On the Exact Space Complexity of Sketching and Streaming Small Norms

Daniel M. Kane[*] Jelani Nelson[†] David P. Woodruff[‡]

Abstract

We settle the 1-pass space complexity of $(1 \pm \varepsilon)$-approximating the L_p norm, for real p with $1 \leq p \leq 2$, of a length-n vector updated in a length-m stream with updates to its coordinates. We assume the updates are integers in the range $[-M, M]$. In particular, we show the space required is $\Theta(\varepsilon^{-2} \log(mM) + \log\log(n))$ bits. Our result also holds for $0 < p < 1$; although L_p is not a norm in this case, it remains a well-defined function. Our upper bound improves upon previous algorithms of [Indyk, JACM '06] and [Li, SODA '08]. This improvement comes from showing an improved derandomization of the L_p sketch of Indyk by using k-wise independence for small k, as opposed to using the heavy hammer of a generic pseudorandom generator against space-bounded computation such as Nisan's PRG. Our lower bound improves upon previous work of [Alon-Matias-Szegedy, JCSS '99] and [Woodruff, SODA '04], and is based on showing a direct sum property for the 1-way communication of the gap-Hamming problem.

1 Introduction

Computing over massive data streams is increasingly important. Large data sets, such as sensor networks, transaction data, the web, and network traffic, have grown at a tremendous pace. It is impractical for most devices to store even a small fraction of the data, and this necessitates the design of extremely space-efficient algorithms. Such algorithms are often only given a single pass over the data, e.g., it may be expensive to read the contents of an external disk multiple times, and in the case of an internet router, it may be impossible to make multiple passes.

Even very basic statistics of a data set cannot be efficiently computed exactly or deterministically in this model, and so algorithms must be both approximate and probabilistic. This model is known as the streaming model and has become popular in the theory community, dating back to the works of Munro and Paterson [34] and Flajolet and Martin [15], and resurging with the work of Alon, Matias, and Szegedy [2]. For a survey of results, see the book by Muthukrishnan [35], or notes from Indyk's course [24].

A fundamental problem in this area is that of norm estimation [2]. Formally, we have a vector $x = (x_1, \ldots, x_n)$ initialized as $x = \vec{0}$, and a stream of m updates, where an update $(i, v) \in [n] \times \{-M, \ldots, M\}$ causes the change $x_i \leftarrow x_i + v$. This model is known as the *turnstile model* of streaming. In this work we consider the following problem:

Problem: Output a $(1 \pm \varepsilon)$-approximation to the value $L_p(x) = \left(\sum_{i=1}^n |x_i|^p\right)^{1/p}$ with success probability 2/3, over the randomness of the algorithm's private coins.

We allow for both multiplicative approximation and probability of error since it is known that linear space is required without these relaxations [2], and here we are concerned with small-space algorithms. Sometimes this problem is posed as estimating $F_p(x) = L_p^p(x)$, which is called the pth frequency moment of x.[1] A large body of work has been done in this area, see, e.g., the references in [24, 35].

It is known that not all L_p norms can be efficiently approximated in a data stream. In particular, [5, 8] show that polynomial space in n, m is required for $p > 2$, whereas space polylogarithmic in these parameters is achievable for $0 < p \leq 2$ [2, 23].[2] In this work, we focus on this feasible regime for p and consider the following question: what *exactly*

[*] Harvard University, Department of Mathematics. dankane@math.harvard.edu. Supported by a National Defense Science and Engineering Graduate (NDSEG) Fellowship.

[†] MIT Computer Science and Artificial Intelligence Laboratory. minilek@mit.edu. Supported by a National Defense Science and Engineering Graduate (NDSEG) Fellowship, and in part by the Center for Massive Data Algorithmics (MADALGO) - a center of the Danish National Research Foundation. Part of this work was done while the author was at the IBM Almaden Research Center.

[‡] IBM Almaden Research Center, 650 Harry Road, San Jose, CA, USA. dpwoodru@us.ibm.com.

[1] Note that for constant p bounded away from 0, which is the focus of our work, $(1 \pm \varepsilon)$-approximating L_p and F_p are equivalent up to a constant factor change in the approximation parameter ε.

[2] When $0 < p < 1$, L_p is not a norm since it does not satisfy the triangle inequality, though it is still well-defined.

is the space complexity of norm estimation for $0 < p \leq 2$? We remark that streaming approximations to L_p in this range area interesting for several reasons. L_1 estimation is used as a subroutine for dynamic earthmover distance approximation [22], approximate linear regression and best rank-k approximation of a matrix (with respect to the L_1 norm) [14], cascaded norm estimation of a matrix [26], and network traffic monitoring [13]. L_2 estimation is useful for database query optimization [1] and network traffic anomaly detection [28]. Both L_1 and L_2 estimation subroutines are used in approximate histogram maintenance [19]. Norm estimation for fractional p was shown useful for mining tabular data in [11] ($p = 0.5$ and $p = 0.25$ were specifically suggested), and L_p estimation for fractional p near 1 is used as a subroutine for estimating empirical entropy, which in turn is again useful for network traffic anomaly detection (see [21] and the references therein). Also, L_p estimation for all $0 < p \leq 2$ is used as a subroutine for weighted sampling in turnstile streams [33].

1.1 Contributions We resolve the space complexity of L_p-estimation for $0 < p \leq 2$ up to constant factors. In particular, the space complexity is $\Theta(\varepsilon^{-2}\log(mM) + \log\log(n))$ bits. For p strictly less than 2, our upper bound is new, and our lower bound is new for all $0 < p \leq 2$. Henceforth, we omit an implicit additive $\log\log n$ which exists in all the L_p space upper and lower bounds (justification is in Section A.1).

1.1.1 Tight upper bounds for L_p-estimation Our first contribution is the first optimal 1-pass space upper bound for L_p-estimation, $0 < p < 2$. In particular, we give an improved derandomization of Indyk's algorithm [23] to use k-wise independence for small k as opposed to Nisan's pseudorandom generator [36] against space-bounded computation. Our improved derandomization allows for an implementation using $O(\varepsilon^{-2}\log(mM))$ bits of space. An algorithm achieving this bound was previously known only for $p = 2$ [2]. In the case of $0 < p < 2$, the previously most space-efficient algorithms are due to Indyk [23] and Li [30], both requiring $O(\varepsilon^{-2}\log(mM)\log(N))$ space with $N = \min\{n, m\}$. A more prudent analysis of the seed length Nisan's generator requires to fool Indyk's algorithm can give a space bound of $O(\varepsilon^{-2}\log(mM) + \log(mM)\log(N))$, but this is still suboptimal. PRGs are a central tool in the design of streaming algorithms, and Indyk's algorithm had become the canonical example of a streaming algorithm for which no derandomization more efficient than via a generic PRG was known. We believe that removing this heavy hammer from norm estimation is an important step forward in improving the derandomization of streaming algorithms.

To see where our improvement comes from, let us recall Indyk's algorithm. That algorithm maintains a linear sketch of the vector x, i.e. the matrix-vector product $Ax = y$ for some $r \times n$ matrix A. In his sketch, $r = \Theta(1/\varepsilon^2)$. The matrix entries $A_{i,j}$ are all i.i.d. from a discretized *p-stable distribution* $\mathcal{D}_p$. The distribution $\mathcal{D}_p$ has the property that for all vectors $x \in \mathbb{R}^n$ and i.i.d. random variables $\{Z_i\}_{i=1}^n$ distributed according to $\mathcal{D}_p$, it holds that $\sum_{i=1}^n Z_i x_i \sim ||x||_p Z$, where $Z \sim \mathcal{D}_p$. His algorithm then returns the median of the $|y_j|$ for $j \in [r]$. Li [30] maintains the same linear sketch, but provides several different estimators which have certain advantages over Indyk's median estimator. The main issue with Indyk's algorithm, and also Li's, is that the number of bits required to store A is $\Omega(N/\varepsilon^2)$, which a polylogarithmic-space algorithm cannot afford. Indyk remedied this problem by using Nisan's PRG [36] to generate the matrix from a short seed, which a small-space algorithm can afford to store. However, the seed length required by Nisan's PRG was an $O(\log(N))$ factor larger than the algorithm's storage, causing the overall space to increase by this factor.

We first show that Indyk's algorithm has a more efficient derandomization. Rather than using Nisan's PRG, we show that it suffices for the entries in each row of A to be k-wise independent for $k = \tilde{O}(1/\varepsilon^p)$, and that the seeds used to generate each row need only be pairwise independent. This statement, together with bounds on the precision required throughout the algorithm, implies $O(\varepsilon^{-2}\log(mM))$ space suffices for $p < 2$. Though, since the time to process a stream update is $O(kr)$, it is desirable to have k as small as possible. We then give an alternative estimator to Indyk's median estimator for which we show $k = O(\log(1/\varepsilon)/\log\log(1/\varepsilon))$ suffices, thus yielding an algorithm which also has optimal space, but with improved update time.

Other work on L_p estimation for $0 < p \leq 2$ includes the work of Ganguly and Cormode [18], which requires a suboptimal $O(\varepsilon^{-(2+p)}\log^{O(1)}(mM))$ bits of space, but at the benefit of requiring $\log^{O(1)}(mM)$ update time independent of ε.

We remark here that our techniques seem possibly applicable to other derandomization questions. For example, using our techniques one can give an alternative proof of one of the key components used in [12] to show that bounded independence fools halfspaces. In particular, for $a \in \mathbb{R}^n$ with $||a||_2 = 1$ and $\theta \in \mathbb{R}$, consider the function $f_{a,\theta}(x) = \mathrm{sgn}(\langle a, x\rangle - \theta)$ where $x \in \{-1, 1\}^n$. The work of [12] showed that $\mathbf{E}_x[f_{a,\theta}(x)]$ is preserved to within $O(\varepsilon)$ even if the x_i are only $k(\varepsilon)$-wise independent for $k(\varepsilon) = O(\varepsilon^{-2}\log^2(1/\varepsilon))$. At a high level, their proof proceeded in two main steps: (a) reduce the general case to *regular* halfspaces, where $|a_i| < \varepsilon$ for all i, then (b) show that $k(\varepsilon)$-wise independence fools regular halfspaces. We show in Section A.5 that our techniques naturally apply to give an alternative proof of (b), with the slightly weaker bound $k(\varepsilon) = \varepsilon^{-2}\log^{2+o(1)}(1/\varepsilon)$.

1.1.2 Tight space lower bounds for L_p-estimation

We show a space lower bound of $\Omega(\min\{N, \varepsilon^{-2}\log(\varepsilon^2 mM)\})$ bits for all $0 < p \le 2$. This is $\Omega(\varepsilon^{-2}\log(mM))$ for $\varepsilon \ge 1/N^{1/2-\delta}$ for any constant $\delta > 0$, matching our upper bound in this range. This is perhaps the most interesting range for ε since the following trivial algorithms are always possible: maintain the entire vector, or entire stream, in memory. Thus when $\varepsilon < 1/\sqrt{N}$, the cheaper of these solutions has cost $O(N\log(mM))$ (Section A.1.1 justifies avoiding logarithmic dependence on n), showing that trivial solutions are already nearly optimal for such small ε.

The previous lower bound was $\Omega(\min\{N, \varepsilon^{-2} + \log N\})$, and is the result of a sequence of work [2, 6, 38]. See [25, 39] for simpler proofs. Given our lower bound and algorithm above, and the L_2-estimation algorithm of Alon, Matias, and Szegedy [2], the space complexity of L_p-estimation is now resolved for all $0 < p \le 2$. Our bound holds even when each coordinate is updated twice, implying that our algorithm (and a previous algorithm of Feigenbaum et al. [13]) is space-optimal even for the simpler problem of L_1-difference estimation. Our lower bound is also the first to have logarithmic dependence on mM (previously only a folklore $\Omega(\log\log(mM))$ bound was known via reduction from the communication complexity of Equality).

Our lower bounds are based upon embedding multiple geometrically-growing hard instances for estimating L_p in an insertion-only stream into a stream, and using the deletion property together with the geometrically-growing property to reduce the problem to solving a single hard instance. More precisely, a hard instance for L_p is based on a reduction from a two-party communication game in which the first party, Alice, receives a string $x \in \{0,1\}^{\varepsilon^{-2}}$, and Bob an index $i \in [\varepsilon^{-2}]$, and Alice sends a single message to Bob who must output x_i with constant probability. This problem, known as Indexing, requires $\Omega(\varepsilon^{-2})$ communication. To reduce it to estimating L_p in an insertion-only stream, there is a reduction [38, 39] through the gap-Hamming problem for which Alice creates a stream $\mathcal{S}_x$ and Bob a stream $\mathcal{S}_i$, with the property that either $L_p(\mathcal{S}_x \circ \mathcal{S}_i) \ge \varepsilon^{-2}/2 + \varepsilon^{-1}/2$, or $L_p(\mathcal{S}_x \circ \mathcal{S}_i) \le \varepsilon^{-2}/2 - \varepsilon^{-1}/2$. Here, "$\circ$" denotes concatenation of two streams. Thus, any 1-pass streaming algorithm which $(1 \pm \varepsilon)$-approximates L_p requires space which is at least the communication cost of Indexing, namely, $\Omega(\varepsilon^{-2})$.

We instead consider the Augmented-Indexing problem. Set $t = \Theta(\varepsilon^{-2}\log(M))$ (a different setting is made to gain the dependence on m). We give Alice a string $x \in \{0,1\}^t$, and Bob both an index $i \in [t]$ together with $x_{i+1}, \ldots, x_t$. This problem requires $\Omega(t)$ bits of communication if Alice sends only a single message to Bob [4, 32]. Alice splits x into $b = \varepsilon^2 t$ equal-sized blocks $X_0, \ldots, X_{b-1}$. In the j-th block she uses the ε^{-2} bits in that block to create a stream $\mathcal{S}_{X_j}$ that is similar to what she would have created in the insertion-only case, but each non-zero item is given frequency roughly $2^{j/p}$, so that it contributes 2^j to F_p. Given i, Bob finds the block j in which i belongs, and creates a stream $\mathcal{S}_i$ as in the insertion-only case, but where each non-zero item is given frequency roughly $-2^{j/p}$. Moreover, Bob can create all the streams $\mathcal{S}_{X_{j'}}$ for blocks j' above block j. Bob inserts all of these latter stream items as deletions, while Alice inserts them as insertions. Thus, when running an F_p-estimation algorithm on Alice's list of streams followed by Bob's, all items in streams $\mathcal{S}_{X_{j'}}$ vanish. Due to the geometrically increasing contribution of blocks to F_p, approximating F_p well on the entire stream corresponds to approximating F_p well on $\mathcal{S}_{X_j} \circ \mathcal{S}_i$, and thus a $(1 \pm \varepsilon)$-approximation algorithm to F_p can be used to solve Augmented-Indexing. Our technique can be viewed as showing a direct sum property for the one-way communication complexity of gap-Hamming.

Variations on our proof can also be used to show lower bounds for $p = 0$, for additive entropy estimation, and for norm estimation in the strict turnstile model (where no x_i can ever be negative). A discussion is in Section 3. Variants of our techniques were also useful for obtaining tight bounds for linear algebra problems in a stream [10] and in compressed sensing [3].

1.2 Notation

For integer $z > 0$, $[z]$ denotes the set $\{1, \ldots, z\}$. All our space bounds are measured in bits. The variables n, m, M denote vector length, stream length, and the maximum absolute value of a frequency update, respectively, and N denotes $\min\{n, m\}$. For a function $f : \mathbb{R} \to \mathbb{R}$ and nonnegative integer ℓ, $f^{(\ell)}$ denotes the ℓth derivative of f, with $f^{(0)} = f$. We also often use $x \approx_\varepsilon y$ to state that $|x - y| = O(\varepsilon)$.

2 Optimal L_p Estimation $(0 < p < 2)$

Here we describe a proof that Indyk's L_p-estimation algorithm can be more efficiently derandomized (Figure 1) to produce a space-optimal algorithm, though with update time $\tilde{O}(\varepsilon^{-2-p})$. We then describe a new algorithm (Figure 2) which maintains space-optimality while having update time only $\tilde{O}(\varepsilon^{-2})$. The tilde notation here

```
1. Pick a random matrix $A \in \mathbb{R}^{r \times n}$ as follows for $r = \Theta(1/\varepsilon^2)$. Each $A_{i,j}$ is distributed according to $\mathcal{D}_p$. For fixed $i$, the $A_{i,j}$ are $k$-wise independent with $k = \Theta(\varepsilon^{-p}\log^{3p}(1/\varepsilon))$. For $i \neq i'$, the seeds used to generate the $\{A_{i,j}\}_{j=1}^n$ and $\{A_{i',j}\}_{j=1}^n$ are pairwise independent.
2. Maintain the vector $y = Ax$ throughout the stream.
3. Output $\mathrm{median}\{|y_i|\}_{i=1}^r / \mathrm{median}(|\mathcal{D}_p|)$.
```

Figure 1: Indyk's derandomized L_p estimation algorithm pseudocode, $0 < p < 2$, assuming infinite precision. Remarks on computing $\mathrm{median}(|\mathcal{D}_p|)$ are in Section A.2.

hides $\log^{O(1)}(1/\varepsilon)$ factors.

We assume $p \in (0, 2)$ is a fixed constant bounded away from 0. Some constants in our asymptotic notation are functions of p. We also assume $||x||_p > 0$; $||x||_p = 0$ is detected when $y = 0$ in both Figure 1 and Figure 2. Finally, we assume $\varepsilon \geq 1/\sqrt{m}$. Otherwise, the trivial solution of keeping the entire stream in memory requires $O(m \log(nM)) = O(\varepsilon^{-2}\log(nM))$ space, which can be made $O(\varepsilon^{-2}\log(mM))$ by the argument in Section A.1.1. Our first main theorem is the following.

Theorem 2.1. *For all $p \in (0, 2)$, the algorithm of Figure 1 can be implemented with limited precision to use space $O(\varepsilon^{-2}\log(mM))$ and output $(1 \pm \varepsilon)||x||_p$ with probability at least $7/8$.*

To understand the first step of Figure 1, we recall the definition of a p-stable distribution.

Definition 2.1. (Zolotarev [41]) *For $0 < p < 2$, there exists a probability distribution $\mathcal{D}_p$ called the p-stable distribution with $\mathbf{E}[e^{itZ}] = e^{-|t|^p}$ for $Z \sim \mathcal{D}_p$. For any n and vector $x \in \mathbb{R}^n$, if $Z_1, \ldots, Z_n \sim \mathcal{D}_p$ are independent, then $\sum_{j=1}^n Z_j x_j \sim ||x||_p Z$ for $Z \sim \mathcal{D}_p$.*

We also state a lemma giving the decay of the density function of $\mathcal{D}_p$, which will be useful later.

Lemma 2.1. (Nolan [37, Theorem 1.12]) *For fixed $0 < p < 2$, the probability density function of the p-stable distribution is $\Theta(|x|^{-p-1})$ for large $|x|$.*

We now prove the following technical lemma, which plays a role in our later analyses.

Lemma 2.2. *There exists an $\varepsilon_0 > 0$ such that the following holds. Let n be a positive integer and $0 < \varepsilon < \varepsilon_0$, $0 < p < 2$ be given. Let $f : \mathbb{R} \to \mathbb{R}$ satisfy $||f^{(\ell)}||_\infty = O(\alpha^\ell)$ for all $\ell \geq 0$, for some α satisfying $\alpha^p \geq \log(1/\varepsilon)$. Let $k = \alpha^p$. Let $a \in \mathbb{R}^n$ satisfy $||a||_p = O(1)$. Let X_i be a $3Ck$-independent family of p-stable random variables for C a suitably large even constant. Let Y_i be a fully independent family of p-stable random variables. Let $X = \sum_i a_i X_i$ and $Y = \sum_i a_i Y_i$. Then $\mathbf{E}[f(X)] = \mathbf{E}[f(Y)] + O(\varepsilon)$.*

Proof. The basic idea of the proof will be to show that the expectation can be computed to within $O(\varepsilon)$ just by knowing that the X_i's are $O(k)$-wise independent. Our main idea is to approximate f by a Taylor series and use our knowledge of the moments of the X_i. The problem is that the tails of p-stable distributions for $p < 2$ are wide, and hence the expectations of the moments are infinite. We circumvent this difficulty via a combination of truncation and approximate inclusion-exclusion.

Define the random variables

$$U_i = \begin{cases} 1 & \text{if } |a_i X_i| > \lambda \\ 0 & \text{otherwise} \end{cases}$$

and

$$X'_i = (1 - U_i)X_i = \begin{cases} 0 & \text{if } |a_i X_i| > \lambda \\ X_i & \text{otherwise} \end{cases},$$

where we set $\lambda = 1/\alpha$. We note a couple of properties of these. First,

$$\mathbf{E}[U_i] = O\left(\int_{|a_i|^{-1}\lambda}^{\infty} x^{-1-p}dx\right) = O\left(|a_i|^p \lambda^{-p}\right)$$

by Lemma 2.1. We would also like to bound the moments of X'_i. We note that $\mathbf{E}[(a_i X'_i)^\ell]$ is 1 for $\ell = 0$, by symmetry is 0 when ℓ is odd, and otherwise is

(2.1)
$$\begin{aligned} O\left(\int_0^{|a_i|^{-1}\lambda} (a_i x)^\ell x^{-p-1}\right) &= O\left(|a_i|^\ell (|a_i|^{-1}\lambda)^{\ell-p}\right) \\ &= O\left(|a_i|^p \lambda^{\ell-p}\right) \end{aligned}$$

where the implied constant above can be chosen to hold independently of ℓ.

For $S \subseteq [n]$, let $\mathbf{1}_S$ be the indicator random variable for the event

$$S = \{i : U_i = 1\}.$$

Then since $\sum_{S \subseteq [n]} \mathbf{1}_S = 1$ for any point in the probability space,

(2.2)
$$\mathbf{E}[f(X)] = \mathbf{E}\left[\sum_{S\subseteq[n]} \mathbf{1}_S \cdot f(X)\right] = \sum_{S\subseteq[n]} \mathbf{E}[\mathbf{1}_S \cdot f(X)].$$

Now let $\mathbf{1}'_S$ be the indicator random variable for the event
$$S \subseteq \{i : U_i = 1\}$$
so that
$$\mathbf{1}_S = \mathbf{1}'_S \cdot \left(\prod_{i\notin S}(1-\mathbf{1}'_{\{i\}})\right) = \sum_{T\subset[n]\setminus S}(-1)^{|T|}\mathbf{1}'_{S\cup T},$$
and define
$$\begin{aligned} & F_{S,T}\left(\overrightarrow{X}\right) \\ &= (-1)^{|T|}\left(\prod_{i\in S\cup T} U_i\right) f\left(\sum_{i\in S} a_iX_i + \sum_{i\notin S} a_iX'_i\right). \end{aligned}$$

Then, by definition of U_i and Eq. (2.2),
$$\mathbf{E}[f(X)] = \mathbf{E}\left[\sum_{S\subseteq[n]}\sum_{T\subseteq[n]\setminus S} F_{S,T}\left(\overrightarrow{X}\right)\right].$$

We will approximate $\mathbf{E}[f(X)]$ as

$$\mathbf{E}\left[\sum_{\substack{S\subseteq[n]\\|S|\le Ck}}\sum_{\substack{T\subseteq[n]\setminus S\\|T|\le Ck}} F_{S,T}\left(\overrightarrow{X}\right)\right]. \tag{2.3}$$

That is, we approximate $\mathbf{E}[f(X)]$ using approximate inclusion-exclusion, by truncating the summation to not include large S, T. Call the function inside the expectation in Eq. (2.3) $F\left(\overrightarrow{X}\right)$. We would like to bound the error in approximating $f(X)$ by $F\left(\overrightarrow{X}\right)$. Fix values of the X_i, and let O be the set of i with $U_i = 1$. We note that
$$F\left(\overrightarrow{X}\right) = \sum_{\substack{S\subseteq O\\|S|\le Ck}}\sum_{\substack{T\subseteq O\setminus S\\|T|\le Ck}}(-1)^{|T|} f\left(\sum_{i\in S} a_iX_i + \sum_{i\notin S} a_iX'_i\right).$$

Notice that other than the $(-1)^{|T|}$ term, the expression inside the sum does not depend on T. This means that if $0 < |O\setminus S| \le Ck$ then the inner sum is 0, since $O\setminus S$ will have exactly as many even subsets as odd ones. Hence if $|O| \le Ck$, we have that
$$\begin{aligned} F\left(\overrightarrow{X}\right) &= \sum_{S=O} f\left(\sum_{i\in S} a_iX_i + \sum_{i\notin S} a_iX'_i\right) \\ &= f\left(\sum_{i\in O} a_iX_i + \sum_{i\notin O} a_iX'_i\right) \\ &= f(X). \end{aligned}$$

Otherwise, after fixing O and S, we can sum over possible values of $t = |T|$ and obtain:
$$\sum_{\substack{T\subseteq O\setminus S\\|T|\le Ck}}(-1)^{|T|} = \sum_{t=0}^{Ck}(-1)^t\binom{|O\setminus S|}{t}.$$

In order to bound this we use the following lemma.

Lemma 2.3. *For integers $A \ge B+1 > 0$ we have that $\sum_{i=0}^{B}(-1)^i\binom{A}{i}$ and $\sum_{i=0}^{B+1}(-1)^i\binom{A}{i}$ have different signs, with the latter sum being* 0 *if $A = B+1$.*

Proof. First suppose that $B < A/2$. We note that since the terms in each sum are increasing in i, each sum has the same sign as its last term, proving our result in this case. For $B \ge A/2$ we note that $\sum_{i=0}^{A}(-1)^i\binom{A}{i} = 0$, and hence letting $j = A - i$, we can replace the sums by $(-1)^{A+1}\sum_{j=0}^{A-B-1}(-1)^j\binom{A}{j}$ and $(-1)^{A+1}\sum_{j=0}^{A-B-2}(-1)^j\binom{A}{j}$, reducing to the case of $B' = A - B - 1 < A/2$. ■

Using Lemma 2.3, we note that $\sum_{t=0}^{Ck}(-1)^t\binom{|O\setminus S|}{t}$ and $\sum_{t=0}^{Ck+1}(-1)^t\binom{|O\setminus S|}{t}$ have different signs. Therefore we have that
$$\left|\sum_{t=0}^{Ck}(-1)^t\binom{|O\setminus S|}{t}\right| \le \binom{|O\setminus S|}{Ck+1} = \binom{|O|-|S|}{Ck+1}.$$

Recalling that $||f||_\infty$ is bounded, we are now ready to bound $\left|F\left(\overrightarrow{X}\right) - f(X)\right|$. Recall that if $|O| \le Ck$, this difference is 0, and otherwise we have that
$$\begin{aligned} \left|F\left(\overrightarrow{X}\right) - f(X)\right| &\le ||f||_\infty \cdot \sum_{\substack{S\subseteq O\\|S|\le Ck}}\binom{|O|-|S|}{Ck+1} \\ &= O\left(\sum_{s=0}^{Ck}\binom{|O|}{s}\binom{|O|-s}{Ck+1}\right) \\ &= O\left(\sum_{s=0}^{Ck}\binom{|O|}{Ck+s+1}\binom{Ck+s+1}{s}\right) \\ &= O\left(\sum_{s=0}^{Ck}2^{Ck+s+1}\binom{|O|}{Ck+s+1}\right). \end{aligned}$$

Therefore we can bound the error as

$$\begin{aligned}\left|\mathbf{E}\left[F\left(\overrightarrow{X}\right)\right]-\mathbf{E}[f(X)]\right| \\ = O\left(\sum_{s=0}^{Ck}2^{Ck+s+1}\mathbf{E}\left[\binom{|O|}{Ck+s+1}\right]\right).\end{aligned}\tag{2.4}$$

We note that

$$\binom{|O|}{Ck+s+1}=\sum_{\substack{I\subseteq[n]\\|I|=Ck+s+1}}\prod_{i\in I}U_i.$$

Hence by linearity of expectation, $(2Ck+1)$-wise independence, and the fact that $s\leq Ck$,

$$\begin{aligned}\mathbf{E}\left[\binom{|O|}{Ck+s+1}\right] &= \sum_{\substack{I\subseteq[n]\\|I|=Ck+s+1}}\mathbf{E}\left[\prod_{i\in I}U_i\right]\\ &= \sum_{\substack{I\subseteq[n]\\|I|=Ck+s+1}}\prod_{i\in I}O(|a_i|^p\lambda^{-p})\\ &= e^{O(Ck)}\sum_{\substack{I\subseteq[n]\\|I|=Ck+s+1}}\left(\prod_{i\in I}|a_i|^p\lambda^{-p}\right).\end{aligned}$$

We note when this sum is multiplied by $(Ck+s+1)!$, these terms all show up in the expansion of $\left(||a||_p^p\lambda^{-p}\right)^{Ck+s+1}$. In fact, for any integer $0\leq t\leq n$,

$$\sum_{\substack{I\subseteq[n]\\|I|=t}}\left(\prod_{i\in I}|a_i|^p\lambda^{-p}\right)\leq\frac{(||a||_p^p\lambda^{-p})^t}{t!}.\tag{2.5}$$

Hence, since $||a||_p=O(1)$,

$$\begin{aligned}\mathbf{E}\left[\binom{|O|}{Ck+s+1}\right] &= \frac{e^{O(Ck)}\lambda^{-p(Ck+s+1)}}{(Ck+s+1)!}\\ &= e^{O(Ck)}\left(\frac{\lambda^{-p}}{Ck}\right)^{(Ck+s+1)}.\end{aligned}$$

Therefore, by choice of λ and Eq. (2.4),

$$\begin{aligned}\left|\mathbf{E}\left[F\left(\overrightarrow{X}\right)\right]-\mathbf{E}[f(X)]\right| &= e^{O(Ck)}\sum_{s=0}^{Ck}\left(\frac{\lambda^{-p}}{Ck}\right)^{(Ck+s+1)}\\ &= C^{-O(Ck)}=O(\varepsilon).\end{aligned}\tag{2.6}$$

Hence it suffices to approximate $\mathbf{E}\left[F\left(\overrightarrow{X}\right)\right]$.

Recall

$$F\left(\overrightarrow{X}\right)=\sum_{\substack{S,T\subseteq[n]\\|S|,|T|\leq Ck\\S\cap T=\emptyset}}F_{S,T}\left(\overrightarrow{X}\right).$$

We will attempt to compute the conditional expectation of $F_{S,T}\left(\overrightarrow{X}\right)$, conditioned on the X_i for $i\in S\cup T$. Note the independence on the X_i's is sufficient that the values of the X_i for $i\in S\cup T$ are fully independent of one another, and that even having fixed these values, the remaining X_i are still Ck-wise independent.

We begin by making some definitions. Let $R=[n]\backslash(S\cup T)$. Having fixed S, T, and the values of X_i for $i\in S\cup T$, set $c=\sum_{i\in S}a_iX_i$ and $X'=\sum_{i\in R}a_iX_i'$. We note that $F_{S,T}\left(\overrightarrow{X}\right)=0$ unless $U_i=1$ for all $i\in S\cup T$, and otherwise that $F_{S,T}\left(\overrightarrow{X}\right)=f(c+X')$. This is because if $U_i=1$ for some $i\in T$, then $X_i'=0$. Let $p_c(x)$ be the Taylor series for $f(c+x)$ about 0, truncated so that its highest degree term is degree $Ck-1$. We approximate $\mathbf{E}[f(c+X')]$ by $\mathbf{E}[p_c(X')]$.

Lemma 2.4. $|\mathbf{E}[f(c+X')]-\mathbf{E}[p_c(X')]|<e^{-Ck}$.

Proof. By Taylor's theorem, the fact that C is even, and our given bounds on $||f^{(Ck)}||_\infty$,

$$|p_c(x)-f(c+x)|\leq\frac{|x|^{Ck}\alpha^{Ck}}{(Ck)!}=\frac{x^{Ck}\alpha^{Ck}}{(Ck)!}.$$

We note that $\mathbf{E}[p_c(X')]$ is determined simply by the independence properties of the X_i since it is a low-degree polynomial in functions of the X_i.

We now attempt to bound the error in approximating $f(c+x)$ by $p_c(x)$. In order to do so we will wish to bound $\mathbf{E}[(X')^{Ck}]$. Let $\ell=Ck$. We have that $\mathbf{E}[(X')^\ell]=\mathbf{E}\left[\left(\sum_{i\in R}a_iX_i'\right)^\ell\right]$. Expanding this out and using linearity of expectation yields a sum of terms of the form $\mathbf{E}\left[\prod_{i\in R}(a_iX_i')^{\ell_i}\right]$, for some non-negative integers ℓ_i summing to ℓ. Let L be the set of i so that $\ell_i>0$. Since $|L|\leq\ell$ which is at most the degree of independence, Eq. (2.1) implies that the above expectation is $\left(\prod_{i\in L}|a_i|^p\lambda^{-p}\right)\lambda^\ell e^{O(|L|)}$. Notice that the sum of the coefficients in front of such terms with a given L is at most $|L|^\ell$. This is because for each term in the product, we need to select an $i\in L$. Eq. (2.5) implies that summing $\prod_{i\in L}|a_i|^p\lambda^{-p}$ over all subsets L of size s, gives at most $\frac{(||a||_p^p\lambda^{-p})^s}{s!}$. Putting everything together:

$$\begin{aligned}\mathbf{E}\left[(X')^\ell\right] &\leq \sum_{s=1}^{\ell}\frac{s^\ell\lambda^{\ell-sp}e^{O(s)}}{s!}\\ &= \sum_{s=1}^{\ell}\exp(\ell\log s-s\log s\\ &\qquad -(\ell-sp)\log(1/\lambda)+O(s)).\end{aligned}\tag{2.7}$$

The summand (ignoring the $O(s)$) is maximized when

$$\frac{\ell}{s}+\log(1/\lambda^p)=\log(s)+1.$$

Rearranging terms and setting $u = \ell \cdot \lambda^p$, this happens when $\ell = s\log(\lambda^p \cdot s) + s$, which occurs for

$$s = \left(1 + O\left(\frac{\log\log(u)}{\log(u)}\right)\right) \cdot \frac{\ell}{\log(u)}.$$

Since the sum is at most ℓ times the biggest term,

$$\begin{aligned}\mathbf{E}\left[(X')^\ell\right] \le \exp\Bigg(\ell \cdot \Bigg(&\log(\ell) - \log\log(u) - \frac{\log(\ell)}{\log(u)} \\ &- \log(1/\lambda) + \frac{\log(1/\lambda^p)}{\log(u)} + O(1)\Bigg)\Bigg).\end{aligned}$$

Therefore we have that

$$\begin{aligned}&|\mathbf{E}[f(c+X')] - \mathbf{E}[p_c(X')]| \le \mathbf{E}\left[\frac{(X')^\ell \alpha^\ell}{\ell!}\right] \\ &\quad \le \exp\Bigg(\ell \cdot \Bigg(\log(\alpha) - \log\log(u) - \frac{\log(\ell)}{\log(u)} \\ &\qquad\qquad - \left(1 - \frac{p}{\log(u)}\right)\log(1/\lambda) + O(1)\Bigg)\Bigg) \\ &\quad = \exp\Bigg(\ell \cdot \Bigg(\log(\alpha) - \log\log(u) - \frac{\log(\ell)}{\log(u)} \\ &\qquad\qquad - \left(\frac{1}{p} - \frac{1}{\log(u)}\right)\log(\ell/u) + O(1)\Bigg)\Bigg) \\ &\quad = \exp\Big(\ell \cdot \Big(\log(\alpha) - \log\log(u) - \log(\ell^{1/p}) \\ &\qquad\qquad + \log(u^{1/p}) + O(1)\Big)\Big) \\ &\quad = \exp\left(-\ell \cdot \left(\log\left(\frac{1}{\alpha \cdot \lambda}\right) + \log\log(u) - O(1)\right)\right) \\ &\quad < e^{-\ell}\end{aligned}$$

with the last inequality holding for C (and hence u) a sufficiently large constant. ■

So to summarize:

$$\mathbf{E}[f(X)] = \mathbf{E}\left[F\left(\overrightarrow{X}\right)\right] + O(\varepsilon).$$

Now,

$$\begin{aligned}\mathbf{E}\left[F\left(\overrightarrow{X}\right)\right] &= \sum_{\substack{S,T\subseteq[n]\\ |S|,|T|\le Ck\\ S\cap T=\emptyset}} \mathbf{E}\left[F_{S,T}\left(\overrightarrow{X}\right)\right] \\ &= \sum_{\substack{S,T\subseteq[n]\\ |S|,|T|\le Ck\\ S\cap T=\emptyset}} (-1)^{|T|} \int_{\{x_i\}_{i\in S\cup T}} \Bigg(\left(\prod_{i\in S\cup T} U_i\right) \times \\ &\qquad\qquad \mathbf{E}[f(c+X')]\Bigg) dX_i(x_i) \\ &= \sum_{\substack{S,T\subseteq[n]\\ |S|,|T|\le Ck\\ S\cap T=\emptyset}} (-1)^{|T|} \int_{\{x_i\}_{i\in S\cup T}} \Bigg(\left(\prod_{i\in S\cup T} U_i\right) \times \\ &\qquad\qquad \left(\mathbf{E}[p_c(X')] \pm e^{-Ck}\right)\Bigg) dX_i(x_i).\end{aligned}$$

We recall that the term involving $\mathbf{E}[p_c(X')]$ is entirely determined by the $3Ck$-independence of the X_i's. We are left with an error of magnitude

$$\begin{aligned}& e^{-Ck} \cdot \left(\sum_{\substack{S,T\subseteq[n]\\ |S|,|T|\le Ck\\ S\cap T=\emptyset}} (-1)^{|T|} \int_{\{x_i\}_{i\in S\cup T}} \left(\prod_{i\in S\cup T} U_i\right) dX_i(x_i)\right) \\ &\le e^{-Ck} \cdot \left(\sum_{\substack{S,T\subseteq[n]\\ |S|,|T|\le Ck\\ S\cap T=\emptyset}} \mathbf{E}\left[\prod_{i\in S\cup T} U_i\right]\right) \\ &\le e^{-Ck} \cdot \left(\sum_{\substack{S,T\subseteq[n]\\ |S|,|T|\le Ck\\ S\cap T=\emptyset}} \left(\prod_{i\in S\cup T} |a_i|^p \lambda^{-p}\right) e^{O(|S|+|T|)}\right).\end{aligned}$$

Letting $s = |S| + |T|$, we change this into a sum over s. We use Eq. (2.5) to upper bound the product. We also note that given $S \cup T$, there are at most 2^s ways to pick S and T. Putting this together and recalling the choice of λ and that $||a||_p = O(1)$, the above is at most

$$\begin{aligned} e^{-Ck}\left(\sum_{s=0}^{2Ck} 2^s \left(\frac{||a||_p^{ps}\lambda^{-ps}}{s!}\right) e^{O(s)}\right) &= e^{-Ck}\left(\sum_{s=0}^{2Ck} \frac{O\left(\frac{1}{\lambda^p}\right)^s}{s!}\right) \\ &< e^{-Ck} \cdot e^{O\left(\frac{1}{\lambda^p}\right)} = O(\varepsilon). \end{aligned} \tag{2.8}$$

Hence $\mathbf{E}[f(X)]$ is determined up to $O(\varepsilon)$. ■

REMARK 2.1. *For constant α in the statement of Lemma 2.2, one can slightly optimize the proof to show that $k = \log(1/\varepsilon)/\log\log(1/\varepsilon)$ suffices. We describe here the necessary changes in the proof. First, we instead set $\lambda = (Ck)^{-1/10}$. In the first inequality where the value of λ is used, namely Eq. (2.6), the desired difference is then $(Ck)^{-O(Ck)} = O(\varepsilon)$. Next, in the proof of Lemma 2.4, the summand in Eq. (2.7) is maximized when $s = O(\ell/\log\ell)$, and straightforward calculations show that the desired difference in the statement of Lemma 2.4 is $O(\varepsilon^2)$. The left hand side of Eq. (2.8) is then at most $O(\varepsilon^2)\cdot e^{O(1/\lambda^p)}$, which is still $O(\varepsilon)$.*

Before we prove Theorem 2.1 we state the following lemma, whose proof is in Section A.3. Here $I_{[a,b]}$ denotes the indicator function of the interval $[a,b]$ ($|a|, |b|$ may be infinite). Our construction of the functions $J^c_{[a,b]}$ below proceeds by a process similar to mollification [17], where one smooths a function by convolving it with a narrow bump function of unit area. The only difference in our construction is that we convolve with the scaled Fourier transform of a bump function, which allows us to obtain better bounds on the high order derivatives of $J^c_{[a,b]}$. We call this slight variation "FT-mollification", to signify that we use the Fourier transform.

LEMMA 2.5. *There exist constants $c', \varepsilon_0 > 0$ such that for all $c > 0$ and $0 < \varepsilon < \varepsilon_0$, and for all $[a,b] \subseteq \mathbb{R}$, there exists a function $J^c_{[a,b]} : \mathbb{R} \to \mathbb{R}$ satisfying:*

i. $||(J^c_{[a,b]})^{(\ell)}||_\infty = O(c^\ell)$ *for all* $\ell \geq 0$.

ii. *For all x such that $a, b \notin [x-\varepsilon, x+\varepsilon]$, and as long as $c > c'\varepsilon^{-1}\log^3(1/\varepsilon)$, $|J^c_{[a,b]}(x) - I_{[a,b]}(x)| < \varepsilon$.*

We now prove Theorem 2.1. We defer analysis of the required precision (and hence the required space) to Section A.6, and here just argue correctness.

Proof (of Theorem 2.1). Consider first the following argument that Indyk's median estimator provides a $(1\pm\varepsilon)$-approximation when $r = \Theta(1/\varepsilon^2)$ and we use a sketch matrix B such that the $B_{i,j}$ are *fully* independent from $\mathcal{D}_p$. The following argument is only slightly different from Indyk's original argument, but is presented in such a way that adapts well to the entries of the sketch matrix having limited independence. Let $z = Bx$ be the sketch when using the fully independent matrix B. Since we scale our final output by median($|\mathcal{D}_p|$), we henceforth argue as if median($|\mathcal{D}_p|$) $= 1$ (remarks on computing median($|\mathcal{D}_p|$) are in Section A.2). The value 1 being the median is equivalent to the statement $\mathbf{E}[I_{[-1,1]}(z_i/||x||_p)] = 1/2$, since $\mathcal{D}_p$ is symmetric. Let μ_p be the probability density function of $\mathcal{D}_p$. Then by compactness, μ_p takes on some minimum value η_p in the interval $[-2,2]$. Furthermore, $\eta_p > 0$ (strict inequality) since $\mu_p > 0$ everywhere (this follows since it is known that μ_p is unimodal with mode zero [40], and is non-zero for large $|x|$ by Lemma 2.1). Then

$$\mathbf{E}\left[I_{[-1+\varepsilon,1-\varepsilon]}\left(\frac{z_i}{||x||_p}\right)\right] \leq \frac{1}{2} - \eta_p\varepsilon = \frac{1}{2} - \Theta(\varepsilon) \tag{2.9}$$

and

$$\mathbf{E}\left[I_{[-1-\varepsilon,1+\varepsilon]}\left(\frac{z_i}{||x||_p}\right)\right] \geq \frac{1}{2} + \eta_p\varepsilon = \frac{1}{2} + \Theta(\varepsilon). \tag{2.10}$$

Now if we let

$$Z = \frac{1}{r}\sum_{i=1}^{r} I_{[-1+\varepsilon,1-\varepsilon]}\left(\frac{z_i}{||x||_p}\right)$$

and

$$Z' = \frac{1}{r}\sum_{i=1}^{r} I_{[-1-\varepsilon,1+\varepsilon]}\left(\frac{z_i}{||x||_p}\right)$$

then $\mathbf{E}[Z] = 1/2 - \Theta(\varepsilon)$, $\mathbf{E}[Z'] = 1/2 + \Theta(\varepsilon)$, and $\mathbf{Var}[Z], \mathbf{Var}[Z'] \leq 1/r = \Theta(\varepsilon^2)$. Writing $r = c'/\varepsilon^2$, Chebyshev's inequality and a union bound imply $Z = 1/2 - \Theta(\varepsilon)$ and $Z' = 1/2 + \Theta(\varepsilon)$ with probability 7/8 for large c'. We conclude by noting median$\{|z_i|\}_{i=1}^r = (1\pm\varepsilon)||x||_p$ when both these events occur.

We now modify the above argument to handle our case where we use the sketch $y = Ax$ with the $A_{i,j}$ only k-wise independent for fixed i, and the seeds used to generate different rows of A being pairwise independent. Note that once we established bounds on $\mathbf{E}[Z]$ and $\mathbf{E}[Z']$ above, concentration of Z, Z' was shown via Chebyshev's inequality, which only required pairwise independence of the sketch matrix rows. Thus, we need only show that $\mathbf{E}[I_{[a,b]}(z_i/||x||_p)] \approx_\varepsilon \mathbf{E}[I_{[a,b]}(y_i/||x||_p)]$. Ideally we would just apply Lemma 2.2, but we cannot do this directly: the function $I_{[a,b]}$ does not have bounded high-order derivatives. We instead argue indirectly using the function $J^c_{[a,b]}$ from Lemma 2.5 with $c = O(\varepsilon^{-1}\log^3(1/\varepsilon))$. Let $Z_i = z_i/||x||_p$ and $Y_i = y_i/||x||_p$. We argue the following chain of inequalities:

$$\mathbf{E}[I_{[a,b]}(Z_i)] \approx_\varepsilon \mathbf{E}[J^c_{[a,b]}(Z_i)] \approx_\varepsilon \mathbf{E}[J^c_{[a,b]}(Y_i)] \approx_\varepsilon \mathbf{E}[I_{[a,b]}(Y_i)].$$

$\mathbf{E}[\mathbf{I_{[a,b]}}(\mathbf{Z_i})] \approx_\varepsilon \mathbf{E}[\mathbf{J^c_{[a,b]}}(\mathbf{Z_i})]$**:** This follows since $I_{[a,b]}$ and $J_{[a,b]}$ are within ε everywhere except for two intervals of length $O(\varepsilon)$. Since $\mathcal{D}_p$ is anticoncentrated (any length-$O(\varepsilon)$ interval contains $O(\varepsilon)$ probability mass) and $||I_{[a,b]}||_\infty, ||J^c_{[a,b]}||_\infty = O(1)$, these intervals contribute $O(\varepsilon)$ to the difference.

$\mathbf{E}[\mathbf{J^c_{[a,b]}}(\mathbf{Z_i})] \approx_\varepsilon \mathbf{E}[\mathbf{J^c_{[a,b]}}(\mathbf{Y_i})]$**:** Apply Lemma 2.2 with $\alpha = O(\varepsilon^{-1}\log^3(1/\varepsilon))$ and vector $x/||x||_p$.

```
1. Pick two random matrices $A \in \mathbb{R}^{r\times n}, A' \in \mathbb{R}^{r'\times n}$ as follows for $r = \Theta(1/\varepsilon^2)$ and $r' = \Theta(1)$. Each $A_{i,j}$ is distributing according to $\mathcal{D}_p$. For fixed $i$, the $A_{i,j}$ are $k$-wise independent with $k = \Theta(\log(1/\varepsilon)/\log\log(1/\varepsilon))$. For $i \neq i'$, the seeds used to generate the $\{A_{i,j}\}_{j=1}^n$ and $\{A_{i',j}\}_{j=1}^n$ are pairwise independent. $A'$ is generated similarly, but with fresh randomness, and with columns needing only be $k'$-wise independent for $k' = \Theta(1)$.
2. Maintain the vectors $y = Ax$ and $y' = A'x$ throughout the stream.
3. Let $y'_{\text{med}} = \text{median}\{|y'_i|\}_{i=1}^{r'}/\text{median}(|\mathcal{D}_p|)$. Output $y'_{\text{med}} \cdot \left(-\ln\left(\frac{1}{r}\sum_{i=1}^r \cos\left(\frac{y_i}{y'_{\text{med}}}\right)\right)\right)^{1/p}$.
```

Figure 2: Our log-cosine L_p estimation algorithm pseudocode, $0 < p < 2$, assuming infinite precision.

$\mathbf{E}[\mathbf{J}^{\mathbf{c}}_{[\mathbf{a},\mathbf{b}]}(\mathbf{Y_i})] \approx_\varepsilon \mathbf{E}[\mathbf{I}_{[\mathbf{a},\mathbf{b}]}(\mathbf{Y_i})]$: We argue as in the first inequality, but now we must show anticoncentration of the Y_i. Suppose for any $t \in \mathbb{R}$ we had a nonnegative function $f_{t,\varepsilon} : \mathbb{R} \to \mathbb{R}$ symmetric about t satisfying:

i. $||f^{(\ell)}_{t,\varepsilon}||_\infty = O(\alpha^\ell)$ for all $\ell \geq 0$, with $\alpha = O(1/\varepsilon)$

ii. $\mathbf{E}[f_{t,\varepsilon}(D)] = O(\varepsilon)$ for $D \sim \mathcal{D}_p$

iii. $f_{t,\varepsilon}(t+\varepsilon) = \Omega(1)$

iv. $f_{t,\varepsilon}(x)$ is strictly decreasing as $|x - t| \to \infty$

By (i), (ii) and Lemma 2.2 we would have $\mathbf{E}[f_{t,\varepsilon}(Y_i)] \approx_\varepsilon \mathbf{E}[f_{t,\varepsilon}(D)] = O(\varepsilon)$. Then, $\mathbf{E}[f_{t,\varepsilon}(Y_i)] \geq f_{t,\varepsilon}(t+\varepsilon) \cdot \mathbf{Pr}[Y_i \in [t-\varepsilon, t+\varepsilon]] = \Omega(\mathbf{Pr}[Y_i \in [t-\varepsilon, t+\varepsilon]])$ by (iii) and (iv), implying anticoncentration in $[t-\varepsilon, t+\varepsilon]$ as desired. We exhibit such functions $f_{t,\varepsilon}$ in Section A.4. ■

In Section 1.1.1 we stated that our techniques could be used to give an alternative proof to [12] that bounded independence fools regular halfspaces. This alternative proof is quite similar to part of the proof of Theorem 2.1, and we give full details in Section A.5.

Now that we have established a space-optimal derandomization of Indyk's algorithm, we give an alternative space-optimal algorithm, but with update time only $O(\varepsilon^{-2}\log(1/\varepsilon)/\log\log(1/\varepsilon))$ (Figure 2).

Lemma 2.6. *Let $B = \Theta(||x||_p)$. Let $A \in \mathbb{R}^{r\times n}$ and $y = Ax$ be as in Figure 2. Then with probability $7/8$ it holds that* $\left|\left(\frac{1}{r}\sum_{i=1}^r \cos\left(\frac{\sum_{j=1}^n A_{i,j}x_j}{B}\right)\right) - e^{-\left(\frac{||x||_p}{B}\right)^p}\right| < \varepsilon.$

Proof. Using that $\cos(x) = (e^{ix} + e^{-ix})/2$ we can calculate $\mathbf{E}[\cos(B'Z)]$ for fixed B' as $e^{-|B'|^p}$, using the Fourier transform of the density function of $\mathcal{D}_p$. Thus, letting $B' = \frac{||x||_p}{B}$ and applying Lemma 2.2 together with the observation of Remark 2.1 on the vector x/B, $\mathbf{E}[\cos(y_i/B)] = e^{-(||x||_p/B)^p} + O(\varepsilon)$. Also, since cos is bounded by 1, $\mathbf{Var}[\frac{1}{r}\sum_{i=1}^r \cos(y_i/B)] \leq 1/r$. The claim follows by Chebyshev's inequality. ■

Theorem 2.2. *For all $p \in (0,2)$, the algorithm of Figure 2 can be implemented with limited precision to use space $O(\varepsilon^{-2}\log(mM))$ and output $(1 \pm \varepsilon)||x||_p$ with probability at least $3/4$.*

Proof. As long as k', r' are chosen to be larger than some constant, y'_{med} is a constant factor approximation to $||x||_p$ by Theorem 2.1 with probability at least 7/8. Conditioned on this, consider $C = (\sum_i \cos(y_i/y'_{\text{med}}))/r$. Note that y'_{med} is independent of the y_i since A, A' were generated independently. Then by Lemma 2.6, with probability at least 7/8, C is within $O(\varepsilon)$ of $e^{-(||x||_p/y'_{\text{med}})^p}$ from which a $(1+O(\varepsilon))$-approximation of $||x||_p$ can be computed as $y'_{\text{med}} \cdot (-\ln(C))^{1/p}$. Note this expression is in fact a $(1+O(\varepsilon))$-approximation since the function $f(z) = e^{-|z|^p}$ is bounded both from above and below by constants for z in a constant-sized interval (in our case, z is $||x||_p/y'_{\text{med}}$), and thus $e^{-(||x||_p/y'_{\text{med}})^p} + O(\varepsilon) = (1+O(\varepsilon))e^{-(||x||_p/y'_{\text{med}})^p}$. Thus,

$$y'_{\text{med}} \cdot (-\ln(C))^{1/p} = ||x||_p + O(\varepsilon) \cdot y'_{\text{med}} = (1+O(\varepsilon))||x||_p.$$

Bounds on precision (and hence space) required are in Section A.6. ■

3 Lower Bound

In this section we prove our lower bound for $(1 \pm \varepsilon)$-multiplicative approximation of F_p for any positive real constant p bounded away from 0 in the turnstile model. We show a lower bound of $\Omega(\min\{N, \varepsilon^{-2}(\log(\varepsilon^2 mM))\})$. Note that if $\varepsilon \geq 1/N^{1/2-\delta}$ for any constant $\delta > 0$, the lower bound becomes $\Omega(\varepsilon^{-2}(\log(mM))$, matching our upper bound. We also describe a folklore lower bound of $\Omega(\log\log n)$ in Section A.1.2. Our lower bound holds for all ranges of the parameters ε, n, m, M varying independently.

Our proof in part uses that Augmented-Indexing requires linear communication in the one-way, one-round model [32] (an alternative proof was also given later in [4, Lemma 2]). We also use a known reduction

[25, 39] from Indexing to Gap-Hamdist. Henceforth all communication games discussed will be one-round and two-player, with the first player to speak named "Alice", and the second "Bob". We assume that Alice and Bob have access to public randomness.

Definition 3.1. *In the* Augmented-Indexing *problem, Alice receives a vector $x \in \{0,1\}^n$, Bob receives some $i \in [n]$ as well as all x_j for $j > i$, and Bob must output x_i. The problem* Indexing *is defined similarly, except Bob receives only $i \in [n]$, without receiving x_j for $j > i$.*

Definition 3.2. *In the* Gap-Hamdist *problem, Alice receives $x \in \{0,1\}^n$ and Bob receives $y \in \{0,1\}^n$. Bob is promised that either $\Delta(x,y) \le n/2 - \sqrt{n}$ (*NO *instance), or $\Delta(x,y) \ge n/2 + \sqrt{n}$ (*YES *instance) and must decide which holds. Here $\Delta(\cdot,\cdot)$ denotes Hamming distance.*

We also make use of the following two theorems.

Theorem 3.1. (Miltersen et al. [32, Lemma 13]) *The randomized one-round, one-way communication complexity of* Augmented-Indexing *with error probability at most $1/3$ is $\Omega(n)$.* ■

Theorem 3.2. ([25], [39, Section 4.3]) *There is a reduction from* Indexing *to* Gap-Hamdist *such that deciding* Gap-Hamdist *with probability at least $11/12$ implies a solution to* Indexing *with probability at least $2/3$. Furthermore, in this reduction the parameter n in* Indexing *is within a constant factor of that for the reduced* Gap-Hamdist *instance.* ■

Theorem 3.3. (Main lower bound) *For any real constant $p > 0$, any one-pass streaming algorithm for $(1 \pm \varepsilon)$-multiplicative approximation of F_p with probability at least $11/12$ in the turnstile model requires $\Omega(\min\{N, \varepsilon^{-2}(1 + p \cdot \log(\lceil \varepsilon^2 mM \rceil))\})$ bits of space.*

Proof. Let $q = 2^{1/p}$ and define $k = \lfloor 1/\varepsilon^2 \rfloor$ and $t = \min\{\lfloor N/k \rfloor, \lfloor \log_q(M) \rfloor + 1\}$. Given an algorithm A providing a $(1 \pm d\varepsilon/2^p)$-multiplicative approximation of F_p with probability at least $11/12$, where $d > 0$ is some small constant to be fixed later, we devise a protocol to decide Augmented-Indexing on strings of length kt, where the number of bits communicated in the protocol is dominated by the space complexity of A. Since $kt = \Omega(\min\{N, \varepsilon^{-2}(1+p \cdot \log(\lceil \varepsilon^2 mM \rceil))\})$ and $2^p = O(1)$ for constant p, the lower bound (ignoring the $\varepsilon^2 m$ term) follows. At the end of the proof, we describe how to obtain our stated dependence on m in the lower bound as well.

Let Alice receive $x \in \{0,1\}^{kt}$, and Bob receive $z \in [kt]$. Alice conceptually divides x into t contiguous blocks where the ith block b_i is of size k. Bob's index z lies in some $b_{i(z)}$, and Bob receives bits x_j that lie in a block b_i with $i > i(z)$. Alice applies the Gap-Hamdist reduction of Theorem 3.2 to each b_i separately to obtain new vectors y_i each of length at most $c \cdot k$ for some constant c for all $0 \le i < t$. Alice then creates a stream from the set of y_i by, for each i and each bit $(y_i)_j$ of y_i, inserting an update $((i,j), \lfloor q^i \rfloor)$ into the stream if $(y_i)_j = 1$. Alice processes this stream with A then sends the state of A to Bob along with the Hamming weight $w(y_i)$ of y_i for all i. Note the size of the universe in the stream is at most $ckt = O(N) = O(n)$.

Now, since Bob knows the bits in b_i for $i > i(z)$ and shares randomness with Alice, he can run the same Gap-Hamdist reduction as Alice to obtain the y_i for $i > i(z)$ then delete all the insertions Alice made for these y_i. Bob then performs his part of the reduction from Indexing on strings of length k to Gap-Hamdist within the block $b_{i(z)}$ to obtain a vector $y(B)$ of length ck such that deciding whether $\Delta(y(B), y_{i(z)}) > ck/2 + \sqrt{ck}$ or $\Delta(y(B), y_{i(z)}) < ck/2 - \sqrt{ck}$ with probability at least $11/12$ allows one to decide the Indexing instance on block $b_{i(z)}$ with probability at least $2/3$. Here $\Delta(\cdot,\cdot)$ denotes Hamming distance. For each j such that $y(B)_j = 1$, Bob inserts $((i(z),j), -\lfloor q^{i(z)} \rfloor)$ into the stream being processed by A. We have so far described all stream updates, and thus the number of updates is at most $2ckt = O(N) = O(m)$. Note the pth moment L'' of the underlying vector of the stream now exactly satisfies

$$L'' = \left\lfloor q^{i(z)} \right\rfloor^p \cdot \Delta(y(B), y_{i(z)}) + \sum_{i<i(z)} w(y_i) \left\lfloor q^i \right\rfloor^p . \tag{3.11}$$

Setting $\eta = \sum_{i<i(z)} w(y_i) \lfloor q^i \rfloor^p$ and rearranging terms, $\Delta(y(B), y_{i(z)}) = (L'' - \eta)/\lfloor q^{i(z)} \rfloor^p$. Recall that in this Gap-Hamdist instance, Bob must decide whether $\Delta(y(B), y_{i(z)}) < ck/2 - \sqrt{ck}$ or $\Delta(y(B), y_{i(z)}) > ck/2 + \sqrt{ck}$. Bob can compute η exactly given Alice's message. To decide Gap-Hamdist it thus suffices to obtain a $(\sqrt{ck}/4)$-additive approximation to $\lfloor q^{i(z)} \rfloor^{-p} L''$. Note $\lfloor q^i \rfloor^p = 2^i \cdot (\lfloor q^i \rfloor / q^i)^p$, and thus $2^i/2^p \le \lfloor q^i \rfloor^p \le 2^i$. Thus, $\lfloor q^{i(z)} \rfloor^{-p} L''$ is upper-bounded by

$$\left\lfloor q^{i(z)} \right\rfloor^{-p} \sum_{i=0}^{i(z)} \left\lfloor q^i \right\rfloor^p \cdot ck \le 2^{p-i(z)}(2^{i(z)+1} - 1)ck \le 2^{p+1}ck,$$

implying our desired additive approximation is guaranteed by obtaining a $(1 \pm \varepsilon')$-multiplicative approximation to L'' for $\varepsilon' = (\sqrt{ck}/4)/(4 \cdot 2^{p+1}ck) = 1/(2^{p+3}\sqrt{ck})$. By choice of k, this is a $(1 \pm O(\varepsilon/2^p))$-multiplicative approximation, which we can obtain from A by setting d

to be a sufficiently small constant. Recalling that A provides this $(1 \pm O(\varepsilon/2^p))$-approximation with probability at least 11/12, we solve Gap-Hamdist in the block $i(z)$ with probability at least 11/12, and thus Indexing in block $i(z)$ with probability at least 2/3 by Theorem 3.2, which is equivalent to solving the original Augmented-Indexing instance. This implies that the total number of bits communicated in this protocol must be $\Omega(kt)$. Now note that the only bits communicated other than the state of A are the transmissions of $w(y_i)$ for $0 \leq i < t$. Since $w(y_i) \leq k$, all Hamming weights can be communicated in $O(t \log(k)) = o(kt)$ bits. Thus, indeed, the communication of the protocol is dominated by the space complexity of A, implying A uses space $\Omega(kt)$.

The above argument yields the lower bound $\Omega(\min\{N, \varepsilon^2 \log(M))$. We can similarly obtain the lower bound $\Omega(\min\{N, \varepsilon^2 \log(\lceil \varepsilon^2 m \rceil))$ by, rather than updating an item in the stream by $f_i = \lfloor q^i \rfloor$ in one update, we update the same item f_i times by 1. The number of total updates in the ith block is then at most $\lfloor q^i \rfloor \cdot k$, and thus the maximum number of blocks of length k we can give Alice to ensure that both the stream length and number of used universe elements is at most N is $t = \min\{\lfloor N/k \rfloor, O(\log(\lceil m/k \rceil))\}$. The proof is otherwise identical. ■

Our lower bound technique also improves lower bounds for $p = 0$, L_p estimation in the strict turnstile model (where we are promised $x_i \geq 0$ always), and additive estimation of entropy. Details are in Section A.7.

Acknowledgments

We thank Venkat Chandar for pointing out prior work on mollification and its similarity with Section A.3. We thank Steven G. Johnson for answering questions about [27]. We thank Ilias Diakonikolas, Piotr Indyk, T.S. Jayram, Swastik Kopparty, John Nolan, and Krzysztof Oleszkiewicz for valuable discussions.

References

[1] Noga Alon, Phillip B. Gibbons, Yossi Matias, and Mario Szegedy. Tracking join and self-join sizes in limited storage. *J. Comput. Syst. Sci.*, 64(3):719–747, 2002.

[2] Noga Alon, Yossi Matias, and Mario Szegedy. The Space Complexity of Approximating the Frequency Moments. *J. Comput. Syst. Sci.*, 58(1):137–147, 1999.

[3] Khanh Do Ba, Piotr Indyk, Eric C. Price, and David P. Woodruff. Lower bounds for sparse recovery. In *Proceedings of the 21st Annual ACM-SIAM Symposium on Discrete Algorithms (SODA), to appear*, 2010.

[4] Ziv Bar-Yossef, T. S. Jayram, Robert Krauthgamer, and Ravi Kumar. The sketching complexity of pattern matching. In *8th International Workshop on Randomization and Computation (RANDOM)*, pages 261–272, 2004.

[5] Ziv Bar-Yossef, T. S. Jayram, Ravi Kumar, and D. Sivakumar. An information statistics approach to data stream and communication complexity. In *Proceedings of the 43rd Symposium on Foundations of Computer Science (FOCS)*, pages 209–218, 2002.

[6] Joshua Brody and Amit Chakrabarti. A multi-round communication lower bound for gap hamming and some consequences. In *Proceedings of the 24th Annual IEEE Conference on Computational Complexity (CCC)*, pages 358–368, 2009.

[7] Amit Chakrabarti, Graham Cormode, and Andrew McGregor. A near-optimal algorithm for computing the entropy of a stream. In *Proceedings of the 18th Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*, pages 328–335, 2007.

[8] Amit Chakrabarti, Subhash Khot, and Xiaodong Sun. Near-optimal lower bounds on the multi-party communication complexity of set disjointness. In *Proceedings of the 18th Annual IEEE Conference on Computational Complexity (CCC)*, pages 107–117, 2003.

[9] John M. Chambers, Colin L. Mallows, and B. W. Stuck. A method for simulating stable random variables. *J. Amer. Statist. Assoc.*, 71:340–344, 1976.

[10] Kenneth L. Clarkson and David Woodruff. Numerical linear algebra in the streaming model. In *Proceedings of the 41st Annual ACM Symposium on Theory of Computing (STOC)*, pages 205–214, 2009.

[11] Graham Cormode, Piotr Indyk, Nick Koudas, and S. Muthukrishnan. Fast mining of massive tabular data via approximate distance computations. In *Proceedings of the 18th International Conference on Data Engineering (ICDE)*, pages 605–, 2002.

[12] Ilias Diakonikolas, Parikshit Gopalan, Ragesh Jaiswal, Rocco Servedio, and Emanuele Viola. Bounded independence fools halfspaces. In *Proceedings of the 50th Annual IEEE Symposium on Foundations of Computer Science (FOCS), to appear*, 2009.

[13] Joan Feigenbaum, Sampath Kannan, Martin Strauss, and Mahesh Viswanathan. An approximate L1-difference algorithm for massive data streams. *SIAM J. Comput.*, 32(1):131–151, 2002.

[14] Dan Feldman, Morteza Monemizadeh, Christian Sohler, and David P. Woodruff. Coresets and sketches for high dimensional subspace problems. In *Proceedings of the 21st Annual ACM-SIAM Symposium on Discrete Algorithms (SODA), to appear*, 2010.

[15] Philippe Flajolet and G. Nigel Martin. Probabilistic counting. In *Proceedings of the 24th Annual IEEE Symposium on Foundations of Computer Science (FOCS)*, pages 76–82, 1983.

[16] Michael L. Fredman, János Komlós, and Endre Szemerédi. Storing a sparse table with O(1) worst case access time. *J. ACM*, 31(3):538–544, 1984.

[17] Kurt Otto Friedrichs. The identity of weak and strong extensions of differential operators. *Transactions of the American Mathematical Society*, 55(1):132–151, 1944.

[18] Sumit Ganguly and Graham Cormode. On estimating frequency moments of data streams. In *Proceedings of the 11th International Workshop on Randomization and Computation (RANDOM)*, pages 479–493, 2007.

[19] Anna C. Gilbert, Sudipto Guha, Piotr Indyk, Yannis Kotidis, S. Muthukrishnan, and Martin Strauss. Fast, small-space algorithms for approximate histogram maintenance. In *Proceedings on 34th Annual ACM Symposium on Theory of Computing (STOC)*, pages 389–398, 2002.

[20] Uffe Haagerup. The best constants in the Khintchine inequality. *Studia Math.*, 70(3):231–283, 1982.

[21] Nicholas J. A. Harvey, Jelani Nelson, and Krzysztof Onak. Sketching and streaming entropy via approximation theory. In *Proceedings of the 49th Annual IEEE Symposium on Foundations of Computer Science (FOCS)*, pages 489–498, 2008.

[22] Piotr Indyk. Algorithms for dynamic geometric problems over data streams. In *Proceedings of the 36th Annual ACM Symposium on Theory of Computing (STOC)*, pages 373–380, 2004.

[23] Piotr Indyk. Stable distributions, pseudorandom generators, embeddings, and data stream computation. *J. ACM*, 53(3):307–323, 2006.

[24] Piotr Indyk. Sketching, streaming and sublinear-space algorithms, 2007. Graduate course notes available at `http://stellar.mit.edu/S/course/6/fa07/6.895/`.

[25] T. S. Jayram, Ravi Kumar, and D. Sivakumar. The one-way communication complexity of gap hamming distance. *Theory of Computing*, 4(1):129–135, 2008.

[26] T.S. Jayram and David P. Woodruff. The data stream space complexity of cascaded norms. In *Proceedings of the 50th Annual IEEE Symposium on Foundations of Computer Science (FOCS), to appear*, 2009.

[27] Steven G. Johnson. Saddle-point integration of C_∞ "bump" functions. Manuscript. Available at `http://math.mit.edu/~stevenj/bump-saddle.pdf`.

[28] Balachander Krishnamurthy, Subhabrata Sen, Yin Zhang, and Yan Chen. Sketch-based change detection: methods, evaluation, and applications. In *Proceedings of the 3rd ACM SIGCOMM Conference on Internet Measurement (IMC)*, pages 234–247, 2003.

[29] Eyal Kushilevitz and Noam Nisan. *Communication Complexity*. Cambridge University Press, 1997.

[30] Ping Li. Estimators and tail bounds for dimension reduction in l_p $(0 < p \leq 2)$ using stable random projections. In *Proceedings of the 19th Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*, pages 10–19, 2008.

[31] Ping Li. Compressed counting. In *Proceedings of the 20th Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*, pages 412–421, 2009.

[32] Peter Bro Miltersen, Noam Nisan, Shmuel Safra, and Avi Wigderson. On data structures and asymmetric communication complexity. *J. Comput. Syst. Sci.*, 57(1):37–49, 1998.

[33] Morteza Monemizadeh and David P. Woodruff. Single pass relative-error l_p sampling with applications. In *Proceedings of the 21st Annual ACM-SIAM Symposium on Discrete Algorithms (SODA), to appear*, 2010.

[34] J. Ian Munro and Mike Paterson. Selection and Sorting with Limited Storage. *Theoretical Computer Science*, 12(3):315–323, 1980.

[35] S. Muthukrishnan. Data Streams: Algorithms and Applications. *Foundations and Trends in Theoretical Computer Science*, 1(2):117–236, 2005.

[36] Noam Nisan. Pseudorandom generators for space-bounded computation. *Combinatorica*, 12(4):449–461, 1992.

[37] John P. Nolan. *Stable Distributions - Models for Heavy Tailed Data*. Birkhäuser, 2010. In progress, Chapter 1 online at `http://academic2.american.edu/∼jpnolan`.

[38] David P. Woodruff. Optimal space lower bounds for all frequency moments. In *Proceedings of the 15th Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*, pages 167–175, 2004.

[39] David P. Woodruff. *Efficient and Private Distance Approximation in the Communication and Streaming Models*. PhD thesis, Massachusetts Institute of Technology, 2007.

[40] Makoto Yamazato. Unimodality of infinitely divisible distribution functions of class L. *Ann. Probability*, 6:523–531, 1978.

[41] Vladimir Mikhailovich Zolotarev. *One-dimensional Stable Distributions*. Vol. 65 of Translations of Mathematical Monographs, American Mathematical Society, 1986.

A Appendix

A.1 The additive $\log\log n$ in L_p bounds

Throughout the paper, we omitted an additive $\log\log n$ both in the statement of upper and lower bounds for L_p estimation. We here discuss why this term arises.

A.1.1 The upper bounds

In Section 1, we stated that stream updates (i, v) come from $[n] \times \{-M, \ldots, M\}$, causing the change $x_i \leftarrow x_i + v$. We note that in fact, we can assume that i comes not from the universe $[n]$, but rather from $[U]$ for $U = O(m^2)$, at the cost of an additive $O(\log\log n)$ in the space complexity of any L_p estimation algorithm. We discuss why this is the case below, using an idea of [16].

Let $\{i_1, \ldots, i_r\}$ be the indices appearing in the stream. Picking a prime q and treating all updates (i, v) as $(i \mod q, v)$, our L_p estimate is unaffected as long as $i_{j_1} \neq i_{j_2} \mod q$ for all $j_1 \neq j_2$. There are at most $r^2/2$ differences $|i_{j_1} - i_{j_2}|$, and each difference is an integer bounded by n, thus having at most $\log n$ prime factors. There are thus at most $r^2 \log n$ prime factors dividing *some* $|i_{j_1} - i_{j_2}|$. If we pick a random prime $q = \tilde{O}(r^2 \log n)$, we can ensure with constant probability arbitrarily close to 1 (by increasing the constant in the "big-Oh") that no indices collide modulo q. Since $r \leq m$, we thus have $q = \tilde{O}(m^2 \log n)$. We then pick a hash function $h : \{0, \ldots, q-1\} \rightarrow [O(m^2)]$ at random from pairwise independent family. With constant probability which can be made arbitrarily high (again by increasing the constant in the "big-Oh"), the mapping $i \mapsto h(i \mod q)$ perfectly hashes the indices appearing in the stream. Storing both h and q requires $O(\log q + \log m) = O(\log m + \log\log n)$ bits.

A.1.2 The lower bounds

An $\Omega(\log\log n)$ lower bound for moment estimation in the turnstile model follows by a simple reduction from the communication problem of Equality in the private coin model. The lower bound holds for streams of length at least 2 and arbitrary M. In Equality, Alice receives $x \in [n]$, and Bob receives $y \in [n]$, and they must decide whether $x = y$ with probability $2/3$. This problem requires $\Omega(\log\log n)$ communication [29]. Given a streaming algorithm A for turnstile F_p estimation for $p \geq 0$ with success probability $2/3$, Alice performs the update $(x, +1)$, sends A's state to Bob, and Bob performs the update $(y, -1)$. Either $x = y$, implying $F_p = 0$, else $F_p \neq 0$. Thus, any multiplicative approximation of F_p gives a solution to Equality, implying A uses $\Omega(\log\log n)$ space.

A.2 Approximating the median of $|\mathcal{D}_p|$

Indyk's algorithm (Figure 1) and our log-cosine algorithm (Figure 2) both require knowledge of median$(|\mathcal{D}_p|)$. It is known that median$(|\mathcal{D}_1|) = 1$, but a closed-form expression for the median is not known for general p.

In fact, we note that a $(1 \pm \varepsilon)$-approximation of this quantity suffices. One method to obtain such a quantity is the following. Let μ_p be the density function of $\mathcal{D}_p$, and let $\bar{\mu}_p$ be the cdf. Let x_p^- be such that $5/8 < \bar{\mu}_p(x_p^-) < 3/4$ and x_p^+ be such that $3/4 < \bar{\mu}_p(x_p^+) < 7/8$; such values can be found during preprocessing numerically with constant precision. Let $\tilde{\delta}_p$ be a value in $[\mu_p(x_p^+)/2, \mu_p(x_p^+)]$, which can also be computed numerically with constant precision. Let median$(|\mathcal{D}_p|) = x_{\mathrm{med,p}}$ so that $x_p^- < x_{\mathrm{med,p}} < x_p^+$. Then we know $\mu_p(x_{\mathrm{med,p}}) > \tilde{\delta}_p$ by unimodality of $\mathcal{D}_p$ with mode zero [40]. Thus, in preprocessing, by selecting $C/((\varepsilon\tilde{\delta}_p x_p^-)^2)$ samples from $\mathcal{D}_p$ and taking the median of the absolute value of the samples, the result is guaranteed to be $x_{\mathrm{med,p}} \pm \varepsilon x_p^- = (1 \pm \varepsilon) x_{\mathrm{med,p}}$ with arbitrarily large constant probability by increasing C, by a Chernoff bound.

A.3 Approximating the indicator function of an interval

Throughout this section we let $b : \mathbb{R} \rightarrow \mathbb{R}$ be the function

$$b(x) = \begin{cases} e^{-\frac{x^2}{1-x^2}} & \text{for } |x| < 1 \\ 0 & \text{otherwise} \end{cases},$$

which is known to be smooth.

We let

$$\hat{b}(t) = \frac{1}{2\pi} \cdot \int_{-\infty}^{\infty} b(x) e^{-itx} dx = \frac{1}{2\pi} \cdot \int_{-1}^{1} b(x) e^{-itx} dx$$

denote the Fourier transform of b.

We now prove a few properties of b and $\hat{b}$. The following is standard and our desired properties may have appeared in the literature before, but we include full proofs since we do not know an appropriate reference.

Lemma A.1. *The following upper bounds hold:*

i. $||b^{(\ell)}||_\infty < e \cdot 32^\ell \ell! \cdot \max\{\ell^{2\ell+2}, 1\}$ *for all* $\ell \geq 0$

ii. $||b^{(\ell)}||_1 \leq 2||b^{(\ell)}||_\infty$ *for all* $\ell \geq 0$

Proof. Item (ii) follows since support$(b^{(\ell)}) = (-1, 1)$.

We now show (i). We need only consider $x \in (-1, 1)$ since $b^{(\ell)}(x) = 0$ for $|x| \geq 1$. Define $A_1(x) = e^{-1/(2-2x)}$ and $A_2(x) = e^{-1/(2+2x)}$ so that $b(x) = e \cdot A_1(x) A_2(x)$. Then $b^{(\ell)}(x)$ is a sum of terms of the form $c_j A_1^{(j)}(x) A_2^{(\ell-j)}(x)$, where the sum of the absolute values of the c_j is $e \cdot 2^\ell$. Also, define $B_1(x) = 1/(2-2x)$ and $B_2(x) = 1/(2+2x)$. Then $A_i'(x) = (-1)^i 2A_i(x) B_i^2(x)$ and $B_i'(x) = (-1)^{i+1} 2B_i^2(x)$ for $i \in$

$\{1, 2\}$. We claim for $k \geq 1$ that

$$A_i^{(k)}(x) = \sum_{j=k+1}^{2k} d_j A_i(x) B_i^j(x)$$

with $|d_j| < 4^k k!$. This holds for $k = 1$. We then have for $k \geq 1$ that

$$\begin{aligned}
A_i^{(k+1)}(x) = & \sum_{j=k+1}^{2k} 2d_j \Big((-1)^i A_i(x) B_i^{j+2}(x) \\
& \qquad + (-1)^{i+1} j A_i(x) B_i^{j+1}(x) \Big) \\
= & \sum_{j=k+2}^{2k+1} 2d_{j-1} \Big((-1)^i A_i(x) B_i^{j+1}(x) \\
& \qquad + (-1)^{i+1} (j-1) A_i(x) B_i^{j}(x) \Big) \\
= & (-1)^i \Bigg(2(k+1) d_{k+1} A_i(x) B_i^{k+2}(x) \\
& \qquad + \sum_{j=k+3}^{2k+2} 2A_i(x) B_i^j(x) (d_{j-2} - (j-1)d_{j-1}) \Bigg)
\end{aligned}$$

The claim then holds since

$$|2(k+1)d_{k+1}| < 2 \cdot 4^k (k+1)! < 4^{k+1}(k+1)!$$

and

$$\begin{aligned}
|2(d_{j-2} - (j-1)d_{j-1})| &< |2(4^k k! + (2k+1)4^k k!)| \\
&\leq 2(2k+2)4^k k! = 4^{k+1}(k+1)!.
\end{aligned}$$

By definition of the A_i and B_i, for $x \in (-1, 1)$:

$$|A_i(x) B_i^j(x)| \leq \sup_{y>0} \; y^j e^{-y} < j^j.$$

The final inequality follows since j is the sole positive root of the derivative of $y^j e^{-y}$. Thus for $k \geq 1$,

$$\sup_{x \in (-1,1)} \; |A_i^{(k)}(x)| < 4^k k! (2k)^{2k} k = k! (4k)^{2k} k.$$

Thus for all $1 \leq j \leq \ell - 1$,

$$\begin{aligned}
|A_1^{(j)}(x) A_2^{(\ell-j)}(x)| &< j!(4j)^{2j} j \cdot (\ell - j)! \\
&\quad \times (4(\ell-j))^{2(\ell-j)} (\ell - j) \\
&\leq \ell! (4\ell)^{2\ell} \ell^2.
\end{aligned}$$

The above inequality is also true for $j \in \{0, \ell\}$ since $\sup_{x \in (-1,1)} |A_i(x)| = e^{-1/4} < 1$, and thus

$$||b^{(\ell)}||_\infty < (e \cdot 2^\ell) \cdot \ell! (4\ell)^{2\ell} \ell^2 = e \cdot 32^\ell \ell! \cdot \ell^{2\ell+2}$$

for $\ell \geq 1$. The inequality holds for $\ell = 0$ by inspection since $||b||_\infty = 1$. ■

Lemma A.2. *For arbitrary $\ell, n \geq 0$ and $t \in \mathbb{R}$, $t \neq 0$,*

$$|\hat{b}^{(\ell)}(t)| < (|t|/64)^{-n} (n+1)^{2n+3} \cdot n! \cdot \frac{\ell!}{([[\ell - n]])!} \cdot \frac{1}{[[\ell - n]] + 1}$$

where $[[x]]$ *denotes* $\max\{0, x\}$.

Proof. We can write

$$\hat{b}^{(\ell)}(t) = \frac{i^{\ell-n} t^{-n}}{2\pi} \cdot \int_{-\infty}^{\infty} \left(\frac{\partial^n}{\partial y^n} y^\ell b(y) \right)(x) e^{-ixt} dx.$$

For $n \leq \ell$, a straightforward induction shows

$$\left(\frac{\partial^n}{\partial y^n} y^\ell b(y) \right)(x) = \sum_{k=0}^{n} c_{k,n,\ell} x^{\ell - n + k} b^{(k)}(x) \tag{A.1}$$

with

$$c_{k,n,\ell} = \frac{\ell!}{(\ell - n + k)!} \binom{n}{k}.$$

In the case $n > \ell$, the summation of Eq. (A.1) begins at $k = n - \ell$, and each $c_{k,n,\ell}$ is upper bounded by the above. Since $b^{(k)}$ is supported on $[-1, 1]$ for all $k \geq 0$,

$$\begin{aligned}
\left| \hat{b}^{(\ell)}(t) \right| &= \frac{|t|^{-n}}{2\pi} \cdot \left| \int_{-1}^{1} \left(\frac{\partial^n}{\partial y^n} y^\ell b(y) \right)(x) dx \right| \\
\leq & \quad \frac{|t|^{-n}}{2\pi} \cdot \sum_{k=[[n-\ell]]}^{n} \left(\frac{\ell!}{(\ell - n + k)!} \binom{n}{k} ||b^{(k)}||_\infty \right. \\
& \qquad \left. \times \int_{-1}^{1} |x|^{\ell - n + k} dx \right) \\
< & \quad |t|^{-n} \sum_{k=[[n-\ell]]}^{n} \left(\frac{\ell!}{(\ell - n + k)!} \binom{n}{k} 32^k k! \right. \\
& \qquad \left. \times \max\{k^{2k+2}, 1\} \cdot \frac{1}{\ell - n + k + 1} \right) \\
\leq & \quad (|t|/64)^{-n} (n+1)^{2n+3} \cdot n! \\
& \qquad \times \frac{\ell!}{([[\ell - n]])!} \cdot \frac{1}{[[\ell - n]] + 1}
\end{aligned}$$

■

Lemma A.3. *For all $\ell \geq 0$, $||\hat{b}^{(\ell)}||_1 = O(1)$.*

Proof. $n = 0$ in Lemma A.2 gives $||\hat{b}^{(\ell)}||_\infty < 1/(\ell + 1)$, while $n = 2$ gives $|b^{(\ell)}(t)| = O((\ell + 1)/t^2)$. Thus,

$$\begin{aligned}
||\hat{b}^{(\ell)}||_1 &= \int_{-\infty}^{\infty} |\hat{b}^{(\ell)}(t)| dt \\
&= \int_{-\ell+1}^{\ell+1} |\hat{b}^{(\ell)}(t)| dt + 2 \int_{\ell+1}^{\infty} |\hat{b}^{(\ell)}(t)| dt \\
&< 2 + 2 \int_{\ell+1}^{\infty} O((\ell + 1)/t^2) \\
&= O(1)
\end{aligned}$$

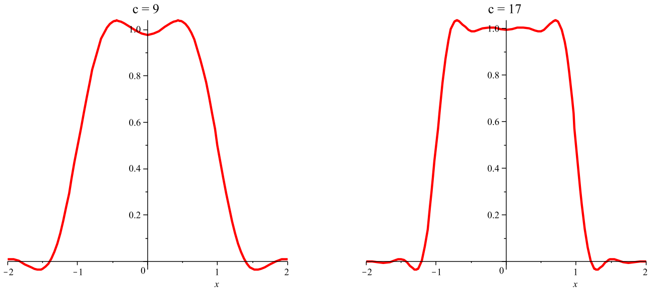

Figure 3: Approximate indicator functions of the interval $[-1,1]$; $J^c_{[-1,1]}$ for $c = 9, 17$.

■

We are now ready to prove the main lemma of this section. We construct good smooth approximations of indicator functions with small high-order derivatives, using FT-mollification (see Figure 3).

Proof (of Lemma 2.5). For a function f let f_c denote the function $f_c(x) = c \cdot f(cx)$. Define

$$J^c_{[a,b]}(x) = (\hat{b}_c * I_{[a,b]})(x) = \int_{c\cdot(x-b)}^{c\cdot(x-a)} \hat{b}(t)dt.$$

To show (i),

$$\begin{aligned}|J^{(\ell)}_{[a,b]}(x)| = |(\hat{b}^{(\ell)}_c * I_{[a,b]})(x)| = c^\ell \cdot |((\hat{b}^{(\ell)})_c * I_{[a,b]})(x)| \\ \leq c^\ell \cdot ||\hat{b}^{(\ell)}||_1 = O(c^\ell)\end{aligned}$$

with the last equality holding by Lemma A.3.

Now we show (ii). Suppose $x \notin [a - \varepsilon, b + \varepsilon]$. Then

$$|J^c_{[a,b]}(x)| \leq \int_{c\varepsilon}^{\infty} |\hat{b}(t)|dt.$$

We may assume without loss of generality that $1/\varepsilon$ is a power of 2 so that $\log(1/\varepsilon)$ is integral. Then applying Lemma A.2 with $\ell = 0$ and $n = \log(1/\varepsilon)$, and assuming $c > 1024n^3/\varepsilon$,

$$\begin{aligned}\int_{c\varepsilon}^{\infty} |\hat{b}(t)|dt &\leq 64^n \cdot (n+1)^{2n+3} \cdot n! \cdot \int_{c\varepsilon}^{\infty} t^{-n}dt \\ &\leq 64^n \cdot (2n)^{3n+3} \cdot (c\varepsilon)^{-n+1} \qquad \text{(A.2)} \\ &= 8 \cdot n^3 \cdot c\varepsilon \cdot (512n^3/(c\varepsilon))^n \\ &= O(n^6/2^n) \\ &< \varepsilon/2\end{aligned}$$

Now consider the case $x \in [a + \varepsilon, b - \varepsilon]$. Note

$$\int_{-\infty}^{\infty} \hat{b}(t)dt = b(0) = 1$$

We then have

$$\begin{aligned}J^c_{[a,b]}(x) &= \int_{c\cdot(x-b)}^{c\cdot(x-a)} \hat{b}(t)dt \\ &\geq \int_{-\infty}^{\infty} \hat{b}(t)dt - \left| \int_{-\infty}^{c\cdot(x-b)} \hat{b}(t)dt \right. \\ &\qquad \left. - \int_{c\cdot(x-a)}^{\infty} \hat{b}(t)dt \right| \\ &\geq 1 - 2\int_{c\varepsilon}^{\infty} |\hat{b}(t)|dt \\ &> 1 - \varepsilon\end{aligned}$$

with the final inequality holding by Eq. (A.2). It can similarly be shown that $J^c_{[a,b]}(x) < 1 + \varepsilon$ by flipping the sign of the second summand, and reversing the direction of inequality, in the second inequality. ■

Remark A.1. *The proof of Lemma 2.5 required $c = O(\varepsilon^{-1}\log^3(1/\varepsilon))$ to satisfy (ii) in the statement of the lemma, with the constant in the big-Oh being rather large. The proof reveals that the rate of decay of $\hat{b}$ is exactly what determined the factor multiplying $1/\varepsilon$ in c: here $O(\log^3(1/\varepsilon))$. Though in fact, for our same exact $\hat{b}$, Johnson [27] showed that $|\hat{b}(t)| = O(|t|^{-3/4}e^{-\sqrt{|t|}})$, with the constant in the big-Oh being fairly small. Plugging Johnson's bound into our proof then shows that $c = O(\varepsilon^{-1}\log^2(1/\varepsilon))$ actually suffices. In fact, [27] also implies that for all $\delta > 0$, if one considers the bump function $b_\alpha(x) = (b(x)/e)^\alpha$ for sufficiently large $\alpha = \alpha(\delta)$, it holds that $|\hat{b_\alpha}(t)| = O(e^{-|t|^{1-\delta}})$. An alternate construction which convolves with $(\hat{b_\alpha})_c$ for large α then only requires $c = \varepsilon^{-1}\log^{1+o(1)}(1/\varepsilon)$.*

A.4 Existence of a good function for showing anticoncentration

Part of the proof of Theorem 2.1 required us to exhibit, for any $t \in \mathbb{R}$ and $0 < \varepsilon < \varepsilon_0$ for some constant $\varepsilon_0 > 0$, a nonnegative function $f_{t,\varepsilon} : \mathbb{R} \to \mathbb{R}$ symmetric about t satisfying the following properties:

i. $||f^{(\ell)}_{t,\varepsilon}||_\infty = O(\alpha^\ell)$ for all $\ell \geq 0$, with $\alpha = O(1/\varepsilon)$

ii. $\mathbf{E}[f_{t,\varepsilon}(D)] = O(\varepsilon)$ for $D \sim \mathcal{D}_p$

iii. $f_{t,\varepsilon}(t + \varepsilon) = \Omega(1)$

iv. $f_{t,\varepsilon}(x)$ is strictly decreasing as $|x - t| \to \infty$

We first prove the following useful lemma.

Lemma A.4. *Let $f : \mathbb{C} \to \mathbb{C}$ be holomorphic on the complex plane satisfying $|f(z)| = e^{O(1+|\Im(z)|)}$ for all $z \in \mathbb{C}$, where $\Im(z)$ denotes the imaginary part of z. Then, for any $x \in \mathbb{R}$, $|f^{(\ell)}(x)| = e^{O(1+\ell)}$.*

Proof. Let $\mathcal{C}$ be the circle of radius ℓ centered at x in the complex plane. By Cauchy's integral formula,

$$\begin{aligned} |f^{(\ell)}(x)| &= \left|\frac{\ell!}{2\pi i}\oint_{\mathcal{C}}\frac{f(z)}{(z-x)^{\ell+1}}dz\right| \\ &\le \frac{\ell!}{2\pi}\int_0^{2\pi}\left|\frac{e^{O(1+|\ell\cdot\sin(t)|)}}{(\ell e^{it})^{\ell+1}}\ell e^{it}dt\right| \\ &\le \frac{\ell! e^{O(1+\ell)}}{2\pi\ell^\ell}\int_0^{2\pi}\frac{1}{|e^{i\ell t}|}dt \\ &\le \frac{e^{O(1+\ell)}}{2\pi}\int_0^{2\pi}dt \\ &= e^{O(1+\ell)}. \end{aligned}$$

■

Now, consider the function

$$f(x) = -\int_{-\infty}^{x}\frac{\sin^4(y)}{y^3}dy = -\int_{-\infty}^{x}\operatorname{sinc}^3(y)\sin(y)dy.$$

Then on $\mathbb{R}$, f is even since it is the integral of an odd function, and it is nonnegative by construction. Now consider $x \in \mathbb{C}$. The integrand is the product of holomorphic functions and is thus holomorphic, so we may apply Cauchy's integral theorem to choose a contour to evaluate the integral. We choose the piecewise linear contour which goes from $-\infty$ to $\Re(x)$ along $\mathbb{R}$, followed by the line from $\Re(x)$ to x, where $\Re(x)$ denotes the real part of x. The former integral is $O(1)$. The latter integral is pointwise bounded in magnitude by $e^{O(1+|\Im(x)|)}$, which can be seen by expanding $\sin(y) = (e^{iy} - e^{-iy})/2$, and thus in total is bounded by $|\Im(x)| \cdot e^{O(1+|\Im(x)|)} = e^{O(1+|\Im(x)|)}$. It thus follows that $||f^{(\ell)}||_\infty = e^{O(1+\ell)}$ on $\mathbb{R}$ by Lemma A.4.

Now define $f_{0,\varepsilon} : \mathbb{R} \to \mathbb{R}$ by $f_{0,\varepsilon}(x) = f(x/\varepsilon)$ (the general construction is given by setting $f_{t,\varepsilon} = f_{0,\varepsilon}(x-t)$, then applying a similar argument as to what follows). This function is nonnegative and symmetric about 0 by the above discussion, and has the desired derivative bounds of (i) by the chain rule and bounds on $||f^{(\ell)}||_\infty$. Let μ_p be the density function of $\mathcal{D}_p$, which is bounded by some absolute constant C_p on $\mathbb{R}$. Note $f_{0,\varepsilon}(x) = O(\varepsilon^2/x^2)$, and recall $\mathcal{D}_p$ is anticoncentrated. Item (ii) is then satisfied since

$$\begin{aligned} \mathbf{E}[f_{0,\varepsilon}(D)] \le f(0)\cdot\mathbf{Pr}[|D| \le \varepsilon] \\ + 2C_p\int_{\varepsilon}^{\infty}O(\varepsilon^2/x^2)dx = O(\varepsilon). \end{aligned}$$

Item (iii) follows since $f_{0,\varepsilon}(\varepsilon) = f(1)$, and (iv) follows by construction.

A.5 Fooling regular halfspaces For $a \in \mathbb{R}^n$ with $||a||_2 = 1$ and $\theta \in \mathbb{R}$, let $f_{a,\theta}(x) = \operatorname{sgn}(\langle a, x\rangle - \theta)$ for $x \in \{-1,1\}^n$. Diakonikolas et al. showed that $\mathbf{E}_x[f_{a,\theta}(x)]$ is preserved to within $O(\varepsilon)$ even if the x_i are only $k(\varepsilon)$-wise independent for $k(\varepsilon) = O(\varepsilon^{-2}\log^2(1/\varepsilon))$. They first provided a reduction to the *regular* case, where $|a_i| < \varepsilon$ for all $i \in [n]$, then resolved the regular case. We give an alternative proof that $\varepsilon^{-2}\log^{2+o(1)}$-wise independence fools regular halfspaces, which is an adaptation of our proof of Theorem 2.1.

Let $x = (x_1, \ldots, x_n)$ have fully independent entries and $y = (y_1, \ldots, y_n)$ have k-wise independent entries in $\{-1,1\}$ for k even. Let $X = \langle a, x\rangle$ and $Y = \langle a, y\rangle$. It suffices to show $\mathbf{E}[I_{[\theta,\infty)}(X)] \approx_\varepsilon \mathbf{E}[I_{[\theta,\infty)}(Y)]$ since $f_{a,\theta}(z) = 2\cdot I_{[\theta,\infty)}(\langle a, z\rangle) - 1$. Similarly to the proof of Theorem 2.1, we show this by proving

$$\begin{aligned} \mathbf{E}[I_{[\theta,\infty)}(X)] &\approx_\varepsilon \mathbf{E}[J^c_{[\theta,\infty)}(X)] \\ &\approx_\varepsilon \mathbf{E}[J^c_{[\theta,\infty)}(Y)] \approx_\varepsilon \mathbf{E}[I_{[\theta,\infty)}(Y)] \end{aligned}$$

with $J^c_{[\theta,\infty)}$ as in Lemma 2.5, $c = O(\varepsilon^{-1}\log^3(1/\varepsilon))$.

$\mathbf{E}[\mathbf{I}_{[\theta,\infty)}(\mathbf{X})] \approx_\varepsilon \mathbf{E}[\mathbf{J}^{\mathbf{c}}_{[\theta,\infty)}(\mathbf{X})]$**:** This follows since (a) $I_{[\theta,\infty)}$ and $J_{[\theta,\infty)}$ are within ε everywhere except for two intervals of length $O(\varepsilon)$, and (b) the random variable X is anticoncentrated (any length-$O(\varepsilon)$ interval contains $O(\varepsilon)$ probability mass) by the Berry-Esséen theorem, by regularity; since $||I_{[\theta,\infty)}||_\infty, ||J^c_{[\theta,\infty)}||_\infty = O(1)$, these intervals contribute $O(\varepsilon)$ to the difference.

$\mathbf{E}[\mathbf{J}^{\mathbf{c}}_{[\theta,\infty)}(\mathbf{X})] \approx_\varepsilon \mathbf{E}[\mathbf{J}^{\mathbf{c}}_{[\theta,\infty)}(\mathbf{Y})]$**:** Unlike in the case of Lemma 2.2, we can directly apply Taylor's theorem since X, Y have finite moments. By Taylor's theorem,

$$J^c_{[\theta,\infty)}(X) = p_{k-1}(X) \pm \frac{||(J^c_{[\theta,\infty)]})^{(k)}||_\infty X^k}{k!},$$

where $\text{degree}(p_{k-1}) = k-1$. Do similarly for Y. Then

$$\text{(A.3)}\quad |\mathbf{E}[J^c_{[\theta,\infty)}(X)] - \mathbf{E}[J^c_{[\theta,\infty)}(Y)]| \le \frac{O(c^k)\cdot\mathbf{E}[X^k]}{k!}$$

since $||(J^c_{[\theta,\infty)})^{(k)}||_\infty = O(c^k)$ and $\mathbf{E}[X^k] = \mathbf{E}[Y^k]$. We have $\mathbf{E}[X^k] \le k^{k/2}$ by Khintchine's inequality [20], and $k! = k^k/2^{O(k)}$. Thus, setting $k = \Omega(c^2)$ makes Eq. (A.3) at most ε. Our setting of c causes k to be $O(\varepsilon^{-2}\log^6(1/\varepsilon))$, but as discussed in Remark A.1 of Section A.3, this can be improved to $\varepsilon^{-2}\log^{2+o(1)}(1/\varepsilon)$.

$\mathbf{E}[\mathbf{J}^{\mathbf{c}}_{[\theta,\infty)}(\mathbf{Y})] \approx_\varepsilon \mathbf{E}[\mathbf{I}_{[\theta,\infty)}(\mathbf{Y})]$**:** Just as in the proof of Theorem 2.1, we argue as in the first inequality, but we now must show anticoncentration of Y. This is done exactly as in the proof of Theorem 2.1, using the same function $f_{t,\varepsilon}$ of Section A.4.

A.6 L_p algorithm precision issues In this section, we deal with the precision issues mentioned in the proofs of Theorem 2.1 and Theorem 2.2 in Section 2. Since the argument is nearly identical in both algorithms, we focus only that of (Figure 2) discussed in Theorem 2.2.

We deal with rounding errors first. We will pick some number $\delta = \Theta(\varepsilon m^{-1})$. We round each $A_{i,j}, A'_{i,j}$ to the nearest multiple of δ. This means that we only need to store the y_i, y'_i to a precision of δ. This does produce an error in these values of size at most $||x||_1\delta \leq |i : x_i \neq 0| \cdot \max(|x_i|) \cdot \delta \leq m||x||_p\delta = \Theta(\varepsilon||x||_p)$. In this case, y'_{med} will differ by at most an additive $\varepsilon||x||_p$, and thus still be within a constant factor of $||x||_p$ with 7/8 probability. Also, $C = (\sum_i \cos(y_i/y'_{\text{med}}))/r$ will be calculated with an additive error of $O(\varepsilon)$ (which is a multiplicative error of $1 + O(\varepsilon)$ since $C = \Theta(1)$) since $(\sum_i \cos((y_i + O(\varepsilon||x||_p))/y'_{\text{med}}))/r = (\sum_i(\cos(y_i/y'_{\text{med}}) + O(\varepsilon))/r$.

Next we need to determine how to sample from these continuous distributions. It was shown by [9], and also used in [23], that a p-stable random variable can be generated by taking θ uniform in $[-\pi/2, \pi/2]$, r uniform in $[0, 1]$ and letting

$$X = f(r, \theta) = \frac{\sin(p\theta)}{\cos^{1/p}(\theta)} \cdot \left(\frac{\cos(\theta(1-p))}{\log(1/r)}\right)^{(1-p)/p}.$$

We would like to know how much of an error is introduced by using values of r and θ only accurate to within δ'. This error is at most δ' times the derivative of f. This derivative is not large except when θ or $(1-p)\theta$ is close to $\pm\pi/2$, or when r is close to 0 or 1. Since we only ever need mr different values of $A_{i,j}$ (and even fewer values for A'), we can assume that with reasonable probability we never get an r or θ closer to these values than $O(m^{-1}\varepsilon^2)$. In such a case the derivative will be bounded by $(m\varepsilon^{-1})^{O(1)}$. Therefore, if we choose r and θ with a precision of $(m^{-1}\varepsilon)^{O(1)}$, we can get the value of X with introducing an error of only δ.

Lastly, we need to consider memory requirements. We consider only the memory requirements for storing A, since that for A' is even less. Our rows must be from a 2-independent family containing $O(\varepsilon^{-2})$ k-independent families of n random variables. Each random variable requires $O(\log(m\varepsilon^{-1}))$ bits. The amount of space needed to pick out an element of this family is $O(k(\log(n) + \log(m\varepsilon^{-1}))) = O(k\log(nm/\varepsilon)) = O(k\log(nm))$ bits (recall $\varepsilon \geq 1/\sqrt{m}$). The dependence here on n can be eliminated using the observation of Section A.1.1, with an additive $O(\log\log n)$ cost. We also need to store the y_i to a precision of δ. Since there are only mr values of $A_{i,j}$ we ever consider in the stream, with large constant probability, none of them is bigger than a polynomial in mr. If this is the case, the maximum value of any y_i is at most $(mM\varepsilon^{-1})^{O(1)}$. Hence each of the $O(1/\varepsilon^2)$ values y_i can be stored in $O(\log(mM\varepsilon^{-1})) = O(\log(mM))$ bits.

A.7 Further lower bounds

Observation A.1. *For vectors u, v of equal length with entries in $\{0, r\}$, let $\Delta(u, v) = |\{i : u_i \neq v_i\}|$ denote their Hamming distance. Let $w(z) = |\{i : z_i \neq 0\}|$ denote the weight of z. Then for any $p \geq 0$,*

$$r^p(2^p - 2)\Delta(u, v) = (2r)^p w(u) + (2r)^p w(v) - 2||u + v||_p^p.$$

We remind the reader that in strict turnstile streams, each frequency x_i is promised to always be nonnegative. We now show the following.

Theorem A.1. *For any real constant $p > 0$, any one-pass streaming algorithm for $(1 \pm \varepsilon)$-multiplicative approximation of F_p with probability at least 11/12 in the strict turnstile model requires $\Omega(\min\{N, |p - 1|^2\varepsilon^{-2}(\log(\varepsilon^2 mM/|p-1|^2))\})$ bits of space.*

Proof (Sketch). The proof is very similar to that of Theorem 3.3, so we only explain the differences. The main difference is the following. In the proof of Theorem 3.3, Bob inserted item $(i(z), j)$ into the stream with frequency $-\lfloor q^{i(z)} \rfloor$ for each j satisfying $y(B)_j = 1$. Doing this may not yield a strict turnstile stream, since $(i(z), j)$ may never have received a positive update from Alice. We instead have Bob insert $(i(z), j)$ with *positive* frequency $\lfloor q^{i(z)} \rfloor$.

Now, after all updates have been inserted into the stream, Observation A.1 implies that the pth frequency moment of the stream is exactly

$$\begin{aligned} L'' = &\frac{(2\lfloor q^{i(z)} \rfloor)^p}{2} w(y_{i(z)}) + \frac{(2\lfloor q^{i(z)} \rfloor)^p}{2} w(y(B)) \\ &- \frac{\lfloor q^{i(z)} \rfloor^p (2^p - 2)}{2} \Delta(y_{i(z)}, y(B)) \\ &+ \sum_{i < i(z)} w(y_i) \lfloor q^i \rfloor^p. \end{aligned}$$

Setting $\eta = \sum_{i<i(z)} w(y_i) \lfloor q^i \rfloor^p$ and rearranging terms,

$$\begin{aligned} \Delta(y_{i(z)}, y(B)) = &\frac{2^{p-1}}{2^{p-1} - 1} w(y_{i(z)}) + \frac{2^{p-1}}{2^{p-1} - 1} w(y(B)) \\ &+ \frac{\lfloor q^{i(z)} \rfloor^{-p} (\eta - L'')}{2^{p-1} - 1}. \end{aligned}$$

Bob knows η, $w(y_{i(z)})$, and $w(y(B))$ exactly. To decide Gap-Hamdist for vectors $y_{i(z)}, y(B)$, it thus suffices to obtain a $((2^{p-1} - 1)/(4\sqrt{ck}))$-additive approximation to $\lfloor q^{i(z)} \rfloor^{-p} L''$. Since $2^{-i(z)}L''$ is upper-bounded in absolute value by $(1 + 2^p)ck$, our desired additive approximation is guaranteed by obtaining a $(1 \pm ((2^{p-1} - 1)\sqrt{ck}/(4 \cdot (1 + 2^p))))$-multiplicative approximation to L''. Since $p \neq 1$ is a constant, $|2^x - 1| = \Theta(|x|)$

as $|x| \to 0$, and $k = \Theta(1/\varepsilon^2)$, this is a $(1 \pm O(|p-1|\varepsilon))$-multiplicative approximation. To conclude, a $(1 \pm O(|p-1|\varepsilon))$-multiplicative approximation to F_p with probability 11/12 gives a protocol for AUGMENTED-INDEXING with success probability 2/3, with Alice having a string of length kt for k, t as in the proof of Theorem 3.3. The theorem follows. ■

The lower bounds of Theorem 3.3 and Theorem A.1 fail to give any improved lower bound over the previously known $\Omega(\min\{N, 1/\varepsilon^2\})$ lower bound for p near (and including) 0. The reason is that we gave items in block j a frequency of roughly $2^{j/p}$, so that contributions to F_p increase geometrically as block ID increases. This fails for, say, $p = 0$, since in this case increasing frequency does not increase contribution to F_0 at all. We fix this issue by, rather than giving items in large blocks a large frequency, instead blow up the universe size. Specifically, we use a proof identical to that of Theorem A.1, but rather than give an index i in block j frequency roughly $2^{j/p}$, we instead create 2^j indices $(i,1), \ldots, (i, 2^j)$ and give them each a frequency of 1. The setting of t, the number of blocks, can then be at most $O(\log(\varepsilon^2 N))$ since $n, m \le 2\varepsilon^{-2} \sum_{j=0}^{t-1} 2^j$, which we require to be at most N. We thus have:

THEOREM A.2. *For any real constant $p \ge 0$, one-pass $(1 \pm \varepsilon)$-multiplicative approximation of F_p with probability at least 11/12 in the strict turnstile model requires $\Omega(|p-1|^2 \varepsilon^{-2} \log(\varepsilon^2 N/|p-1|^2))$ bits of space.*

The decay of our lower bounds in the strict turnstile model as $p \to 1$ is necessary since Li gave an algorithm in this model whose dependence on ε becomes subquadratic as $p \to 1$ [31]. Furthermore, when $p = 1$ there is a simple, deterministic $O(\log(mM))$-space algorithm for computing F_1: maintain a counter.

Our technique also improves the known lower bound for additively estimating the entropy of a stream in the strict turnstile model.[3] The proof combines ideas of [7] with our technique of embedding geometrically-growing hard instances. By entropy of the stream, we mean the empirical probability distribution on $[n]$ obtained by setting $p_i = x_i/||x||_1$.

THEOREM A.3. *Any algorithm for ε-additive approximation of H, the entropy of a stream, in the strict turnstile model with probability at least 11/12 requires space $\Omega(\varepsilon^{-2} \log(N)/\log(1/\varepsilon))$.*

Proof. We reduce from AUGMENTED-INDEXING, as in Theorem 3.3. Alice receives a string of length $s = \log N/(2\varepsilon^2 \log(1/\varepsilon))$, and Bob receives an index $z \in [s]$. Alice conceptually divides her input into $b = \varepsilon^2 s$ blocks, each of size $1/\varepsilon^2$, and reduces each block using the INDEXING→GAP-HAMDIST reduction of Theorem 3.2 to obtain b GAP-HAMDIST instances with strings $y_1, \ldots, y_b$, each of length $\ell = \Theta(1/\varepsilon^2)$. For each $1 \le i \le b$, and $1 \le j \le \ell$ Alice inserts universe elements $(i, j, 1, (y_i)_j), \ldots, (i, j, \varepsilon^{-2i}, (y_i)_j)$ into the stream and sends the state of a streaming algorithm to Bob.

Bob identifies the block $i(z)$ in which z lands and deletes all stream elements associated with blocks with index $i > i(z)$. He then does his part in the INDEXING→GAP-HAMDIST reduction to obtain a vector $y(\mathrm{Bob})$ of length ℓ. For all $1 \le j \le \ell$, he inserts the universe elements $(i(z), j, 1, y(\mathrm{Bob})_j), \ldots, (i(z), j, \varepsilon^{-2i(z)}, y(\mathrm{Bob})_j)$ into the stream.

The number of stream tokens from block indices $i < i(z)$ is $A = \varepsilon^{-2} \sum_{i=0}^{i(z)-1} \varepsilon^{-2i} = \Theta(\varepsilon^{-2i(z)})$. The number of tokens in block $i(z)$ from Alice and Bob combined is $2\varepsilon^{-(2i(z)+2)}$. Define $B = \varepsilon^{-2i(z)}$ and $C = \varepsilon^{-2}$. The L_1 weight of the stream is $R = A + 2BC$. Let Δ denote the Hamming distance between $y_{i(z)}$ and $y(\mathrm{Bob})$ and H denote the entropy of the stream.

We have:

$$\begin{aligned} H &= \frac{A}{R}\log(R) + \frac{2B(C-\Delta)}{R}\log\left(\frac{R}{2}\right) + \frac{2B\Delta}{R}\log(R) \\ &= \frac{A}{R}\log(R) + \frac{2BC}{R}\log(R) - \frac{2BC}{R} + \frac{2B\Delta}{R} \end{aligned}$$

Rearranging terms gives

$$\Delta = \frac{HR}{2B} + C - C\log(R) - \frac{A}{2B}\log(R) \qquad \text{(A.4)}$$

To decide the GAP-HAMDIST instance, we must decide whether $\Delta < 1/2\varepsilon^2 - 1/\varepsilon$ or $\Delta > 1/2\varepsilon^2 + 1/\varepsilon$. By Eq. (A.4) and the fact that Bob knows A, B, C, and R, it suffices to obtain a $1/\varepsilon$-additive approximation to $HR/(2B)$ to accomplish this goal. In other words, we need a $2B/(\varepsilon R)$-additive approximation to H. Since $B/R = \Theta(\varepsilon^2)$, it suffices to obtain an additive $\Theta(\varepsilon)$-approximation to H. Let $\mathcal{A}$ be a streaming algorithm which can provide an additive $\Theta(\varepsilon)$-approximation with probability at least 11/12. Recalling that correctly deciding the GAP-HAMDIST instance with probability 11/12 allows one to correctly decide the original AUGMENTED-INDEXING instance with probability 2/3 by Theorem 3.2, and given Theorem 3.1, $\mathcal{A}$ must use at least $\log(N)/(\varepsilon^2 \log(1/\varepsilon))$ bits of space. As required, the length of the vector being updated in the stream is at most $\sum_{i=1}^{s} \varepsilon^{-2i} = O(N) = O(n)$, and the length of the stream is exactly twice the vector length, and thus $O(N) = O(m)$. ■

[3] The previous proof of [7] though has the advantage of even holding in the weakest insertion-only model, i.e. no negative frequency updates.