Adaptive threshold sampling

Daniel Ting
Tableau Software
Seattle, Washington
dting@tableau.com

ABSTRACT

Sampling is a fundamental problem in computer science and statistics. However, for a given task and stream, it is often not possible to choose good sampling probabilities in advance. We derive a general framework for adaptively changing the sampling probabilities via a collection of thresholds. In general, adaptive sampling procedures introduce dependence amongst the sampled points, making it difficult to compute expectations and ensure estimators are unbiased or consistent. Our framework address this issue and further shows when adaptive thresholds can be treated as if they were fixed thresholds which samples items independently. This makes our adaptive sampling schemes simple to apply as there is no need to create custom estimators for the sampling method.

Using our framework, we derive new samplers that can address a broad range of new and existing problems including sampling with memory rather than sample size budgets, stratified samples, multiple objectives, distinct counting, and sliding windows. In particular, we design a sampling procedure for the top-K problem where, unlike in the heavy-hitter problem, the sketch size and sampling probabilities are adaptively chosen.

ACM Reference Format:
Daniel Ting. 2019. Adaptive threshold sampling. In *2019 International Conference on Management of Data (SIGMOD '19), June 30-July 5, 2019, Amsterdam, Netherlands.* ACM, New York, NY, USA, 13 pages. https://doi.org/10.1145/3299869.3319897

1 INTRODUCTION

Sampling is a fundamental problem in computer science and statistics. By reducing the amount of data processed, it can significantly improve performance and lower costs, or it can ensure that the data processed fits within a system's resource constraints. Of particular interest are random sampling without replacement procedures as they do not sample redundant information.

Before observing the data, it is not possible to choose appropriate sampling probabilities. For example, given a weighted stream of unknown length and some memory budget B, it is not possible to choose a sampling probability ahead of time which ensures it satisfies the budget B. Furthermore, to guarantee the budget is satisfied, each item's inclusion decreases the available budget and affects the appropriate inclusion probability of other items.

SIGMOD '19, June 30-July 5, 2019, Amsterdam, Netherlands

ACM ISBN 978-1-4503-5643-5/19/06.
https://doi.org/10.1145/3299869.3319897

This dependence causes several difficulties when designing sampling procedures and deriving estimators from the sample. In particular, the dependence often makes it intractable to compute the inclusion probability for each item or sets of items. When these sampling probabilities cannot be estimated, the sample is almost useless in data analysis. No good estimates can be obtained since estimators must be able to adjust the contribution of each item based on its inclusion probability.

These difficulties can extend to designing sampling procedures. When the sample size is not fixed, items can be drawn independently with weight w_i with probability $\pi_i \propto w_i$. For example, the Conditional Poisson Sampling scheme is one that draws a fixed size sample and has desirable properties; however, no known algorithm can efficiently draw Conditional Poisson samples.

We propose a framework, *adaptive threshold sampling*, that addresses all of these challenges. We use it to solve novel problems and improve existing solutions. In this framework, samples are easy to draw; sample sizes and probabilities can change on the fly; and good estimators can be derived even though samples are dependent. This framework mimics drawing independent (Poisson) samples. Each item x is associated with an independent random value R_x and a threshold T_x. The item is included in the sample if $U_x < T_x$. However, we adjust the threshold in a data and sample dependent way to obtain desirable properties for the sample.

Our methodological contributions revolve around making it easy to build thresholds where the resulting sample is easy to analyze. We establish conditions when the threshold T_x can be treated *as if* it was a fixed threshold that yields an independent sample. This simplifies analysis of the sample since one can simply apply an existing unbiased estimator for independent samples. Deriving and analyzing a custom estimator based on the true sampling distribution becomes unnecessary. For cases where the conditions do not hold, we introduce a more general notion of *threshold recalibration* that makes it easy to compute expectations and derive new unbiased estimators for a broader class of thresholding rules. We also provide methods for building and composing thresholds and for merging sampling. We also prove an empirical process convergence result that further simplifies the development of good estimators and sampling designs. It extends our theory for unbiased estimators and shows when consistent estimators for independent samples remain consistent when applied to adaptive thresholding samples. It also provides justification for heuristic thresholding schemes that do not satisfy our conditions for unbiased estimation but satisfy an asymptotic convergence condition which may be easier to verify.

Our contributions to applications exploits our methodological contributions to develop new or improved sketches and estimators on a range of new and existing problems. For example, we double the effectiveness of the state-of-the-art in sampling from sliding windows [14] even though we use exactly the same sketch

to construct the sample. We improve merge procedures for distinct counting sketches. We handle sampling budget constraints given variable length items. We provide a solution to a novel top-k problem, a more challenging variation of the well-studied heavy hitter problem where the top-k items by frequency must be returned regardless of how small their frequency may be.

1.1 Related work

A long line of work has studied the bottom-k sample [7, 8, 12, 23, 24], showing both how it can be used to draw weighted samples and deriving good estimators for sums based on the samples as well as estimates of their variance. These bottom-k sampling methods can be considered an adaptive threshold sampling method where the threshold ensures the sample size is exactly k. Extensions [6, 9] study combining bottom-k samples.

Little work has generalized the bottom-k thresholding rule while providing a unbiased estimators. However, in distinct counting applications using $Uniform(0, 1)$ priorities, the Theta sketch [11] provides a general 1-goodness condition. Thresholding rules that satisfy this condition ensure that cardinality estimates are unbiased. For uniform sampling, a generalized thresholding rule has been used for distributed sampling [30].

Our contributions consist of both providing a framework for adaptive threshold sampling as well as using it to address novel problems and improve existing methods. Novel problems include sampling with fixed memory budgets, multi-stratified sampling, and top-k queries. These problems are more difficult variations of existing problems. Examples of existing sketches and sampling methods that are improved include multi-objective sampling [6], sampling in sliding windows [14], distinct counting [11, 15, 28]. Our applications are provided simply as examples of our framework's usefulness. Bottom-k methods have been used in numerous other applications such as set similarity [5], networking [25], time-decayed sampling [10], and distributed sampling [13, 18, 30]

2 SAMPLING DESIGN AND ESTIMATION

We first review sampling without replacement and and the use of fixed thresholds for drawing samples. One key contribution of our paper is under what conditions can a random, adaptive threshold be treated as if it was a fixed threshold.

The notion of weighted or unequal probability sampling is at the core of many sampling problems. Sampling more informative items with higher probability leads to lower variance estimates. Choosing an appropriate measure of informativeness allows for accurate estimation for problems ranging from simple aggregates to complex machine learning models [**?**]. Unequal probability sampling can also arise other situations, such as in stratified sampling where strata can have different sampling probabilities or in distributed sampling where nodes may make independent choices about the sampling rate.

2.1 Poisson sampling with thresholds

Of particular interest are sampling schemes where the inclusion decisions for each item is independent of all others. That is

$$Z_i \sim Bernoulli(\pi_i). \tag{1}$$

Such schemes are called Poisson Sampling designs. This is not to be confused with drawing Poisson distributed random variables. In this case, the desired inclusion probability π_i for each x_i must be known in advance.

Drawing such a sample can be performed using a fixed threshold. Associate an independent auxiliary variable R_i for each item along with a fixed threshold T_i. The item x_i is included if $R_i < T_i$. If R_i is continuous and F_i is its cumulative distribution function (CDF), then the probability x_i is sampled is $p(R_i < T_i) = F_i(T_i)$. Choosing a threshold such that $\pi_i = F_i(T_i)$ yields a sample from the desired sampling design. We call the variable R_i the priority of x_i and denote the inclusion of x_i by $Z_i = 1(R_i < T_i)$. From this, it is easy to see how the threshold T_i can be used to adjust the inclusion probabilities.

2.2 Sampling challenges and estimation

However, unequal probability sampling can lead to challenges in sampling and estimation. For example, when there are memory budget or sample size constraints, Poisson sampling can violate the constraints since there is some non-zero probability that all items are sampled. It is thus natural to consider samplers that draw fixed size samples. The natural extension of Poisson sampling to fixed sizes is Conditional Poisson Sampling (CPS), which is obtained from a Poisson Sampling design conditional on the sample size being exactly k. The Conditional Poisson Sampling design has the attractive property of being the maximum entropy sampling procedure for a set of inclusion probabilities. However, there is no efficient algorithm known for drawing a CPS sample or for computing the inclusion probabilities [27].

Estimating quantities of interest can be even more problematic. Good estimators are crucial as a sample is almost useless without them. In unequal probability sampling designs, an unbiased estimate of the population total S is given by the Horvitz-Thompson (HT) estimator,

$$\hat{S} = \sum_i x_i \frac{Z_i}{\pi_i} \tag{HT}$$

where Z_i indicates if the item x_i is included in the sample and $\pi_i = p(Z_i = 1)$ is the inclusion probability. This also provides a solution to the subset sum problem [12] by zeroing any value x_i that is not in the desired subset. The fundamental problem for Conditional Poisson sampling and other sampling schemes with dependence between items' inclusion is that this dependence makes inclusion probabilities and good estimators difficult to derive.

If inclusion probabilities can be computed, one can search for the optimal sampling design given this estimator. When the inclusion probabilities $\pi_i \propto x_i$, each term in the HT estimator is constant. Furthermore, if the sample size is fixed at k, the HT estimator itself is constant, and thus has minimal variance. A sample that draws elements with probability proportional to some size x_i is called a probability proportion to size (PPS) sample.

2.3 Adaptive threshold sampling

Designing good sampling procedures is further complicated by the fact that the desired inclusion probabilities are often not known in advance. They may depend on unknown properties of data stream such as length or the variance of the data.

Our solution to the challenge of adaptively choosing the desired probabilities is to replace the fixed threshold generating an independent Poisson sampling scheme with a an adaptive threshold that can depend on the data. For example, adaptively choosing the threshold can ensure that the sample fits inside a memory budget regardless of the size of the stream.

Mathematically, we define an adaptive threshold $T_i = \tau_i(\mathbf{R}|\mathcal{D})$ to be a function τ_i of the data $\mathcal{D}$ and priorities $\mathbf{R}$ that determine the sample. Like Conditional Poisson Sampling, this dependence between the threshold and the priorities makes it difficult to compute the inclusion probability $p(Z_i = 1) = p(R_i < T_i)$ needed by the HT estimator and can make unbiased estimators difficult to generate. For example, suppose the data consists of individuals' demographic information. In an extreme case, consider the threshold $T_i := \min\{R_j : gender_j = \text{Female}\}$. Such a sample is grossly biased as it would exclude all females. Generating an unbiased estimator of the population is impossible for this sampling scheme.

Given the challenges introduced when samples are not drawn independently, we are interested in the following questions. If the thresholds depend on the data, when can an unbiased or consistent estimator still be derived from the sample? Furthermore, when can the adaptive thresholds be treated as if they are fixed thresholds? An unbiased estimator for the fixed threshold would then automatically give an unbiased estimator for the adaptive threshold.

2.4 Function classes

Although the dependence between items makes computing arbitrary expectations difficult, we will show that for restricted classes of functions, the expectations can be easy to compute. We consider two function classes.

Our first goal is to examine when one can simply apply an unbiased estimator for a fixed threshold sampler to get an unbiased estimate for an adaptive threshold sampler. Given a adaptive thresholding scheme that allows for such a substitution, we seek to find the most general form of an estimator that allows the substitution. We first note that when sampling, any estimator is naturally restricted to be a function on the sample. These take the form of a polynomial in the Z. For the first class of functions, consider those of the form

$$\hat{\theta}(\mathbf{Z}, T) = \sum_{\lambda \in \Lambda_0} \beta_\lambda(\mathbf{x}_\lambda, T_\lambda) \prod_{i \in \lambda} Z_i. \tag{2}$$

Here, $\Lambda_0 \subset \mathcal{P}([n])$ where $\mathcal{P}([n])$ is the powerset of the indices $[n] := \{1, \ldots, n\}$. A subscript of λ selects the indices in λ. Since the priorites R determine Z and T we also write the estimator as $\hat{\theta}(R, T)$ or $\hat{\theta}(R)$. This form as a polynomial is particularly useful as the linearity of expectations allows each monomial term's expectation to be computed separately.

By further restricting the class of functions, we can capture a wider range of adaptive thresholding schemes where one has existing unbiased estimators. The second function class, the important set of pseudo-HT estimators, are of the form

$$\hat{\theta}(\mathbf{R}) = \sum_{\lambda \in \Lambda_0} h_\lambda(\mathbf{x}_\lambda) \prod_{i \in \lambda} \frac{\tilde{Z}_i^\lambda}{F_i(T_i^\lambda)}. \tag{3}$$

Rather than each item having a single threshold, an item's threshold can change depending on the term λ in the sum. Since the thresholds can be different, the inclusion indications $\tilde{Z}^\lambda$ can be as well.

We note that the restriction to pseudo-HT estimators is mild. For any i.i.d. sample from a distribution G, any estimable parameter of the distribution G is equal to $\mathbb{E}h(X_1, \ldots, X_m)$ for some finite m and symmetric function h [16]. This form, which is that of a U-statistic, is an even more restricted form that that of an HT-estimator.

2.5 Threshold recalibration

When using an adaptive threshold sampling rule, the main difficulty in computing expectations is due to the interdependence of the threshold and priorities. Our main idea is that, if we consider a monomial term $\Pi_{i \in \lambda} Z_i$, the true adaptive threshold T can be replaced with an alternative threshold T^λ that is independent of the priorities R_λ.

This can also be seen as creating an alternative threshold sampling procedure for every monomial term. Since each monomial corresponds to a subset λ of all the items, the alternative threshold can adapt to the data using only items not indexed by λ. As long as the alternative thresholds are not larger than the original ones, $\tilde{T}_\lambda^\lambda \le T_\lambda$, the sample using the alternative can be computed from the sample using the original threshold. We are particularly interested in the case where the alternative and original thresholds are equal, $\tilde{T}_\lambda^\lambda = T_\lambda$.

These alternative thresholding rules are created in the following manner. Let $\mathcal{R}^\lambda(R) = \{r : r_i = R_i \, \forall r \notin \lambda\}$ be the set of priorities where items *not* in λ have priorities equal to those in R. Define the *recalibrated thresholding rule* and threshold with respect to λ by

$$\tilde{\tau}_i^\lambda(R_{-\lambda}) = \inf_r \{\tau_i(r) : r_j = R_j \, \forall j \notin \lambda\}$$
$$\tilde{T}^\lambda = \tilde{\tau}^\lambda(R_{-\lambda}) \le T_\lambda$$

In other words, given a set of priorities, we ignore the values of priorities R_λ and find the smallest threshold with the given values of $R_{-\lambda}$. When the thresholding rule τ is non-decreasing, the alternative threshold is obtained by setting the priority R_i of every item $i \in \lambda$ to the smallest possible value. This yields inclusion indicators $\tilde{Z}_i^\lambda = 1(R_i < T_i^\lambda)$.

Although the true inclusion probability typically remains intractable to compute, computing a conditional inclusion probability is easy.

Lemma 1. *The conditional inclusion probability given the recalibrated threshold is*

$$p(\prod_{i \in \lambda} \tilde{Z}_i = 1 \mid \tilde{T}^\lambda) = \prod_{i \in \lambda} 1(R_i < \tilde{T}_i^\lambda) = \prod_{i \in \lambda} F_i(\tilde{T}_i^\lambda),$$

This provides the ability to estimate any statistic as long as the sample size is large enough.

Theorem 2. *Given a statistic $\theta = \sum_{\lambda \in \Lambda_0} h_\lambda(x_\lambda)$, the pseudo-HT estimator using recalibrated inclusion indicators $\tilde{Z}$ and thresholds $\tilde{T}$*

$$\hat{\theta}(\tilde{Z}, \tilde{T}) = \sum_{\lambda \in \Lambda_0} h_\lambda(x_\lambda) \prod \frac{\tilde{Z}_i^\lambda}{F_i(\tilde{T}_i^\lambda)} \tag{4}$$

is an unbiased estimator for θ.

Proof. Conditioning on $\tilde{T}^\lambda$ and applying the tower rule gives $\mathbb{E}\left(\prod \frac{\tilde{Z}_i^\lambda}{F_i(\tilde{T}_i^\lambda)}\right) = 1$. □

The natural consequence of the lemma given an unbiased estimator for sums.

Corollary 3. *The following conditional HT-estimator based on the alternative thresholds is unbiased:*

$$\hat{\theta}_{HT}(R, \tilde{T}) := \sum_{i=C} x_i \frac{\tilde{Z}_i^i}{F_i(\tilde{T}_i^i)}.$$

2.5.1 Example: Bottom-k sketches and priority sampling. Priority sampling and other bottom-k sampling procedures choose the threshold to be the $(k+1)^{th}$ smallest priority $R_{(k+1)}$. Since the thresholding rule is non-decreasing, we can recalibrate the threshold for any item x_i in the sample by changing the priority R_i to $-\infty$. These priorities were already smaller than the $(k+1)^{th}$ smallest priority. Changing them to $-\infty$ does not affect the threshold, so $\tilde{T}_i^i = T_i$. Thus, the estimator remains unchanged after substituting with the recalibrated thresholds, and

$$\hat{\theta}_{HT}(R, T) = \hat{\theta}_{HT}(R, \tilde{T}) \tag{5}$$

is an unbiased estimator of the sum as long as $F_i(T_i) > 0$ almost surely for all $i \in C$.

The main differences when deriving the unbiased estimator using threshold recalibration compared to existing methods are that (1) threshold recalibration provides a constructive procedure for updating the thresholds while existing derivations must first propose a new thresholding rule that is then verified to match the existing one, and (2) we allow the alternative thresholds for differ for every monomial term. However, we are most interested in the case where the original thresholds and the recalibrated ones are the same. In this case, the original adaptive thresholds can be treated as fixed thresholds for the relevant class of functions.

2.6 Threshold substitutability

Suppose the recalibrated thresholds $\tilde{T}^\lambda$ are equal to the original ones whenever the subset λ is in the sample. Formally, $\tilde{Z}_i^\lambda = 1 \, \forall i \in \lambda \implies \tilde{T}_\lambda^\lambda = T_\lambda$. We call thresholds that satisfy this *substitutable* thresholds. If a threshold only satisfies this when $|\lambda| \leq d$ then we call it a d-substitutable threshold. Substitutable thresholds have several attractive properties. Most importantly, unbiased estimators derived under a fixed thresholding scheme are also unbiased under the true adaptive sampling scheme under modest regularity conditions. Thus, substitutable thresholds can be treated almost like fixed thresholds.

Theorem 4 (Threshold substitution). *Let T be a substitutable threshold. Suppose the estimator $\hat{\theta}$ in the form given in 2. Then*

$$\hat{\theta}(R, T) = \hat{\theta}(R, \tilde{T}). \tag{6}$$

This also holds if τ is d-substitutable and $\hat{\theta}$ is additionally at most a d degree polynomial in Z.

Proof. Consider a term $\beta_\lambda(x_\lambda, T_\lambda) \prod_{i\in\lambda} Z_i$. If $\prod_{i\in\lambda} Z_i = 1$ then the threshohold $\tilde{T}_\lambda^\lambda = T_\lambda$ does not change. Thus, $\tilde{Z}_\lambda^\lambda = Z_\lambda$ as well, and $\beta_\lambda(x_\lambda, T_\lambda) \prod_{i\in\lambda} Z_i = \beta_\lambda(x_\lambda, \tilde{T}_\lambda^\lambda) \prod_{i\in\lambda} \tilde{Z}_i^\lambda$. Otherwise, if $\prod_{i\in\lambda} Z_i = 0$ then $\prod_{i\in\lambda} \tilde{Z}_i^\lambda = 0$ since recalibrated thresholds are always less than or equal to he original one. □

Corollary 5. *Let T be a substitutable threshold. Suppose $\hat{\theta}(R, t)$ is in the form given in 2 and is an unbiased estimator of θ for any value of $t \in Range(T)$. Then $\hat{\theta}(R, T)$ is an unbiased estimator of θ.*

Proof. $\mathbb{E}\hat{\theta}(R, T) = \mathbb{E}\hat{\theta}(R, \tilde{T}) = \mathbb{E}(\hat{\theta}(R, \tilde{T})|\tilde{T}) = \mathbb{E}(\theta|\tilde{T}) = \theta$ □

The definition of a substitutable thresholding rule requires verifying that the original thresholds equal recalibrated thresholds with respect to every possible subset of indices. We now provide a simpler condition to verify substitutability that recalibrates with respect to singletons.

Theorem 6 (Substitutability from singletons). *Let τ be a non-decreasing adaptive thresholding rule generating the threshold T. If for any $i \in \{1, \ldots, n\}$, $\tilde{T}_\lambda^i = T_\lambda$ whenever $\prod_{j\in\lambda} Z_j = 1$, then T is a substitutable threshold.*

Proof. Let $\tau(R) = T$ be the thresholding function for T. We must verify $\tilde{T}_\lambda^\lambda = T$ for all subsets λ with non-zero probability of being selected. Without loss of generality assume the subset to be verified is $\lambda = \{1, \ldots, k\}$. If $\prod_{j\in\lambda} Z_j = 1$, then $R_j < T_j$ for all $j \in \lambda$. Using induction, we have $T = \tau(r_1, r_2, \ldots, r_k, R_{k+1}, \ldots, R_n)$ whenever $r_j < T_j$ for all $j \in \lambda$. Since the recalibrated threshold $\tilde{T}^\lambda$ is simply the infimum over the coordinates indexed by λ, it follows that $\tilde{T}^\lambda = T$. □

2.6.1 Variance of the HT estimator. The value of threshold substitution is that it makes it trivial to obtain unbiased estimators under adaptive threshold sampling. One can simply use an existing estimator for simple, Poisson sampling designs. We illustrate the ease in estimating the variance of an HT estimator using our framework. In comparison, the priority sampling paper [12] required a one and a half page derivation.

Section 2.5.1 showed that the bottom-k and priority sampling threshold satisfies the conditions of Corollary 6. Thus, it is substitutable. The variance of the HT estimator $\hat{\theta}_t$ under fixed threshold sampling with threshold t and an unbiased estimator of this variance are given by

$$Var(\hat{\theta}_t) = \sum_{i\in C} \left(\frac{1 - F_i(t)}{F_i(t)}\right) x_i^2 \tag{7}$$

$$\widehat{Var}(\hat{\theta}_t) := \sum_{i\in C} \left(\frac{1 - F_i(t)}{F_i(t)^2}\right) Z_i x_i^2 \tag{8}$$

Since the squared error $(\hat{\theta}(Z, T) - \theta)^2$ is in the function class given by equation 2, the variance estimator for the HT estimator on fixed thresholds is also unbiased for the adaptive bottom-k threshold. That is, $Var(\hat{\theta}_T) = Var(\hat{\theta}_{\tilde{T}}) = \mathbb{E}Var(\hat{\theta}_{\tilde{T}}|\tilde{T}) = \mathbb{E}\widehat{Var}(\hat{\theta}_{\tilde{T}}|\tilde{T})$ due to the tower rule and unbiasedness of $\hat{\theta}$.

2.7 Sequential thresholding rules

The last class of thresholds and functions we consider consists of thresholds whose values can be determined in a sequential manner. Although these thresholds may not be substitutable, we show they can yield unbiased HT-estimators. We motivate these with the following example.

Example (1-substitutable threshold): Suppose a data stream is processed using a bottom-k sketch. Instead of storing only items in the final state of the bottom-k sketch, suppose an item is stored as long as it was in the bottom-k sketch at some point in the stream. This allows aggregates to be computed over time windows $[0, t]$ for any time t. In this case, the threshold rule τ_i is a function of the priorities $R_1, \ldots, R_i$. τ is trivially 1-substitutable. However, the threshold is not 2-substitutable. To see this, consider the state of the sketch after processing the entire stream, and let x_j be the last item included in the sample. If the data stream is sufficiently large, then the threshold T_j must be equal to a priority R_i of an item that appeared earlier in the stream and was included sample but was kicked out of the sketch before the end of the stream. If that earlier priority R_i changes, then the later threshold T_j also changes. Thus, while 1-substitutability allows us to use the HT estimator for sums, it does not allow us to compute variances.

However, despite being non-substitutable, we show this threshold can still be treated like a fixed threshold for pseudo-HT estimators. We show that if there is an ordering of the data such that the thresholding choices are made sequentially, then a pseudo-HT estimator is unbiased. That is, we must show $\mathbb{E}\prod_{i\in\lambda} Z_i/F_i(T_i) = 1$ for any $\lambda \subset \Lambda_0$ where $p(\prod_{i\in\lambda} Z_i = 1) > 0$.

Theorem 7. *Given a permutation $\rho_1, \ldots, \rho_n$ of $[n]$, define the sequential sample $S^j = \{\rho_k : Z_{\rho_k} = 1, k \le j\}$. If the recalibrated thresholds $\tilde{T}_k^{S^j} = T_k$ for all $k \in \{\rho_{n-j+1}, \ldots, \rho_n\}$ with $Z_{\rho_k} = 1$, then*

$$\mathbb{E}\prod_{i\in\lambda}\frac{Z_i}{F_i(T_i)} = 1 \tag{9}$$

for any λ such that $p(\prod_{i\in\lambda} Z_i = 1) > 0$.

Proof.

$$\begin{aligned}
&\int \prod_{i\in\lambda}\frac{Z_i}{F_i(T_i)} dF_{s_k}(R_{s_k})\cdots dF_{s_1}(R_{s_1}) \\
&= \int \prod_{i\in\lambda\setminus\{s_k\}}\frac{Z_i}{F_i(T_i)}\frac{F_k(T_k)}{F_k(T_k)} dF_{s_{k-1}}(R_{s_{k-1}})\cdots dF_{s_1}(R_{s_1}) \\
&\cdots = 1
\end{aligned}$$

□

In the case where a sequence is ordered by priority, sequential rules can also be used to create substitutable thresholds.

Lemma 8. *Consider the sequence $R_{i_1} > R_{i_2} > \cdots > R_{i_n}$. If M is a stopping time with respect to the filtration defined by this sequence, then the rule $\tau(\mathbf{R}) = R_{i_M}$ is a substitutable threshold.*

2.8 Building and composing thresholds

Thus far, we have described methods for taking an existing thresholding rule and modifying it to allow for unbiased estimation.

We now examine how thresholds can be composed and samples can be merged. A simplified interpretation of our result states that taking the maximum of a 1-substitutable thresholding rules yields another 1-substitutable rule. Taking the minimum of fully or d-substitutable rules yields another fully or d-substitutable rule respectively. However, for a specialized case of sequential thresholding rules where every threshold is chosen with respect to the same underlying sequence of priorities, taking the maximum preserves full or d-substitutability. This specialized case includes bottom-k sketches.

Theorem 9. *Let T^1, T^2 be 1-substitutable thresholds on a data set and priorities $\mathcal{D}$. Denote the indices of the corresponding sampled items by $\mathcal{I}^1, \mathcal{I}^2$. Consider the dataset $\mathcal{D}'$ formed by the sampled items and their priorities. Consider a thresholding rule τ' where τ'_i is a function of $\mathcal{D}'$ and T_i^1, T_i^2. Denote the corresponding threshold by T'_i. If τ' is a 1-substitutable thresholing rule for the data set $\mathcal{D}'$ and $T' \le \max\{T^1, T^2\}$ then it is 1-substitutable threshold on the original dataset. Likewise, if T^1, T^2 are substitutable, τ' is a function of T^1, T^2 and $\mathcal{D}'$ and is substitutable, and $T' \le \min\{T^1, T^2\}$, then T' is substitutable.*

Proof. For 1-substitutability, note that changing the priority R_i for any $i \in \mathcal{I}^1 \cup \mathcal{I}^2$ does not change T_i^1, T_i^2. Hence, it also cannot change T'_i. Similarly, substitutability holds when T^1, T^2 are substitutable. □

2.9 Priority-threshold duality

Another useful property of adaptive threshold sampling is that adjusting priorities is equivalent to adjusting thresholds. An item with a priority distribution F_i and per item threshold T_i is included if $R_i = F_i^{-1}(U_i) < T_i$. Equivalently, it is included if a random uniform random variable $U_i < F_i(T_i)$ is less than the pseudo-inclusion probability. Thus, the adaptive threshold sampling framework can be used in cases where the importance of an item and its priority distribution can change over time.

For instance, in time-decayed sampling [10] with exponential decay, the weight of an item $w_i(t) = \overline{w}_i \exp(-t)$ decreases exponentially with time t after the item appear at time t_i^0. Thus, one can build a sampling method which uses adaptively chooses a threshold $T(t)$ given time varying weights $w_i(t)$. An item is included in the sample at time t if its priority $R_i(t) = U_i/w_i(t) < T(t)$ However, it is inconvenient in practice to change the weight of existing points. Changing the threshold to increase exponentially instead, allows the priorities to remain fixed. An item is included if $\overline{R}_i = U_i/\overline{w}_i < exp(t)T(t)$.

3 APPLICATIONS AND SAMPLING DESIGNS

Although the definitions and theorems in section 2.3 are subtle and abstract, we now show that they are powerful, allowing us to easily generate sampling schemes that solve novel problems†, improve existing sketches*, and unify the theory for multiple sampling methods‡. We label the sections with the respective symbol †, *, or ‡ to note the contribution our paper makes.

3.1 Variable item sizes†

Bottom-k sketches ensure the sample size is always k. However, the memory usage of the sample can vary if items are of different sizes. If there is a memory budget B, to guarantee that the sample fits within the budget, the parameter k must set conservatively to B/L_{max} where L_{max} is the size of the largest item. This is highly inefficient if the size of the largest item is much bigger than that of the average item. Another thresholding rule τ simply takes as many items as possible that fit within the memory budget. When

processing items from lowest to highest priority, the threshold is the priority of the first item which causes the budget to be exceeded. Like a bottom-k sketch, the actual values of the smaller priorities are irrelevant and can be set to 0, so the threshold is substitutable. Thus, as long as the budget $B \geq L_{max}$ so that every item has a non-zero change of being selected, the usual HT estimator provides estimates of subset sums, and if $B \geq 2\,L_{max}$ the usual HT variance estimator provides unbiased estimates of its variance.

For example, the items in the 2020 Kaggle data science survey can vary in size since the survey contains text free responses and cases where the respondent does not finish the survey. As a string, the maximum length of an item is 5113 characters while the average length is 1265. A bottom-k sample that is guaranteed to fit within a budget constraint is expected to be $1/4th$ the size of an adaptive threshold sample that utilizes the entire budget.

3.2 Sliding windows*

Oftentimes, only recent items in a data stream are of interest. A sliding window sampler draws a uniform sample from points that arrive in the time interval $(t - \Delta, t]$ where t is the current time and Δ is the length of the time window. When the arrival rate of items changes over time, it can be impossible to draw a fixed size sample in bounded space [4, 14]. The state-of-the-art for drawing uniform samples from sliding windows in bounded space is given by Gemulla and Lehner (G&L) [14]. We show this is an instance of adaptive threshold sampling, but a highly inefficient one. Our framework immediately yields improved thresholds that double the number of usable points with zero modifications to the sketch.

At time t, the G&L scheme consists of one set of expired samples $X(t)$ that occur in the time window $(t - 2\Delta, t - \Delta]$ and another set of current examples $C(t)$ in the window $(t - \Delta, t]$. Although G&L do not describe their procedure as a thresholding scheme, we can describe it as one consisting of two stages, one which samples non-uniformly to build candidate points and a one which provides the final uniform sample.

The initial threshold $T_n(t_n)$ for an item x_n at time t_n is 1 if there are fewer than k current examples. Otherwise, $T_n(t_n)$ is the k^{th} smallest priority of the current sample $C_-(t_n)$ just before time t_n and the new priority R_n. $T'_n(t_n) = \max_{i \in C^-(t_n) \cup \{R_n\}} R_i$. If there are ever more than k current examples, the largest priority item is discarded by adjusting the threshold of all the current examples, $T_i(t_n) = \min\{T_i^-(t_n), T_n(t_n)\}$. An example that falls out of the current window is moved from the current to expired examples, and any expired item that is two window lengths or more from the current time is discarded. This threshold determines whether or not an item is stored but does not ensure that it is a uniform draw from a sliding window.

The G&L scheme assigns a final threshold T_{GL} equal to the k^{th} smallest priority of the combined current and expired examples. This is guaranteed to return a uniform sample from the current time window, although the size of the sample is not fixed. In this case, the item corresponding to the threshold can be included in the sample due to symmetry. The priority need not be strictly less than the threshold.

This bottom-k threshold, however, results in an inefficient threshold that discards half of the useful points. We note that the initial thresholding rule consists of two parts, a sequential sampling rule, and a sequence of minimum operations on the thresholds. Our adaptive thresholding framework immediately provides a much improved threshhold. The sequential rule generates substitutable thresholds and taking the min of thresholds preserves substitutability. Thus, taking another min of all the thresholds in the current examples yields another substitutable threshold and a uniform sample. That is $T_{improved}(t) = \min_{i \in C(t)} T_i(t_n)$. Furthermore, the per-item threshold of the current examples can be calculated from the expired examples and earlier current examples. Computing the improved threshold require no additional storage, and the adaptive sampling framework provides the improvement for free.

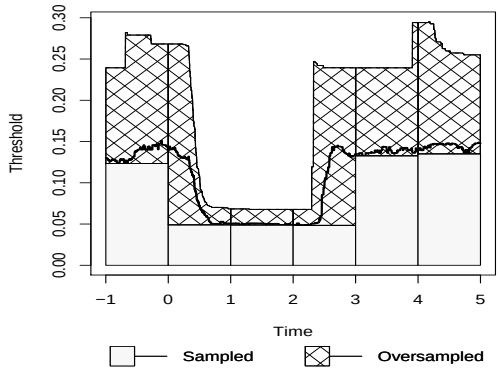

Figure 1: The top line shows the true marginal sampling probabilities which our method recovers. The middle line shows the conservative estimate used by the G&L scheme, and the boxes at the bottom show the thresholds used by a set of sliding windows. The hatched area above each box shows the amount of oversampling used to draw that sliding window sample. Not that even if the true thresholds were used, the scheme must still significantly oversample when the item arrival rate changes.

Figure 1 shows the evolution of the per item threshholds $T_i(t_i)$ over time and the much smaller threshold used by G&L to construct a sliding window sample. Figure 2 shows the behavior when there is a spike in the item arrival rate. Not only does our adaptive threshold sampling framework yield nearly twice as many samples when the item arrival rate is steady, it recovers from the spike faster.

3.3 Adaptive sampling for top-k and disaggregated subset sums†

Adaptive thresholding can also be used to modify a frequent item sketch into a top-k sketch that also supports further aggregations. Given a parameter m, a frequent item sketch returns all items where the proportion of times each appears is $> 1/m$. This can be done with $O(m)$ space using the Misra-Gries [22] or equivalent Space-saving sketch [21]. The top-k problem requires returning the top k items by frequency. There is no guarantee on the minimum proportion of times these top-k items appear. This makes the top-k problem a more challenging variation of the frequent item problem. While frequent item sketches such as can be used for the top-k problem, they require knowing the appropriate size parameter m

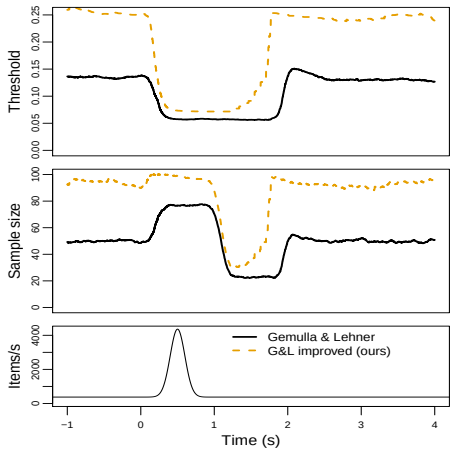

Figure 2: Our improvement on G&L draws more samples (middle) while recovering more quickly from a spike in the item arrival rate (bottom). This is because G&L gives an underestimate of the threshold (top).

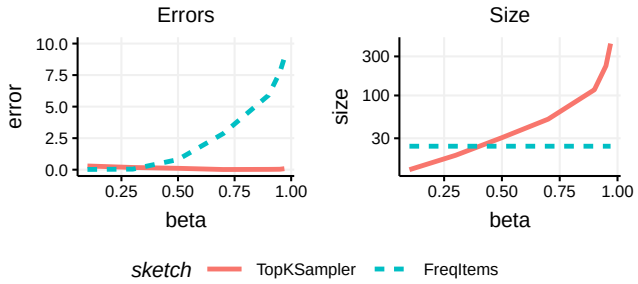

Figure 3: Comparison of the fixed size FrequentItems sketch and our adaptive Top-K sampler as the distribution of frequent items changes. Smaller β values have a few dominant heavy hitters while larges values yield more evenly distributed frequencies. Left: The average number of incorrect items in the top-k items returned by each sketch. Right: The size of the sketch in number of items.

in advance. It can also be useful for the sketch to support further aggregations. For example, one may wish to analyze web page impressions and find the frequently viewed pages. One may then wish to further aggregrate pages by topic. This disaggregated subset sum problem is addressed by [20, 29].

This problem can be seen as an adaptive sampling procedure which learns to downsample infrequent items, leaving the frequent ones. By virtue of being an adaptive threshold sampling procedure, it automatically supports unbiased estimation of counts and further aggregations using the HT-estimator. This leads to the following adaptive threshold sampling procedure. For each point x_t in a data stream, assign a $Uniform(0,1)$ priority R_t. Maintain a variable length list where each entry consists of an item x_i, its priority R_i, a threshold T_i, and a count v_i of the times it appears after entering into the sample. An unbiased estimate of the count of an item is the Horvitz-Thompson estimate $\hat{c}_i = 1/T_i + v_i$. We define the adaptive threshold $T(t)$ at time t to be the smallest priority such that at least k items in the sample have estimated count $\hat{c}_i > 1/T(t)$. This splits the items in the sample into infrequent items with $\hat{c}_i \le 1/T(t)$ and k frequent items with $\hat{c}_i > 1/T(t)$. Whenever the adaptive threshold $T(t)$ is updated, only infrequent items are updated. Those with priority $R_i \ge T(t)$ are discarded. All others update their threshold to $T(t)$ and their counter to $v_i = 0$.

This procedure can be seen as a thresholding based variation of Unbiased Space-Saving [29]. In both, infrequent items form a random sample of the items not assigned to a frequent item counter. Frequent items start as infrequent items and simply maintain a count of the number of times each occurred after entering the sample. It is easy to see that this thresholding rule is substitutable. For any subset of items in the sample, changing their priorities to 0 has no effect on the sample or thresholds. Thus, like Unbiased Space-Saving, it can be used for unbiased estimation for the disaggregated subset sum problem.

We compare our adaptive procedure with the FrequentItems sketch in Apache Datasketches [1, 2]. This FrequentItems sketch is a variation of the Misra-Gries sketch [21, 22] that allows for faster updates. For each sketch, we query for the top-k items in a stream with $k = 10$ and record the number of errors in the result. Since we wish to compare performance as the distribution of the heavy hitters changes, we use a synthetic Pitman-Yor$(1, \beta)$ preferential attachment process that is able to generate both light tailed and heavy tailed behavior. It is commonly used in Bayesian cluster models. Larger values of $\beta \in [0, 1)$ result in heavier tails. More precisely, in the Pitman-Yor process, the t^{th} item in the stream is a new item with probability $(1 + \beta C_t)/t$ where C is the number of unique items seen already. Otherwise, it is equal to the j^{th} unique item with probability $(n_{tj} - \beta)/t$ where n_{tj} is the number of times unique item j has been seen amongst the first $t - 1$ items.

Figure 3 shows that our procedure can accurately capture the top-k items without prior knowledge of the distribution. It does so by appropriately adjusting the size of the sample so that it captures the top-k items with high probability. For distributions where the the frequent items are well-separated from the remaining items, our adaptive sampler requires even less space than the FrequentItems sketch. For distributions where not all frequent items are well separated from the infrequent items, the FrequentItems sketch performs poorly. Our adaptive procedure, on the other hand, is able to adjust its size to the data and capture the frequent items. For FrequentItems we take the size to be 0.75 times the size of the allocated hash table.

3.4 Distinct counting for weighted samples‡

The subset sum and distinct count problems are often treated as separate problems. However, they can be addressed using a single weighted, coordinated priority sample. For any substitutable threshold $\mathbf{T}$, the distinct count estimator is simply $\hat{N} = \sum_i Z_i / F_i(w_i T_i)$ while the subset sum estimator for a subset of indices $\mathcal{A}$ is $\hat{S}(\mathcal{A}) = \sum_{i \in \mathcal{A}} w_i Z_i / F_i(w_i T_i)$. This extends the Theta sketch framework

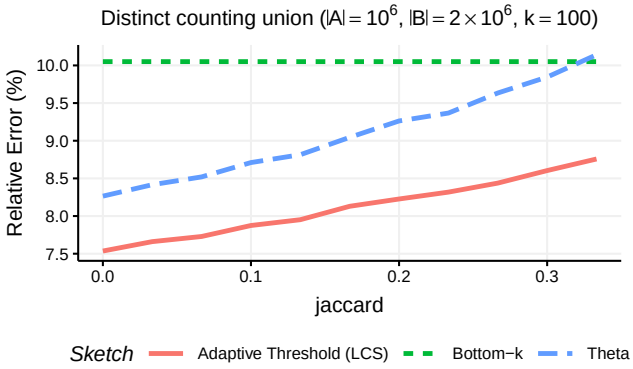

Figure 4: Relative error $SD(\hat{N} - N)/N$ as a function of the Jaccard similarity. Our adaptive thresholding procedure is the same as LCS in this case and improves upon the basic bottom-k sketch and the Theta sketch except when $A \subset B$.

[11] to non-uniform priorities and weighted samples and allows for per item thresholds.

3.5 Improved merges for distinct counting[‡]

The framework also can be used to improve merge procedures for distinct counting sketches based on coordinated samples, such as the MinCount/bottom-k sketch [3, 11, 15]. Given two coordinated samples with 1-substitutable thresholds T^A and T^B for sets A and B, one simply needs to produce a new 1-substitutable threshold with $T'_i \leq \max\{T^A_i, T^B_i\}$ to produce another distinct counting sketch. This generalizes the LCS sketch of [9] which specifically takes the max of bottom-k thresholds. Our use of arbitrary 1-substitutable thresholds also allows merges to be chained together. Figure 4 illustrates the improvement when taking the union of sets A and B of size $|A| = 10^6$, $|B| = 2 \times 10^6$, and varying Jaccard similarity.

3.6 Frequent items for distinct counting[†]

A common database aggregation computes a distinct count of an item grouped by some attributes, for example, the number of distinct users that saw an ad grouped by time and demographic variables. Although each distinct counting sketch is small, the group by operation can create tens of millions of sketches, resulting in a large memory footprint. Oftentimes, many of these groups contain only a few items. Previous approaches have used counter sharing [31**?**] and sparse representations for small counters [17]. We propose a novel approach use subsampling.

At any point in time, rather than maintaining a bottom-k sketch for every group, maintain a bottom-k sketch of item hashes for only the current top m groups where m is a user chosen parameter. For these m groups, denote the threshold for group g by T_g. For any item in the dataset. If the item belongs to group g, it uses the threshold T_g. Otherwise, it uses $\max_g T_g$. In effect, we adjust the sampling rate to be the appropriate sampling rate for the top m groups. Equivalently, the tolerated error for a small group is raised from being a percentage of the small group's size to a percentage of the heavy hitters' group sizes.

While using a bottom-k sketch for every group forces the number of sketches to grow linearly with the number of groups, in this frequent item sketch for groups, many small groups will not have any sampled items. This alleviates the problem of having too many counters.

3.7 Multi-Stratified sampling[†]

Suppose a data set of users can be stratified in two ways, by country or by age. We wish to draw a single sample which is both a stratified sample by country and a stratified sample by age and fits within a budget of B items. We first show how to generate a stratified sample, and then how to control the budget.

First, maintain a bottom-k threshold $\tau_c^{(country)}$ for each country c and $\tau_a^{(age)}$ for each age a. A user x_i with country c_i and age a_i has a per item threshold of $\tau'(\mathbf{R}) = \max\{\tau_{c_i}^{(country)}(\mathbf{R}), \tau_{a_i}^{(age)}(\mathbf{R})\}$. Since the bottom-k thresholding rules are sequential rules on the same sorted priorities, the resulting threshold is substitutable.

This sample has variable size $m \in [k \max\{n_c, n_a\}, k(n_c + n_a)]$ where n_c and n_a are the number of distinct countries and ages. We wish to ensure the sample contains exactly B items. We modify the stopping rule for the thresholds so that k, the number of items per stratum, is dynamically selected. For a set of thresholds, choose a stratum with the most number of elements below its threshold. Decrement its threshold to the next smaller priority. Since an item belongs to more than one stratum, one for country and one for age, this may not decrease the total sample size. Continue decreasing the thresholds until the desired sample size is reached.

3.8 Multi-objective samples[*]

When using samples, the importance of an item may depend on the query. For example, an analyst may be interested in either profit or revenue. Queries on profit ideally utilize a weighted sample with weight proportional to each item's profit. Similarly, queries on revenue ideally weight by revenue. The existing approach by Cohen [6] combines two coordinated bottom-k sketches, one for profit and another for revenue, to obtain a sketch that has size $\leq 2k$. This ensures the combined sketch never does worse than an individual sketch. However, as more objectives are added, each objective's sketch must be made smaller. If there is a budget constraint B and c different objectives, each objective can only be allocated a sketch with space B/c. When sketches have high overlap, the size of the combined sketch can be much less than the budget. For example, if every objective assigns the same weight for each item, then the priorities and hences, sketches for every objective are the same and only B/c of the budget B is used after combining the sketches.

3.9 Variance sized samples[†]

Priority sampling provides a relative error guarantee on the sum provided that the weights are proportional to the values in the sum. It guarantees that the variance of the error ϵ is bounded by $V(\epsilon) \leq S^2/(k-1)$ where S is the true sum and k is the sample size [26]. However, one may wish to have a guarantee on the absolute

error $V(\epsilon) \leq \delta^2$ or the weights may not be proportional to the weights.

We instead set a stopping time which stops at the first threshold T where the estimated variance matches the desired error, $V(\hat{S}_T) \geq \delta^2$. Here, $\hat{S}_t$ is the HT estimator with a fixed threshold t. The unbiased estimate of the variance of the HT estimator $\hat{V}(\hat{S}_t) = \sum_i 1(R_i < t) x_i^2 \frac{1-wt}{wt} 1(wt < 1)$ is discontinuous only at jumps when the threshold is equal to some priority. At all other threshold values, it increases continuously as t decreases. Thus, $\mathbb{E}V(\hat{S}_T) = \delta^2$.

We note that while processing a stream or data file, it is typically impossible to verify that a threshold is a stopping time with just the information in the sample. The stopping time may be a larger threshold that includes additional points that are not in the sample. However, since the true variance of an estimator is strictly increasing as the threshold decreases, it is reasonable to expect that all thresholds where the estimated variance equals the desired variance are close together. By slightly oversampling, one is likely able to recover the true stopping time.

3.10 Early stopping in approximate query processing[†]

Rather than explicitly constructing samples, another way to use adaptive threshold sampling is to store all of the data but sort items by their priorities. Given a user specified standard error δ, an approximate query processing system can provide an answer by using the variance sized sampling scheme given above to determine how many items need to be read and stop processing once it has read enough.

This can also be combined with other methods to define a physical layout of the data that is appropriate for sampling. For example, one can combine it with multi-objective sampling to get a layout that is useful for multiple queries. Suppose both revenue and quantity are metrics of interest in a dataset. Denote their values for row i by s_i and q_i respectively and their priorities by $S_i = U_i/s_i$ and $Q_i = U_i/q_i$. One can generate a file block consisting of a bottom-k sample ordered by S_i and a bottom-k sample of the remaining items ordered by Q_i. Repeating this procedure on the remaining items generates a physical layout of the data which only needs to read m blocks to get a weighted sample of size at least mk.

4 ASYMPTOTIC BEHAVIOR

The framework we have described thus far covers unbiased estimation. However, this excludes a number of important estimators that are consistent but not unbiased, such as maximum likelihood estimators, quantile estimators, and more generally, any estimator which is the maximizer of an objective function. We now extend our framework to justify such estimators as well as some heuristic thresholding rules in asymptotic settings. We also show that in asymptotic settings where the inclusion probabilities go to 0, all priority distributions are asymptotically equivalent to the weighted uniform distribution used in priority sampling.

While the practical implication of this section are significant, it is largely theoretical and requires knowledge of empirical process theory. For this reason, we first provide a brief overview of the main ideas that requires minimal prerequisite knowledge and states our main results in a easily understandable way at the cost of rigor. We then provide an overview for the theory we develop and techniques we use before stating and proving our results.

4.1 Basic Overview

The main idea is to consider adaptive thresholds that converge to or are close to a fixed threshold with high probability and show it what ways they can be treated as if they were actually fixed thresholds. For example, if a thresholding rule generates a single threshold T that is used for all points, the threshold T approximates a fixed threshold t_0 if $T \in (t_0 - \epsilon, t_0 + \epsilon)$ with high probability (w.h.p.) for some small $\epsilon \geq 0$. If $\hat{\theta}_t$ is an estimator for θ for fixed thresholds t. When applied to an adaptive threshold, the error of $\hat{\theta}_T$ is bounded by the worst case error $|\hat{\theta}_T - \theta| \leq \sup_{|h|<\epsilon} |\hat{\theta}_{t+h} - \theta|$ for estimators on that range of thresholds.

The difficulty it that we must show that the estimate at all perturbed thresholds $\hat{\theta}_{t+h}$ are close to $\hat{\theta}_t$. This requires a notion of continuity which cannot be obtained from a pointwise variance calculation at a specific threshold. Empirical process theory allows us to do just that.

We apply it to analyze the class of M-estimators, that is estimators that maximize an objective function formed by the sum over independent random variables. They are of the form

$$J_n(\theta) = \mathbb{E}_n f_\theta(X) = \sum_{i=1}^{n} f_\theta(X_i) \tag{10}$$

for some function f and n i.i.d. random draws from some distribution $X_i \sim P$, and they include maximum likelihood estimators, quantile estimators, regression estimators under L_2 or some other loss, as well as many neural networks. We extend this to obtain an objective $J_n(\theta; t)$ that depends on both the parameter θ as well as the threshold t. Empirical process theory allows us to show, under the appropriate rescaling, this objective asymptotically converges to a Gaussian process as more data is encountered. Crucially, this Gaussian process has continuous path. In other words, when treated as a function of t, the Gaussian process is a random, continuous function. Thus, convergence of the objective to a continuous function can be used to obtain convergence of the estimators.

We are particularly interested in the case where the sample size grows sub-linearly with the data. Here we define an appropriate scaling of the thresholds and show that regardless of how many parameters are in a priority distribution, if all priorities are non-negative and the priority density is non-zero around 0, the resulting threshold sampler is asymptotically equivalent to a simple weighted sampling procedure where the priority distribution are just parameterized by a univariate weight. This means any priority distribution can replicated using $R_i \sim Uniform(0, 1/w_i)$ as the priority distribution and appropriate choices of weights w_i. This results in an asymptotically equivalent sampling distribution. This convergence can only be proved when the class of functions $\{f_\theta\}_\theta$ used in the objective and the range of possible thresholds is not too complex. Heuristically, this means they cannot be chosen to overfit the data too much.

Together these provides our main results. Loosely speaking, the first states that if (1) an estimator $\hat{\theta}_t$ on deterministic thresholds

is consistent and (2) an adaptive threshold T converges to a deterministic threshold in an appropriate asymptotic regime and is not overly complex, then the estimator $\hat{\theta}_T$ on the adaptive threshold is also consistent. The second implies that if an adaptive threshold sample grows sub-linearly with the data and priorities are non-negative, the precise choice of priority distribution does not matter. Any priority distribution is asymptotically equivalent to using $R_i \sim Uniform(0, 1/w_i)$ for an appropriate set of weights w_i.

4.2 Technical Overview

We now present an overview with the technical details. We consider the extension of an objective function to include a threshold as a parameter. For any priority distribution and fixed threshold t, it is easy to see that substituting the Horvitz-Thompson estimator of the objective for the original objective is will yield another M-estimator. The HT-estimator of the objective with threshold t can be expressed as an empirical expectation after reweighting by

$$\hat{J}_n(\theta; t) = \mathbb{E}_n f_\theta(X_i) w(R, t(X_i)) \quad \text{where} \tag{11}$$
$$w(R, t(X_i)) = \frac{1(R_i < t(X_i))}{F_i(t(X_i))}.$$

Since our asymptotic results requires an infinite sequence of points, we assume the points X_i are i.i.d. draws from some distribution P_x. The corresponding priority $R_i \sim F(\cdot|X_i)$ is drawn from some conditional distribution that depends only on X_i. The threshold $t(X_i)$ is also a function of a data point.

We show that the suitably rescaled objective converges to a Gaussian process, but rather than being indexed by just the parameter of interest θ, it is indexed by both θ and the threshold t. This convergence holds when both the function class $\{f_\theta\}_\theta$ and the class of thresholds $\mathcal{T}$ are not too complex.

This allows us to prove consistency of M-estimators under adaptive thresholding schemes. If a threshold T_n converges in probability to a deterministic threshold t, then the objective under the random threshold $\hat{J}(\theta; T)$ and fixed threshold $\hat{J}(\theta; t)$ converge to the same limit under the continuity of Gaussian processes. Hence, if an estimator is consistent under the deterministic threshold, it is also consistent under the random threshold.

Note that the threshold in this case need not be substitutable, nor does one need to be able to recalibrate it. Heuristic thresholding rules may be used as long as they converge to an appropriate limit.

We consider two asymptotic regimes, one where the adaptive thresholds converge to fixed thresholds and sample sizes grow linearly with the data, and one where the sample sizes grow sublinearly. The first case is straightforward and relies only on the closure properties of bounded uniform entropy integral function classes to prove a Donsker result that $n^{-1/2}(J_n(\theta, t) - J(\theta)) \xrightarrow{p} GP_{\theta,t}$. For the latter, we not only need to rescale the objective but also the thresholds to obtain convergence. In this regime, we show that the shape of the priority distribution is asymptotically irrelevant. The asymptotic distribution is only affected by the scaling of priorities.

Given convergence of the objective, we can prove our main result

Theorem 10. *Let $\hat{\theta}$ be an M-estimator for θ_0 under some distribution P_{θ_0}, and suppose its objective $J_n(\theta) = \mathbb{E}_n f_\theta(X_i)$ satisfies the conditions of the M-estimator consistency theorem 2.12 in [19]. Suppose there is a sequence of constants $c_n \to c_0 \geq 0$ such that $c_n n \to \infty$ and thresholds $T^{(n)} \in \mathcal{T}$ with $c_n T^{(n)} \xrightarrow{p} T$. If $\{f_\theta\}_\theta$ and $\mathcal{T}$ satisfy the conditions of theorem 12 then the HT-estimate of the objective $\hat{J}_n(\theta, T^{(n)})$ yields a consistent estimator $\hat{\theta}^{(n)}_{T^{(n)}}$ of θ_0.*

Proof. Under these assumptions, theorems 12 and 11 and the continuous mapping theorem show that the HT-estimate of the objective also satisfies the the conditions of the M-estimator consistency theorem. Hence, $\hat{\theta}^{(n)}_{T^{(n)}}$ is consistent. □

The basic setup for proving convergence to an empirical process starts with a class of functions $\mathcal{F}$ and an empirical measure $\mathbb{P}_n$ for n random draws from some measure P. This empirical measure takes a function $f \in \mathcal{F}$ and maps it to the empirical mean

$$\mathbb{P}_n f = \mathbb{E}_{\mathbb{P}_n} f(R_0) = n^{-1} = \sum_{i=1}^{n} f(R_i) \tag{12}$$

where $R_i \sim P$. Since this is a mean, under mild regularity conditions, for any single function f, the empirical mean $\mathbb{P}_n f$ converges to a Gaussian random variable by the central limit theorem. Empirical processes theory allows one to show that, if the class of functions $\mathcal{F}$ is not too complex, then

5 EMPIRICAL PROCESSES

We briefly review some theory for empirical processes to unfamiliar readers. We refer interested readers to [19] for more details. Our goal is to show for a class of functions $\Phi = \{f_\theta\}_\theta$ and threshold functions $\mathcal{T}$, $\sqrt{n}\left(\hat{J}_n(\theta, t) - J(\theta)\right)$ converges weakly to a Gaussian process $\Psi_{\theta,t}$ with $\mathrm{Cov}(\Psi_{\theta,t}, \Psi_{\theta',t'}) = \mathrm{Cov}(f_\theta(X) w_t(R, X), f_{\theta'}(X) w_{t'}(R, X))$ for all $f_\theta, f_{\theta'}$ and $t, t' \in \mathcal{T}$ as $n \to \infty$. We can write this more succinctly as $\sqrt{n}\left(\hat{J}_n(\theta, t) - J(\theta)\right) \rightsquigarrow \Psi_{\phi,t}$ in $\ell^\infty(\Phi \times \mathcal{T})$.

Proving this convergence can be done in two steps. First, the finite dimensional distributions and the covariance can be shown to be Gaussian using the usual central limit theorem. Second, the empirical process $\Psi^{(n)}_{\phi,t} := \sqrt{n}\left(\hat{J}_n(\theta, t) - J(\theta)\right)$ is shown to be asymptotically tight. Asymptotic tightness ensures that the sample paths of the process are appropriately smooth. Asymptotic tightness is the main challenge for establishing Donsker results.

In empirical process theory, a sufficient condition for proving asymptotic tightness is that the function class of interest is not too complex. There are multiple measures of complexity such as the VC-dimension, bracketing entropy , and uniform entropy. The corresponding conditions that ensure asymptotic tightness are finite VC-dimension, finite bracketing entropy integral, and bounded uniform entropy integral with integrable envelope. For each of these measures of complexity, there are known, broad classes of functions that satisfy these conditions. Donsker preservation theorems show that certain transformations of these classes preserve the conditions on the complexity. New classes of interest can often be verified to be Donsker by showing they are contained in a class that is appropriately transformed from these broad, base classes.

Of these conditions, we are most interested in function classes with finite VC-dimension, also known as a VC-class, and those with bounded uniform entropy integral with integrable envelope, known as a BUEI-class with integrable envelope. The envelope G for a class $\mathcal{G}$ is the function such that $G(x) := \sup_{g \in \mathcal{G}} |g(x)|$. VC-dimension

is the most restrictive of the notions of complexity. Any VC-class is also a BUEI-class. As the most restrictive measure of complexity, it allows for the greatest range of transformations. In particular, it allows for composition with monotone functions. BUEI-classes are of interest because a product of BUEI-classes remains a BUEI-class.

5.1 Donsker result for fixed thresholds

We first consider the case where an adaptive threshold converges to a fixed one. In this case, the sample size grows linearly with the data.

Theorem 11. *Let Φ be a BUEI-class of functions on a data point, and let $\mathcal{T}$ be a class of threshold functions that has finite VC-dimension. Further assume that $F(t(x)) > \epsilon$ for some $\epsilon > 0$ and all $x \in \mathcal{X}$. Then, $f_\theta(X)w(R, t(X))$ has bounded uniform entropy integral. Hence,*

$$\hat{J}_n(\theta, t) = \mathbb{E}_n f_\theta(X) w(R, t(X)) \tag{13}$$

$$\sqrt{n}(\hat{J}_n(\theta, t) - J(\theta)) \rightsquigarrow \Psi^0_{\theta,t} \tag{14}$$

where $\Psi^0_{\theta,t}$ is a Gaussian process indexed by $f_\theta \in \Phi$ and $t \in \mathcal{T}$.

Proof. Since $\mathcal{T}$ has finite VC-dimension, the composition rules for VC-classes (lemma 9.9 in [19]) give that both $1(r - t(x) < 0)$ and $F(t(x))$ generate VC-classes since the indicator and any CDF are monotone. Likewise, $1/F(t(x))$ is a VC-class. Furthermore, it has a measurable envelope $H(x) = 1/\epsilon$. Since VC-classes have BUEI if there exists a envelope and BUEI classes are closed under multiplication (theorem 9.15 in [19]), the class of functions $\{g(x)w_t(r, x) : g \in \mathcal{G}, t \in \mathcal{T}\}$ has BUEI with envelope $G \cdot H$. This establishes asymptotic tightness of the process $\{\hat{\theta}_{gt}\}$. Since Donsker classes are preserved under multiplication by a bounded, measurable function (corollary 9.31 in [19]), $\mathcal{G} \cdot w_t$ is P-Donsker for any fixed t. Since they are also closed under finite sums, the finite dimensional distributions of Θ converge to multivariate Gaussian distributions. Asymptotic tightness and the central limit theorem on finite dimensional distributions imply that Θ is P-Donsker. □

5.2 Donsker result on sub-linear samples

In the above result, the resulting sample sizes must grow linearly with the data. We are interested in the case where the sample sizes still grows to infinity, but the inclusion probabilities go to 0. We show that under appropriate conditions, this is sufficiently well-approximated by a two-step procedure which downsamples the uniformly to obtain a sublinear number of data points and then applies threshold sampling where the inclusion probabilities are bounded away from 0 as above.

Theorem 12. *Consider priorities R_i taking values in the non-negative reals. Further suppose their conditional CDFs $F(\cdot|x)$ have a linear expansion near 0*

$$\Delta(r) := \sup_{x \in Supp(P_x)} |F(r|x) - w_x r| = O(r^2) \quad \text{if } r \geq 0 \tag{15}$$

for some weights w_x for all $x \in Supp(P_x)$ such that $M := \sup_x w_x < \infty$ and $\inf_x w_x > 0$.

Let Φ be a class of functions $\Phi = \{f_\theta\}$ and $\mathcal{T}$ a class of thresholds $\mathcal{T}$ satisfying the conditions of theorem 11. Further assume that the classes are uniformly bounded with $\|t\|_\infty < T$ for all $t \in \mathcal{T}$. Furthermore, consider alternative priorities $\dot{R}_i|X_i = x \sim Uniform(0, 1/w_x)$ and let $\cdot J_n(\theta, t)$ be the estimated objective using these priorities.

Then, for a sequence $c_n \to 0$ such that $c_n n \to \infty$, the processes

$$G^{(n)}_{\theta,t} := (c_n MTn)^{1/2} \left(\hat{J}_n(\theta, c_n t) - J(\theta)\right) \rightsquigarrow \Psi_{\theta,t} \tag{16}$$

$$\dot{G}^{(n)}_{\theta,t} := n^{1/2} \left(\dot{J}_n(\theta, t/MT) - J(\theta)\right) \rightsquigarrow \dot{\Psi}_{\theta,t} \tag{17}$$

as $n \to \infty$ where $\Psi_{\theta,t}, \dot{\Psi}_{\theta,t}$ are Gaussian processes and $\Psi_{\theta,t} \overset{d}{=} \dot{\Psi}_{\theta,t}$.

We first outline the proof. Rather than dealing with a fixed function class as before, these conditions require on proving a Donsker result when the class of thresholds is changing with n. Bounded uniform entropy conditions also exist for this case which ensure convergence to a Gaussian process. However, when directly applied, the decreasing thresholds lead to envelopes whose integrals go to ∞. To handle that, we show that the random process can be generated in two stages: a uniform sampling stage that draws a sample with sublinear size and on the order of nc_n and a second stage where the thresholds do not go to 0. The boundedness conditions ensure that Donsker preservation results can be applied on this second stage and obtain convergence to a Gaussian process. We then compare the process using priorities R_i with the approximating process using $\dot{R}_i$ and show that their first two moments are the same, and hence if they converge to Gaussian processes, the must have the same distribution.

Proof. First, note that for any threshold t, inclusion of an item X_i is equivalent to $U_i < F(c_n t(X_i)|X_i)$ for some $U_i \sim Uniform(0, 1)$. Since $F(c_n t(x)|x) = w_x c_n t(x) + o(w_x c_n t(x)) < c_n MT + c_n \epsilon$ eventually for any $\epsilon > 0$, this inclusion event can be rewritten as $U_i < c_n \gamma_\epsilon := c_n MT + c_n \epsilon$ and, if it passes this first stage, $\frac{U_i}{c_n \gamma_\epsilon} < \frac{F(c_n t(x)|x)}{c_n \gamma_\epsilon}$. Note that this second stage uses the conditional distribution $\frac{U_i}{c_n \gamma_\epsilon} | U_i < c_n \gamma_\epsilon \sim Uniform(0, 1)$. The first stage of the inclusion event is a uniform Poisson sample. It thins the data by drawing i.i.d. $Bernoulli(c_n \gamma_\epsilon)$ draws. The resulting sample of size C_n is still from the same base distribution P as the original data. Thus, we have that

$$\begin{aligned}
\hat{J}_n(\theta, c_n t) &= \mathbb{E}_n f_\theta(X) \frac{1(U < F(c_n t(X)|X))}{F(c_n t(X)|X)} \\
&= \frac{C_n}{n} \frac{1}{c_n \gamma_\epsilon} \mathbb{E}_{C_n} f_\theta(X) \frac{1\left(U < \frac{F(c_n t(X)|X)}{c_n \gamma_\epsilon}\right)}{\frac{F(c_n t(X)|X))}{c_n \gamma_\epsilon}}.
\end{aligned}$$

Since $C_n \sim Binomial(n, c_n \gamma_\epsilon)$ and $nc_n \to \infty$, it follows that $\frac{C_n}{n} \frac{1}{c_n \gamma_\epsilon} \xrightarrow{p} 1$ and can be ignored by Slutzky's lemma.

Thus, we can instead examine the convergence of the process $\mathbb{E}_{C_n} \phi(X) \frac{1(U < F(c_n t(X)|X)/c_n \gamma_\epsilon)}{F(c_n t(X)|X))/c_n \gamma_\epsilon}$. Let $\mathcal{V}$ be the VC-index of $\mathcal{T}$. The threshold $\{F(c_n t(X)|X))/c_n \gamma_\epsilon : t \in \mathcal{T}\}$ is a VC-class with index $\leq \mathcal{V}$ since scalar transformations do not change the VC-index and monotone transformations composed with a VC-class do not increase the VC-index. Furthermore, since $F(c_n t(X)|X = x))/c_n \gamma_\epsilon \to w_x t(x)/MT \leq 1$ and Φ is a BUEI-class that is uniformly bounded by some constant H, the empirical expectations are taken over function classes that are uniformly bounded by H. Together these imply

that

$$\sqrt{C_n}\left(\mathbb{E}_{C_n} f_\theta(X) \frac{1\left(U < \frac{F(c_n t(X)|X)}{c_n\gamma_\epsilon}\right)}{\frac{F(c_n t(X)|X))}{c_n\gamma_\epsilon}} - J(\theta)\right) \tag{18}$$

converges to a Gaussian process limit with mean 0 as long as its finite dimensional marginals converge to multivariate Gaussians. The law of large numbers gives that $nc_n\gamma_\epsilon/C_n \to 1$. Since ϵ can be arbitrarily small, equation 16 is proved.

Now consider the process which replaces $F(r|x)$ with the approximation $\dot{F}(r|x) = w_x r$. We wish to show that the mean and covariances of the process in equation 18 remain the same after the substitution. Since the HT-estimator is unbiased regardless of the choice of F, the mean is 0. Let $Z_n = 1\left(U < \frac{F(c_n t(X)|X)}{c_n\gamma_\epsilon}\right)$ and $\dot{Z}_n$ be the same with F replaced by $\dot{F}$.

Note that $F(c_n r|x) = c_n w_x r + c_n^2 O(r^2)$. The difference of the inclusion variables inversely weighted by its pseudo-inclusion probability is

$$\begin{aligned}\Delta_n &= \frac{Z_n}{F(c_n t(X)|X))/c_n\gamma_\epsilon} - \frac{\dot{Z}_n}{\dot{F}(c_n t(X)|X))/c_n\gamma_\epsilon} \\ &= \frac{\gamma_\epsilon}{w_X t(X)}\left(Z_n(1 + O(c_n t(X))) - \dot{Z}_n\right)\end{aligned}$$

This gives $|\Delta_n| \le \gamma_\epsilon(w_X t(X)(w_X t(X) + O(c_n t(X))^{-1}$ if $Z_n = \dot{Z}_n$ and $|\Delta_n| \le \gamma_\epsilon w_X t(X) \min\{1, 1 + O(c_n t(X))/(w_X t(X))\}$ otherwise. Since $P(Z_n \ne \dot{Z}_n) = O(c_n t(X))$, the variance $Var(\Delta_n) = O(c_n) \to 0$ as $n \to \infty$. Since Φ has integrable envelope, $\|f_\theta\|^2$ is bounded. The Cauchy-Schwartz inequality gives that covariances of the processes using F and $\dot{F}$ are equal. Thus, $\hat{J}$ in equation 16 can be replaced by $\cdot J$ while still converging to the same limit.

Finally, note that when priorities are from $\dot{F}$, the two-stage sampling trick provides a means to rescale the threshold. This gives $\cdot J_n(\theta, \alpha\delta t) \stackrel{d}{=} \cdot J_B(\theta, \delta t)$ where $B \sim Binomial(n, \alpha)$. Using $\dot{F}$ in place of F and rescaling the threshold in equation 16 by $c_n MT$ yields equation 17.

In the thinning stage, the role of ϵ is simply to ensure that $F(c_n t(X)|X)$ can be upper bounded by $c_n(MT + \epsilon)$. Any value of ϵ that yields an upper bound yields exactly the same process in the end. Since $\dot{F}(c_n t(X)|X) \le MT$ already, ϵ can be set to 0.

□

Although the condition on the CDFs requiring a linear expansion near 0 may appear restrictive, we note that it is satisfied under reasonable settings. If the priorities are drawn from conditional densities $f(\cdot|X)$ that are differentiable in a neighborhood $[0, \delta)$ with $\delta > 0$ and if $f(0|\cdot)$ is both upper bounded and bounded away from 0, then a Taylor expansion ensures the condition is satisfied. Furthermore, as long as the rate at which items are sampled does not depend on x, the original priorities can be transformed into one in which the CDF has a linear expansion.

Lemma 13. *Consider priorities R_i taking values in the non-negative reals. Further suppose their conditional distributions $F(\cdot|x)$ are continuous in a neighborhood of 0 with $F(0|x) = 0$ and there exists constants w_x with $\inf_x w_x > 0$ such that for $\delta \to 0^+$*

$$\sup_{x,y}\left|\frac{F(\delta|x)}{F(\delta|y)} - \frac{w_x}{w_y}\right| = o(1). \tag{19}$$

There exist independently drawn priorities with conditional distribution $\dot{R}_i|X_i = x \sim Uniform(0, 1/w_x)$ and a monotone transformation ρ such that $p(1(\dot{R}_i < t) \ne 1(\rho(R_i) < t)) = o(t)$.

Proof. The conditions imply $\sup_x |F(\delta|x) - \alpha_x\eta(\delta)| = o(\eta(\delta))$ by taking $\eta(\delta) = \alpha_y^{-1}F(\delta|y)$ for some fixed y. Since η is increasing and continuous in a neighborhood of 0, for sufficiently small b we can simply rescale the priorities to obtain $\tilde{R}_i = b\eta^{-1}(R_i)$ if $R_i \in [0, b)$ and $\tilde{R}_i = R_i$ otherwise. This defines the function ρ. Denote the CDF of a $Uniform(0, 1/w_x)$ distribution as $\dot{F}(r|x) = w_x r$. The priorities $\tilde{R}_i = \tilde{F}^{-1}(U_i|X_i)$ and $\dot{R}_i = \dot{F}^{-1}(U_i|X_i)$ can be obtained from the inverse probability transform for their corresponding CDFs. Thus, the indicators are not equal only if $\dot{F}(t|X_i) \le U_i < \tilde{F}(t|X_i)$ or $\tilde{F}(t|X_i) \le U_i < \dot{F}(t|X_i)$. Each of these happen with probability less than $\sup_x |\tilde{F}(t|x) - \dot{F}(t|x)| = o(t)$. □

6 OTHER APPLICATIONS OF ASYMPTOTIC THEORY

While we have already shown that the asymptotic theory justifies the re-use of consistent estimators for Poission sampling, we can also use it to justify the use of heuristically constructed thresholds.

Consider again the problem of building a sample that can provide an absolute variance guarantee. Recall that in section 3.9, it is not sufficient to choose a threshold where the estimated variance is equal to the target variance. Additional points must be sampled to ensure that the chosen threshold is the largest of such thresholds. This added layer of complexity can be removed by applying the asymptotic theory to show that the heuristically constructed threshold without oversampling still leads to consistent estimators.

Our Donsker result show that the HT variance estimate in after centering and rescaling $\widehat{Var}(\hat{\theta}_t) \approx Var(\hat{\theta}_t)) + \Psi_t/\sqrt{n}$ where Ψ_t is a zero-mean Gaussian process. Since $Var(\hat{\theta}_t))$ is increasing with t and the error term $\Psi_t/\sqrt{n}$ is decreasing with n, if the variance estimator's variance is not too large, then a Gaussian maximal inequality can be used to show the heuristically chosen threshold without the additional oversampling step is close to the threshold with the desired variance.

7 FUTURE WORK AND CONCLUSION

We have provided a general framework for building sampling schemes that adapt to the data on the fly and satisfy system constraints while behaving similarly to fixed thresholds. This simplifies creating new sampling schemes while making the resulting sampled data sets easy to analyze. This framework unifies a long line of research on bottom-k sampling [**?**].

We demonstrate its usefulness by providing sampling schemes that can solve a variety of existing and new problems. These include engineering problems of fitting within system and budgetary constraints, usability problems that allow users to tune the desired accuracy for AQP results at query time, and novel problems such as top-k queries. This flexibility in the framework can make it more useful in practice, by making it easier to design systems and satisfy user requirements. At the same time, it ensures that the same estimators used for fixed thresholds can also be used for adaptive thresholds, making it easy to code just one set of estimators while

the underlying sampling schemes can be easily changed. Furthermore, these are just a few examples that apply the framework. Future work expands on other potential applications that can be solved with this sampling framework.

REFERENCES

[1] D. Anderson, P. Bevan, K. Lang, E. Liberty, L. Rhodes, and J. Thaler. A high-performance algorithm for identifying frequent items in data streams. *Internet Measurement Conference*, 2017.

[2] Apache Software Foundation. Datasketches.

[3] K. Beyer, R. Gemulla, P. J. Haas, B. Reinwald, and Y. Sismanis. Distinct-value synopses for multiset operations. *Communications of the ACM*, 52(10):87–95, 2009.

[4] V. Braverman, R. Ostrovsky, and C. Zaniolo. Optimal sampling from sliding windows. In *PODS*. ACM, 2009.

[5] A. Z. Broder. On the resemblance and containment of documents. In *Proceedings. Compression and Complexity of SEQUENCES 1997 (Cat. No. 97TB100171)*, pages 21–29. IEEE, 1997.

[6] E. Cohen. Multi-objective weighted sampling. In *Hot Topics in Web Systems and Technologies (HotWeb), 2015 Third IEEE Workshop on*, pages 13–18. IEEE, 2015.

[7] E. Cohen, N. Duffield, H. Kaplan, C. Lund, and M. Thorup. Stream sampling for variance-optimal estimation of subset sums. In *SODA*. Society for Industrial and Applied Mathematics, 2009.

[8] E. Cohen and H. Kaplan. Summarizing data using bottom-k sketches. In *PODC*, 2007.

[9] E. Cohen and H. Kaplan. Leveraging discarded samples for tighter estimation of multiple-set aggregates. In *ACM SIGMETRICS Performance Evaluation Review*, volume 37, pages 251–262. ACM, 2009.

[10] G. Cormode, F. Korn, and S. Tirthapura. Time-decaying aggregates in out-of-order streams. In *PODS*, pages 89–98. ACM, 2008.

[11] A. Dasgupta, K. J. Lang, L. Rhodes, and J. Thaler. A framework for estimating stream expression cardinalities. In *19th International Conference on Database Theory*, 2016.

[12] N. Duffield, C. Lund, and M. Thorup. Priority sampling for estimation of arbitrary subset sums. *Journal of the ACM (JACM)*, 54(6):32, 2007.

[13] P. S. Efraimidis and P. G. Spirakis. Weighted random sampling with a reservoir. *Information Processing Letters*, 97(5):181–185, 2006.

[14] R. Gemulla and W. Lehner. Sampling time-based sliding windows in bounded space. In *SIGMOD*. ACM, 2008.

[15] F. Giroire. Order statistics and estimating cardinalities of massive data sets. *Discrete Applied Mathematics*, 157(2):406–427, 2009.

[16] P. Halmos. The theory of unbiased estimation. *The Annals of Mathematical Statistics*, 17(1):34–43, 1946.

[17] S. Heule, M. Nunkesser, and A. Hall. Hyperloglog in practice: Algorithmic engineering of a state of the art cardinality estimation algorithm. In *EDBT*, 2013.

[18] M. Kolonko and D. Wäsch. Sequential reservoir sampling with a nonuniform distribution. *ACM Transactions on Mathematical Software*, 32(2):257–273, June 2006.

[19] M. R. Kosorok. *Introduction to empirical processes and semiparametric inference*. Springer, 2008.

[20] J. Li, Z. Li, Y. Xu, S. Jiang, T. Yang, B. Cui, Y. Dai, and G. Zhang. Wavingsketch: An unbiased and generic sketch for finding top-k items in data streams. *KDD*, 2020.

[21] A. Metwally, D. Agrawal, and A. El Abbadi. Efficient computation of frequent and top-k elements in data streams. In *ICDT*, 2005.

[22] J. Misra and D. Gries. Finding repeated elements. *Science of computer programming*, 2(2):143–152, 1982.

[23] B. Rosén. Asymptotic theory for order sampling. *Journal of Statistical Planning and Inference*, 62(2):135–158, 1997.

[24] B. Rosén. On sampling with probability proportional to size. *Journal of Statistical Planning and Inference*, 62(2):159–191, 1997.

[25] N. T. Spring and D. Wetherall. A protocol-independent technique for eliminating redundant network traffic. In *Proceedings of the conference on Applications, Technologies, Architectures, and Protocols for Computer Communication*, pages 87–95, 2000.

[26] M. Szegedy. The dlt priority sampling is essentially optimal. In *STOC*, pages 150–158. ACM, 2006.

[27] Y. Tillé. *Sampling algorithms*. Springer, 2006.

[28] D. Ting. Towards optimal cardinality estimation of unions and intersections with sketches. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1195–1204. ACM, 2016.

[29] D. Ting. Data sketches for disaggregated subset sum and frequent item estimation. In *Proceedings of the 2018 International Conference on Management of Data*, pages 1129–1140. ACM, 2018.

[30] S. Tirthapura and D. Woodruff. Optimal random sampling from distributed streams revisited. *Distributed Computing*, pages 283–297, 2011.

[31] Q. Xiao, S. Chen, M. Chen, and Y. Ling. Hyper-compact virtual estimators for big network data based on register sharing. *SIGMETRICS Perform. Eval. Rev.*, 43(1):417–428, June 2015.