Mach Learn (2008) 72: 5–19
DOI 10.1007/s10994-008-5054-x

Sketching information divergences

Sudipto Guha · Piotr Indyk · Andrew McGregor

Received: 28 August 2007 / Revised: 31 January 2008 / Accepted: 27 March 2008 /
Published online: 14 May 2008

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 in the data-stream model, a concept related to dimensionality reduction, such that these measures can be approximated from the sketches. Related problems for ℓ_p distances are reasonably well understood via a series of results by Johnson and Lindenstrauss (Contemp. Math. 26:189–206, 1984), Alon et al. (J. Comput. Syst. Sci. 58(1):137–147, 1999), Indyk (IEEE Symposium on Foundations of Computer Science, pp. 202–208, 2000), and Brinkman and Charikar (IEEE Symposium on Foundations of Computer Science, pp. 514–523, 2003). In contrast, almost no analogous results are known to date about constructing sketches for the information divergences used in statistics and learning theory.

Our main result is an impossibility result that shows that no small-space sketches exist for the multiplicative approximation of any commonly used f-divergences and Bregman divergences with the notable exceptions of ℓ_1 and ℓ_2 where small-space sketches exist. We then present data-stream algorithms for the additive approximation of a wide range of information divergences. Throughout, our emphasis is on providing general characterizations.

Editors: Claudio Gentile, Nader H. Bshouty.

Part of this work originally appeared in the Twentieth Annual Conference on Learning Theory, 2007.
This research was supported by in part by an Alfred P. Sloan Research Fellowship and by NSF Awards CCF-0430376, and CCF-0644119.

S. Guha
Dept. of Computer and Information Sciences, University of Pennsylvania, Philadelphia, PA 19104, USA
e-mail: sudipto@cis.upenn.edu

P. Indyk
Computer Science and Artificial Intelligence Lab, Massachusetts Institute of Technology, Cambridge, MA 02139, USA
e-mail: indyk@theory.lcs.mit.edu

A. McGregor (✉)
Information Theory & Applications Center, University of California, San Diego, CA 92109, USA
e-mail: andrewm@ucsd.edu

Keywords Information divergences · Data stream model · Sketches · Communication complexity · Approximation algorithms

1 Introduction

In recent years, the data-stream model has enjoyed significant attention because of the need to process massive data sets (e.g. Henzinger et al. 1999; Alon et al. 1999; Feigenbaum et al. 2002). A streaming computation is a sublinear space algorithm that reads the input in sequential order and any item not explicitly remembered is subsequently 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 empirical probability distributions defined by a stream of updates as follows.

Definition 1 For a data stream $S = \langle a_1, \ldots, a_m \rangle$ where $a_i \in \{p, q\} \times [n]$ we define *empirical distributions* p and q as follows. Let $m(p)_i = |\{j : a_j = \langle p, i \rangle\}|$, $m(p) = |\{j : a_j = \langle p, \cdot \rangle\}|$ and $p_i = m(p)_i / m(p)$. Similarly for q.

One of the cornerstones in the theory of data stream algorithms has been the result of Alon et al. (1999). They showed that it is possible to estimate $\ell_2(p, q) := \|p - q\|_2$ (the Euclidean distance) up to a $(1 + \epsilon)$ factor using only $\text{poly}(\epsilon^{-1}, \log n)$ space. The algorithm can, in retrospect, be viewed in terms of the famous embedding result of Johnson and Lindenstrauss (1984). This result implies that for any two vectors p and q and a $k \times n$ matrix A whose entries are independent Normal$(0, 1)$ random variables (scaled appropriately),

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

with high probability 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, updateable 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 (2003) proved that there was no analog of the Johnson-Lindenstrauss result for ℓ_1, Indyk (2000) demonstrated that $\ell_1(p, q)$ could also be estimated in $\text{poly}(\epsilon^{-1}, \log n)$ space by using Cauchy$(0, 1)$ random variables rather than Normal$(0, 1)$ random variables. The results extended to all ℓ_p measures with $0 < p \leq 2$ using stable distributions. Over a sequence of papers (Saks and Sun 2002; Chakrabarti et al. 2003; Cormode et al. 2003; Bar-Yossef et al. 2004; Indyk and Woodruff 2005; Bhuvanagiri et al. 2006; Cormode and Ganguly 2007) ℓ_p and Hamming distances have become well understood. Concurrently several methods of creating summary representations of streams have been proposed (Broder et al. 2000; Charikar et al. 2002; Cormode and Muthukrishnan 2005) for a variety of applications; 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[1] 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 measures (1) f-divergences, which are used in statistical tests and (2) Bregman divergences which are used in finding optimal models via mathematical programming.

Definition 2 (f-divergences) Let p and q be two n-point distributions. A convex function $f : (0, \infty) \to \mathbb{R}$ such that $f(1) = 0$ gives rise to an f-divergence,

$$\mathcal{D}_f(p, q) = \sum_{i \in [n]} p_i f(q_i/p_i),$$

where we define $0 f(0/0) = 0$, $a f(0/a) = a \lim_{u \to 0} f(u)$, and $0 f(a/0) = a \lim_{u \to \infty} f(u)/u$ if these limits exist.

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 (e.g. Cover and Thomas 1991). Several of these divergences appear as exponents of error probabilities for optimal classifiers, e.g., in Stein's Lemma. Results of Csiszár (1991), Liese and Vajda (1987), and Amari (1985) 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 the sub-streams defining p and 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 where $f(u) = |u - 1|$ and is often referred to as the variational distance in this context. However, ℓ_1 distances do not capture the "marginal" utilities of evidence and in innumerable cases Kullback–Leibler where $f(u) = -\log(u)$, Hellinger where $f(u) = (\sqrt{u} - 1)^2$, Triangle where $f(u) = (1 - u)^2/(1 + u)$, and Jensen–Shannon divergences where $f(u) = -(u + 1)\log(1/2 + u/2) + u \log u$ 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" (Breiman 1999) and "AnyBoost" (Mason et al. 1999) family of algorithms fall into a constrained convex programming framework introduced earlier by Bregman (1967). Friedman et al. (2000) established the connection between boosting algorithms and logistic loss, and subsequently over a series of papers (Lafferty et al. 1997; Lafferty 1999; Kivinen and Warmuth 1999; Collins et al. 2002), 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 et al. (2005).

[1] Several of the "distances" used are not metric, and we henceforth use the more appropriate term "divergence."

Definition 3 (Decomposable Bregman divergences) Let p and q be two n-point distributions. A strictly convex function $F : (0, 1] \to \mathbb{R}$ gives rise to a Bregman divergence,

$$\mathcal{B}_F(p, q) = \sum_{i \in [n]} \big(F(p_i) - F(q_i) - (p_i - q_i)F'(q_i)\big).$$

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. (1997) 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 principal 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 the 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, in Section 3, 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 cannot be approximated multiplicatively in the data-stream 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 $x_i - y_i$ the measure d cannot be sketched.
- For all f-divergences where f is twice differentiable and f'' is strictly positive, no polynomial factor approximation of $\mathcal{D}_f(p, q)$ is possible in $o(n)$ bits of space. Note that for ℓ_1, which can be sketched, the function $f(u) = |u - 1|$ and therefore f'' is not defined at 1.
- For all Bregman divergences $\mathcal{B}_F$ for which F is twice differentiable and there exists $\rho, z_0 > 0$ such that,

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

  no polynomial factor approximation of $\mathcal{B}_F$ is possible in $o(n)$ bits of 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, $F(z) = z^2$ and therefore F'' is constant everywhere.

Then, in Sect. 4, we consider finding additive approximations. We say an algorithm returns an (ϵ, δ)-additive-approximation for a real number Q if it outputs a value $\hat{Q}$ such that $|\hat{Q} - Q| \le \epsilon$ with probability at least $(1 - \delta)$ over its internal coin tosses. (ϵ, δ)-additive-approximation algorithms that took two passes over the data stream were presented in Guha

et al. (2006). In this paper, we show sharp characterizations about what can be achieved in a single pass. We show the following:

- If $\mathcal{D}_f$ is bounded, then there is an (ϵ, δ)-additive-approximation for $\mathcal{D}_f$ using $O(\epsilon^{-2}\tau(\epsilon)\log\delta^{-1}(\log n + \log m))$ bits of space where $\tau(\cdot)$ is a function determined by the derivative of f, e.g., $\tau(\epsilon) = O(1)$ for Triangle and $\tau(\epsilon) = O(\epsilon^{-1})$ for Hellinger. Complementing this, any $(\epsilon, 1/4)$-additive-approximation of $\mathcal{D}_f$ requires $\Omega(\epsilon^{-2})$ bits of space. Any $(\epsilon, 1/4)$-additive-approximation of an unbounded $\mathcal{D}_f$ requires $\Omega(n)$ bits of space for any ϵ.
- If F and F'' are bounded in the range $[0, 1]$, then there is an (ϵ, δ)-additive-approximation for $\mathcal{B}_F$ using $O(\epsilon^{-2}\log\delta^{-1}(\log n + \log m))$ bits of space. If $F(0)$ or $F'(0)$ are unbounded, any $(\epsilon, 1/4)$-additive-approximation of an unbounded $\mathcal{B}_F$ requires $\Omega(n)$ bits of space for any ϵ.

2 Preliminaries

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

2.1 f-divergences

We start by defining a *conjugate* $f^*(u) = uf(1/u)$. We can then write,

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

The following simple lemma from Guha et al. (2006) will allow us to make certain assumptions about f without loss of generality.

Lemma 4 *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. $\mathcal{D}_f = \mathcal{D}_g$.
2. *g is positive and if f is differentiable at 1 then $g'(1) = 0$.*
3. *If $\mathcal{D}_f$ is bounded then $g(0) = \lim_{u\to 0} g(u)$ and $g^*(0) = \lim_{u\to 0} g^*(u)$ exists.*

Proof For $p = (1/2, 1/2)$ and $q = (0, 1)$,

$$D_f(p, q) = (f(0) + f(2))/2 \quad \text{and} \quad D_f(q, p) = 0.5 \lim_{u\to 0} uf(1/u) + f(0.5).$$

Hence, if D_f is bounded then $f(0) = \lim_{u\to 0} f(u)$ and $f^*(0) = \lim_{u\to 0} f^*(u) = \lim_{u\to 0} uf(1/u)$ exist. Let $c = \lim_{u\to 1^-} \frac{f(1)-f(u)}{1-u}$. If f is differentiable then $c = f'(1)$. Otherwise, this limit still exists because f is convex and defined on $(0, \infty)$. Then $g(u) = f(u) - c(u - 1)$ satisfies the necessary conditions. □

For example, the Hellinger divergence can be realized by either $f(u) = (\sqrt{u} - 1)^2$ or $f(u) = 2 - 2\sqrt{u}$. Henceforth, we assume f is non-increasing in the range $[0, 1]$ and non-decreasing in the range $[1, \infty)$.

The next lemma shows that, if we are willing to tolerate an additive approximation, we may make certain assumptions about the derivative of f. This is achieved by approximating f by a straight line for very small and very large values.

Lemma 5 *Given a bounded $\mathcal{D}_f$ with f differentiable (w.l.o.g., f is unimodal and minimized at 1) and $\epsilon \in (0,1)$, let*

$$u_0(\epsilon) = \max\left\{u \in (0,1] : f(u)/f(0) \geq 1-\epsilon,\ f^*(u)/f^*(0) \geq 1-\epsilon\right\}$$

and define g:

$$g(u) = \begin{cases} f(u) & \text{for } u \in (u_0, 1/u_0), \\ f(0) - u(f(0) - f(u_0))/u_0 & \text{for } u \in [0, u_0], \\ uf^*(0) - (f^*(0) - f^*(u_0))/u_0 & \text{for } u \in [1/u_0, \infty). \end{cases}$$

Then, $\mathcal{D}_g(p,q)(1-\epsilon) \leq \mathcal{D}_f(p,q) \leq \mathcal{D}_g(p,q)$ and

$$\max_u |g'(u)| \leq \max\left(\epsilon f(0)/u_0,\ f^*(0)\right) \quad \text{and} \quad \max_u |g^{*\prime}(u)| \leq \max\left(\epsilon f^*(0)/u_0,\ f(0)\right).$$

Proof Because f, f^*, g, g^* are non-increasing in the range $[0,1]$, for all $u \in [0, u_0]$,

$$1 \leq \frac{g(u)}{f(u)} \leq \frac{f(0)}{f(u)} \leq \frac{f(0)}{f(u_0)} \quad \text{and} \quad 1 \leq \frac{g^*(u)}{f^*(u)} \leq \frac{f^*(0)}{f^*(u)} \leq \frac{f^*(0)}{f^*(u_0)}. \tag{1}$$

The first claim follows from equation 1 and the assumption that

$$\min\left(f(u_0)/f(0),\ f^*(u_0)/f^*(0)\right) \geq 1-\epsilon.$$

To bound the derivatives note that $g(u)$ and $g^*(u)$ are convex and hence the absolute value of the derivative is maximized at $u = 0$ or $u \to \infty$. The second claim follows by taking the derivative at these points and bounding $f(u_0) \geq (1-\epsilon)f(0)$ and $f^*(u_0) \geq (1-\epsilon)f^*(0)$. □

Note that $\lim_{u\to 0}|g'(u)|$ is bounded whereas $\lim_{u\to 0}|f'(u)|$ need not be bounded. For example, for the Hellinger divergence, $f(u) = (\sqrt{u}-1)^2$ and therefore $f'(u) = (\sqrt{u}-1)/\sqrt{u}$ which is unbounded as u tends to 0.

2.2 Bregman divergences

Similar to Lemma 4, the following lemma demonstrates that, without loss of generality, we may make various assumptions about the F that defines a Bregman divergence.

Lemma 6 *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. *$G(z) \geq 0$ for $x \in (0,1]$ and G is increasing in the range $(0,1]$.*
3. *$\lim_{u\to 0^+} G'(u) = 0$ and $\lim_{u\to 0^+} G(u) = 0$.*

Proof The function $G(z) = F(z) - F'(0)z - F'(0)$ satisfies the necessary conditions. □

3 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. In essence the results shows that it is impossible to approximate $d_\phi(p,q) = \sum_i \phi(p_i, q_i)$ in small space if $\phi(\alpha, \alpha+\delta)$ can vary significantly for different values of the "shift" α.

Theorem 7 (Shift Invariance Theorem) *Let $\phi : [0,1]^2 \to \mathbb{R}^+$ be such that $\phi(x,x) = 0$ for all $x \in [0,1]$ and for all sufficiently large n there exists $a, b, c \in \mathbb{N}$ such that,*

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

where $t = an/4 + bn + cn/2$. Then any algorithm returning an estimate of $d_\phi(p,q) = \sum_{i \in [5n/4]} \phi(p_i, q_i)$ within a factor α 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. This remains true even if the algorithm may take a constant number of passes over the stream.

The factor n on the right-hand side of (2) is only necessary if we wish to prove a $\Omega(n)$ space lower bound and thereby rule out sub-linear space algorithms. More generally, if n is replaced by some $w \leq n$ then the lower bound would become $\Omega(w)$. However, the above formulation will be sufficient for the purposes of proving results on the estimation of information divergences.

Proof The proof is by a reduction from the communication complexity of the SET-DISJOINTNESS problem. An instance of this problem consists of two binary strings, $x, y \in \{0,1\}^n$ such that $\sum_i x_i = \sum_i y_i = n/4$. We consider two players, Alice and Bob, such that 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., $x \cdot y = 0$ (where the inner product is taken over the reals). It is known that determining if $x \cdot y = 0$ with probability at least $3/4$ requires $\Omega(n)$ bits to be communicated (Razborov 1992).

However, suppose that there exists a streaming algorithm $\mathcal{A}$ that takes P passes over a stream and uses W working memory to α-approximate $d_\phi(p,q)$ with probability $3/4$. We will show that this algorithm gives rise to a $(2P-1)$-round protocol for SET-DISJOINTNESS that only requires $O(PW)$ bits to be communicated and therefore $W = \Omega(n/P)$.

We will assume that $\phi(a/t, (a+c)/t) \geq \phi((a+c)/t, a/t)$. If $\phi(a/t, (a+c)/t) \leq \phi((a+c)/t, a/t)$ then the proof follows by reversing the roles of the p and q that we now define. Consider the multi-sets,

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

This defines the following frequencies:

$$m(p)_i = \begin{cases} a & \text{if } x_i = 1 \text{ and } i \in [n], \\ b & \text{if } x_i = 0 \text{ and } i \in [n], \\ b + c & \text{if } n < i \leq 5n/4, \end{cases} \quad \text{and}$$

$$m(q)_i = \begin{cases} a & \text{if } (x_i, y_i) = (1, 0) \text{ and } i \in [n], \\ b & \text{if } (x_i, y_i) = (0, 0) \text{ and } i \in [n], \\ a + c & \text{if } (x_i, y_i) = (1, 1) \text{ and } i \in [n], \\ b + c & \text{if } (x_i, y_i) = (0, 1) \text{ and } i \in [n], \\ b & \text{if } n < i \leq 5n/4. \end{cases}$$

Consequently,

$$\begin{aligned} d_\phi(p, q) = (x \cdot y)\phi\left(a/t, (a + c)/t\right) + (n/4 - x \cdot y)\phi\left(b/t, (b + c)/t\right) \\ + (n/4)\phi\left((b + c)/t, b/t\right), \end{aligned}$$

where $t = m(p) = m(q) = an/4 + bn + cn/2$ and therefore,

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

Hence any α-approximation of $d_\phi(p, q)$ determines if $x \cdot y = 0$. Alice and Bob can emulate $\mathcal{A}$ on $S_A(x) \cup S_B(y)$ in the natural way: 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. If the algorithm returns an α-approximation for $d_\phi(p, q)$ then Bob can successfully infer if $x \cdot y = 0$ from the approximation. □

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. It is also possible to prove a simpler result for single pass algorithms using a reduction from the communication complexity of the INDEX problem, a variant of the SET-DISJOINTNESS problem in which Bob's string has weight one. In this case the result states that if there exist $a, b, c \in \mathbb{N}$ such that

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

where $t = an/4 + 3bn/4 + b + c$, then any single-pass α-approximation of $\sum_{i \in [n+1]} \phi(p_i, q_i)$ requires $\Omega(n)$ bits of space.

We next present two corollaries of Theorem 7. These characterize the f-divergences and Bregman divergences that can not be sketched. We 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. In both cases the result follows by finding conditions on f or F such that it is possible to appeal to Theorem 7.

Corollary 8 (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 $o(n)$ bits of space.*

The idea behind the proof is to establish that $\frac{b}{t} f(b/(b+1))$ can be made very small compared to $\frac{1}{t} f(1/2)$ for large b if $f'(u)$ tends to zero as u tends to 1. A sufficient condition for this to happen will be if f is twice differentiable and f'' is strictly positive. For ℓ_1, which can be sketched, the function $f(u) = |u-1|$ and therefore f'' is not defined at 1.

Proof We first note that by Lemma 4 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,

$$\phi(b/t, (b+c)/t) = (b/t) f(1+1/b) = (b/t)\left[f(1) + \frac{1}{b} f'(1) + \frac{1}{2!b^2} f''(1+\gamma) \right]$$

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,

$$\phi(b/t, (b+c)/t) \leq \frac{f''(1)+1}{2tb} = \frac{f''(1)+1}{2f(2)b} t^{-1} f(2) \leq \frac{8}{\alpha^2 n} \phi(a/t, (a+c)/t).$$

Similarly we can show that for sufficiently large n, $\phi((b+c)/t, b/t) \leq 8\phi(a/t, (a+c)/t)/(\alpha^2 n)$. Then, appealing to Theorem 7 we get the required result. □

Corollary 9 (Bregman Divergences) *Given a Bregman divergence $\mathcal{B}_F$, if F is twice differentiable and there exists $\rho, z_0 > 0$ such that,*

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

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

then no polynomial factor approximation of $\mathcal{B}_F$ is possible in $o(n)$ bits of space.

This condition effectively states that $F''(z)$ vanishes or diverges monotonically, and polynomially fast, as $z \to 0$. Note that for ℓ_2^2, which can be sketched, $F(z) = z^2$ and therefore F'' is constant everywhere.

Proof By the Mean-Value Theorem, for any $t, r \in \mathbb{N}$, there exists $\gamma(r) \in [0, 1]$ such that,

$$\begin{aligned} \phi(r/t, (r+1)/t) + \phi(r/t + 1/t, r/t) &= t^{-1}(F'(r/t + 1/t) - F'(r/t)) \\ &= t^{-2} F''((r + \gamma(r))/t). \end{aligned}$$

Therefore, for any $a, b \in \mathbb{N}, c = 1$ and $t = an/4 + bn + n/2$,

$$\frac{\max(\phi(\frac{a}{t}, \frac{a+c}{t}), \phi(\frac{a+c}{t}, \frac{a}{t}))}{\phi(\frac{b+c}{t}, \frac{b}{t}) + \phi(\frac{b}{t}, \frac{b+c}{t})} \geq \frac{1}{2} \frac{F''((a+\gamma(a))/t)}{F''((b+\gamma(b))/t)}. \tag{3}$$

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 $\forall\, 0 \leq z_2 \leq z_1 \leq z_0,\ F''(z_1)/F''(z_2) \leq (z_2/z_1)^{\rho}$ then set $a = 1$ and $b = (\alpha n)^{1/\rho}$. In both cases we deduce that the RHS of (3) is greater than $\alpha^2 n/4$. Hence, appealing to Theorem 7, we get the required result. □

4 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. We first present a general algorithmic result based on an extension of a technique first used by Alon et al. (1999). We then prove two general lower bounds. In the subsequent sections, we consider f-divergences and Bregman divergences in particular.

Theorem 10 *For $\phi : [0,1]^2 \to \mathbb{R}$ such that $\phi(0,0) = 0$, there exists an (ϵ, δ)-additive-approximation for $d_\phi(p,q)$ using $O(\tau\epsilon^{-2}\log\delta^{-1}(\log n + \log m))$ bits of space where*

$$\tau = 4 \max_{x,y\in[0,1]} \left(\left| \frac{\partial}{\partial x}\phi(x,y) \right| + \left| \frac{\partial}{\partial y}\phi(x,y) \right| \right).$$

The algorithm does not need to know $m(p)$ or $m(q)$ in advance.

Proof We will describe a basic estimator that can be computed in small space without prior knowledge of $m(p)$ or $m(q)$. We will then argue that the estimator is correct in expectation. Finally, we show that, by averaging a small number of independent basic estimators, we may return a sufficiently accurate estimator with the necessary probability.

Let $d \in_R \{p,q\}$ and $j_d \in_R [m(d)]$ where $\in_R$ denotes an element being chosen uniformly from the relevant set. Let $a_j = \langle d,k\rangle$ be the j_d-th element in the stream of the form $\langle d,\cdot\rangle$ and compute

$$r := I[d=p]\cdot r_0 + |\{\ell > j : a_\ell = \langle p,k\rangle\}|\cdot m(q),$$
$$s := I[d=q]\cdot s_0 + |\{\ell > j : a_\ell = \langle q,k\rangle\}|\cdot m(p)$$

where $r_0 \in_R [m(q)]$ and $s_0 \in_R [m(p)]$.

Note that r and s can be computed without prior knowledge of $m(p)$ and $m(q)$ since all that needs to be computed before we reach the end of the stream is j_d, $|\{\ell > j : a_\ell = \langle p,k\rangle\}|$, and $|\{\ell > j : a_\ell = \langle q,k\rangle\}|$. To do this we set $j_d = 1$ when we see the first element of the form $\langle d,\cdot\rangle$ and start computing $|\{\ell > j : a_\ell = \langle p,k\rangle\}|$, and $|\{\ell > j : a_\ell = \langle q,k\rangle\}|$. On seeing the i-th element of the form $\langle d,\cdot\rangle$ we reset j_d to i with probability $1/i$ and start computing $|\{\ell > j : a_\ell = \langle p,k\rangle\}|$, and $|\{\ell > j : a_\ell = \langle q,k\rangle\}|$ afresh. Note that the probability that j_d is set to i at the end of the algorithm is

$$\frac{1}{i}\left(1-\frac{1}{i+1}\right)\left(1-\frac{1}{i+2}\right)\cdots\left(1-\frac{1}{m(d)}\right) = \frac{1}{i}\cdot\frac{i}{i+1}\cdot\frac{i+1}{i+2}\cdots\frac{m(d)-1}{m(d)} = \frac{1}{m(d)},$$

and hence this procedure computes j_d, $|\{\ell > j : a_\ell = \langle p,k\rangle\}|$, and $|\{\ell > j : a_\ell = \langle q,k\rangle\}|$ as required.

Given r and s we define the basic estimator as

$$X(r,s) = \begin{cases} 2m^*(\phi(r/m^*, s/m^*) - \phi(r/m^* - 1/m^*, s/m^*)) & \text{if } d = p \\ 2m^*(\phi(r/m^*, s/m^*) - \phi(r/m^*, s/m^* - 1/m^*)) & \text{if } d = q \end{cases}$$

where $m^* = m(p)m(q)$.

Note that $\Pr[k=i] = (p_i + q_i)/2$ and that, because of a telescoping property of the appropriate sum,

$$E[X(r,s)|k=i] = 2m^*\left(\frac{\phi(m(p)_i m(q)/m^*, m(q)_i m(p)/m^*)}{m(p)m(q)_i + m(q)m(p)_i}\right) = \frac{2\phi(p_i,q_i)}{p_i+q_i}.$$

To see this consider the sub-stream consisting of the elements of the form $\langle \cdot, i\rangle$, e.g.,

$$\langle p,i\rangle, \langle q,i\rangle, \langle q,i\rangle, \langle p,i\rangle, \langle q,i\rangle, \langle p,i\rangle.$$

and expand $E[X(r,s)|k=i]$ as follows:

$$\begin{aligned}
E[X(r,s)|k=i] = \gamma\Bigg(& \sum_{i\in[m(q)]} X(2m(q)+i, 3m(p)) + \sum_{i\in[m(p)]} X(2m(q), 2m(p)+i) \\
& + \sum_{i\in[m(p)]} X(2m(q), m(p)+i) + \sum_{i\in[m(q)]} X(m(q)+i, m(p)) \\
& + \sum_{i\in[m(p)]} X(m(q), i) + \sum_{i\in[m(q)]} X(i,0)\Bigg) \\
= 2m^*\gamma\big(& \phi(2/m(q), 3/m(q)) - \phi(2/m(q), 2/m(q)) \\
& + \phi(2/m(p), 2/m(q)) - \phi(2/m(q), 1/m(q)) \\
& + \phi(2/m(p), 1/m(q)) - \phi(1/m(q), 1/m(q)) \\
& + \phi(1/m(p), 1/m(q)) - \phi(1/m(q), 0/m(q)) \\
& + \phi(1/m(p), 0/m(q)) - \phi(0/m(q), 0/m(q))\big) \\
= & \frac{2\phi(p_i,q_i)}{p_i+q_i}.
\end{aligned}$$

where $\gamma = \frac{1}{m(p)_i m(q) + m(p)m(q)_i}$.

Therefore $E[X(r,s)] = \sum_i \phi(p_i,q_i)$ as required. Furthermore,

$$|X(r,s)| \le 2\max\left\{\max_{x\in[\frac{r-1}{m^*},\frac{r}{m^*}]}\left|\frac{\partial}{\partial x}\phi(x, s/m^*)\right|, \max_{y\in[\frac{s-1}{m^*},\frac{s}{m^*}]}\left|\frac{\partial}{\partial y}\phi(r/m^*, y)\right|\right\} \le \tau.$$

Hence, by an application of the Chernoff bound, averaging $O(\tau\epsilon^{-2}\log\delta^{-1})$ independent basic estimators gives an (ϵ,δ)-additive-approximation. □

We next prove a lower bound on the space required for additive approximation by any single-pass algorithm. The proof uses a reduction from the one-way communication complexity of the GAP-HAMMING problem (Woodruff 2004). It is widely believed that a similar lower bound exists for multi-round communication (e.g. McGregor 2007, Question 10 (R. Kumar)) and, if this is the case, it would imply that the lower-bound below also applies to algorithms that take a constant number of passes over the data.

Theorem 11 *Any $(\epsilon, 1/4)$-additive-approximation of $d_\phi(p,q)$ requires $\Omega(\epsilon^{-2})$ bits of space if,*

$$\exists a,b>0, \forall x, \quad \phi(x,0) = ax, \qquad \phi(0,x) = bx, \qquad \phi(x,x) = 0.$$

Proof The proof is by a reduction from the communication complexity of the GAP-HAMMING problem. An instance of this problem consists of two binary strings, $x, y \in \{0,1\}^n$ such that $\sum_i x_i = \sum_i y_i = cn$ for some constant c. We consider two players, Alice and Bob, such that Alice knows the string x and Bob knows the string y. Alice sends a single message to Bob with the goal of Bob then being able to determine $d_H(x,y)$, the Hamming distance between x and y, up to an additive $\sqrt{n}$ term. It is known that achieving this with probability at least $3/4$ requires $\Omega(n)$ bits to be communicated (Woodruff 2004).

However, suppose that there exists a single-pass algorithm $\mathcal{A}$ using W working memory that returns an $(\epsilon, 1/4)$-additive-approximation for $d_\phi(p,q)$. We will show that this algorithm gives rise to a one-way protocol for GAP-HAMMING for $n = \lfloor \epsilon^{-2} \rfloor$ that only requires $O(W)$ bits to be communicated and therefore $W = \Omega(n)$.

Consider the sets $S_A(x) = \bigcup_{i:x_i=1}\{\langle p, i\rangle\}$ and $S_B(y) = \bigcup_{i:y_i=1}\{\langle q, i\rangle\}$. Then,

$$d_\phi(p,q) = \frac{a|\{i : x_i = 1, y_i = 0\}|}{cn} + \frac{b|\{i : x_i = 0, y_i = 1\}|}{cn} = d_H(x,y)\frac{a+b}{2cn}.$$

Therefore an $\epsilon(a+b)/(4c)$-additive-approximate determines $d_H(x,y)$ up to additive $\sqrt{n}$.

Alice and Bob can emulate $\mathcal{A}$ on $S_A(x) \cup S_B(y)$ in the natural way: Alice runs $\mathcal{A}$ on $S_A(x)$, communicates the memory state of $\mathcal{A}$ and then Bob runs $\mathcal{A}$ initiated with this memory state on $S_B(x)$. If the algorithm returns an $\epsilon(a+b)/(4c)$-additive-approximation for $d_\phi(p,q)$ then Bob can successfully infer $d_H(x,y)$ up to an additive $\sqrt{n}$. □

Finally in this section we demonstrate that no $o(n)$ space, constant pass algorithm can return any additive approximation if d_ϕ is unbounded.

Theorem 12 *Any $(\epsilon, 1/4)$-additive-approximation of $d_\phi(p,q) = \sum_{i\in[n]} \phi(p_i, q_i)$ requires $\Omega(n)$ space if either $\phi(x,0)$ or $\phi(0,x)$ is unbounded for all $x > 0$ and bounded otherwise. This applies even if one of the distributions is known to be uniform.*

Proof The proof is by a reduction from the communication complexity of the SET-DISJOINTNESS problem and is almost identical to the proof of Theorem 7. The only difference is that $S_A(x) = \bigcup_{i:x_i=0}\{\langle q, i\rangle\}$, $S_B(y) = \bigcup_{i:y_i=0}\{\langle q, i\rangle\}$, and we assume that p is the uniform distribution. If $\phi(1/n, 0)$ is unbounded then $d_\phi(p,q)$ is finite if and only if $x \cdot y = 0$. Otherwise, if $\phi(0, 1/n)$ is unbounded then $d_\phi(q,p)$ is finite if and only if $x \cdot y = 0$. □

4.1 Additive approximation for f-divergences

In this section we show that $\mathcal{D}_f(p,q)$ can be additively approximated up to any additive $\epsilon > 0$ if and only if $\mathcal{D}_f$ is bounded.

Theorem 13 *There exists a one-pass, $O(\epsilon^{-2}\tau(\epsilon)\log\delta^{-1}(\log n + \log m))$-space, (ϵ,δ)-additive-approximation for any bounded f-divergence where,*

$$\tau(\epsilon) = O(\epsilon/u_0) \quad \text{where } u_0 = \max\left\{u \in (0,1] : f(u)/f(0) \geq 1-\epsilon,\ f^*(u)/f^*(0) \geq 1-\epsilon\right\}.$$

For example, $\tau(\epsilon) = O(1)$ for Triangle and $\tau(\epsilon) = O(\epsilon^{-1})$ for Hellinger. The algorithm does not need to know $m(p)$ or $m(q)$ in advance.

Proof We appeal to Theorem 10 and note that,

$$\max_{x,y\in[0,1]}\left(\left|\frac{\partial}{\partial x}\phi(x,y)\right|+\left|\frac{\partial}{\partial y}\phi(x,y)\right|\right)=\max_{x,y\in[0,1]}(|f(y/x)-(y/x)f'(y/x)|+|f'(y/x)|)$$
$$\leq 2\max_{u\geq 0}(|f^{*\prime}(u)|+|f'(u)|).$$

By Lemma 5, we may bound the derivatives of f and f^* in terms of the additive approximation error ϵ. This gives the required result. □

We complement Theorem 13 with the following result which follows from Theorems 11 and 12.

Theorem 14 *Any $(\epsilon, 1/4)$-additive-approximation of an unbounded $\mathcal{D}_f$ requires $\Omega(n)$ bits of space. This applies even if one of the distributions is known to be uniform. Any $(\epsilon, 1/4)$-additive-approximation of a bounded $\mathcal{D}_f$ requires $\Omega(\epsilon^{-2})$ bits of space.*

4.2 Additive approximation for Bregman divergences

In this section we prove a partial characterization of the Bregman divergences that can be additively approximated.

Theorem 15 *There exists a one-pass, $O(\epsilon^{-2}\log\delta^{-1}(\log n+\log m))$-space, (ϵ,δ)-additive-approximation of a Bregman divergence if F and F'' are bounded in the range $[0,1]$. The algorithm does not need to know $m(p)$ or $m(q)$ in advance.*

Proof We appeal to Theorem 10 and note that,

$$\max_{x,y\in[0,1]}\left(\left|\frac{\partial}{\partial x}\phi(x,y)\right|+\left|\frac{\partial}{\partial y}\phi(x,y)\right|\right)=\max_{x,y\in[0,1]}\left(\left|F'(x)-F'(y)\right|+|x-y|F''(y)\right).$$

We may assume this is constant by convexity of F and the assumptions of the theorem. The result follows. □

The next theorem follows immediately from Theorem 12.

Theorem 16 *If $F(0)$ or $F'(0)$ is unbounded then an $(\epsilon, 1/4)$-additive-approximation of $\mathcal{B}_F$ requires $\Omega(n)$ bits of space even if one of the distributions is known to be uniform.*

5 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 presented algorithms and lower bounds for the additive approximation of information divergences.

Our first open question concerns multiplicative approximation of information divergences in the *aggregate data-stream model* in which all elements of the form $\langle r, i\rangle$ appear

consecutively for each $i \in [n], r \in \{p, q\}$. It is easy to $(1+\epsilon)$ multiplicatively approximate the Hellinger divergence in this model using $O(\epsilon^{-2} \operatorname{polylog} m)$ bits of 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 $O(\operatorname{polylog} m)$ space. How much space is required to find an $(1+\epsilon)$-approximation?

Our second open question concerns additive approximation in the *distributed data-stream model*. In this model, the data-stream defining p and q is partitioned into multiple sub-streams and each sub-stream is observed at a different location. After the sub-streams have been processed, a message is sent from each location to some central authority who returns an approximation of $d_\phi(p, q)$. While the lower-bounds we presented also apply in this model, the additive-approximation algorithms we presented required the assumption that the entire stream was observed at a single location. Is additive approximation possible in the distributed model?

Acknowledgements We would like to thank the anonymous referees for various helpful suggestions regarding the presentation of our results.

References

Alon, N., Matias, Y., & Szegedy, M. (1999). The space complexity of approximating the frequency moments. *Journal of Computer and System Sciences*, *58*(1), 137–147.

Amari, S. (1985). *Differential-geometrical methods in statistics*. New York: Springer.

Bar-Yossef, Z., Jayram, T. S., Kumar, R., & Sivakumar, D. (2004). An information statistics approach to data stream and communication complexity. *Journal of Computer and System Sciences*, *68*(4), 702–732.

Bhuvanagiri, L., Ganguly, S., Kesh, D., & Saha, C. (2006). Simpler algorithm for estimating frequency moments of data streams. In *ACM-SIAM symposium on discrete algorithms* (pp. 708–713).

Bregman, L. M. (1967). The relaxation method of finding the common point of convex sets and its application to the solution of problems in convex programming. *USSR Computational Mathematics and Mathematical Physics*, *7*(1), 200–217.

Breiman, L. (1999). Prediction games and arcing algorithms. *Neural Computation*, *11*(7), 1493–1517.

Brinkman, B., & Charikar, M. (2003). On the impossibility of dimension reduction in ℓ_1. In *IEEE symposium on foundations of computer science* (pp. 514–523).

Broder, A. Z., Charikar, M., Frieze, A. M., & Mitzenmacher, M. (2000). Min-wise independent permutations. *Journal of Computer and System Sciences*, *60*(3), 630–659.

Chakrabarti, A., Khot, S., & Sun, X. (2003). Near-optimal lower bounds on the multi-party communication complexity of set disjointness. In *IEEE conference on computational complexity* (pp. 107–117).

Chakrabarti, A., Cormode, G., & McGregor, A. (2007). A near-optimal algorithm for computing the entropy of a stream. In *ACM-SIAM symposium on discrete algorithms* (pp. 328–335).

Charikar, M., Chen, K., & Farach-Colton, M. (2002). Finding frequent items in data streams. In *International colloquium on automata, languages and programming* (pp. 693–703).

Collins, M., Schapire, R. E., & Singer, Y. (2002). Logistic regression, AdaBoost and Bregman distances. *Machine Learning*, *48*(1–3), 253–285.

Cormode, G., & Ganguly, S. (2007). On estimating frequency moments of data streams. In *International workshop on randomization and approximation techniques in computer science*.

Cormode, G., & Muthukrishnan, S. (2005). An improved data stream summary: the count-min sketch and its applications. *Journal of Algorithms*, *55*(1), 58–75.

Cormode, G., Datar, M., Indyk, P., & Muthukrishnan, S. (2003). Comparing data streams using hamming norms (how to zero in). *IEEE Transactions on Knowledge and Data Engineering*, *15*(3), 529–540.

Cover, T. M., & Thomas, J. A. (1991). *Elements of information theory. Wiley series in telecommunications* Wiley, New York.

Csiszár, I. (1991). Why least squares and maximum entropy? An axiomatic approach to inference for linear inverse problems. In *Annals of statistics* (pp. 2032–2056).

Feigenbaum, J., Kannan, S., Strauss, M., & Viswanathan, M. (2002). An approximate L^1 difference algorithm for massive data streams. *SIAM Journal on Computing*, *32*(1), 131–151.

Friedman, J., Hastie, T., & Tibshirani, R. (2000). Additive logistic regression: a statistical view of boosting. *Annals of Statistics*, *28*, 337–407.

Guha, S., McGregor, A., & Venkatasubramanian, S. (2006). Streaming and sublinear approximation of entropy and information distances. In *ACM-SIAM symposium on discrete algorithms* (pp. 733–742).

Henzinger, M. R., Raghavan, P., & Rajagopalan, S. (1999). Computing on data streams. In *External memory algorithms* (pp. 107–118).

Indyk, P. (2000). Stable distributions, pseudorandom generators, embeddings and data stream computation. In *IEEE symposium on foundations of computer science* (pp. 189–197).

Indyk, P., & Woodruff, D. P. (2005). Optimal approximations of the frequency moments of data streams. In *ACM symposium on theory of computing* (pp. 202–208).

Johnson, W. B., & Lindenstrauss, J. (1984). Extensions of Lipschitz mapping into Hilbert space. *Contemporary Mathematics*, *26*, 189–206.

Kivinen, J., & Warmuth, M. K. (1999). Boosting as entropy projection. In *Conference on learning theory* (pp. 134–144).

Lafferty, J. D. (1999). Additive models, boosting, and inference for generalized divergences. In *Conference on learning theory* (pp. 125–133).

Lafferty, J. D., Pietra, S. D., & Pietra, V. J. D. (1997). Statistical learning algorithms based on Bregman distances. In *Proceedings of Canadian workshop on information theory*.

Liese, F., & Vajda, F. (1987). Convex statistical distances. *Teubner-Texte zur Mathematik* (Band 95), Leipzig.

Mason, L., Baxter, J., Bartlett, P., & Frean, M. (1999). Functional gradient techniques for combining hypotheses. In *Advances in large margin classifiers*. Cambridge: MIT Press.

McGregor, A. (Ed.). (2007). *Open problems in data streams and related topics*. Available at: www.cse.iitk.ac.in/users/sganguly/data-stream-probs.pdf.

Nguyen, X., Wainwright, M. J., & Jordan, M. I. (2005). Divergences, surrogate loss functions and experimental design. In *Proceedings of NIPS*.

Razborov, A. A. (1992). On the distributional complexity of disjointness. *Theoretical Computer Science*, *106*(2), 385–390.

Saks, M. E., & Sun, X. (2002). Space lower bounds for distance approximation in the data stream model. In *ACM symposium on theory of computing* (pp. 360–369).

Woodruff, D. P. (2004). Optimal space lower bounds for all frequency moments. In *ACM-SIAM symposium on discrete algorithms* (pp. 167–175).