value in the range $(0,1)$. Let $\gamma(\epsilon) < \epsilon/16$ be such that

$$\forall u \leq \gamma, |f(u) - \lim_{u\to 0} f(u)| \leq \epsilon/16 \text{ and } \forall u \geq \frac{1}{\gamma}, \left|\frac{1}{u}f(u) - \lim_{u\to\infty}\frac{1}{u}f(u)\right| \leq \epsilon/16 .$$

1. Use Universe-Sampling to compute p_i, q_i for $i \in S$ where S is a random subset of $[n]$ of size $3\epsilon^{-2}n(\rho+\gamma^2\rho)\ln(2\delta^{-1})$ where $\rho = 1/\sqrt{n}$.
2. Use MG-Sketches to compute $(\tilde{p}_i)_{i\in[n]}$ and $(\tilde{q}_i)_{i\in[n]}$ such that
$$p_i - \gamma^2\rho \leq \tilde{p}_i \leq p_i \quad \text{and} \quad q_i - \gamma^2\rho \leq \tilde{q}_i \leq q_i .$$
3. Return,
$$\sum_{i\in T}\tilde{p}_i f(\tilde{q}_i/\tilde{p}_i) + \frac{n}{|S|}\sum_{i\in S\setminus T} p_i f(q_i/p_i)$$
where $T = \{i : \max\{\tilde{p}_i, \tilde{q}_i\} \geq \rho\}$.

Fig. 1. Additive Approximation of Some f-Divergences

what follows we assume that this value is 1. The idea behind the algorithm is a combination of Universe-Sampling and MG-Sketching. With MG-Sketches we can identify all $i \in [n]$ such that either p_i or q_i is larger than some threshold. For the remaining i it is possible to show that $p_i f(q_i/p_i)$ is small enough such that estimating the contribution of these terms by Universe-Sampling yields the required result. See Figure 1 for a detailed description of the algorithm.

Lemma 4. *$\max\{p_i, q_i\} \leq \rho + \gamma^2\rho$ for $i \notin T$ and $\max\{p_i, q_i\} \geq \rho$ for $i \in T$. Furthermore,*

$$\left|\sum_{i\in T} p_i f(q_i/p_i) - \sum_{i\in T}\tilde{p}_i f(\tilde{q}_i/\tilde{p}_i)\right| \geq \epsilon/2.$$

Proof. The first part of the lemma follows from the properties of the MG-Sketch discussed in Section 3. Let $\Delta(p_i)$, $\Delta(q_i/p_i)$, $\Delta(f(q_i/p_i))$, and $\Delta(q_i)$ be the absolute errors in $p_i, q_i/p_i, f(q_i/p_i), q_i$ respectively and note that,

$$|\tilde{p}_i f(\tilde{q}_i/\tilde{p}_i) - p_i f(q_i/p_i)| \leq f(q_i/p_i)\Delta(p_i) + p_i\Delta(f(q_i/p_i)) + \Delta(p_i)\Delta(f(q_i/p_i)) .$$

There are four cases to consider.

1. $p_i \geq \rho$ and $q_i \leq \gamma\rho/2$. Then $\Delta(f(q_i/p_i)) \leq \epsilon/8$ since $q_i/p_i, \tilde{q}_i/\tilde{p}_i \leq \gamma$ and f is non-increasing in the range $(0,1)$. Therefore,
$$|\tilde{p}_i f(\tilde{q}_i/\tilde{p}_i) - p_i f(q_i/p_i)| \leq \gamma^2\rho f(q_i/p_i) + \epsilon p_i/16 + \gamma^2\rho\epsilon/16 + \leq \epsilon p_i/8 .$$
2. $q_i \geq \rho$ and $p_i \leq \gamma\rho/2$. Similar to case 1.
3. $p_i \geq \rho$, $q_i \geq \gamma\rho/2$ and $p_i > q_i$. First note that $\Delta(q_i/p_i) \leq (2\gamma+\gamma^2)q_i/p_i$.
$$\begin{aligned}|\tilde{p}_i f(\tilde{q}_i/\tilde{p}_i) - p_i f(q_i/p_i)| &\leq \gamma^2 p_i + p_i(2\gamma+\gamma^2)q_i/p_i + \gamma^2 p_i(2\gamma+\gamma^2)q_i/p_i \\ &\leq \epsilon p_i/16 + \epsilon q_i/8 .\end{aligned}$$

4. $q_i \geq \rho$, $p \geq \gamma\rho/2$, and $p_i < q_i$. Similar to case 3.

Therefore summarizing all cases, the additive error per-term is at most $\epsilon(q_i + p_i)/4$. Summing over all $i \in T$ establishes the second part of the lemma.

Theorem 3. *There exists an (ϵ, δ)-additive-approximation for any bounded f-divergence using $O_\epsilon(\sqrt{n} \log n \log \delta^{-1})$ space.*

Proof. The space use is immediate from the algorithm. The main observation to prove correctness is that for each $i \notin T$, $p_i, q_i \leq \rho + \gamma^2\rho$, and hence $p_i f(q_i/p_i) \leq \rho + 2\gamma^2\rho$. Hence, by an application of the Chernoff bound,

$$\Pr\left[\left|\frac{n}{|S|}\sum_{i\in S\setminus T} p_i f\left(\frac{q_i}{p_i}\right) - \sum_{i\in[n]\setminus T} p_i f\left(\frac{q_i}{p_i}\right)\right| \geq \epsilon/2\right] \leq 2\exp\left(-\frac{|S|\epsilon^2}{3n(\rho + 2\gamma^2\rho)}\right).$$

This is at most δ for our value of $|S|$. Appealing to Lemma 4 yields the result.

Some f-divergences can be additively approximated in significantly smaller space. For example, the Jensen–Shannon divergence can be rewritten as

$$JS(p, q) = \ln 2\left(2H\left(\frac{p+q}{2}\right) - H(p) - H(q)\right) \quad ,$$

where H is the entropy. There exists a single-pass (ϵ, δ)-additive-approximation of entropy in the streaming model [10]. This yields the following theorem.

Theorem 4. *There exists a single-pass (ϵ, δ)-additive-approximation of the JS-divergence using $O(\epsilon^{-2} \log^2 n \log^2 m \log \delta^{-1})$ space.*

Finally in this section we show that the space bound of $O_\epsilon(\sqrt{n} \log n \log \delta^{-1})$ can be improved if we knew one of the distributions, e.g., if we had a prior distribution and were trying to estimate a fit.

Theorem 5. *We can (ϵ, δ)-additively-approximate any f–divergence $\mathcal{D}_f(p, q)$ in space $O(\epsilon^{-2} \log n \log \delta^{-1})$ if $\mathcal{D}_f$ is bounded and one of p or q is known in advance.*

Proof. Let p be the known distribution and let q be defined by the stream. We may assume that $f(1) = f'(1) = 0$. Therefore,

$$\mathcal{D}_f(p, q) = \sum_{q_i<p_i} p_i f\left(\frac{q_i}{p_i}\right) + \sum_{q_i>p_i} q_i f^*\left(\frac{p_i}{q_i}\right) \quad .$$

We consider each term separately. To approximate the first term the algorithm picks i with respect to the known distribution p and then computes q_i. The basic estimator is

$$g(i) = \begin{cases} 0 & \text{if } q_i > p_i \\ f(q_i/p_i) & \text{if } q_i \leq p_i \end{cases} \quad .$$

Note that $E[g(i)] = \sum_{i:q_i<p_i} p_i f(q_i/p_i)$ and $0 \le g(i) \le f(0)$. Hence, applying Chernoff bounds we can $(\epsilon/2, \delta/2)$-additively-approximate $E[g(i)]$ with $O(\epsilon^{-2}\log\delta^{-1})$ basic estimators.

To approximate the second term we use an AMS-Sketch. Specifically, the algorithm picks a random i in the stream q and computes, r_i, the number of times i occurs after it was picked. The basic estimator is,

$$h(r_i) = \begin{cases} 0 & \text{if } p_i \ge r_i/m \\ f^*(0) & \text{if } p_i = 0 \\ r_i f^*(\frac{mp_i}{r_i}) - (r_i-1)f^*(\frac{mp_i}{r_i-1}) & \text{otherwise} \end{cases} .$$

Note that $E[h(r_i)] = \sum_{i:q_i>p_i} q_i f(\frac{p_i}{q_i})$ and $0 \le h(i) \le f^*(0)$ by Lemma 2. Hence, applying Chernoff bounds we can $(\epsilon/2, \delta/2)$-additively-approximate $E[g(i)]$ with $O(\epsilon^{-2}\log\delta^{-1})$ basic estimators.

5.3 Lower bound for Bregman Divergences

Theorem 6. *If $\max\{\lim_{u\to 0} F(u), \lim_{u\to 0} F'(u)\}$ is unbounded then $(\epsilon, 1/4)$-additive-approximation of $\mathcal{B}_F$ requires $\Omega(n)$ space. This applies even if one of the distributions is known to be uniform.*

Proof. We refer the reader to the lower bounds template discussed in Section 3. Let $(x, y) \in \mathbb{F}_2^n \times \mathbb{F}_2^n$ be an instance of Set-Disjointness. Let q be determined by Alice and Bob as in Theorem 2. and let p be the uniform distribution. If $\lim_{u\to 0} F(u)$ is unbounded then $\mathcal{B}_F(q, p)$ is finite iff $x.y = 0$. If $\lim_{u\to 0} F(u)$ is bounded but $\lim_{u\to 0} F'(u)$ is unbounded then $\mathcal{B}_F(p, q)$ is finite iff $x.y = 0$.

5.4 Upper bound for Bregman Divergences

In this section we show the matching upper bounds to Theorem 6. In the section we assume that $F(0), F'(0)$ and $F'(1)$ are defined. Recall from Lemma 3 that we may assume that $F(0) = F'(0) = 0$. This makes F monotone increasing over $[0, 1]$. Note that this transformation preserves $F'(1)$ to be a constant. As with the f–divergences, any multiple of an Bregman divergence is another Bregman divergence and hence the best we can hope for is an approximation which is dependent on $F'(1)$. In what follows we assume that this value is 1.

Theorem 7. *Assuming $F(0), F'(0), F'(1)$ exist we can approximate $\mathcal{B}_F(p, q)$ for any two unknown streams p, q upto additive ϵ in $O_\epsilon(\log n \log\delta^{-1})$ space.*

Proof. Write $\mathcal{B}_F$ as $\mathcal{B}_F(p, q) = \sum_i F(p_i) - \sum_i F(q_i) - \sum_i p_i F'(q_i) + \sum_i q_i F'(q_i)$. We show how to estimate each term with probability at least $1 - \delta/4$ up to an additive $\epsilon/4$ term. Because $0 \le m(p)[F(X_j/m(p)) - F((X_j - 1)/m(p))] \le F'(1) = 1$,

$$\Pr\left[\left|\widetilde{\sum_i F(p_i)} - \sum_i F(p_i)\right| > \epsilon/4\right] \le 2\exp(-|S|\epsilon^2/48) \le \delta/4 .$$

Algorithm $\mathcal{B}$-Est(p,q): Let $\epsilon_2 = \epsilon/12$. Let $\gamma(\epsilon_2) \leq \epsilon_2$ be such that, $\forall u \in (0,1]$, $|F'(u+\gamma) - F'(u)| \leq \epsilon_2$ and let $\epsilon_1 = \gamma\epsilon_2$.

1. Use AMS-Sketches to estimate $\sum_i F(p_i)$: Choose a random subset of $S \subset [m(p)]$ of size $48\epsilon^{-2}\ln(4\delta^{-1})$. For each $j \in S$,
   (a) Let $e(j) = i$ where $a_j^p = \langle p, i\rangle$.
   (b) Let $X_j = |\{k : a_k = a_j, k \geq j\}|$
   Let
   $$\widetilde{\sum_i F(p_i)} = \frac{m(p)}{|S|}\sum_j [F\left(\frac{X_j}{m(p)}\right) - F\left(\frac{X_j - 1}{m(p)}\right)]$$
   and define $\widetilde{\sum_i F(q_i)}$ analogously.
2. Use MG-Sketches to compute $(\tilde{p}_i)_{i\in[n]}$ and $(\tilde{q}_i)_{i\in[n]}$ such that
   $$p_i - \epsilon_1 \leq \tilde{p}_i \leq p_i \quad \text{and} \quad q_i - \epsilon_2 \leq \tilde{q}_i \leq q_i \ .$$
3. Return,
   $$\widetilde{\sum_i F(p_i)} - \widetilde{\sum_i F(q_i)} - \sum_i (\tilde{p}_i - \tilde{q}_i)F'(\tilde{q}_i) \ .$$

Fig. 2. Additive Approximation of Some Bregman Divergences

The calculation for the second term is similar. To bound the remaining terms, since $p_i \geq \tilde{p}_i \geq \max\{p_i - \epsilon_1, 0\}$ and $q_i \geq \tilde{q}_i \geq \max\{q_i - \epsilon_1, 0\}$, we get that $F'(q_i) \geq F'(\tilde{q}_i) \geq \max\{F'(q_i) - \epsilon_2, 0\}$ and $\sum_i p_i F'(q_i) \geq \sum_i \tilde{p}_i F'(\tilde{q}_i)$. Hence,

$$\begin{aligned}
\sum_i \tilde{p}_i F'(\tilde{q}_i) &\geq \sum_i \max\{p_i - \epsilon_1, 0\}\max\{F'(q_i) - \epsilon_2, 0\} \\
&\geq \sum_i p_i F'(q_i) - \sum_{i:\epsilon_1 < p_i, q_i < \gamma} \epsilon_1 F'(q_i) - \sum_{i:\epsilon_1 < p_i, q_i \geq \gamma} \epsilon_1 F'(q_i) - \epsilon_2 \\
&\geq \sum_i p_i F'(q_i) - \epsilon_2 - \frac{\epsilon_1}{\gamma} - \epsilon_2 \\
&\geq \sum_i p_i F'(q_i) - 3\epsilon_2 \geq \sum_i p_i F'(q_i) - \epsilon/4 \ .
\end{aligned}$$

The calculation for the fourth term is entirely similar.

6 Conclusions and Open Questions

We presented a partial characterization of the information divergences that can be multiplicatively approximated in the data stream model. This characterization was based on a general result that suggests that any distance that is sketchable has certain "norm-like" properties.

We then considered additive-approximation of f-divergences and Bregman divergences. In particular, we showed that all bounded f-divergences can be approximated up to an additive ϵ term in a single pass using $O_\epsilon(\sqrt{n}\,\mathrm{polylog}\, n)$

space. In two passes, $O(\text{polylog}\, n)$-space is known to be sufficient [22]. As was noted, there does exists a single-pass, $O_\epsilon(\text{polylog}\, n)$-space additive approximation for the Jensen-Shannon divergence. This begs the question whether there exist single-pass $O_\epsilon(\text{polylog}\, n)$-space algorithms for all bounded f-divergences?

A final open question relates to multiplicative approximation of information divergences in the *aggregate data stream model* in which all elements of the form $\langle p, \cdot \rangle$ appear consecutively. It is easy to $(1+\epsilon)$ multiplicatively approximate the Hellinger divergence in this aggregate model using $O(\epsilon^{-2}\,\text{polylog}\, n)$ space by exploiting the connection between the Hellinger divergence and the L_2 distance. The Jensen-Shannon divergence is constant factor related to Hellinger and therefore there exists a constant factor approximation to Jensen-Shannon in $O(\text{polylog}\, n)$ space. How much space is required to find an $(1+\epsilon)$-approximation?

References

1. N. Alon, Y. Matias, and M. Szegedy. The space complexity of approximating the frequency moments. *Journal of Computer and System Sciences*, 58(1):137–147, 1999.
2. S.-I. Amari. *Differential-geometrical methods in statistics*. Springer-Verlag, New York, 1985.
3. S.-I. Amari and H. Nagaoka. *Methods of Information Geometry*. Oxford University and AMS Translations of Mathematical Monographs, 2000.
4. 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*, pages 708–713, 2006.
5. P. Bose, E. Kranakis, P. Morin, and Y. Tang. Bounds for frequency estimation of packet streams. In *SIROCCO*, pages 33–42, 2003.
6. L. M. Bregman. The relaxation method of finding the common point of convex sets and its application to the solution of problems in convex programming. *U.S.S.R. Computational Mathematics and Mathematical Physics*, 7(1):200–217, 1967.
7. L. Breiman. Prediction games and arcing algorithms. *Neural Computation*, 11(7):1493–1517, 1999.
8. B. Brinkman and M. Charikar. On the impossibility of dimension reduction in l_1. In *IEEE Symposium on Foundations of Computer Science*, pages 514–523, 2003.
9. A. Z. Broder, M. Charikar, A. M. Frieze, and M. Mitzenmacher. Min-wise independent permutations. *J. Comput. Syst. Sci.*, 60(3):630–659, 2000.
10. 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.
11. 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*, pages 107–117, 2003.
12. M. Charikar, K. Chen, and M. Farach-Colton. Finding frequent items in data streams. In *International Colloquium on Automata, Languages and Programming*, pages 693–703, 2002.
13. M. Collins, R. E. Schapire, and Y. Singer. Logistic regression, Adaboost and Bregman distances. *Machine Learning*, 48(1-3):253–285, 2002.

14. 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(3):529–540, 2003.
15. G. Cormode and S. Muthukrishnan. An improved data stream summary: the count-min sketch and its applications. *J. Algorithms*, 55(1):58–75, 2005.
16. T. M. Cover and J. A. Thomas. *Elements of Information Theory*. Wiley Series in Telecommunications. John Wiley & Sons, New York, NY, USA, 1991.
17. I. Csiszár. Why least squares and maximum entropy? an axiomatic approach to inference for linear inverse problems. *Ann. Statist.*, pages 2032–2056, 1991.
18. S. Dasgupta and A. Gupta. An elementary proof of a theorem of johnson and lindenstrauss. *Random Struct. Algorithms*, 22(1):60–65, 2003.
19. E. D. Demaine, A. López-Ortiz, and J. I. Munro. Frequency estimation of internet packet streams with limited space. In *ESA*, pages 348–360, 2002.
20. J. Feigenbaum, S. Kannan, M. Strauss, and M. Viswanathan. An approximate L^1 difference algorithm for massive data streams. *SIAM Journal on Computing*, 32(1):131–151, 2002.
21. S. Guha and A. McGregor. Space-efficient sampling. In *AISTATS*, pages 169–176, 2007.
22. S. Guha, A. McGregor, and S. Venkatasubramanian. Streaming and sublinear approximation of entropy and information distances. In *ACM-SIAM Symposium on Discrete Algorithms*, pages 733–742, 2006.
23. M. R. Henzinger, P. Raghavan, and S. Rajagopalan. Computing on data streams. *External memory algorithms*, pages 107–118, 1999.
24. P. Indyk. Stable distributions, pseudorandom generators, embeddings and data stream computation. *IEEE Symposium on Foundations of Computer Science*, pages 189–197, 2000.
25. P. Indyk and D. P. Woodruff. Optimal approximations of the frequency moments of data streams. In *ACM Symposium on Theory of Computing*, pages 202–208, 2005.
26. R. T. Jerome Friedman, Trevor Hastie. Additive logistic regression: a statistical view of boosting. *Annals of Statistics*, 28:337–407, 2000.
27. W. B. Johnson and J. Lindenstrauss. Extensions of Lipshitz mapping into Hilbert Space. *Contemporary Mathematics, Vol 26*, pages 189–206, May 1984.
28. J. Kivinen and M. K. Warmuth. Boosting as entropy projection. In *COLT*, pages 134–144, 1999.
29. J. D. Lafferty. Additive models, boosting, and inference for generalized divergences. In *COLT*, pages 125–133, 1999.
30. J. D. Lafferty, S. D. Pietra, and V. J. D. Pietra. Statistical learning algorithms based on bregman distances. In *Canadian Workshop on Information Theory*, 1997.
31. F. Liese and F. Vajda. Convex statistical distances. *Teubner-Texte zur Mathematik, Band 95, Leipzig*, 1987.
32. L. Mason, J. Baxter, P. Bartlett, and M. Frean. Functional gradient techniques for combining hypotheses. In *Advances in Large Margin Classifiers*. MIT Press, 1999.
33. J. Misra and D. Gries. Finding repeated elements. *Sci. Comput. Program.*, 2(2):143–152, 1982.
34. X. Nguyen, M. J. Wainwright, and M. I. Jordan. Divergences, surrogate loss functions and experimental design. *Proceedings of NIPS*, 2005.
35. A. A. Razborov. On the distributional complexity of disjointness. *Theor. Comput. Sci.*, 106(2):385–390, 1992.
36. M. E. Saks and X. Sun. Space lower bounds for distance approximation in the data stream model. *ACM Symposium on Theory of Computing*, pages 360–369, 2002.