Count-Min: Optimal Estimation and Tight Error Bounds using Empirical Error Distributions

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

ABSTRACT

The Count-Min sketch is an important and well-studied data summarization method. It can estimate the count of any item in a stream using a small, fixed size data sketch. However, the accuracy of the Count-Min sketch depends on characteristics of the underlying data. This has led to a number of count estimation procedures which work well in one scenario but perform poorly in others. A practitioner is faced with two basic, unanswered questions. Given an estimate, what is its error? Which estimation procedure should be chosen when the data is unknown?

We provide answers to these questions. We derive new count estimators, including a provably optimal estimator, which best or match previous estimators in all scenarios. We also provide practical, tight error bounds at query time for all estimators and methods to tune sketch parameters using these bounds.

The key observation is that the full distribution of errors in each counter can be empirically estimated from the sketch itself. By first estimating this distribution, count estimation becomes a statistical estimation and inference problem with a known error distribution. This provides both a principled way to derive new and optimal estimators as well as a way to study the error and properties of existing estimators.

ACM Reference Format:
Daniel Ting. 2018. Count-Min: Optimal Estimation and Tight Error Bounds using Empirical Error Distributions. In *KDD '18: The 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, August 19–23, 2018, London, United Kingdom.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3219819.3219975

1 INTRODUCTION

The Count-Min sketch has proven to be one of the most effective sketches for obtaining approximate counts for pointwise queries and for computing approximate inner products. Having such a summarization has become increasingly important in the current world of huge streaming datasets. For example, the ad prediction and reporting problem often relies on computing historical click and ad impression counts for every ad, across billions of users broken down by country, IP, and other dimensions [25]. The combinatorial number of possible breakdowns grows the number of counters needed for exact aggregations to the trillions and beyond, making storage and computation intractable.

However, there are notable cases where the sketch performs suboptimally or poorly. For example, when there are few heavy hitters and a large number of items, the Count-Min sketch can be highly biased and perform poorly compared to the Count sketch [3]. This has led to a number of attempts [16], [19], [10], [20], [4] to improve estimation from the Count-Min sketch in these regimes. In all cases, these methods can be shown to perform suboptimally in some regimes or for some sketch parameter settings and often worse than the basic Count-Min estimator. As a result, it is unclear to a practitioner which method to choose. Although several empirical studies [24], [7] have attempted to address this issue, choosing the best method has required a priori knowledge of the unseen data's properties. A second issue with the Count-Min sketch is that there is no practical estimate of the error that can be reported for a query. Although it has a probabilistic error guarantee, this guarantee is extremely loose and of little to no practical use.

This paper introduces methods that provide better accuracy than existing methods under all regimes and provide tight, practical error bounds. This takes the guesswork out of count estimation. Our approach treats count estimation from the Count-Min sketch as a statistical estimation problem where the irrelevant counts are modeled as error terms.

The key idea is that the *distribution* of these error terms can be estimated from the sketch itself. Equipped with an error distribution, we develop two classes of estimators: ones which use the full likelihood information and ad-hoc estimators with some good properties. All existing estimators are shown to be from the latter class. For these estimators, we show that bootstrap methods can be used to debias a wide class of estimators and obtain tight confidence intervals for them. For likelihood based methods, we propose two estimators: a maximum likelihood estimator and a Bayesian estimator. The Bayesian estimator, while more computationally expensive, is proved to be optimal even when the sketch is of fixed depth. The more practical maximum likelihood estimator is empirically shown to outperform all other methods in all scenarios. Key to the likelihood based methods is a non-parametric estimate of the error distribution. We show this can be accomplished with log-concave density estimation. This estimator has attractive properties as it requires no tuning parameters and yields a concave log-likelihood function that ensures maximum likelihood estimation is fast and easy. This log-likelihood generates robust count estimators even when the assumption of log-concavity is false.

In addition to the practical improvements motivated by theory, our work also advances our understanding of the Count-Min sketch

KDD '18, August 19–23, 2018, London, United Kingdom

ACM ISBN 978-1-4503-5552-0/18/08.
https://doi.org/10.1145/3219819.3219975

and related sketches. We serve as a brief survey of existing estimation algorithms and summarize the techniques used. We show that unlike existing methods which exploit only one or two techniques, our method is able to exploit all of them to obtain better results.

The paper is structured as follows. First, we review the Count-Min sketch and define the empirical error distribution relative to a pointwise query. Next, we give a brief survey of existing work on improving estimation for the sketch, provide insights into how they work, and show they can be generalized in natural ways. Section 4 then introduces the bootstrap and shows how simple statistics can be converted into unbiased estimators for the count and gives procedures to construct tight error bounds. As simple statistics may not make full use of the information in the data, section 5 shows that the likelihood can be non-parametrically estimated from the data and proposes estimators based on it. Section 7 provides empirical results on real and synthetic data to show that our estimators are indeed the most accurate in a variety of settings and that the error bounds are tight

Throughout the paper we rely heavily on statistical estimation theory and concepts that we unfortunately do not have sufficient space to cover in detail. These concepts are the full distribution based counterparts to the tail probability and concentration inequality driven theory common in the sketching literature.

2 COUNT-MIN

The Count-Min sketch compresses and aggregates a large and possibly unknown number of $(item, count)$ tuples into a finite sketch of $r \times k$ numeric counters. It allows for two basic types of queries: 1) pointwise queries which provide an estimate of the aggregated count for any item or set of items, and 2) inner product queries which provide for an estimate for $u^T v$ for count vectors u and v indexed by distinct items. These two basic queries can be used to formulate more complex queries. For example, aggregated counts for range queries can be constructed out of pointwise queries that expand numeric valued items into membership in a set of dyadic ranges [8]. We focus only on pointwise queries in this paper though some of our techniques can translate to the inner product case.

The Count-Min summarization technique can be decomposed into the construction of the sketch and the estimation procedure for count queries. In this paper, we focus on improvements to estimation and not on sketch construction. For clarity, we will refer to the construction as the Count+ summarization and the estimator as the Min estimator. Here, the plus sign represents the one-sided errors for the sketch.

An $r \times k$ Count+ sketch consists of two parts: a hash based projection and replication. The first hashes each item to one of k counters. The k vector of observed counters is obtained by summing the counts in each bin. The second part simply replicates this process r times with independent hashes. r and k are often referred to as the depth and width of the sketch.

More precisely, given a hash function h, the (item, count) pair (x_i, c_i) updates the counter vector $\mathbf{V}$ by the rule

$$V_{h(x_i)}^{(new)} = V_{h(x_i)}^{(old)} + c_i. \tag{1}$$

Symbol	Definition
$\mathbf{n}$	Vector of all counts indexed by item
$\hat{n}_x$	Estimated count for item x
$\mathcal{I}$	Set of indices that x or S are hashed to
$r \times k$	(# of replicates) $\times$ (counters per replicate)
$h^{(a)}$	Hash function for replicate a
$\mathbf{V}$	Count-Min counters
$V_i^{(a)}$, $V_{(a,i)}$	i^{th} counter in replicate a
ϵ	Vector of errors (relative to some item x)
F, $\mathbb{F}$	True and empirical distribution of errors
$\mathbf{M}$, $M^{(a)}$	Projection matrix for the sketch and for replicate a

Table 1: Table of symbols

This process is repeated r times to obtain independent identically distributed (i.i.d.) vectors $\mathbf{V^{(a)}}$ using independent hashes $h^{(a)}$ for $a = 1, \ldots, r$.

Estimation from this sketch is simple and relies on the fact that counts are non-negative. Denote the vector of true counts indexed by item by $\mathbf{n}$. For any of the k-vectors $V^{(a)}$, the counter $V_{h^{(a)}(x)}^{(a)}$ is an upper bound on the total count n_x for item x. The original Min-estimator for the Count+ sketch takes the minimum over the r replicates

$$\hat{N}_x = \min_a V_{h^{(a)}(x)}^{(a)} \geq n_x. \tag{2}$$

Several simple observations can be made from this construction and estimator. Only the counters that an item is hashed to contain any information about its count. Removing an item and its count from the Count+ summarization yields vectors of exchangeable error terms where the error terms are all non-negative. The Min estimator is biased as it cannot underestimate the count. More formally, for any replicate $V^{(a)}$,

$$V_i^{(a)} = n_x 1(h^{(a)}(x) = i) + \epsilon_i^{(a)} \tag{3}$$

where the $\epsilon_i^{(a)} \geq 0$ are identically distributed and exchangeable.

These observations motivate our basic strategy. Take counters which only contain error terms. Use them to empirically estimate a non-centered, non-negative error distribution. An item's counters plus the error distribution for those counters provides all the available information to estimate the item's count. Statistical estimation techniques then yield count and error estimates. Furthermore, when the estimated error distribution is correct, an optimal estimator can be derived.

2.1 Linear algebra of the Count-Min sketch

The Count+ summarization is an example of a linear sketch. In other words, the sketch is a linear transformation of the underlying counts $\mathbf{n}$. Specifically, each replicate is a random projection $M^{(a)}$ of the counts $\mathbf{n}$ where the construction of $M^{(a)}$ does not depend on $\mathbf{n}$. This may be expressed as

$$V^{(a)} = M^{(a)}\mathbf{n} \tag{4}$$

where $M^{(a)}$ is a $k \times d$ random binary matrix with precisely 1 non-zero value per column. More explicitly, $M_{ix}^{(a)} = 1$ if $h^{(a)}(x) = i$ and

0 otherwise. For succinctness in notation we denote the concatenation of the $V^{(a)}$ as simply $\mathbf{V}$ and likewise for $\mathbf{M}$. We also write $V_i^{(a)}$ by $\mathbf{V}_{a,i}$ and similarly for $\mathbf{M}$.

When only a subset S is of interest, the sketch has the form,

$$\mathbf{V} = \mathbf{M}_{\cdot,S}\mathbf{n}_S + \epsilon(S) \tag{5}$$

$$\epsilon(S) = \mathbf{M}_{\cdot,S^c}\mathbf{n}_{S^c}. \tag{6}$$

Note that the equation representing the counters $\mathbf{V}$ has the same form as a linear regression problem where $\mathbf{M}_{\cdot,S}$ are the known covariates and $\mathbf{n}_S$ are the unknown regression coefficients. The error terms $\epsilon(S)$ are defined relative to the queried items S. It differs from typical linear regression problems in that the errors are not centered to have mean zero, and the distribution of the errors is not known or assumed. For notational convenience, we will simply write ϵ for the error term as S is always clear from the context.

2.2 Empirical error distributions

Thus far, we have defined the form of the statistical model modulo specification of the error distribution. While typical statistical modeling tasks require strong assumptions on the error distribution, the Count+ summarization allows the error distribution to be non-parametrically estimated from the sketch itself. For any pointwise query for an item x, only the r counters that x hashes to provide information about the count n_x. The remaining $r(k-1) \gg r$ counters are draws from an error distribution. This large sample of observed errors is the empirical error distribution.

The significance of having an accurate empirical version of the true error distribution is that it allows us to improve estimation and generate tight error bounds. When its functional form is also estimated, it reduces the count estimation problem to a familiar problem of parameter estimation with a known error distribution where the statistical machinery for optimal and efficient estimation can be applied as we do in section 5.

3 EXISTING WORK

We first examine existing work on improving the Min estimator and then show how our work encompasses all previous estimators. The estimation techniques for the Count+ summary can be categorized into four basic ideas:

(1) Bias reduction
(2) Linear Regression
(3) Support constraints
(4) Robust objective choice

Each existing estimator exploits only one or two of these ideas. For example, the Min estimator exploits only the non-negative support constraint of the error distribution. The Median estimator exploits only a robust L_1 objective choice.

3.1 Debiasing

Most prior work, [10], [16], [4], focus on debiasing the estimator under different choices of objectives. We describe this debiasing operation with a more general procedure and list the choices made by each procedure. This allows us to extend debiasing to a large class of base estimators, such as any quantile.

Let $\mathcal{I}$ be the set of (replicate number, index) $\in \{1,\ldots,r\} \times \{1,\ldots,k\}$ pairs that item x is hashed to. Let T be some function on a set of r counters so that

$$T(V_{\mathcal{I}}) = n_x + T(\epsilon_{\mathcal{I}}). \tag{7}$$

We refer to this as the *translation property* in this paper. Obvious examples of T include the mean, minimum, median, and any quantile. These are also all special cases of maximizers of the form $T(V_{\mathcal{I}}) = \arg\max_\theta J(V_{\mathcal{I}} - \theta)$. For the mean, $J(x) = \|x\|_2^2$, and for the median, $J(x) = \|x\|_1$. Any maximizer of this form has the translation property.

For any T satisfying this property, $T(V_{\mathcal{I}}) - \mu$ is an unbiased estimate for n_x when $\mu = \mathbb{E}T(\epsilon_{\mathcal{I}})$. This yields a general method for constructing a debiased estimator. 1) Choose a function T with the translation property, and 2) find an empirical estimate of the bias μ.

For the hCount* estimator [16], T remains the minimum. To estimate the bias, they explicitly query for a small set of items that are known to have count 0 and take the average of the corresponding estimates. For the CMM estimators [10], T is taken to be the median. Rather than explicitly querying to find noise counters, they use counters that do not to contain the query key to estimate the bias. Since $\mathbb{E}T(\epsilon_{\mathcal{I}}) \approx \mathbb{E}T(\epsilon_{\mathcal{I}'})$ regardless of the sizes of $\mathcal{I}$ and $\mathcal{I}'$, the resulting estimate is nearly unbiased.

Bias Aware estimation [4] proposes other debiased Median and Mean estimators for T. They differ from other debiasing methods since they use information not contained in the sketch itself. Rather than directly applying the mean or median to the set $V_{\mathcal{I}}$ of relevant counters, they compute "debiased counters" $\tilde{V}_i = V_i - \beta(W_i - 1)$ where W_i is the number of items hashed to counter V_i and β is a per item bias estimate. By construction the error terms $T(\tilde{V}_{\mathcal{I}} - n_x)$ have mean 0. However, computing this requires knowing and being able to iterate over the universe of distinct items.

3.2 Regression and Support Constraints

When multiple items counts are estimated together, one item's estimate can reduce the error for another item when there is a hash collision. More formally, equation 5 shows that adding elements to the set S of desired item counts reduces the number of items mapping to the error term. When the added items are heavy hitters, this can substantially reduce the magnitude of the error. The choice of regression model is thus dictated by what one knows about the universe of items and the unknown error distribution

Under the assumption that the error distribution is normal and only a subset S of items are known, one recovers the linear least squares method of [19]. This is equivalent to the solution of the maximization problem

$$\hat{n}_s = \arg\max_\theta \|V - \mathbf{M}_{\cdot,S}\theta\|_2^2. \tag{8}$$

If the entire universe of items is known, the Counter Braids estimation algorithm [20] is guaranteed to be no worse than the Min estimator and can often recover the exact counts. The Counter Braids estimator does so via a message passing algorithm that provides deterministic upper and lower bounds on the estimated counts. We show in full version of this paper that this algorithm can be formulated as a standard optimization problem. It is a cutting plane algorithm [17] for finding the feasible set for an optimization problem, and the feasible set exploits only the non-negative support of error distributions.

Exploiting ideas from both methods yields the general class of regression based procedures that solve the *constrained* optimization problem

$$\hat{n}_S = \underset{\theta \geq 0 \, s.t. \, \mathbf{M}_{\mathcal{I},S}\theta \leq \mathbf{V}_{\mathcal{I}}}{\arg\min} J(\mathbf{V}_{\mathcal{I}} - \mathbf{M}_{\mathcal{I},S}\theta) \tag{9}$$

where J is some loss function. Section 5 will show that an estimated log-likelihood function yields a good loss function.

3.3 Our methods

When the problem is fully modeled by a statistical model, the four techniques for improving estimation in the previous section can be simplified into two: linear regression and modeling the error distribution. The error distribution encodes the bias, support, and optimal objective function to use for count estimation while regression incorporates joint knowledge of multiple counts. Furthermore, the uncertainty of estimates can be inferred from the error distribution. It yields the exact sampling distribution of an estimator and corresponding tight confidence intervals (CIs).

We propose two methods based on non-parametric modeling of the error distribution. First, we propose a class of bootstrap estimators which do not require explicit estimation of the error distribution. This class of estimators can be based off statistics that are fast and easy to compute and implement. It covers all existing debiased estimators and allows for the easy generation of new estimators with good properties. Second, we propose full likelihood based estimators based on a non-parametric estimate of the error density or mass function. These methods ensure all information contained in the sketch can be exploited to yield optimal estimation and can incorporate regression techniques to exploit information about the universe of items.

4 BOOTSTRAP ESTIMATORS

The bootstrap [13] is a technique that yields properties of an estimator T by resampling the *existing* data rather than needing to draw a new independent sample. In particular, it can be used to estimate the bias and variance of an estimator. In our case, the näive bootstrap will not work since there are only a small number r of relevant counters to resample. However, given any statistic T with the translation property, the translation property allows the $r(k-1)$ error counters to be used to estimate the sampling distribution of the statistic: $T(V_{\mathcal{I}}) \overset{d}{=} n_x + T(\epsilon_{\mathcal{R}})$ where $\mathcal{R}$ is an set of r indices spanning the r replicates and $\overset{d}{=}$ indicates equality of distributions. From this,it is easy to debias T by taking $T(V_{\mathcal{I}}) - \mathbb{E}T(\epsilon_{\mathcal{R}}) = n_x + \delta_{\mathcal{I}}$ where $\delta_{\mathcal{I}}$ are now zero-mean error terms. Likewise, a confidence interval for 0 based on $T(\epsilon_{\mathcal{R}})$ can be translated to a confidence interval for n_x. This results in unbiased estimates as shown in theorem 4.1 and tight confidence intervals 4.3.

Theorem 4.1. *Let T be any function that satisfies the translation property. Consider an item x and the collection of indices $\mathcal{I}(x)$ that x is hashed to. Consider the empirical distribution of the counters excluding those in $\mathcal{I}(x)$, and denote expectation under this distribution by $\mathbb{E}_{\mathcal{I}(x)^c}$. Let Y_r be r i.i.d. draws from this distribution. Then,*

$$\hat{n}_x = T(V_{\mathcal{I}(x)}) - \mathbb{E}_{\mathcal{I}(x)^c}T(Y_r) \tag{10}$$

is an unbiased estimator for the count n_x.

Proof. Denote by $\epsilon = \mathbf{V} - \mathbf{M}_{\cdot,x}n_x$ the vector of error terms for item x. By independence of the replicates, the error terms $\epsilon_i \sim F$ for $i \in C$ are i.i.d. from some distribution F whenever each i belongs to a different replicate. Choose C to contain one index per replicate and such that $C \cap \mathcal{I}(x) = \emptyset$. It follows that $\mathbb{E}T(\epsilon_{\mathcal{I}(x)}) = \mathbb{E}T(\epsilon_C) = \mathbb{E}_{\mathcal{I}(x)^c}T(Y_r)$. Hence, $\mathbb{E}\hat{n}_x = n_x + \mathbb{E}T(\epsilon_{\mathcal{I}(x)}) - \mathbb{E}T(\epsilon_C) = n_x$. □

While this theorem constructs an unbiased estimator out of any base statistic T that satisfies the translation property, we note it is possible for the resulting estimate to be negative. When the true counts are always non-negative, it is sensible to truncate the estimate at 0 to ensure all estimates are non-negative as well. This results in a slightly biased estimator. We apply this truncation to all estimators, and hence refer to them as debiased and not unbiased estimators. We also point out that the base statistic T cannot be a truncated statistic as truncated statistics cannot have the translation property.

4.1 Tight error estimation

Although the Count-Min sketch has been useful for estimating counts, the problem of returning a practical error bound has not been addressed before. Figure 1 shows that in the heavy tailed regime where the Count-Min sketch performs well, the existing Markov inequality based confidence intervals are an order of magnitude wider. We show that the empirical error distribution and bootstrap not only yield practical error bounds, but that these bounds are tight.

Previous analyses derive probabilistic bounds using Markov's inequality, $P(\hat{n}_x^{Min} > n_x + c\mathbb{E}\epsilon_1) = P(\epsilon_1 > c\mathbb{E}\epsilon_1)^r \leq c^{-r}$ where r is the sketch depth and ϵ_1 is the noise in a single counter. Since $\mathbb{E}\epsilon_1 \leq n_{tot}/k$, this yields a one-sided $1-\alpha$ confidence interval with width $w = n_{tot}\alpha^{-1/r}/k$ where k is the sketch width. This bound is poor when data is heavily skewed since an item with large count can have an arbitrarily large effect on the bound but is highly unlikely to affect the estimate for any given item. While this bound has been improved [9], [6], the improvement depends on knowledge of the top heavy hitters or a strong assumption that the count distribution is Zipf.

By comparison the bootstrap on a T with the translation property gives not an inequality, but the equality

$$P(T(V_{\mathcal{I}}) \geq n_x + w) = P(T(\epsilon_I) \geq w) \tag{11}$$

which can be computed almost exactly from the empirical error distribution without any additional knowledge of the heavy hitters or distributional assumptions.

More precisely, theorem 4.2 shows the bootstrap can be used to construct confidence intervals that have the correct finite sample coverage in all situations. A confidence interval $R(V)$ for n_x at level $1-\alpha$ is a probabilistic error bound which guarantees that $P(n_x \in R(V)) \geq 1-\alpha$. These intervals may be one-sided like those obtained from Markov's inequality, symmetric, or asymmetric. Corollary 4.3 shows any of the bootstrap confidence intervals are tight in the sense that any procedure that always produces shorter intervals must violate the desired error bound.

Theorem 4.2. *Let u_q be the q quantile of the empirical distribution $\mathbb{G}$ of $T(Y_r)$. The interval $[T(V_{\mathcal{I}(x)}) - u_b, T(V_{\mathcal{I}(x)}) - u_a]$ is a $(b-a)$ confidence interval for the count n_x. The coverage of the interval is*

$\mathbb{G}(u_b) - \underline{\mathbb{G}}(u_a)$ *where* $\underline{\mathbb{G}}(y)$ *denotes the probability a draw from the empirical distribution is strictly less than* y *rather than less than or equal to* y.

Proof. By symmetry of the errors, $P(T(\epsilon_{\mathcal{I}(x)}) \in [u_a, u_b]) = \mathbb{E}(\mathbb{G}(u_b) - \underline{\mathbb{G}}(u_a)) \geq (b-a)$. Substituting $T(\epsilon_{\mathcal{I}(x)}) = T(V_{\mathcal{I}(x)}) - n_x$ and rearranging gives the desired result. □

Corollary 4.3. *Any shorter interval* $[T(V_{\mathcal{I}(x)}) - \tilde{u}_b, T(V_{\mathcal{I}(x)} - \tilde{u}_a]$ *with* $[\tilde{u}_a, \tilde{u}_b] \subsetneq [u_a, u_b]$ *has coverage strictly less than* $b - a$.

Proof. $\mathbb{G}(\tilde{u}_b) - \underline{\mathbb{G}}(\tilde{u}_a) < (b - a)$ □

In addition to more precise error bounds, figure 1 also shows the actual coverage of the confidence intervals matches or exceeds the desired coverage while being as tight as possible. The coverage exceeds the desired coverage primarily when the intervals are narrow. In these cases, we verified that excess coverage is due to the discrete jumps in probability in a discrete distribution. Attempts to shorten the intervals yielded insufficient coverage. For example, reducing the intervals by 0.5 on each side and effectively turning the interval from a closed to open interval for discrete counts, reduced the empirical coverage for a 90% CI for the MLE estimator from 0.91 to a less than advertised coverage of 0.88. Thus, the empirical results verify the theory which states they are tight as possible.

In addition to returning error bounds at query time, section 6 demonstrates how tight confidence intervals can be converted to power calculations. This provides a way to optimally tune the sketch parameters based on a small pilot estimate of the count distribution. Figure 3 illustrates the improvement in sketch error obtainable by optimizing the sketch parameters.

4.2 Computation

Bootstrap quantities can pose some computational difficulty as they are typically calculated via Monte Carlo simulation. However, in some cases, the quantities can often be easily computed from the empirical distribution [14]. In particular, the mean and the distribution of any quantile or order statistic can be easily approximated. The order statistic $X_{(i)}$ of a set of items $X_1, \ldots, X_r$ is the i^{th} smallest value in that set. For example the Min estimator is the order statistic equal to the smallest value in a set of r values.

This can be done by relating the distribution of the order statistics from F distributed random variables to those of $Uniform(0,1)$ random variables. Recall that the inverse c.d.f. transform generates a F distributed random variable from a $Uniform(0,1)$ random variable via $Y_i = F^{-1}(U_i)$ for $U_i \sim Uniform(0,1)$. Since F is monotone, the order statistic $Y_{(i)} = F^{-1}(U_{(i)})$. The distribution of $U_{(i)}$ is well-known and is $U_{(i)} \sim Beta(i, r-i+1)$.

When applied to debiasing operations, this gives $\mathbb{E}Y_{(i)} \approx F^{-1}(\mathbb{E}U_{(i)}) = F^{-1}(i/(r+1))$. In particular, the Min estimator can be debiased using the estimated bias $\mu = \mathbb{F}^{-1}(1/(r+1))$ where $\mathbb{F}$ is the empirical distribution of the errors. More importantly, an *exact* confidence interval can be computed directly from $\mathbb{F}$ by using an outer confidence interval [21].

For the Min estimator, a "one-sided" 95% confidence interval for the error is $[0, \mathbb{F}^{-1}(b_{95})]$ where $b_{95} = Beta_r^{-1}(0.95)$ is the 0.95 quantile of a $Beta(1, r)$ distribution. This leads to algorithm 2 which debiases the Min-estimator and provides a confidence interval. We refer to this as a "one-sided" confidence interval since the upper bound cannot be violated. A two-sided interval for the Min or any quantile estimator can be similarly estimated. For the i^{th} order statistic, compute a $1-\alpha$ confidence interval $[a, b]$ for $U_{(i)}$. Theorem 4.2 gives that $[T(V_{\mathcal{I}}) - \mathbb{F}^{-1}(b), T(V_{\mathcal{I}}) - \mathbb{F}^{-1}(a)]$ is a $1-\alpha$ confidence interval for the estimate. For implementation purposes, note that T is the base estimator prior to debiasing.

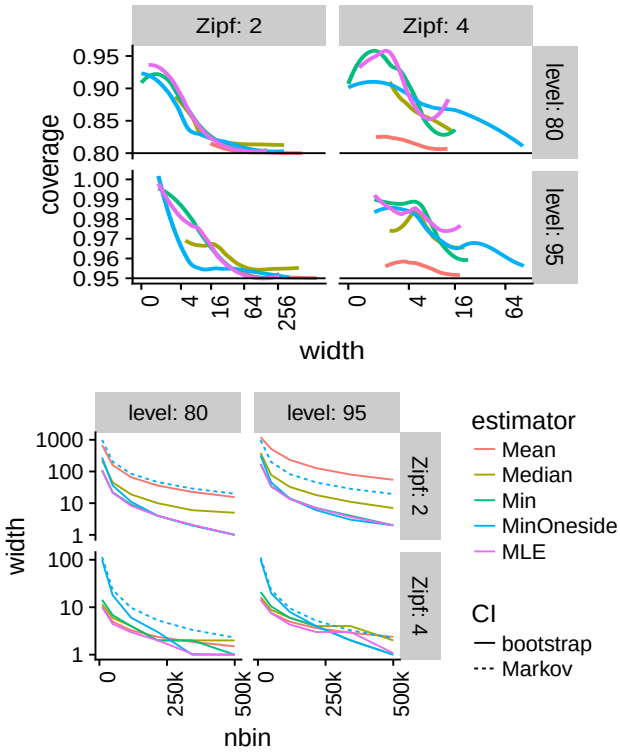

Figure 1: Top: The confidence intervals deliver the promised coverage. Overcoverage is due to the discreteness of the data and probabilites when the accuracy is high. Bottom: The existing probabilistic bound based on Markov's inequality is extremely poor, typically being off by an order of magnitude on a heavy tailed distribution. Except for the "one-sided" interval, all intervals are the two-sided intervals given in algorithm 1.

Even when the bootstrap quantities cannot be directly computed from the distribution of error counters, they can be computed just once and applied to all count estimates. Since quantiles are always robust and most estimators T that we consider are also robust to large errors, there is little difference in estimating the bias μ and interval $[u_a, u_b]$ using all counters rather than only the counters that do not contain a given item. This yields algorithm 1 which debiases an estimator and returns a confidence interval.

5 LIKELIHOOD BASED ESTIMATION

In contrast to bootstrap methods, likelihood based methods must explicitly estimate the error distribution. The benefit is that the

Algorithm 1 Bootstrap debiasing with Confidence Interval

function PRE-PROCESS ERRORS(T, V, a, b)
  for i=1,…,k **do**
    $Y_i = \{V_{(j,i)} : j = 1, \ldots, r\}$
    $Z_i \leftarrow T(Y_i)$
  end for
  Let $\mathbb{G}$ be the empirical c.d.f. of the $\{Z_i\}$
  $(u_a, u_b) \leftarrow (\mathbb{G}^{-1}(a), \mathbb{G}^{-1}(b))$
  return $\mu = \mathbb{E}_{\mathbb{G}} Z$ and $[u_a, u_b]$
end function
function DEBIASED-ESTIMATOR(x, T, V, μ, u_a, u_b)
  $\mathcal{I} \leftarrow \{(i, h^{(i)}(x)) : i = 1, \ldots, r\}$
  $\hat{n}_{raw} \leftarrow T(V_{\mathcal{I}})$
  return $\hat{n}_x = \max\{0, \hat{n}_{raw} - \mu\}$ and $[\hat{n}_{raw} - u_b, \hat{n}_{raw} - u_a]$
end function

Algorithm 2 Debiased Min estimator with Confidence Interval

function DEBIASED COUNT-MIN(x, V, ℓ)
  $\mathcal{I} \leftarrow \{(a, h^{(a)}) : a = 1, \ldots, r\}$
  $\hat{n}_{min} \leftarrow \min_{i \in \mathcal{I}} V_i$
  $\mu \leftarrow k^{th}$ smallest value of V (i.e. $\mathbb{F}^{-1}(1/r)$).
  $b \leftarrow BetaCDF^{-1}(\ell, 1, r)$
  $u_b \leftarrow (b \cdot r \cdot k)^{th}$ smallest value of V (i.e. $\mathbb{F}^{-1}(b)$).
  return $\hat{n}_x = \max\{\hat{n}_{min} - \mu, 0\}$ and
    CI $[\max\{\hat{n}_{min} - u_b, 0\}, \hat{n}_{min}]$.
end function

statistical machinery for efficient estimation and inference can then be applied.

The setup of likelihood based inference is as follows. Denote the unknown true error distribution's cumulative distribution function (c.d.f.) as F and its density or mass function as f. When Y is drawn from a distribution with c.d.f. F, we write $Y \sim F$. In the case of a pointwise query for a single item, the distribution of a counter $V_{a,h^{(a)}(x)} \sim F(\cdot - n_x)$. Estimating the count n_x is a parametric estimation problem from the one-parameter location family $\{F(\cdot - \theta)\}_{\theta \geq 0}$ of distributions.

As can be seen from above, the estimation problem depends primarily on a good estimate of the error distribution. Unlike the bootstrap case, the functional form of the error distribution is needed for count estimation. We show how this can be estimated non-parametrically and without any additional tuning parameters. This allows the easy application of maximum likelihood estimation as well as Bayes optimal estimation under moderate assumptions. Furthermore, the likelihood based approaches provide a framework for performing joint estimation of counts via regression to obtain even more accurate estimates.

5.1 Log-concave density estimation

To ensure good performance under all possible count distributions, we use a non-parametric estimate of the error distribution under the modest assumption that the distribution of the log-errors are log-concave. The concavity has the added benefit that the continuous relaxation of the maximum likelihood objective is easily maximized by standard concave maximization algorithms. Furthermore, unlike other non-parametric methods such as kernel density estimation, a log-concave density has a consistent maximum likelihood estimator [11] that requires no tuning of parameters such as the bandwidth.

Log-concave densities cover many common distributions. These include the Poisson, Binomial, Exponential, Normal, Negative-Binomial, among others. We remark that heavy tailed distributions with probability $f(y) \propto y^{-\alpha}$ for large y have a log density or log mass function that is log-convex in the tails rather than concave. In this case, we compute a log-concave projection of the trimmed density which results in linearly decaying tails. The resulting objective function is a *robust* objective which can perform well even when the assumptions are not met. This robustness is illustrated in the long version of this paper.

We also note that in many commonly used distributions where the log-concavity assumption is invalid, the density or mass function is monotone decreasing. Though non-parametric density estimators for decreasing densities exist, they are unnecessary for count estimation since theorem 5.1 shows any decreasing density yields the Min estimator as the MLE. For any decreasing density with unbounded support, a log-concave density estimator will also recover a decreasing density as its estimate by theorem 5.2.

We are not aware of precise statements on the computational complexity of the log-concave density estimation algorithms. However, the final estimate of the log density is always a linear spline. Estimating the density with a spline is an optimization problem with constraints equal to the number of knots. We find that our final solutions typically have a small number of knots, 10 to 40, so that fitting the density is inexpensive.

THEOREM 5.1. *Let ϵ_i be i.i.d. non-negative random variables from some decreasing density or mass function $f(x)$ with support $[0, \infty)$ or the non-negative integers $\mathbb{N}$. The maximum likelihood estimator for n given $V_i = n_i + \epsilon_i$ is $\hat{n} = \min_i V_i$.*

PROOF. This trivially follows from comparing the likelihood at $\hat{n}$ to any other point. □

THEOREM 5.2. *Let f be a decreasing probability mass function with finite entropy and $\hat{f}$ be its log-concave projection. It follows that $\hat{f}$ is decreasing.*

PROOF. Given in full version of the paper. □

5.2 Maximum likelihood estimation

When the error density f is known, a standard estimation technique is maximum likelihood estimation. The maximum likelihood estimate (MLE) for the count n_x is given by

$$\hat{n}_x = \arg\max_{\theta} \sum_{i \in \mathcal{I}} \log f(V_i - \theta) \tag{12}$$

where $\mathcal{I}$ is the set of counters that item x hashes to. Although the problem of estimating the count is a non-regular estimation problem where standard asymptotic efficiency arguments for maximum likelihood do not apply, maximum likelihood estimation can still be shown to be asymptotically efficient in regimes where the Min estimator does not achieve the best asymptotic rate [15]. This helps explain why its performance dominates other methods. Figure 2 empirically shows the MLE is always the best estimator identifies

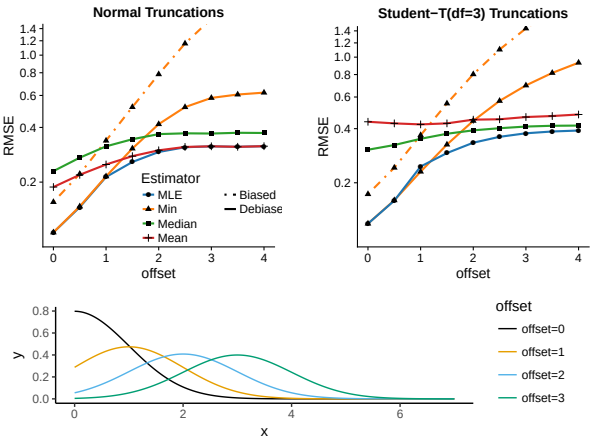

Figure 2: Performance of Debiased estimators under different error distribution shapes. We consider shifting the mode of a distribution (lower figure) and changing the heaviness of the tails. The MLE is always the best or nearly the best estimator. For a truncated normal (left), the Min estimator is optimal in the regime when the mode is near 0, and the Mean is optimal when it is far. For a heavy-tailed t-distribution (right), no simple statistic can match the MLE when the mode is away from 0.

regimes in which it returns significantly more accurate results than any simple statistic.

In practical scenarios where the sketch is shallow and there are a small number of replicates, the asymptotic regime is not reached, and the MLE can be biased. However, since the estimator is of the form given in section 3.1, it can be debiased by the bootstrap procedure in section 4. Empirical results show this additional debiasing step is important for obtaining the best performing estimator in all scenarios as shown in figure 4. Computation in this case can be moderately expensive, however, as there is no analytic form for the sampling distribution of the estimator, unlike for the Min- or other quantile estimators.

5.3 Bayesian estimation

For finite samples, knowledge of the likelihood yields optimal Bayes estimators given a prior and loss function. Given a prior distribution π for the unknown count n_x and error density f, the posterior distribution for N_x is given by

$$p(n_x|V_{\mathcal{I}}) \propto \pi(n_x) \prod_{i \in \mathcal{I}} f(V_i - n_x) \tag{13}$$

where $\mathcal{I}$ is the set of indices x hashes to. Replacing the error density f with its estimate $\hat{f}$ gives an estimated posterior. Given a loss $L(\theta, n_x)$, the optimal Bayesian estimator is the minimizer

$$\hat{n}_x = \arg\min_{\theta} \int L(\theta, n_x) p(n_x|V_{\mathcal{I}}) dn_x. \tag{14}$$

This leads to the optimality result in theorem 5.3. In simple terms, it states that if the number of replicates and average number of distinct items per counter stays the same but the number of error counters goes to infinity, then the Bayes optimal estimator using

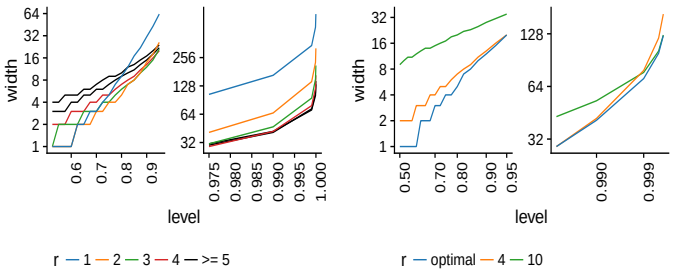

Figure 3: Tuning sketch depths r to optimize the confidence interval width given a *Negative-Binomial*$(30, 0.01)$ count distribution and fixed memory budget. Left: Shallow, wide sketches outperform deep sketches except at high confidence levels. Right: Two previously suggested depth settings are compared to the optimized one for each level. Optimized parameters can yield much narrower confidence intervals.

the approximate posterior converges to the true optimal estimator in probability.

Theorem 5.3. *Let $\{N_i\}_i$ be a sequence of infinitely exchangeable counts with bounded marginal mass function g. Consider a sequence of Count+ summaries on the first Poisson(d) counts where $d \to \infty$ such that $d/k \to \lambda > 0$ and sketch parameters r is fixed and $k \to \infty$. Let f_λ be the mass function of a Compound-Poisson(λ, g) and F_λ be its c.d.f.. Let n_x^{opt} be the optimal Bayes estimator given in equation 14 using a bounded loss function and $\hat{n}_x^{est}$ be the estimator using the approximate posterior obtained by estimating f using the maximum likelihood log-concave density estimator and an atomic mass at 0. Assume f_λ is log-concave and has finite entropy. Further assume that the objective $J(\theta) = \int L(\theta, y) p(y|V) dy$ has a well separated maximum with probability 1. That is, given the maximizer θ_0, if $J(\theta_i) \to J(\theta_0)$ then $\theta_i \to \theta_0$. Then,*

$$\hat{n}_x^{est} - n_x^{opt} \xrightarrow{p} 0. \tag{15}$$

Proof. Given in the full version of the paper. □

We note that this optimality result is a strong finite sample result, as only r counters contain an item's count, rather than an asymptotic optimality result or an even weaker rate result that is typical in the literature. Only finitely many replicates are observed for each item of interest.

6 TUNING SKETCH PARAMETERS

Although our methods take the guesswork out of what estimation procedure to choose, the sketch creator must still choose the number of replicates r and the number of counters per replicate k, or width. The original Count-Min paper [8] suggests choosing these to minimize the space required to achieve a desired error guarantee. For the guarantee, $P(\hat{n}_x \leq n_x + \epsilon \|\mathbf{n}\|_1) < \delta$, their error bound yields the suggestion $r = \lceil \log(1/\delta) \rceil$ and $m = \lceil e/\epsilon \rceil$. It has been suggested [6] that typically $r \approx 10 - 30$ in practice but can be as low as 4 [7] without obvious ill-effects. Several industry implementations such as the RedisLabs module [23] choose a default of $r = 10$.

The previous suggestion finds the smallest sketch that will guarantee a certain confidence level and interval width based on a loose confidence bound. The same can be applied to our tight confidence intervals. We demonstrate how this can be done efficiently without trial and error by using the counter distribution of a sketch. It is easy to show that the counter distribution for a sketch is asymptotically *Compound-Poisson*(λ, g) where λ is the mean number of distinct items per counter, and g is the distribution of item counts.

We first consider the natural case where there is a fixed memory budget $B = rk$, and one desires the smallest interval width. Since the asymptotic theory suggests the region where the Min estimator is optimal or near optimal is the best regime, it is sensible to minimize the width of the Min estimator's interval. Let F_λ be the distribution function of a *Compound-Poisson*(λ, g) distribution and $Beta_r$ be the distribution function of a $Beta(1, r)$ random variable. Given a desired confidence level ℓ for the one-sided confidence interval, the choice of r is

$$\hat{r}_\ell = \arg\min_\rho F^{-1}_{\rho \cdot d/B}(Beta^{-1}_\rho(\ell)). \tag{16}$$

where d is the number of distinct items.

This is easily computed from a single $1 \times B$ Count+ summary and without knowledge of the number of distinct items. The summary provides the error distribution F_{λ_0} where the rate $\lambda_0 = d/B$ and a corresponding density estimate of f_{λ_0}. The superposition theorem for Poisson processes [18] easily gives that the error distribution for any choice of parameters $r \times k$ can be computed as the convolutional power $f^{*r}_{\lambda_0}$. This can be efficiently computed using a Fast-Fourier transform. Figure 3 illustrates how the interval width changes with r for a range of confidence levels and fixed memory budget.

Furthermore, the underlying data can be downsampled using coordinated or bottom-k sampling [5] to estimate error distributions with even smaller rates. This allows one to explore the confidence interval widths for a range of sketch sizes as well.

As an illustration of how this can be applied in a database system, consider the Google N-gram viewer which deals with the canonical natural language processing task of computing counts of n-grams. An n-gram is of a sequence of n words. For example, "An n-gram consists" is a 3-gram. The number of n-grams and possible pointwise queries is very large. One study [26] found there were on the order of 10^{10} unique 5-grams in 100 million English web pages out of which $\approx 10^9$ appeared at least 5 times. Naively tuning parameters is costly. It requires computing a large number of exact counts as well as repeatedly computing a sketch and estimated counts for a large number of parameter settings. Our method shows that no true counts need to be computed, the error is obtained by a single quantile calculation, and only one sketch needs to be computed for all parameter settings.

Even when prior information about the error or count distribution is unavailable, the asymptotic theory provides guidance on how to choose the sketch parameters as wider sketches tend to be closer to the "super-efficient" regime where the Min estimator is nearly optimal.

7 EMPIRICAL RESULTS

We test our MLE estimator in a variety of real and synthetic situations. It is shown to match or best other estimators in all situations. We also empirically show that our confidence intervals provide the correct coverage. A comparison of these tight bounds with prior bounds shows that they are orders of magnitude better.

For synthetic simulations, we use the family of Zipf-Mandelbrot, or discrete power law, distributions. These distributions have probability mass function given by $p(x) \propto (a + x)^{-\alpha}$ on the positive integers. Here a is some offset that adjusts the mass near 1 with smaller values having a larger mass at 1, and α controls the tail behavior with smaller values having heavier tails. For $\alpha = 2$, the distribution has infinite variance. We always consider a universe with $d = 10^6$ items.

For real world datasets, we used a network and a natural language processing dataset. For network data, we used the CAIDA Anonymized OC48 Internet Traces dataset [1]. In 15 minutes of network traffic there were 21.8 million packets from 1.6 million distinct source addresses and ports. We use a Count+ summary to estimate the number of packets for each source. For natural language processing data, we used the Google N-grams dataset [22] for all 2-grams starting with the letters 'ta'. There are 1.4 million distinct 2-grams out of a total of 713 million.

We used the R package **logcondens** [12] to perform log-concave density estimation though we note there is a corresponding package **logcondiscr** [2] for discrete distribution. We chose the continuous valued density estimation package so that the resulting objective function is continuous and can be easily solved by a standard real-valued optimizer.

Although we do not consider timings for our simulation to be representative for practical implementations as R is slow, we report that count estimation for 2000 counts for a sketch of size 8×10^6 took roughly 4 ms per count on a 2.4Ghz CPU when running on a single thread. Each count estimate used roughly 16 evaluations of the objective function when using the function **optimize** which does not make use of known gradient or Hessian information.

To compare the sketches, we use the root mean squared error and the relative efficiency. The relative efficiency of estimator ϕ_1 to ϕ_2 on random data X is

$$RelativeEfficiency(\phi_1, \phi_2) = \frac{\mathbb{E}\|\phi_2(X) - \theta\|_2^2}{\mathbb{E}\|\phi_1(X) - \theta\|_2^2} \tag{17}$$

where θ are the true values being estimated. For unbiased estimators of real valued θ this computes the ratio of the variances, and under regular assumptions where the variance scales inversely to sample size, the relative efficiency of β represents needing β times more data for estimator ϕ_1 to achieve the same error as ϕ_2

We compare the following estimators: the Min, Debiased Min, Debiased Mean, Debiased Median, MLE, and Debiased MLE estimators. Of these, the MLE estimators are the only completely new estimators. Other estimators benefit from our computational simplification when applicable. For all these estimators, the tight confidence intervals are from our new bootstrap procedure. For each sketch, we estimate the counts for the top 2000 heavy hitters. In simulations, the sketch sizes range in depth from 2 to 16 replicates and width from 10^4 to 5×10^5 counters per replicate. Figure 4 shows the empirical error and efficiency under the real and synthetic scenarios. The debiased MLE estimator is clearly the best estimator under all scenarios.

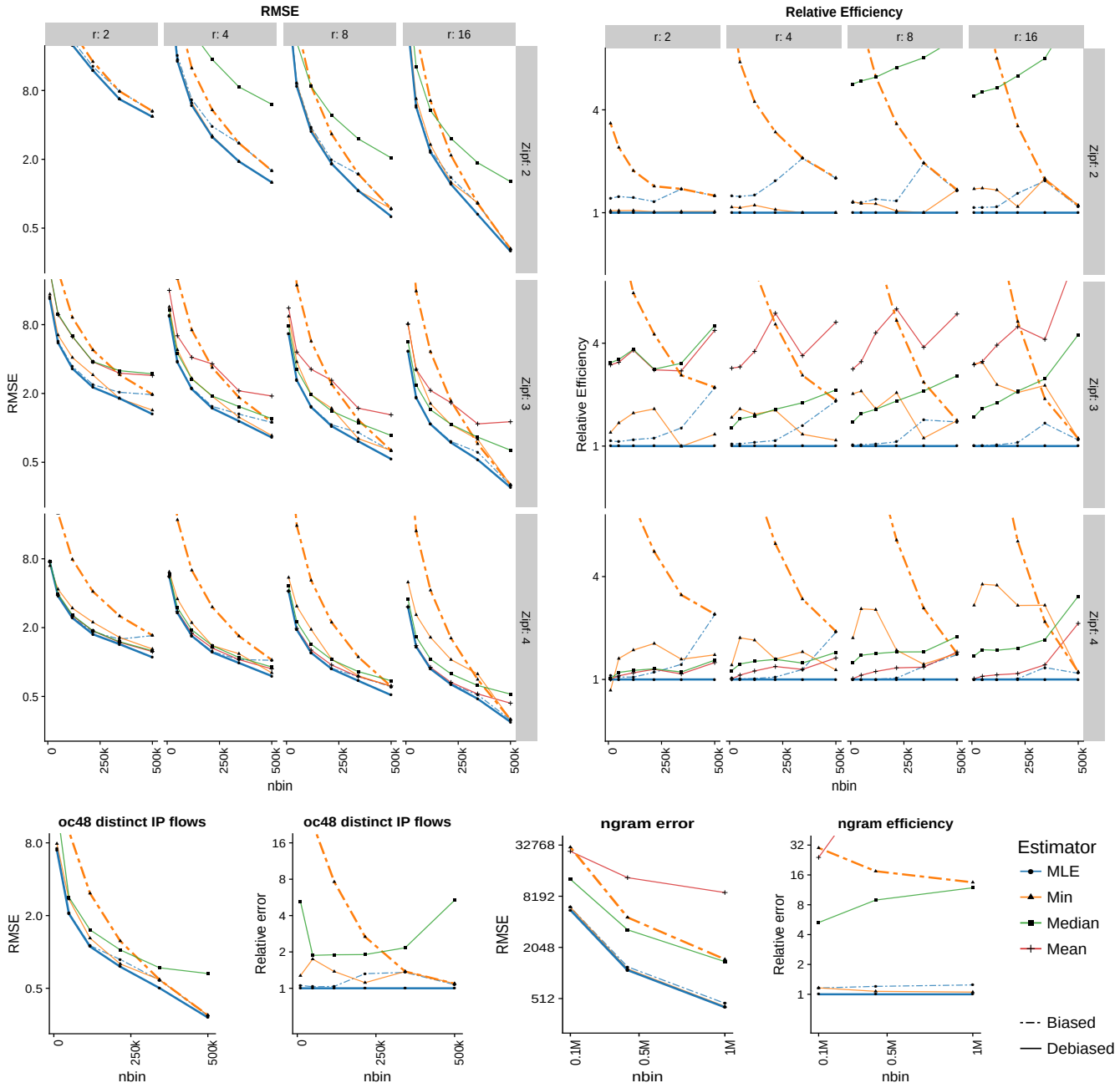

Figure 4: The figures on the top show the performance of different estimators over a range of distribution skews and sketch parameters while the bottom figures are on real world datasets. The Debiased MLE estimator is the most accurate estimator in all scenarios. The Debiased Min estimator is competitive when there are heavy tails and particularly in the real datasets, but the basic Count-Min estimator (orange dashed) is significantly worse than the Debiased MLE (solid blue) estimator. For the n-gram data, the relative efficiency of the Debiased Min estimator is 1.1 times worse than the Debiased MLE estimator while it is 1.3 times worse for the oc48 data when counting the number of distinct ip flows. The difference was negligible when counting raw ip flows (not shown). Estimators that do not appear on the plots (such as the Mean estimator on the $Zipf(2)$ data) have very large errors that do not fit within the axes.

Figure 1 shows the coverage of the corresponding confidence intervals for each of the estimators. They match the desired confidence levels at all levels in a multitude of settings. The resulting error bounds are orders of magnitude better than those available from theoretical analysis. The long version of this paper also provides experimental results showing the improvement obtained by joint estimation of the counts using regression with an empirically estimated error distribution.

8 DISCUSSION

We discuss the applicability of our techniques to other counting sketches and address computational issues that arise with using empirical error distributions and likelihood based estimators.

The same idea of empirically estimating an error distribution to improve count estimation can be applied to other counting sketches and modifications of the Count+ summary. It is straightforward to apply to linear sketches such as the Count sketch [3] and modifications to the Count+ summary that preserve linearity, for example the time adaptive Ada-sketch [25]. However, that there is little reason to prefer the Count summary over the Count+ summary for pointwise queries. The Count summary is the same as the Count+ summary except item x's counter is randomly incremented by either $-n_x$ or n_x rather than n_x. Thus, the error terms are necessarily more noisy than those in a Count+ summary, and estimation should not be expected to be better when exploiting the full likelihood.

Non-linear sketches such as the Conservative Update Count-Min (CU-CM) sketch result in summaries where the error terms are no longer exchangeable and cannot be used to improve estimation. As a result, there is no debiasing operation nor confidence intervals for it. Hence, it necessarily has poor performance in the same regimes where the standard Count-Min sketch is biased and has poor performance. In these regimes, the error and bias of CU-CM grows linearly $O(\lambda)$ with the number of items λ in the sketch while the error of the Count+MLE estimator will grow with the standard deviation $O(\sqrt{\lambda})$.

Thus far, estimation of the empirical error distribution has been assumed to have manageable computational cost. This is aided by the fact that if a sketch does not change, then the error distribution only needs to be estimated once. This may not be the case in streaming settings. Furthermore, in extremely high throughput situations, the maximum likelihood estimator may also be relatively expensive to compute in comparison to simple estimators like the Min, Mean, and Median. These problems may be alleviated in two ways. First, the estimated error distribution can be updated infrequently. If the empirical distribution is updated only when it can differ by δ so that $\|\mathbb{F}_n - \hat{F}_{current}\|_\infty < \delta$, then the number of times the estimated error distribution is updated is logarithmic in the stream size. The amortized cost of adding a count to the sketch goes to 0. Second, rather than using the MLE estimator, the tight error bounds can be used to periodically select the best simple estimator. Thus, the estimator can smoothly transition from the regime where the Min estimator is optimal to ones where the Mean or some quantile estimator is better.

9 CONCLUSION

This paper addresses a number of practical problems for counting sketches and advances our understanding of the mechanisms by which they work. We provide two distinct primary contributions. 1) We give the first method that produces practical and tight error estimates for a pointwise query, and 2) we derive improved and optimal estimators that make full use of the information contained in the sketch. Besides their immediate contributions to counting sketches, we show they help solve other problems facing a practitioner including which sketch and which count estimator to use and how to select optimal sketch tuning parameters.

REFERENCES

[1] The caida ucsd anonymized passive oc48 internet traces dataset 2003-04-24. http://www.caida.org/data/passive/passive_oc48_dataset.xml.

[2] F. Balabdaoui, H. Jankowski, K. Rufibach, and M. Pavlides. Asymptotics of the discrete log-concave maximum likelihood estimator and related applications. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 75(4):769–790, 2013.

[3] M. Charikar, K. Chen, and M. Farach-Colton. Finding frequent items in data streams. In *International Colloquium on Automata, Languages, and Programming*, pages 693–703. Springer, 2002.

[4] J. Chen and Q. Zhang. Bias-aware sketches. *VLDB*, 2017.

[5] E. Cohen and H. Kaplan. What you can do with coordinated samples. In *RANDOM*, 2013.

[6] G. Cormode, M. Garofalakis, P. J. Haas, and C. Jermaine. Synopses for massive data: Samples, histograms, wavelets, sketches. *Foundations and Trends in Databases*, 4(1–3):1–294, 2012.

[7] G. Cormode and M. Hadjieleftheriou. Finding frequent items in data streams. *VLDB*, 1(2):1530–1541, 2008.

[8] G. Cormode and S. Muthukrishnan. An improved data stream summary: the count-min sketch and its applications. *Journal of Algorithms*, 55(1):58–75, 2005.

[9] G. Cormode and S. Muthukrishnan. Summarizing and mining skewed data streams. In *SIAM International Conference on Data Mining*. SIAM, 2005.

[10] F. Deng and D. Rafiei. New estimation algorithms for streaming data: Count-min can do more, 2007.

[11] L. Dümbgen, K. Rufibach, et al. Maximum likelihood estimation of a log-concave density and its distribution function: Basic properties and uniform consistency. *Bernoulli*, 15(1):40–68, 2009.

[12] L. Dümbgen, K. Rufibach, et al. logcondens: Computations related to univariate log-concave density estimation. *Journal of Statistical Software*, 2010.

[13] B. Efron et al. Bootstrap methods: Another look at the jackknife. *The Annals of Statistics*, 7(1):1–26, 1979.

[14] B. Efron and R. J. Tibshirani. *An introduction to the bootstrap*. CRC press, 1994.

[15] P. Hall. On estimating the endpoint of a distribution. *The Annals of Statistics*, pages 556–568, 1982.

[16] C. Jin, W. Qian, C. Sha, J. X. Yu, and A. Zhou. Dynamically maintaining frequent items over a data stream. In *Proceedings of the twelfth international conference on Information and knowledge management*, pages 287–294. ACM, 2003.

[17] J. Kelley, Jr. The cutting-plane method for solving convex programs. *Journal of the Society for Industrial and Applied Mathematics*, 8(4):703–712, 1960.

[18] J. F. C. Kingman. *Poisson processes*. Wiley Online Library, 1993.

[19] G. M. Lee, H. Liu, Y. Yoon, and Y. Zhang. Improving sketch reconstruction accuracy using linear least squares method. In *Internet Measurement Conference*, 2005.

[20] Y. Lu, A. Montanari, B. Prabhakar, S. Dharmapurikar, and A. Kabbani. Counter braids: a novel counter architecture for per-flow measurement. *SIGMETRICS*, 2008.

[21] J. S. Meyer. Outer and inner confidence intervals for finite population quantile intervals. *Journal of the American Statistical Association*, 82(397):201–204, 1987.

[22] J.-B. Michel, Y. K. Shen, A. P. Aiden, A. Veres, M. K. Gray, J. P. Pickett, D. Hoiberg, D. Clancy, P. Norvig, J. Orwant, et al. Quantitative analysis of culture using millions of digitized books. *science*, 331(6014):176–182, 2011.

[23] RedisLabs. Count-min sketch. https://github.com/RedisLabsModules/countminsketch, 2017.

[24] F. Rusu and A. Dobra. Statistical analysis of sketch estimators. In *SIGMOD*, 2007.

[25] A. Shrivastava, A. C. König, and M. Bilenko. Time adaptive sketches (ada-sketches) for summarizing data streams. *SIGMOD*, 2016.

[26] S. Yang, H. Zhu, A. Apostoli, and P. Cao. N-gram statistics in english and chinese: similarities and differences. In *Semantic Computing, 2007. ICSC 2007. International Conference on*, pages 454–460. IEEE, 2007.