Random Sampling for Data Intensive Computations

Dinkar Vasudevan*
Hamilton Institute
Ireland
dinkar.vasudevan@nuim.ie

Milan Vojnović
Microsoft Research
Cambridge, United Kingdom
milanv@microsoft.com

April 2010

Technical Report
MSR-TR-2009-8

Microsoft Research
Microsoft Corporation
One Microsoft Way
Redmond, WA 98052
http://www.research.microsoft.com

Abstract – We consider estimation of arbitrary range partitioning of data values and ranking of frequently occurring items based on random sampling, within small number of samplings and prescribed accuracy. These problems arise in the context of parallel-processing of massive datasets, e.g. performed in data centers of Internet-scale cloud services and large-scale scientific computations. The range partitioning is a basic block of parallel-processing systems based on the paradigm of map and reduce.
For the range partitioning, we consider a direct estimation method based on constructing an arbitrary-height histogram and characterize the estimation error. This approach provides substantial savings in constructing unbalanced range partitionings with respect to a standard approach based on equi-height histograms; our results extend previous work restricted to equi-height histograms. For the problem of ranking of frequently occurring items, we use a lumping of small frequency items that enables us to obtain tighter bounds that are independent of the total number of distinct items in a dataset. The analysis deploys the framework of large deviations that is well suited to typically large scale of data in the considered applications.
We demonstrate tightness and benefits of our sampling methods using a large data set of an operational cloud service that involves data at a scale of hundreds of billions of records. Our results provides insights and inform design of practical sampling methods.

1. INTRODUCTION

Cloud computing is considered to open a new era in computing and has been recently gaining quite some attention with major computer software and service companies deploying data centres consisting of hundredes of thousands of (commodity-hardware) machines [13, 20]. In this context but also other (e.g. scientific computations e.g. [27]), data-intensive compution are carried out on a daily basis by owners of individual applications over massive amounts of data, often in the order of tera (10^{12}) or even peta (10^{15}) bytes. For example, such computations are routinely run in production data centers of providers of Internet online services and involve computations such as ranking of web pages, computation of product recommendations for e-commerce or media content consumption, social networking, and other online services. It is important to note that computations are typically performed by *multiple* owners of individual applications even within a single organization who thus compete for typically large, albeit limited resources of data centers. It is crucial to enable efficient processing with respect to processing time and the use of resources in data centers.

Several proposals have been recently made for parallel processing of massive data sets, e.g. Mapreduce [11], Dryad [21], SCOPE [5], most of which are also deployed in production envioronments. Computations based on this paradigm can also be performed by users through commercially available services (e.g. Amazon Elastic Mapreduce [1]) and are also available through open-source (e.g. Hadoop [19]). In these systems, large volumes of data are typically processed by multiple tasks that process individual pieces of the data in parallel across machines of a data center. The basic component that facilitates this process is partitioning of data, in particular, range partitioning where data is partitioned over consecutive ranges with respect to values of the underlying data set. Another basic component is identification of frequently occuring data items on which specific computations are performed. The sheer volume of data necessiates solving these basic tasks approximately based on a sample of the data set. It is crucial to perform these basic computations with given accuracy and a small number of samples as this implies time and storage efficiency.

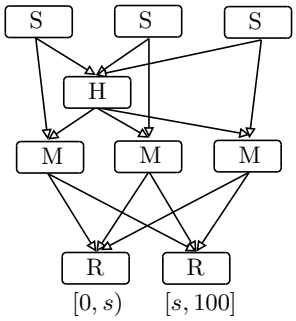

Figure 1: An example range partitioning in map-reduce scenario: the histogram block (H) estimates a range partitioning $[0, s)$, $[s, 100]$ using a sample from data sources (S) which is used by map tasks (M) to send output data to designated reduce tasks (R).

In this paper we consider two problems that are used as basic elements in variety of computations **(P1)** approximate construction of *arbitrary-height range partitioning* and **(P2)** approximate identification of frequently occuring data items. For these basic computational tasks we ask the following question:

> **(Q)** How many samples are necessary and sufficent for finding a solution within given error tolerance?

Our results on arbitrary-height range partitioning provide a direct and practical method for constructing arbitrary-height histograms. We show that this method provides substantial reduction of the sampling costs over an indirect approach that is based on equi-height histograms. To the best of our knowledge, our work provides first results on arbitrary-height range partitioning; previous work was restricted to equi-height histograms (Chaudhuri et al [7]). Furthermore, to the best of our knowledge, we provide first characterization results for several natural variants of identifying frequently occuring items which yield constructive algorithms for the underlying identification tasks. The basic computational tasks (P1) and (P2) are used in various computations and is thus important to understand the computational complexity of their solving; for concreteness, in the following we describe the use of range partitioning in a concrete application scenario.

Application scenario of map and reduce. Common to popular systems [11, 21] is a parallel-processing paradigm

based on the phases of map and reduce. First, data is chopped into pieces (e.g. blocks of 250M Byte) which is typically done by an underlying network file system. Second, computation over these individual data blocks is assigned to so called map tasks that are distributed across machines. Third, upon completion of the map tasks the results are sent to so called reduce tasks that are also distributed across machines and that perform the data aggregation. A computation typically consists of several such map and reduce phases which is application specific. It is important that each reduce task is designated for a specific part of the output of the map tasks. See Figure 1 for a typical example.

The partitioning of data over map and reduce tasks needs to ensure load balancing so that the amount of workload assigned to a task matches the processing capabilities of the machine that runs the task. This is important as for many applications the processing finishes and the results are of use only when the last task completes. Furthermore, for parallel-processing based on the map-reduce paradigm a new map-reduce phase is initiated only after the preceding has completed. Besides load balancing there may be additional requirements such that each task processes only the data from a specific range which requires *range partitioning*. It is noteworthy that there exist systems where users are allowed to implement their own data partitioning, e.g. [29].

The need for efficient arbitrary-height range partitioning. While in many scenarios one would want to distribute ranges across tasks evenly, there are several scenarios where an unbalanced range partitioning would need to be performed. For example, this may be for the following reasons:

• *heterogeneous hardware resources* – e.g. a system of commodity machines that differ in their processing capabilities (e.g. CPU and disk I/O); such heterogeneous systems may be commonplace in smaller enterprises, home environments, or even operating systems (e.g. Barrel fish [2]);

• *quantile-specific computations* – e.g. additional attributes are computed for the top 15% of most visited web pages or 10% most popular songs because they are expected to be served to most of users;

• *concurrent processing* – e.g. machines have same specifications but differ in the current CPU and I/O disk load because of other applications that run on them.

The standard approach commonly deployed in practice is based on estimating equi-height range parititiong [4]. While this approach works well for balanced partitioning in symmetric scenarios where data is partitioned evenly over tasks, it may perform poorely for constructing unbalanced partitioning. For example, consider the following range partitioning over two ranges according to the partitioning 20% and 80% – using the standard approach, first, an equi-height histogram of 5 bins is constructed and then one bin is assigned to the 20%-portion range and the remaining four bins are assigned to the 80%-portion range. We will see that this method can be grossly inefficient as it may require to estimate equi-height histograms of small widths and much larger number of bins than in fact needed; thus, in the $(20\%, 80\%)$ example, we need to estimate a histogram of 5 bins while the end goal is to partition data in 2 bins.

Frequently occuring items. We further consider identification of frequently occurring items in a dataset. Specifically, we consider three ranking objectives that are increasing in the amount of information about most frequently occurring items: (1) *top-k* where the goal is to identify a set of k most frequently occurring items; (2) *top-k with ranking* whose objective is in addition to rank the items in the top-k set in decreasing order with respect to their frequencies; and (3) *top-k with frequencies* where the goal is also to report frequencies of the items in the top-k set. These top-k problems arise in many applications where a computational task requires identification of frequently occurring items such as identification of most popular news, most played songs, or most watched videos, which then can be used as an end result or to condition further processing of an underlying more complex processing task.

It is important to note that the design of efficient sampling methods is a key element for data-intensive computing systems where applications need to process massive amounts of data within short time. Furthermore, the online aspect of the estimation is important as one may not have a prior information about the statistics of an underlying data set.

Summary of our Results. We summarize our contributions in the following points:

• We provide a characterization of the probability of error in constructing an arbitrary-height range partitioning of data values within given accuracy. The results characterize the trade-off between sample size and the accuracy and tell how much samples suffices for given level of accuracy. In particular, it is found that the minimum height of a bin plays a key role in determining the probability of error. These results extend previous research [7] that considered equi-height histograms and are of practical importance for efficient estimation of unbalanced data partitions. The results are established using the methods of large deviations which is a natural setting in view of the large scale of data in typical scenarios of data-intensive computations.

• We characterize the probability of error for each of the top-k ranking objectives. The results provide insights into how much more sampling is needed for given level of accuracy as we increase the amount of information that is being inferred by each of the ranking objectives. We use a novel approach based on the lumping of small frequency items to improve and provide tighter bounds on the error probability. This yields bounds that are independent of total number of distinct items in a data set, but depend only on the frequencies of the items in the top-k set. We also consider sequential sampling algorithms for our top-k set problems which do not require any a prior information.

• We provide numerical results based on real-world data and simulations. The data is from an operational cloud-service that involves tens of datasets of hundreds of billions of rows. Our numerical results are used to support the following points: (1) bounds established in this paper are near to empirically observed; (2) sampling methods yield sub-

stantial savings with respect to the number of samplings; (3) range partitioning based on estimating arbitrary-height histograms can yield substantial savings with respect to the number of samplings compared to the standard approach based on estimating equ-height histograms; (4) lumping of the small frequency items can yield substantial gains.

Outline of the Paper. Section 2 discusses related work. In Section 3 we present our results on the estimation of range partitioning. Section 4 contains results for each of the top-k ranking objectives considered in this paper. In Section 5, we show our numerical results. Finally, we conclude in Section 6. Some of the proofs are deferred to Appendix.

2. RELATED WORK

Sampling-based approaches for estimation of various statistics have been considered in the context of database systems and data streams by many, e.g. [15, 18, 6, 24, 7, 16, 17, 25, 23, 28, 8, 22]. Here we only discuss work that is closely related to ours.

Related to our range partitioning problem is the line of work on estimating quantiles in one pass through a data set [16, 17] where space and sample complexities are for identifying the item of a value of rank k is studied. The work that is perhaps most closely related to ours is that of Chaudhuri et al [7] that establishes results for equi-height range partitioning (or histograms). We admit the same definition of the error in constructing a range partitioning as in [7] which has quite some pratical appeal. Our results generalize the results to arbitrary-height range partitioning which yields a direct and practical method for efficient construction of arbitrary-height range partitions which in general may not be feasible by using equi-height histograms.

An early work on using random sampling to identify frequent items in a set is that of Gibbons and Matias [15] (therein referred to as "hot list queries"). They introduce two techniques of constructing and incrementally maintaining random samples (therein refered to as concise and counting samples) in one pass of the dataset. However, no explicit tradeoffs are provided between the sample size (or footprint) and the probability of failure in returning the top k set. Furthermore, the problem formulation in [15] is a little different from ours in that the objective is to identify items with frequencies greater or equal to $\max(p_k, \epsilon)$ where p_k is the frequency of the k-th most frequent item and ϵ an input (error tolerance) parameter. The scheme may therefore return fewer than k items. Our objective is that of identifying exactly k items with frequency greater than $p_k - \epsilon$. Another work using sparse sampling is [7] where the authors describe one pass algorithms for the related problem of identifying elements whose frequency exceeds a particular threshold. Similar objective as for our top-k problem was considered by Mannor and Tsitsiklis [26] but the authors therein focus on a specific problem of approximately identifying the most frequent item.

Finally, further related work is that on randomized hashing schemes [6, 9, 23] that were used for the related problem of estimating frequency moments [14]. The scheme involves constructing hashes from the data stream where each hash keeps count of some randomly chosen subset of elements of the data stream. All these schemes, however, assume a prior knowledge (and number) of all the distinct items in the dataset which is used to construct the hash functions. Such a strong assumption is not required for random sampling based schemes where the sketch is simply defined as the set of sampled items. Furthermore, random sampling-based schemes may be more practical to implement in distributed environments than random hashing schemes – for example, the distributed scheme [23] requires consistency on the hash functions used by individual nodes and non-trivial in-network aggregations of the hash values.

3. RANGE PARTITIONING

We consider the following range partitioning problem. Let $X = \{x_1, x_2, \ldots, x_n\}$ be a set of distinct values of n items $1, 2, \ldots, n$. The x_i's take value in an ordered domain; without loss of generality, let the items be numbered so that $x_1 < x_2 < \cdots < x_n$. Given is a distribution $P = (p_1, p_2, \ldots, p_k)$ that defines the height of each bin of a k-bin histogram of values. We want to determine a range partition $s_0 < s_1 < \cdots < s_k$ such that fraction p_i of items fall in the interval $(s_{i-1}, s_i]$. See Figure 2 for an illustration. More precisely, we define s_i for $i = 1, \ldots, k-1$ by

$$s_i = \min \left\{ x_j \in X : j \geq \lceil \sum_{l=1}^{i} p_l n \rceil \right\}.$$

and define $s_0 = x_1$ and $s_k = x_n$. If the distribution P is such that every p_i takes values in $\{\frac{1}{n}, \frac{2}{n}, \ldots, \frac{n}{n}\}$, then we can write

$$p_i = \frac{|\{x_j \in X : s_{i-1} < x_j \leq s_i\}|}{n}.$$

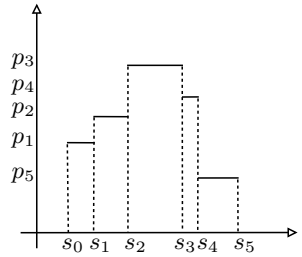

Figure 2: A histogram for $k = 5$.

We estimate the range partition as follows: (1) we construct a random sketch $\hat{X}$ by taking all the values observed in t samplings with replacement of the set X; (2) we sort the items in $\hat{X}$ by value and identify the range partition $\hat{s}_0 \leq \hat{s}_1 \leq \cdots \leq \hat{s}_k$ with respect to the set $\hat{X}$ and distribution P. In particular $\hat{s}_0 = x_1$ and $\hat{s}_k = x_n$ and for $i = 1, \ldots, k-1$,

$$\hat{s}_i = \min \left\{ \hat{x}_j \in \hat{X} : j \geq \lceil \sum_{l=1}^{i} p_l t \rceil \right\}.$$

We now define the heights of the bins $Q = (q_1, q_2, \ldots, q_k)$ induced by the separators $\hat{s}_0, \hat{s}_1, \ldots, \hat{s}_k$ of the random sketch. The height q_i is the fraction of items in the data set X that falls in the range $(\hat{s}_{i-1}, \hat{s}_i]$, i.e.

$$q_i = \frac{|\{x_j \in X : \hat{s}_{i-1} < x_j \leq \hat{s}_i\}|}{n}.$$

DEFINITION 3.1. *A range partition induced by the separators $\hat{s}_0 \leq \hat{s}_1 \leq \cdots < \hat{s}_k$ is said to be ϵ-approximation with respect to distribution P if*

$$|q_i - p_i| \leq \epsilon p_i \text{ for every } i = 1, \ldots, k.$$

For an ϵ-approximate range partitioning, the relative difference of the actual height q_i and the desired height p_i is at most ϵ. The above approximation objective was originally proposed in [7] and was used for analysis of equi-height histograms. The objective is nautural and is compelling for applications to bound the deviations of the bin heights relative with respect to the bin heights.

Throughout the paper, we denote with $D(P||Q)$ the Kullback-Leibler divergence between two distributions P and Q, which is defined as

$$D(P||Q) = \sum_{i=1}^{m} p_i \log \frac{p_i}{q_i}.$$

With a slight abuse of notation, we will denote with $D(x||y)$ for $x, y \in [0, 1]$ the Kullback-Lieber divergence between two binary distributions $(x, 1-x)$ and $(y, 1-y)$.

Let p_e denote the probability that the estimated range partitioning is not ϵ-approximation with respect to the distribution P. We prove the following bounds on p_e.

THEOREM 3.1. *For every $\epsilon < (1 - p_1)/p_1$, the probability of error satisfies*

$$p_e \leq kn^2 e^{-tD(\pi||\pi(1+\epsilon))} \tag{1}$$

where π is the minimum height $\pi = \min_i p_i$ For small ϵ, we have that

$$p_e \leq kn^2 e^{-t(\frac{\epsilon^2 \pi}{2(1-\pi)} + O(\epsilon^3))} \tag{2}$$

The upper-bound (1) has asymptotically exact error exponent, i.e.

$$\lim_{t \to \infty} \frac{1}{t} \log p_e = -D(\pi||\pi(1+\epsilon)).$$

PROOF. The error probability p_e is given by

$$p_e = \Pr(|q_i - p_i| > \epsilon p_i, \text{ for some } i).$$

The following chain of inequalities hold:

$$\begin{aligned}
p_e &\overset{(a)}{\leq} \sum_{i=1}^{k} \Pr(|q_i - p_i| > \epsilon p_i) \\
&\overset{(b)}{=} \sum_{i=1}^{k} \sum_{\hat{s}_{i-1}, \hat{s}_i : |q_i - p_i| > \epsilon p_i} \Pr(|q_i - p_i| > \epsilon p_i) \\
&\overset{(c)}{\leq} \sum_{i=1}^{k} \sum_{\hat{s}_{i-1}, \hat{s}_i : |q_i - p_i| > \epsilon p_i} e^{-tD(p_i||q_i)} \\
&\overset{(d)}{\leq} \sum_{i=1}^{k} n^2 e^{-t \min_{q_i : |q_i - p_i| > \epsilon p_i} D(p_i||q_i)} \\
&\overset{(e)}{=} \sum_{i=1}^{k} n^2 e^{-t \min_i (D(p_i||p_i(1+\epsilon)), D(p_i||p_i(1-\epsilon)))} \\
&\overset{(f)}{=} kn^2 e^{-t(D(\pi||\pi(1+\epsilon)), D(\pi||\pi(1-\epsilon)))} \\
&\overset{(g)}{=} kn^2 e^{-tD(\pi||\pi(1+\epsilon))}
\end{aligned} \tag{3}$$

The inequality (a) follows by taking the union bound for the probabilities of the errors resulting from deviations of the individual bins. The equality (b) follows by splitting the term $\Pr(|q_i - p_i| > \epsilon p_i)$ into a summation of disjoint probabilities corresponding to given separator pair $\hat{s}_{i-1}, \hat{s}_i$. The inequality (c) holds due to the following reason: from definition, a fraction p_i of elements in the sketch are contained in $(\hat{s}_{i-1}, \hat{s}_i]$ while a fraction q_i of elements in the underlying set X are contained in $(\hat{s}_{i-1}, \hat{s}_i]$. Let us represent the values in X which lie in $(\hat{s}_{i-1}, \hat{s}_i]$ by 1 and the values which do not lie in $(\hat{s}_{i-1}, \hat{s}_i]$ by 0. Then, the event $\{|q_i - p_i| > \epsilon p_i\}$ for a given choice of $\hat{s}_{i-1}, \hat{s}_i$ implies that sampling uniformly t times from a Bernoulli(q_i) distribution results in a sketch with a portion p_i of elements lying in $(\hat{s}_{i-1}, \hat{s}_i]$ and a portion $1 - p_i$ of elements lying outside the interval, i.e., a sketch with empirical distribution which is Bernoulli(p_i). By Sanov's theorem [12], this probability is upper bounded by $e^{-tD(p_i||q_i)}$.

The inequality (d) follows from an upper bound on the number of different separator pairs $(\hat{s}_{i-1}, \hat{s}_i)$ that we need to consider. Since each separator can assume one of n values, the number of terms in the inner summation is bounded by n^2 and (d) follows by replacing every divergence term with it's minimum value. The equality (e) follows by choosing q_i as close to p_i while violating the guarantee, i.e., choosing $q_i = p_i(1 + \epsilon)$ or $q_i = p_i(1 - \epsilon)$. [1] It can be shown that the divergence terms are minimized by the p_i with minimum value and thus by defining $\pi = \min_i p_i$, (f) follows.

The small ϵ approximation to (3) gives (2).

Finally, the last statement follows from the tightness of the exponent of the error probability (Sanov's theorem [12]). □

COROLLARY 3.1. *For given P, $0 < \epsilon < (1 - p_1)/p_1$ and $0 < \delta < 1$, in order for $p_e \leq \delta$ to hold it suffices to take t*

[1] The reader may note that since the $q_i's$ have granularity $1/n$, it may not attain a specific value, so the equality may actually be an inequality.

samples with

$$t \geq \frac{2(1-\pi)}{\pi\epsilon^2}\left(\log\frac{1}{\delta} + 2\log n + \log k\right)(1+o(\epsilon)), \qquad (4)$$

where $\pi = \min_i p_i$.

REMARK 3.1. *The error typically occurs due to deviation in the estimate of the smallest-height bin, i.e., a bin of height* π*, by a relative factor of* ϵ*. This is reflected in (4) where* t *is increasing with* $1/\pi$*.*

REMARK 3.2. *The above corollary extends previous work by Chaudhuri, Motwani and Narasayya [7] to arbitrary-height histograms. In [7], the authors also consider* ϵ*-approximate range partitioning but only for balanced histograms, i.e.,* P *uniform. For this special case of* $p_i = 1/k$*, Equation (4) boils down to*

$$t \geq \frac{2(k-1)}{\epsilon^2}\left(\log\frac{1}{\delta} + 2\log n + \log k\right)(1+o(\epsilon)).$$

3.1 The Benefit over Indirect Approach

We now address the need and benefits of having estimates of unbalanced range partitions in contrast to having estimates of only balanced ranged partitions. The approach used in practice consists of the following steps which we refer to as an indirect approach. First, a balanced range partitioning is estimated. Then, this range partitioning is used to assign ranges to individual machines. If the goal is a balanced range partitioning, i.e. each machine is assigned an equal portion of items, then the indirect approach solves the problem – if there are k machines, a balanced partitioning over k bins is estimated, which are then allocated to machines according to a one-to-one mapping.

The indirect approach can be used to produce unbalanced range partitionings when the distribution P is such that for every bin i, $p_i = \frac{M_i}{M}$ for positive integers M_i and M such that $M_1 + M_2 + \cdots + M_k = M$. The indirect approach can then be applied by first estimating a *balanced* partitioning into k' intermediate bins which are then grouped to from a range partitioning over k bins with respect to the distribution P where we set $M/\gcd(M_1, M_2, \ldots, M_k)$. This approach to unbalanced range partitioning using balanced partitions can be grossly more expensive with respect to the required number of samplings than by directly estimating the unbalanced histogram. The minimum height of a bin by the direct approach is $\pi = \min(M_1, M_2, \ldots, M_k)/M$ while for the indirect approach it is $\pi' = \gcd(M_1, M_2, \ldots, M_k)/M$. From our analysis (refer Corollary 3.1) it follows that for any given probability of error δ and small ϵ, the number of of samplings under the indirect approach is at least a factor $(1/\pi')/(1/\pi)$ more than that under the direct approach, i.e.

$$\frac{\frac{1}{\pi'}}{\frac{1}{\pi}} = \frac{\min(M_1, M_2, \ldots, M_k)}{\gcd(M_1, M_2, \ldots, M_k)}. \qquad (5)$$

This measures how much more samplings it is required by the indirect approach relative to the direct approach and can be shown in many cases to assume large values. In the following we provide an example and provide more comprehensive evidence in Section 5.

Suppose $M = 100$ and $M_1 \in \{1, 2, \ldots, 50\}$. With the indirect approach we construct an equi-height histogram with k' bins where $k' = 100/\gcd(M_1, M_2)$ For example, if $M_1 = 40$, then $\gcd(M_1, M_2) = 20$ and $k' = 5$; if $M_1 = 45$, then $\gcd(M_1, M_2) = 5$ and $k' = 20$. In Figure 3, we show the ratio $\min(M_1, M_2)/\gcd(M_1, M_2)$ and observe that it can assume large values.

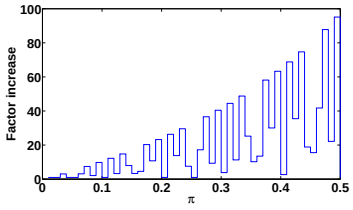

Figure 3: Factor increase lower bounds the ratio of the sufficient number of samples of the indirect and the direct approach.

4. RANKING BY FREQUENCY

We consider a set of items $X = \{x_1, x_2, \ldots, x_n\}$ where each item from this set is an element of a set of m distinct items, $\mathcal{X} = \{1, 2, \ldots, m\}$. For example, the set X could represent a database column or a data stream and the set $\mathcal{X}$ represents all the distinct rows in the database column or the data stream. For each $i \in \mathcal{X}$, let n_i denote the number of occurences of item i in the set X. Define the distribution P on the alphabet $\mathcal{X}$ as $p_i = n_i/n$ for every $i \in \mathcal{X}$. Without loss of generality, let us assume that items in $\mathcal{X}$ are labeled such that

$$p_1 \geq p_2 \geq \cdots \geq p_m.$$

We would like to identify a set of $1 \leq k \leq m$ most frequently occuring items. Specifically, we consider the following problems:

1. **Top-k**: identify a set of k most frequent items;

2. **Top-k with ranking**: identify a set of k most frequent items and order them in decreasing order of their frequencies;

3. **Top-k with frequencies**: identify and estimate the frequencies of k most frequent items.

Note that the above described problems are in increasing order of demand, i.e., top-k with frequencies provides additional information as compared to top-k with ranking, which in turn provides additional information as compared to top-k. In the following we define each of the top-k ranking objectives more formally, after we introduce some notation that we use throughout the paper.

Let $\mathcal{T}$ be a set that contains k items with frequencies greater or equal to p_k. Let t denote size of the sketch derived by sampling with replacement from the dataset X and let $Q = (q_i, i \in \mathcal{X})$ denote the frequencies of items in the sketch. The sketch is sorted in decreasing order of the frequencies Q and the top k most frequently occuring items in the sketch are reported as the top k set. In the case that their frequen-

cies are also required, the empirical frequencies are reported along with the top k elements.

Let $\mathcal{B}$ be a set that contains items with frequencies less than $p_k - \epsilon$, i.e. an item i is in the set $\mathcal{B}$ if $p_i < p_k - \epsilon$; these are the items that should not be reported as an answer.

DEFINITION 4.1 (TOP-K). *A set of items S is said to be an ϵ-approximate top-k set (resp. relative ϵ-approximate), if S contains any k distinct items from $\mathcal{X}$ and $p_i \geq p_k - \epsilon$ (resp. $p_i \geq (1-\epsilon)p_k$), for every item $i \in S$.*

DEFINITION 4.2 (TOP-K WITH RANKING). *A set of items S is said to be an ϵ-approximate top-k ranked set (resp. relative ϵ-approximate), if S is a top k set and furthermore item j is ranked above item i in S whenever $p_i < p_j - \epsilon$ (resp. $p_i < p_j(1-\epsilon)$).*

DEFINITION 4.3 (TOP-K WITH FREQUENCIES). *A set of items S along with the frequencies $(q_i, i \in S)$ is said to be ϵ-approximate top-k with frequencies if S contains k distinct items from $\mathcal{X}$ such that*

1. *$p_i \geq p_k - \epsilon$ for every $i \in S$, and*
2. *$|q_i - p_i| < \frac{\epsilon}{2}$, for every $i \in S$.*

Similarly, S, $(q_i, i \in S)$ is said to be relative ϵ-approximate top-k with frequencies if item 1 holds with ϵ replaced with ϵp_k and item 2 holds with ϵ replaced with ϵp_i.

DEFINITION 4.4. *A uniformly sampled sketch of size t provides a $1-\delta$ guarantee if probability of error in reporting the ϵ-approximation is within δ.*

The Lumping of Small Frequency Items. Our objective is to derive tradeoffs between the number of samplings required and the error probability of reconstructing the top-k sets. In particular, we show that the number of samplings *does not scale* with the alphabet size m. To obtain tighter bounds on the error probabilities, we use a technique of "lumping" elements of small frequencies, described in the following. We use the notation ϵ_k where $\epsilon_k = \epsilon$ and $\epsilon_k = \epsilon p_k$ in the ϵ-approximate and relative ϵ-approximate case, respectively. We define a lumped distribution $\tilde{P}$ from P as follows: we iteratively merge any two items with frequencies smaller than $\frac{p_k - \epsilon_k}{2}$ into "super-items" until either none or one (super-)item is left with frequency smaller than $(p_k - \epsilon_k)/2$. This defines a new set of distinct items $\tilde{\mathcal{X}}$ that has at most one item with frequency smaller than $(p_k - \epsilon_k)/2$. We emphasize here that $\tilde{P}$ and $\tilde{\mathcal{X}}$ constructed so are not necessarily unique. All items in P with frequencies larger than $\frac{p_k - \epsilon_k}{2}$ remain unmerged and hence retain their identity and frequency as in $\tilde{P}$. Items in P with frequencies smaller than $\frac{p_k - \epsilon_k}{2}$ map to super-items in $\tilde{P}$. Clearly, the size of $\tilde{\mathcal{X}}$ is bounded as:

$$|\tilde{\mathcal{X}}| \leq \frac{2}{p_k - \epsilon_k} + 1. \tag{6}$$

We will use the distribution $\tilde{P}$ to provide tighter bounds on the probability of error in the following sections. Roughly speaking, this enables us to improve a pre-factor in the probability of error from order m to order $1/p_k$ which in many cases in practice can be orders of magnitude smaller. We will also see that the lumping typically either does not effect or effects the error exponent only slightly, thus the error exponent remains tight or nearly tight.

In the following sections, we describe the tradeoffs between the error probability and the amount of sampling. The proofs of the theorems are in the Appendix.

4.1 Top-k

An error may occur in reporting a correct ϵ-approximate top-k if for an item $j \in \mathcal{B}$ and an item $i \in \mathcal{T}$, the observed frequencies q_j and q_i are such that $q_j \geq q_i$. Therefore, we consider the following probability:

$$p_e = \Pr\left(\bigcup_{i \in \mathcal{T}, j \in \mathcal{B}} \{q_j \geq q_i\} \right).$$

The following result characterizes the error probability p_e and absolute error tolerance.

THEOREM 4.1. *For every $0 < \epsilon < p_k$,*

$$\begin{aligned} p_e &\leq k(m-k)\left(1 - (\sqrt{p_k} - \sqrt{p_{l(k)}})^2\right)^t && (7) \\ &\leq k(m-k)\left(1 - (\sqrt{p_k} - \sqrt{p_k - \epsilon})^2\right)^t && (8) \end{aligned}$$

where $l(k)$ is an item in $\mathcal{B}$ with maximum frequency. Furthermore, the bound (7) has asymptotically tight exponent, i.e.

$$\lim_{t \to \infty} \frac{1}{t} \log p_e = \log\left(1 - (\sqrt{p_k} - \sqrt{p_{l(k)}})^2\right). \tag{9}$$

The result tells us that typical error occurs at the bottom of the top-k where an item with the smallest frequency in the set $\mathcal{T}$ gets sampled less than an item with maximum frequency smaller than $p_k - \epsilon$. The key parameters that determine the exponent of the error probability are the frequencies p_k and the largest frequency smaller or equal to $p_k - \epsilon$. For the bound (8) the key parameter is the frequency of the k-th most frequent item in the data set X.

With the lumping of small frequency items. Using the lumping of small frequency items we can reduce the pre-factor in the bound (8) as shown in the following result.

THEOREM 4.2. *For every $0 < \epsilon < p_k$,*

$$p_e \leq kK\left(1 - (\sqrt{p_k} - \sqrt{p_k - \epsilon})^2\right)^t \tag{10}$$

where

$$K = \min\left(\frac{2}{p_k - \epsilon}, m - k\right).$$

In comparison with Theorem 4.2, the pre-factor of the bound on the error probability is reduced from order $k(m-k)$ to $O(k \min(1/p_k, m-k))$.

We have the following corollary that characterizes the sampling cost.

COROLLARY 4.1. *For every $0 < \epsilon < p_k$ and $0 < \delta < 1$, suppose that the sketch is of size t such that*

$$t \geq \frac{4p_k}{\epsilon^2}\left(\log\frac{1}{\delta} + \log(kK)\right)(1+o(\epsilon)) \qquad (11)$$

then, the probability of error p_e is less or equal to δ.

The lumping argument provides tighter bounds to the error probability (alternatively sampling costs) whenever $p_k > \frac{2}{m-k} + \epsilon$ which in many cases of practical interest would hold true. In Section 5 we show cases from practice where p_k is orders of magnitude larger than $\frac{2}{m-k} + \epsilon$ which provides substantial reduction of the sketch size.

Relative error tolerance. For relative error tolerance we have the following results.

THEOREM 4.3. *The same statements hold as in Theorem 4.2 by replacing ϵ with ϵp_k and K replaced by K' where*

$$K' = \min\left(\frac{2}{p_k(1-\epsilon)}, m-k\right).$$

COROLLARY 4.2. *For every $0 < \epsilon < 1$, $0 < \delta < 1$ suppose that the sketch is of size t such that*

$$t \geq \frac{4}{\epsilon^2 p_k}\left(\log\frac{1}{\delta} + \log(kK')\right)(1+o(\epsilon)) \qquad (12)$$

then, the probability of error p_e is less or equal to δ.

Notice that these results derive from those under absolute error tolerance by replacing ϵ with ϵp_k. By doing so, we obtain the sampling cost that scales inversely proportional to p_k and the gain of the same order by using the lumping of small frequency items.

4.2 Top-k with ranking

For the top k ranking problem, we have the following result.

THEOREM 4.4. *For every $0 < \epsilon < p_k$,*

$$\begin{aligned} p_e &\leq (k(m-k)+k)\left(1 - \min_{i\leq k}(\sqrt{p_i} - \sqrt{p_{l(i)}})^2\right)^t \qquad (13)\\ &\leq (k(m-k)+k)\left(1 - (\sqrt{p_1} - \sqrt{p_1-\epsilon})^2\right)^t. \qquad (14)\end{aligned}$$

Furthermore, the bound (13) has asymptotically tight exponent, i.e.

$$\lim_{t\to\infty}\frac{1}{t}\log p_e = \log\left(1 - \min_{i\leq k}(\sqrt{p_i} - \sqrt{p_{l(i)}})^2\right). \qquad (15)$$

The results are different from that for top-k problem (Theorem 4.1) in that, (1) the pre-factors in (13) and (14), since for the top-k with ranking we need to compare pairs of items within the set $\mathcal{T}$, (2) the term $\min_{i\leq k}(\sqrt{p_i} - \sqrt{p_{l(i)}})^2$ in (13) and (15). The latter difference is due to the stricter need of ranking in accordance with Definition 4.2. From (15), the typical error occurs due to item $l(i)$ being sampled more often than an item i in $\mathcal{T}$ that minimizes the gap $\sqrt{p_i} - \sqrt{p_{l(i)}}$.

With the lumping of small frequency items. We have the following result whose proof is along the same lines as for Theorem 4.2 and is omitted.

THEOREM 4.5. *For every $0 < \epsilon < p_k$,*

$$p_e \leq kL\left(1 - (\sqrt{p_1} - \sqrt{p_1-\epsilon})^2\right)^t$$

where

$$L = \min\left(\frac{2}{p_k - \epsilon} + 1, m-k+1\right).$$

In comparison to Theorem 4.4, the reduction of the prefactor is from order $k(m-k+1)$ to $O(k\min(1/p_k, m-k+1))$. The number of samplings for the top-k ranking problem is given by

COROLLARY 4.3. *For every $0 < \epsilon < p_k$ and $0 < \delta < 1$, suppose that the sketch is of size t such that*

$$t \geq \frac{4p_1}{\epsilon^2}\left(\log\frac{1}{\delta} + \log(kL)\right)(1+o(\epsilon)) \qquad (16)$$

then, the probability of error p_e is less or equal to δ.

Notice that the sampling cost of top-k with ranking with absolute error tolerance is a factor p_1/p_k more than that for top-k (Corollary 4.1). In Section 5 we quanitify how much more samplings is needed if we require ranking of the top-k elements using real-world datasets.

Relative error tolerance. For top-k with ranking under relative error tolerance we have the following result:

THEOREM 4.6. *For the top-k ranking problem with relative error tolerance of ϵ, the probability of error p_e satisfies*

$$p_e \leq kL'\left(1 - (\sqrt{p_k} - \sqrt{p_k(1-\epsilon)})^2\right)^t.$$

where $L' = \min\left(\frac{2}{p_k(1-\epsilon)} + 1, m-k+1\right)$.

Notice that unlike to top-k with ranking with absolute error tolerance (Theorem 4.4) it is the frequency p_k of the k-th most frequenty item that is a key parameter, not the frequency p_1 of the most frequent item.

COROLLARY 4.4. *For a given ϵ, δ, we have that for $p_e \leq \delta$ it suffices to take t samples with*

$$t \geq \frac{4}{\epsilon^2 p_k}\left(\log\frac{1}{\delta} + \log kL'\right)(1+o(\epsilon)).$$

Interestingly, under relative error tolerance, the sampling complexity of the top-k with ranking is the same as for top-k (Corollary 4.2) although the underlying identification task is more demanding. This happens as the typical error in both cases occurs at the bottom of the top-k set.

4.3 Top-k with Frequencies

For the top-k with frequencies recall that we want every declared item i to have frequency p_i greater than $p_k - \epsilon$ and, moreover, we want the empirical frequency q_i to be within $\frac{\epsilon}{2}$ of the frequency p_i. Note that this ensures that from the observed frequencies q_i and q_j of two declared items i and j, we can correctly order them with respect to their frequencies p_i and p_j provided the gap between the latter two frequencies is sufficently large. Specifically, if $p_i > p_j + \epsilon$, then if we observe $q_i > q_j$ then it follows that $p_i > p_j$ with probability at least $1 - \delta$.

THEOREM 4.7. *For the error probability p_e of the top-k with frequencies we have*

$$p_e \leq 2m \exp\left(-t \min_i D_i\right) \tag{17}$$

where

$$D_i = D(\min(p_i, 1-p_i) + \frac{\epsilon}{2} || \min(p_i, 1-p_i)). \tag{18}$$

Furthermore,

$$\lim_{t\to\infty} \frac{1}{t} \log p_e = -\min_i D_i. \tag{19}$$

Note that $D_i = D(p_i + \frac{\epsilon}{2}||p_i)$ in all the cases but for the largest frequency item $i = 1$ if $p_i > 1/2$. We make the following observations:

FACT 4.1. $\min_i D_i$ *is achieved by an item i with frequency* $\min(p_i, 1-p_i)$ *greater or equal to $p_{l(k)}$.*

However, if ϵ is sufficiently small then $\min_i D_i$ is achieved by the item with highest frequency. More precisely,

FACT 4.2. *If $\epsilon \leq 2(1-2p_1)$ then $\min_i D_i$ is achieved by item 1 and $\min_i D_i = D(p_1 + \frac{\epsilon}{2}||p_1)$.*

In this case, typical error occurs due to oversampling of the item with highest frequency.

It follows that for small ϵ, we have

$$\lim_{t\to\infty} \frac{1}{t} \log p_e \leq \frac{\epsilon^2}{8p_1(1-p_1)}(1 + o(\epsilon)). \tag{20}$$

With the lumping of small frequency items. For the top-k with frequencies, we have the following result:

THEOREM 4.8.

$$p_e \leq 2M \exp\left(-t \min_i D_i\right) \tag{21}$$

where

$$D_i = \min\Big(D(\min(p_i, 1-p_i) + \frac{\epsilon}{2}|| \min(p_i, 1-p_i)),$$
$$D(\min(p_k, 1-p_k) - \frac{\epsilon}{2}|| \min(p_k, 1-p_k) - \epsilon)\Big)$$

and

$$M = \min\left(\frac{2}{p_k - \epsilon} + 1, m\right).$$

The result yields a reduction of the pre-factor in the bound on the probability of error from order m to $O(\min(1/p_k, m))$ which in many cases can provide significant reductions (see Section 5). The slight difference in the above exponent compared to (18) is due to accounting for a "super-item" close to $p^*_{\frac{\epsilon}{2}}$ (see Appendix).

COROLLARY 4.5. *Given ϵ, δ we have that $p_e \leq \delta$ provided that the number of samplings t satisfies*

$$t \geq \frac{8p_1(1-p_1)}{\epsilon^2}\left(\log\frac{1}{\delta} + \log(2M)\right)(1 + o(\epsilon)) \tag{22}$$

Notice that in comparison to top-k (Corollary 4.1) and top-k with ranking (Corollary 4.3), for given error probability, the top-k with frequencies requires approximately a factor $2p_1/p_k$ and a factor 2 more samplings, respectively.

Relative error tolerance. Under relative error tolerance, we have the following result:

THEOREM 4.9. *The error probability for the top-k set with frequencies and relative tolerance of ϵ satisfies*

$$p_e \leq 2M' \exp\Big(-tD\Big(p_k(1-\epsilon)(1+\frac{\epsilon}{2})||p_k(1-\epsilon)\Big)\Big) \tag{23}$$

where

$$M' = \min\Big(\frac{2}{p_k(1-\epsilon)} + 1, m\Big).$$

COROLLARY 4.6. *For a given ϵ, δ, we have that for $p_e \leq \delta$ it suffices to take t samples with*

$$t \geq \frac{8(1-p_k)}{\epsilon^2 p_k}\Big(\log\frac{1}{\delta} + \log 2M'\Big)(1 + o(\epsilon)).$$

Notice that unlike to under absolute error tolerance (Corollary 4.5), the key parameter is the frequency of k-th most frequent item. In comparison to top-k (Corollary 4.2)and top-k with ranking (Corollary 4.4) under relative error tolerance, for same error probability the top-k with frequencies requires about twice as many samples.

4.4 Comparison of the Sampling Complexity

In this section we summarize the comparison on the sampling complexity of different top-k ranking objectives. In Table 1 we provide a summary of the sampling complexity for the various top-k problems for both absolute and relative error guarantees.

We make the following observations:

(i) The sampling cost t varies inversely with the square of the tolerance ϵ and directly with the logarithm of the error

Table 1: Sampling complexity of various top-k problems.

Ranking	Sampling cost (abs tol)	Sampling cost (rel tol)
top-k	$\frac{4p_k}{\epsilon^2}\left(\log\frac{1}{\delta}+\log(kK)\right)$	$\frac{4}{p_k\epsilon^2}\left(\log\frac{1}{\delta}+\log(kK')\right)$
top-k with ranking	$\frac{4p_1}{\epsilon^2}\left(\log\frac{1}{\delta}+\log(kL)\right)$	$\frac{4}{p_k\epsilon^2}\left(\log\frac{1}{\delta}+\log(kL')\right)$
top-k with frequencies	$\frac{8p_1(1-p_1)}{\epsilon^2}\left(\log\frac{1}{\delta}+\log(2M)\right)$	$\frac{8(1-p_k)}{p_k\epsilon^2}\left(\log\frac{1}{\delta}+\log(2M')\right)$

probability δ. In other words, the error probability exponentially decays with the number of samples.

(ii) The sampling cost increases with increasing order of demand of the top-k set problem. In particular, for absolute error guarantees for the top-k set problem, t depends on the k-th largest frequency p_k, while for the top-k ranking and frequency problems, t depends on the largest frequency p_1. As observed before, this is due to the qualitative difference in the typical error events: In the former case, typical error occurs due to over sampling an item from $\mathcal{B}$ as compared to an item from $\mathcal{T}$, while in the latter case, typical error is due to deviations of the item with largest frequency.

(iii) In contrast, for relative error guarantees, the sampling costs varies inversely with the k-th largest frequency p_k for all variants of the problem. The reason is, for a given relative tolerance, deviations are more likely to occur for elements with smaller frequencies. In the present case, typical error is due to the violation of relative error guarantees by an element with k-th largest frequency.

(iv) We also note that the sampling costs do not scale a $\log(m)$ but instead as $\log(O(\min(1/p_k, m))$. The saving is due to the lumping arguments used.

4.5 Sequential Sampling

In the preceding text we considered the sample size for top-k estimation problems for given probability of error. The sample sizes that we derived depend on some parameters of the underlying distribution P; for example, on the frequency of the k-th most frequent item p_k. These could be seen as nuisance parameters in the same vein as variance is a nuisance parameter in estimating a confidence interval. The derived sample sizes can be applied for top-k estimation provided that one has a prior information about the values of the nuisance parameters; e.g. using a two-phase procedure where in the first phase the nuisance parameters are estimated, and then in the second phase, top-k estimation is performed for the number of samples determined by the previously-estimated nuisance parameters. In this section we outline a procedure that does this all at once using a well known sequential sampling approach where unknown parameters are estimated on-line and the sample size is determined by a stopping rule that depends on the observations.

In order to derive a stopping rule we use the posterior distribution of the error event; let $p_e(y_1,\ldots,y_t)$ denote the posterior distribution given the observed samples $y_1,\ldots,y_t$. The estimation is stopped at a number of samples T when the posterior distribution $p_e(y_1,\ldots,y_T)$ is less than or equal to given δ. We use standard Bayesian approach to characterize the posterior probability of error where a prior distribution is admitted for P. As standard, let P' be a vector by taking all but one coordinate of the distribution P; say the omitted coordinate is for item m. Let Q' be the corresponding vector of the observed frequencies. By asymptotic Bayesian theory [3, 10], we have that for large t, $Z = P' - Q'$ is asymptotically a multivariate Gaussian random variable with zero mean and covariance matrix Σ where Σ is the inverse of the Fisher's information matrix

$$I(P) = \left[-\frac{\partial^2}{\partial p_i \partial p_j}\sum_{s=1}^{t}\log(\Pr(X_s = y_s|P))\right]$$

that is evaluted at the maximum-likelihood estimate Q'. Since in our case $X_1,\ldots,X_t$ is a sequence of independent and identically distributed random variables where the distribution of X_s is multinomial, it is not difficult to show that $\Sigma = [\sigma_{i,j}]$ where $\sigma_{i,i} = \frac{1}{t}q_i(1-q_i)$ and $\sigma_{i,j} = -\frac{1}{t}q_iq_j$ for $i \neq j$. We use the fact that for large sample size, the distribution of the random variable Z is asymptotically Gaussian to evaluate the posterior probability of error.

Top-k. The posterior probability of error, given the observed samples $y_1,\ldots,y_t$ is

$$p_e(y_1,\ldots,y_t) = \Pr(\cup_{i\in\mathcal{T},j\in\mathcal{B}}\{p_j \geq p_i\}|Q)$$

where $\mathcal{T}$ and $\mathcal{B}$ are defined as earlier but with respect to the distribution Q. We use

$$\Pr(\cup_{i\in\mathcal{T},j\in\mathcal{B}}\{p_j \geq p_i\}|Q) \leq |\mathcal{T}||\mathcal{B}|\max_{i\in\mathcal{T},j\in\mathcal{B}}\Pr(\{p_j \geq p_i\}|Q).$$

Now, we consider the event $\{p_j \geq p_i\}$ for any given item i and item j, conditional on Q. We have that $p_j - p_i$ is asymptotically Gaussian with mean $q_j - q_i$ and variance $\mathrm{E}((Z_j - Z_i)^2)$. It is not hard to derive that

$$\mathrm{E}((Z_j - Z_i)^2) = \frac{1}{t}[q_i + q_j - (q_i - q_j)^2]. \tag{24}$$

Let z_x be the $(1-x)$-quantile of a normal random variable, i.e. $1 - \Phi(z_x) = x$ where $\Phi()$ is the cumulative distribution function of a normal random variable. Under assumption that Z is a multivariate Gaussian random variable we have that the posterior probability of error is less than or equal to δ provided it holds

$$\frac{q_i - q_j}{\sqrt{\mathrm{E}((Z_i - Z_j)^2)^2}} \geq z_\Delta \text{ for every } i \in \mathcal{T}, j \in \mathcal{B}$$

where $\Delta|\mathcal{T}||\mathcal{B}| = \delta$. Using (24) note that the last inequality is equivalent to

$$\begin{aligned} t &\geq \max_{i\in\mathcal{T},j\in\mathcal{B}}\frac{q_i+q_j}{(q_i-q_j)^2}z_\Delta^2 \\ &= \frac{q_k+q_{l(k)}}{(q_k-q_{l(k)})^2}z_\Delta^2. \end{aligned} \tag{25}$$

This yields a stopping rule for the top-k problem; see Figure 1 for a pseudo code of the sequential algorithm. Note

that (25) yields similar number of samples as that in Theorem 4.1.[2]

Algorithm 1 Sequential sampling for top-k

Input: δ, ϵ
Procedure: $f_k(Q) = \frac{q_k + q_{l(k)}}{(q_k - q_{l(k)})^2}$
Init: $Q = 0$, $T = 1$
while 1 **do**
    $x \leftarrow$ sample of an item from the dataset X
    $Q \leftarrow \left(1 - \frac{1}{T}\right) Q$
    $q_x \leftarrow q_x + \frac{1}{T}$
    if $T > f_k(Q) z_\Delta^2$ **then** break
    end if
    $T \leftarrow T + 1$
end while
Output: a set of k items with largest frequencies q_i

Top-k with Frequencies. We preserve notation with the change $\Delta(Q) := \frac{\delta(q_k - \epsilon)}{2}$. Same stopping rule applies as for top-k but with $f_k(Q) = \frac{\max_i q_i(1-q_i)}{\epsilon^2}$. The stopping rule is derived by choosing t so that the posteriori error probability given by $p_e = \Pr(\cup_i |p_i - q_i| > \frac{\epsilon}{2} | Q)$ is bounded by δ.

5. NUMERICAL RESULTS

In this section we present numerical results that we obtained by simulations and using frequencies of items that we inferred from datasets of an operational cloud service. The goals of this section are: (1) to demonstrate tightness of our bounds by comparison to empirical counterparts; (2) to demonstrate sampling gains by using direct construction of arbitrary-height range partitioning; (3) to demonstrate sampling gains of approximate solving of top-k problems and evaluate the sampling complexity with respect to k; (4) evaluate the gain of using the lumping of small frequency items; (5) demonstrate the validity of the sequential sampling algorithm for top-k set problems.

Data. Our data consists of a database of an operational cloud service for provisioning of media content that is used by hundreds of thousands of users and involves tens of millions of media items. The database consists of tens of tables and more than a hundred of data columns; the datasets that we use in our analysis consists of more than fifty columns from this database whose number of rows span the range of a few million rows to hundreds of billions. These datasets would be representative of typical online services where various events are recorded over time associated to individual users and product items.

Range partitioning. For range partitioning, we use simulations on artificial datasets as the results apply to any set of items whose values are from an ordered set of items. In Fig. 4 we show the estimated probability of error versus the sample size alongside with our bounds (Corollary 4.1) for three-bin range partitioning. The results confirm tightness of our bounds and illustrate the fact that the probability of error is larger for smaller minimum height of the partitioning.

[2]To see this, note that $\frac{q_k + q_{l(k)}}{(q_k - q_{l(k)})^2} \leq 2q_k/\epsilon^2$ and $z_\Delta^2 \leq 2\log(\frac{1}{\Delta})$ and using these we recover the same number of samples as in Corollary 4.1.

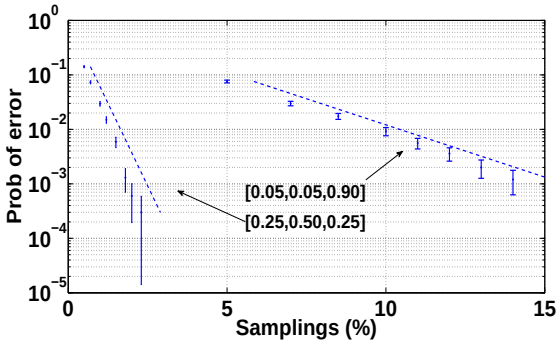

Figure 4: Probability of error vs. sample size for three-bin range partitionings; the sample size is given as the percentage of the total number of items in the dataset.

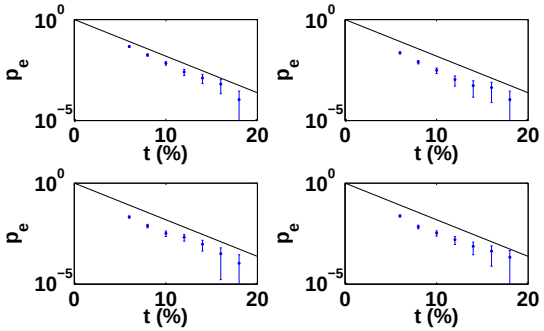

Figure 5: Range partitioning for different histogram heights P but common minimum height $\min_i p_i$.

The latter is further illustrated in Fig. 5 where we observe same decay rate of the probability of error with the sample size for different range partitions $P = [0.05, 0.05, 0.9]$, $[0.05, 0.25, 0.7]$, $[0.05, 0.35, 0.6]$, and $[0.05, 0.45, 0.5]$ that all have common minimum bin-height $\min_i p_i = 0.05$.

In Fig. 6 we evaluate the benefit of using the direct estimation method for arbitrary-height range partitioning over the indirect approach. Recall that in Section 3.1 we already showed that the indirect approach requires the number of samplings that is at least a factor given in Eq. (5) of that under the direct approach. In Fig. 6 we evaluate (5) for different partitionings as follows. Given positive integers M and k we sample partitions of M into k bins uniformly at random from the set of all possible partitions of M into k partitions. For each given M and k we take 10^4 such samples and compute the mean and standard deviation of the values (5). Fig. 6 indicates that the benefit of the direct approach is larger for larger values of M and smaller number of the bins k and it can be substantial.

Top-k. We first compare our bounds for the sample size of the top-k set problems. In Fig. 7 we show the estimated probability of error versus the sample size for top-1 set and top-10 set problems along with analytical bounds (Theorem 4.2). The results confirm that error probabilites from simulations are indeed within the bounds (plotted in broken lines). We confirm that the required sample size for top-k set problem with given absolute error tolerance ϵ and proba-

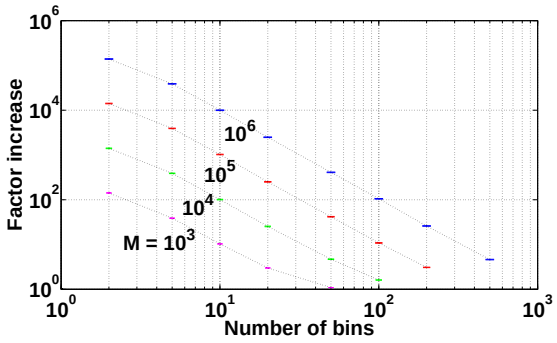

Figure 6: The benefit of the direct estimation method over the indirect approach for range partitioning.

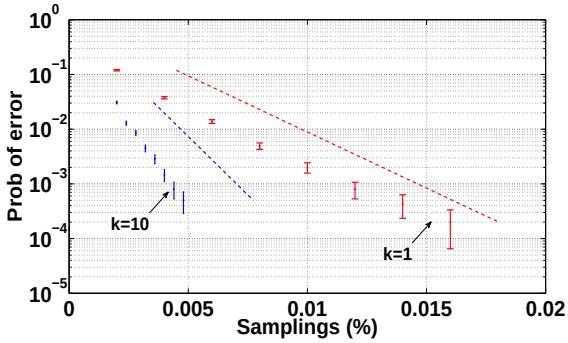

Figure 7: Probability of error vs. sample size for top-1 set and top-10 set problems with absolute error tolerance $\epsilon = 1.8 \cdot 10^{-3}$. The corresponding frequencies are $p_1 = 6.9 \cdot 10^{-3}$ and $p_{10} = 3.3 \cdot 10^{-3}$.

bility of error scales as p_k. Indeed, in Fig. 7 observe that for given probability of error, the top-1 set requires about twice as many samples as the top-10 set problem which conforms to the value of the frequency p_1 being about twice the value of the frequency p_{10}.

A similar plot is shown in Fig. 8 for the top-1 set with frequencies problem and for two distinct values of the absolute error tolerance ϵ. The results confirm validity of our bounds (Theorem 4.7) and further demonstrate that the required sample size for given probability of error scales as $1/\epsilon^2$; this is indicated in the figure as the required sample size for given probability of error under $\epsilon = 0.9 \cdot 10^{-3}$ is about four times of that under $\epsilon = 1.8 \cdot 10^{-3}$.

The next set of plots quantify the difference between $1/p_k$ and m for practical data sets. The distribution of the values of the ratio $(1/p_k)/m$ over a range of datasets is plotted in Fig. 9. It is observed that the median values are smaller than $1/100$ for a wide range of values of k. Since our sampling costs scale as $\log(\min(1/p_k, m))$, the bounds we obtain from lumping arguments are tighter by a factor of 5 as compared to the bounds from standard arguments which is a significant reduction.

The sampling costs that are established in Section 4 for different top-k problems depend on the frequencies of items

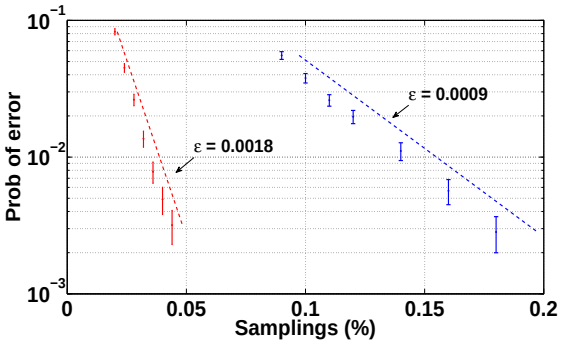

Figure 8: Probability of error vs. sample size for top-1 set problem with absolute error tolerances ϵ as indicated. The frequency of the most frequent item in the dataset is $p_1 = 6.9 \cdot 10^{-3}$.

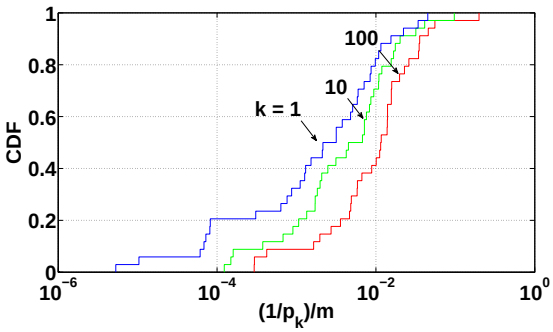

Figure 9: CDF of the values of the ratio $(1/p_k)/m$ for $k = 1$, 10, and 100.

in a dataset. This can provide a useful guideline in understanding the sampling costs for the different variants of the top k problem. In particular, we have shown that for absolute tolerance guarantees, the ratio of sampling complexity for the top-k set problem and the top-k ranking (or frequency) problem varies as p_k/p_1. In Fig. 10 we show the ratio p_k/p_1 versus k for three different datasets. In particular, we observe the value of p_k/p_1 of about one half for $k = 10$ which indicates that solving the top-10 set problem with frequencies can require twice as many samples as compared to identifying just the top-10 set.

Finally, we demonstrate validity of the sequential sampling algorithm that we described in Section 4.5. Specifically, we consider top-k set problem for $k = 1$, absolute error tolerance $\epsilon = 1.8 \cdot 10^{-3}$, and estimate the probability of error based on 20,000 simulation repetitions. Fig. 11 shows the probability versus the sample size alongside with input probability of error δ. The simulation and analytical results are in conformance.

6. CONCLUSION

We considered the estimation of arbitrary-height range partitioning and identification of frequently occuring items, two basic computational tasks that are of interest in the context of data-intentive compuations; the results provide insights into sampling complexity and simple and practical sampling algorithms.

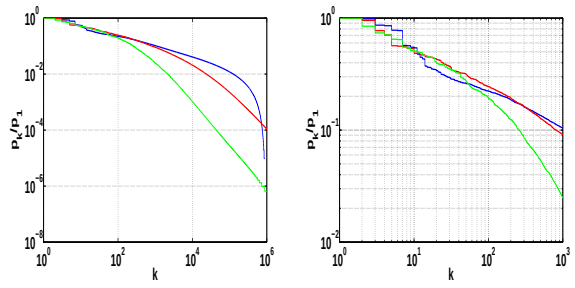

Figure 10: p_k/p_1 vs. k; the right graph is a zoomed version of the left graph.

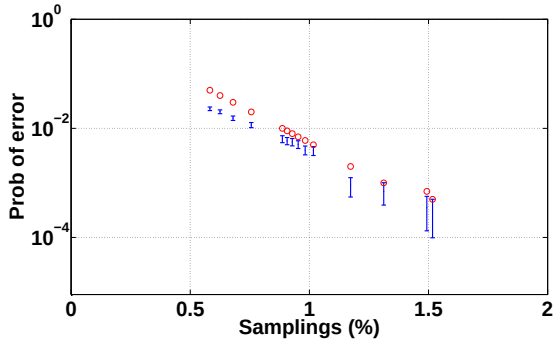

Figure 11: Sequential sampling for top-k. The circles correspond to the input probability of error δ.

7. REFERENCES

[1] Amazon. Elastic mapreduce. http://aws.amazon.com/elasticmapreduce/, 2009.

[2] A. Baumann, P. Barham, P.-E. Dagand, T. Harris, R. Isaacs, S. Peter, T. Roscoe, A. Schupbach, and A. Sighania. The multikernel: A new os architecture for scalable multicore systems. In *Proc. of SOSP'09*, 2009.

[3] J. O. Berger. *Statistical Decision Theory and Bayesian Analysis*. Springer, 2 edition, 1985.

[4] M. Budiu. Private communication, 2009.

[5] R. Chaiken, B. Jenkins, P. Larson, and B. Ramsey. Scope: Easy and efficient parallel processing of massive data sets. In *Proc. of ACM PVLDB'08, Aucland, New Zeland*, August 2008.

[6] M. Charikar, K. Chen, and M. Farach-Colton. Finding frequent items in data streams. In *Proc. of the 29th Int'l Colloquium on Automata, Languages and Programming*, volume 2380, pages 683–703, 2002.

[7] S. Chaudhuri, R. Motwani, and V. Narasayya. Random sampling for histogram construction: how much is enough? In *Proc. of ACM SIGMOD*, volume 27, pages 436–447, June 1998.

[8] E. Cohen and H. Kaplan. Leveraging discarded samples for tighter estimation of multiple-set aggregates. In *Proc. of ACM Sigmetrics'09, Seattle, WA*, June 2009.

[9] G. Cormode and S. Muthukrishnan. An improved data stream summary: The count-min sketch and its applications. *LATIN 2004: Theoretical Informatics*, 2976:29–38, 2004.

[10] D. R. Cox and D. V. Hinkley. *Theoretical Statistics*. Chapman & Hall/CRC, 1974.

[11] J. Dean and S. Ghemawat. Mapreduce: Simplified data processing on large clusters. In *Proc. of OSDI'04 (Sixth Symp. on Operating System Design and Implementation), San Francisco, CA*, December 2004.

[12] A. Dembo and O. Zeitouni. *Large Deviations Techniques and Applications*. Springer, 2 edition, 1998.

[13] The Economist. Clash of the clouds. October 17th-23RD, 2009.

[14] P. Flajolet and N. Martin. Probabilistic counting. In *Proc. of FOCS '83*, pages 76–82, 1983.

[15] P. B. Gibbons and Y. Matias. New sampling-based summary statistics for improving approximate query answers. In *Proc. of ACM SIGMOD '98*, pages 331–342, 1998.

[16] M. Greenwald and S. Khanna. Space-efficient online computation of quantile summaries. In *Proc. of ACM SIGMOD '01*, pages 56–66, 2001.

[17] S. Guha and A. McGregor. Approximate quantiles and the order of the stream. In *Proc. of ACM PODS'06*, June 2006.

[18] S. Guha and A. McGregor. Space-efficient sampling. In *Proc. of AISTATS*, pages 169–176, 2007.

[19] hadoop. Hadoop. http://hadoop.apache.org/, 2009.

[20] Q. Hardy. A zure thing? Forbes, November 2009.

[21] M. Isard, M. Budiu, Y. Yu, A. Birrell, and D. Fetterly. Dryad: Distributed data-parallel programs from sequential building blocks. In *Proc. of Eurosys '07*, 2007.

[22] A. Kirsch, M. Mitzenmacher, A. Pietracaprina, G. Pucci, E. Upfal, and F. Vandin. A rigorous statistical approach for identifying significant itemsets. In *Proc. of IEEE ICDM'08*, 2008.

[23] F. Kuhn, T. Locher, and S. Schmid. Distributed computation of the mode. In *Proc. of ACM Symp. on PODC*, pages 15–24, 2008.

[24] G. S. Manku and R. Motwani. Approximate frequency counts over data streams. In *Proc. of VLDB '02*, pages 346–357, 1996.

[25] G. S. Manku, S. Rajagopalan, and B. G. Lindsay. Random sampling techniques for space efficient online computation of order statistics of large datasets. In *Proc. of ACM SIGMOD '99*, volume 28, pages 251–262, June 1999.

[26] S. Mannor and J. N. Tsitsiklis. The sample complexity of exploration in the multi-armed bandit problem. *Journal of Machine Learning Research*, 5:623–648, 2004.

[27] U.S. Department of Energy. Data-intensive computing to large science discoveries. http://www.eurekalert.org/features/doe/2005−12/dnnl−, 2009.

[28] A. Das Sarma, S. Gollapudi, and R. Panigrahy. Estimating pagerank on graph streams. In *Proc. of ACM PODS '08*, pages 69–78, 2008.

[29] Yahoo! Yahoo! hadoop tutorial. http://developer.yahoo.com/hadoop/tutorial/, 2009.

APPENDIX

Proof of Theorem 4.1

The error probability p_e satisfies the following

$$p_e = \Pr\left(\cup_{i\in\mathcal{T},j\in\mathcal{B}}\{q_j \geq q_i\}\right) \leq \sum_{i\in\mathcal{T},j\in\mathcal{B}} \Pr(q_j \geq q_i) \quad (26)$$

where the inequality is by the union bound.

For the probability of the event $\{q_j \geq q_i\}$, for any $i \in \mathcal{T}$ and $j \in \mathcal{B}$, using Chernoff's bound we have the following lemma

whose proof is in Appendix.

LEMMA .1. *For any $i \in \mathcal{T}$ and $j \in \mathcal{B}$ the following holds*

$$Pr(q_j \geq q_i) \leq \left(1 - (\sqrt{p_i} - \sqrt{p_j})^2\right)^t.$$

We use the following uniform bound for the probabilities $\Pr(q_j \geq q_i)$, $i \in \mathcal{T}$, $j \in \mathcal{B}$,

$$\begin{aligned}\max_{i\in\mathcal{T},j\in\mathcal{B}} \Pr(q_j \geq q_i) &\leq \max_{i\in\mathcal{T},j\in\mathcal{B}} \left(1 - (\sqrt{p_i} - \sqrt{p_j})^2\right)^t \\ &= \left(1 - (\sqrt{p_k} - \sqrt{p_{l(k)}})^2\right)^t. \qquad (27)\end{aligned}$$

The last equality follows since $\sqrt{p_i} - \sqrt{p_j}$ is minimized by choosing i as large as possible within the set $\mathcal{T}$ and j as small as possible within the set $\mathcal{B}$. Thus, $p_i = p_k$ and $p_j = p_{l(k)}$ where $l(k)$ is the item with largest frequency in $\mathcal{B}$.

From (26) and (27), we have

$$\begin{aligned}p_e &\leq \sum_{i\in\mathcal{T},j\in\mathcal{B}} \Pr(q_j \geq q_i) \\ &\leq |\mathcal{T}||\mathcal{B}|\left(1 - (\sqrt{p_k} - \sqrt{p_{l(k)}})^2\right)^t \\ &\leq k(m-k)\left(1 - (\sqrt{p_k} - \sqrt{p_{l(k)}})^2\right)^t\end{aligned}$$

where the last inequality is because $|\mathcal{T}| = k$ and $|\mathcal{B}| \leq m-k$.

The proof is completed by noting that as we use Chernoff's bound, by Cramer's theorem the error exponent is tight. □

Proof of Lemma .1

Let $I_x(y)$ be an indicator defined by $I_x(y) = 1$ if $y = x$ and $I_x(y) = 0$, otherwise. Suppose $(y_1, y_2, \ldots, y_t)$ are the items of the sketch. Then,

$$\Pr(q_j \geq q_i) = \Pr\left(\sum_{k=1}^{t} Z_k \geq 0\right). \qquad (28)$$

Note that $Z_k := I_j(y_k) - I_i(y_k)$, $k = 1, 2, \ldots, t$, is a sequence of independent and identically distributed random variables. By Chernoff's bound we have, for any real number x and any positive integer t,

$$\Pr\left(\sum_{k=1}^{t} Z_k \geq xt\right) \leq \exp\left(-t \sup_{\theta>0} \Lambda(\theta, x)\right) \qquad (29)$$

where

$$\Lambda(\theta, x) = \theta x - \log \mathrm{E}(e^{\theta Z_1}).$$

Now, note

$$e^{\theta Z_1} = \begin{cases} e^{\theta} & \text{with prob. } p_j \\ e^{-\theta} & \text{with prob. } p_i \\ 1 & \text{with prob. } 1 - p_i - p_j. \end{cases}$$

It follows

$$\mathrm{E}\left(e^{\theta Z_1}\right) = 1 - p_i - p_j + e^{\theta} p_j + e^{-\theta} p_i$$

and

$$\Lambda(\theta, x) = \theta x - \log\left(1 - p_i - p_j + e^{\theta} p_j + e^{-\theta} p_i\right). \qquad (30)$$

Let θ_x be the maximizer in (29) for given x. We are interested in $\Lambda(\theta_x, x)$ for $x = 0$. We will show that the following holds

$$\Lambda(\theta_0, 0) = -\log\left(1 - (\sqrt{p_i} - \sqrt{p_j})^2\right) \qquad (31)$$

The maximizer θ_x satisfies $\frac{\partial}{\partial\theta}\Lambda(\theta, x) = 0$ at $\theta = \theta_x$. Combining with

$$\frac{\partial}{\partial\theta}\Lambda(\theta, x) = x - \frac{p_j e^{\theta} - p_i e^{-\theta}}{1 - p_i - p_j + p_j e^{\theta} + p_i e^{-\theta}}$$

we obtain

$$(1-x)p_j e^{2\theta_x} - x(1 - p_i - p_j)e^{\theta_x} - (1+x)p_i = 0.$$

This is a quadratic equation for e^{θ_x} whose solution for $x = 0$ is $e^{\theta_0} = \sqrt{\frac{p_i}{p_j}}$. Plugging this in (30) for $x = 0$, we obtain (31). This completes the proof.

Theorem 4.2

Whenever an item in $\mathcal{B}$ is sampled more often than an item $i \in \mathcal{T}$ with respect to the distribution P, this implies that the corresponding super-item in $\tilde{B}$ is sampled more often than item i with respect to the distribution $\tilde{P}$. Therefore,

$$\begin{aligned}p_e &= \Pr_P\left(\cup_{i\in\mathcal{T},j\in\mathcal{B}}\{q_j \geq q_i\}\right) \\ &\leq \Pr_{\tilde{P}}\left(\cup_{i\in\mathcal{T},j\in\tilde{\mathcal{B}}}\{q_j \geq q_i\}\right) \qquad (32)\end{aligned}$$

where $\Pr_P(\cdot)$ and $\Pr_{\tilde{P}}(\cdot)$ are the probabilities under the sampling from the distribution P and the distribution $\tilde{P}$, respectively. The rest of the proof follows the same steps as that of Theorem 4.1. □

Proof of Theorem 4.6

From the arguments as for the absolute error guarantees, it holds that

$$p_e \leq kL'\left(1 - \min_{i\leq k}(\sqrt{p_i} - \sqrt{p_{l(i)}})^2\right)^t$$

where $l(i)$ is the element with maximum frequency less than $p_i(1-\epsilon)$. Note that

$$(\sqrt{p_i} - \sqrt{p_{l(i)}})^2 \geq \frac{\epsilon^2 p_i^2}{(\sqrt{p_i} + \sqrt{p_i(1-\epsilon)})^2}.$$

Therefore, it holds that

$$\begin{aligned}p_e &\leq kL'\left(1 - \min_{i\leq k}(\sqrt{p_i} - \sqrt{p_{l(i)}})^2\right)^t \\ &\leq kL'\left(1 - \min_{i\leq k}\frac{\epsilon^2 p_i^2}{(\sqrt{p_i} + \sqrt{p_i(1-\epsilon)})^2}\right)^t \\ &= kL'\left(1 - \frac{\epsilon^2 p_k^2}{(\sqrt{p_k} + \sqrt{p_k(1-\epsilon)})^2}\right)^t \\ &= kL'\left(1 - (\sqrt{p_k} - \sqrt{p_k(1-\epsilon)})^2\right)^t.\end{aligned}$$

□

Proofs of Fact 4.1 and Fact 4.2

We use the following lemma.

LEMMA .2. *For every fixed $a \in [0, \frac{1}{2}]$, $D(x+a||x)$ achieves a minimum over $x \in [0, 1-a]$ at a point p_a^* that lies in the interval $[\frac{1-a}{2}, \frac{1}{2}]$.*

PROOF. Let $f(x) := D(x+a||x)$ for $x \in [0, 1-a]$. We have

$$f'(x) = \log\left(\frac{(x+a)(1-x)}{(1-x-a)x}\right) - \frac{a}{x(1-x)}.$$

Let

$$f_1(x) := \frac{(1-x)(x+a)}{x(1-(x+a))} \text{ and } f_2(x) := e^{\frac{a}{x(1-x)}}.$$

Condition $f'(x) = 0$ is equivalent to $f_1(x) = f_2(x)$.

On the one hand, function $f_1(x)$ is non-negative and tends to infinity at 0 and $1-a$. It is readily checked that $f_1(x)$ has a unique minimum at the point $(1-a)/2$. On the other hand, function $f_2(x)$ over $[0,1]$ has a unique minimum at $\frac{1}{2}$, is symmetric around $1/2$, and tends to infinity at 0 and 1. Therefore, it must be that $f_1(x)$ and $f_2(x)$ intersect at a point in the interval $[\frac{1-a}{2}, \frac{1}{2}]$ which completes the proof. □

Proof of Fact 1: We will establish that for $p^*_{\frac{\epsilon}{2}}$ defined in Lemma .2 the following holds,

$$p^*_{\frac{\epsilon}{2}} \geq \min(p_k, 1-p_k) - \epsilon$$

from which the result follows. By Lemma .2, $p^*_{\frac{\epsilon}{2}} \geq \frac{1}{2} - \frac{\epsilon}{4}$ which we use in the following. For $k > 1$, $\min(p_k, 1-p_k) = p_k$ and thus suffices that $p_k - \epsilon \leq \frac{1}{2} - \frac{\epsilon}{4}$ which indeed holds as $p_k \leq 1/2$. For $k = 1$ and $p_1 \leq \frac{1}{2}$, the same argument applies. Finally, for $k = 1$ and $p_1 > \frac{1}{2}$, $\min(p_k, 1-p_k) = 1-p_1$ hence it suffices that $1 - p_1 - \epsilon \leq \frac{1}{2} - \frac{\epsilon}{4}$ which indeed holds under $p_1 > \frac{1}{2}$. □

Proof of Fact 2: Note that $\min_i D_i$ is achieved for an item i such that $\min(p_i, 1-p_i)$ is either first to the left or first to the right of the point $p^*_{\frac{\epsilon}{2}}$. Hence, if $\min(p_i, 1-p_i) \leq p^*_{\frac{\epsilon}{2}}$ for every i, then D_i is smallest for an item i with largest $\min(p_i, 1-p_i)$, which if $p_1 < 1/2$ is item 1. In view of the lower bound on $p^*_{\frac{\epsilon}{2}}$ in Lemma .2, D_i is smallest for $i = 1$ provided that $p_1 \leq (1 - \frac{\epsilon}{2})/2$. Eq. (20) follows. □

Proof of Theorem 4.7

For the probability of error p_e we have

$$\begin{aligned} p_e &= \Pr\left(\cup_{i=1}^m \left\{|q_i - p_i| \geq \frac{\epsilon}{2}\right\}\right) \\ &\leq m \max_i \Pr\left(|q_i - p_i| \geq \frac{\epsilon}{2}\right). \end{aligned}$$

Note

$$\begin{aligned} &\Pr\left(|q_i - p_i| \geq \frac{\epsilon}{2}\right) \\ \leq\ &\Pr\left(q_i - p_i \geq \frac{\epsilon}{2}\right) + \Pr\left(q_i - p_i \leq -\frac{\epsilon}{2}\right). \end{aligned}$$

Indeed, q_i is the empirical mean based on t samples from a Bernoulli distribution with mean p_i. By Chernoff's bound, we have, for $\epsilon \leq 2(1-p_i)$,

$$\Pr\left(q_i - p_i \geq \frac{\epsilon}{2}\right) \leq \exp\left(-tD(p_i + \frac{\epsilon}{2}||p_i)\right)$$

and, for $\epsilon \leq 2p_i$,

$$\Pr\left(q_i - p_i \leq -\frac{\epsilon}{2}\right) \leq \exp\left(-tD(p_i - \frac{\epsilon}{2}||p_i)\right).$$

Therefore, we have (17) with

$$D_i := \min\left(D(p_i + \frac{\epsilon}{2}||p_i), D(p_i - \frac{\epsilon}{2}||p_i)\right).$$

We note the following fact whose proof is provided in Appendix.

FACT .1.

$$D_i = \begin{cases} D(p_i + \frac{\epsilon}{2}||p_i), & p_i \leq \frac{1}{2} \\ D(p_i - \frac{\epsilon}{2}||p_i), & p_i > \frac{1}{2}. \end{cases}$$

Noting that $D(p_i - \frac{\epsilon}{2}||p_i) = D(1 - p_i + \frac{\epsilon}{2}||1-p_i)$, it follows that D_i satisfies (18).

Since we used Chernoff's bounds the error exponent is tight so (19) holds. □

Proof of Fact .1

The claim follows once we establish the following: for every fixed $y \in [0, \frac{1}{2}]$,

$$D(y+x||y) \leq D(y-x||y), \text{ for every } x \in [0, y]. \quad (33)$$

Note that this also means that if $y \in [\frac{1}{2}, 1]$, then $D(y + x||y) \geq D(y-x||y)$, for every $x \in [0, 1-y]$. This follows from (33) by using the substitution $z = 1 - y$ and using the correspondences $D(y+x||y) = D(z-x||z)$ and $D(y-x||y) = D(z+x||z)$.

Let $f(x) := D(y+x||y) - D(y-x||y)$ for $x \in [0, y]$. It is not difficult to show that

$$f'(x) = \log\left(\frac{(1-y)^2}{y^2}\frac{y^2 - x^2}{(1-y)^2 - x^2}\right).$$

Function $f(x)$ is equal to 0 at $x = 0$. Therefore, (33) follows if $f'(x) \leq 0$ for every $x \in [0, y]$. But note that this is equivalent to saying

$$\frac{y^2 - x^2}{(1-y)^2 - x^2} \leq \frac{y^2}{(1-y)^2}$$

which indeed holds as equality holds for $x = 0$ and the left-hand side is non-increasing with x under our assumption $y \leq \frac{1}{2}$.

Proof of Theorem 4.9

The error probability p_e is upper bounded as

$$\begin{aligned} p_e \leq \Pr\Bigg(&\bigcup_{i: p_i > p_k(1-\epsilon)} \left\{|q_i - p_i| \geq p_i \frac{\epsilon}{2}\right\} \\ &\bigcup_{i: p_i \leq p_k(1-\epsilon)} \left\{|q_i - p_i| \geq p_k \frac{\epsilon}{2}\right\}\Bigg). \end{aligned}$$

Observe that in the above expression, lower frequency elements ($p_i \leq p_k - \epsilon$) can deviate upto absolute amount $\frac{\epsilon p_k}{2}$. This condition along with the restriction on the deviation of the counts of the top frequency ($p_i \geq p_k$) elements ensures that the count of lower frequency elements does not exceed the count of the higher frequency elements. For $k \geq 2$ or for $k = 1$ and $p_1 < 0.5$, it can be easily checked that the error probability is bounded by (23). □