Composable Sketches for Functions of Frequencies: Beyond the Worst Case

Edith Cohen 1 2  Ofir Geri 3 4  Rasmus Pagh 1 5 6

Abstract

Recently there has been increased interest in using machine learning techniques to improve classical algorithms. In this paper we study when it is possible to construct compact, composable sketches for weighted sampling and statistics estimation according to functions of data frequencies. Such structures are now central components of large-scale data analytics and machine learning pipelines. However, many common functions, such as thresholds and pth frequency moments with \( p > 2 \), are known to require polynomial size sketches in the worst case. We explore performance beyond the worst case under two different types of assumptions. The first is having access to noisy advice on item frequencies. This continues the line of work of Hsu et al. (ICLR 2019), who assume predictions are provided by a machine learning model. The second is providing guaranteed performance on a restricted class of input frequency distributions that are better aligned with what is observed in practice. This extends the work on heavy hitters under Zipfian distributions in a seminal paper of Charikar et al. (ICALP 2002). Surprisingly, we show analytically and empirically that “in practice” small polylogarithmic-size sketches provide accuracy for “hard” functions.

1. Introduction

Composable sketches, also known as mergeable summaries (Agarwal et al., 2013), are data structures that support summarizing large amounts of distributed or streamed data with small computational resources (time, communication, and space). Such sketches support processing additional data elements and merging sketches of multiple datasets to obtain a sketch of the union of the datasets. This design is suitable for working with streaming data (by processing elements as they arrive) and distributed datasets, and allows us to parallelize computations over massive datasets. Sketches are now a central part of managing large-scale data, with application areas as varied as federated learning (McMahan et al., 2017) and statistics collection at network switches (Liu et al., 2016; 2019).

The datasets we consider consist of elements that are key-value pairs \((x, v)\) where \( v \geq 0 \). The frequency \( w_x \) of a key \( x \) is defined as the sum of the values of elements with that key. When the value of all elements is 1, the frequency is simply the number of occurrences of a key in the dataset. Examples of such datasets include search queries, network traffic, user interactions, or training data from many devices. These datasets are typically distributed or streamed.

Given a dataset of this form, one is often interested in computing statistics that depend on the frequencies of keys. For example, the statistics of interest can be the number of keys with frequency greater than some constant (threshold function), or the second frequency moment \( \sum_x w_x^2 \), which can be used to estimate the skew of the data. Generally, we are interested in statistics of the form

\[
\sum_x L_x f(w_x)
\]

where \( f \) is some function applied to the frequencies of the keys and the coefficients \( L_x \) are provided (for example as a function of the features of the key \( x \)). An important special case, popularized in the seminal work of (Alon et al., 1999), is computing the \( f \)-statistics of the data:

\[
\|f(w)\|_1 = \sum_x f(w_x).
\]

One way to compute statistics of the form (1) is to compute a random sample of keys, and then use the sample to compute estimates for the statistics. In order to compute low-error estimates, the sampling has to be weighted in a way that depends on the target function \( f \); each key \( x \) is weighted by \( f(w_x) \). Since the problem of computing a weighted sample is more general than computing \( f \)-statistics, our focus in this work will be on composable sketches for weighted sampling according to different functions of frequencies.
When estimating statistics from samples, the accuracy depends on the sample size and how much the sampling probabilities "suit" the statistics we are estimating. In order to minimize the error, the sampling probability of each key \( x \) should be (roughly) proportional to \( f(w_x) \). This leads to a natural and extensively-studied question: for which functions \( f \) can we design efficient sampling sketches?

The literature and practice are ripe with surprising successes for sketching, including small (polylogarithmic size) sketch structures for estimating the number of distinct elements \((\text{Flajolet & Martin, 1985; Flajolet et al., 2007})\) \((f(w) = I_{w>0})\), frequency moments \((f(w) = w^p)\) for \( p \in [0, 2] \) \((\text{Alon et al., 1999; Indyk, 2001})\), and computing \( \ell_p \) heavy hitters \((p \leq 2)\) \((\text{an} \ell_p \text{-heavy hitter is a key} x \text{ with} w > T \text{ for} x \geq 0, \text{where an} \ell_p \text{-heavy hitter is a key} x \text{ with} w_p > 0 \text{ for} p \leq 2 \text{, where an} \ell_p \text{-heavy hitter is a key} x \text{ with} w > T \text{ for} x \geq 0, \text{where an} \ell_p \text{-heavy hitter is a key} x \text{ with} w > T \text{ for} x \geq 0)\) \((\text{Misra & Gries, 1982; Charikar et al., 2002; Manku & Motwani, 2002})\), and computing \( \ell_p \) heavy hitters \((p \leq 2)\) \((\text{an} \ell_p \text{-heavy hitter is a key} x \text{ with} w > T \text{ for} x \geq 0, \text{where an} \ell_p \text{-heavy hitter is a key} x \text{ with} w > T \text{ for} x \geq 0, \text{where an} \ell_p \text{-heavy hitter is a key} x \text{ with} w > T \text{ for} x \geq 0)\) \((\text{Cormode & Muthukrishnan, 2005; Metwally et al., 2005})\). Here \( I_x \) is the indicator function that is 1 if the predicate \( \sigma \) is true, and 0 otherwise. A variety of methods now support sampling via small sketches for rich classes of functions of frequencies \((\text{Cohen, 2018; McGregor et al., 2016; Jayaram & Woodruff, 2018; Cohen & Geri, 2019})\), including the moments \( f(w) = w^p \) for \( p \in [0, 2] \) and the family of concave sublinear functions.

The flip side is that we know of lower bounds that limit the performance of sketches using small space for some fundamental tasks \((\text{Alon et al., 1999})\). A full characterization of functions \( f \) for which \( f \)-statistics can be estimated using polylogarithmic-size sketches was provided in \((\text{Braverman & Ostrovsky, 2010})\). Examples of "hard" functions are thresholds \( f(w) = I_{w>T} \) \((\text{counting the number of keys with frequency above a specified threshold value} T)\), threshold weights \( f(w) = w I_{w>T} \), and high frequency moments \( f(w) = w^p \) with \( p > 2 \). Estimating the \( p \)-th frequency moment \( \sum w^p \) for \( p > 2 \) is known to require space \( \Omega(n^{1-2/p}) \) \((\text{Alon et al., 1999; Li & Woodruff, 2013})\), where \( n \) is the number of keys. These particular functions are important for downstream tasks: Threshold aggregates characterize the distribution of frequencies, and high moment estimation is used in the method of moments, graph applications \((\text{Eden et al., 2019})\), and for estimating the cardinality of multi-way self-joins \((\text{Alon et al., 2002})\) \((\text{a} p \text{-th moment is used for estimating a} p \text{-way join})\). Beyond the worst case. Much of the discussion of sketching classified functions into "easy" and "hard". For example, there are known efficient methods for sampling according to \( f(w) = w^p \) for \( p \leq 2 \), while for \( p > 2 \), even the easier task of computing the \( p \)-th moment is known to require polynomial space. However, the hard data distributions used to establish lower bounds for some functions of frequency are arguably not very realistic. Real data tends to follow nice distributions and is often (at least somewhat) predictable. We study sketching where additional assumptions allow us to circumvent these lower bounds while still providing theoretical guarantees on the quality of the estimates. We consider two distinct ways of going beyond the worst case: 1) accessing advice models, and 2) making natural assumptions on the frequency distribution of the dataset.

For the sampling sketches described in this paper, we use a notion of overhead to capture the discrepancy between the sampling probabilities used in the sketch and the "ideal" sampling probabilities of weighted sampling according to a target function of frequency \( f \). An immensely powerful property of using sampling to estimate statistics of the form \((1)\) is that the overhead translates into a multiplicative increase in sample/sketch size, without compromising the accuracy of the results (with respect to what an ideal "benchmark" weighted sample provides). This property was used in different contexts in prior work, e.g., \((\text{Frieze et al., 2004; Cohen et al., 2009})\), and we show that it can be harnessed for our purposes as well. For the task of estimating \( f \)-statistics, we use a tailored definition of overhead, that is smaller than the overhead for the more general statistics \((1)\).

Advice model. The advice model for sketching was recently proposed and studied by Hsu et al. \((2019)\). The advice takes the form of an oracle that is able to identify whether a given key is a heavy hitter. Such advice can be generated, for example, by a machine learning model trained on past data. The use of the "predictability" of data to improve performance was also demonstrated in \((\text{Kraska et al., 2018; Indyk et al., 2019})\). A similar heavy hitter oracle was used in \((\text{Jiang et al., 2020})\) to study additional problems in the streaming setting. For high frequency moments, they obtained sketch size \( O(n^{1/2-1/p}) \), a quadratic improvement over the worst-case lower bound.

Here we propose a sketch for sampling by advice. We assume an advice oracle that returns a noisy prediction of the frequency of each key. This type of advice oracle was used in the experimental section of \((\text{Hsu et al., 2019})\) in order to detect heavy hitters. We show that when the predicted \( f(w_x) \) for keys with above-average contributions \( f(w_x) \) is approximately accurate within a factor \( C \), our sample has overhead \( O(C) \). That is, the uncertainty in the advice translates to a factor \( O(C) \) increase in the sketch size but does not impact the accuracy.
Frequencies-functions combinations. Typically, one designs sketch structures to provide guarantees for a certain function \( f \) and any set of input frequencies \( w \). The performance of a sketch structure is then analyzed for a worst-case frequency distribution. The analysis of the advice model also assumes worst-case distributions (with the benefits that comes from the advice). We depart from this and study sketch performance for a combination \((F, W, h)\) of a family of functions \( F \), a family \( W \) of frequency distributions, and an overhead factor \( h \). Specifically, we seek sampling sketches that produce weighted samples with overhead at most \( h \) with respect to \( f(w) \) for every function \( f \in F \) and frequency distribution \( w \in W \). By limiting the set \( W \) of input frequency distributions we are able to provide performance guarantees for a wider set \( F \) of functions of frequency, including functions that are worst-case hard. We particularly seek combinations with frequency distributions \( W \) that are typical in practice. Another powerful property of the combination formulation is that it provides multi-objective guarantees with respect to a multiple functions of frequency \( F \) using the same sketch (Cohen, 2015; 2018; Liu et al., 2016).

The performance of sketch structures on “natural” distributions was previously considered in a seminal paper by Charikar et al. (2002). The paper introduced the Count Sketch structure for heavy hitter detection, where an \( \ell_2 \) \( \varepsilon \)-heavy hitter is defined to be a key with \( w_x^2 \geq \varepsilon \|w\|_2^2 \). They also show that for Zipf-distributed data with parameter \( \alpha \geq 1/2 \), a count sketch of size \( O(k) \) can be used to find the \( k \) heaviest keys (a worst-case hard problem) and that an \( \ell_1 \) sample can only identify the heaviest keys for Zipf parameter \( \alpha \geq 1 \).

We significantly extend these insights to a wider family of frequency distributions and to a surprisingly broad class of functions of frequencies. In particular we show that all high moments \( (p \geq 2) \) are “easy” as long as the frequency distribution has an \( \ell_1 \) or \( \ell_2 \) \( \varepsilon \)-heavy hitter. In this case, an \( \ell_1 \) or \( \ell_2 \) sample with overhead \( 1/\varepsilon \) can be used to estimate all high moments. We also show that in a sense this characterization is tight in that if we allow all frequencies, we meet the known lower bounds. It is very common for data sets in practice to have a most frequent key that is an \( \ell_1 \) or \( \ell_2 \) \( \varepsilon \)-heavy hitter. This holds in particular for Zipf or approximate Zipf distributions.

Moreover, we show that Zipf frequency distributions have small universal sketches that apply to any monotone function of frequency (including thresholds and high moments). Zipf frequencies were previously considered in the advice model (Aamand et al., 2019). Interestingly, we show that for these distribution a single small sketch is effective with all monotone functions of frequency, even without advice. In these cases, universal sampling is achieved with off-the-shelf polylogarithmic-size sketches such as \( \ell_p \) samples for \( p \leq 2 \) and multi-objective concave-sublinear samples (Cohen et al., 2012; Cohen, 2018; McGregor et al., 2016; Jayaram & Woodruff, 2018).

Empirical study. We complement our analysis with an empirical study on multiple real-world datasets including datasets studied in prior work on advice models (Pass et al., 2006; CAIDA, 2016; Paranjape et al., 2017; Maciá-Fernández et al., 2018). (Additional discussion of the datasets appears in the supplementary material.) We apply sampling by advice, with advice based on models from prior work or direct use of frequencies from past data. We then estimate high moments from the samples. We observe that sampling-by-advice was effective on these datasets, yielding low error with small sample size. We also observed however that \( \ell_2 \) and \( \ell_1 \) samplers were surprisingly accurate on these tasks, with \( \ell_2 \) samplers generally outperforming sampling by advice. This surprisingly good performance of these simple sampling schemes is suggested from our analysis.

We compute the overhead factors for some off-the-shelf sampling sketches on multiple real-world datasets with the objectives of \( \ell_p \) sampling \( (p > 2) \) and universal sampling. We find these factors to be surprisingly small. For example, the measured overhead of using \( \ell_2 \) sampling for the objective of \( \ell_p \) sampling is in the range \([1.2, 18]\). For universal sampling, the observed overhead is lower with \( \ell_1 \) and with multi-objective concave sublinear samples than with \( \ell_2 \) sampling and is in \([93, 800]\), comparing very favorably with the alternative of computing a full table. Finally, we use sketches to estimate the distribution of rank versus frequency, which is an important tool for optimizing performance across application domains (for network flows, files, jobs, or search queries). We find that \( \ell_1 \) samples provide quality estimates, which is explained by our analytical results.

2. Preliminaries

We consider datasets where each data element is a (key, value) pair. The keys belong to a universe denoted by \( \mathcal{X} \) (e.g., the set of possible users or words), and each key may appear in more than one element. The values are positive, and for each key \( x \in \mathcal{X} \), we define its frequency \( w_x \geq 0 \) to be the sum of values of all elements with key \( x \). The data elements may appear as a stream or be stored in a distributed manner. We denote the number of active keys (keys with frequency greater than 0) by \( n \).

We are interested in sketches that produce a weighted sample of keys according to some function \( f \) of their frequencies, where roughly key \( x \) is sampled with probability proportional to \( f(w_x) \). We use \( f(w) \) as a shorthand for the vector...
of all values \( f(w_x) \) (in any fixed order).

**Estimates from a sample.** We work with sampling schemes that produce a random subset \( S \subseteq \mathcal{X} \) of the keys in the dataset. For each key \( x \in S \) we have its frequency \( w_x \) and can compute its probability \( p_x \) to be sampled. From such a sample, we can compute for each key \( x \) the inverse probability estimate (Horvitz & Thompson, 1952) of \( f(w_x) \) defined as

\[
\hat{f}(w_x) = \begin{cases} 
\frac{f(w_x)}{p_x} & \text{if } x \in S \\
0 & \text{if } x \notin S
\end{cases}.
\]

These unbiased per-key estimates can be summed to obtain unbiased estimates of the \( f \)-statistics of a domain \( H \subset \mathcal{X} \):

\[
\sum_{x \in H} \hat{f}(w_x) := \sum_{x \in H} \hat{f}(w_x) = \sum_{x \in S} \hat{f}(w_x).
\]

The last equality follows because \( \hat{f}(w_x) = 0 \) for keys not in the sample. We can similarly estimate other statistics that are linear in \( \hat{f}(w_x) \), e.g., \( \sum_{x \in \mathcal{X}} L_x f(w_x) \) (for coefficients \( L_x \)).

**Benchmark variance bounds.** We measure performance with respect to that of a “benchmark” weighted sampling scheme where each key \( x \) is sampled with probability proportional to \( f(w_x) \). Recall that for “hard” functions \( f \) these schemes can not be implemented with small sketches. These benchmark schemes include (i) probability proportional to size (pps) with replacement where we have \( k \) independent draws where key \( x \) is selected with probability \( f(w_x)/\|f(w)\|_1 \), (ii) pps without replacement (ppswor (Rosén, 1972; 1997; Cohen & Kaplan, 2008)), or (iii) priority sampling (Ohlsson, 1998; Duffield et al., 2007). With these schemes, the variance for key \( x \) is upper bounded by

\[
\text{Var}[\hat{f}(w_x)] \leq \frac{1}{k} f(w_x)\|f(w)\|_1 
\]

Consequently, the variance of the sum estimator for the \( f \)-statistics of a domain \( H \) is bounded by (due to nonpositive covariance shown in earlier works):

\[
\text{Var} \left[ \sum_{x \in H} \hat{f}(w_x) \right] \leq \frac{1}{k} \sum_{x \in H} f(w_x)\|f(w)\|_1.
\]

The variance on the estimate of \( \|f(w)\|_1 \) is bounded by

\[
\text{Var}[\|f(w)\|_1] \leq \frac{1}{k} \|f(w)\|_1^2.
\]

With these “benchmark” schemes, if we wish to get multiplicative error bound (normalized root mean squared error) of \( \varepsilon \) for estimating \( \|f(w)\|_1 \) we need sample size \( k = O(\varepsilon^{-2}) \). We note that the estimates are also concentrated in the Chernoff sense (Duffield et al., 2007; Cohen, 2015).

We refer to the probability vector

\[
p_x := \frac{f(w_x)}{\|f(w)\|_1}
\]

as pps sampling probabilities for \( f(w) \). When \( f(w) = w^p \) (for \( p > 0 \)) we refer to sampling with the respective pps probabilities as \( \ell_p \) sampling.

**Emulating a weighted sample.** Let \( p \) be the base pps probabilities for \( f(w) \). When we use a weighted sampling scheme with weights \( q \neq p \) then the variance bound (3) does not apply. We will say that weighted sampling according to \( q \) emulates weighted sampling according to \( p \) with overhead \( s \) if for all \( k \) and for all \( H \), a sample of size \( ks \) provides the variance bound (3) (and the respective concentration bounds).

**Lemma 2.1.** The overhead of emulating weighted sample according to \( p \) using weighted sampling according to \( q \) is at most

\[
h(p, q) := \max \frac{p_x}{q_x}.
\]

**Proof.** We first bound the variance of \( \hat{f}(w_x) \) for a key \( x \) with weighted sampling by \( q \). Consider a weighted sample of size \( k \) according to base probabilities \( q_x \). Then for all \( x \),

\[
\text{Var}[\hat{f}(w_x)] \leq \frac{1}{k} (f(w_x))^2 \left( \frac{1}{p_x} - 1 \right) \leq \frac{1}{k} (f(w_x))^2 \frac{1}{p_x} \frac{p_x}{q_x} \leq \frac{1}{k} f(w_x)\|f(w)\|_1 \frac{p_x}{q_x}.
\]

The upper bound on the variance for a domain \( H \) is:

\[
\frac{1}{k} \left( \sum_{x \in H} f(w_x) \right) \|f(w)\|_1 \max \frac{p_x}{q_x}.
\]

Thus for any \( H \), the variance bound (4) is larger than the benchmark bound (3) by at most a factor of \( h(p, q) \).

Note that the inaccuracy in the probabilities (using \( q \) instead of \( p \)) is compensated for by a larger sample size without compromising accuracy of the estimate.
Remark 2.2. The emulation overhead can be interpreted as providing an upper bound over all possible estimation tasks that the emulated sample could be used for. This definition of overhead is tight if we wish to transfer guarantees for all subsets $H$: Consider a (subset that is a single) key $x$ and sample size $k$ such that $q_x \leq p_x < 1/k$. The variance when sampling according to $q$ is $\frac{(f(w_x))^2}{(1/(kq_x))} - 1 = \frac{(f(w_x))^2}{(kq_x)} \approx \frac{1}{k} f(w_x) \frac{f(w_x)}{p_x} q_x = \frac{1}{k} f(w_x) \frac{f(w_x)}{p_x} q_x$. This is a factor of $p_x/q_x$ larger than the variance when sampling according to $p_x = f(w_x)/\|f(w)\|_1$, which is $\approx (f(w_x))^2/(kp_z) = \frac{1}{k} f(w_x) \|f(w)\|_1$.

Overhead for estimating $f$-statistics. If we are only interested in estimates of the full $f$-statistics $\|f(w)\|_1$, the overhead reduces to the expected ratio $E_{x \sim \mathbb{P}}[p_x/q_x]$ instead of the maximum ratio.

Corollary 2.3. Let $p$ be the base pps probabilities for $f(w)$. Consider weighted sampling of size $k$ according to $q$. Then,

$$\text{Var} \left[ \|f(w)\|_1 \right] \leq \frac{\|f(w)\|_1^2}{k} \sum_x p_x \frac{p_x}{q_x} = \frac{\|f(w)\|_1^2}{k} E_{x \sim \mathbb{P}}[p_x/q_x].$$

3. The Advice Model

In this section, we assume that in addition to the input, we are provided with an oracle access to an “advice” model. When presented with a key $x$, the advice model returns a prediction $a_x$ for the total frequency of $x$ in the data. For simplicity, we assume that predictions are the same for all queries with the same key. This model is similar to the model used in a recent paper about frequency estimation (Hsu et al., 2019).

A detailed description of our sketch structure and the corresponding estimators (including proofs) is provided in the supplementary material. At a high level, our sampling sketch takes size parameters $(k_h, k_p, k_u)$, maintains a set of at most $k_h + k_p + k_u$ keys, and collects the exact frequencies $w_x$ for these stored keys. The primary component of the sketch is a weighted-sample-by-advice of size $k_p$. Our sketch stores keys according to two additional criteria in order to provide robustness to the prediction quality of the advice:

- The top-$k_h$ keys by advice. This provides tolerance to inaccuracies in the advice for these heaviest keys. Since these keys are included with probability 1, they will not contribute to the error.
- A uniform sample of $k_u$ keys. This allows keys that are “below average” in their contribution to $\|f(w)\|_1$ to be represented appropriately in the sample, regardless of the accuracy of the advice. This provides robustness to the accuracy of the advice on these very infrequent keys and ensures they are not undersampled. Moreover, this ensures that all active keys ($w_x > 0$), including those with potentially no advice ($a_x = 0$), have a positive probability of being sampled. This is necessary for unbiased estimation.

We provide an unbiased estimator that smoothly combines the different sketch components and provides the following guarantees:

Lemma 3.1. Suppose the advice model is such that for some $c_p, c_u \geq 0$ and $h \geq 0$, all keys $x$ that are active ($w_x > 0$) and not in the $h$ largest advice values of active keys ($a_x \leq \{a_y \mid w_y > 0\}_{(n-h+1)}$) satisfy

$$\frac{f(w_x)}{\|f(w)\|_1} \leq \max\{c_p \frac{f(a_x)}{\|f(a)\|_1}, c_u \frac{1}{n}\}.$$ 

Then for all $k \geq 1$, a sample with $(k_h, k_p, k_u) = (h, \lceil kc_p \rceil + 2, \lceil kc_u \rceil + 2)$ satisfies the variance bound (3) for all $H$.

In particular, if our advice is approximately accurate, say $f(w_x) \approx f(a_x) \leq c_p \cdot f(w_x)$, the overhead when sampling by advice is $c_p$.

Corollary 3.2. Let $f$ be such that $f(w_x) \leq f(a_x) \leq c_p f(w_x)$ then for all $k \geq 1$, with sample size $(k_h, k_p, k_u) = (0, \lceil kc_p \rceil + 2, 0)$ we have $\text{Var} \left[ \|f(w)\|_1 \right] \leq \frac{1}{k} \|f(w)\|_1^2$.

Experiments. We evaluate the effectiveness of “sampling by advice” for estimating the frequency moments with $p = 3, 7, 10$ on datasets from (Pass et al., 2006; Paranjape et al., 2017). We use advice models from prior work (Hsu et al., 2019) based on a machine learning algorithms applied to past data and advice based directly on frequencies in past data. Some representative results are reported in Figure 1 and additional results and more details are provided in the supplementary material. The results reported for sampling by advice are with $k_h = 0$ and two choices of balance between the ppswor sample based on the advice and the uniform sample: $k_p = k_u$ and $k_u = 32$. We also report performance of ppswor ($\ell_1$ sampling without replacement), $\ell_2$ sampling (with and without replacement), and the benchmark upper bound $(1/\sqrt{k})$.

We observe that for the third moment, sampling by advice did not perform significantly better than ppswor (and sometimes performed even worse). For higher moments, however, sampling by advice performed better than ppswor when the sample size is small. With replacement $\ell_2$ sampling was more accurate than advice (without replacement $\ell_2$ sampling performs best). Our analysis in the next section explains the perhaps surprisingly good performance of $\ell_1$ and $\ell_2$ sampling schemes.
4. Frequencies-Functions Combinations

In this section, we analyze performance for inputs that comes from restricted families of frequency distributions $W$. Restricting the family $W$ of possible inputs allows us to extend the family $F$ of functions of frequency that can be efficiently emulated with a small overhead.

Specifically, we will consider sampling sketches and corresponding combinations $(W, F, h)$ of frequency vectors $W$, functions of frequency $F$, and overhead $h \geq 1$ so that for every frequency distribution $w \in W$ and frequency function $f \in F$, our sampling sketch emulates a weighted sample of $f(w)$ with overhead $h$. We will say that our sketch supports the combination $(W, F, h)$.

Recall that emulation with overhead $h$ means that a sampling sketch of size $h\varepsilon^{-2}$ (i.e., holding this number of keys or hashes of keys) provides estimates with NRMSE $\varepsilon$ for $f(w)$ for any $f \in F$ and that these estimates are concentrated in the Chernoff bound sense. Moreover, for any $f \in F$ we can estimate statistics of the form (1) with the same guarantees on accuracy as provided by a dedicated weighted sample according to $f$.

We study combinations supported by off-the-shelf sampling schemes that can be implemented with small (e.g., polylogarithmic) size sketches: \( \ell_q \) sampling for $q \leq 2$, ppswor (\( \ell_1 \) sampling without replacement) (Gibbons & Matias, 1998; Estan & Varghese, 2002; Cohen et al., 2012), \( \ell_2 \) samplers (with replacement) (Indyk, 2001; McGregor et al., 2016; Jayaram & Woodruff, 2018), and multi-objective concave-sublinear sampling (Cohen, 2018). We will use the notation $w = \{w_i\}$ for dataset frequencies, where $w_i$ is the frequency of the $i$th most frequent key.

We report results of experiments on datasets listed in Table 1 with details provided in the supplementary material.

4.1. Emulating an \( \ell_p \) Sample by an \( \ell_q \) Sample

We express the overhead of emulating \( \ell_p \) sampling by \( \ell_q \) sampling ($p \geq q$) in terms of the properties of the frequency distribution. Recall that \( \ell_p \) sampling (and estimates of $p$-th frequency moments) can be implemented with polylogarithmic size sketches for $p \leq 2$ but requires polynomial size sketches in the worst case when $p > 2$.

**Lemma 4.1.** Consider a dataset with frequencies $w$ (in nonincreasing order). For $p \geq q$, the overhead of emulating \( \ell_p \) sampling by \( \ell_q \) sampling is bounded by

\[
\frac{\|w\|_p^q}{\|w\|_p^q} \leq \frac{\|w\|_q^p}{\|w\|_q^p}.
\]

\[\text{(5)}\]

**Proof.** The sampling probabilities for key $i$ under \( \ell_p \) sampling and \( \ell_q \) sampling are $w_i^p/\|w\|_p^p$ and $w_i^q/\|w\|_q^q$, respectively. Then, the overhead of emulating \( \ell_p \) sampling by \( \ell_q \) sampling is

\[
\max_i w_i^p/\|w\|_p^p = \max_i w_i^{p-q} \cdot \|w\|_q^p/\|w\|_q^p = w_i^{p-q} \cdot \|w\|_q^p/\|w\|_q^p.
\]

We can obtain a (weaker) upper bound on the overhead, expressed only in terms of $q$, that applies to all $p \geq q$:

**Corollary 4.2.** The overhead of emulating \( \ell_p \) sampling using \( \ell_q \) sampling (for any $p \geq q$) is at most $\|w\|_q^q/\|w\|_q^p$.

**Remark 4.3.** Emulation works when $p \geq q$. When $q > p$, the maximum in the overhead bound (see proof of Lemma 4.1) is incurred on the least frequent key, with frequency $w_{n}$. We therefore get a bound of $\|w\|_q^q/\|w\|_q^p$ and Corollary 4.2 does not apply.

4.2. Frequency Distributions with a Heavy Hitter

We show that for distributions with an \( \ell_q \) heavy hitter, \( \ell_q \) sampling emulates \( \ell_p \) sampling for all $p \geq q$ with a small overhead.

**Definition 4.4.** Consider frequencies $w$. An \( \ell_q \) $\phi$-heavy hitter is defined to be a key such that $w_i^q \geq \phi \cdot \|w\|_q^q$.

We rephrase Corollary 4.2 in terms of a presence of a heavy hitter:

**Corollary 4.5.** Let $w$ be a frequency vector with a $\phi$-heavy hitter under \( \ell_q \). Then for $p \geq q$, the overhead of using \( \ell_q \) sample to emulate an \( \ell_p \) sample is at most $1/\phi$.

We are now ready to specify combinations $(W, F, h)$ of frequency vectors $W$, functions of frequency $F$, and overhead $h \geq 1$ that are supported by \( \ell_q \) sampling.
Theorem 4.6. For any $q > 0$ and $\phi \in (0, 1]$, an $\ell_q$-sample supports the combination

$$ W := \{ w \text{ with an } \ell_q \text{-heavy hitter} \} $$

$$ F := \{ f(w) = w^p \mid p \geq q \}^+ $$

$$ h := 1/\phi $$

where the notation $\overline{F}_+$ is the closure of a set $F$ of functions under nonnegative linear combinations.

In particular, if the input distribution has an $\ell_q$-heavy hitter then $\ell_q$ sampling of size $\varepsilon^{-2}/\phi$ emulates an $\ell_p$ sampling of size $\varepsilon^{-2}$ for any $p > q$.

**Table 1. Datasets**

<table>
<thead>
<tr>
<th>Dataset</th>
<th>$n/10^9$</th>
<th>$\ln n$</th>
<th>$\ell_1$ HH</th>
<th>$\ell_2$ HH</th>
<th>$\alpha$</th>
</tr>
</thead>
<tbody>
<tr>
<td>SO</td>
<td>2.23</td>
<td>14.6</td>
<td>0.00016</td>
<td>0.00010</td>
<td>0.053_2</td>
</tr>
<tr>
<td>AOL</td>
<td>2.30</td>
<td>14.6</td>
<td>0.00153</td>
<td>0.00099</td>
<td>0.107_36</td>
</tr>
<tr>
<td>CAIDA</td>
<td>1.07</td>
<td>13.9</td>
<td>0.00333</td>
<td>0.00322</td>
<td>0.048_94</td>
</tr>
<tr>
<td>UGR</td>
<td>7.38</td>
<td>18.2</td>
<td>0.11188</td>
<td>0.04010</td>
<td>0.850_109</td>
</tr>
</tbody>
</table>

Table 1 reports properties and the relative $\ell_1$ and $\ell_2$ weights of the most frequent key for our datasets. We can see that the most frequent key is a heavy hitter with $1/\phi \leq 21$ for $\ell_2$ and $1/\phi \leq 625$ for $\ell_1$ which gives us upper bounds on the overhead of emulating any $\ell_p$ sample ($p \geq 2$). Table 3 reports (for $p = 3, 7, 10$) the overhead of emulating the respective $\ell_p$ sample and the (smaller) overhead of estimating the $p$th moment. We can see that high moments can be estimated well from $\ell_2$ and with a larger overhead from $\ell_1$ samples.

**Certified emulation.** The quality guarantees of a combination $(W, F, h)$ are provided when $w \in W$. In practice, however, we may compute samples of arbitrary dataset frequencies $w$. Conveniently, we are able to test the validity of emulation by considering the most frequent key in the sample: For an $\ell_q$ sample of size $k$ we can compute $r \leftarrow \max x \in S w_x / \|w\|_q^2$ and certify that our sample emulates $\ell_p$ samples ($p > q$) of size $kr$. If $kr$ is small, then we do not provide meaningful accuracy but otherwise we can certify the emulation with sample size $kr$. When the input $w$ has an $\ell_q$-heavy hitter then an $\ell_q$ sample of size $k$ will include it with probability at least $1 - e^{-\kappa_0}$ and the result will be certified. Note that the result can only be certified if there is a heavy hitter.

**Tradeoff between $W$ and $F$.** If $w$ has an $\ell_q$-heavy hitter then it has an $\ell_p$-heavy hitter for every $p \geq q$. This means that for moments $p \geq 2$, an $\ell_2$ sample supports a larger set $W$ of frequencies than an $\ell_1$ sample, including also those with an $\ell_2$-heavy hitter but not an $\ell_1$-heavy hitter. The $\ell_1$ sample however supports a larger family $F$ that includes moments with $p \in [1, 2]$. Note that for a fixed overhead and $\ell_q$ sampling, the set $F$ of supported functions decreases with $q$ whereas $W$ increases with $q$.

### 4.3. Zipfian and sub-Zipfian Frequencies

Zipfian distributions are a very common model for frequency distributions in practice. We explore supported combinations with frequencies that are (approximately) Zipf.

**Definition 4.7.** We say that the frequencies $w$ where $\|w\|_1 = n$ are Zipf($\alpha$) if for all $i$, $w_i / w_1 = 1/\alpha$.

Values $\alpha \in [1, 2]$ are common in practice. The best-fit Zipf parameter for the datasets we studied is reported in Table 1 and the frequency distribution (sorted by rank) is shown in Figure 2. We can see that our datasets are approximately Zipf (which would be an approximate straight line) and for all but one we have $\alpha \in [1.3, 1.5]$.

We now define a broader class of approximately Zipfian distributions.

**Definition 4.8.** Frequencies $w$ are subZipf($\alpha, c, n$) if for all $i$, $w_i / w_1 \leq ci^{-\alpha}$.

Note that Zipf($\alpha, n$) is sub-Zipf with the same $\alpha$ and $c = 1$. We show that sub-Zipf frequencies (and in particular Zipf frequencies) have heavy hitters:

**Lemma 4.9.** For subZipf($\alpha, c, n$) frequencies, and $q$ such that $q \alpha \geq 1$, the frequency vector has an $\ell_q$ $c\beta$ subZipf-$\alpha$ heavy hitter, where $H_{n, \alpha} := \sum_{i=1}^n i^{-\alpha}$ are the generalized harmonic numbers.

Table 2 lists supported combinations that include these approximately Zipfian distributions. We see that for these approximately Zipf distributions, $\ell_1$ or $\ell_2$ samples emulate $\ell_p$ samples with small overhead.

### 4.4. Experiments on Estimate Quality

The overhead factors reported in Table 3 are in a sense worst-case upper bounds (for the dataset frequencies). Figure 4 reports the actual estimation error (normalized root mean square error) for high moments for representative datasets as a function of sample size. The estimates are with $\text{ppswor}$ ($\ell_1$ sample with replacement) and $\ell_2$ samples with and without replacement. Additional results are reported in the supplementary material. We observe that actual accuracy is significantly better than even the benchmark bounds.
Finally we consider estimating the full distribution of frequencies, that is, the curve that relates frequency of keys to their rank. We do this by estimating the actual rank of each key in the sample (using an appropriate threshold function of frequency). Representative results are reported in Figure 3 for ppswor and for with-replacement \( \ell_2 \) sampling (additional results are reported in the supplementary material). We used a sample of size \( k = 32 \) or \( k = 1024 \) for each set of estimates. We observe that generally the estimates are fairly accurate even with a small sample size (despite threshold function requiring large sketches in the worst case). We see that \( \ell_2 \) samples are accurate for the frequent keys but often have no representatives from the tail whereas the without replacement \( \ell_1 \) samples are more accurate on the tail.

5. Universal Samples

We study combinations where the emulation is universal, that is, \( F \) includes the set \( M \) of all monotone non-decreasing functions of frequency. Interestingly, there are sampling probabilities that provide universal emulation for any \( w \):

**Lemma 5.1.** (Cohen, 2015) Consider the probabilities \( q_i \) where the \( i \)th most frequent key has \( q_i = \frac{1}{i H_n} \). Then a weighted sample by \( q \) is a universal emulator with overhead at most \( H_n \).

This universal sampling, however, can not be implemented with small (polylogarithmic size) sketches. This because \( M \) includes functions that require large (polynomial size) sketches such as thresholds \( \{w \geq T \} \) and high moments \( (p > 2) \). We therefore aim for small sketches that provide universal emulation to a restricted \( W \).
For particular sampling probabilities \( q \) and frequencies \( w \) we consider the universal emulation overhead to be the overhead factor that will allow the sample to emulate weighted sampling with respect to \( f(w) \) for any \( f \in M \).

\[
\max_{f \in M} \max_i f(w_i)/\|f(w)\|_1 q_i.
\]

(6)

Interestingly, the universal emulation overhead of \( q \) does not depend on the particular \( w \).

**Lemma 5.2.** The universal emulation overhead of \( q \) is

\[
\max_i 1/(iq_i)
\]

and is always at least \( H_n \). This is tight even when \( W \) contains a single \( w \), as long as frequencies are distinct \( w_i > w_{i+1} \) for all \( i \).

We can similarly consider for sampling probabilities \( q \) the universal estimation overhead which is the overhead needed for estimating all (full) monotone \( f \)-statistics. As discussed in Section 2, the estimation is a weaker requirement than emulation (only applies to the full statistics) and hence for any particular \( q \) the estimation overhead can be lower than the emulation overhead. The estimation overhead, however, is still at least \( H_n \).

**Lemma 5.3.** The universal estimation overhead for estimating all monotone \( f \)-statistics for \( q \) is

\[
\max_i \frac{1}{i^2} \sum_{j=1}^i \frac{1}{q_i}.
\]

Proof. The overhead with frequencies \( w \) is

\[
\max_{f \in M} \sum_i f(w_i)^2/\|f(w)\|_1 q_i.
\]

(7)

It suffices to consider \( f \) that are threshold functions. The expression for the threshold at \( w_1 \) has \( f(w_j)/\|f(w)\|_1 = 1/i \) for \( j \leq i \) and 0 otherwise. We get that the sum is \( \frac{1}{i^2} \sum_{j=1}^i \frac{1}{q_i} \). The claim follows from taking the maximum over all threshold functions.

In our context, the probabilities \( q \) are not something we directly control but rather emerge as an artifact of applying a certain sampling scheme to a dataset with certain frequencies \( w \). We will explore the universal overhead of \( q \) we obtain when applying off-the-shelf schemes to Zipf frequencies and to our datasets.

For Zipf(\( \alpha \)) frequencies (\( \alpha \geq 1/2 \), \( \ell_p \) sampling with \( p = 1/\alpha \) is a universal emulator with (optimal) overhead \( H_n \). Interestingly, for \( \alpha \geq 1/2 \), this is attained by \( \ell_p \) sampling with \( p \leq 2 \), which has polylogarithmic size sketches.

Note that we match here a different \( \ell_p \) sample for each possible Zipf parameter \( \alpha \) of the data frequencies. A sampling scheme that emulates \( \ell_p \) sampling for a range \( [p_1, p_2] \) of \( p \) values with some overhead \( h \) will be a universal emulator with overhead \( h H_n \) for Zipf(\( \alpha \)) for \( \alpha \in [1/p_2, 1/p_1] \). One such sampling scheme with polylogarithmic-size sketches was provided in (Cohen, 2018; 2015). The sample emulates all concave sublinear functions that include capping functions \( f(w) = \min(w, T) \) for \( T > 0 \) and low moments with \( p \in [0, 1] \) with \( O(\log n) \) overhead.

Table 3 reports the universal overhead on our datasets with \( \ell_1, \ell_2 \), and multi-objective concave-sublinear sampling probabilities. We observe that while \( \ell_2 \) sampling emulates high moments extremely well, its universal overhead is very large due to poor emulation of “slow growth” functions. The better universal overhead of \( \ell_1 \) and concave-sublinear samples is \( h \in [143, 700] \) and is practically meaningful as it is in the regime where \( h \varepsilon^{-2} \ll n \).

6. Conclusion

We propose a framework where performance and statistical guarantees of sampling schemes are analyzed in terms of supported frequencies-functions combinations. We demonstrate analytically and empirically that sketches originally designed to sample according to “easy” functions of frequency on “hard” frequency distributions turn out to be accurate for sampling according to “hard” functions of frequency on “practical” frequency distributions. In particular, on “practical” distributions we can accurately approximate high frequency moments (\( p \geq 2 \)) and the rank versus frequency distribution using small composable sketches.

Acknowledgments

We are grateful to the authors of (Hsu et al., 2019), especially Chen-Yu Hsu and Ali Vakilian, for sharing their data, code, and predictions with us. We thank Ravi Kumar and Robert Krauthgamer for helpful discussions. The work of Rasmus Pagh was supported by Investigator Grant 16582, Basic Algorithms Research Copenhagen (BARC), from the VILLUM Foundation.

References


Composable Sketches Beyond the Worst Case


