High Probability Frequency Moment Sketches

Sumit Ganguly
Indian Institute of Technology, Kanpur, India
sganguly@cse.iitk.ac.in

David P. Woodruff
Carnegie Mellon University, School of Computing, Pittsburg, USA
dwoodruff@cs.cmu.edu

Abstract

We consider the problem of sketching the $p$-th frequency moment of a vector, $p > 2$, with multiplicative error at most $1 \pm \epsilon$ and with high confidence $1 - \delta$. Despite the long sequence of work on this problem, tight bounds on this quantity are only known for constant $\delta$. While one can obtain an upper bound with error probability $\delta$ by repeating a sketching algorithm with constant error probability $O(\log(1/\delta))$ times in parallel, and taking the median of the outputs, we show this is a suboptimal algorithm! Namely, we show this is a suboptimal algorithm! Namely, we show optimal upper and lower bounds of $\Theta(n^{1-2/p} \log(1/\delta) + n^{1-2/p} \log^{2/p}(1/\delta) \log n)$ on the sketching dimension, for any constant approximation. Our result should be contrasted with results for estimating frequency moments for $1 \leq p \leq 2$, for which we show the optimal algorithm for general $\delta$ is obtained by repeating the optimal algorithm for constant error probability $O(\log(1/\delta))$ times and taking the median output. We also obtain a matching lower bound for this problem, up to constant factors.

1 Introduction

The frequency moments problem is a very well-studied and foundational problem in the data stream literature. In the data stream model, an algorithm may use only sub-linear memory and a single pass over the data to summarize a data stream that appears as a sequence of incremental updates. A data stream may be viewed as a sequence of records of the form $((i_1, v_1), (i_2, v_2), \ldots, (i_m, v_m))$, where, $i_j \in [n] = \{1, 2, \ldots, n\}$ and $v_j \in \mathbb{R}$. The record $(i_j, v_j)$ changes the $i_j$-th coordinate $x_{i_j}$ of an underlying $n$-dimensional vector $x$ to $x_{i_j} + v_j$. Equivalently, for $i \in [n]$, $x_i = \sum_{j: i_j = i} v_j$. Note that $v_j$ may be positive or negative, which corresponds to the so-called turnstile model in data streams. Also, the $i$-th coordinate of $x$ is sometimes referred to as the frequency of item $i$, though note that it can be negative in the turnstile model. The $p$-th moment of $x$ is defined to be $F_p = \sum_{i \in [n]} |x_i|^p$, for a real number $p \geq 0$, which for $p \geq 1$ corresponds to the $p$-th power of the $\ell_p$-norm $\|x\|_p$ of $x$.

The $F_p$ estimation problem with approximation parameter $\epsilon$ and failure probability $\delta$ is: design an algorithm that makes one pass over the input stream and returns $\hat{F}_p$ such that $\Pr[|\hat{F}_p - F_p| \leq \epsilon F_p] \geq 1 - \delta$. Such an algorithm is also referred to as an $(\epsilon, \delta)$-approximation of $F_p$. This is a problem that is among the ones that has received the most attention.
We start the upper bound timeline with [19], since that is the first work which achieved an exponent of $1 - 2/p$ for $n$. For earlier work which achieved worse exponents for $n$, see [1, 12, 14, 15].

<table>
<thead>
<tr>
<th>$F_p$ Algorithm</th>
<th>Sketching Dimension</th>
</tr>
</thead>
<tbody>
<tr>
<td>[19]</td>
<td>$O(n^{1-2/p}e^{-O(1)}\log^{O(1)}(n\log(1/\delta)))$</td>
</tr>
<tr>
<td>[7]</td>
<td>$O(n^{1-2/p}e^{-2-4/p}\log n \log(M)\log(1/\delta))$</td>
</tr>
<tr>
<td>[31]</td>
<td>$O(n^{1-2/p}e^{-O(1)}\log^{O(1)}(n\log(1/\delta)))$</td>
</tr>
<tr>
<td>[3]</td>
<td>$O(n^{1-2/p}e^{-2-6/p}\log n \log(1/\delta))$</td>
</tr>
<tr>
<td>[8]</td>
<td>$O(n^{1-2/p}e^{-2-4/p}\log n \cdot g(p,n)\log(1/\delta))$</td>
</tr>
<tr>
<td>[2]</td>
<td>$O(n^{1-2/p}\log N^c e^{-O(1)}\log(1/\delta))$</td>
</tr>
<tr>
<td>[16], Best upper bound</td>
<td>$O(n^{1-2/p}\log^{O(1)}(1/\delta) + n^{1-2/p}e^{-4/p}\log n \log(1/\delta))$</td>
</tr>
</tbody>
</table>

in the data stream literature, and we only give a partial list of work on this problem [1, 2, 3, 4, 6, 7, 8, 10, 12, 14, 15, 16, 20, 19, 25, 26, 27, 30, 31, 34].

We study the class of algorithms based on linear sketches, which store only a sketch of the input vector $x$ and a (possibly randomized) matrix $A$. This model is well-studied, both for the problem of estimating norms and frequency moments [4, 18, 30, 32], and for other problems such as estimating matrix norms [29], and matching size [5, 28]. The efficiency is measured in terms of the sketching dimension which is the maximum number of rows of a matrix $S$ used by the algorithm. Since the algorithm is randomized, it may choose different $S$ based on its randomness, so the maximum is taken over its randomness. Linear sketches are particularly useful for data streams since given an update $(i_j, v_j)$, one can update $S x$ as $S(x + v_j e_{i_j}) = Sx + Sv_j e_{i_j}$, where $e_{i_j}$ is the standard unit vector in the $i_j$-th direction. They are also used in distributed environments, since given $S \cdot x$ and $618 S \cdot y$, one can add these to obtain $S \cdot (x + y)$, the sketch of $x + y$.

When $0 < p \leq 2$, one can achieve a sketching dimension of $O(\epsilon^{-2} \log(1/\delta))$ independent of $n$ [1, 25, 27], while for $p = 0$ the sketching dimension is $O(\epsilon^{-2}(\log(1/\epsilon) + \log \log n)\log(1/\delta))$ [26]. For $p = 2$ there is a sketching lower bound of $\Omega(\epsilon^{-2} \log(1/\delta))$ [24], which implies an optimal algorithm for general $\delta$ is to run an optimal algorithm with error probability $1/3$ and take the median of $O(\log(1/\delta))$ independent repetitions. As a side result, we show in the full version a lower bound of $\Omega(\epsilon^{-2} \log(1/\delta))$ for any $1 \leq p < 2$, which shows this strategy of amplifying the success probability by $O(\log(1/\delta))$ independent repetitions is also optimal for any $1 \leq p < 2$.

Perhaps surprisingly, for $p > 2$, the sketching dimension needs to be polynomial in $n$, as first shown in [32], with the best known lower bounds being $\Omega(n^{1-2/p} \log n)$ [4] for constant $\epsilon$ and $\delta$, and $\Omega(n^{1-2/p}e^{-2})$ for constant $\delta$ [30]. Regarding upper bounds, we present the long list of bounds in Table 1. The best known upper bound is $O(n^{1-2/p}e^{-2}\log(1/\delta) + n^{1-2/p}e^{-4/p}\log n \log(1/\delta))$ [16]. This is tight only when $\epsilon$ and $\delta$ are constant, in which case it matches [4], or when $\delta$ is constant and $\epsilon < 1/\log(\log n)$, since it matches [30].

### 1.1 Our Contributions

In this work, we show optimal upper and lower bounds of $\Theta(n^{1-2/p} \log(1/\delta) + n^{1-2/p} \log^{O(1)}(1/\delta) \log n)$ on the sketching dimension for $F_p$-estimation, for any $p > 2$, and for any constant $\epsilon$. Our upper bound shows, perhaps surprisingly, that the optimal bound is not to run $O(\log(1/\delta))$ independent repetitions of a constant success probability algorithm and report the median of the outputs. Indeed, such an algorithm would give a worse $O(n^{1-2/p} \log(1/\delta) \log n)$ sketching dimension.
Figure 1 Shelf structure and level sets for each shelf index $j$ whose contribution to $F_p$ is estimated accurately.

For general $\epsilon$, our upper bound is $O(n^{1-2/p-2}/(1/\delta)) + n^{1-2/p-4/p}\log^2(1/\delta)\log n)$ and our lower bound is $\Omega(n^{1-2/p-2}(1/\delta) + n^{1-2/p-2/p}(1/\delta)\log n)$, which differ by at most an $\epsilon^{-2/p}$ factor. Our results thus come close to resolving the complexity for general $\epsilon$ as well.

Our results should be contrasted to $1 \leq p \leq 2$, for which the optimal sketching dimension for such $p$ is $\Theta(\epsilon^{-2}\log(1/\delta))$, and so for these $p$ it is optimal to run $O(\log(1/\delta))$ independent repetitions of a constant probability algorithm. Here we strengthen the $\Omega(\epsilon^{-2}\log(1/\delta))$ bound for $p = 2$ of [24] by showing the same bound for $1 \leq p \leq 2$.

### 1.1.1 Overview of Upper Bound

In order to obtain a confidence of $1 - \delta$, we use the $d = \lceil \log(1/\delta) \rceil$th moment of an estimate $\hat{F}_p$ of $F_p$. Since we are unable to use the $d$th moment of the Taylor polynomial estimator of [17], we employ a different estimator $X_i$ for estimating individual coordinates $|x_i|$ and use it as $X_i^p$ to estimate $|x_i|^p$. This estimator is based on (a) using random $q$th roots of unity for sketches instead of standard Rademacher variables, and (b) taking the average of the estimates from those tables where the item does not collide with the set of top-$k$ estimated heavy hitters.

**The Shelf Structure.** The algorithm uses two structures, namely, a GHSS-like structure from [17] and a new shelf structure, which is our main algorithmic novelty (both formally defined later). The shelf structure is necessary when the failure probability is $\delta = n^{-\Theta(1)}$; otherwise, for $\delta = n^{-\Theta(1)}$, somewhat surprisingly the GHSS structure of [17] alone suffices with parameter $C = n^{1-2/p}(1/\delta)/\log(1/\delta)$ and number of measurements $O(C\log n)$, which requires an intricate $d$-th moment analysis of the GHSS structure.

The shelf structure is partitioned into shelves, indexed from $j = 0, \ldots, J$, for a value $J$ which is specified below. Each shelf consists of a pair of CountSketch like structures, $HH_j$ and $AvgEst_j$. The number of buckets in the tables of the $j$th shelf is $H_j$ and the number of tables in the $j$th shelf of the $HH_j$ structure is $w_j$ and of the $AvgEst_j$ structure is $2w_j$. We set $H_J = \Theta(n^{1-2/p-2})$ and $w_J = \Theta(1/\delta)$, while $H_0 = \Theta(n^{1-2/p-4/p}\log^2(1/\delta))$ and $w_0 = s = \Theta(\log n)$. 

- $U_0 = \frac{n^{1-2/p-2}}{\log^2(1/\delta)}$ buckets per table
- $U_1 = U_0/2$ tables
- $w_0 = a \cdot w_0$
- $H_0 = O(1/\delta)$ tables
The input vector $x$ is provided as input to all the shelves' structures. The table height $H_j = H_0b^j$ decays geometrically with parameter $0 < b < 1$ and the table width $w_j = w_0a^j$ increases geometrically with parameter $a > 1$. Note that the parameters $a$ and $b$ determine $J$. By requiring that $|1 - ab| = \Omega(1)$, we ensure that the total number of measurements of the shelf structure is $\sum_{j=0}^{J} H_j w_j = O(H_0w_0 + H_Jw_J)$, no matter which value of $J$ we choose. For the shelf structure, frequency-wise thresholds are defined as $U_j = O(\hat{F}_2/H_j)^{1/2}$, for $j = 0, 1, \ldots, J$. The shelf frequency group corresponding to shelf $j$ is $S_j = [U_j, U_{j+1})$, where, $U_{J+1} = \infty$ and $U_0 = T_0$. We sometimes conflate $S_j$ with the set of items whose frequency belongs to $S_j$. The frequency group $G_0$ is defined as $[T_0, U_1]$ and coincides with $S_0$. See Figure 1.

So why a shelf structure? Suppose for simplicity that $\epsilon$ is a constant. Consider a vector $x$ which has a constant number of “large” coordinates of value $\Theta(n^{1/p})$, and $\Theta(n)$ remaining “small” coordinates of absolute value $O(1)$. Then we need to find all the large coordinates to accurately estimate $F_p$ up to a small constant factor. This is well-known to be possible with $\Theta(n^{1-2/p})$ buckets in the $J$-th shelf, since with probability $1 - \delta$, each of the large coordinates will not collide with any other large coordinate in more than a small constant fraction of tables. Note that in each table, in each bucket containing a large coordinate, the “noise” in the bucket from small coordinates will be $Cn^{1/p}$ for an arbitrarily small constant $C > 0$ with constant probability, and so this will happen in most buckets containing a large coordinate in most tables with probability $1 - \delta$.

However, now consider a vector $x$ which has $\Theta(\log(1/\delta))$ “large-ish” coordinates of value $\Theta(n^{1/p}/\log^{1/p}(1/\delta))$, and $\Theta(n)$ remaining “small” coordinates of absolute value $O(1)$, as before. Then we again need to find most of the “large-ish” coordinates to accurately estimate $F_p$ up to a constant factor. We also cannot subsample and try to estimate how many large-ish coordinates there are from a subsample. Indeed, since there are only $O(\log(1/\delta))$ total large-ish coordinates, sub-sampling would not accurately estimate this total with probability at least $1 - \delta$. However, to find these “large-ish” coordinates, we need to increase the number of buckets from $\Theta(n^{1-2/p})$ to $\Theta(n^{1-2/p} \cdot \log^{2/p}(1/\delta))$ just so that in a bucket containing one of these coordinates, with constant probability the noise will not be too large. But if we then want this to happen for a $1 - \delta$ fraction of tables, we still need $\Theta(\log(1/\delta))$ tables, which gives overall $\Theta(n^{1-2/p} \cdot \log^{1+2/p}(1/\delta))$ measurements, which is above our desired total of $O(n^{1-2/p}(\log(1/\delta) + \log(n) \log^{2/p}(1/\delta)))$ measurements.

So what went wrong? The key idea in our analysis is to relax the requirement of trying to recover all the large-ish coordinates with probability $1 - \delta$. Suppose instead of $\Theta(\log(1/\delta))$ tables we just use $\Theta(\log n)$ tables. Then with probability $1 - 1/n$, there may be two large-ish coordinates which collide and cancel with each other in every single table, and we have no way of recovering them. However, we are able to show that with probability $1 - \delta$, only $O(\log(1/\delta)/\log n)$ large-ish coordinates will fall into this category, and neglecting this roughly $(1 - 1/\log n)$ fraction of the large-ish coordinates will not affect our estimate of $F_p$ by more than a constant factor. And indeed, our 0-th shelf has exactly $\Theta(n^{1-2/p} \cdot \log^{2/p}(1/\delta))$ buckets and $\Theta(\log n)$ tables, so is exactly suited for finding these large-ish coordinates. In general, we can show that one of our shelves will be able to handle every vector with coordinates of magnitude between the large and large-ish coordinates. Again, by choosing the shelf structure carefully, the total number of measurements is dominated by that in the zero-th plus the $J$-th shelf, giving us $O(n^{1-2/p}(\log(1/\delta) + \log(n) \log^{2/p}(1/\delta)))$ total measurements, and explaining where the $\log^{2/p}(1/\delta)$ in the upper bound comes from.
The Non-Large-ish Coordinates. Our shelves are designed to estimate the contribution to $F_p$ from all coordinates of absolute value at least $\Theta(n^{1/p} / \log^{1/p}(1/\delta))$. For coordinates of smaller value, we can now afford to sub-sample and apply the same 0-th shelf structure to estimate their contribution to $F_p$. We apply the ghss structure, which is analogous to the structure presented in [17] and has $L + 1$ levels corresponding to $l = 0, \ldots, L$, and consists of a pair of CountSketch like structures $\text{HH}_l$ and $\text{AvgEst}_l$ at each level. The sub-sampling technique and the associated frequency-wise thresholds and frequency groups are defined analogously (with new parameters) to [17].

A notable difference with [17] is that the $\text{AvgEst}$ structures in the ghss and shelf structures use complex qth roots of unity and return the average of table estimates instead of the median of table estimates used by CountSketch, which are novelties in this context, though have been used for other data stream problems [23]. We have that $E[X_p^i] = |x_i|^p(1 \pm n^{-\Omega(1)})$ for our estimator $X_i$ of $|x_i|$, and thus $X_p^i$ provides a nearly unbiased estimator of $|x_i|^p$. Additionally, we use averaging in the definition of $X_i$ instead of the median to allow for a tractable, though intricate calculation of the $d$-th moment of the sum of the $p$-th powers of $X_i$.

1.1.2 Overview of Lower Bounds

We give an overview for the case of constant $\epsilon$. In both cases we start by applying Yao’s minimax principle for which we fix $S$ and then design a pair of distributions $\alpha$ and $\beta$ which must be distinguished by an $(\epsilon, \delta)$-approximation algorithm for $F_p$. We can also assume the rows of $S$ are orthonormal, since a change of basis to the row space of $S$ can always be applied in post-processing.

Our $\Omega(n^{1-2/p}e^{-2/p}(\log^{2/p}1/\delta)) \log n)$ bound. This is our technically more involved lower bound. We first upper bound the variation distance using the $\chi^2$-divergence as in [4] and work only with the latter. We let $\alpha = N(0, I_n)$ be an $n$-dimensional isotropic Gaussian distribution, while $\beta$ is a distribution formed by sampling an $x \sim N(0, I_n)$, together with a random subset $T \subset [n]$ of size $O(\log(1/\delta))$, and outputting $z = x + \sum_{i \in T}(Cn^{1/p}/t^{1/p})e_i$, where $e_i$ is the $i$-th standard unit vector and $C > 0$ is a constant. For $y \sim \alpha$ and $z \sim \beta$, one can show that with probability $1 - O(\delta)$, one has that $\|z\|_p^p$ is a constant factor larger than $\|y\|_p^p$, since $\|y\|_p^p$ and $\|x\|_p^p$ are concentrated at $\Theta(n)$, while $\sum_{i \in T} C^p n / t = \Theta(n)$.

A common technique in upper bounds, including our own, is the notion of subsampling, whereby a random fraction of roughly $1/2^t$ of the $n$ coordinates are sampled, for each value of $i \in O(\log n)$, and information is then gathered for each $i$ and combined into an overall estimate of $F_p$. We choose our hard distributions so that subsampling does not help. Indeed, if one subsamples half of the coordinates of $z \sim \beta$, with probability $\Omega(\delta)$ all of the coordinates in $T$ will be removed, at which point $z$ is indistinguishable from $y \sim \alpha$. Therefore, our pair of distributions suggests itself as being hard for $(\Theta(1), \delta)$-approximate $F_p$ algorithms.

What drives our analysis is conditioning our distributions on an event $\mathcal{G}$ which only happens with probability $\Omega(\delta)$. Note that for any algorithm which can distinguish samples from $\alpha$ from those from $\beta$ with probability at least $1 - \delta$, it must still have probability $9/10$, say, of distinguishing the distributions given an event $\mathcal{G}$ which occurs for samples drawn from $\beta$. The event $\mathcal{G}$ corresponds to every $i \in T$ having the property that the corresponding column $S_i$ of our sketching matrix $S$ has squared length at most $2r/n$, where $r$ is the number of rows of $S$. By a Markov bound, half of the columns of $S$ have this property, and since $T$ has size $O(\log 1/\delta)$, with probability $\Omega(\delta)$, event $\mathcal{G}$ occurs.

We analyze the $\chi^2$-divergence of the distributions $\alpha$ and $\beta$ conditioned on $\mathcal{G}$. One technique helpful for this is an equality that we show in the full version, which states that for...
a distribution on $\mathbb{R}^n$, that $\chi^2(N(0, I_n) \ast p, N(0, I_n)) = \mathbb{E}[e^{\langle X, X' \rangle}] - 1$, where $X$ and $X'$ are independently drawn from $p$. This equality was used in [4, 29, 35] among other places. In our case, the inner product of $X$ and $X'$ corresponds to an inner product $P$ of two independent random sums of $t$ columns of $S$, restricted to only those columns with squared length at most $2\sqrt{r}/n$. Let the $t$ columns forming $X$ be denoted by $T$ and the $t$ columns forming $X'$ be denoted by $U$.

Critical to our analysis is bounding $\mathbb{E}[P^j]$ for large powers of $j$, as shown in the lemma in the full version. One can think of indexing the rows of $S^T S$ by $T$ and the columns of $S^T S$ by $U$, where $S^T S$ is an $n \times n$ matrix. Let $M$ denote the resulting submatrix. The inner product of interest is then $e_T^T Me_U$, where $e_T = \sum_{i \in T} e_i$ and $e_U = \sum_{i \in U} e_i$.

Our bound, given in the above-referred to lemma in the full version, is very sensitive to minor changes. Indeed, if instead of showing $\mathbb{E}[P^j] \leq \left( \frac{j^2}{n^2} \right) \cdot \left( \frac{16r}{n} \right)^3$, we had shown $\mathbb{E}[P^j] \leq \left( \frac{j^2}{n^2} \right) \cdot \left( \frac{16r}{n} \right)^3$ or $\mathbb{E}[P^j] \leq \left( \frac{j^2}{n^2} \right) \cdot \left( \frac{16r \log n}{n} \right)^3$, our resulting bound for the $\chi^2$-divergence would be larger than 1. For instance, a natural approach is to instead consider $e_T = \sum_{i \in T} \sigma_i e_i$ and $e_U = \sum_{i \in U} \sigma_i e_i$, where the $\sigma_i$ are independent random signs (i.e., $\Pr[\sigma_1 = 1] = \Pr[\sigma_1 = -1] = 1/2$), which would correspond to redefining the distribution $\beta$ above to sample $z = x + \sum_{i \in T} (Cn^{1/\ell}/p^{1/\ell}) \sigma_i e_i$. Without further conditioning the $\sigma_i$ variables, the $\chi^2$-divergence can be as large as $n^{\Theta(\log(1/\delta))}$. This is because with probability roughly $2^{-2^j}$, over the choice of the $\sigma_i$, one has $\sum_{i \in T} \sigma_i e_i$ and $\sum_{i \in U} \sigma_i e_i$ both being very well aligned with the top singular vector of $M$ (if say, $S$ were a random matrix with orthonormal rows), at which point our desired inner product is too large. Instead, by setting all $\sigma_i = 1$, that is, by considering $e_T = \sum_{i \in T} e_i$ and $e_U = \sum_{i \in U} e_i$ as we do, we rule out this possibility.

We prove our lemma by expanding $\mathbb{E}[P^j]$ into a sum of products, each having the form $\prod_{i=1}^j [(S_{a_i}, S_{b_i})]$ where the $S_{a_i}, S_{b_i}$ are columns of $S$. One thing that matters in such products is the multiplicities of duplicate columns that appear in a product. We split the summation by what we call $y$-patterns. We can think of a $y$-pattern as a partition of $\{1, 2, \ldots, j\}$ into $y$ non-empty pieces. We can also define a $z$-pattern as a partition of $\{1, 2, \ldots, j\}$ into $z$ non-empty pieces. We analyze the expectation for a particular pair $P, Q$, where $P$ is a $y$-pattern and $Q$ is a $z$-pattern for some $y, z \in \{1, 2, \ldots, \ell\}$, that is, we only sum over pairs of $j$-tuples $a_1, \ldots, a_j$ and $b_1, \ldots, b_j$ for which for each non-empty piece $\{d_1, \ldots, d_{\ell}\}$ in $P$, where $d_i \in \{1, 2, \ldots, j\}$ for all $i$ and $\ell \leq j$, we have $a_{d_1} = a_{d_2} = \cdots = a_{d_\ell}$. Similarly for each $\{e_1, \ldots, e_m\}$ in $Q$, where $e_i \in \{1, 2, \ldots, j\}$ for all $i$ and $m \leq j$, we have $b_{e_1} = b_{e_2} = \cdots = b_{e_m}$. We also require if $d, d' \in \{1, 2, \ldots, j\}$ are in different pieces of $P$, then $a_d = a_{d'}$. Similarly, if $e, e' \in \{1, 2, \ldots, j\}$ are in different pieces of $Q$, then $b_e = b_{e'}$. Thus, each pair of $j$-tuples is valid for exactly one pair $P, Q$ of patterns.

The valid pairs of $j$-tuples for $P$ and $Q$ define a bipartite multi-graph as follows. In the left partition we create a node for each non-empty piece of $P$, and in the right partition we create a node for each non-empty piece of $Q$. We include an edge from a node $a$ in the left to a node $b$ in the right if $i \in a$ and $i \in b$ for some $i \in \{1, 2, \ldots, j\}$. If there is more than one such $i$, we include an edge with multiplicity corresponding to the number of such $i$. This bipartite graph only depends on $P$ and $Q$. We consider a maximum matching in this multi-graph, and we upper bound the contribution of valid pairs for $P$ and $Q$ based on that matching. By summing over all pairs $P, Q$, we obtain our bound on $\mathbb{E}[P^j]$.

**Our $\Omega(n^{1-2/\rho}e^{-2} \log(1/\delta))$ bound.** This bound uses the same distributions $\alpha$ and $\beta$ as in [30], where an $\Omega(n^{1-2/\rho}e^{-2})$ bound was shown, but we strengthen it to hold for general $\delta$. To do so, we use an exact characterization of the variation distance between multi-variate...
We need a fact about the distance between a Gaussian location mixture to a Gaussian.

We first describe our lower bounds in a little more detail. Due to space constraints, we present a highly abridged version without proofs here (see full version). We defer the problem to distinguishing an isotropic Gaussian from an isotropic Gaussian plus a small multiple of a fixed column of $S$, which typically has small norm since $S$ has orthonormal rows. We then apply a bound as derived above (see full version).

Let $z \sim N(0, I_n)$ be the distribution of $z$, $\bar{S} \cdot y$ with $y \sim \alpha'$ and $\bar{\beta}'$ be the distribution of $S \cdot z$ for $z \sim \bar{\beta}'$.

Our $\Omega(e^{-2} \log(1/\delta))$ bound for $1 \leq p < 2$. This lower bound uses similar techniques to our lower bound of $\Omega(\frac{n^{1-2/p}e^{-2} \log(1/\delta))}$, but considers distinguishing an isotropic Gaussian $N(0, I_n)$ from an $N((1+\epsilon)I_n)$ random variable. Here we set $n = \Theta(e^{-2} \log(1/\delta))$, and show the $p$-norms of samples from the two distributions differ by a $(1+\epsilon)$-factor with probability $1-\delta$. Using that $S$ has orthonormal rows, the images of the two distributions under our sketching matrix $S$ correspond to $N(0, I_r)$ and $N((1+\epsilon)I_r)$, where $r$ is the number of rows of $S$. The result then follows by using the product structure of Hellinger distance.

## 2 Our Lower Bounds

We first describe our lower bounds in a little more detail. Due to space constraints, we present a highly abridged version without proofs here (see full version). We defer both our $\Omega(\frac{n^{1-2/p}e^{-2} \log(1/\delta)})$ lower bound for $p > 2$ and our $\Omega(e^{-2} \log(1/\delta))$ lower bound for $1 \leq p < 2$ entirely to the full version. Here we focus on our lower bound of $\Omega(\frac{n^{1-2/p}e^{-2} /p \log^{2/p}(1/\delta))} \log n)$ for $p > 2$. See also Section 1 for an overview of all of our lower bounds.

We assume $\delta$-Bound 4, which is that $\log(1/\delta) \leq (\frac{n^{1-2/p}e^{-2} /p (\log^{2/p}(1/\delta))} \log n)^{1/4} n^{-c'}$, for a sufficiently small constant $c'>0$. Since $p>2$ is an absolute constant, independent of $n$, this just states that $\delta \geq 2^{-n^{c''}}$ for a sufficiently small constant $c''>0$. There are other bounds - $\delta$-Bound 1, $\delta$-Bound 2, and $\delta$-Bound 3 - see the full version, but these are not assumptions but rather implied by relations between the various parameters (e.g., otherwise the $\Omega(\frac{n^{1-2/p}e^{-2} \log(1/\delta))}$ lower bound is stronger).

Let $p$ and $q$ be probability density functions of continuous distributions. The $\chi^2$-divergence from $p$ to $q$ is $\chi^2(p,q) = \int_x \left( \frac{p(x)}{q(x)} - 1 \right)^2 q(x) dx$.

**Fact 1.** ([33], p.90) For any two distributions $p$ and $q$, we have $D_{TV}(p,q) \leq \sqrt{\chi^2(p,q)}$.

We need a fact about the distance between a Gaussian location mixture to a Gaussian distribution.

**Fact 2.** (p.97 of [21]) Let $p$ be a distribution on $\mathbb{R}^n$. Then $\chi^2(N(0, I_n) * p, N(0, I_n)) = E[e^{X'X'}]-1$, where $X$ and $X'$ are independently drawn from $p$.

Let $T$ be a sample of $t \overset{\text{def}}{=} \log_2(1/\sqrt{\delta})$ coordinates $i \in [n]$ without replacement.

**Case 1:** Suppose $y \sim N(0, I_n)$, and let $\alpha'$ be the distribution of $y$.

**Case 2:** Let $z = x + \sum_{i \in T} \frac{C' y' e^{-x_i}}{t \sqrt{\pi}} e_i$, where $x \sim N(0, I_n)$ and $E_{x \sim N(0, I_{n-t})}$, and $C' > 0$ is a sufficiently large constant. Let $\bar{\beta}'$ be the distribution of $z$.

In the full version we show that for the sketching algorithm to be correct, $D_{TV}(\bar{\alpha}', \bar{\beta}') \geq 1 - 2\delta$, where $\bar{\alpha}'$ is the distribution of $S \cdot y$ for $y \sim \alpha'$ and $\bar{\beta}'$ is the distribution of $S \cdot z$ for $z \sim \bar{\beta}'$. 
Fix an $r \times n$ matrix $S$ with orthonormal rows. Important to our proof will be the existence of a subset $W$ of $n/2$ of the columns for which $\|S_i\|^2 \leq 2r/n$ for all $i \in W$. To see that $W$ exists, consider a uniformly random column $S_i$ for $i \in [n]$. Then $E[\|S_i\|^2] = r/n$ and so by Markov’s inequality, at least a $1/2$-fraction of columns $S_i$ satisfy $\|S_i\|^2 \leq 2r/n$. We fix $W$ to be an arbitrary subset of $n/2$ of these columns.

Suppose we sample $t$ columns of $S$ without replacement, indexed by $T \subset [n]$. Let $G$ be the event that the set $T$ of sampled columns belongs to the set $W$.

► Lemma 3. $\Pr[G] \geq \sqrt{\delta}$.

Let $\alpha_G = \tilde{\alpha} \mid G$ and $\beta_G = \tilde{\beta} \mid G$. By the triangle inequality, $1 - 2\delta \leq D_{TV}(\tilde{\alpha}, \tilde{\beta}) \leq \Pr[G]D_{TV}(\alpha, \beta) + 1 - \Pr[G] \leq \frac{\sqrt{2}}{\sqrt{2}} D_{TV}(\alpha, \beta) + 1 - \frac{\sqrt{2}}{\sqrt{2}}$, which implies that $1 - 4\sqrt{\delta} \leq D_{TV}(\alpha, \beta)$. We can assume $\delta$ is less than a sufficiently small positive constant, and so it suffices to show for sketching dimension $r = o(n^{1-2/p} \log^{2/p} 1/\delta \log n)$, that $D_{TV}(\alpha, \beta) \leq 1/2$. By Fact 1, it suffices to show $\chi^2(\alpha, \beta) \leq 1/4$.

Since $S$ has orthonormal rows, $\alpha'$ is distributed as $N(0, I_t)$. Note that, by definition of $\alpha$, we in fact have $\alpha' = \alpha_G$ since conditioning on $G$ does not affect this distribution. On the other hand, $\beta_G$ is a Gaussian location mixture, that is, it has the form $N(0, I_t) * p$, where $p$ is the distribution of a random variable chosen by sampling a set $T$ subject to event $G$ occurring and outputting $\sum_{i \in T} C_{i+T}^{\ell(p)} E_{m,i-S_i}$. We can thus apply Fact 2 and it suffices to show for

$$r = o(n^{1-2/p} \log^{2/p} 1/\delta \log n)$$

that $E[e^{x_T u}] = \sum_{\alpha, \beta < \infty} E[x_T u]^{\alpha_x \beta_y}$.

The final equality uses that $E_{m-t} = \Theta(n^{1/p})$ and here $O(1)^{2j}$ denotes an absolute constant raised to the $2j$-th power. We can think of $T$ as indexing a subset of rows of $S^T S$ and $U$ indexing a subset of columns. Let $M$ denote the resulting $t \times t$ submatrix of $S^T S$. Then $\sum_{i \in T} S_{i, j \in U} = \sum_{i, j \in [t]} M_{i, j} \leq \sum_{i, j \in [t]} |M_{i, j}| \overset{\text{def}}{=} P$, and we seek to understand the value of $E[P]$ for integers $j \geq 1$.

The following lemma is the key to the argument; its proof is described in Section 1. The proof is based on defining $g$-patterns and looking at matchings in an associated bipartite multi-graph.

► Lemma 4. For integers $j \geq 1$, $E[P] \leq \left(\frac{7}{16}\right)^j \cdot \left(\frac{16r}{n}\right)^j$.

Given the previous lemma, by $\delta$-Bound4, we have $\frac{7}{16} \leq \frac{1}{2^{1/2}}$, and therefore Lemma 4 establishes that $E[P] \leq \frac{1}{2^{1/2}} \cdot \left(\frac{16r}{n}\right)^j$. We thus have, $E\left[\exp\left\{\frac{(c)^{\ell(p)} E_{m,i-S_i}}{\ell(p)} \left(\sum_{i \in T} S_{i, j \in U} S_j\right)\right\}\right] = $
As outlined earlier, the algorithm uses two level-based structures, namely, \( \text{GHSS} \), which is similar to the \( \text{GHSS} \) structure presented in [17], and the shelf structure. The shelf structure is needed only when \( \delta = n^{-\Theta(1)} \), otherwise, the \( \text{GHSS} \) structure suffices. The \( \text{GHSS} \) has \( L + 1 \) levels, corresponding to \( l = 0, 1, \ldots, L \), and the shelf structure has \( J \) shelves numbered \( 0, 1, \ldots, J \). In particular, shelf 0 is identical to \( \text{GHSS} \) level 0.

\section{Algorithm}

As outlined earlier, the algorithm uses two level-based structures, namely, \( \text{GHSS} \), which is similar to the \( \text{GHSS} \) structure presented in [17], and the shelf structure. The shelf structure is needed only when \( \delta = n^{-\Theta(1)} \), otherwise, the \( \text{GHSS} \) structure suffices. The \( \text{GHSS} \) has \( L + 1 \) levels, corresponding to \( l = 0, 1, \ldots, L \), and the shelf structure has \( J \) shelves numbered \( 0, 1, \ldots, J \). In particular, shelf 0 is identical to \( \text{GHSS} \) level 0.

\subsection{Estimating \( F_p \)}

\textbf{GHSS structure.} Corresponding to each \( \text{GHSS} \) level \( l \in \{0, 1, \ldots, L - 1\} \), a pair of \textit{CountSketch} like structures named \( \text{HH}_l = \text{HH}(C_l, s) \) (denoting that the number of buckets per table is \( 16C_l \) and number of independent repetitions is \( s \)) and \( \text{AvgEst}_l = \text{AvgEst}(C_l, 2s) \) are kept. Here, \( s = \Theta(\log(n)) \), recall \( C = n^{1-2/p}(\epsilon^2 \log(1/\delta))/\log(n) + \epsilon^{-4/p}\log^{2/p}(1/\delta)) \), \( C = C_0 = \Theta(p^2 n^{1-2/p}\epsilon^{-4/p}\log^{2/p}(1/\delta)) \) and \( C_l = C_0 \alpha^l \), for \( l = 0, 1, 2, \ldots, L - 1 \), where, \( \alpha = 1 - (1 - 2/p)\nu \) and \( \nu \) is a constant. The number of levels is \( L = \lceil \log_{2\epsilon}(n/C) \rceil \). The final level \( L \) of the \( \text{GHSS} \) structure uses an \( \ell_2/\ell_1 \) deterministic sparse-recovery algorithm [9, 13]. We will show that the number of items that are subsampled into level \( L \) is \( O(C_L) \) with probability \( 1 - O(\delta) \) and therefore from [9, 13], by using \( O(C_L \log(n/C)) \) measurements, these item frequencies are recovered deterministically. Following [17], the \( \text{GHSS} \) structure subsamples the stream hierarchically using independent random hash functions \( g_1, \ldots, g_L : [n] \rightarrow \{0, 1\} \).

\textbf{HH and \textit{AvgEst}} structures. The \( \text{HH}(C_l, s) \) is a \textit{CountSketch} structure [11]. The \( \text{AvgEst}(C_l, 2s) \) structure is similar, except that instead of Rademacher sketches, it uses random \( q \)-th roots of unity sketches, where, \( q = O(\log(1/\delta) + \log n) \). At level \( l \) and for table indexed \( r \in [2s] \), the corresponding hash function is \( h_r : [n] \rightarrow [16C_l] \), and the sketch for bucket index \( b \) is given by \( T_r[b] = \sum_{i \in [n]} x_i \omega_r(i) \), where, \( \{\omega_r(i)\}_{i \in [n]} \) is a random family of \( q \)-th roots of unity that is \( O(\log(1/\delta) + \log n) \)-wise independent. The hash functions across the tables and distinct levels, and the seeds of the family of the random roots of unity, are independent.

\textbf{Shelf structure.} The shelves, indexed from \( j = 0, \ldots, J \), each also consist of an analogous pair of structures, namely, \( \text{HH}(H_j, w_j) \) and \( \text{AvgEst}(H_j, 2w_j) \), where, \( O(H_j) \) is the number of buckets per hash table in these structures, and there are \( O(w_j) \) independent repetitions per structure. The \( \text{AvgEst} \) structures of the shelves also use sketches using \( q \)-th roots of unity, instead of Rademacher sketches. In particular, \( H_0 = C_0 \) and \( w_0 = s \), ensuring that shelf 0 coincides with level 0 of \( \text{GHSS} \). Further, \( H_J = \Theta(n^{1-2/p}\epsilon^{-2}) \) and \( w_J = O(\log(1/\delta)) \). There are two cases, namely, (1) \( H_J = \Omega(H_0) \), or, (2) \( H_J = o(H_0) \). In the first case, \( J = 1 \) and there are only two shelves, considerably simplifying the analysis. The other case \( H_J = o(H_0) \) is more interesting. Here, we let \( H_j = H_0 b^j \), for a parameter \( b < 1 \) and \( b = \Omega(1) \). The table

\[ E[e^{x'r'}] = 1 + \frac{1}{n^{1/\delta}} \sum_{j \geq 1} \frac{O(1)^{\ell_j} (\epsilon^2 \log(1/\delta))^{\ell_j}}{\ell_j!} \cdot \left( x' r' \right)^j \leq 1 + \frac{1}{n^{1/\delta}} \sum_{j \geq 1} \frac{O(1)^{\ell_j} (\epsilon^2 \log(1/\delta))^{\ell_j}}{\ell_j!} \leq 1 + \frac{1}{n^{1/\delta}} \cdot e^{(\log n)} \leq 1 + \frac{1}{\delta}, \text{ since } c > 0 \text{ is an arbitrarily small constant independent of the constant in the } n^{\Omega(1)}. \]

The proof is complete.

For \( 1 \leq p < 2 \), we now show that the sketching dimension is \( \Omega(\epsilon^{-2} \log(1/\delta)) \), which as discussed in Section 1, matches known upper bounds up to a constant factor.

\begin{theorem}
The sketching dimension for \( (\epsilon, \delta) \)-approximating \( F_p \) for \( 1 \leq p < 2 \) is \( \Omega(\epsilon^{-2} \log(1/\delta)) \).
\end{theorem}
Frequency groups, thresholds, estimates and samples. Let $B = \Theta(C)$ and $\bar{\epsilon} = (B/C)^{1/2} = \Theta(1/p)$. Let $\hat{F}_2$ be an estimate for $F_2 = \|x\|^2$ satisfying $F_2 \leq \hat{F}_2 \leq (1 + O(1/p))F_2$ with probability $1 - O(\delta)$. Define frequency thresholds for GHSs as follows: $T_0 = (\hat{F}_2/B)^{1/2}$, $T_l = (2\alpha)^{-1/2}T_0$ and $Q_l = T_l(1 - \bar{\epsilon})$, for $l \in \{0, L - 1\}$. Let $Q_{L} , T_{L} = 0^+$ (i.e., $\alpha \geq T_{L} \text{ iff } a > 0$). For shelf $j = 0, \ldots, J$, let $E_j = x^2H_j$. For shelf $j$, define the frequency threshold $U_j = (\hat{F}_2/E_j)^{1/2}$ and let $U_{j+1} = \infty$. For GHS level indices $l = 0, \ldots, L - 1$, let $\hat{x}_{il}$ denote the estimate for $x_i$ obtained using HH$_1$, and (overloading notation), for shelf indices, $j = 0, \ldots, J$, let $\hat{x}_{ij}$ denote the estimate for $x_i$ obtained from the HH structure of shelf $j$. $\hat{x}_{il}$ denotes the estimate returned from the $\ell_2/\ell_1$ sparse recovery structure at level $L$.

Discovering Items. We say that $i$ is discovered at shelf $j \in [J]$, provided, $(1 - \bar{\epsilon})U_j \leq |\hat{x}_{ij}| \leq (1 + \bar{\epsilon})U_{j+1}$ and $j \in [J]$ is the highest numbered shelf with this property. We say that $i$ is discovered at GHSS level $l \in \{0, \ldots, L\}$, if $l$ is not discovered at any shelf indexed $j \in [J]$, and $l$ is the smallest level such that $T_l(1 - \bar{\epsilon}) < \hat{x}_{il} \leq T_{l-1}(1 + \bar{\epsilon})$. If $i$ is discovered at shelf $j$, then, $i$ is included in the shelf sample $\hat{S}_j$. If $i$ is discovered at level $l \in \{0, 1, \ldots, L\}$ and $|\hat{x}_{il}| \geq T_l$, then, $i$ is included in the level sample $\hat{G}_l$. If $i$ is discovered at level $l$ and $T_l(1 - \bar{\epsilon}) < |\hat{x}_{il}| < T_l$ then, $i$ is placed in $\hat{G}_{l+1}$ iff the random toss of an unbiased coin $K$, lands heads; and upon tails, it is not placed in any sample group. The GHSs level sampling scheme is similar to [17].

The average estimator and nocollison. For each item $i$ included in a group sample $\hat{G}_l$ or shelf sample $\hat{S}_j$, an estimate for $x_i$, $|\hat{x}_{ij}|$, is obtained using the corresponding AvgEst structure of that level or shelf, provided the event NOCOLLISION(i) succeeds. If $i$ is sampled into $\hat{G}_l$, then NOCOLLISION(i) holds if there is a set $R_l(i) \subset [28]$ of table indices of the AvgEst$_l$ structure such that for each $r \in R_l(i), \hat{x}_{ir}$ does not collide under the hash function $h_{r,i}$ with any of the items that are the top-$C_l$ absolute estimated frequencies using HH$_1$. An analogous definition holds if $i$ is included in the $j$th shelf sample. Assuming NOCOLLISION(i) holds, $x_i$ is defined as the average of the estimates obtained from the tables whose indices are in the set $R_l(i)$ (resp. $R_j(i)$ if $i$ was discovered in shelf $j$), that is, $\hat{x}_i = (1/|R_l(i)|)\sum_{r \in R_l(i)}\hat{x}_{ir}$. Further, we check whether $1 - \bar{\epsilon}T_l \leq \hat{x}_i \leq (1 + \bar{\epsilon})T_{l-1}$ (resp. $1 - \bar{\epsilon}U_j \leq \hat{x}_i \leq (1 + \bar{\epsilon})U_{j+1}$, if $i$ is in shelf $j$ sample), otherwise, $i$ is dropped from the sample.

Estimating $F_p$. The estimate for the $p$th frequency moment, $\hat{F}_p$, is the sum of the contribution from the shelf samples $\hat{S}_j, j \in [J]$, and the contribution from the sample groups $\hat{G}_l, l = 0, \ldots, L$. For an item $i \in \hat{G}_l$, let $\hat{d}_l(i)$ be the level at which an item $i$ is discovered. Let $\hat{F}^{\text{shelf}}_p = \sum_{j=1}^{J} \sum \{X_j^p | i \in \hat{S}_j, (1 - \bar{\epsilon})U_j \leq |X_j| \leq (1 + \bar{\epsilon})U_{j+1}\}$, and $\hat{F}^{\text{ghss}}_p = \sum_{l=0}^{L} 2^l \sum \{X_j^p | i \in \hat{G}_l, \hat{d}_l(i) < l, (1 - \bar{\epsilon})T_l \leq X_i \leq (1 + \bar{\epsilon})T_{l-1}\} + 2^L \sum_{\hat{d}_l(i) = L} |\hat{x}_i|^p$. The final estimate is $\hat{F}_p = \hat{F}^{\text{shelf}}_p + \hat{F}^{\text{ghss}}_p$.

3.2 Analysis

Notation. Let $F_2^{\text{hes}}(k)$ be the sum of the squares of all coordinates except the top-$k$ absolute coordinates. For a GHSs level $l \in [L], F_2^{\text{hes}}(l, k)$ is the random $k$-residual second moment of the frequency vector in the sampled sublist $S_l$. Let (1) GOODF$_2 \equiv F_2 \leq \hat{F}_2 \leq (1 + 0.001/(2p))F_2$, (2) SMALLRES$_{2} \equiv F_2^{\text{hes}}(2C_1, l) \leq 1.5F_2^{\text{hes}}(((2\alpha)K_C^2)) / 2^l - 1$, $l = 0, 1, \ldots, L$, (3) SMALLRES$_{\forall l \in \{0, 1, \ldots, L\}} \subseteq \text{SMALLRES}_2$, (4) GOODLEVEL$_2 \equiv \{\hat{f}_l \leq f_j\}$ and $\forall \not \in S_l, (\hat{f}_l, \hat{f}_L = 0)$. We condition the analysis on the “good event” $\mathcal{G} \equiv \text{GOODF}_2 \land \text{SMALLRES} \land \text{GOODLEVEL}_2$, that we show holds with probability $1 - \min(O(\delta), n^{-\Omega(1)})$. widths increase geometrically as $w_j = w_0a^j$, for a parameter $a > 1$. The total measurements used by the shelf structure is $\sum_{j=0}^{J} H_jw_j = H_0w_0\sum_{j=0}^{J} (ab)^j = O(\max(H_0w_0, H_jw_j))$, provided, $|1 - ab| = \Omega(1)$, or, $|\ln(ab)| = \Omega(1)$. The entire stream $S$ is provided as input to each of the shelves $j = 0, 1, \ldots, J$. 

High Probability Frequency Moment Sketches
Lemma 6. $\mathcal{G}$ holds with probability $1 - \min(O(\delta), n^{-\Omega(1)})$.

The range of item frequencies is subdivided into frequency groups, so that each item belongs to exactly one shelf frequency group or to exactly one GHSS frequency group. The frequency group corresponding to the shelf $j$ is $[U_j, U_{j+1})$, for $j = 1, \ldots, J$, where $U_{J+1} = \infty$ and $U_0 = T_0$. The frequency group corresponding to level $l$ of GHSS is $[T_l, T_{l-1})$, where, $T_L = 0$ and $T_{-1} = U_1$. Let $S_j$ (resp. $G_l$) denote the set of items whose frequency belongs to the frequency group corresponding to shelf $j$ (resp. group $l$). A few other events are used in the analysis. If $i \in G_l$, then, $\Pr[\text{NOCOLLISION}(i)] \geq 1 - \exp(-\Theta(\log n))$ as shown in [17] (Lemma 30). If $i \in S_j$, $\Pr[\text{NOCOLLISION}(i)] \geq 1 - \exp(-\Theta(w_j))$. We condition on the following events.

\begin{align*}
\text{(5)} & \quad \text{GOODEST}(i) \equiv \forall l \in [0, \ldots, L], i \in S_l \Rightarrow |\hat{x}_i - x_i| \leq \left(\frac{F_2^\text{res}(2C_l, l)}{C_l}\right)^{1/2} \\
\text{(6)} & \quad \text{ACCUEST}(i) \equiv \forall l \in [0, \ldots, L], i \in S_l \Rightarrow |\hat{x}_i - x_i| \leq \left(\frac{F_2^\text{res}\left((2\alpha)^l C\right)}{(2(2\alpha)^l C)}\right)^{1/2} .
\end{align*}

As shown in [17], (a) GOODEST(i) and ACCUEST(i) each hold with probability $1 - n^{-\Omega(1)}$, and, (b) GOODEST(i) ∧ SMALLRES imply the event ACCUEST(i). For an item $i$ that is discovered at some shelf $j$, GOODEST(i) is the same as ACCUEST(i) and is defined as $|\hat{x}_ij - x_i| \leq \left(\frac{F_2^\text{res}(U_j)}{U_j}\right)^{1/2}$ and holds with probability $1 - \exp(-\Theta(w_j))$.

Lemma 7 extends the approximate 2-wise independence property of the sampling scheme of [17] to an approximate $d$-wise independence property.

Lemma 8 bounds $|X_i - E[X_i]|$ using the 2dth moment method.

Lemma 9 ([22]). $|E[X_i^p] - |x_i|^p| \leq |x_i|^p n^{-\Omega(1)}$.

For $i \in [n]$, let $x_{il}$ be an indicator variable that is 1 iff $i \in S_l$. Let $X_i$ denote $|\hat{x}_i|$ when $l_d(i) = L$ and otherwise, let its meaning be unchanged. Let $z_{il}$ be an indicator variable that is 1 if $i \in G_l$ and 0 otherwise. Define $\tilde{F}_p = \sum_{i \in [n]} Y_i$, where, $Y_i = \sum_{l_d(i) = L} z_{il}^p X_i^p$. Let $H = \mathcal{G} \cap \text{GOODEST} \cap \text{NO_COLLISION}$ and $G' = \text{marg}(G_0) \cup \bigcup_{l=1}^L G_l$.

Lemma 10. Let $B \geq O(n^{-1/4} \log^{1/3}(1/\delta))$. For integral $0 \leq d_1, d_2 \leq \lceil\log(1/\delta)\rceil$, we have, $E[\sum_{i \in G(H)}(Y_i - E[Y_i | H])]d_1 \sum_{i \in G(H)}(Y_i - E[Y_i | H])d_2 \leq \left(\frac{\epsilon F_p}{2d_2}\right)^{d_1 + d_2}$.

3.3 Analysis for the case $\delta \geq n^{-O(1)}$

For the case $\delta = n^{-O(1)}$, the shelf structure is not needed. Redefine the group $G_0$ to correspond to the frequency range $[T_0, \infty)$. The lemmas in this section assume that the family $\{\omega_l(i)\}_{i \in [n]}$ are $O(\log(1/\delta) + \log(n))$-wise independent, and independent across $l, r$ and all hash functions are also $O(\log(1/\delta) + \log(n))$-wise independent.
Lemma 11. Let $1 \leq e, g \leq \lfloor \log(1/\delta) \rfloor, l \in \text{mid}(G_0)$ and $|x_l| \geq \left( \frac{F_{p, e}^{\text{ghss}}(G)}{c_{ghss}} \right)^{1/2}$. Then, $E \left[ (Y_l - E[Y_l | \mathcal{H}])^e \right| Y_l, H \right] \right) \right| \mathcal{H} \right)$ is real and is at most $\left( \frac{a|x_l|^p - b|x_l|^q}{\rho^c} \right)^{c+g/2}$ for some constant $a$. Further, $E \left[ (Y_l - E[Y_l | \mathcal{H}])^e \right| \mathcal{H} \right) \right| \mathcal{H} \right) \leq |x_l|^p n^{O(c)}$.

The calculation of the $d$th central moment for the contribution to $\hat{F}_p$ from the items in mid$(G_0)$ requires an upper bound on the following combinatorial sums. $Q(S_1, S_2) = \sum_{q=1}^{\min(S_1, S_2)} \sum_{\delta_1 + \cdots + \delta_q = S_1} \sum_{\delta_1 + \cdots + \delta_q = S_2} \left( \binom{S_1}{\delta_1} \binom{S_2}{\delta_2} \right)$, and $R(S) = \sum_{\delta_1 + \cdots + \delta_q = S} \sum_{\delta_1 + \cdots + \delta_q = S} \sum_{\delta_1 + \cdots + \delta_q = S} \left( \binom{S}{\delta_1} \binom{S}{\delta_2} \right)$.

Lemma 12. $Q(S_1, S_2) \leq R(S_1 + S_2) \leq (16e(S_1 + S_2)F_{2p-2})^{(S_1 + S_2)}$.

Lemma 13. Let $C \geq O(n^{1-2/p} / \log(n))e^{-2 \log(1/\delta)}$. Then, for $0 \leq d_1, d_2 \leq \log(1/\delta)$, $E \left[ \left( \sum_{i \in \text{mid}(G_0)} \left( Y_i - E[Y_i | \mathcal{H}] \right) \right)^{d_1} \left( \sum_{i \in \text{mid}(G_0)} \left( Y_i - E[Y_i | \mathcal{H}] \right) \right)^{d_2} \right| \mathcal{H} \right) \right| \mathcal{H} \right) \right| \mathcal{H} \right) \right| \mathcal{H} \right) \leq \left( \frac{d_1 + d_2}{d_1 d_2} \right)$. For $\epsilon > 0$, the threshold parameter $\delta$ is an absolute constant.

Lemma 14. Let $C \geq K n^{1-2/p} \log(1/\delta) / \log(n) + L n^{1-2/p} e^{-4/p} \log^{2/p}(1/\delta)$, where $K, L$ are constants. Then, for $d = \lfloor \log(1/\delta) \rfloor$, $E \left[ \left( \sum_{i \in S} \left( Y_i - E[Y_i | \mathcal{H}] \right) \right)^d \right| \mathcal{H} \right) \right| \mathcal{H} \right) \right| \mathcal{H} \right) \right| \mathcal{H} \right) \leq \left( \frac{d_1 + d_2}{d_1 d_2} \right)$. It follows that $Pr \left[ \left| \hat{F}_p - F_p \right| \geq \frac{e}{2} \right] F_p \leq \delta$.

3.4 Analysis for the case $\delta = n^{-\omega(1)}$

We now extend the analysis for failure probability $\delta$ smaller than $n^{-\Theta(1)}$ and up to $\delta = 2^{-n^{\Theta(1)}}$. For the GHSS structure, NOCOLLISION and GOODEST may hold only with probability $1 - n^{-\Theta(1)}$. We first show that the number of items that fail to satisfy NOCOLLISION or GOODEST is at most $O(\log(1/\delta)/\log(n))$ with probability $1 - O(\delta)$. The following lemmas assume the parameter sizes for $B, C, C_l, H_J$ and $H_j$ as described earlier.

Lemma 16. With probability $1 - O(\delta)$, the number of elements for which GOODEST or NOCOLLISION fails is at most $O(\log(1/\delta)/\log(n))$.

Thus, it is possible that legitimate items are not discovered, or are dropped due to collisions, or mistakenly classified and their contribution added to samples. Let Error$^{\text{ghss}}$ denote the total contribution of such items to $\hat{F}_p^{\text{ghss}}$ and let Error$^{\text{sht}}$ denote the error arising in the estimate of $\hat{F}_p^{\text{sht}}$ due to analogous errors. As described earlier, we mainly emphasize the more interesting and complicated case when $H_j = o(H_0)$ (otherwise, $J = 1$).

Lemma 17. Error$^{\text{ghss}} \leq O(e^2 F_p / \log(n))$ and Error$^{\text{sht}} \leq O(\max(e^2 F_p / (\log(n)), O(e^2 F_p))$, each with probability $1 - \delta / n^{O(1)}$.

We first prove a refinement of Lemma 11.

Lemma 18. [Refinement of Lemma 11.] Let $1 \leq e, g \leq \lfloor \log(1/\delta) \rfloor$, $l \in S_j$, and $\log(1/\delta) = \omega(\log(n))$. Assume that ACCUSER$(l)$ holds and $H_j \geq \Omega(p^2 E_j)$ and $|x_l| \geq (F_2/E_j)^{1/2}$. Then, $E \left[ \left( \left( 1 + \frac{x_l}{|x_l|} \right)^{|x_l|} - 1 \right)^p \left( \left( 1 + \frac{x_l}{|x_l|} \right)^{|x_l|} - 1 \right)^q \right| \mathcal{H} \right)$ is real and bounded above by $e^h |x_l|^{-h} (F_2/E_j)^{h/2} \left( \min \left( \frac{h}{h/2}, 1 \right) \right)^{h/2}$, where $h = e + g$ and $c$ is an absolute constant.
Lemma 19 considers the $2d$th central moment of the contribution to $\hat{F}_p$ from all but the outermost shelf, and from the set of outermost shelf items denoted $S_j$, separately. Let $S' = S_1 \cup \ldots \cup S_j$.

- **Lemma 19.** Let $0 \leq d_1, d_2 \leq \lceil \log(1/\delta) \rceil$ and integral and $c_1, c_2$ be constants. Then,
  $$
  \mathbb{E} \left( \sum_{i \in S'} (Y_i - \mathbb{E}[Y_i | H]) \right)^{d_1} \left( \sum_{i \in S_j} (Y_i - \mathbb{E}[Y_i | H]) \right)^{d_2} | H | \leq (c_1 \epsilon F_p)^{d_1 + d_2} \cdot \mathbb{E} \left( \sum_{i \in S_j} (Y_i - \mathbb{E}[Y_i | H]) \right)^{d_1} \left( \sum_{i \in S_j} (Y_i - \mathbb{E}[Y_i | H]) \right)^{d_2} | H | \leq (c_2 \epsilon F_p)^{d_1 + d_2} \cdot
t
  $$

Combining Lemmas 10, 13 and 19 with Lemma 17, we obtain the following.

- **Lemma 20.** \exists constant $c$ s.t. for $1 \leq d \leq \lceil \log(1/\delta) \rceil$, \begin{align*}
  \mathbb{E} \left( \sum_{i \in [n]} (Y_i - \mathbb{E}[Y_i | H]) \right)^d 
  \left( \sum_{i \in [n]} (Y_i - \mathbb{E}[Y_i | H]) \right)^d | H | \leq (c \epsilon F_p)^{2d}.
  \end{align*}

- **Theorem 21.** For each $0 < \epsilon < 1$ and $7/8 \geq \delta > 2^{-n^{1/2}}$ there is a sketching algorithm that $(\epsilon, \delta)$-approximates $F_p$ with sketching dimension $O(n^{1-2/p}(\epsilon^{-2} \log(1/\delta)) + \epsilon^{-2/p} \log^{2/p}(1/\delta) \log(n))$ and update time (per stream update) $O((\log n) \log(1/\delta))$.

References

Yi Li and David Woodruff. “A Tight Lower Bound for High Frequency Moment Estimation with Small Error”. In *Proceedings of International Workshop on Randomization and Computation (RANDOM)*, 2013.


