Better Cardinality Estimators for HyperLogLog, PCSA, and Beyond

Dingyu Wang
wangdy@umich.edu
University of Michigan
Ann Arbor, Michigan, USA

Seth Pettie
pettie@umich.edu
University of Michigan
Ann Arbor, Michigan, USA

ABSTRACT

Cardinality Estimation (aka Distinct Elements) is a classic problem in sketching with many applications in databases, networking, and security. Although sketching algorithms are fairly simple, analyzing the cardinality *estimators* is notoriously difficult, and even today the analyses of state-of-the-art sketches like HyperLogLog and PCSA are not very accessible.

In this paper we introduce a new class of estimators called τ-*Generalized-Remaining-Area* estimators, as well as a dramatically simpler approach to analyzing estimators. The estimators of Durand and Flajolet [11], Flajolet et al. [15], and Lang [24] can be seen as τ-GRA estimators for integer values of τ. By using *fractional* values of τ we derive improved estimators for HyperLogLog and PCSA whose variance comes very close to the Cramér-Rao lower bounds.

We also derive τ-GRA-based estimators for the class of Curtain sketches introduced by Pettie, Wang, and Yin [29], which can be seen as a hybrid of HyperLogLog and PCSA with a more attractive simplicity-accuracy tradeoff than both.

CCS CONCEPTS

• **Theory of computation → Sketching and sampling**.

KEYWORDS

cardinality estimation, data summary

ACM Reference Format:
Dingyu Wang and Seth Pettie. 2023. Better Cardinality Estimators for HyperLogLog, PCSA, and Beyond. In *Proceedings of the 42nd ACM SIGMOD-SIGACT-SIGAI Symposium on Principles of Database Systems (PODS '23), June 18–23, 2023, Seattle, WA, USA.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3584372.3588680

1 INTRODUCTION

The Problem. A stream $\mathbf{x} = (x_1, \ldots, x_n)$ of elements from a universe $[U]$ is received one at a time. We wish to maintain a small *sketch* S, whose size is independent of n, so that we can return an estimate $\hat{\lambda}$ to the *cardinality* $\lambda = |\{x_1, \ldots, x_n\}|$. Because $\mathbf{x}$ may be partitioned among many machines and processed separately, it is desirable that the resulting sketches be *mergeable*. For this reason we only consider sketches whose state S depends only on the *set* $\{x_1, \ldots, x_n\}$, i.e., it is insensitive to duplicates and is not a function of the *order* in which elements are processed. See [28] for a longer discussion of mergeability and [9, 29, 32] for *non*-mergeable cardinality sketching.

The Model. The Cardinality Estimation/Distinct Elements problem is studied under two models, each with its own conventions. In the RandomOracle model it is assumed that we have access to a uniformly random hash function $h : [U] \to [0, 1]$. By mapping $\mathbf{x}$ to $h(\mathbf{x}) = (h(x_1), \ldots, h(x_n))$, the state of the sketch S can be updated according to a deterministic transition function. In particular, the distribution of the state of S depends only on the cardinality λ, not $\mathbf{x}$. By convention, estimators for sketches in the RandomOracle model are unbiased (or close to unbiased), and their efficiency is measured by the *relative* variance $\lambda^{-2}\mathbb{V}(\hat{\lambda})$, or equivalently, the standard error $\lambda^{-1}\sqrt{\mathbb{V}(\hat{\lambda})}$.[1] The leading constants in the space usage and variance are typically stated explicitly. See [4, 7, 11, 13–16, 18, 24–28].

In the Standard model we can generate independent random bits, but must explicitly store any hash functions. By convention, the estimators in this model come with an (ϵ, δ)-guarantee (rather than bias and variance guarantees), i.e., $\mathbb{P}(\hat{\lambda} \notin [(1-\epsilon)\lambda, (1+\epsilon)\lambda]) \le \delta$. The space depends on ϵ, δ, U, and is expressed in big-Oh notation, often with large hidden constants. In this model $\Theta(\epsilon^{-2} \log \delta^{-1} + \log U)$ bits of space is necessary and sufficient. See Jayram and Woodruff [21] and Alon, Matias, and Szegedy [1] for the lower bound and Błasiok [5] for the upper bound. See also [2, 3, 17, 20, 22] for other results in the Standard model.

In this paper we assume the RandomOracle model. The sketches used in practice (HyperLogLog, PCSA, k-Min, etc.) all originate in the RandomOracle model and despite being implemented with imperfect hash functions, their empirical behavior closely matches their theoretical analysis [19, 24, 31].

Sketches and Estimators. In 1983 Flajolet and Martin [16] developed the first non-trivial sketch called Probabilistic Counting with Stochastic Averaging (PCSA). A PCSA sketch S_{PCSA} consists of an array of m bit vectors or *subsketches*. The random oracle produces a pair $(h, g)(x)$, where $h(x) \in [m]$ is a uniformly random subsketch index and $g(x) \in \mathbb{Z}^+$ is equal to k with probability 2^{-k}. The bit $S_{\text{PCSA}}(j, k)$ is 1 if there exists an x_i in the stream with $h(x_i) = j$ and $g(x_i) = k$, and 0 otherwise. Define $z(j) = \min\{k : S_{\text{PCSA}}(j, k) = 0\}$ to be the position of the least significant zero in the jth subsketch. Each $z(j)$ is individually a decent estimate of $\log(\lambda/m)$. Flajolet and

[1] We use $\mathbb{P}$, $\mathbb{E}$, and $\mathbb{V}$ for probability mass, expectation, and variance.

PODS '23, June 18–23, 2023, Seattle, WA, USA

ACM ISBN 979-8-4007-0127-6/23/06…$15.00
https://doi.org/10.1145/3584372.3588680

Martin [16] analyzed the "first zero" estimator for PCSA, namely

$$\hat{\lambda}_{\mathrm{FM}}(S_{\mathrm{PCSA}}) \propto m \cdot 2^{\frac{1}{m}\sum_{j=1}^{m} z(j)}$$

and proved it has relative variance about $0.6/m$ and hence standard error about $0.78/\sqrt{m}$. It suffices to keep $\log U$ bits per subsketch, so PCSA requires $m \log U$ bits. Although the "first zero" has better concentration than the "last one," the latter is much cheaper to store. In 2003 Durand and Flajolet [11] implemented this idea in the LogLog sketch S_{LL}, which requires only $m \log\log U$ bits. Here $S_{\mathrm{LL}}(j) = \max\{k : S_{\mathrm{PCSA}}(j,k) = 1\}$. Durand and Flajolet proved that the estimator

$$\hat{\lambda}_{\mathrm{DF}}(S_{\mathrm{LL}}) \propto m \cdot 2^{\frac{1}{m}\sum_{j=1}^{m} S_{\mathrm{LL}}(j)}$$

has relative variance about C_{DF}/m and standard error about $\sqrt{C_{\mathrm{DF}}/m} \approx 1.3/\sqrt{m}$, where $C_{\mathrm{DF}} = \frac{2\pi^2+\log^2 2}{12} < 1.69$.[2] This estimator can be regarded as taking the *geometric mean* of individual estimates $2^{S_{\mathrm{LL}}(1)}, \ldots, 2^{S_{\mathrm{LL}}(m)}$. In 2007, Flajolet, Fusy, Gandouet, and Meunier [15] proposed a better estimator for LogLog based on the *harmonic* mean:

$$\hat{\lambda}_{\mathrm{FFGM}}(S_{\mathrm{LL}}) \propto m^2 \cdot \left(\sum_{j=1}^{m} 2^{-S_{\mathrm{LL}}(j)}\right)^{-1}$$

and called the resulting sketch HyperLogLog. It has relative variance roughly C_{FFGM}/m and standard error $\sqrt{C_{\mathrm{FFGM}}/m} \approx 1.04/\sqrt{m}$, where $C_{\mathrm{FFGM}} = 3\ln 2 - 1 \approx 1.07944$. (The constants C_{DF} and C_{FFGM} are, in fact, limiting constants as $m \to \infty$.)

Optimal Cardinality Sketching. The sketches above consist of m subsketches, where the memory scales linearly with m, and the relative variance with m^{-1}. The most reasonable way to measure the *overall efficiency* of a sketch is by its memory-variance product (MVP). Scheuermann and Mauve [30] experimented with *compressed* versions of PCSA and (Hyper)LogLog,[3] and found Compressed-PCSA to be slightly MVP-superior to Compressed-HyperLogLog. Lang [24] also experimented with these compressed sketches, but used *maximum likelihood* estimators (MLE) instead. He found that using MLE, Compressed-PCSA is *substantially* better than Compressed-HyperLogLog. In general, the MLE $\hat{\lambda}_{\mathrm{MLE}}(S)$ of a sketch S is the λ^* that maximizes the probability of seeing S, conditioned on $\lambda = \lambda^*$ being the true cardinality. The MLE is cumbersome to compute and update. Lang [24] also found that a simple "coupon collector" estimator based on counting the number of 1s in a PCSA sketch gives better estimates than Flajolet and Martin's original estimator $\hat{\lambda}_{\mathrm{FM}}$.

$$\hat{\lambda}_{\mathrm{Lang}}(S_{\mathrm{PCSA}}) \propto m \cdot 2^{\frac{1}{m}\sum_{j=1}^{m}\sum_{k\geq 1} S_{\mathrm{PCSA}}(j,k)}.$$

Lang [24] argued informally that the relative variance of $\hat{\lambda}_{\mathrm{Lang}}$ should be about $(\log^2 2)/m$, which agreed with his experiments.

One annoying feature of all the sketches cited above is that their relative variance (and bias) are not fixed but *multiplicatively periodic* with period factor 2. The magnitude of these periodic functions is tiny, but *independent* of m. Pettie and Wang [28] gave a generic "smoothing" mechanism to get rid of this periodic behavior. They studied the optimality of sketches under the memory-variance product (MVP), where both "memory" and "variance" are interpreted as taking on their information-theorically optimum values. They defined the *Fish-number* of a sketch in terms of (1) its Fisher information, which controls the variance of an optimal estimator (e.g., MLE is asymptotically optimal), and (2) its Shannon entropy, which controls its memory under optimal compression. They found closed form expressions for the entropy and Fisher information of base-q variants of PCSA and LogLog, and discovered that q-PCSA has Fish-number $H_0/I_0 \approx 1.98$ for all q, and q-LogLog has a Fish-number strictly larger than H_0/I_0, but that it tends to H_0/I_0 in the limit, as $q \to \infty$. Here H_0 and I_0 are precisely defined constants.[4] The Fishmonger sketch of [28] is a smoothed, entropy compressed version of PCSA with an MLE estimator, which achieves $1/\sqrt{m}$ standard error with $(1+o(1))mH_0/I_0$ bits of space. Moreover, they give circumstantial evidence that Fishmonger is optimal, i.e., no sketch can achieve Fish-number (memory-variance product) better than H_0/I_0. For example, to achieve 1% standard error, [28] indicates that one needs $(H_0/I_0)/(0.01)^2$ bits, which is about 2.42 kilobytes.

1.1 Dartboards and Remaining Area

Ting [32] introduced a very intuitive *visual* way to think about cardinality sketches he called the *area cutting process*. Pettie, Wang, and Yin [28, 29] described a constrained version of Ting's process they called the *Dartboard* model. The elements of this model are as follows:

- **Dartboard and Darts.** The *dartboard* is a unit square $[0,1]^2$. When an element (dart) $x \in [U]$ arrives, it is *thrown* at a point $h(x) \in [0,1]^2$ in the dartboard determined by the random oracle h.
- **Cells and States.** The dartboard is partitioned into a countable set C of *cells*. Every cell may be *occupied* or *free*. The *state* of the sketch is defined by the set $\sigma \subseteq C$ of occupied cells. The *state space* is some subset of 2^C.
- **Occupation Rules.** If a dart is thrown at an occupied cell, the state does not change. If a dart is thrown at a free cell c, and the current state is σ, the new state is $f(\sigma, c) \supseteq \sigma \cup \{c\}$ in the state space.

Note that the state transition function $f(\sigma, c)$ may force a cell to become *occupied* even though it contains no dart, which occurs in (Hyper)LogLog, for example. See Figure 1. It was observed [28, 32] that the Dartboard model includes all mergeable sketches, and even some non-mergeable ones like the S-Bitmap [8].

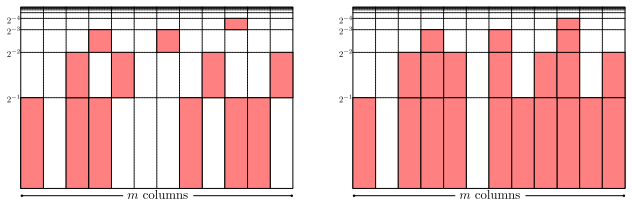
A useful summary statistic of state σ is its *remaining area* $\mathrm{RA}(\sigma) = \sum_{c \in C\setminus\sigma} |c|$, where $|c|$ is the size of cell c. In other words, the remaining area is the total size of all free cells, or equivalently, the probability that the sketch changes upon seeing the next *distinct* element. Remaining area plays a key role in the (non-mergeable) Martingale sketches of [9, 29, 32]. It also gives us a less fancy way to describe the HyperLogLog estimator without mentioning harmonic means: $\hat{\lambda}_{\mathrm{FFGM}}(S_{\mathrm{LL}}) \propto m\,(\mathrm{RA}(S_{\mathrm{LL}}))^{-1}$. Estimating the cardinality proportional to the reciprocal of the remaining area is

[2] All logarithms are natural unless specified otherwise.

[3] It is straightforward to show that the entropy of both sketches is $O(m)$ bits.

[4] $I_0 = \pi^2/6$ measures the Fisher information and $H_0 = \frac{1}{\log 2} + \sum_{k=1}^{\infty} \frac{1}{k}\log_2(1+1/k)$ the Shannon entropy of a PCSA sketch.

reasonable for *any* sketch. This is the optimal estimator for k-Min-type sketches [7, 26], and as we will see, superior to Flajolet and Martin's original $\hat{\lambda}_{\text{FM}}$ estimator for PCSA.



(a) (b)

Figure 1: The cell partition used by PCSA and (Hyper)LogLog. (a) A possible state of PCSA. Occupied (red) cells are precisely those containing darts. (b) The corresponding state of (Hyper)LogLog. Occupied (red) cells contain a dart, or lie below a cell in the same column that contains a dart.

Generalized Remaining Area. Rather than have each cell $c \notin \sigma$ contribute $|c|$ to the remaining area, we could let it contribute $|c|^\tau$ for some fixed exponent $\tau > 0$. The resulting summary statistic is called *τ-generalized remaining area*.

$$\tau\text{-GRA}(\sigma) = \sum_{c \in C \setminus \sigma} |c|^\tau.$$

Note that 0-GRA counts the number of free cells, which we regard as equivalent to counting the number of occupied cells, as is done explicitly by $\hat{\lambda}_{\text{Lang}}$.

2 RELATED WORK

One weakness of HyperLogLog is its poor performance on small cardinalities $\lambda = \tilde{O}(m)$. Heule et al. [19] proposed improvements to [15]'s estimator on small cardinalities, as well as some more efficient sketch encodings when λ is small. Ertl [12] experimented with maximum likelihood estimation (MLE) for HyperLogLog sketches, which behaves well at all cardinalities.

Łukasiewicz and Uznański [25] developed a HyperLogLog-like sketch that, in our terminology, samples $g(x)$ from a *Gumbel* distribution rather than a *geometric* distribution. As the maximum of several Gumbel-distributed variables is Gumbel-distributed, this resulted in a simpler analysis relative to [15].

It is well known that the entropy of HyperLogLog is $O(m)$. Durand [10] gave a prefix-free code for (Hyper)LogLog with expected length $3.01m$, and Pettie and Wang [28] gave a precise expression for the entropy of (Hyper)LogLog, which is about $2.83m$. Xiao et al. [34] proposed lossy compressions of HyperLogLog to $4m$ and even $3m$ bits, but their variance calculation is incorrect; see [28] for a discussion of the problems of lossy compression in this context. Very recently Karppa and Pagh [23] presented a lossless compression of HyperLogLog to $(1 + o(1))m \log\log\log U$ bits ('HyperLogLogLog') while still allowing fast update times.

Pettie, Wang, and Yin [28] proposed a class of *Curtain* sketches that combine elements of LogLog and PCSA while being easily compressible, but they only analyzed them in the *non*-mergeable setting of [9, 32]. Ohayon [27] analyzed the most practical (and mergeable) Curtain$(2, \infty, 1)$ sketch, and found it to be substantially more efficient than HyperLogLog in terms of memory-variance product. In particular, its limiting variance is C/m, $C = \frac{41 \log 2}{16} - 1 \approx 0.776$ while using only m more bits than HyperLogLog or any lossless compression thereof, e.g. [31] or [23].

2.1 New Results

A conceptual contribution of this paper is the introduction of the τ-GRA summary statistic. The main technical contribution is a relatively simple analysis of the limiting relative variance of τ-GRA-based estimators for PCSA, (Hyper)LogLog, and Curtain(d) sketches [29]. A Curtain(d) sketch uses the same cell-partition as PCSA/LogLog. A cell is occupied iff it is hit by a dart, or a cell at least $d + 1$ spots above it in its column is hit by a dart, so PCSA=Curtain(∞) and LogLog=Curtain(0). The space complexity for Curtain(d) is just dm bits more than LogLog.

Our analysis has several benefits.

A Unified View. HyperLogLog is based on 1-GRA and, if properly interpreted, LogLog is based on 0-GRA. Moreover, Lang's "coupon collector" estimator $\hat{\lambda}_{\text{Lang}}$ for PCSA is based on 0-GRA. Our analysis confirms Lang's back-of-the-envelope calculations that $\hat{\lambda}_{\text{Lang}}$ has limiting relative variance $(\log^2 2)/m$.

Simplicity. We use two techniques to dramatically simplify the analysis of τ-GRA-based estimators. The first, which has been used before [15, 16, 28, 29], is to consider a "Poissonized" dartboard model, which allows us to avoid issues with small cardinalities and infinitesimal negative correlations between cells. The second is a smoothing operation similar to the one introduced in [28]. The combined effect of Poissonization and smoothing is to make the sketch truly scale-invariant at every cardinality, without any periodic behavior.

Efficiency. A *statistically* optimal estimator for PCSA or LogLog meets the Cramér-Rao lower bound, which depends on the Fisher information of the given sketch; see [28]. It is known [6, 33] that the maximum likelihood estimator $\hat{\lambda}_{\text{MLE}}$ meets the Cramér-Rao lower bound asymptotically, as $m \to \infty$, but MLE is not particularly simple to update as the sketch changes. The limiting relative variance of HyperLogLog's $\hat{\lambda}_{\text{FFGM}}$ is $(3 \log 2 - 1)/m \approx 1.07944/m$, plus a tiny periodic function. Pettie and Wang's analysis [28, Lemmas 4,5] shows that the Cramér-Rao lower bound for (Hyper)LogLog is $\frac{\log 2}{\pi^2/6 - 1}/m \approx 1.07475/m$, which does not leave much room for improvement! In contrast, there is a wider gap between the limiting variance of PCSA's $\hat{\lambda}_{\text{DF}}$, namely $0.6/m$, or Lang's improvement $\hat{\lambda}_{\text{Lang}}$, namely $(\log^2 2)/m \approx 0.48/m$, and the Cramér-Rao lower bound [28, Theorem 3] of $\frac{\pi^2}{6 \log 2}/m \approx 0.42138/m$. By choosing the optimal τs, our τ-GRA-based estimators achieve relative variance $1.0750/m$ for the LogLog sketch and $0.435532/m$ for the PCSA sketch, in both cases *nearly closing the gap* between the best known explicit estimators and the Cramér-Rao lower bound. HyperLogLog is simple and widely deployed, and PCSA is notable for being the most efficient sketch (in its compressed state) [24, 28, 31]. However, the most attractive sketch in

terms of simplicity of implementation and statistical efficiency are the Curtain(d) sketches of [29]. We prove the limiting relative variance of Curtain(1) sketches is $0.77275/m$ that of Curtain(2) sketches is $0.61699/m$, the latter being at least a 14.26% improvement over HyperLogLog at the same space footprint; see Appendix B for more details.

Figure 2 illustrates the efficiency of τ-GRA-based estimators relative to other estimators. See Table 1 for a symbolic summary of this data.

3 POISSONIZATION AND SMOOTHING

Suppose we have an estimator E_λ at cardinality λ. Ideally, for a statistic to be a *measurement* of cardinality the relative error should distribute identically for any cardinality, i.e., E_λ should be *scale-invariant*.

Definition 1 (scale-invariance)**.** Let E_λ be an estimator of λ. We say E_λ is *scale-invariant* if for any $\lambda > 0$, $\frac{E_\lambda}{\lambda} \sim E_1$.[5]

Note that scale-invariance implies unbiasedness and constant relative variance that is independent of λ. Since $\mathbb{E}E_\lambda = \lambda\mathbb{E}E_1$, if $\mathbb{E}E_1 \neq 1$ then we can replace E_λ with $\frac{E_\lambda}{\mathbb{E}E_1}$ to make it an unbiased estimate of λ. Moreover, $\mathbb{V}(E_\lambda) = \lambda^2\mathbb{V}(E_1)$, where $\mathbb{V}(E_1)$ is some fixed constant independent of λ.

Much of the simplicity and elegance of our analysis relies on beginning from *this* definition of strict scale-invariance. Unfortunately, in the real world the PCSA and HyperLogLog sketches are only *approximately* scale-invariant, stemming from two causes mentioned earlier: the "edge effects" when $\lambda = \tilde{O}(m)$ is small or $\lambda = \tilde{\Omega}(U)$ is very large, and the periodic behavior due to base-2 discretization.

We consider a *smoothed*, *Poissonized*, and *infinite* dartboard model to make the task of variance analysis dramatically simpler.

Definition 2 (Smoothed, Poissonized, Infinite model)**.** The dartboard model and cell partition of PCSA and (Hyper)LogLog are changed as follows.

- **Smoothing.** The sketch consists of m subsketches; these correspond to the columns in Figure 1. Pick a vector $\mathbf{R} = (R_1, \ldots, R_m)$ of *offsets*. Cell j in column i now covers the vertical interval $(2^{-j-R_i}, 2^{-(j-1)-R_i}]$. We will normally pick each $R_i \in [0, 1)$ uniformly at random.
- **Infinite Dartboard.** Rather than index cells by $\mathbb{Z}^+$, index them by $\mathbb{Z}$, i.e., the dartboard has unit width and infinite height. For example, cell -5 covers the vertical interval $(2^{5-R_i}, 2^{6-R_i}]$.
- **Poissonization.** In the usual dartboard, the probability that a cell c remains free at cardinality λ is $(1-|c|)^\lambda \to e^{-|c|\lambda}$ and the correlation between cells vanishes as $\lambda \to \infty$. For simplicity, these asymptotic properties can be achieved even for small λ with *Poissonization*. Informally speaking, with Poissonization, for each insertion, instead of throwing *one* dart at the board, darts *appear* on the board memorylessly with density 1. Formally speaking, for every new insertion, a Poisson point process on the infinite board with density 1 is added to the board, where each point in the process corresponds to a dart. Thus, after λ insertions, the darts on the board form a Poisson point process with density λ. By construction, for any λ—even $\lambda = 1$—the cells are independent and a cell c will remain free with probability precisely $e^{-|c|\lambda}$.

Smoothing eliminates periodic behavior, and the combination of Poissonization and the Infinite Dartboard makes the distribution of the sketch scale invariant for all λ.[6] From this point on, the smoothed, Poissonized, and infinite dartboard model is assumed.

4 ESTIMATION BY GENERALIZED REMAINING AREA

Cardinality estimation can be viewed as a *point estimation* problem where the number of subsketches is the number of independent samples/observations. Classically, one can produce i.i.d. estimates $\left(E_\lambda^{(i)}\right)_{i\in[1,m]}$ of λ with each subsketch and then use the sample mean as the combined estimator. A more general framework is to produce estimates $\left(E_{\lambda;f}^{(i)}\right)_{i\in[1,m]}$ of $f(\lambda)$ for some monotonic function f, then take the sample mean $\frac{1}{m}\sum_{i=1}^m E_{\lambda;f}^{(i)}$, which is concentrated around $f(\lambda)$. Thus we can recover an estimator of λ by applying f^{-1} to the sample mean. This process is summarized as follows.

$$\boxed{\begin{array}{c}E_{\lambda;f}^{(1)}, E_{\lambda;f}^{(2)}, \ldots, E_{\lambda;f}^{(m)}\\ (m \text{ independent estim. of } f(\lambda))\end{array}}$$

$$\xrightarrow{\text{sample mean}} \boxed{\begin{array}{c}\frac{1}{m}\sum_{i=1}^m E_{\lambda;f}^{(i)}\\ (\text{concentrated estim. of } f(\lambda))\end{array}}$$

$$\xrightarrow{\text{apply } f^{-1}} \boxed{\begin{array}{c}f^{-1}\left(\frac{1}{m}\sum_{i=1}^m E_{\lambda;f}^{(i)}\right)\\ (\text{concentrated estim. of } \lambda)\end{array}}$$

An important example is its application to the *remaining area*. The remaining area (of one subsketch) offers a natural estimate for λ^{-1}. One can get a concentrated estimation for λ^{-1} using the sample mean of remaining areas of the subsketches and then take the reciprocal to get a concentrated estimator for λ. This is exactly what the HyperLogLog estimator $\hat{\lambda}_{\text{FFGM}}$ does.

The remaining area estimates λ^{-1} and in general, the τ-GRA estimates $\lambda^{-\tau}$. Let $A_{\lambda;\tau}$ be the τ-generalized remaining area of one subsketch and $A_{\lambda;\tau}^{(1)}, A_{\lambda;\tau}^{(2)}, \ldots, A_{\lambda;\tau}^{(m)}$ be m i.i.d. copies. Thus by the same process, we get a generic estimator $\hat{\lambda}_{\tau;m}$ based on τ-GRA. $\hat{\lambda}_{\tau;m} \propto \left(\frac{1}{m}\sum_{i=1}^m A_{\lambda;\tau}^{(i)}\right)^{-\tau^{-1}}$. For any sketch, it turns out that the induced estimator $\hat{\lambda}_{\tau;m}$ is scale-invariant if the τ-GRA statistic itself is *τ-scale-invariant*. Refer to Appendix A for proofs.

[5] For two random variables X and Y, we write $X \sim Y$ if X and Y has the same distribution.

[6] Our justification for these changes is that they make the analysis simpler, and it does not really matter whether they are implemented in practice once λ is not too small. For example, w.h.p., there is no way to detect whether we are in a unit or infinite dartboard once $\lambda = \Omega(m \log m)$ as all cells indexed by $\mathbb{Z} - \mathbb{Z}^+$ will be occupied. Moreover, as $\lambda \to \infty$ the distribution of the true dartboard converges toward the Poissonized dartboard. Smoothing eliminates the tiny periodic behavior of the estimator, but these effects are too small to worry about unless the magnitude of this periodic function is close to the desired variance, in which case smoothing *should* be implemented in practice.

Sketch & Estimator		Limiting Relative Variance	Citation
PCSA	Flajolet & Martin 1983		[16]
First Zero ($\hat{\lambda}_{\text{FM}}$)		$\approx 0.6/m + \theta(\lambda)$	[16]
Coupon Collector ($\hat{\lambda}_{\text{Lang}}$)		$\approx (\log^2 2)/m + \theta(\lambda) \approx 0.48/m$	[24]
Smoothed 0-GRA		$(\log^2 2)/m$	**Theorem 5**
Smoothed 1-GRA		$\frac{3\ln 2}{4}/m \approx 0.51986/m$	**Theorem 5**
Smoothed τ-GRA	$\tau = 0.343557$	$\approx 0.435532/m$	**Theorem 5**
MLE / Cramér-Rao Lower Bound		$\frac{\pi^2}{6\log 2}/m \approx 0.42138/m$	[28]
LogLog	Durand and Flajolet 2003		[11]
Geometric Mean ($\hat{\lambda}_{\text{DF}}$)		$\frac{2\pi^2+\log^2 2}{12}/m + \theta(\lambda) \approx 1.69/m$	[11]
Harmonic Mean ($\hat{\lambda}_{\text{FFGM}}$)		$(3\log 2 - 1)/m + \theta(\lambda) \approx 1.07944/m$	[15]
Smoothed 0-GRA		$\frac{2\pi^2+\log^2 2}{12}/m \approx 1.69/m$	**Theorem 3**
Smoothed 1-GRA		$(3\log 2 - 1)/m \approx 1.07944/m$	**Theorem 3**
Smoothed τ-GRA	$\tau = 0.889897$	$\approx 1.07507/m$	**Theorem 3**
MLE / Cramér-Rao Lower Bound		$\frac{\log 2}{\pi^2/6-1}/m \approx 1.07475/m$	[28]

Table 1: Relative variance as $m, \lambda \to \infty$. All $\theta(\lambda)$ functions are multiplicatively periodic with period 2, which have a small magnitude independent of m. The "smoothing" mechanism (Section 3) eliminates periodic behavior.

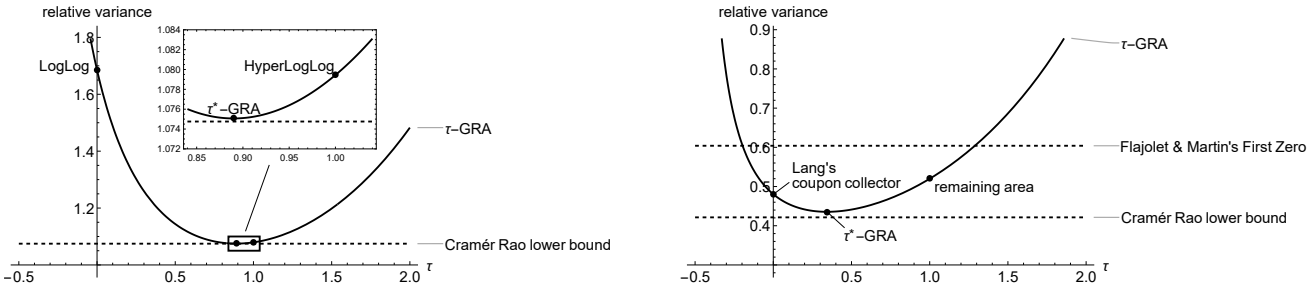


Figure 2: Left: Relative variance of estimators for the LogLog sketch. The τ-GRA estimator attains minimum variance at $\tau^* = 0.88989$, which comes within 0.02% of the Cramér-Rao lower bound. As a comparison, HyperLogLog is 0.4% over the bound. Right: Relative variance of estimators for the PCSA sketch. The τ-GRA estimator attains minimum variance at $\tau^* = 0.343557$, which comes within 3% of the Cramér-Rao lower bound.

Definition 3 (τ-scale-invariance)**.** Let $A_{\lambda;\tau}$ be the τ-generalized remaining area of a sketch. We say $A_{\lambda;\tau}$ is τ-scale-invariant if $A_{\lambda;\tau} \sim \lambda^{-\tau} A_{1;\tau}$ for any $\lambda > 0$.

Theorem 1. *If $A_{\lambda;\tau}$ is τ-scale-invariant, then $\hat{\lambda}^*_{\tau;m} = \left(\frac{1}{m}\sum_{i=1}^m A^{(i)}_{\lambda;\tau}\right)^{-\tau^{-1}}$ is a scale-invariant estimator for λ.*

We prove the following useful theorem that expresses the asymptotic mean and variance of $\hat{\lambda}^*_{\tau;m}$ by the mean and variance of $A_{1;\tau}$ as $m \to \infty$. Although $\hat{\lambda}^*_{\tau;m}$ is scale-invariant, it is not yet normalized to be unbiased; the estimator $\hat{\lambda}_{\tau;m}$ will be the unbiased version of $\hat{\lambda}^*_{\tau;m}$. The asymptotic relative variance after normalization is also given in the theorem.

Theorem 2. *If $A_{1;\tau}$ is τ-scale-invariant with finite variance, we have for any $\lambda > 0$,*

(1) $\lim_{m\to\infty} \mathbb{E}\hat{\lambda}^*_{\tau;m} = \lambda(\mathbb{E}A_{1;\tau})^{-\tau^{-1}}$.

(2) $\lim_{m\to\infty} m\lambda^{-2}\mathbb{V}(\hat{\lambda}^*_{\tau;m}) = \tau^{-2}(\mathbb{E}A_{1;\tau})^{-2\tau^{-1}-2}\mathbb{V}(A_{1;\tau})$.

(3) *For any $\lambda > 0$, the normalized estimator $\hat{\lambda}_{\tau;m} = (\mathbb{E}A_{1;\tau})^{\tau^{-1}}\hat{\lambda}^*_{\tau;m}$ is asymptotically unbiased and has limit relative variance $\lim_{m\to\infty} m\lambda^{-2}\mathbb{V}(\hat{\lambda}_{\tau;m}) = \tau^{-2}(\mathbb{E}A_{1;\tau})^{-2}\mathbb{V}(A_{1;\tau})$.*

Theorems 1 and 2 give us a simple 3-step recipe for calculating the limiting relative variance of τ-GRA-based estimators.

(1) Calculate the mean $\mu = \mathbb{E}A_{1;\tau}$ and the variance $\sigma^2 = \mathbb{V}(A_{1;\tau})$ of the τ-generalized remaining area at density 1.

(2) By Theorem 1, the induced estimator $\hat{\lambda}^*_{\tau;m} = \left(\frac{1}{m}\sum_{i=1}^m A^{(i)}_{\lambda;\tau}\right)^{-\tau^{-1}}$ is a scale-invariant estimator for λ, but possibly biased.

(3) After normalization, we get the estimator $\hat{\lambda}_{\tau;m} = \mu^{\tau^{-1}}\hat{\lambda}^*_{\tau;m}$ which is asymptotically unbiased. By Theorem 2, its relative variance is asymptotically $\tau^{-2}\mu^{-2}\sigma^2/m$.

Results. We apply the 3-step recipe to analyze the relative variance of τ-GRA-based estimators for LogLog and Curtain(d) sketches. See Appendix B for proofs of Theorems 3 and 4.

Theorem 3. *[τ-GRA for the LogLog sketch] Let the offset vector $(R_i) \in [0,1)^m$ be selected uniformly at random. Let $X_\lambda^{(i)}$ be the integer index of the highest one in the ith subsketch after λ insertions. Then for any $\tau > 0$,*

$$\hat{\lambda}_{\tau;m} = m\left(\Gamma(\tau)\frac{1-2^{-\tau}}{\log 2}\right)^{\tau^{-1}}\left(\frac{1}{m}\sum_{i=1}^{m} 2^{-\tau(R_i+X_\lambda^{(i)})}\right)^{-\tau^{-1}}$$

is a scale-invariant estimator for λ that is asymptotically unbiased. The asymptotic normalized relative variance is

$$\lim_{m\to\infty} m\lambda^{-2}\mathbb{V}(\hat{\lambda}_{\tau;m}) = \tau^{-2}\left(\frac{\Gamma(2\tau)\log 2}{\Gamma(\tau)^2}\cdot\frac{1+2^{-\tau}}{1-2^{-\tau}} - 1\right).$$

The celebrated estimator $\hat{\lambda}_{\text{FFGM}}$ of HyperLogLog corresponds to setting $\tau = 1$ in Theorem 3, with $\Gamma(2) = \Gamma(1) = 1$ and the leading constant of the variance being $3\log 2 - 1 \approx 1.07944$. As the τ-mean converges toward the geometric mean as $\tau \to 0$, Durand and Flajolet's estimator $\hat{\lambda}_{\text{DF}}$ corresponds to $\tau \to 0$, with the limiting relative variance constant being $\lim_{\tau\to 0}\tau^{-2}\left(\frac{\Gamma(2\tau)\log 2}{\Gamma(\tau)^2}\cdot\frac{1+2^{-\tau}}{1-2^{-\tau}} - 1\right) = \frac{2\pi^2+\log^2 2}{12} \approx 1.68497$, matching Durand and Flajolet [10, 11]. By numerical optimization, the minimal variance 1.07507 is obtained at $\tau^* = 0.889897$. This comes quite close to the Cramér-Rao lower bound for LogLog sketches, which Pettie and Wang [28] computed to be $\frac{\log 2}{\pi^2/6-1} \approx 1.07475$.

Theorem 4. *[τ-GRA for the Curtain sketch with $d = 1, 2$] Let the offset vector $(R_i) \in [0,1)^m$ be selected uniformly at random. Let $X_\lambda^{(i)}$ be the integer index of the highest cell hit in the ith subsketch after λ insertions. Let $A^{(i)}$ and $B^{(i)}$ be indicators for whether cells $X_\lambda^{(i)} - 1$ and $X_\lambda^{(i)} - 2$ have been hit (and are occupied) in the ith subsketch.*

- *When $d = 1$, for any $\tau > 0$,*

$$\hat{\lambda}_{\tau;m} = \frac{2}{3}m\left(\Gamma(\tau)\frac{1}{\log 2}\right)^{\tau^{-1}}\left(\frac{1}{m}\sum_{i=1}^{m}\left(\frac{1}{2^\tau-1}2^{-\tau(R_i+X_\lambda^{(i)})} + (1-A^{(i)})2^{-\tau(R_i+X_\lambda^{(i)}-1)}\right)\right)^{-\tau^{-1}}$$

  is a scale-invariant estimator for λ that is asymptotically unbiased. The asymptotic normalized relative variance is

$$\lim_{m\to\infty} m\lambda^{-2}\mathbb{V}(\hat{\lambda}_{\tau;m}) = \frac{1}{\tau^2\Gamma(\tau)^2}\left((\log 2)\Gamma(2\tau)(1+2\frac{2^{-2\tau}}{1-2^{-\tau}} + 2\cdot 3^{2\tau}2^{-5\tau}) - \Gamma(\tau)^2\right).$$

- *When $d = 2$, for any $\tau > 0$,*

$$\hat{\lambda}_{\tau;m} = \frac{4}{5}m\left(\Gamma(\tau)\frac{1}{\log 2}\right)^{\tau^{-1}}\left(\frac{1}{m}\sum_{i=1}^{m}\left(\frac{1}{2^\tau-1}2^{-\tau(R_i+X_\lambda^{(i)})} + (1-A^{(i)})2^{-\tau(R_i+X_\lambda^{(i)}-1)} + (1-B^{(i)})2^{-\tau(R_i+X_\lambda^{(i)}-2)}\right)\right)^{-\tau^{-1}}$$

  is a scale-invariant estimator for λ that is asymptotically unbiased. The asymptotic normalized relative variance is

$$\lim_{m\to\infty} m\lambda^{-2}\mathbb{V}(\hat{\lambda}_{\tau;m}) = \frac{1}{\tau^2\Gamma(\tau)^2}\left((\log 2)\Gamma(2\tau)\left(1+2\frac{2^{-3\tau}}{1-2^{-\tau}} + 2(7/5)^{-2\tau}2^{-\tau} + 2(6/5)^{-2\tau}2^{-2\tau}\right) - \Gamma(\tau)^2\right).$$

4.1 Optimal Choice of d

The *natural* way to implement Curtain(d) is to begin with an implementation of (Hyper)LogLog and then supplement it with a $(d\times m)$-bit matrix of *indicators* for the d cells under the highest cell hit in each subsketch. If $U = 2^{64}$, and one dedicates $\log\log U = 6$ bits per LogLog-subsketch, then a $(\log\log U + d)$-bit-per-subsketch implementation of Curtain(d) will be superior even for rather large values of $d \geq 3$.

A common implementation of HyperLogLog [31] is to store the *minimum* index of any subsketch, and then store each subsketch index as a 4-bit *offset*, with $\{0, 1, \ldots, 14\}$ being offsets, and the value 15 indicating that the true offset is ≥ 15 and stored in a separate *exception list*. For reasonable values of m, the length of the exception list is small, and barely influences the space or update time of the data structure. If one begins with a 4-bit implementation of LogLog, Table 2 indicates that the optimal choice of d is 2, using a τ-GRA-based estimator with $\tau = 0.7551$.

Ohayon [27] analyzed the Curtain(1) sketch with the usual 1-GRA (remaining area) based estimator. He called this sketch ExtendedHyperLogLog. A good parameterization of Theorem 4 is to set $d = 2$ and $\tau = 0.7551$, yielding relative variance $0.61699/m$. Compared to a common 4-bit-per-column implementation of HyperLogLog [31], a 6-bit-per-column implementation of Curtain$(d = 2)$ is superior in terms of memory-variance product. See Table 2 for more on these details.

d	τ	standard error	note
0	0	1.4513%	LogLog [11]
0	1	1.1616%	HyperLogLog [15]
0	0.8899	1.1592%	optimal τ-GRA estimator
1	1	1.1012%	ExtendedHyperLogLog [27]
1	0.8941	1.0988%	optimal τ-GRA estimator
2	0.7551	1.0756%	optimal τ-GRA estimator
3	0.6151	1.0794%	optimal τ-GRA estimator

Table 2: 4KB memory is assumed, which is 32000 bits. Assume the baseline HyperLogLog implementation uses 4 bits per subsketch. A Curtain(d) implementation uses d additional bits per subsketch. This table compares algorithms with the *same memory*. For example, with 4KB, there are $32000/4 = 8000$ subsketches for HyperLogLog while if $d = 1$, then each subsketch needs 5 bits and there are only $32000/5 = 6400$ subsketches. The standard error (i.e., square root of the relative variance) is calculated from Proposition 1 and Theorem 2. The calculated standard errors of LogLog, HyperLogLog and ExtendedHyperLogLog match their respective authors' result as $m \to \infty$.

We also analyze τ-GRA-based estimators for PCSA. Theorem 5 is proved in Appendix C.

Theorem 5. *[τ-GRA for the PCSA sketch] Let $A^{(i)}_{\lambda;\tau}$ be the τ-generalized remaining area of the ith subsketch with uniform offsetting, and $A = \sum_{i=1}^{m} A^{(i)}_{\lambda;\tau}$ be the τ-GRA. Then for any $\tau > 0$, $\hat{\lambda}_{\tau;m} = m\left(\frac{\Gamma(\tau)}{\log 2}\right)^{\tau^{-1}}\left(\frac{A}{m}\right)^{-\tau^{-1}}$ is a scale-invariant estimator for λ that is asymptotically unbiased. The asymptotic normalized relative variance is*

$$\lim_{m\to\infty} m\lambda^{-2}\mathbb{V}(\hat{\lambda}_{\tau;m}) = \frac{(1-2^{-2\tau})\Gamma(2\tau)\log 2}{\tau^2\Gamma(\tau)^2}.$$

As $\tau \to 0$, $\hat{\lambda}_{\tau;m}$ is essentially counting the number of free cells (0s in the sketch), which corresponds to Lang's [24] "coupon collector" estimator $\hat{\lambda}_{\text{Lang}}$ that counts occupied cells (1s in the sketch). The limiting variance of this estimator is $\lim_{\tau\to 0}\frac{(1-2^{-2\tau})\Gamma(2\tau)\log 2}{\tau^2\Gamma(\tau)^2} = \log^2 2 \approx 0.480453$, which confirms Lang's [24] back-of-the-envelope calculation that it should be $\log^2 2$. By numerical optimization, the minimal variance 0.435532 is obtained at $\tau^* = 0.343557$. This comes very close to the Cramér-Rao lower bound of $\frac{\pi^2}{6\log 2} \approx 0.42138$ for PCSA sketches, as computed in [28].

REFERENCES

[1] Noga Alon, Yossi Matias, and Mario Szegedy. 1999. The Space Complexity of Approximating the Frequency Moments. *J. Comput. Syst. Sci.* 58, 1 (1999), 137–147. https://doi.org/10.1006/jcss.1997.1545

[2] Ziv Bar-Yossef, T. S. Jayram, Ravi Kumar, D. Sivakumar, and Luca Trevisan. 2002. Counting Distinct Elements in a Data Stream. In *Proceedings 6th International Workshop on Randomization and Approximation Techniques (RANDOM) (Lecture Notes in Computer Science, Vol. 2483)*. 1–10. https://doi.org/10.1007/3-540-45726-7_1

[3] Ziv Bar-Yossef, Ravi Kumar, and D. Sivakumar. 2002. Reductions in streaming algorithms, with an application to counting triangles in graphs. In *Proceedings 13th Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*. 623–632.

[4] Kevin S. Beyer, Rainer Gemulla, Peter J. Haas, Berthold Reinwald, and Yannis Sismanis. 2009. Distinct-value synopses for multiset operations. *Commun. ACM* 52, 10 (2009), 87–95. https://doi.org/10.1145/1562764.1562787

[5] Jarosław Błasiok. 2020. Optimal Streaming and Tracking Distinct Elements with High Probability. *ACM Trans. Algorithms* 16, 1 (2020), 3:1–3:28. https://doi.org/10.1145/3309193

[6] G. Casella and R. L. Berger. 2002. *Statistical Inference, 2nd Ed.* Brooks/Cole, Belmont, CA.

[7] Philippe Chassaing and Lucas Gerin. 2006. Efficient estimation of the cardinality of large data sets, In Proceedings of the 4th Colloquium on Mathematics and Computer Science Algorithms, Trees, Combinatorics and Probabilities. *Discrete Mathematics & Theoretical Computer Science.*

[8] Aiyou Chen, Jin Cao, Larry Shepp, and Tuan Nguyen. 2011. Distinct Counting With a Self-Learning Bitmap. *J. Amer. Statist. Assoc.* 106, 495 (2011), 879–890. https://doi.org/10.1198/jasa.2011.ap10217

[9] Edith Cohen. 2015. All-Distances Sketches, Revisited: HIP Estimators for Massive Graphs Analysis. *IEEE Trans. Knowl. Data Eng.* 27, 9 (2015), 2320–2334. https://doi.org/10.1109/TKDE.2015.2411606

[10] Marianne Durand. 2004. *Combinatoire analytique et algorithmique des ensembles de données. (Multivariate holonomy, applications in combinatorics, and analysis of algorithms).* Ph. D. Dissertation. Ecole Polytechnique X.

[11] Marianne Durand and Philippe Flajolet. 2003. Loglog Counting of Large Cardinalities. In *Proceedings 11th Annual European Symposium on Algorithms (ESA) (Lecture Notes in Computer Science, Vol. 2832)*. Springer, 605–617. https://doi.org/10.1007/978-3-540-39658-1_55

[12] Otmar Ertl. 2017. New Cardinality Estimation Methods for HyperLogLog Sketches. *CoRR* abs/1706.07290 (2017). arXiv:1706.07290 http://arxiv.org/abs/1706.07290

[13] Cristian Estan, George Varghese, and Michael E. Fisk. 2006. Bitmap algorithms for counting active flows on high-speed links. *IEEE/ACM Trans. Netw.* 14, 5 (2006), 925–937. https://doi.org/10.1145/1217709

[14] Philippe Flajolet. 1990. On adaptive sampling. *Computing* 43, 4 (1990), 391–400. https://doi.org/10.1007/BF02241657

[15] Philippe Flajolet, Éric Fusy, Olivier Gandouet, and Frédéric Meunier. 2007. HyperLogLog: the analysis of a near-optimal cardinality estimation algorithm, In Proceedings of the 18th International Meeting on Probabilistic, Combinatorial, and Asymptotic Methods for the Analysis of Algorithms (AofA). *Discrete Mathematics & Theoretical Computer Science*, 127–146.

[16] Philippe Flajolet and G. Nigel Martin. 1985. Probabilistic Counting Algorithms for Data Base Applications. *J. Comput. Syst. Sci.* 31, 2 (1985), 182–209. https://doi.org/10.1016/0022-0000(85)90041-8

[17] Phillip B. Gibbons and Srikanta Tirthapura. 2001. Estimating simple functions on the union of data streams. In *Proceedings 13th Annual ACM Symposium on Parallel Algorithms and Architectures (SPAA)*. 281–291. https://doi.org/10.1145/378580.378687

[18] Frédéric Giroire. 2009. Order statistics and estimating cardinalities of massive data sets. *Discret. Appl. Math.* 157, 2 (2009), 406–427. https://doi.org/10.1016/j.dam.2008.06.020

[19] Stefan Heule, Marc Nunkesser, and Alexander Hall. 2013. HyperLogLog in practice: algorithmic engineering of a state of the art cardinality estimation algorithm. In *Proceedings 16th International Conference on Extending Database Technology (EDBT)*. 683–692. https://doi.org/10.1145/2452376.2452456

[20] Piotr Indyk and David P. Woodruff. 2003. Tight Lower Bounds for the Distinct Elements Problem. In *Proceedings 44th IEEE Symposium on Foundations of Computer Science (FOCS), October 2003, Cambridge, MA, USA, Proceedings*. 283–288. https://doi.org/10.1109/SFCS.2003.1238202

[21] T. S. Jayram and David P. Woodruff. 2013. Optimal Bounds for Johnson-Lindenstrauss Transforms and Streaming Problems with Subconstant Error. *ACM Trans. Algorithms* 9, 3 (2013), 26:1–26:17. https://doi.org/10.1145/2483699.2483706

[22] Daniel M. Kane, Jelani Nelson, and David P. Woodruff. 2010. An optimal algorithm for the distinct elements problem. In *Proceedings 29th ACM Symposium on Principles of Database Systems (PODS)*. 41–52. https://doi.org/10.1145/1807085.1807094

[23] Matti Karppa and Rasmus Pagh. 2022. HyperLogLogLog: Cardinality Estimation With One Log More. In *Proceedings 28th ACM Conference on Knowledge Discovery and Data Mining (KDD)*. 753–761. https://doi.org/10.1145/3534678.3539246

[24] Kevin J. Lang. 2017. Back to the Future: an Even More Nearly Optimal Cardinality Estimation Algorithm. *CoRR* abs/1708.06839 (2017). arXiv:1708.06839

[25] Aleksander Łukasiewicz and Przemysław Uznański. 2022. Cardinality estimation using Gumbel distribution. In *Proceedings 30th European Symposium on Algorithms (ESA)*.

[26] Jérémie Lumbroso. 2010. An optimal cardinality estimation algorithm based on order statistics and its full analysis. In *Proceedings of the 21st International Meeting on Probabilistic, Combinatorial, and Asymptotic Methods in the Analysis of Algorithms (AofA)*. 489–504.

[27] Tal Ohayon. 2021. ExtendedHyperLogLog: Analysis of a new Cardinality Estimator. *CoRR* abs/2106.06525 (2021). arXiv:2106.06525

[28] Seth Pettie and Dingyu Wang. 2021. Information theoretic limits of cardinality estimation: Fisher meets Shannon. In *Proceedings 53rd Annual ACM Symposium on Theory of Computing (STOC)*. 556–569. https://doi.org/10.1145/3406325.3451032

[29] Seth Pettie, Dingyu Wang, and Longhui Yin. 2021. Non-Mergeable Sketching for Cardinality Estimation. In *Proceedings 48th International Colloquium on Automata, Languages, and Programming (ICALP) (LIPIcs, Vol. 198)*. Schloss Dagstuhl - Leibniz-Zentrum für Informatik, 104:1–104:20. https://doi.org/10.4230/LIPIcs.ICALP.2021.104

[30] Björn Scheuermann and Martin Mauve. 2007. Near-Optimal Compression of Probabilistic Counting Sketches for Networking Applications. In *Proceedings of the 4th International Workshop on Foundations of Mobile Computing (DIALM-POMC)*.

[31] The Apache Foundation. 2019. Apache DataSketches: A software library of stochastic streaming algorithms. https://datasketches.apache.org/. (2019). https://datasketches.apache.org/

[32] Daniel Ting. 2014. Streamed approximate counting of distinct elements: beating optimal batch methods. In *Proceedings 20th ACM Conference on Knowledge Discovery and Data Mining (KDD)*. 442–451. https://doi.org/10.1145/2623330.2623669

[33] A. W. van der Vaart. 1998. *Asymptotic Statistics.* Cambridge University Press. https://doi.org/10.1017/CBO9780511802256

[34] Qingjun Xiao, Shigang Chen, You Zhou, and Junzhou Luo. 2020. Estimating Cardinality for Arbitrarily Large Data Stream With Improved Memory Efficiency. *IEEE/ACM Trans. Netw.* 28, 2 (2020), 433–446. https://doi.org/10.1109/TNET.2020.2970860

A PROOFS FROM SECTION 3

A.1 Proof of Theorem 1

Theorem 1. *If $A_{\lambda;\tau}$ is τ-scale-invariant, then $\hat{\lambda}^*_{\tau;m} = \left(\frac{1}{m}\sum_{i=1}^{m} A^{(i)}_{\lambda;\tau}\right)^{-\tau^{-1}}$ is a scale-invariant estimator for λ.*

Proof. By default, $\hat{\lambda}^*_{\tau;m}$ is the estimator at cardinality λ. When needed, we use $\hat{\lambda}^*_{\tau;m}[\lambda']$ to indicate that it is being evaluated on a sketch with cardinality λ'. By the τ-scale-invariance of $A_{\lambda;\tau}$, we have $A_{\lambda;\tau} \sim \lambda^{-\tau} A_{1;\tau}$. Thus

$$\hat{\lambda}^*_{\tau;m}[\lambda] \sim \left(\frac{1}{m}\sum_{i=1}^{m} \lambda^{-\tau} A^{(i)}_{1;\tau}\right)^{-\tau^{-1}} = \left(\frac{1}{m}\lambda^{-\tau}\sum_{i=1}^{m} A^{(i)}_{1;\tau}\right)^{-\tau^{-1}} = \lambda \cdot \hat{\lambda}^*_{\tau;m}[1].$$

□

A.2 Proof of Theorem 2

Theorem 2. *If $A_{1;\tau}$ is τ-scale-invariant with finite variance, we have for any $\lambda > 0$,*

(1) $\lim_{m\to\infty} \mathbb{E}\hat{\lambda}^*_{\tau;m} = \lambda(\mathbb{E}A_{1;\tau})^{-\tau^{-1}}$.

(2) $\lim_{m\to\infty} m\lambda^{-2}\mathbb{V}(\hat{\lambda}^*_{\tau;m}) = \tau^{-2}(\mathbb{E}A_{1;\tau})^{-2\tau^{-1}-2}\mathbb{V}(A_{1;\tau})$.

(3) *For any $\lambda > 0$, the normalized estimator $\hat{\lambda}_{\tau;m} = (\mathbb{E}A_{1;\tau})^{\tau^{-1}}\hat{\lambda}^*_{\tau;m}$ is asymptotically unbiased and has limit relative variance $\lim_{m\to\infty} m\lambda^{-2}\mathbb{V}(\hat{\lambda}_{\tau;m}) = \tau^{-2}\left(\mathbb{E}A_{1;\tau}\right)^{-2}\mathbb{V}(A_{1;\tau})$.*

Proof. By scale-invariance, it suffices to consider the case $\lambda = 1$. Let $X = A_{1;\tau}$ and $Y_m = \frac{1}{m}\sum_{i=1}^m A^{(i)}_{1;\tau}$ be the mean of m copies of X. Define $f(x) = x^{-\tau^{-1}}$. Then $\hat{\lambda}^*_{\tau;m} = f(Y_m)$. Since we consider the case as $m \to \infty$, by the central limit theorem, Y_m is asymptotically normal around $\mathbb{E}X$. With high probability we have $Y_m \in (\mathbb{E}X - \frac{\log m}{\sqrt{m}}, \mathbb{E}X + \frac{\log m}{\sqrt{m}})$. Consider the first order approximation in this small neighborhood.

$$\begin{aligned} f(x) &= f(\mathbb{E}X) + (x - \mathbb{E}X)f'(\mathbb{E}X) + O((x-\mathbb{E}X)^2) \\ &= (\mathbb{E}X)^{-\tau^{-1}} - (x-\mathbb{E}X)\tau^{-1}(\mathbb{E}X)^{-\tau^{-1}-1} + O((x-\mathbb{E}X)^2). \end{aligned} \tag{1}$$

Note that $\mathbb{E}Y_m = \mathbb{E}X$ and $\mathbb{V}(Y_m) = \frac{1}{m}\mathbb{V}(X) = O(\frac{1}{m})$. Then we have

$$\begin{aligned} \mathbb{E}f(Y_m) &= (\mathbb{E}X)^{-\tau^{-1}} - (\mathbb{E}Y_m - \mathbb{E}X)\tau^{-1}(\mathbb{E}X)^{-\tau^{-1}-1} + O(\mathbb{V}(Y_m)) \\ &= (\mathbb{E}X)^{-\tau^{-1}} + O(\tfrac{1}{m}). \end{aligned} \tag{2}$$

Turning now to the variance, by (1) and (2)

$$\begin{aligned} \mathbb{V}(f(Y_m)) &= \mathbb{E}\left(f(Y_m) - \mathbb{E}f(Y_m)\right)^2 \\ &= \mathbb{E}\left((Y_m - \mathbb{E}X)\tau^{-1}(\mathbb{E}X)^{-\tau^{-1}-1} + O(\tfrac{1}{m})\right)^2 \\ &= \mathbb{V}(Y_m)\tau^{-2}(\mathbb{E}X)^{-2\tau^{-1}-2} + O(\tfrac{1}{m^2}), \end{aligned}$$

where we note that $\mathbb{E}(Y_m - \mathbb{E}X)^2 = \mathbb{V}(Y_m)$ and $\mathbb{E}(Y_m - \mathbb{E}X) = 0$. As $\mathbb{V}(Y_m) = \frac{1}{m}\mathbb{V}(X)$, this implies that the normalized variance is

$$m\mathbb{V}(f(Y_m)) = \mathbb{V}(X)\tau^{-2}(\mathbb{E}X)^{-2\tau^{-1}-2} + O(\tfrac{1}{m}). \tag{3}$$

We can now obtain a strictly unbiased estimator $(\mathbb{E}\hat{\lambda}^*_{\tau;m}[1])^{-1}\hat{\lambda}^*_{\tau;m}$, where $\hat{\lambda}^*_{\tau;m}[1]$ is the output of the estimator at cardinality (density) 1. We do not know precisely what $\mathbb{E}\hat{\lambda}^*_{\tau;m}[1]$ is, but $\lim_{m\to\infty}\mathbb{E}\hat{\lambda}^*_{\tau;m}[1] = (\mathbb{E}A_{1;\tau})^{-\tau^{-1}}$, so $\hat{\lambda}_{\tau;m} = (\mathbb{E}A_{1;\tau})^{\tau^{-1}}\hat{\lambda}^*_{\tau;m}$ is asymptotically unbiased, establishing Part (1). Part (2) follows from Part (1) and Eqn (3). Finally, observe that $\mathbb{V}(\hat{\lambda}_{\tau;m}) = (\mathbb{E}A_{1;\tau})^{2\tau^{-1}}\mathbb{V}(\hat{\lambda}^*_{\tau;m})$. Since $\lim_{m\to\infty} m\lambda^{-2}\mathbb{V}(\hat{\lambda}^*_{\tau;m}) = \tau^{-2}\left(\mathbb{E}A_{1;\tau}\right)^{-2\tau^{-1}-2}\mathbb{V}(A_{1;\tau})$, we have

$$\lim_{m\to\infty} m\lambda^{-2}\mathbb{V}(\hat{\lambda}_{\tau;m}) = \tau^{-2}\left(\mathbb{E}A_{1;\tau}\right)^{-2}\mathbb{V}(A_{1;\tau}),$$

proving Part (3). □

B GENERALIZED REMAINING AREA FOR LOGLOG AND CURTAIN SKETCHES

In a LogLog sketch we store the index of the highest cell in each of the m subsketches (columns in Figure 1) that has been hit by a dart. Every cell at or below the highest hit cell is implicitly *occupied*, regardless of whether it has been hit. A Curtain(d) sketch is like a LogLog sketch, except that we explicitly maintain, for each subsketch, the hit/unhit status of the d cells below the highest hit cell; those at least $d+1$ cells below the highest hit cell are implicitly occupied. In other words, the cell with index i in its subsketch is *unoccupied* if and only if *both*

(1) it has not been hit by a dart, and

(2) cells $i+d+1, i+d+2, \ldots$ in the same subsketch have all not been hit by a dart.

Let us focus on one subsketch with offset R. To simplify notation, let $X(t)$ be a "fresh" binary random variable for each t such that

$$X(t) = \begin{cases} 1, & \text{with probability } e^{-t} \\ 0, & \text{with probability } 1 - e^{-t}. \end{cases}$$

Note that the height of the ith cell is $2^{-(i+R)}$. Thus $X_i = X(2^{-(i+R)}\lambda/m)$ is the emptiness indicator of the ith cell at cardinality λ. The τ-GRA is then defined as

$$A_{\lambda;\tau} = \sum_{i=-\infty}^{\infty} X_i \mathbb{1}\left\{X_j = 1, \forall j > i+d\right\} 2^{-\tau(i+R)}.$$

We now prove that $A_{\lambda;\tau}$ is τ-scale-invariant.

Lemma 1. *For any $\tau > 0$, $A_{\lambda;\tau}$ is a τ-scale-invariant estimator for $\lambda^{-\tau}$.*

Proof. We need to prove that, for any $\tau, \lambda > 0$, $A_{\lambda;\tau} \sim \lambda^{-\tau}A_{1;\tau}$. Now note that,

$$A_{\lambda;\tau} = \sum_{i=-\infty}^{\infty} X(2^{-(i+R)}\lambda/m)\, \mathbb{1}\left\{X(2^{-(j+R)}\lambda/m) = 1, \forall j > i+d\right\} 2^{-\tau(i+R)}.$$

Since R is uniform over $[0,1)$ and we are summing over $\mathbb{Z}$, this sum is invariant under shifts, e.g., by $\log_2\lambda$. Continuing,

$$\begin{aligned} A_{\lambda;\tau} &= \lambda^{-\tau}\sum_{i=-\infty}^{\infty} X(2^{-(i+R)}\lambda/m)\, \mathbb{1}\left\{X(2^{-(j+R)}\lambda/m) = 1, \forall j > i+d\right\} (\lambda 2^{-(i+R)})^{\tau} \\ &\sim \lambda^{-\tau}\sum_{i=-\infty}^{\infty} X(2^{-(i+R)}/m)\, \mathbb{1}\left\{X(2^{-(j+R)}/m) = 1, \forall j > i+d\right\} (2^{-(i+R)})^{\tau} \\ &= \lambda^{-\tau}A_{1;\tau}. \end{aligned}$$

□

Recall that Γ is the continuous extension of the factorial function, with $\Gamma(n+1) = n!$ when $n \in \mathbb{N}$. Its integral form is $\Gamma(z) = \int_0^\infty u^{z-1}e^{-u}du$.

Proposition 1. *For any $\tau > 0$,*

$$m^{-\tau}\,\mathbb{E}A_{1;\tau} = \frac{1}{\log 2}\frac{1}{(2^{-d}+1)^\tau}\Gamma(\tau), \quad \textit{and}$$

$$m^{-2\tau}\,\mathbb{E}A_{1;\tau}^2 = (2^{-d}+1)^{-2\tau}(\log 2)^{-1}\Gamma(2\tau)(1+2\frac{2^{-\tau(d+1)}}{1-2^{-\tau}}) + 2\sum_{h=1}^{d}(2^{-d}+1+2^{-h})^{-2\tau}2^{-\tau h}(\log 2)^{-1}\Gamma(2\tau).$$

Proof. Note that R is picked beforehand. Conditioning on a certain value of R, X_i and X_j are independent if $i \neq j$.

$$\begin{aligned}\mathbb{E}A_{\lambda,\tau} &= \int_0^1 \sum_{i=-\infty}^{\infty} \mathbb{E}X(2^{-(i+r)}\lambda/m) \\ &\qquad \mathbb{E}\mathbb{1}\left\{X(2^{-(j+r)}\lambda/m) = 1, \forall j > i+d\right\} 2^{-\tau(i+r)}\,dr \\ &= \int_0^1 \sum_{i=-\infty}^{\infty} e^{-\frac{\lambda}{m}2^{-(i+r)}(2^{-d}+1)}(2^{-(i+r)})^\tau\,dr \\ &= \int_{-\infty}^{\infty} e^{-\frac{\lambda}{m}2^{-x}(2^{-d}+1)}(2^{-x})^\tau\,dx \\ &= \left(\frac{m}{\lambda(2^{-d}+1)}\right)^\tau \int_{-\infty}^{\infty} e^{-\frac{\lambda}{m}2^{-x}(2^{-d}+1)}(\frac{\lambda}{m}2^{-x}(2^{-d}+1))^\tau\,dx\end{aligned}$$

shifting x and we have

$$= \left(\frac{m}{\lambda(2^{-d}+1)}\right)^\tau \int_{-\infty}^{\infty} e^{-2^{-x}}(2^{-x})^\tau\,dx$$

setting $y = x\log 2$

$$\begin{aligned}&= \frac{1}{\log 2}\left(\frac{m}{\lambda(2^{-d}+1)}\right)^\tau \int_{-\infty}^{\infty} e^{-e^{-y}}(e^{-y})^\tau\,dy \\ &= \frac{1}{\log 2}\left(\frac{m}{\lambda(2^{-d}+1)}\right)^\tau \Gamma(\tau).\end{aligned}$$

Note that by setting $z = e^{-y}$, $\int_{-\infty}^{\infty} e^{-e^{-y}}(e^{-y})^\tau\,dy = \int_0^\infty e^{-z}z^{\tau-1}\,dz = \Gamma(\tau)$. This is generalized in the following lemma.

Lemma 2. *Let $a, b, q, \tau > 0$. Then*

$$\int_{-\infty}^{\infty} e^{-aq^{-x}}(bq^{-x})^\tau\,dx = a^{-\tau}b^\tau(\log q)^{-1}\Gamma(\tau).$$

Proof. Set $y = (x - \log a)\log q$. □

Now to understand the second moment, we consider the expectations of the product of all the pairs of the terms in the sum $A_{\lambda;\tau} = \sum_{i=-\infty}^{\infty} X_i\mathbb{1}\left\{X_j = 1, \forall j > i+d\right\}2^{-\tau(i+R)}$. Consider pair $i \le j$ where $j - i = h$.

- If $h = 0$, since X_is are indicators, thus
$$\begin{aligned}&\mathbb{E}(X_i\mathbb{1}\{X_k = 1, \forall k > i+d\})^2 \\ &= \mathbb{E}(X_i\mathbb{1}\{X_k = 1, \forall k > i+d\}) \\ &= e^{-\frac{\lambda}{m}2^{-(i+r)}(2^{-d}+1)}.\end{aligned}$$
- If $h \in [1, d]$, then
$$\begin{aligned}&\mathbb{E}(X_iX_j\mathbb{1}\{X_k = 1, \forall k > i+d\}\,\mathbb{1}\{X_t = 1, \forall t > j+d\}) \\ &= \mathbb{E}(X_iX_j\mathbb{1}\{X_k = 1, \forall k > i+d\}) \\ &= e^{-\frac{\lambda}{m}2^{-(i+r)}(2^{-d}+2^{-h}+1)}.\end{aligned}$$
- If $h \in (d, \infty)$, then
$$\begin{aligned}&\mathbb{E}(X_iX_j\mathbb{1}\{X_k = 1, \forall k > i+d\}\,\mathbb{1}\{X_t = 1, \forall t > j+d\}) \\ &= \mathbb{E}(X_i\mathbb{1}\{X_k = 1, \forall k > i+d\}) \\ &= e^{-\frac{\lambda}{m}2^{-(i+r)}(2^{-d}+1)}.\end{aligned}$$

Combining these three cases, use linearity of expectation, we have

$$\begin{aligned}\mathbb{E}A_{\lambda,\tau}^2 = \int_0^1 \Bigg(&\sum_{i=-\infty}^{\infty} e^{-\frac{\lambda}{m}2^{-(i+r)}(2^{-d}+1)}\left(2^{-(i+r)}\right)^{2\tau} \\ &+2\sum_{i=-\infty}^{\infty}\sum_{h=1}^{d} e^{-\frac{\lambda}{m}2^{-(i+r)}(2^{-d}+1+2^{-h})}\left(2^{-(i+r)}q^{-(i+h+r)}\right)^\tau \\ &+2\sum_{i=-\infty}^{\infty}\sum_{h=d+1}^{\infty} e^{-\frac{\lambda}{m}2^{-(i+r)}(2^{-d}+1)}\left(2^{-(i+r)}2^{-(i+h+r)}\right)^\tau\Bigg)\,dr \\ = \int_{-\infty}^{\infty} &e^{-\frac{\lambda}{m}2^{-x}(2^{-d}+1)}(2^{-x})^{2\tau}\,dx \\ &+2\sum_{h=1}^{d}\int_{-\infty}^{\infty} e^{-\frac{\lambda}{m}2^{-x}(2^{-d}+1+2^{-h})}\left(2^{-x}2^{-(x+h)}\right)^\tau\,dx \\ &+2\sum_{h=d+1}^{\infty}\int_{-\infty}^{\infty} e^{-\frac{\lambda}{m}2^{-x}(2^{-d}+1)}\left(2^{-x}2^{-(x+h)}\right)^\tau\,dx\end{aligned}$$

Apply Lemma 2 to each term and we have

$$\begin{aligned}&= \left(\frac{\lambda}{m}(2^{-d}+1)\right)^{-2\tau}(\log 2)^{-1}\Gamma(2\tau) \\ &\quad+2\sum_{h=1}^{d}\left(\frac{\lambda}{m}(2^{-d}+1+2^{-h})\right)^{-2\tau}2^{-\tau h}(\log 2)^{-1}\Gamma(2\tau) \\ &\quad+2\sum_{h=d+1}^{\infty}\left(\frac{\lambda}{m}(2^{-d}+1)\right)^{-2\tau}2^{-\tau h}(\log 2)^{-1}\Gamma(2\tau) \\ &= \left(\frac{\lambda}{m}(2^{-d}+1)\right)^{-2\tau}(\log 2)^{-1}\Gamma(2\tau)\left(1+2\frac{2^{-\tau(d+1)}}{1-2^{-\tau}}\right) \\ &\quad+2\sum_{h=1}^{d}\left(\frac{\lambda}{m}(2^{-d}+1+2^{-h})\right)^{-2\tau}2^{-\tau h}(\log 2)^{-1}\Gamma(2\tau).\end{aligned}$$

□

Setting $d = 0$, this leads to, for example, the following *generalized HyperLogLog theorem.*

Theorem 3. *[τ-GRA for the LogLog sketch] Let the offset vector $(R_i) \in [0,1)^m$ be selected uniformly at random. Let $X_\lambda^{(i)}$ be the integer index of the highest one in the ith subsketch after λ insertions. Then for any $\tau > 0$,*

$$\hat{\lambda}_{\tau;m} = m\left(\Gamma(\tau)\frac{1-2^{-\tau}}{\log 2}\right)^{\tau^{-1}}\left(\frac{1}{m}\sum_{i=1}^{m}2^{-\tau(R_i+X_\lambda^{(i)})}\right)^{-\tau^{-1}}$$

is a scale-invariant estimator for λ that is asymptotically unbiased. The asymptotic normalized relative variance is

$$\lim_{m\to\infty} m\lambda^{-2}\mathbb{V}(\hat{\lambda}_{\tau;m}) = \tau^{-2}\left(\frac{\Gamma(2\tau)\log 2}{\Gamma(\tau)^2}\cdot\frac{1+2^{-\tau}}{1-2^{-\tau}}-1\right).$$

Proof. Note that for a subsketch with highest occupied X_λ, the τ-GRA is calculated as

$$A_{\lambda;\tau} = \sum_{i>X_\lambda} 2^{-\tau(i+R)} = 2^{-\tau(X_\lambda+R)}\frac{1}{2^\tau-1},$$

where R is the random offset. Thus to conform with the style of HyperLogLog, we choose $A'_{\lambda;\tau} = (2^\tau-1)A_{\lambda;\tau}$ as the τ-GRA. By Prop 1 with $d=0$, we have the normalized first and second moments as

$$m^{-\tau}\mathbb{E}A' = \frac{1-2^{-\tau}}{\log 2}\Gamma(\tau), \quad \text{and} \quad m^{-2\tau}\mathbb{E}A'^2 = \frac{\Gamma(2\tau)(1-2^{-2\tau})}{\log 2}$$

The variance is implied from the first two moments. Apply Theorem 2 and we get the result. □

Remark 1. The celebrated estimator $\hat{\lambda}_{\text{FFGM}}$ of HyperLogLog corresponds to $\tau=1$. Inserting $\tau=1$ to the variance formula, we have $\Gamma(2)=\Gamma(1)=1$ and the leading constant of the variance is $3\log 2-1\approx 1.07944$. The bias term at $\tau=1$ is $\frac{1}{2\log 2}$, which match the constants from Flajolet et al. [15] as $m\to\infty$.

Remark 2. Note that for any $x_1, x_2, \ldots, x_m > 0$,

$$\lim_{\tau\to 0}\left(\frac{1}{m}\sum_{i=1}^m x_1^{-\tau}\right)^{-\tau^{-1}} = \left(\prod_{i=1}^m x_i\right)^{m^{-1}},$$

i.e., the τ-mean converges towards the geometric mean as $\tau\to 0$. In other words, Durand and Flajolet's estimator $\hat{\lambda}_{\text{DF}}$ for LogLog corresponds to 0-GRA. We have the normalized relative variance[7]

$$\lim_{\tau\to 0}\tau^{-2}\left(\frac{\Gamma(2\tau)\log 2}{\Gamma(\tau)^2}\cdot\frac{1+2^{-\tau}}{1-2^{-\tau}}-1\right) = \frac{2\pi^2+\log^2 2}{12}\approx 1.68497,$$

which matches the limiting constant calculated by Durand and Flajolet [10, 11].

See Figure 2 for a visualization of the relative variance of the τ-GRA estimators for the LogLog sketch. By numerical optimization, the minimal variance 1.07507 is obtained at $\tau^*=0.889897$. This comes quite close to the Cramér-Rao lower bound for LogLog sketches, which Pettie and Wang [28] computed to be $\frac{\log 2}{\pi^2/6-1}\approx 1.07475$.

For demonstration, one can insert $d=1,2$ to Proposition 1 and apply Theorem 2 to obtain the following theorem.

Theorem 4. *[τ-GRA for the Curtain sketch with $d=1,2$] Let the offset vector $(R_i)\in[0,1)^m$ be selected uniformly at random. Let $X_\lambda^{(i)}$ be the integer index of the highest cell hit in the ith subsketch after λ insertions. Let $A^{(i)}$ and $B^{(i)}$ be indicators for whether cells $X_\lambda^{(i)}-1$ and $X_\lambda^{(i)}-2$ have been hit (and are occupied) in the ith subsketch.*

- *When $d=1$, for any $\tau>0$,*

$$\hat{\lambda}_{\tau;m} = \frac{2}{3}m\left(\Gamma(\tau)\frac{1}{\log 2}\right)^{\tau^{-1}}\left(\frac{1}{m}\sum_{i=1}^m\left(\frac{1}{2^\tau-1}2^{-\tau(R_i+X_\lambda^{(i)})}+(1-A^{(i)})2^{-\tau(R_i+X_\lambda^{(i)}-1)}\right)\right)^{-\tau^{-1}}$$

is a scale-invariant estimator for λ that is asymptotically unbiased. The asymptotic normalized relative variance is

$$\lim_{m\to\infty} m\lambda^{-2}\mathbb{V}(\hat{\lambda}_{\tau;m}) = \frac{1}{\tau^2\Gamma(\tau)^2}\left((\log 2)\Gamma(2\tau)(1+2\frac{2^{-2\tau}}{1-2^{-\tau}}+2\cdot 3^{2\tau}2^{-5\tau})-\Gamma(\tau)^2\right).$$

- *When $d=2$, for any $\tau>0$,*

$$\hat{\lambda}_{\tau;m} = \frac{4}{5}m\left(\Gamma(\tau)\frac{1}{\log 2}\right)^{\tau^{-1}}\left(\frac{1}{m}\sum_{i=1}^m\left(\frac{1}{2^\tau-1}2^{-\tau(R_i+X_\lambda^{(i)})}+(1-A^{(i)})2^{-\tau(R_i+X_\lambda^{(i)}-1)}+(1-B^{(i)})2^{-\tau(R_i+X_\lambda^{(i)}-2)}\right)\right)^{-\tau^{-1}}$$

is a scale-invariant estimator for λ that is asymptotically unbiased. The asymptotic normalized relative variance is

$$\lim_{m\to\infty} m\lambda^{-2}\mathbb{V}(\hat{\lambda}_{\tau;m}) = \frac{1}{\tau^2\Gamma(\tau)^2}\left((\log 2)\Gamma(2\tau)\left(1+2\frac{2^{-3\tau}}{1-2^{-\tau}}+2(7/5)^{-2\tau}2^{-\tau}+2(6/5)^{-2\tau}2^{-2\tau}\right)-\Gamma(\tau)^2\right).$$

C GENERALIZED REMAINING AREA FOR THE PCSA SKETCH

Consider a PCSA sketch with m subsketches. Due to Poissonization, the sketch consists of a set of independent indicator variables corresponding to whether each cell has been hit by at least one dart. As the section above, let $X(t)$ be a "fresh" binary random variable such that

$$X(t) = \begin{cases} 1, & \text{with probability } e^{-t} \\ 0, & \text{with probability } 1-e^{-t}. \end{cases}$$

Consider one subsketch with offset R. Cell i has height $2^{-(i+R)}$ and width $1/m$. At cardinality λ the number of points in the cell is Poisson$(m^{-1}\lambda 2^{-(i+R)})$. Thus the bit vector representing this subsketch distributes identically with $(X(m^{-1}\lambda 2^{-(i+R)}))_{i\in\mathbb{Z}}$. The τ-generalized remaining area for the PCSA sketch is then defined as follows.

$$A_{\lambda;\tau} = \sum_{i\in\mathbb{Z}} X(m^{-1}\lambda 2^{-(i+R)})2^{-(i+R)\tau}.$$

Lemma 3. *For any $\tau>0$, $A_{\lambda;\tau}$ is a τ-scale-invariant estimator for $\lambda^{-\tau}$.*

Proof. We need to prove that for any $\lambda>0$, $A_{\lambda;\tau}\sim\lambda^{-\tau}A_{1;\tau}$. Note that

$$\begin{aligned} A_{\lambda;\tau} &= \sum_{i\in\mathbb{Z}} X(m^{-1}\lambda 2^{-(i+R)})2^{-\tau(i+R)} \\ &= \lambda^{-\tau}\sum_{i\in\mathbb{Z}} X(m^{-1}2^{-(i+R-\log_2\lambda)})2^{-\tau(i+R-\log_2\lambda)} \end{aligned}$$

[7] This limit calculation is done in the algebraic system *Mathematica*.

Note that because R is uniform over $[0,1)$ and we are summing over $\mathbb{Z}$, this sum is invariant under shifts, e.g., by $\log_2 \lambda$. Continuing,

$$A_{\lambda;\tau} \sim \lambda^{-\tau} \sum_{i\in\mathbb{Z}} \mathbb{1}\left\{X(m^{-1}2^{-(i+R)}) = 0\right\} 2^{-\tau(i+R)} = \lambda^{-\tau} A_{1;\tau}.$$

□

In contrast to our smoothing of (Hyper)LogLog, it actually *does* matter that we use the uniform offset vector $\mathbf{R} = (0, 1/m, \ldots, (m-1)/m)$ rather than random offsets. Random offsets would introduce subtle correlations between cells in the same column, and increase the variance by some tiny constant. Uniform offsets have the property that there is a cell of size $2^{-i/m}$ for every $i \in \mathbb{Z}$, so the conceptual organization of cells into columns is no longer relevant.

Proposition 2. *Let $A^{(1)}_{1;\tau}, A^{(2)}_{1;\tau}, \ldots, A^{(m)}_{1;\tau}$ be the τ-GRA of m subsketches with uniform offsetting. For $\tau > 0$,*

$$\lim_{m\to\infty} m^{-1-\tau} \sum_{i=1}^{m} \mathbb{E}(A^{(i)}_{1;\tau}) = \frac{\Gamma(\tau)}{\log 2}, \quad \text{and}$$
$$\lim_{m\to\infty} m^{-1-2\tau} \sum_{i=1}^{m} \mathbb{V}(A^{(i)}_{1;\tau}) = \frac{(1-2^{-2\tau})\Gamma(2\tau)}{\log 2}.$$

Proof. First note the following identity. Assume $q > 1, \tau > 0, \lambda > 0$. Then

$$\psi(\lambda, q, \tau) = \int_{-\infty}^{\infty} e^{-q^{-x}\lambda} q^{-\tau x}\, dx = \int_0^{\infty} e^{-t} (t/\lambda)^{\tau} (t \log q)^{-1}\, dt$$
$$= \frac{1}{\log q} \lambda^{-\tau} \Gamma(\tau).$$

Here $t = q^{-x}$. After uniform offsetting, a PCSA sketch with m subsketches have cells of size $2^{-i/m}$ for all $i \in \mathbb{Z}$. Thus we we have

$$\lim_{m\to\infty} m^{-1-\tau} \sum_{i=1}^{m} \mathbb{E}\left(A^{(i)}_{1;\tau}\right)$$
$$= \lim_{m\to\infty} m^{-1} \sum_{i\in\mathbb{Z}} \mathbb{E}\left(X(m^{-1}2^{-i/m})(m^{-1}2^{-(i/m)})^{\tau}\right)$$

Setting $h(t) = \mathbb{E}\left(X(2^{-t})(2^{-t})^{\tau}\right) = e^{-2^{-t}} 2^{-\tau t}$, the sum becomes $m^{-1} \sum_{i\in\mathbb{Z}} h(i/m + \log_2 m)$. Since we are summing over $\mathbb{Z}$, the shift $\log_2 m$ in the argument affects the sum vanishingly as $m \to \infty$. Thus $\lim_{m\to\infty} m^{-1} \sum_{i\in\mathbb{Z}} h(i/m + \log_2 m) = \lim_{m\to\infty} m^{-1} \sum_{i\in\mathbb{Z}} h(i/m) = \int_{-\infty}^{\infty} h(t)\, dt$. Thus,

$$\lim_{m\to\infty} m^{-1-\tau} \sum_{i=1}^{m} \mathbb{E}\left(A^{(i)}_{1;\tau}\right)$$
$$= \int_{-\infty}^{\infty} e^{-2^{-t}} 2^{-\tau t}\, dt = \psi(1,2,\tau) = \frac{\Gamma(\tau)}{\log 2}.$$

Note that by Poissonization, cells are independent and thus all co-variances are zero.

$$\lim_{m\to\infty} m^{-1-2\tau} \sum_{i=1}^{m} \mathbb{V}(A^{(i)}_{1;\tau})$$
$$= \lim_{m\to\infty} m^{-1} \sum_{i\in\mathbb{Z}} \mathbb{V}\left(X(m^{-1}2^{-i/m})(m^{-1}2^{-i/m})^{\tau}\right)$$

by the same limiting argument laid out above, this is equal to

$$= \int_{-\infty}^{\infty} \mathbb{V}(X(2^{-t})2^{-\tau t})\, dt$$
$$= \int_{-\infty}^{\infty} e^{-2^{-t}} 2^{-2\tau t} - e^{-2\cdot 2^{-t}} 2^{-2\tau t}\, dt$$
$$= \psi(1,2,2\tau) - \psi(2,2,2\tau) = \frac{\Gamma(2\tau)}{\log 2}(1-2^{-2\tau}).$$

□

Theorem 5. *[τ-GRA for the PCSA sketch] Let $A^{(i)}_{\lambda;\tau}$ be the τ-generalized remaining area of the ith subsketch with uniform offsetting, and $A = \sum_{i=1}^{m} A^{(i)}_{\lambda;\tau}$ be the τ-GRA. Then for any $\tau > 0$, $\hat{\lambda}_{\tau;m} = m\left(\frac{\Gamma(\tau)}{\log 2}\right)^{\tau^{-1}} \left(\frac{A}{m}\right)^{-\tau^{-1}}$ is a scale-invariant estimator for λ that is asymptotically unbiased. The asymptotic normalized relative variance is*

$$\lim_{m\to\infty} m\lambda^{-2}\mathbb{V}(\hat{\lambda}_{\tau;m}) = \frac{(1-2^{-2\tau})\Gamma(2\tau)\log 2}{\tau^2\Gamma(\tau)^2}.$$

Proof. This follows directly from Theorem 2 and Proposition 2. □

Remark 3. The remaining area estimator (1-GRA) has normalized relative variance $\frac{(1-2^{-2})\Gamma(2)}{1^2\Gamma(1)^2}\log 2 = \frac{3}{4}\log 2 \approx 0.51986$, which is better than Flajolet and Martin's original "first zero" estimator $\hat{\lambda}_{\mathrm{FM}}$.

Remark 4. As τ goes to 0, $\hat{\lambda}_{\tau;m}$ is essentially counting the number of free cells (0s in the sketch), which corresponds to Lang's [24] "coupon collector" estimator $\hat{\lambda}_{\mathrm{Lang}}$ that counts occupied cells (1s in the sketch). The limiting variance of this estimator is[8]

$$\lim_{\tau\to 0} \frac{(1-2^{-2\tau})\Gamma(2\tau)\log 2}{\tau^2\Gamma(\tau)^2} = \log^2 2 \approx 0.480453,$$

which confirms Lang's [24] back-of-the-envelope calculation that it should be $\log^2 2$.

See Figure 2 for a visualization of the relative variance of the τ-GRA estimators for the PCSA sketch. By numerical optimization, the minimal variance 0.435532 is obtained at $\tau^* = 0.343557$. This comes very close to the Cramér-Rao lower bound of $\frac{\pi^2}{6\log 2} \approx 0.42138$ for PCSA sketches, as computed in [28].

[8] This calculation is done in the algebraic system *Mathematica*.