Proceedings of Machine Learning Research vol 201:1–48, 2023 34th International Conference on Algorithmic Learning Theory

Testing Tail Weight of a Distribution Via Hazard Rate

Maryam Aliakbarpour MARYAMA@ALUM.MIT.EDU
Boston University / Northeastern University, 805 Columbus Ave, Boston, MA 02120

Amartya Shankha Biswas AMARTYASHANKHA@GMAIL.COM
Massachusetts Institute of Technology, 32 Vassar St., Cambridge, MA 02139

Kavya Ravichandran RAVICHANDRAN.KAVYA@GMAIL.COM
Toyota Technological Institute at Chicago[*]*, 6045 S. Kenwood Ave., Chicago, IL 60637*

Ronitt Rubinfeld RONITT@CSAIL.MIT.EDU
Massachusetts Institute of Technology, 32 Vassar St., Cambridge, MA 02139

Editors: Shipra Agrawal and Francesco Orabona

Abstract

Understanding the shape of a distribution of data is of interest to people in a great variety of fields, as it may affect the types of algorithms used for that data. We study one such problem in the framework of *distribution property testing*, characterizing the number of samples required to to distinguish whether a distribution has a certain property or is far from having that property. In particular, given samples from a distribution, we seek to characterize the tail of the distribution, that is, understand how many elements appear infrequently. We develop an algorithm based on a careful bucketing scheme that distinguishes light-tailed distributions from non-light-tailed ones with respect to a definition based on the hazard rate, under natural smoothness and ordering assumptions. We bound the number of samples required for this test to succeed with high probability in terms of the parameters of the problem, showing that it is polynomial in these parameters. Further, we prove a hardness result that implies that this problem cannot be solved without any assumptions.

1. Introduction

Testing properties of a data distribution is a fundamental problem with applications in many scientific endeavors. The goal of distribution testing is to efficiently discern whether observed data confirm a hypothesized model or not. Such problems have been studied in asymptotic statistics for over a century (Pearson (1900); Neyman et al. (1933)), where the aim is to provide tests with vanishing error as the number of samples goes to infinity (i.e., asymptotically). In the past two decades, the field of *property testing of distributions* has sought to characterize what error can be achieved with finitely many samples (i.e., non-asymptotic setting) (Batu et al. (2000, 2013)). The objective of the problem here is to distinguish whether a distribution has some property (null hypothesis) or is far from having that property (alternative hypothesis) with high constant probability using as few samples as possible. Finite-sample guarantees have been given for testing a wide range of distribution properties including uniformity, monotonicity, low-error representation by a k-histogram function, support size estimation, and many others Batu et al. (2004); Diakonikolas et al. (2019); Canonne (2016, 2020); Alon et al. (2007); Acharya and Daskalakis (2015); Chan et al. (2014); Valiant (2011); Rubinfeld and Vasilyan; Aliakbarpour et al. (2019).

[*] Work predominantly done while author was a student at Massachusetts Institute of Technology

Much of the work in distribution property testing makes no assumptions on the distributions being tested and focuses on distributions over discrete domains. However, strong lower bounds have been given for many of these problems, involving dependence on the size of the support of the distribution. Thus, for continuous domains, some assumptions on the distributions are required for the problem to be tractable (e.g., Adamaszek et al. (2010)). We focus on this latter case.

We are interested in the shape of the "tail" of a distribution, that is, the behavior of the distribution as it moves away from the mean. In some distributions, the frequencies of elements far from the mean drop very quickly ("light-tailed"), while in others, the frequencies of large elements drop more slowly ("heavy-tailed"). Harchol-Balter (1999) has shown that the performances of policies for scheduling computing jobs vary dramatically depending on whether the distribution of the job workloads are heavy-tailed or light-tailed. The shape of the distribution has influenced the design and analysis of learning algorithms – e.g., the classification algorithm of Wang et al. (2017), the generalization bound of Feldman (2020), and the frequency estimation result of Hsu et al. (2018) (where the latter two are for specific subclasses of heavy-tailed distributions). In this work, we seek to characterize the tail of the distribution, specifically to decide whether it is substantially "heavy."

"Heavy-Tailed" distributions It is non-trivial to give a single unifying definition of heavy-tailed distributions, since the definition needs to accommodate a wide range of behaviors, including distributions whose tails fall off at irregular rates. We discuss the plethora of definitions found in the literature in Appendix A. Though not equivalent, these definitions are united by the fact that the point of reference is the exponential distribution, meaning that a distribution whose tail decays more slowly than the exponential is considered "heavy-tailed." [1] In this work, we adapt Klugman et al.'s definition of heavy-tailed that is based on the property that the hazard rate of a heavy-tailed distribution is decreasing Klugman et al. (2004); the characterization is of similar structure to other definitions, admits a clean description, and reflects the idea that heavy tails decay more slowly than exponential tails.

Our Setting We consider distributions over continuous and unbounded domains in the non-parametric setting, that is, not assuming that distributions belong to any specific class. We access such distributions via independent and identically-distributed samples.

In order to make the problem tractable, two kinds of technical assumptions are essential: smoothness and monotonicity. The first type of assumption, *smoothness*, is typical in learning theory and non-parametric statistics. Smoothness limits the behavior of the characteristic function of a distribution, protecting against adversarial behavior on small regions of a continuous domain not seen by a finite set of samples: without such assumptions, for any finite number of samples, one could construct two distributions that look the "same" when we draw a finite set of samples but differ on tiny intervals of the domain that are not detected by those samples. The second kind of assumption we require arises from the fact that distribution having a tail implicitly relies on the distribution decaying. Namely, we assume that the distributions we consider are monotone decreasing (or more broadly, unimodal). These conditions are discussed in detail in Section 2.

[1] Distributions with regularly varying tails represent a subset of distributions whose tails decay more slowly than exponential.

Our Contributions In this work, we begin by giving a parametrized definition of heavy-tailed distributions, based on an extension of the hazard rate definition from Klugman et al. (2004). [2] In our definition, a distribution is *light-tailed* if it has non-decreasing hazard rate throughout the domain. On the other hand, a heavy-tailed distribution must show a behavior "far" from light-tailed distributions at least in some interval of the domain:

Definition 1 (informal statement) *A distribution is called* (α, ρ)-Heavy-tailed *if the hazard rate decreases by at least rate* α *on a contiguous portion of the domain that contains at least* ρ *of the probability mass. If the hazard rate is non-decreasing, the distribution is called* light-tailed*.*

For the formal definition, see Definition 3. This parametrization allows for a fine-grained characterization of tail shape and might allow for more nuanced algorithm design. Further, we give a hardness result that shows that we cannot solve the problem in this domain without structural assumptions (Section F), justifying the need for a mild condition on the contiguousness of heavy-tailed regions.

With this definition in mind, we seek to design an efficient algorithm that, given finitely many samples from a distribution, determines whether they come from a light-tailed or (α, ρ)-Heavy-tailed distribution. The main result of our work is a theorem stating the number of samples (up to constant factors) that suffice to perform this task, presented here informally:

Theorem 2 (informal statement) *We can distinguish between light-tailed distributions and* (α, ρ)*-heavy-tailed ones with a number of samples that depends polynomially on smoothness parameters,* α*, and* ρ*.*

Finally, we run experiments on synthetic and real-world data to show the feasibility of our algorithm in practice. We also show that our algorithm outperforms a naive one in its ability to detect a subtly heavy-tailed distribution. Synthetic data experiments can be found in Section 6 and the rest in Appendix G and Appendix H.

Our Approach Our algorithm is based on a simple but fundamental observation about the rate of dropoff as it relates to the weight of a distribution's tail: in the tail of a heavy-tailed distribution, we would expect that the distance in the support required to accumulate a fixed amount of weight would not change too much since it drops more slowly, whereas in a light tailed distribution, this quantity is much more drastic. Based on this, we introduce a *proxy quantity* in Section 3.2 (which we will call S) that measures how long it takes the distribution to accumulate some amount of weight relative to how long it took previously, acting as a proxy for calculating the derivative of the hazard rate.

The main challenge here is that we cannot compute S from the samples directly since it is a function of the density function of the distribution. Therefore, we present a test statistic, called $\hat{S}$, to approximate S from samples. The algorithm makes use of a bucketing scheme that partitions the domain of the distribution into buckets (intervals) that contain equal probability mass. The algorithm then uses the lengths of these intervals to calculate $\hat{S}$ and compares it to a threshold that separates light-tailed distributions from heavy-tailed ones.

[2] The hazard rate of a distribution at a given point in the support is the value of the PDF divided by the amount of mass left in the tail from that point.

We determine the sample complexity for the algorithm in the defined setting by showing that if we draw “enough” samples, $\hat{S}$ is an accurate estimate of S (Theorem 9) and lies on the correct side of the threshold for any underlying distribution. In calculating $\hat{S}$, we incur two main sources of error: first, we approximate derivatives involved in computation of S by the discrete derivative/difference quotient; second, our algorithm uses order statistics from the samples to define the buckets, which introduces error in the estimation of the lengths of the buckets. Our assumption about smoothness allows us to analyze the former, and for the latter, we evaluate the concentration of order statistics.

It is worth noting that precise estimation of bucket endpoints from samples is challenging, since a small amount of probability mass could lie in a very large interval. Indeed, this is the most technically-interesting portion of the analysis and includes results on finite-sample concentration of order statistics, which, to the best of our knowledge, are novel. We discuss the details of the sample complexity and success probability of this result and argue the correctness of the algorithm in Section 3, Section 4, and Section E. To our knowledge, this is the first algorithm with finite sample guarantees to test the shape of the tail weight of a distribution with unbounded support.

Indeed, the novelty of our finite-sample guarantee is underscored by our result showing that without some structural assumptions, no finite sample algorithm can succeed. In particular, in Section 5, we develop two classes of distributions, one light-tailed and one that is heavy-tailed over a large but non-contiguous portion of the support, that cannot be distinguished with finitely many samples. The family of heavy-tailed distributions involves embedding hard instances into non-contiguous small parts of a light-tailed distribution, thereby “fooling” any finite-sample algorithm into classifying distributions from this class as light-tailed distributions.

Summary of our contributions:

- We give a novel parametrized definition for heavy-tailed distributions.
- We develop an algorithm for testing heavy-tailed distributions in the property testing setting which uses finitely many samples. We theoretically prove the correctness of our algorithm.
- As a byproduct of our result, we develop new results on the concentration of the ordered statistics drawn from an arbitrary distributions while using finitely many samples which may be of independent interest.
- We show intractability of this problem without assumptions by presenting two classes of distributions, one light-tailed and one heavy-tailed, that are indistinguishable by any finite-sample algorithm.
- We provide experimental results which confirm the performance of our algorithm.

Related Work Testing the tail weight of a distribution has been studied in the asymptotic theory of statistics by Bryson (1974), where the weight of the tail is characterized via a statistic referred to as *conditional mean exceedance*, and a proxy for it is used to distinguish between the families of Lomax and exponential distributions. Several works in the asymptotic theory literature address properties of distributions with monotone hazard rate Barlow et al. (1963) and algorithms that test whether a distribution has monotone hazard rate (a similar condition to the one used in this work), with guaranteed asymptotic convergence Hall et al. (2005); Gijbels and Heckman (2004). While these works operate in the asymptotic regime, we address the finite sample setting.

The mostly closely related work to ours comes from the theoretical computer science community. The task of distinguishing whether a *discrete* distribution has monotone hazard rate[3] (MHR), which characterizes "light-tailed" distributions, or is far in L_1 distance from such a distribution, has been considered in Acharya et al. (2015); Canonne et al. (2018a).

In these methods, sample complexity is dependent on the domain size; the method of Acharya et al. (2015) relies on a linear program to learn the weight associated with each discrete element. Thus, these methods do not easily translate to finite sample guarantees over continuous domains. Since the sample complexity of this algorithm scales with domain size, it is not finite in our setting. In contrast, we give a finite sample guarantee, but the result is incomparable due to the assumptions we make. For more details, see Appendix B.

2. Preliminaries

Notation We say that a distribution has Probability Density Function (PDF) $f(x)$ and Cumulative Density Function (CDF) $F(x)$. Some families of distributions we reference are: Lomax $\left(f(x) = \frac{\alpha}{\lambda}\left[1+\frac{x}{\lambda}\right]^{-(\alpha+1)}\right)$, exponential $\left(f(x) = \lambda e^{-\lambda x}\right)$, and half of a Gaussian $\left(f(x) = 2/\sigma\sqrt{2\pi})e^{-\frac{1}{2}\left(\frac{x}{\sigma}\right)^2}\right)$, all on $[0,\infty)$. For more details regarding them, see Appendix C. A set of n samples drawn from a distribution with PDF f is denoted $x_1, x_2, ..., x_n$. The i^{th} order statistic, the random variable representing the i^{th} smallest sample of this set of n samples, is denoted $X_{(i)}$. If we use m buckets, then the bucket endpoints in terms of the order statistic are $X_{\left(\frac{n}{m}\right)}, X_{\left(\frac{2n}{m}\right)} \ldots$.

Defining "Heavy-Tailed" While several definitions of "heavy-tailed" have been proposed, we consider definitions based on the hazard rate of the distribution. The hazard rate (HR) is defined in Su and Tang (2003) as $HR(x) = \frac{f(x)}{1-F(x)}$. The textbook of Klugman et al. (2004) defines a distribution with hazard rate that is a decreasing function of x to be "heavy-tailed." We consider this definition because it succinctly captures what many other definitions address obliquely – having a tail that decays more slowly than exponential. Thus, this definition captures the fact that in a heavy-tailed distribution, the probability mass at a given point relative to the tail decreases further into the distribution. This leads us to introduce the following parameterization of how heavy-tailed a distribution is.

Definition 3 *We say a distribution with PDF $f(x)$ and CDF $F(x)$ is (α, ρ)-*Heavy-Tailed *with $\alpha > 0, 0 < \rho < 1$ if the derivative of the hazard rate is negative with magnitude at least α on some interval $[x_1, x_2]$, where at least ρ of the probability mass lies in that interval. That is, if $H_f'(x) < -\alpha \ \forall x \in [x_1, x_2]$ s.t. $F(x_2) - F(x_1) \geq \rho$, we call $f(x)$ (α, ρ)-Heavy-Tailed.*

We use this definition in Claim 10, showing that our algorithm will detect any (α, ρ)-heavy-tailed distribution with a sample complexity dependent on both. The gap between light-tailed distributions and an (α, ρ)-heavy-tailed distribution is related directly to α, so as $\alpha \to 0$, the heavy-tailed distributions become harder to distinguish from light-tailed ones. Meanwhile, the parameter ρ considers what percent of the mass lies in a heavy-tailed region. A wide class of distributions of interest are (α, ρ)-heavy-tailed for some α, ρ. For example,

[3] Acharya et al. (2015); Canonne et al. (2018a) use "monotone hazard rate" to refer to monotone *increasing* hazard rate.

the general class of distributions with CDF $F(x) = 1 - e^{-\gamma x^m}$, where $0 < m < 1$ and PDF $f(x) = \lambda m x^{m-1} e^{-\lambda x^m}$ is $(m \cdot (1-m), 1 - e^{-1})$-heavy-tailed.

Consequent to the definition of heavy-tailed found in Klugman et al. (2004), we have the following definition for light-tailed. Exponential and half-Gaussian distributions are light tailed according to this definition.

Definition 4 *We say a distribution with PDF $f(x)$, CDF $F(x)$ is* light-tailed *if its hazard rate $\frac{f(x)}{1-F(x)}$ is non-decreasing, or equivalently, if the derivative of the hazard rate is non-negative.*

Defining "Well-Behaved" distributions As discussed earlier when describing our setting, we consider continuous distributions on $R_+ = [0, \infty)$ that are monotone decreasing, that is, $f(x_1) \geq f(x_2)$ when $x_1 < x_2$, so that there is a clear notion of the "tail"[4]. Likewise, we require boundedness and smoothness conditions to make the problem tractable but not trivial. Indeed, in Appendix F, we show a hardness result implying that without any regularity conditions, no finite sample algorithm can succeed at this problem. In the literature, it is common to assume Lipschitz density function. However, bounding the derivative of the density implies that functions that drop too quickly cannot be considered, significantly limiting the kinds of functions we could test. Thus, instead, we assume that the inverse CDF of distributions satisfy continuity and Lipschitz conditions. Lipschitzness assumptions are also needed due to the finite sample regime: with infinitely many samples, we could rely on continuity alone. We formalize and unify these constraints in the following definition, and in this paper we consider distributions satisfying this property.

Definition 5 *Let f be a distribution on the range $R_+ = [0, \infty)$ with CDF F. We say f is* well-behaved *if the following conditions hold:*

- *f is a non-increasing continuous function over R_+, and $f'(x)$ exists for every $x \in R_+$.*
- *For all x in the domain, $f(x)$ is bounded by a constant β.*
- *For every $x \in R_+$, $F'(x)$ exists, and it is equal to $f(x)$.*
- *The first and second derivatives of $F^{-1}(x)$ are Lipschitz with parameters B_1, B_2, respectively, on the domain until the very end (where it cannot be Lipschitz anyway[5]). We will call the domain over which it is Lipschitz $[0, 1-\zeta]$ and explain what ζ must be in Section 4.*

The method we present in this paper relies on accurately estimating quantiles, and this is where the particular smoothness assumptions above are used in our work. It is not clear whether this assumption or reliance on quantiles is necessary in general, and this is interesting to investigate in future work.

Problem Statement We assume we have have sample access to the distribution: that is, when we query the distribution, we are returned one iid sample from the distribution. Consider class $\mathcal{L}$, the class of light-tailed distributions and class $\mathcal{H}_{\alpha,\rho}$, the class of (α, ρ)-heavy-tailed distributions. Our goal is to develop an algorithm to determine whether a

[4] Note that the tester can also handle a unimodal distribution by determining where the decreasing region is. The increasing region could be handled by reflecting it across the mode.

[5] For a distribution over an unbounded domain, the CDF asymptotically approaches 1 as the domain value reaches ∞. Thus, its inverse must change in an unbounded way as $x \to 1$.

distribution comes from $\mathcal{L}$ or $\mathcal{H}_{\alpha,\rho}$ with probability 0.9 using finitely many samples, with sample complexity dependent on ϵ, α, ρ and parameters of the distribution.

3. A Proxy for the Behavior of the Tail

In this section, we motivate and define the Equal Weight Bucketing Scheme, upon which we build our test statistic and algorithm. We first determine the endpoints of the buckets and derive their lengths and rates of change of lengths based on the PDF and CDF. Then, we provide a proxy quantity based on this bucketing scheme that is equivalent to testing the hazard rate condition for heavy-tailedness. Finally, we explain how we can calculate this test statistic from samples by describing the bucketing scheme in terms of order statistics.

3.1. Defining a Bucketing Scheme

Dividing the support of a distribution into buckets is a well-known approach in distribution testing and is used widely in the literature Birge (1987); Batu et al. (2004); Canonne et al. (2018b); Batu et al. (2000). In essence, by considering the distribution on these buckets, we often can see trends that are less susceptible to sampling noise. In order to capture the drop rate of the distribution, we propose an equal-weight bucketing scheme.

We derive the expressions for the continuous notion of the Equal Weight Bucketing Scheme. The "weight" refers to probability mass. The left endpoint of a bucket starting at y with weight dy is given by $\mathbb{I}(y)$ and the length of that bucket (rate of change of endpoints) is given by by $\mathbb{L}(\cdot)$. This continuous notion allows us to derive a proxy quantity that reflects the hazard rate condition. The endpoints of the buckets are determined by the inverse of the CDF, giving us that in general, $\mathbb{I}(y) := F^{-1}(y)$ The derivative $d\mathbb{I}/dy$ gives us the length of the intervals. By the chain rule, the lengths and derivatives of lengths of the intervals are:

$$\mathbb{L}(y) := \frac{d}{dy}F^{-1}(y) = \frac{1}{F'(F^{-1}(y))} = \frac{1}{f(F^{-1}(y))} \quad \text{and} \quad \frac{d}{dy}\mathbb{L}(y) := \frac{-f'(F^{-1}(y))}{f(F^{-1}(y))^3} \tag{1}$$

3.2. Derivation of Proxy Quantity

Next, we present the proxy quantity that verifies the hazard rate condition for light-tailedness and (α, ρ)-heavy-tailed. This quantity is based on partitioning the support into intervals on which the density incurs equal weight. The full proof can be found in Appendix D.1.

Theorem 6 *For $\mathbb{L}(x)$ (defined in Equation 1) for a well-behaved function f, define $S(z) := \frac{\mathbb{L}(z)}{\frac{d}{dz}\mathbb{L}(z)}$:*

- *If for all $z \in [0, 1]$: $S(z) > 1 - z$, then the underlying distribution is light-tailed by Definition 4.*
- *Otherwise, if $S(z) < 1 - z - \frac{\alpha(1-z)^2}{\beta^3 B_1}$ for $z \in [z_0, z_0 + \rho]$, for any z_0, then it is (α, ρ)-heavy-tailed.*

Definition 7 *We refer to the distance in the proxy quantity between the lightest (α, ρ)-heavy-tailed distribution and the heaviest light-tailed distribution (exponential), a lower bound on which is $\frac{\alpha(1-z)^2}{\beta^3 B_1}$, as the* gap.

This is the gap in the proxy quantity metric between the two classes of distributions we hope to distinguish between. Since our eventual algorithm will rely on samples, the gap gives us some slack with which to handle error.

3.3. Test Statistic in Terms of Buckets

Here, we explain how we convert the proxy quantity into something we can calculate from knowing bucket endpoints. We approximate the derivative by the difference quotient. For well-behaved distributions, this approximation does not incur too much error due to Lipschitz-ness on the domain of interest. Detailed proofs can be found in Appendix D.

When the derivative of a function g is $B-Lipschitz$, and the derivative $g'(y)$ is monotone, then approximating $g'(y)$ by the difference quotient $\frac{g(y+\Delta y)-g(y)}{\Delta y}$ incurs no more than $B\Delta y$ additive error. This yields the following lemma, used to relax derivatives in S to discrete derivatives.

Lemma 8 *When the derivative of a function g is B_1-Lipschitz, the second derivative is B_2-Lipschitz, the derivative $g'(y)$ and the second derivative $g''(y)$ are both monotone, then approximating $g''(y)$ by:*

1. *estimating $\tilde{g}'(y) = \frac{g(y+\Delta y_1)-g(y)}{\Delta y_1}$, and $\tilde{g}'(y+\Delta y') = \frac{g(y+\Delta y_2+\Delta y_1)-g(y+\Delta y_2)}{\Delta y_1}$,*
2. *and estimating $\tilde{g}''(y) = \frac{\tilde{g}'(y+\Delta y_2)-\tilde{g}'(y)}{\Delta y_2}$.*

incurs no more than $2B_1\frac{\Delta y_1}{\Delta y_2} + B_2\Delta y_2$ additive error.

According to Fact 3.3, we can get bounded approximation error while approximating the derivative with steps of size $1/k$, which we can make small by setting k appropriately. However, in order for the second derivative approximation discussed in Lemma 8 to also be small, we need to consider buckets at two different levels of granularity, so we set $\Delta y_1 = \frac{1}{k^2}$ and $\Delta y_2 = \frac{1}{k}$. This gives us additive error of $\frac{B_1}{k}$ in the numerator and additive error of $\frac{2B_1+B_2}{k}$ in the denominator, which we can make small by setting k appropriately (see Corollary 16). Thus, we can approximate the proxy quantity by a discrete equivalent without incurring too much error (quantified in Lemma 11). Accordingly, we define:

$$\tilde{S} \coloneqq \frac{\tilde{L}_1}{(\tilde{L}_2 - \tilde{L}_1)/\Delta y_2}\,,\ \text{where}\ \tilde{L}_1 \coloneqq \frac{\mathbb{I}(y+\Delta y_1) - \mathbb{I}(y)}{\Delta y_1}\ \text{and}\ \tilde{L}_2 \coloneqq \frac{\mathbb{I}(y+\Delta y_1+\Delta y_2) - \mathbb{I}(y+\Delta y_2)}{\Delta y_1}\,. \tag{2}$$

3.4. Test Statistic in Terms of Order Statistic

In this section, we extend the formulation of the statistic in Equation 2 to show how we calculate it from samples. We set $y = i/k$. Approximating the derivative as the difference divided by the length, we get that the aforementioned tester can be written as follows in terms of the order statistic:

$$\hat{S}[i] = \frac{\left(X_{\left(\frac{ik+1}{k^2}\cdot(n+1)\right)} - X_{\left(\frac{i}{k}\cdot(n+1)\right)}\right)}{\left(k\left(X_{\left(\frac{(i+1)k+1}{k^2}\cdot(n+1)\right)} - X_{\left(\frac{i+1}{k}\cdot(n+1)\right)} - X_{\left(\frac{ik+1}{k^2}\cdot(n+1)\right)} + X_{\left(\frac{i}{k}\cdot(n+1)\right)}\right)\right)}\,.$$

In order to calculate the endpoints of the equal weight buckets, we draw four sets of samples, sort them, and then determine the bucket endpoints by considering the samples at indices $\frac{i}{k^2} \cdot n; i \in \{0, 1, ...k^2\}$. In any given calculation of the statistic, we need four of these order statistics; using different splits for each results in independence.

The test statistic is sensitive to the endpoints due to reliance on the length of buckets; since a small amount of mass could lie in an interval with very long length, this concentration of the endpoints is challenging to show (addressed in Appendix E).

4. Main Result

In this section, we present our main result (Theorem 9), the algorithm that gives us that upper bound, and discuss an overview of the proof.

Theorem 9 *There exists an algorithm that distinguishes between (α, ρ)-heavy-tailed distributions and light-tailed distributions requiring $\Theta\left(\max\left\{\frac{\beta^3 B_1}{\alpha\rho^2}, k\right\} \cdot k^2 \log k \sqrt{\sqrt{B_1} + 1}\right)$ samples with success probability 9/10, where $k = \max\left\{\Theta\left(\frac{\beta^4 B_1(2B_1+B_2)}{\alpha\,\rho^2}\right), \frac{4}{\rho}\right\}$.[6] Such an algorithm is given in Algorithm 1.*

Algorithm 1 (α, ρ)-Heavy-Tailed Test

```
Draw four sets of $n$ samples $\mathcal{R}^{(1)}, \mathcal{R}^{(2)}, \mathcal{R}^{(3)}, \mathcal{R}^{(4)} \leftarrow$ from the distribution
Sort the samples $\mathcal{R}^{(1)}, \mathcal{R}^{(2)}, \mathcal{R}^{(3)}, \mathcal{R}^{(4)}$
Split each $R^{(l)}$ into $k^2$ equal weight buckets.
$\forall \;\; i, j < k$, determine the interval endpoint $I^{(l)}[i, j]$ corresponding to the $(i \cdot k + j)^{th}$ bucket in $R^{(l)}$ (which is order statistic $X_{(i \cdot k+j)}$).
Calculate $L_1[i] = I^{(1)}[i, 1] - I^{(2)}[i, 0]$ and $L_2[i] = I^{(3)}[i, 1] - I^{(4)}[i, 0]$.
Calculate $L'[i] = \frac{L_1[i+1]-L_2[i]}{(1/k)}$.
Calculate the statistic $\hat{S}[i] = \frac{L_1[i]}{L'[i]}$ for $1 < i < k$.
if $\hat{S}[i] < 1 - \frac{i}{k} - \frac{1}{2} gap(\alpha)$ for any $i \in \{2, 3, \ldots, k-1\}$ then
    PASS.
else
    FAIL.
end
```

Algorithm At a high level, the algorithm draws samples, uses them to estimate the statistic, $\hat{S}$, for each bucket, and compares the statistics with the threshold defined in Theorem 6 to determine tail weight. More specifically, the algorithm first computes the order statistics as described in Section 3.4. Subsequently, the algorithm calculates the lengths L and the change in lengths L' of the buckets, and computes a test statistic $S[i]$ for $i \in \{2, \cdots, k-1\}$. The statistic is calculated for every k^{th} bucket within the k^2 buckets. We do not consider the first and last buckets, since the function is not Lipschitz there and so the approximations to the derivative will not be close to the true derivative. The parameter ζ as described in Definition 5, thus, must be $\leq 1/2k$, which allows us to safely use all but the last $k-1$ buckets of length k^2. If each test statistic lies above the threshold, the underlying distribution is declared light-tailed. On the other hand, if any of them lies below the threshold, the distribution is declared heavy-tailed.

Claim 10 *If $f(x)$ is (α, ρ)-Heavy-Tailed, then Alg. 1 will pass it with high probability. Moreover, if the underlying distribution is light-tailed then Alg. 1 will fail it with high probability.*

[6] This can be increased to probability $1 - \delta$ by repeating the algorithm $\log 1/\delta$ times using the standard amplification technique.

Proof Overview The proof breaks down into stages that correspond to the different approximations we make to get from the proxy quantity to the final test statistic. We will use S to denote the proxy quantity, $\tilde{S}$ to denote the statistic approximated by discrete derivatives, and $\hat{S}$ is used to denote the empirical statistic that calculated using order statistics of the samples (Figure 2).

Our analysis proceeds through four stages: (1) correctness of proxy S; (2) $\tilde{S}$ close to S; (3) $\hat{S}$ close to $\tilde{S}$; (4) setting k, n to satisfy theorem.

Here, we provide the lemmas that quantify the error incurred from S to $\tilde{S}$ and $\tilde{S}$ to $\hat{S}$. Further details and analysis of how to satisfy the conditions of these lemmas, are discussed in Section E.

Part 1/4: $\boldsymbol{S}$ is an accurate proxy.

The proxy quantity derived in Theorem 6 considered with respect to the threshold (halfway across the gap) gives a test which accurately determines whether a set of samples came from a light-tailed distribution or a distribution that is (α, ρ)-heavy-tailed. An (α, ρ)-heavy-tailed distribution has hazard rate decreasing at least by α over a region of the PDF with probability mass ρ, which gives us an expression for how far the statistic must be from the original threshold $1 - i/k$ in order for the hazard rate to be decreasing by at least α. Further, if the region of mass ρ lies in at least two buckets, then the proxy quantity will detect it. Thus, if a distribution is light-tailed, then *all* $k - 3$ of the calculated proxy quantities calculated will lie above $1 - i/k$; if even one of the proxies lies below $1 - i/k - gap$, then the distribution is (α, ρ)-heavy-tailed. See Section 3 and Appendix D.1 for detailed discussion.

Part 2/4: $\tilde{\boldsymbol{S}}$ is close to $\boldsymbol{S}$.

In the next step, we show that approximating the derivatives in the proxy quantity by the respective difference quotients causes the test statistic to incur bounded error. Recall that the proxy we are using is $S = N/D = \left(\frac{d}{dy} F^{-1}(y)\right) / \left(\frac{d^2}{dy^2} F^{-1}(y)\right)$, which is approximated by $\tilde{S} = \tilde{N}/\tilde{D} = \tilde{L}_1/((\tilde{L}_2 - \tilde{L}_1)/\Delta y_2)$ as in Eq. 2.

If the error incurred in $\tilde{N}, \tilde{D}$ by approximating the derivatives as above has value ϵ' (we show this condition is met in Section E due to the Lipschitzness conditions), then either we can estimate the value of the proxy quantity within a bounded additive error (and the proxy quantity is small) or the value of the proxy quantity is greater than 1 (and we know we are in the light-tailed case).

Lemma 11 (Additive Bound for $\tilde{S}$) *Given a parameter $\epsilon < 1$, if $|\tilde{N} - N|$ and $|\tilde{D} - D|$ are at most $\epsilon' := \epsilon/(6\beta)$, then either $|\tilde{S} - S| < \epsilon$ or both $S, \tilde{S}$ are at least one.*

PART 3/4: $\hat{\boldsymbol{S}}$ (calculated from order statistics) is close to $\tilde{\boldsymbol{S}}$.

We must next show that we can approximate $\tilde{S}$ as defined in the previous section by $\hat{S}$, computed from the order statistics of a set of samples. First, if we estimate $\tilde{L}_1$ and $\tilde{L}_2$ accurately up to a multiplicative $(1 \pm \epsilon')$ factor, and obtain $\hat{L}_1$ and $\hat{L}_2$, then $\tilde{S}$ can be approximated by $\hat{S} := \frac{\hat{L}_1}{k \cdot \left(\hat{L}_2 - \hat{L}_1\right)}$.

Lemma 12 (Multiplicative Bound for $\hat{S}$) *Suppose we have $\hat{L}_1$ and $\hat{L}_2$, the estimates of $\tilde{L}_1$ and $\tilde{L}_2$ with a multiplicative factor of $\epsilon' = \min\left(\Theta\left(\epsilon/((1+\epsilon)\cdot k)\right), \Theta\left(1/k^2\right)\right)$. Then, one of the following cases holds:*

1. *$\hat{S} > 1 - 2/k$ and $\tilde{S} > 1$, implying they come from a light-tailed region of the distribution.*
2. *$(1-\epsilon)\cdot\tilde{S} \leq \hat{S} \leq (1+\epsilon)\cdot\tilde{S}$.*

To show that we can get the multiplicative estimates $\hat{L}_1, \hat{L}_2$, we need to show that the order statistics concentrate well so that the estimates for the end points are not too far off from their theoretical values. For this, we construct a map between samples from the uniform distribution on $[0,1]$ and an arbitrary distribution with CDF F, showing that concentration of the order statistics of a set of samples from the former implies concentration of order statistics of a set of samples from the latter. This concentration can be expressed in terms of additive error (Lemma 17), and this can be translated to multiplicative error (Lemma 18). We discuss this in more detail in Section E.

Part 4/4: Errors can be set to satisfy theorem.

Finally, we must limit the errors we incur in Parts 2 and 3 to determine how many buckets and samples we require. For this, we note that after incurring both the derivative approximation error and the sampling error, the test statistic $\hat{S}$ must still be on the same side of the threshold (partway across the gap), as S. We ensure that we split the distribution into sufficiently many buckets that no more than a quarter of the gap is crossed due to derivative approximation error. Further, we draw enough samples that no more a tenth of the gap is crossed due to sampling error. Thus, even when both errors are incurred, the test statistic remains on the correct side of the threshold.

5. Hardness Result

In this section, we show that in the absence of assumptions, this problem cannot be solved with finitely many samples. In particular, for any number of samples m, we present two classes of distributions, one light-tailed and the other heavy-tailed, that are indistinguishable using m samples. To start, we consider a slightly different definition for heavy-tailed-ness. We then show that it is hard to distinguish these distributions from light-tailed ones.

Definition 13 *We say a distribution p is (α, ρ)-scattered-heavy-tailed if the hazard rate is decreasing by rate $\geq \alpha$ over measurable intervals of the domain with probability mass $\geq \rho$.*

High level idea: We construct two classes of distributions that are hard to distinguish with few samples: $\mathcal{C}_L$ (light-tailed), and $\mathcal{C}_H$ (heavy-tailed). The class of light-tailed distributions contains only one member that is an exponential distribution with $f_{\exp}(x) = e^{-x}$. We construct the class of heavy-tailed distributions via a randomized process as follows: We start off by the same distribution $f_{\exp}(x) = e^{-x}$. We split the domain of $f_{\exp}$ into s *chunks* such that the probability mass in every chunk is equal to $1/s$. Then, we select roughly $\rho' = \Theta(\rho)$ fraction of these chunks randomly and embed a heavy-tailed distribution, namely f_H, in (some of) those selected chunks. The construction has two key properties: First, the probability mass of a chunk remains the same even when the alteration happens. Second, if we draw one sample from a chunk, we cannot tell whether it is altered or not.

When we alter a chunk, we randomly replace $f_{\exp}$ by a heavy-tailed piece f_H or another partial PDF $\overline{f}_H$. We simply define $\overline{f}_H$ such that the mixture of f_H and $\overline{f}_H$ each with probability a half gives us exactly $f_{\exp}$. Thus, if we receive one sample from a chunk that comes from a random $\mathcal{C}_H$, it is impossible to tell whether the chunk is altered or not. It is worth noting that this process generates a class of distributions, $\mathcal{C}_H$, that depends on a parameter s. We may also $\mathcal{C}_H(s)$ to denote it. To complete our proof, we show that for any algorithm that uses m samples, there is a sufficiently large s such that it is very unlikely to have more than one sample per chunk. Thus, $\mathcal{C}_L$ and $\mathcal{C}_H(s)$ are indistinguishable when we use m samples. This fact implies that no algorithm that uses finitely many samples can distinguish a light tailed distribution from a distribution that is heavy on measurable subset of the domain with mass ρ unless we make further assumptions including that the heavy-tailed part might need to be contiguous. We state the result formally in the theorem below, and the proof is in Appendix F.

Theorem 14 *For any integer m, there is no algorithm that receives m samples from p, a monotone and continuous distribution, and can distinguish whether p is light-tailed or $(\alpha, \rho)-$scattered-heavy-tailed for $\alpha < 0.0043$ and $\rho < 0.5$ with probability more than $0.5 + o(1)$.*

6. Experiments

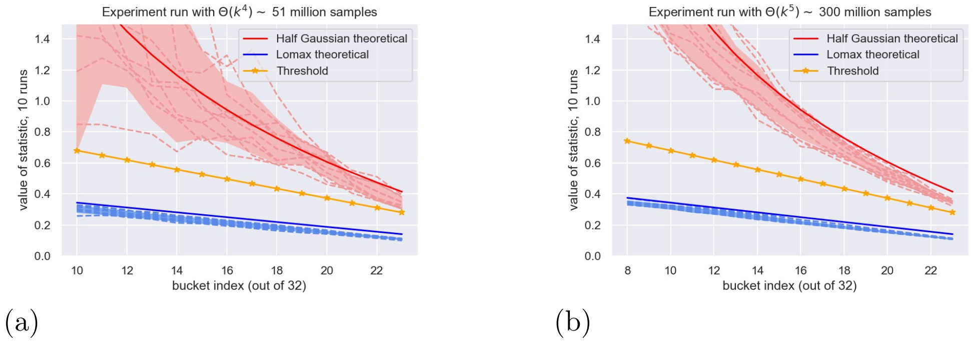
We present experiments that validate our theoretical results. Through experiments on synthetic data, we demonstrate that our statistic distinguishes between Gaussian (light-tailed) and Lomax (heavy-tailed) distributions. The algorithm we run has weaker theoretical guarantees, as we do not consider both granularities of bucketing during derivative approximation.

Data and Methods Points are sampled from (half) Gaussian and Lomax distributions using built-in functions in `numpy.random`. We sample $n = \Theta(k^4) \approx 51$ million and $n = \Theta(k^5) \approx 300$ million points for $k = 32$.

Results We find that the test statistic distinguishes between Gaussian and Lomax over the required range of bucket indices. In Figure 1, the dashed red lines refer to the calculated value of the test statistic for a Gaussian distribution and blue lines for the Lomax. One standard deviation away from the mean is shaded in the appropriate color. From Figure 1, we note: (1) the value of the test statistic calculated from samples is very close to the proxy quantity (S); (2) in part (a) (approx. 51 million samples), we can distinguish Gaussian (red) and Lomax (blue) well over a large range of buckets, but we see greater variability between runs (spread of the dashed lines) than in part (b) (300 million samples, same number of buckets), where the calculated statistic concentrates better. In Appendix H, we explore what happens when we use substantially fewer samples, observing that we require some minimum number of samples per bucket but start to see separation between the same two distributions with a relatively small number of buckets. Thus, in situations where it is not essential that we get provable guarantees, the same algorithm could work with fewer samples.

Acknowledgments

MA is supported by NSF grants CNS-2120667, CNS-2120603, CCF 1934846, and BU's Hariri Institute for Computing. ASB is supported by NSF grants CCF-2006664, IIS-1741137, CCF-1733808, and



(a) (b)

Figure 1: These plots depict the results of 10 runs of Algorithm 2 on samples drawn from a half Gaussian distribution and 10 runs of it on samples drawn from a Lomax. Experiments plotted in (a) used $n = \Theta(k^4)$ samples, and those in (b) used $n = \Theta(k^5)$ samples. With more samples, we are able to distinguish the two distributions over a broader range of buckets. Depicted are the proxy quantities (solid), lines representing the 10 runs for each distribution (dashed), and shading one standard deviation above and below the mean. The orange line represents the threshold, calculated based on $\alpha = 1/4$ for Lomax and appropriate settings for β, B_1 based on the distributions we considered.

Fintech@CSAIL. KR was supported in part by the National Science Foundation under grant CCF-1815011, by the NSF-Simons Funded Collaboration on the Mathematics of Deep Learning, and by the Defense Advanced Research Projects Agency under cooperative agreement HR00112020003. The views expressed in this work do not necessarily reflect the position or the policy of the Government and no official endorsement should be inferred. RR is supported by NSF grants CCF-2006664, DMS-2022448, IIS-1741137, CCF-1733808, CCF-1740751, CCF-2006664, and Fintech@CSAIL.

References

Jayadev Acharya and Constantinos Daskalakis. Testing poisson binomial distributions. In *Proceedings of the 2015 Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*, pages 1829–1840, 2015. doi: 10.1137/1.9781611973730.122. URL https://epubs.siam.org/doi/abs/10.1137/1.9781611973730.122.

Jayadev Acharya, Constantinos Daskalakis, and Gautam Kamath. Optimal testing for properties of distributions. In *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'15, pages 3591–3599, Cambridge, MA, USA, 2015. MIT Press. URL http://dl.acm.org/citation.cfm?id=2969442.2969640.

Michał Adamaszek, Artur Czumaj, and Christian Sohler. Testing monotone continuous distributions on high-dimensional real cubes. In *Proceedings of the 2010 Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*, pages 56–65, 2010. doi: 10.1137/1.9781611973075.6. URL https://epubs.siam.org/doi/abs/10.1137/1.9781611973075.6.

Maryam Aliakbarpour, Themis Gouleakis, John Peebles, Ronitt Rubinfeld, and Anak Yodpinyanee. Towards testing monotonicity of distributions over general posets. In *Proceedings of the Thirty-Second Conference on Learning Theory, COLT*, pages 34–82, 2019.

Noga Alon, Alexandr Andoni, Tali Kaufman, Kevin Matulef, Ronitt Rubinfeld, and Ning Xie. Testing k-wise and almost k-wise independence. In *Proceedings of the Thirty-Ninth Annual ACM Symposium on Theory of Computing*, STOC '07, page 496–505, New York, NY, USA, 2007.

Association for Computing Machinery. ISBN 9781595936318. doi: 10.1145/1250790.1250863. URL https://doi.org/10.1145/1250790.1250863.

Richard E. Barlow, Albert W. Marshall, and Frank Proschan. Properties of probability distributions with monotone hazard rate. *Annals of Mathematical Statistics*, 34(2):375–389, 1963. ISSN 0003-4851. URL https://www.jstor.org/stable/2238381. Publisher: Institute of Mathematical Statistics.

Tugkan Batu, Lance Fortnow, Ronitt Rubinfeld, Warren D. Smith, and Patrick White. Testing that distributions are close. In *41st Annual Symposium on Foundations of Computer Science, FOCS 2000, 12-14 November 2000, Redondo Beach, California, USA*, pages 259–269. IEEE Computer Society, 2000. doi: 10.1109/SFCS.2000.892113. URL https://doi.org/10.1109/SFCS.2000.892113.

Tugkan Batu, Ravi Kumar, and Ronitt Rubinfeld. Sublinear algorithms for testing monotone and unimodal distributions. In *Proceedings of the thirty-sixth annual ACM symposium on Theory of computing - STOC '04*, page 381. ACM Press, 2004. ISBN 978-1-58113-852-8. doi: 10.1145/1007352.1007414. URL http://portal.acm.org/citation.cfm?doid=1007352.1007414.

Tuğkan Batu, Lance Fortnow, Ronitt Rubinfeld, Warren D. Smith, and Patrick White. Testing closeness of discrete distributions. *Journal of the ACM*, 60(1):1–25, 2013. ISSN 0004-5411, 1557-735X. doi: 10.1145/2432622.2432626. URL https://dl.acm.org/doi/10.1145/2432622.2432626.

Lucien Birge. On the risk of histograms for estimating decreasing densities. *Ann. Statist.*, 15(3):1013–1022, 09 1987. doi: 10.1214/aos/1176350489. URL https://doi.org/10.1214/aos/1176350489.

Maurice C Bryson. Heavy-tailed distributions: properties and tests. *Technometrics*, 16(1):61–68, 1974.

Clément L. Canonne. Are few bins enough: Testing histogram distributions. In *Proceedings of the 35th ACM SIGMOD-SIGACT-SIGAI Symposium on Principles of Database Systems*, PODS '16, page 455–463, New York, NY, USA, 2016. Association for Computing Machinery. ISBN 9781450341912. doi: 10.1145/2902251.2902274. URL https://doi.org/10.1145/2902251.2902274.

Clément L Canonne, Ilias Diakonikolas, Themis Gouleakis, and Ronitt Rubinfeld. Testing shape restrictions of discrete distributions. *Theory of Computing Systems*, 62(1):4–62, 2018a.

Clément L Canonne, Themis Gouleakis, and Ronitt Rubinfeld. Sampling correctors. *SIAM Journal on Computing*, 47(4):1373–1423, 2018b.

Clément L. Canonne. A survey on distribution testing: Your data is big. but is it blue? *Theory of Computing*, 2020.

Siu-On Chan, Ilias Diakonikolas, Paul Valiant, and Gregory Valiant. Optimal algorithms for testing closeness of discrete distributions. In *Proceedings of the 2014 Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*, pages 1193–1203, 2014. doi: 10.1137/1.9781611973402.88. URL https://epubs.siam.org/doi/abs/10.1137/1.9781611973402.88.

V. P. Chistyakov. A theorem on sums of independent positive random variables and its applications to branching random processes. *Theory of Probability & Its Applications*, 9(4):640–648, 1964. doi: 10.1137/1109088. URL https://doi.org/10.1137/1109088.

Ilias Diakonikolas, Themis Gouleakis, John Peebles, and Eric Price. Collision-based testers are optimal for uniformity and closeness. *Chicago Journal of Theoretical Computer Science*, 2019(1), May 2019.

Vitaly Feldman. Does learning require memorization? a short tale about a long tail. In Konstantin Makarychev, Yury Makarychev, Madhur Tulsiani, Gautam Kamath, and Julia Chuzhoy, editors, *Proccedings of the 52nd Annual ACM SIGACT Symposium on Theory of Computing, STOC 2020, Chicago, IL, USA, June 22-26, 2020*, pages 954–959. ACM, 2020. doi: 10.1145/3357713.3384290. URL https://doi.org/10.1145/3357713.3384290.

Irène Gijbels and Nancy Heckman. Nonparametric testing for a monotone hazard function via normalized spacings. *Journal of Nonparametric Statistics*, 16(3):463–477, 2004. ISSN 1048-5252, 1029-0311. doi: 10.1080/10485250310001622668. URL http://www.tandfonline.com/doi/abs/10.1080/10485250310001622668.

Peter Hall, Ingrid Van Keilegom, et al. Testing for monotone increasing hazard rate. *The Annals of Statistics*, 33(3):1109–1137, 2005.

Mor Harchol-Balter. The effect of heavy-tailed job size distributions on computer system design. In *Proc. of ASA-IMS Conf. on Applications of Heavy Tailed Distributions in Economics, Engineering and Statistics*, page 17, 1999.

Chen-Yu Hsu, Piotr Indyk, Dina Katabi, and Ali Vakilian. Learning-based frequency estimation algorithms. In *International Conference on Learning Representations*, 2018.

Stuart Klugman, Harry Panjer, and Gordon Willmot. *Loss Models: From Data to Decisions*. John Wiley & Sons, Inc., second edition edition, 2004.

Hongzi Mao, Malte Schwarzkopf, Shaileshh Bojja Venkatakrishnan, Zili Meng, and Mohammad Alizadeh. Learning scheduling algorithms for data processing clusters. In *Proceedings of the ACM Special Interest Group on Data Communication*, pages 270–288. ACM, 2019.

Olivier Marchal, Julyan Arbel, et al. On the sub-gaussianity of the beta and dirichlet distributions. *Electronic Communications in Probability*, 22, 2017.

Thomas Mikosch. *Regular variation, subexponentiality and their applications in probability theory*, volume 99. Eindhoven University of Technology Eindhoven, The Netherlands, 1999.

Jerzy Neyman, Egon Sharpe Pearson, and Karl Pearson. Ix. on the problem of the most efficient tests of statistical hypotheses. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, 231(694-706):289–337, 1933. doi: 10.1098/rsta.1933.0009. URL https://royalsocietypublishing.org/doi/abs/10.1098/rsta.1933.0009.

Karl Pearson. On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 50(302):157–175, 1900. doi: 10.1080/14786440009463897. URL https://doi.org/10.1080/14786440009463897.

Ronitt Rubinfeld and Arsen Vasilyan. Monotone probability distributions over the boolean cube can be learned with sublinear samples. page 34 pages. doi: 10.4230/LIPICS.ITCS.2020.28. URL https://drops.dagstuhl.de/opus/volltexte/2020/11713/. Artwork Size: 34 pages Medium: application/pdf Publisher: Schloss Dagstuhl - Leibniz-Zentrum fuer Informatik GmbH, Wadern/Saarbruecken, Germany Version Number: 1.0.

Chun Su and Qi-he Tang. Characterizations on heavy-tailed distributions by means of hazard rate. *Acta Mathematicae Applicatae Sinica*, 19(1):135–142, 2003. ISSN 0168-9673, 1618-3932. doi: 10.1007/s10255-003-0090-6. URL http://link.springer.com/10.1007/s10255-003-0090-6.

TPC. TPC-h benchmark. URL http://tpc.org/tpch/default5.asp. Retrieved from: http://tpc.org/tpch/default5.asp.

Paul Valiant. Testing symmetric properties of distributions. *SIAM Journal on Computing*, 40(6): 1927–1968, 2011. doi: 10.1137/080734066. URL https://doi.org/10.1137/080734066.

Roman Vershynin. *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge university press, 2018.

Yu-Xiong Wang, Deva Ramanan, and Martial Hebert. Learning to model the tail. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 7029–7039. Curran Associates, Inc., 2017. URL http://papers.nips.cc/paper/7278-learning-to-model-the-tail.pdf.

Appendix A. Alternate Definitions of Heavy-Tailedness

It is non-trivial to give a single unifying definition of heavy-tailed distributions, since the definition needs to accommodate a wide range of behaviors, including distributions whose tails fall off at irregular rates. Further, no notion of having a "heavy" tail is absolute – there must be a point of reference. Throughout the literature, a plethora of non-equivalent conditions on the tail of a distribution are used to define what it means for the distribution to be considered "heavy-tailed." Some examples include increasing conditional mean exceedance Bryson (1974), decreasing hazard rate Klugman et al. (2004), the lack of finite exponential moments, having tails that decay more slowly than exponential Mikosch (1999); Chistyakov (1964), infinite variance Harchol-Balter (1999), regularly varying tails Mikosch (1999), and definitions that build on these conditions Su and Tang (2003). One uniting factor of these definitions is that the point of reference is the exponential distribution, meaning that a distribution whose tail decays more slowly than the exponential is considered "heavy-tailed." [7] In this work, we focus on Klugman et al.'s definition of heavy-tailed that is based on the property that the hazard rate of a heavy-tailed distribution is decreasing Klugman et al. (2004); the characterization is of similar structure to the definitions that Bryson (1974); Su and Tang (2003) use, admits a clean description, and reflects the idea that heavy tails decay more slowly than exponential tails, thereby adequately classifying known families of distributions into heavy-tailed versus light-tailed – e.g., Lomax (heavy), exponential (light), and Gaussian (light).

In this section, we provide alternate definitions for the notion of "heavy-tailed." We use the usual notation that a distribution has Probability Density Function (PDF) $f(x)$ and Cumulative Density Function (CDF) $F(x)$.

A.1. Increasing Conditional Mean Exceedance

The conditional mean exceedance (CME) is defined in Bryson (1974) to be:

$$CME(x) = \mathbb{E}\left[(X - x)|X \geq x\right] = \frac{\int_x^\infty (s-x)f(s)ds}{1-F(x)} \tag{3}$$

According to Bryson (1974), a distribution is considered heavy-tailed if the CME is increasing for all x. The CME essentially captures the shape of tail by considering how the expectation of the tail of the distribution evolves throughout the tail. Thus, if it is increasing, the further into the distribution we go, the more that remains.

This definition considers an exponential distribution the canonical "medium-tailed" distribution and the Lomax distribution the canonical "heavy-tailed" distribution. In practical settings, this metric is valuable, as it represents "decreasing failure rate" Harchol-Balter (1999): the longer a job has been alive (as x increases), the longer it is expected to stay alive ($\mathbb{E}(X - x|X \geq x)$ increasing). This definition can also be framed as the reciprocal of the hazard rate of the equilibrium distribution function of a distribution Su and Tang (2003).

A.2. Infinite Variance

Harchol-Balter says that heavy-tailed distributions are ones that have infinite variance Harchol-Balter (1999). From samples, the variance can never really be infinite; in those cases, very large variance signify heavy-tailedness. This definition is difficult to quantify.

This definition considers a subset of Lomax and Pareto distributions "heavy-tailed."

[7] Distributions with regularly varying tails represent a subset of distributions whose tails decay more slowly than exponential.

A.3. Regular Variation

The tail of a distribution captures how the PDF behaves as $x \to \infty$. We introduce Mikosch (1999)'s notion of regular variation as a way to capture this behavior.

Definition 15 *A positive measurable function f is called* regularly varying (at infinity) with index $\alpha \in \mathbb{R}$ *if*

- *It is defined on some neighborhood $[x_0, \infty)$ of infinity.*
- $\lim\limits_{x \to \infty} \frac{f(tx)}{f(x)} = t^\alpha$ *for all $t > 0$.*

If $\alpha = 0$, then f is said to be slowly varying (at infinity)

One class of heavy-tailed distributions is that which have regularly varying tails. This includes distributions of the type x^α, etc. Note that these are related to generalized Pareto distributions. Thus, this class of distributions is narrower in scope than the set of heavy-tailed distributions considered by other definitions.

A.4. Moment Generating Function Infinite

According to Su and Tang (2003), a distribution is considered heavy-tailed if it has the moment generating function (MGF) is not bounded, i.e., the distribution has no finite exponential moments.. That is, if

$$\int\limits_0^\infty e^{tx} f(x) dx = \infty \qquad \forall t > 0\,, \tag{4}$$

then we consider a distribution heavy-tailed. This definition captures the idea that the tail of a distribution must be heavier than exponential in order for it to be considered "heavy-tailed."

A.5. Relating Definitions

The characterization using the moment generating function in the previous section considers any distribution of the form $e^{-g(x)}$ where $g(x)$ is asymptotically smaller than linear "heavy-tailed." Indeed, this is true of the CME and HR characterizations, as well, and Mikosch (1999) notes that distributions that decay more slowly than the exponential are heavy-tailed. For the general class of distributions with CDF $F(x) = 1 - e^{-g(x)}$ and PDF $f(x) = g'(x)e^{-g(x)} = e^{-(g(x) - \ln(g'(x)))}$, the CME characterization, the HR characterization, and the moment-generating function characterization all consider the same families of distributions heavy-tailed and light-tailed. In light of this, we choose to consider the hazard rate definition, as it is the simplest expression of this notion.

Appendix B. Related Work Discussion

In the works by Acharya et al. (2015) and Canonne et al. (2018b), algorithms are given to perform the task of testing monotone hazard rate with sample complexity that has dependence on the domain size that is nearly square root.

In them, the domain size is finite and no assumptions on the monotonicity, continuity, or Lipschitzness of the distributions are made. The paradigm for testing MHR in Acharya et al. (2015); Canonne et al. (2018a) can be broken into the three steps: first, approximately learning the underlying distribution from samples (assuming that it is MHR), second, testing whether the learned distribution is in fact close to the original distribution, and third, testing whether the learned distribution is in the class or far from the class. The sample complexity $(O(n/\epsilon^2 + \log(n/\epsilon)/\epsilon^4))$ is dominated by the cost of the second step, while the third step requires no new samples and is solely computational, via a solution to a linear programming problem. As the variables for the linear program represent the probability mass function, this step cannot easily translate to distributions over continuous domains. Since the sample complexity of this algorithm scales with domain size, it is not finite in our setting. In contrast, we give a finite sample guarantee, but the result is incomparable due to the assumptions we make.

Appendix C. PDF, CDF of Some Common Continuous Distributions

In this section, we define the various distributions that are referenced throughout the paper as are relevant in our setting.

Exponential The cutoff between heavy-tailed and light-tailed distributions according to both the CME and HR definitions is the exponential distribution. The continuous exponential distribution requires parameter $\lambda > 0$ has probability density function (PDF) $f(x) = \lambda e^{-\lambda x}; x \in [0, \infty)$, cumulative density function (CDF) $F(x) = 1 - e^{-\lambda x}$, and quantile function $F^{-1}(x) = \frac{-1}{\lambda} \ln 1 - x$.

Lomax The Lomax distribution requires parameters $\alpha > 0, \lambda > 0$ and has PDF$f(x) = \frac{\alpha}{\lambda}\left[1 + \frac{x}{\lambda}\right]^{-(\alpha+1)}$, CDF $F(x) = 1 - \left[1 + \frac{x}{\lambda}\right]^{-\alpha}$, and quantile function $F^{-1}_{X,L}(x) = \lambda\left(\frac{1}{(1-x)^{1/\alpha}} - 1\right)$. We consider this the canonical "heavy-tailed" distribution.

Half-Gaussian The half-Gaussian distribution requires parameter σ and has PDF:

$$f(x) = \frac{2}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x}{\sigma}\right)^2}; x \in [0, \infty) \tag{5}$$

and CDF:

$$F(x) = \operatorname{erf}\left(\frac{x}{\sigma\sqrt{2}}\right) \tag{6}$$

The quantile function is:

$$F^{-1}(x) = \sigma\sqrt{2}\operatorname{erf}^{-1}(x) \tag{7}$$

We consider this the canonical "light-tailed" distribution.

Appendix D. Test Statistic Analysis

D.1. Proof of Theorem 6

Theorem 6 *For $\mathbb{L}(x)$ (defined in Equation 1) for a well-behaved function f, define $S(z) := \frac{\mathbb{L}(z)}{\frac{d}{dz}\mathbb{L}(z)}$:*

- *If for all $z \in [0, 1]$: $S(z) > 1 - z$, then the underlying distribution is light-tailed by Definition 4.*
- *Otherwise, if $S(z) < 1 - z - \frac{\alpha(1-z)^2}{\beta^3 B_1}$ for $z \in [z_0, z_0 + \rho]$, for any z_0 , then it is (α, ρ)-heavy-tailed.*

Proof Let $f(y)$ be the PDF of a distribution and $F(y)$ the CDF. From Equation 1, when considering the ratio of the length of the buckets to the rate of change of the derivatives, we have that:

$$\frac{L(y)}{\frac{d}{dy}L(y)} = \frac{L(y)}{-f'(F^{-1}(y))f(F^{-1}(y))^{-2}L(y)} = \left.\frac{-f(F^{-1}(y))^2}{f'(F^{-1}(y))}\right|_{y=z} = \frac{-f(F^{-1}(z))^2}{f'(F^{-1}(z))} \tag{8}$$

First, in order for the distribution to be light-tailed, $\frac{d}{dy}HR(f(y)) > 0$, which implies:

$$\frac{d}{dy}\frac{f(y)}{1-F(y)} = \frac{(1-F(y))f'(y) - f(y)(-f(y))}{(1-F(y))^2} > 0 \tag{9}$$

$$\Rightarrow (1-F(y))f'(y) - f(y)(-f(y)) > 0 \tag{10}$$

$$\Rightarrow \frac{-f(y)^2}{f'(y)} > 1 - F(y)\bigg|_{y=F^{-1}(z)} \tag{11}$$

$$\Rightarrow \frac{-f(F^{-1}(z))^2}{f'(F^{-1}(z))} = S(z) > 1 - F\left(F^{-1}\left(z\right)\right) = 1 - z \tag{12}$$

Equation 11 is a result of our assumption that $f(y)$ is monotone decreasing, and so its derivative must be negative. This derivation gives us a condition on the tail of the distribution which, if satisfied, implies that the distribution is light-tailed.

Now, we consider the case where the hazard rate decreases by at least α. In this case:

$$\frac{d}{dy}\frac{f(y)}{1-F(y)} = \frac{(1-F(y))f'(y) - f(y)(-f(y))}{(1-F(y))^2} < -\alpha \tag{13}$$

$$\Rightarrow (1-F(y))f'(y) - f(y)(-f(y)) < -\alpha(1-F(y))^2 \tag{14}$$

$$\Rightarrow \frac{-f(y)^2}{f'(y)} < 1 - F(y) + \frac{\alpha(1-F(y))^2}{f'(y)}\bigg|_{y=F^{-1}(z)} \tag{15}$$

$$\Rightarrow \frac{-f(F^{-1}(z))^2}{f'(F^{-1}(z))} = S(z) < 1 - z + \frac{\alpha(1-z))^2}{f'(F^{-1}(z))} = 1 - z + \frac{\alpha(1-z)^2}{f'(F^{-1}(z))} < 1 - z - \frac{\alpha(1-z)^2}{\beta^3 B_1} \tag{16}$$

Thus, if the proxy quantity is greater than $1-z$, then the distribution is light-tailed; otherwise, if the proxy quantity lies below $1 - z - \frac{\alpha(1-z)^2}{\beta^3 B_1}$ over a domain of size ρ, then the distribution is (α, ρ)-heavy-tailed.

By imposing the condition from Definition 3 on the hazard rate, we were able to recover an expression where one term denotes the usual threshold and there is an additive term, dependent on α, that specifies what the **gap** is between the threshold and the lightest heavy-tailed distribution for which we can test. In particular:

$$S(z) \leq \frac{-\alpha(1-z)^2}{-f'(F^{-1}(z))} + (1-z) \Rightarrow \text{gap}(\alpha, z) = \frac{-\alpha(1-z)^2}{-f'(F^{-1}(z))} \leq \frac{-\alpha(1-z)^2}{\beta^3 B_2}, z \in \left\{\frac{i}{k}, \frac{i+2}{k}\right\}$$

In our analysis, when we refer to the gap, this is what we are referencing.

■

D.2. Proof of Fact 3.3

When the derivative of a function g is $B-$*Lipschitz*, and the derivative $g'(y)$ is monotone, then approximating $g'(y)$ by the difference quotient $\frac{g(y+\Delta y) - g(y)}{\Delta y}$ incurs no more than $B\Delta y$ additive error.

Proof By the intermediate value theorem, there exists some point $y' \in [y, y+\Delta y]$, s.t.,

$$g'(y') = \frac{g(y+\Delta y) - g(y)}{\Delta y}$$

$$|g'(y') - g'(y)| \leq |g'(y+\Delta y) - g'(y)| \leq B \cdot \Delta y$$

■

D.3. Proof of Lemma 8

Lemma 8 *When the derivative of a function g is B_1-Lipschitz, the second derivative is B_2-Lipschitz, the derivative $g'(y)$ and the second derivative $g''(y)$ are both monotone, then approximating $g''(y)$ by:*

1. *estimating $\tilde{g}'(y) = \frac{g(y+\Delta y_1)-g(y)}{\Delta y_1}$, and $\tilde{g}'(y+\Delta y') = \frac{g(y+\Delta y_2+\Delta y_1)-g(y+\Delta y_2)}{\Delta y_1}$,*
2. *and estimating $\tilde{g}''(y) = \frac{\tilde{g}'(y+\Delta y_2)-\tilde{g}'(y)}{\Delta y_2}$.*

incurs no more than $2B_1\frac{\Delta y_1}{\Delta y_2} + B_2\Delta y_2$ additive error.

Proof We apply Fact 3.3 once to obtain $(B_1\Delta y_1)$-additive approximations $\tilde{g}'(y)$ and $\tilde{g}'(y+\Delta y)$. These approximations are then used to compute:

$$\tilde{g}''(y) = \frac{\tilde{g}'(y+\Delta y_2) - \tilde{g}'(y)}{\Delta y_2}$$

which is a $2B_1\frac{\Delta y_1}{\Delta y_2}$-additive approximation to the true $[g'(y+\Delta y_2) - g'(y)]/\Delta y_2$. Now, we can apply Fact 3.3 once again to obtain the final estimate, which will have $2B_1\frac{\Delta y_1}{\Delta y_2} + B_2\Delta y_2$ additive error. ■

Appendix E. Proof of Main Theorem

Theorem 9 *There exists an algorithm that distinguishes between (α,ρ)-heavy-tailed distributions and light-tailed distributions requiring $\Theta\left(\max\left\{\frac{\beta^3 B_1}{\alpha\rho^2}, k\right\} \cdot k^2 \log k\sqrt{\sqrt{B_1}+1}\right)$ samples with success probability 9/10, where $k = \max\left\{\Theta\left(\frac{\beta^4 B_1(2B_1+B_2)}{\alpha\,\rho^2}\right), \frac{4}{\rho}\right\}$.*[8] *Such an algorithm is given in Algorithm 1.*

Proof

As we introduced earlier, we have the quantity S that serves as a proxy for the tail weight of a distribution. For a light-tailed distribution, the proxy quantity calculated at each bucket will lie above the threshold, whereas for an (α,ρ)-heavy-tailed distribution, the proxy quantity will lie below the threshold for at least one of the buckets, where the threshold is dependent on α, how quickly the hazard rate is decreasing. However, we cannot calculate S directly without full knowledge of the distribution, so we consider a relaxation $\tilde{S}$, in which derivatives are approximated by difference quotients. Finally, we draw samples from the distribution to approximate $\tilde{S}$, and we call this approximation from samples $\hat{S}$ (Figure 2). We show that $\hat{S}$ requires the stated number of samples to distinguish between (α,ρ)-heavy-tailed and light-tailed distributions. In order to prove the theorem, then, there are four steps:

1. **Show that S is a correct proxy for the tail weight.** The proxy quantity stated as a condition on the tail of a distribution is equivalent to the condition on the tail of a distribution arising from the hazard rate definition (Theorem 6). The gap implemented captures that the hazard rate must decrease by at least α in order for a distribution to be (α,ρ) heavy-tailed. See Section 3 and Appendix D.1 for detailed discussion.
2. **Show that $\tilde{S}$ is close to S when there are "enough" buckets.** If we can estimate the numerator and the denominator of the proxy quantity accurately to an additive error, then we can estimate the proxy quantity correctly to an additive error if the proxy quantity is small; if it is big, then we know we are in the light-tailed region (Lemma 11). Indeed, we can accurately estimate the numerator and denominator accurately to an additive error while approximating derivatives with difference quotients due to the Lipschitzness (Fact 3.3, Corollary 8).

[8] This can be increased to probability $1-\delta$ by repeating the algorithm $\log 1/\delta$ times using the standard amplification technique.

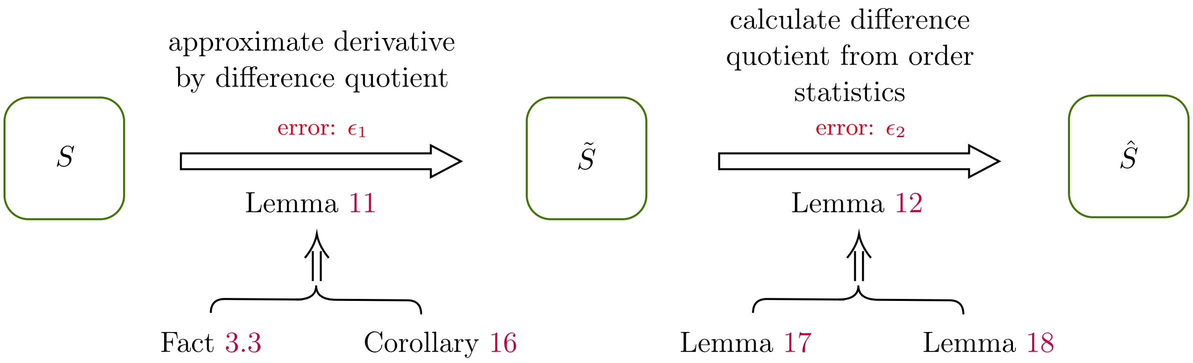


Figure 2: The proxy quantity S is a probe for the tail weight of the distribution based on the hazard rate. By approximating the derivatives in the proxy by difference quotients, we derive $\hat{S}$ while incurring error ϵ_1. An analysis of how this error is incurred and what its value is given in the statements noted.

3. **Show that $\hat{S}$ calculated from order statistics is close to $\tilde{S}$ when there are enough samples.** If we get a good multiplicative approximation of the lengths of the two buckets we are interested in, then we have a good multiplicative approximation to the statistic (Lemma 12). We have multiplicative approximations for the lengths of the buckets because we can show that the order statistics used to estimate the endpoints concentrate well (Lemma 17), which gives us an additive estimate for the error, and then we translate this to a multiplicative error (Lemma 18).

4. **Determine how to limit the errors incurred in each step by explicitly calculating the number of buckets and number of samples.** In order to do this, we note that after incurring both the derivative approximation error and the sampling error, the test statistic $\hat{S}$ must still be on the "correct" side, that is, the same side of the threshold as S. The threshold is halfway between the lower bound for light-tailed distributions and the upper bound for (α, ρ)-heavy-tailed distributions. We ensure that we split the distribution into sufficiently many buckets that no more than a quarter of the gap is crossed due to approximation error. Further, we draw enough samples that no more a quarter of the gap is crossed due to sampling error. Thus, even when both errors are incurred, the test statistic remains on the correct side of the threshold.

Part 1/4: S is an accurate proxy.

The proxy quantity derived in Theorem 6 considered with respect to the threshold gives a test which accurately determines whether a set of samples came from a light-tailed distribution or a distribution that is (α, ρ)-heavy-tailed. An (α, ρ)-heavy-tailed distribution has hazard rate decreasing at least by α over a region of the PDF with probability mass ρ, which gives us an expression for how far the statistic must be from the original threshold $1 - i/k$ in order for the hazard rate to be decreasing by at least α. Further, if the region of mass ρ lies in at least two buckets, then the proxy quantity will detect it. Thus, if a distribution is light-tailed, then *all* $k - 3$ of the calculated proxy quantities calculated will lie above $1 - i/k$; if even one of the proxies lies below $1 - i/k - \text{gap}$, then the distribution is (α, ρ)-heavy-tailed.

More formal details for this can be found in Section 3 and Appendix D.1.

Part 2/4: $\tilde{S}$ is close to S.

In the next step, we show that approximating the derivatives in the proxy quantity by the respective difference quotients causes the test statistic to incur bounded error. In particular, we show how the error in the statistic is related to the number of buckets (which is directly related to how well we approximate derivatives; in the limit as $k \to \infty$, we recover the exact derivative). To this end, we first show that if we can estimate the numerator and the denominator of the proxy quantity to known additive errors, we can estimate the proxy quantity within an additive error, and we give the relationship between the two errors (Lemma 11). Then, to show that we can estimate the numerator and denominator of $\tilde{S}$ well, we show that the difference quotient approximation to the derivative incurs bounded error under stated Lipschitzness conditions (Fact 3.3, Corollary 8).

Recall that the proxy we are using is

$$S = N/D = \left(\frac{d}{dx}F^{-1}(x)\right) / \left(\frac{d^2}{dx^2}F^{-1}(x)\right),$$

which is being approximated by

$$\tilde{S} = \tilde{N}/\tilde{D}; \quad \tilde{N} := \frac{F^{-1}\left(x + \frac{1}{k^2}\right) - F^{-1}\left(x\right)}{1/k^2} \tag{17}$$

$$\tilde{D} := \frac{\left(F^{-1}\left(x + \frac{1}{k} + \frac{1}{k^2}\right) - F^{-1}\left(x + \frac{1}{k^2}\right)\right) - \left(F^{-1}\left(x + \frac{1}{k^2}\right) - F^{-1}\left(x\right)\right)}{1/k^3}. \tag{18}$$

We show that if the error incurred in $\tilde{N}, \tilde{D}$ in approximating the derivatives by difference quotients has value ϵ', then we can estimate the value of the proxy quantity within an additive error of $\epsilon := 6\beta\epsilon'$, or the value of the proxy quantity is greater than 1.

Lemma 11 (Additive Bound for $\tilde{S}$) *Given a parameter $\epsilon < 1$, if $|\tilde{N} - N|$ and $|\tilde{D} - D|$ are at most $\epsilon' := \epsilon/(6\beta)$, then either $|\tilde{S} - S| < \epsilon$ or both $S, \tilde{S}$ are at least one.*

Proof We start by noting that since $N = \frac{1}{f(F^{-1}(x))}$ and the PDF is bounded by β, $N > 1/\beta = 6\epsilon'/\epsilon > 6\epsilon'$. The last inequality is due to the assumption that $\epsilon < 1$.

First, assume that $S \geq \tilde{S}$. We consider the lowest possible value of $\tilde{S} = \tilde{N}/\tilde{D}$ and take the difference from the true value of S.

$$S - \tilde{S} \leq \frac{N}{D} - \frac{N - \epsilon'}{D + \epsilon'} = \frac{\epsilon' N + \epsilon' D}{D^2 + \epsilon' D} \leq \frac{\epsilon'(N + D)}{D^2} = \frac{\epsilon' S(S+1)}{N}. \tag{19}$$

We analyze the difference in two different cases:

- **Case 1: $S \leq 2$.** In this case, the above difference is at most $\epsilon' \cdot 2 \cdot 3/N < \epsilon$ since N is at least $6\epsilon'/\epsilon$.
- **Case 2: $S > 2$.** In this case, we have $N/2 > D$ which implies

$$\frac{N - \epsilon'}{D + \epsilon'} > \frac{N - \epsilon'}{N/2 + \epsilon'} = 1 + \frac{N/2 - 2\epsilon'}{N/2 + \epsilon'} > 1.$$

The last inequality is due to the fact that $N > 6\epsilon'$.

Next, we assume that $S < \tilde{S}$. We consider the largest possible value of $\tilde{S}$ and take the difference from the true value of S as before:

$$\tilde{S} - S = \frac{N + \epsilon'}{D - \epsilon'} - \frac{N}{D} = \epsilon' \frac{N + D}{D^2 - \epsilon' D} = \epsilon' \frac{S + 1}{D - \epsilon'} \,. \tag{20}$$

If $S > 1$, both S and $\tilde{S}$ are larger than one, and we are done. Otherwise, $D > N > 1/\beta$, which implies that $D - \epsilon' > 1/\beta - 1/(6\beta) > 1/(2\beta)$, and therefore, we can conclude that:

$$\tilde{S} - S \le \epsilon' \frac{S + 1}{D - \epsilon'} < 4\epsilon'\beta < 6\epsilon'\beta = \epsilon \,.$$

Thus, the statement of the lemma is concluded. ■

Now, we must show that we can estimate N, D by $\tilde{N}, \tilde{D}$ respectively within additive error ϵ'. First, we bound the additive error arising from the gradient approximation. When the derivative of a function g is $B-Lipschitz$, and the derivative $g'(y)$ is monotone, then approximating $g'(y)$ by the difference quotient $\frac{g(y+\Delta y)-g(y)}{\Delta y}$ incurs no more than $B\Delta y$ additive error. As discussed in 3.3, this allows us to determine the error bound for the difference quotient approximations of derivatives in the proxy quantity.

Lemma 8 *When the derivative of a function g is B_1-Lipschitz, the second derivative is B_2-Lipschitz, the derivative $g'(y)$ and the second derivative $g''(y)$ are both monotone, then approximating $g''(y)$ by:*

1. *estimating $\tilde{g}'(y) = \frac{g(y+\Delta y_1)-g(y)}{\Delta y_1}$, and $\tilde{g}'(y + \Delta y') = \frac{g(y+\Delta y_2+\Delta y_1)-g(y+\Delta y_2)}{\Delta y_1}$,*
2. *and estimating $\tilde{g}''(y) = \frac{\tilde{g}'(y+\Delta y_2)-\tilde{g}'(y)}{\Delta y_2}$.*

incurs no more than $2B_1 \frac{\Delta y_1}{\Delta y_2} + B_2 \Delta y_2$ additive error.

For reasons discussed in Section 3.3, we set $\Delta y_1 = 1/k^2$, and $\Delta y_2 = 1/k$, meaning that the incurred error in the second derivative is $\frac{B_1+B_2}{k}$. We set this to be less than $\epsilon' = \frac{\epsilon}{6\beta}$ and find the value of k for which this happens: $k > \frac{6\beta(2B_1+B_2)}{\epsilon}$. Thus, we have the following corollary that specifies the number of buckets necessary to incur no more than ϵ error in estimating the proxy quantity when it is not greater than 1.

Corollary 16 *If $k > 6\beta\frac{2B_1+B_2}{\epsilon}$, and the estimate $\tilde{S}$ of the statistic S is computed as described above, then either both $S, \tilde{S} > 1$, or $|\tilde{S} - S| < \epsilon$.*

PART 3/4: $\hat{S}$ (calculated from order statistics) is close to $\tilde{S}$.

We must next show that we can approximate $\tilde{S}$ as defined in the previous section by the order statistics of a set of samples. First, we prove that if we can estimate the lengths of intervals accurately up to multiplicative error, we can estimate the test statistic accurately up to multiplicative error (Lemma 12). Then, we show that we can achieve good multiplicative estimates of the lengths of the intervals using the following stages:

1. We construct a map between order statistics of samples from a uniform distribution, denoted $Y_{(i)}$, (which we know concentrate well and in fact are *sub-Gaussian*) and order statistics of samples from an arbitrary distribution (denoted $X_{(i)}$).
2. Concentration of $Y_{(i)}$ implies concentration of $X_{(i)}$ up to accounting for the gap between $F(\mathbb{E}[X_{(i)}])$ and $\mathbb{E}[F(X_{(i)})]$ (Lemma 17).

3. The additive error between realization of order statistic and the quantity in the test statistic (which we get from the sub-Gaussianity parameter) can be converted to a multiplicative error in terms of the length of the bucket (Lemma 18).

One can view $\tilde{S}$ in terms of two quantities:

$$\tilde{L}_1 \coloneqq \frac{F^{-1}(y + 1/k^2) - F^{-1}(y)}{1/k^2}$$

$$\tilde{L}_2 \coloneqq \frac{F^{-1}(y + 1/k + 1/k^2) - F^{-1}(y + 1/k)}{1/k^2}\,.$$

In terms of the previous notation,

$$\tilde{N} = \tilde{L}_1 \qquad \tilde{D} = \frac{\tilde{L}_2 - \tilde{L}_1}{1/k}\,,$$

$$\text{giving us that: } \tilde{S} = \frac{\tilde{L}_1}{k(\tilde{L}_2 - \tilde{L}_1)} = \frac{\tilde{N}}{\tilde{D}}\,.$$

In the following lemma, we show that if we estimate $\tilde{L}_1$ and $\tilde{L}_2$ accurately up to a multiplicative factor, and obtain $\hat{L}_1$ and $\hat{L}_2$, then $\tilde{S}$ can be approximated by

$$\hat{S} \coloneqq \frac{\hat{L}_1}{k \cdot \left(\hat{L}_2 - \hat{L}_1\right)}\,.$$

Lemma 12 (Multiplicative Bound for $\hat{S}$) *Suppose we have $\hat{L}_1$ and $\hat{L}_2$, the estimates of $\tilde{L}_1$ and $\tilde{L}_2$ with a multiplicative factor of $\epsilon' = \min\left(\Theta\left(\epsilon/((1+\epsilon)\cdot k)\right), \Theta\left(1/k^2\right)\right)$. Then, one of the following cases holds:*

1. *$\hat{S} > 1 - 2/k$ and $\tilde{S} > 1$, implying they come from a light-tailed region of the distribution.*
2. *$(1-\epsilon)\cdot\tilde{S} \leq \hat{S} \leq (1+\epsilon)\cdot\tilde{S}$.*

Proof We analyze two separate cases.

Case 1: $\boldsymbol{\tilde{L}_2 - \tilde{L}_1 \leq \frac{1}{k+1}\tilde{L}_2}$. Since $\tilde{L}_2 - \tilde{L}_1 > 0$ (due to montonicity of f), we have that $1 \geq \tilde{L}_1/\tilde{L}_2 \geq k/(k+1)$. Thus, we have that:

$$\tilde{S} = \frac{\tilde{L}_1}{k(\tilde{L}_2 - \tilde{L}_1)} \geq \frac{\tilde{L}_1}{\frac{k}{k+1}\tilde{L}_2} \geq 1\,. \tag{21}$$

In turn, since we have $(1-\epsilon')\tilde{L}_1 \leq \hat{L}_1 \leq (1+\epsilon')\tilde{L}_1$ and similarly for $\hat{L}_2$:

$$\hat{S} = \frac{\hat{L}_1}{k(\hat{L}_2 - \hat{L}_1)} \geq \frac{(1-\epsilon')\tilde{L}_1}{k((1-\epsilon')(\tilde{L}_2 - \tilde{L}_1) + 2\epsilon'\tilde{L}_2)} \geq \frac{(1-\epsilon')\frac{k}{k+1}\tilde{L}_2}{k(\frac{1-\epsilon'}{k+1} + 2\epsilon')\tilde{L}_2} = \frac{1-\epsilon'}{1-\epsilon'+2\epsilon'(k+1)} \tag{22}$$

$$= 1 - \frac{2\epsilon'(k+1)}{1-\epsilon'+2\epsilon' k}\,. \tag{23}$$

When $\epsilon' \leq 1/(k^2-k+1)$, $\hat{S} \geq 1 - 2/k$. Since $\tilde{S} > 1$ in this case, as well, the test statistic falls clearly in light-tailed territory.

This completes the first case.

Case 2: $\boldsymbol{\tilde{L}_2 - \tilde{L}_1 > \frac{1}{k+1}\tilde{L}_2}$, which can be rearranged to give $\tilde{L}_2 - \tilde{L}_1 \leq \tilde{L}_2 < (k+1)(\tilde{L}_2 - \tilde{L}_1)$. We first show that $\hat{S}$ cannot be too small. In fact, we have:

$$\hat{S} \geq \frac{(1-\epsilon')\tilde{L}_1}{k((1+\epsilon')\tilde{L}_2 - (1-\epsilon')\tilde{L}_1)} = \frac{(1-\epsilon')\tilde{L}_1}{k((1-\epsilon')(\tilde{L}_2 - \tilde{L}_1) + 2\epsilon'\tilde{L}_2)} \tag{24}$$

$$\geq \frac{(1-\epsilon')\tilde{L}_1}{k((1-\epsilon')(\tilde{L}_2 - \tilde{L}_1) + 2\epsilon'(k+1)(\tilde{L}_2 - \tilde{L}_1))} \geq \frac{1-\epsilon'}{1+\epsilon'(2k+1)}\tilde{S} \geq (1-\epsilon)\tilde{S}\,. \tag{25}$$

where the last inequality holds if we set $\epsilon' \leq \epsilon/(1 + (1-\epsilon)(2k+1))$. We can also find an upper bound for $\hat{S}$:

$$\hat{S} \leq \frac{(1+\epsilon')\tilde{L}_1}{k\left((1-\epsilon')\tilde{L}_2 - (1+\epsilon')\tilde{L}_1\right)} \leq \frac{(1+\epsilon')\tilde{L}_1}{k\left((1+\epsilon')(\tilde{L}_2 - \tilde{L}_1) - 2\epsilon'\tilde{L}_2\right)} \tag{26}$$

$$\leq \frac{(1+\epsilon')\tilde{L}_1}{k\left((1+\epsilon')(\tilde{L}_2 - \tilde{L}_1) - 2\epsilon'(k+1)(\tilde{L}_2 - \tilde{L}_1)\right)} \leq \frac{(1+\epsilon')}{1-\epsilon'(2k+1)}\tilde{S} \leq (1+\epsilon)\tilde{S}\,. \tag{27}$$

To satisfy this, we can set $\epsilon' \leq \epsilon/(1 + (1+\epsilon)(2k+1))$. Note that this is always smaller than the condition for ϵ' derived using the lower bound.

Thus, by setting $\epsilon' = \min\left(\frac{\epsilon}{1+(1+\epsilon)(2k+1)}, \frac{1}{k^2}\right)$, we get that either we are in Case (1), where $\tilde{S} > 1, \hat{S} > 1 - 2/k$ or we can get an $1 \pm \epsilon$ multiplicative approximation. ■

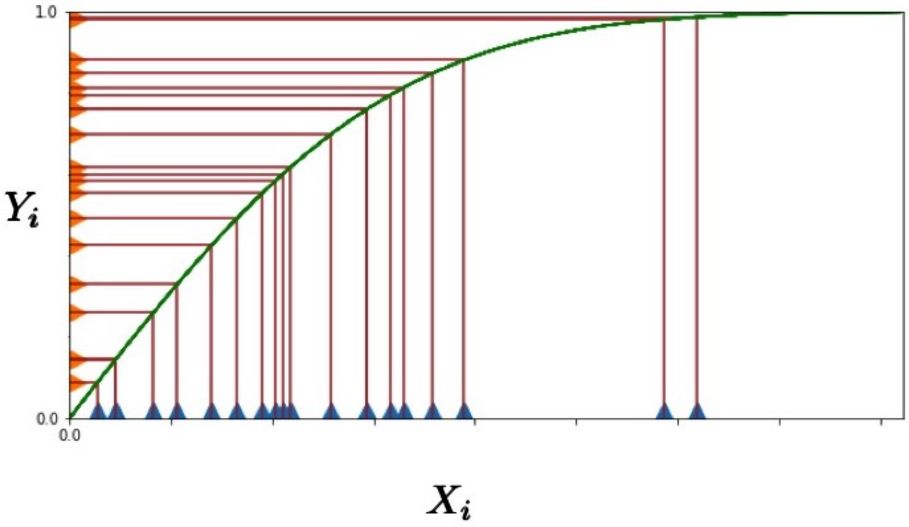


Figure 3: In order to analyze the concentration of order statistic, we construct a map between samples drawn from a uniform distribution (orange points on vertical axis) and samples drawn from an arbitrary distribution (blue points on horizontal axis) with CDF F (green curve). This map preserves several important properties as discussed in the text of the paper.

Now that we know that we can accurately estimate $\tilde{S}$ by $\hat{S}$ if we have good estimates of the values of $\hat{L}_1, \hat{L}_2$, we will next show that we are able to estimate $\hat{L}_1, \hat{L}_2$ accurately, proceeding in 3 stages as discussed above.

1. Mapping to uniform distribution To prove the concentration of the $X_{(i)}$s, we transform the samples from f to samples from a uniform distribution over $[0, 1]$ (Figure 3). The mapping is selected in such a way that it does not change the order of the elements. This fact implies that the order statistic with rank i, $X_{(i)}$ will be mapped to the order statistic with the same rank from the samples from the uniform distribution. Then, we show the concentration of order statistics according to the uniform distribution to establish the result.

Formally, we map every sample X to $Y = F(X)$. Since we assumed that f is well-behaved, this mapping is a bijection. If we draw a random X from f, it is not too hard to see that $Y = F(X)$ comes from a uniform distribution over $[0, 1]$. In particular, for every $y \in [0, 1]$, we have:

$$\mathbf{Pr}_Y[Y \leq y] = \mathbf{Pr}_X\big[X \leq F^{-1}(y)\big] = F(F^{-1}(y)) = y\,.$$

By the monotonicity of F, it is not hard to see that if we transform a set of n samples from f, then $X_{(i)}$ will be mapped to the i-th element in the sorted list of the Y's, denoted by $Y_{(i)}$. On the other hand, one can view $Y_{(i)}$ as the i-th order statistic among n samples from a uniform distribution over $[0, 1]$. In Marchal et al. (2017), Marchal and Arbel have proved that $Y_{(i)} - \mathbf{E}\big[Y_{(i)}\big]$ is a sub-Gaussian random variable. We have used this fact to show the concentration of $X_{(i)}$'s around $F^{-1}(\mathbf{E}\big[Y_{(i)}\big]) = F^{-1}(i/(n+1))$. With appropriate choice i and n, we can show the order statistic concentrates around the endpoint of the buckets as desired.

2. Concentration of $X_{(i)}$ Given this mapping, we can show that the order statistics from an arbitrary distribution concentration around $F^{-1}(i/(n+1))$. To do this, we use the sub-Gaussianity

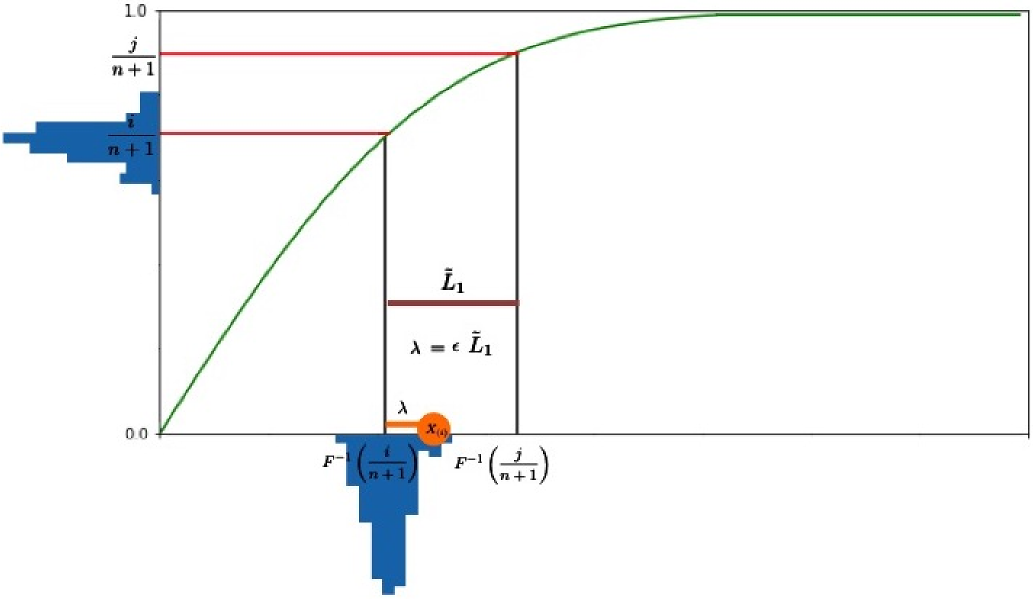

1.0
$\frac{j}{n+1}$
$\frac{i}{n+1}$
$\tilde{L}_1$
$\lambda = \epsilon\, \tilde{L}_1$
λ
$X_{(i)}$
0.0
$F^{-1}\left(\frac{i}{n+1}\right)$
$F^{-1}\left(\frac{j}{n+1}\right)$

Figure 4: In Lemma 17, we show that since $Y_{(i)}$ concentrates around $i/(n+1)$ (as evidenced by the blue histogram on the vertical axis), $X_{(i)}$ must also concentrate around $F^{-1}(i/(n+1))$ (as evidenced by the blue histogram on the horizontal axis). Then, in Lemma 18, we show that the distance that the order statistic (orange circle) falls from $F^{-1}(i/(n+1))$ (orange bar labeled λ) can converted to a multplicative error in terms of the length of the interval $\hat{L}_1$ (maroon bar).

of $Y_{(i)}$ around $i/(n+1)$, transform this to a statement about $X_{(i)}$ related to $F^{-1}(i/(n+1))$, which is the endpoint of a bucket, and analyze terms separately. To this end, we present Lemma 17 and its proof.

Lemma 17 *Suppose f is a well-behaved distribution. For $\epsilon \leq \frac{B_1 f\left(F^{-1}\left(\frac{i}{n+1}\right)\right)+k}{k^2 f\left(F^{-1}\left(\frac{i}{n+1}\right)\right)}$, we have:*

$$\mathbf{Pr}\left[\left|X_{(i)} - F^{-1}\left(\frac{i}{n+1}\right)\right| \geq \epsilon\right] \leq 2\exp\left(-\epsilon^2/\lambda^2\right)\,, \quad \lambda := \frac{B_1}{2\,(n+2)\,\left(\sqrt{4B_1+1}-1\right)\, f(F^{-1}\left(\frac{i}{n+1}\right))}\,.$$

Proof As we explained earlier, we map the samples from f, $x_1, x_2, \ldots, x_n$, to $y_1, y_2, \ldots, y_n$ where $y_i = F(x_i)$. We have shown that one can view y_i's as random samples from the uniform distribution over $[0,1]$. Since the mapping preserve the order, the i-th order statistic $X_{(i)}$ is mapped to the i-th order statistic among y_i's, which we denote by $Y_{(i)}$. Now, we show the concentration of the order statistics from the uniform distribution. Note that the distribution over the order statistic is a beta distribution with parameters $\alpha := i$ and $\beta := n-i+1$:

$$\mathbf{Pr}\left[Y_{(i)} = y\right] = n \cdot \binom{n-1}{i-1} \cdot y^{i-1} \cdot (1-y)^{n-i}\,.$$

Marchal and Arbel have shown that a (centered) random variable drawn from the beta distribution is a sub-Gaussian random variable with parameter $\sigma^2 = 1/4(\alpha+\beta+1) = 1/4(n+2)$ (Theorem 2.1 in Marchal et al. (2017)). In other words, for any $\lambda \in \mathbb{R}$, we have:

$$\mathbf{E}\big[\exp\left(\lambda\cdot\left(Y_{(i)}-\mathbf{E}\big[Y_{(i)}\big]\right)\right)\big]\le\exp\left(\frac{\lambda^2}{32\,(n+2)^2}\right).$$

Then, using equivalent definitions of sub-Gaussian random variables (See Proposition 2.5.2 in Vershynin (2018).), we obtain:

$$\mathbf{Pr}\big[\big|Y_{(i)}-\mathbf{E}\big[Y_{(i)}\big]\big|\ge t\big]\le 2\,\exp\left(-16\,(n+2)^2\,t^2\right)\qquad\forall t\ge 0\,.$$

In the next step, using the Lipschitzness of F^{-1}, we relate the probability that X deviates from $F^{-1}\left(\mathbf{E}\big[F\left(X_{(i)}\right)\big]\right)$ to the probability that $Y_{(i)}$ deviates from its expectation. Then, we use the sub-Gaussianity of $Y_{(i)}$'s to obtain the desired bound. Let T denote $F^{-1}\left(\mathbf{E}\big[F\left(X_{(i)}\right)\big]\right)=F^{-1}\left(\mathbf{E}\big[Y_{(i)}\big]\right)=F^{-1}(i/n+1)$. Our goal here is to find a bound of the following form for some parameter λ which we determine later.

$$\mathbf{Pr}\left[\left|X_{(i)}-F^{-1}\left(\frac{i}{n+1}\right)\right|\ge\epsilon\right]\le 2\exp\left(-\epsilon^2/\lambda^2\right).$$

Using the definition of the mapping, we have:

$$\mathbf{Pr}\left[\left|X_{(i)}-F^{-1}\left(\frac{i}{n+1}\right)\right|\ge\epsilon\right]=\mathbf{Pr}\big[\big|F^{-1}\left(Y_{(i)}\right)-F^{-1}\left(\mathbf{E}\big[Y_{(i)}\big]\right)\big|\ge\epsilon\big]\tag{28}$$

Using the mean value theorem, we know that there exists a y between $Y_{(i)}$ and $\mathbf{E}\big[Y_{(i)}\big]$ such that:

$$F^{-1\prime}(y)\cdot\left(Y_{(i)}-\mathbf{E}\big[Y_{(i)}\big]\right)=F^{-1}\left(Y_{(i)}\right)-F^{-1}\left(\mathbf{E}\big[Y_{(i)}\big]\right).$$

To bound the probability in eq. (28) via the above identity, we need to find an upper bound on $F^{-1\prime}(y)$. Note that since F is concave, $F^{-1\prime}(y)$ is equal to $1/f(F^{-1}(y))$, and it is an increasing function. Now, we consider a few different cases based on the interval that y belongs to. More precisely, for some parameter $0<\delta<\Theta(1/k)$ which we define later, we define the following events:

- E_1 := indicates the event where $Y_{(i)}-\mathbf{E}\big[Y_{(i)}\big]\le 0$.
- E_2 := indicates the event where $0<Y_{(i)}-\mathbf{E}\big[Y_{(i)}\big]<\delta$.
- E_3 := indicates the event where $Y_{(i)}-\mathbf{E}\big[Y_{(i)}\big]\ge\delta$.

Then, we rewrite the probability in eq. (28) in terms of conditional probabilities as follows:

$$\begin{aligned}\mathbf{Pr}\left[\left|X_{(i)}-F^{-1}\left(\frac{i}{n+1}\right)\right|\ge\epsilon\right]&=\mathbf{Pr}\left[\left|F^{-1\prime}(y)\cdot\left(Y_{(i)}-\mathbf{E}\big[Y_{(i)}\big]\right)\right|\ge\epsilon\right]\\&=\mathbf{Pr}\left[\left|F^{-1\prime}(y)\cdot\left(Y_{(i)}-\mathbf{E}\big[Y_{(i)}\big]\right)\right|\ge\epsilon\;\middle|\;E_1\right]\cdot\mathbf{Pr}[E_1]\\&+\mathbf{Pr}\left[\left|F^{-1\prime}(y)\cdot\left(Y_{(i)}-\mathbf{E}\big[Y_{(i)}\big]\right)\right|\ge\epsilon\;\middle|\;E_2\right]\cdot\mathbf{Pr}[E_2]\\&+\mathbf{Pr}\left[\left|F^{-1\prime}(y)\cdot\left(Y_{(i)}-\mathbf{E}\big[Y_{(i)}\big]\right)\right|\ge\epsilon\;\middle|\;E_3\right]\cdot\mathbf{Pr}[E_3]\,.\end{aligned}\tag{29}$$

Below, we bound each of the above terms:

1. **First term:** In the case that E_1 holds, y is at most $\mathbf{E}\big[Y_{(i)}\big]$. Using the monotonicity of f, we have ${F^{-1}}'(y)$ is at most $1/f\left(F^{-1}\left(\mathbf{E}\big[Y_{(i)}\big]\right)\right) = 1/f(F^{-1}(i/n+1))$. This bound implies that

$$\begin{aligned}
&\mathbf{Pr}\Big[\,\Big|{F^{-1}}'(y)\cdot\big(Y_{(i)}-\mathbf{E}\big[Y_{(i)}\big]\big)\Big|\geq\epsilon\ \Big|\ E_1\Big]\cdot\mathbf{Pr}[E_1]\\
&\quad\leq \mathbf{Pr}\left[\,Y_{(i)}-\mathbf{E}\big[Y_{(i)}\big]\leq -\epsilon\cdot f\left(F^{-1}\left(\frac{i}{n+1}\right)\right)\ \middle|\ E_1\right]\cdot\mathbf{Pr}[E_1]\\
&\quad= \mathbf{Pr}\left[Y_{(i)}-\mathbf{E}\big[Y_{(i)}\big]\leq -\epsilon\cdot f\left(F^{-1}\left(\frac{i}{n+1}\right)\right)\right]\\
&\quad- \mathbf{Pr}\left[\,Y_{(i)}-\mathbf{E}\big[Y_{(i)}\big]\leq -\epsilon\cdot f\left(F^{-1}\left(\frac{i}{n+1}\right)\right)\ \middle|\ \overline{E_1}\right]\cdot\mathbf{Pr}\big[\overline{E_1}\big]\,,
\end{aligned}$$

where the bar in $\overline{E_1}$ indicates the complement event of E_1. Note that $Y_{(i)}-\mathbf{E}\big[Y_{(i)}\big]$ cannot be smaller than a negative quantity if E_1 does not happen. Thus, the last term above is zero, and we have:

$$\begin{aligned}
&\mathbf{Pr}\Big[\,\Big|{F^{-1}}'(y)\cdot\big(Y_{(i)}-\mathbf{E}\big[Y_{(i)}\big]\big)\Big|\geq\epsilon\ \Big|\ E_1\Big]\cdot\mathbf{Pr}[E_1]\\
&\qquad\qquad\leq \mathbf{Pr}\left[Y_{(i)}-\mathbf{E}\big[Y_{(i)}\big]\leq -\epsilon\cdot f\left(F^{-1}\left(\frac{i}{n+1}\right)\right)\right]
\end{aligned}\tag{30}$$

2. **Second term:** In the case that E_2 holds, y is at most $\mathbf{E}\big[Y_{(i)}\big]+\delta$. Note that since δ is smaller than $1/k$, and we assume that the function is Lipschitz in the bucket which starts at $\mathbf{E}\big[Y_{(i)}\big]$, then we can use the Lipschitzness assumption as follows:

$${F^{-1}}'(y)\leq {F^{-1}}'\left(\mathbf{E}\big[Y_{(i)}+\delta\big]\right)\leq 1/f\left(F^{-1}\left(\frac{i}{n+1}\right)\right)+\delta\,B_1\,.$$

Now, we have:

$$\begin{aligned}
&\mathbf{Pr}\Big[\,\Big|{F^{-1}}'(y)\cdot\big(Y_{(i)}-\mathbf{E}\big[Y_{(i)}\big]\big)\Big|\geq\epsilon\ \Big|\ E_2\Big]\cdot\mathbf{Pr}[E_2]\\
&\quad\leq \mathbf{Pr}\left[\,Y_{(i)}-\mathbf{E}\big[Y_{(i)}\big]\geq \frac{\epsilon}{\frac{1}{f\left(F^{-1}\left(\frac{i}{n+1}\right)\right)}+\delta B_1}\ \middle|\ E_2\right]\cdot\mathbf{Pr}[E_2]\,.
\end{aligned}\tag{31}$$

We will use the above bound later and combine it with the bound for the third term.

3. **Third term:** In the case where E_3 holds, we do not have an upper bound for ${F^{-1}}'(y)$. However, we exploit the fact that this case happens with a small probability.

$$\mathbf{Pr}\Big[\,\Big|{F^{-1}}'(y)\cdot\big(Y_{(i)}-\mathbf{E}\big[Y_{(i)}\big]\big)\Big|\geq\epsilon\ \Big|\ E_3\Big]\cdot\mathbf{Pr}[E_3]\leq\mathbf{Pr}[E_3]=\mathbf{Pr}\big[Y_{(i)}-\mathbf{E}\big[Y_{(i)}\big]>\delta\big]\,.\tag{32}$$

4. **Combining the bounds for the last two terms:** Now we combine the two bounds above in eq. (31) and eq. (32). Note that if δ is smaller than $\epsilon/(1/f(F^{-1}\left(\frac{i}{n+1}\right))+\delta B_1)$, then the right hand side in eq. (31) is equal to:

$$\mathbf{Pr}\left[\,Y_{(i)}-\mathbf{E}\big[Y_{(i)}\big]\geq \frac{\epsilon}{\frac{1}{f\left(F^{-1}\left(\frac{i}{n+1}\right)\right)}+\delta B_1}\geq\delta\ \middle|\ 0<Y_{(i)}-\mathbf{E}\big[Y_{(i)}\big]<\delta\right]\cdot\mathbf{Pr}[E_2]=0\,.$$

Thus, when delta is small, after combining the last two bounds, we get:

$$\begin{aligned}
&\mathbf{Pr}\Big[\Big|{F^{-1}}'(y)\cdot\big(Y_{(i)}-\mathbf{E}\big[Y_{(i)}\big]\big)\Big|\geq\epsilon\;\Big|\;E_2\Big]\cdot\mathbf{Pr}\big[E_2\big]\\
&\quad+\mathbf{Pr}\Big[\Big|{F^{-1}}'(y)\cdot\big(Y_{(i)}-\mathbf{E}\big[Y_{(i)}\big]\big)\Big|\geq\epsilon\;\Big|\;E_3\Big]\cdot\mathbf{Pr}\big[E_3\big]\\
&\quad\leq\mathbf{Pr}\big[Y_{(i)}-\mathbf{E}\big[Y_{(i)}\big]\geq\delta\big]
\end{aligned}$$

On the other hand, if δ is at least $\epsilon/(1/f(F^{-1}\left(\frac{i}{n+1}\right))+\delta B_1)$, then by combining eq. (31) and eq. (32), we have:

$$\begin{aligned}
&\mathbf{Pr}\Big[\Big|{F^{-1}}'(y)\cdot\big(Y_{(i)}-\mathbf{E}\big[Y_{(i)}\big]\big)\Big|\geq\epsilon\;\Big|\;E_2\Big]\cdot\mathbf{Pr}\big[E_2\big]\\
&\quad+\mathbf{Pr}\Big[\Big|{F^{-1}}'(y)\cdot\big(Y_{(i)}-\mathbf{E}\big[Y_{(i)}\big]\big)\Big|\geq\epsilon\;\Big|\;E_3\Big]\cdot\mathbf{Pr}\big[E_3\big]\\
&\quad\leq\mathbf{Pr}\left[Y_{(i)}-\mathbf{E}\big[Y_{(i)}\big]\geq\frac{\epsilon}{\frac{1}{f(F^{-1}(\frac{i}{n+1}))}+\delta B_1}\;\middle|\;E_2\right]\cdot\mathbf{Pr}\big[E_2\big]+\mathbf{Pr}\big[E_3\big]\\
&\quad\leq\mathbf{Pr}\left[Y_{(i)}-\mathbf{E}\big[Y_{(i)}\big]\geq\frac{\epsilon}{\frac{1}{f(F^{-1}(\frac{i}{n+1}))}+\delta B_1}\;\middle|\;E_2\right]\cdot\mathbf{Pr}\big[E_2\big]\\
&\quad\underbrace{+\mathbf{Pr}\left[Y_{(i)}-\mathbf{E}\big[Y_{(i)}\big]\geq\frac{\epsilon}{\frac{1}{f(F^{-1}(\frac{i}{n+1}))}+\delta B_1}\;\middle|\;E_3\right]}_{=1}\cdot\mathbf{Pr}\big[E_3\big]\\
&\quad=\mathbf{Pr}\left[Y_{(i)}-\mathbf{E}\big[Y_{(i)}\big]\geq\frac{\epsilon}{\frac{1}{f(F^{-1}(\frac{i}{n+1}))}+\delta B_1}\;\middle|\;Y_{(i)}-\mathbf{E}\big[Y_{(i)}\big]>0\right]\cdot\mathbf{Pr}\big[\mathbf{E}\big[Y_{(i)}\big]-Y_{(i)}<0\big]\\
&\quad=\mathbf{Pr}\left[Y_{(i)}-\mathbf{E}\big[Y_{(i)}\big]\geq\frac{\epsilon}{\frac{1}{f(F^{-1}(\frac{i}{n+1}))}+\delta B_1}\right].
\end{aligned}$$

Note that the first probability in the last line above is exactly one, since we are conditioning on the fact that $Y_{(i)}-\mathbf{E}\big[Y_{(i)}\big]$ is at least δ. Now, we get:

$$\begin{aligned}
&\mathbf{Pr}\Big[\Big|{F^{-1}}'(y)\cdot\big(Y_{(i)}-\mathbf{E}\big[Y_{(i)}\big]\big)\Big|\geq\epsilon\;\Big|\;E_2\Big]\cdot\mathbf{Pr}\big[E_2\big]\\
&\quad+\mathbf{Pr}\Big[\Big|{F^{-1}}'(y)\cdot\big(Y_{(i)}-\mathbf{E}\big[Y_{(i)}\big]\big)\Big|\geq\epsilon\;\Big|\;E_3\Big]\cdot\mathbf{Pr}\big[E_3\big]\\
&\quad=\mathbf{Pr}\left[Y_{(i)}-\mathbf{E}\big[Y_{(i)}\big]\geq\frac{\epsilon}{\frac{1}{f(F^{-1}(\frac{i}{n+1}))}+\delta B_1}\;\middle|\;\mathbf{E}\big[Y_{(i)}\big]-Y_{(i)}<0\right]\cdot\mathbf{Pr}\big[\mathbf{E}\big[Y_{(i)}\big]-Y_{(i)}<0\big]\\
&\quad=\mathbf{Pr}\left[Y_{(i)}-\mathbf{E}\big[Y_{(i)}\big]\geq\frac{\epsilon}{\frac{1}{f(F^{-1}(\frac{i}{n+1}))}+\delta B_1}\right].
\end{aligned}$$

Now, to obtain the best bound we solve for $\delta=\epsilon/(1/f(F^{-1}\left(\frac{i}{n+1}\right))+\delta B_1)$, and set δ to the positive solution of this quantity. Note that since ϵ is bounded from above, it is not hard to show that δ is at most $1/k$ as required in the beginning. Then, we achieve:

$$\begin{aligned}
&\mathbf{Pr}\left[\,\left|{F^{-1}}'(y)\cdot\left(Y_{(i)}-\mathbf{E}\left[Y_{(i)}\right]\right)\right|\geq\epsilon\;\middle|\;E_2\right]\cdot\mathbf{Pr}\left[E_2\right]\\
&\quad+\mathbf{Pr}\left[\,\left|{F^{-1}}'(y)\cdot\left(Y_{(i)}-\mathbf{E}\left[Y_{(i)}\right]\right)\right|\geq\epsilon\;\middle|\;E_3\right]\cdot\mathbf{Pr}\left[E_3\right]\\
&\quad\leq\mathbf{Pr}\left[Y_{(i)}-\mathbf{E}\left[Y_{(i)}\right]\geq\delta\right]\\
&\quad\leq\mathbf{Pr}\left[Y_{(i)}-\mathbf{E}\left[Y_{(i)}\right]\geq\frac{\sqrt{4\,f\left(F^{-1}\left(\frac{i}{n+1}\right)\right)^2\,B_1\,\epsilon+1}-1}{2\,f\left(F^{-1}\left(\frac{i}{n+1}\right)\right)\,B_1}\right]\\
&\quad\leq\mathbf{Pr}\left[Y_{(i)}-\mathbf{E}\left[Y_{(i)}\right]\geq\frac{\left(\sqrt{4\,f\left(F^{-1}\left(\frac{i}{n+1}\right)\right)^2\,B_1+1}-1\right)\cdot\epsilon}{2\,f\left(F^{-1}\left(\frac{i}{n+1}\right)\right)\,B_1}\geq\frac{\sqrt{4\,B_1+1}-1}{2\,B_1}\cdot\epsilon\cdot f\left(F^{-1}\left(\frac{i}{n+1}\right)\right)\right].
\end{aligned}$$

Now, we put all the pieces in eq. (29) back together. We combine the above bound with the bound we have obtained earlier in eq. (30) and get the following:

$$\begin{aligned}
\mathbf{Pr}\left[\left|X_{(i)}-F^{-1}\left(\frac{i}{n+1}\right)\right|\geq\epsilon\right]&\leq\mathbf{Pr}\left[Y_{(i)}-\mathbf{E}\left[Y_{(i)}\right]\leq-\,\epsilon\cdot f\left(F^{-1}\left(\frac{i}{n+1}\right)\right)\right]\\
&\quad+\mathbf{Pr}\left[Y_{(i)}-\mathbf{E}\left[Y_{(i)}\right]\geq\frac{\sqrt{4\,B_1+1}-1}{2\,B_1}\cdot\epsilon\cdot f\left(F^{-1}\left(\frac{i}{n+1}\right)\right)\right]\\
&\leq\mathbf{Pr}\left[\left|Y_{(i)}-\mathbf{E}\left[Y_{(i)}\right]\right|\geq\frac{\sqrt{4\,B_1+1}-1}{2\,B_1}\cdot f\left(F^{-1}\left(\frac{i}{n+1}\right)\right)\cdot\epsilon\right],
\end{aligned}$$

where the last inequality is due to the fact that $\sqrt{4\,B_1+1}-1/(2B_1)$ is always smaller than one. Now, we use the sub-Gaussianity of $Y_{(i)}-\mathbf{E}\left[Y_{(i)}\right]$, and get

$$\begin{aligned}
\mathbf{Pr}\left[\left|X_{(i)}-F^{-1}\left(\frac{i}{n+1}\right)\right|\geq\epsilon\right]&\leq 2\,\exp\left(-\frac{16\,(n+2)^2\,\left(\sqrt{4\,B_1+1}-1\right)^2\,f\left(F^{-1}\left(\frac{i}{n+1}\right)\right)^2}{4\,B_1^2}\cdot\epsilon^2\right).\\
&\leq 2\exp\left(-\epsilon^2/\lambda^2\right)
\end{aligned}$$

where the last line holds when we set the parameter λ to be

$$\lambda\coloneqq\frac{B_1}{2\,(n+2)\,\left(\sqrt{4B_1+1}-1\right)\,f\left(F^{-1}\left(\frac{i}{n+1}\right)\right)}\,. \tag{33}$$

■

3. Converting error in $X_{(i)}$ to error in statistic $\hat{S}$: We can convert the additive error given in Lemma 17 into a multiplicative error in terms of the difference between $F^{-1}\left(j/(n+1)\right)-F^{-1}\left(i/(n+1)\right)$. This quantity represents the length of an interval in $\tilde{S}$. In Lemma 18, we show how to convert the additive error from Lemma 17 to a multiplicative error in terms of the length of the interval.

Lemma 18 *For a sufficiently large parameter $\tau>1$ and number of buckets $k>2B_1\beta$, suppose we have $n=O(log(k)\cdot\sqrt{\sqrt{B_1}+1\cdot\tau})$ samples from a well-behaved distribution f. Then, we have:*

$$\mathbf{Pr}\left[\left|X_{(i)}-F^{-1}\left(\frac{i}{n+1}\right)\right|\geq\frac{2}{\tau(j-i)/(n+1)}\left(F^{-1}\left(\frac{j}{n+1}\right)-F^{-1}\left(\frac{i}{n+1}\right)\right)\right]\leq\frac{0.1}{4\,k^2}$$

Proof For a single interval, we have from Fact 3.3 that:

$$\frac{F^{-1}\left(\frac{j}{n+1}\right) - F^{-1}\left(\frac{i}{n+1}\right)}{\frac{j-i}{n+1}} \geq \frac{1}{f\left(F^{-1}\left(\frac{i}{n+1}\right)\right)} - B_1 \frac{j-i}{n+1} \tag{34}$$

Note that this holds regardless of whether $j > i$ or $j < i$. We set $\frac{j-i}{n+1} \leq \frac{1}{2B_1\beta}$, which gives us: $\frac{B_1}{k} \leq \frac{1}{2\beta} \leq \frac{1}{2f(F^{-1}(\frac{i}{n+1}))}$. Plugging this back in, we get:

$$\frac{F^{-1}\left(\frac{j}{n+1}\right) - F^{-1}\left(\frac{i}{n+1}\right)}{\frac{j-i}{n+1}} \geq \frac{1}{2f(F^{-1}\left(\frac{i}{n+1}\right))} \tag{35}$$

This implies then, that we can convert the additive bound to a multiplicative one. In particular, applying the bounds from Lemma 17, this gives us that:

$$\mathbf{Pr}\left[\left|X_{(i)} - F^{-1}\left(\frac{i}{n+1}\right)\right| \geq \frac{2}{\tau\frac{j-i}{n+1}}\left(F^{-1}\left(\frac{j}{n+1}\right) - F^{-1}\left(\frac{i}{n+1}\right)\right) \geq \frac{1}{\tau \cdot f(F^{-1}\left(\frac{i}{n+1}\right))}\right] \leq \frac{0.1}{4\,k^2} \tag{36}$$

■

The above result shows that by union bound we can make sure all the order statistics we use are estimated with small error with high probability.

With this, we have all the pieces required to prove Theorem 9. We know that the proxy quantity is correct (Theorem 6). We have shown that approximating the derivatives by difference quotients incurs bounded error (call this ϵ_1) (Lemma 11), and we have now shown that we incur bounded error when using order statistics to approximate the test statistic (call this error ϵ_2) (Lemma 12). In the final section, we will show how many buckets and samples we require for the algorithm to succeed with high probability.

Part 4/4: Errors can be set to satisfy theorem.

We start by giving a corollary that helps us translate the result of Lemma 18 to one that is easy to use for analyzing the number of samples required.

Corollary 19 *If we draw $n = O(\frac{k \log k}{\epsilon}\sqrt{\sqrt{B_1}+1})$ samples, with probability at least 9/10, $\hat{L}$ concentrates around $\tilde{L}$ as defined previously claim. That is,*

$$(1-\epsilon)\cdot\tilde{L} \leq \hat{L} \leq (1+\epsilon)\cdot\tilde{L}\,.$$

Proof In order to prove this, we apply Lemma 18. As a corollary of that lemma, if $k \geq 2B_1\beta$ and $n = O(\log k\sqrt{\sqrt{B_1}+1\tau})$, we have that:

$$\begin{aligned}
&\mathbf{Pr}[]\left|X_{(i(n+1)/k)} - F^{-1}\left(\frac{n+1}{k}\frac{i}{n+1}\right)\right| \geq \frac{2}{\tau\cdot 1/k}\left(F^{-1}\left(\frac{n+1}{k}\frac{i+1}{n+1}\right) - F^{-1}\left(\frac{n+1}{k}\frac{i}{n+1}\right)\right) \leq \frac{0.1}{4\,k^2}\\
\Leftrightarrow\ &\mathbf{Pr}[]\left|X_{(i(n+1)/k)} - F^{-1}\left(\frac{i}{k}\right)\right| \geq \frac{2}{\tau\cdot 1/k}\left(F^{-1}\left(\frac{i+1}{k}\right) - F^{-1}\left(\frac{i}{k}\right)\right) \leq \frac{0.1}{4\,k^2}
\end{aligned}$$

and

$$\mathbf{Pr}[\,]\left|X_{((i+1)(n+1)/k)} - F^{-1}\left(\frac{i+1}{k}\right)\right| \geq \frac{2}{\tau \cdot -1/k}\left(F^{-1}\left(\frac{i}{k}\right) - F^{-1}\left(\frac{i+1}{k}\right)\right) \leq \frac{0.1}{4\,k^2}$$
$$\Leftrightarrow \mathbf{Pr}[\,]\left|X_{((i+1)(n+1)/k)} - F^{-1}\left(\frac{i+1}{k}\right)\right| \geq \frac{2}{\tau \cdot 1/k}\left(F^{-1}\left(\frac{i+1}{k}\right) - F^{-1}\left(\frac{i}{k}\right)\right) \leq \frac{0.1}{4\,k^2}$$

Thus, with probability at least $1 - \frac{0.2}{4k^2}$,

$$F^{-1}\left(\frac{i}{k}\right) - \frac{2}{\tau \cdot 1/k}\left(F^{-1}\left(\frac{i+1}{k}\right) - F^{-1}\left(\frac{i}{k}\right)\right) \leq X_{(i(n+1)/k)} \tag{37}$$
$$\leq F^{-1}\left(\frac{i}{k}\right) + \frac{2}{\tau \cdot 1/k}\left(F^{-1}\left(\frac{i+1}{k}\right) - F^{-1}\left(\frac{i}{k}\right)\right) \tag{38}$$
$$F^{-1}\left(\frac{i+1}{k}\right) - \frac{2}{\tau \cdot 1/k}\left(F^{-1}\left(\frac{i+1}{k}\right) - F^{-1}\left(\frac{i}{k}\right)\right) \leq X_{((i+1)(n+1)/k)} \tag{39}$$
$$\leq F^{-1}\left(\frac{i+1}{k}\right) + \frac{2}{\tau \cdot 1/k}\left(F^{-1}\left(\frac{i+1}{k}\right) - F^{-1}\left(\frac{i}{k}\right)\right) . \tag{40}$$

And so with probability at least $1 - \frac{0.2}{4k^2}$:

$$X_{((i+1)(n+1)/k)} - X_{(i(n+1)/k)} \leq F^{-1}\left(\frac{i+1}{k}\right) + \frac{2}{\tau \cdot 1/k}\left(F^{-1}\left(\frac{i+1}{k}\right) - F^{-1}\left(\frac{i}{k}\right)\right) \tag{41}$$
$$- \left(F^{-1}\left(\frac{i}{k}\right) - \frac{2}{\tau \cdot 1/k}\left(F^{-1}\left(\frac{i+1}{k}\right) - F^{-1}\left(\frac{i}{k}\right)\right)\right) \tag{42}$$
$$= \left(1 + \frac{4k}{\tau}\right)\left(F^{-1}\left(\frac{i+1}{k}\right) - F^{-1}\left(\frac{i}{k}\right)\right) \tag{43}$$

and

$$X_{((i+1)(n+1)/k)} - X_{(i(n+1)/k)} \geq F^{-1}\left(\frac{i+1}{k}\right) - \frac{2}{\tau \cdot 1/k}\left(F^{-1}\left(\frac{i+1}{k}\right) - F^{-1}\left(\frac{i}{k}\right)\right) \tag{44}$$
$$- \left(F^{-1}\left(\frac{i}{k}\right) + \frac{2}{\tau \cdot 1/k}\left(F^{-1}\left(\frac{i+1}{k}\right) - F^{-1}\left(\frac{i}{k}\right)\right)\right) \tag{45}$$
$$= \left(1 - \frac{4k}{\tau}\right)\left(F^{-1}\left(\frac{i+1}{k}\right) - F^{-1}\left(\frac{i}{k}\right)\right) \tag{46}$$

This tells us that if we draw $n = O(\frac{k \log k}{\epsilon}\sqrt{\sqrt{B_1} + 1})$ samples, then with probability at least $1 - \frac{0.2}{4k^2}$:

$$(1 - \epsilon)\,\tilde{L} \leq \hat{L} \leq (1 + \epsilon)\,\tilde{L}\,.$$

■

Now, we analyze exactly how much error we can incur while still remaining on the correct side of the threshold. In particular, given that both approximating the derivative and sampling introduce errors, we seek to draw enough samples that the test statistic calculated from samples lies on the same side of the threshold as the proxy quantity. We have essentially two sources of error that we can control: the first is ϵ_1, the additive error from the derivative approximation; the second is ϵ_2, the multiplicative error from sampling. Keeping ϵ_1 and ϵ_2 small enough to not cross the threshold with high probability entails using enough buckets and sufficiently many samples per bucket. Based on the earlier derivation of the absolute value of the gap, we have that:

$$\left|\frac{\alpha(1-z_2)^2}{f'(F^{-1}(z_2))}\right| \geq \frac{\alpha(1-z_2)^2}{f(F^{-1}(z_2))^3 B_1} \qquad \text{because} \quad |F^{-1''}(y)| = \left|\frac{-f'(F^{-1}(x))}{f(F^{-1}(x))^3}\right| \leq B_1 \tag{47}$$

Let us define $\gamma_i \coloneqq \frac{\alpha(1-i/k)^2}{f(F^{-1}(i/k))^3\, B_1} \geq \gamma \coloneqq \frac{\alpha\rho^2}{\beta^3 B_1}$, a lower bound on the actual gap. We use this quantity in our analysis. Thus, the lower bound of the test statistic calculated for a light-tailed distribution is:

$$(1-\epsilon_2)(S-\epsilon_1) = S - \epsilon_1 - \epsilon_2\, S + \epsilon_1\epsilon_2 \geq 1 - \frac{i}{k} - \frac{1}{10}\gamma$$

First, we know that for a light-tailed distribution, $S \geq 1 - \frac{1}{k}$, so we require that $\epsilon_1\epsilon_2 - \epsilon_1 - \epsilon_2\, S \geq \frac{-\gamma}{10}$. Let us first consider the case where $S \leq 1$, Then, if $\epsilon_1 \leq \gamma/20$ and $\epsilon_2 \leq \gamma/20$, we have that:

$$-\frac{\gamma}{20} - \frac{\gamma}{20} \cdot S + \frac{\gamma^2}{20} \geq -\frac{\gamma}{20} - \frac{\gamma}{20} = -\frac{\gamma}{10}\,.$$

This tells us that it suffices to have $\epsilon_1 = \gamma/20$ and $\epsilon_2 = \gamma/20$. Thus, we select the number of buckets and the number of samples in such a way as to achieve $\epsilon_1, \epsilon_2 \leq \frac{\gamma}{20}$. In particular, from Corollary 16, we know that if $k \geq \Theta(\beta(2B_1+B_2)/\gamma)$, then $|\tilde{S} - S| < \gamma/20$. Next, Lemma 12 tells us that in this case, we need estimates of $\tilde{L}$ to a multiplicative factor of $1 \pm \epsilon'$, where $\epsilon' = \Theta(\gamma/((20+\gamma)k))$. From the claim above, we have that with $n = \Theta\left(\frac{(k^2 \log k)\sqrt{\sqrt{B_1}+1}}{\gamma}\right)$, we achieve this guarantee.

Next, if we are still in the light-tailed case and $S \geq 1$, then regardless of ϵ_2 , Lemma 12 tells us we require estimates of $\tilde{L}$ to $1 \pm \epsilon'$, where $\epsilon' = \Theta(1/k^2)$. Again, from the claim above, we have that with $n = \Theta\left(k^3 \log k \sqrt{\sqrt{B_1}+1}\right)$, we can achieve the desired guarantee.

Having addressed the light-tailed case, let us consider the $(\alpha,\rho)-$heavy-tailed case. We have that:

$$\hat{S} \leq (1+\epsilon_2)(S+\epsilon_1)\,, \text{which we want to be} \leq 1 - \frac{i}{k} - \frac{9\gamma}{10}\,.$$

Since the distribution is α,ρ-heavy tailed, we know that $S \leq 1 - \frac{i}{k} - \gamma$. Thus, we set ϵ_1 such that $S + \epsilon_1 \leq 1 - \frac{i}{k} - \frac{19\gamma}{20} = 1 - \frac{i}{k} - \gamma + \frac{\gamma}{20}$, i.e., we set $\epsilon_1 = \gamma/20$. Now, $\epsilon_2 = \gamma/20$ suffices, since:

$$\left(1+\frac{\gamma}{20}\right)\left(1-\frac{i}{k}-\frac{19\gamma}{20}\right) \leq 1 - \frac{i}{k} + \frac{\gamma}{20}\left(1-\frac{i}{k}\right) - \left(1+\frac{\gamma}{20}\right)\frac{19\gamma}{20} \leq 1 - \frac{i}{k} + \frac{\gamma}{20} - \frac{19\gamma}{20} = 1 - \frac{i}{k} - \frac{9\gamma}{10}$$

as desired. We apply the union bound over the k^2 order statistics we compute to get that all simultaneously concentrate to $1 \pm \epsilon_2$ multiplicative error with probability ≥ 0.1. Thus, the number of samples required is $n = \Theta\left(\frac{k^2 \log k\sqrt{\sqrt{B_1}+1}}{\gamma}\right)$, and the number of required buckets is $k \geq \Theta(\frac{\beta(2B_1+B_2)}{\gamma})$.

In the worst case, $\gamma \geq \Theta(\alpha\rho^2\beta^{-3}B_1^{-1})$, and so $k \geq \frac{\beta^4 B_1(2B_1+B_2)}{\alpha\,\rho^2}$. Then,

$$n = \max\left\{\frac{\beta^3 B_1 k^2 \log k\sqrt{\sqrt{B_1}+1}}{\alpha\,\rho^2}, k^3 \log k\sqrt{\sqrt{B_1}+1}\right\}\,.$$

Equivalently,

$$n = \max\left\{\frac{\beta^3 B_1}{\alpha\rho^2}, k\right\} \cdot k^2 \log k \sqrt{\sqrt{B_1}+1}\,.$$

■

Appendix F. Hardness Result

In this section, we show for any number of samples m, we can construct two classes of distributions, one light-tailed and the other heavy-tailed, such that they are indistinguishable using m samples. Our lower bound implies that there is no algorithm to distinguish the class of light-tailed and heavy-tailed distributions unless further assumptions is made. Here we give an alternative definition of heavy tails, which does not require the heavy-tailed part be contiguous. Then we show that it is hard to distinguish these type of heavy-tailed distributions from the light-tailed ones.

Definition 20 *We say a distribution p is (α, ρ)-scattered-heavy-tailed if the hazard rate is decreasing by rate at least α over measurable intervals of the domain with probability mass at least ρ.*

High level idea: We construct two classes of distributions that are hard to distinguish with few samples: light tailed distributions, $\mathcal{C}_L$, and heavy-tailed distribution $\mathcal{C}_H$. The class of light-tailed distributions contains only one member that is an exponential distribution with $f_{\exp}(x) = e^{-x}$. We construct the class of heavy-tailed distributions via a randomized process as follows: We start off by the same distribution $f_{\exp}(x) = e^{-x}$. We split the domain of $f_{\exp}$ into s *chunks* such that the probability mass in every chunk is equal to $1/s$. Then, we select roughly $\rho' = \Theta(\rho)$ fraction of these chunks randomly and embed a heavy-tailed distribution, namely f_H, in (some of) those selected chunks. The construction has two key properties: First, the probability mass of a chunk remains the same even when the alteration happens. Second, if we draw one sample from a chunk, we cannot tell whether it is altered or not. The key idea which leads to this property is in the embedding. When we alter a chunk, we randomly replace $f_{\exp}$ by a heavy-tailed piece f_H or another partial PDF $\overline{f}_H$. We simply define $\overline{f}_H$ such that the mixture of f_H and $\overline{f}_H$ each with probability a half gives us exactly $f_{\exp}$. Thus, if we receive one sample from a chunk that comes from a random $\mathcal{C}_H$, it is impossible to tell whether the chunk is altered or not. It is worth noting that this process generates a class of distributions, $\mathcal{C}_H$, that depends on a parameter s. We may also $\mathcal{C}_H(s)$ to denote it. To complete our proof, we show that for any algorithm that uses m samples, there is a sufficiently large s such that it is very unlikely to have more than one sample per chunk. Thus, $\mathcal{C}_L$ and $\mathcal{C}_H(s)$ are indistinguishable when we use m samples. This fact implies that no algorithm that uses finitely many samples can distinguish a light tailed distribution from a distribution that is heavy on measurable subset of the domain with mass ρ unless we make further assumptions including that the heavy-tailed part might need to be contiguous.

Theorem 21 *For any integer m, there is no algorithm that receives m samples from p , a monotone and continuous distribution, and can distinguish whether p is light-tailed distribution or a $(\alpha, \rho)-$scattered-heavy-tailed for $\alpha < 0.0043$ and $\rho < 0.5$ with probability more than $0.5 + o(1)$.*

Proof By way of contradiction, let us assume there exists such algorithm, namely $\mathcal{A}$, which uses m samples and can distinguish if light-tailed distributions from heavy-tailed ones with probability at least $0.5 + \delta$. Here, we assume that δ is in $(0, 1/2]$ and $1/\delta$ is a constant. (i.e., δ is sufficiently away from zero.) To prove our theorem: we construct two classes of distributions: $\mathcal{C}_L$ and $\mathcal{C}_H(s)$ for a sufficiently large parameter s. The main goal is to show that if we draw m samples from the distribution in each of these classes, the outcome should be very similar, thus no algorithm can tell the difference better than guessing.

Construction of classes: As we mentioned earlier, $\mathcal{C}_L$ contains a single exponential distribution with the pdf $f_{\exp}(x) = e^{-x}$. We construct the class of heavy-tailed distribution via a randomized process as follows: We start off with the same distribution $f_{\exp}(x) = e^{-x}$. We split the domain of $f_{\exp}$ into s *chunks* such that the probability mass in every chunk is equal to $1/s$. Then, for $i \leq 3s/4$, we select chunk i with probability $\rho' = 2\rho$ randomly and embed a heavy-tailed distribution, namely f_H, in those selected chunks. When we alter a chunk, we randomly replace $f_{\exp}$ by a heavy-tailed

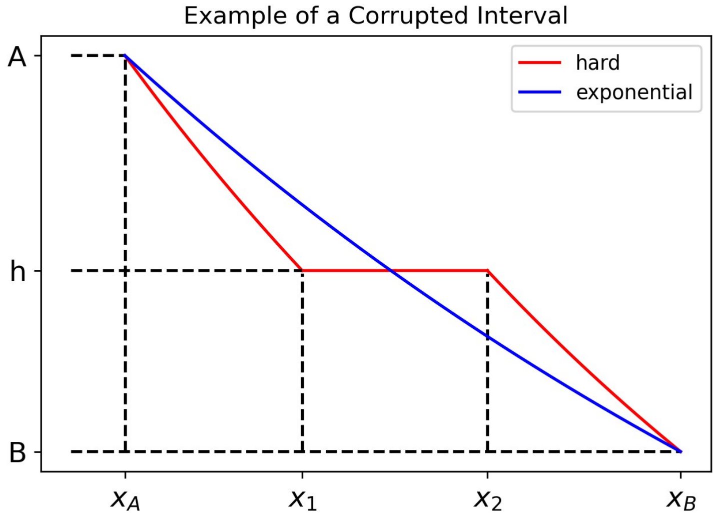


Figure 5: An example of a corrupted interval.

piece f_H or another partial PDF $\overline{f}_H$. We simply define $\overline{f}_H$ such that the mixture of f_H and $\overline{f}_H$ each with probability a half gives us exactly $f_{\exp}$. $\mathcal{C}_H$ consists of the distributions that result from this corruption process. Below are the details of a single alteration of the exponential in a chunk by f_H. The alteration on one chunk via f_H has the following properties:

- The interval has the same mass that some exponential distribution (light-tailed) would have on that interval (i.e., $1/s$), but the hazard rate is decreasing on most of the mass of that interval. We call this the *fooling* region of the interval.
- In place of the exponential distribution on that interval, we use a fast-dropping exponential, followed by a uniform, followed by another fast-dropping exponential (see Figure 5).
- Note that once the starting point and the amount of weight of the interval are fixed, for a fixed parameter $\beta \in (1, 2)$ denoting the drop rate of the fast-dropping exponential, the rest of the construction is determined. We fix this parameter to $\beta = 1.5$, but a different choice would lead to a similar analysis with slightly different constants.
- Note that since we are choosing $i \leq (3/4)s$, x_B for the rightmost chunk we could select is $\ln(4)$.

Construction of f_H: We start off by a chunk of weight $C := 1/s$ starting at $x = x_A$ and the distribution $f_{\exp}(x) = e^{-x}$. Note that this uniquely defines the end point of the chunk x_B. Clearly, we have:

$$f_{\exp}(x_A) = A\,, \qquad f_{\exp}(x_B) = B\,, \text{ and } \qquad \int\limits_{x_A}^{x_B} f_{\exp}(x)dx = A - B = C\,.$$

Letting $x_1 \in (x_A, x_B)$ denote the point at which the distribution switches from exponential to uniform and $x_2 \in (x_A, x_B)$ denote the point where the distribution switches back from uniform to exponential in the altered chunk, f_H is given by:

$$f_{\mathrm{H}}(x) = \begin{cases} Ae^{-\beta(x-x_A)} & x \in [x_A, x_1) \\ h & x \in [x_1, x_2) \\ Be^{-\beta(x-x_B)} & x \in [x_2, x_B) \end{cases} \tag{48}$$

From the fact that this construction is continuous, we have that $h = Ae^{-\beta(x_1-x_A)} = Be^{-\beta(x_2-x_B)}$. We use that to solve for x_1, x_2:

$$x_1 = x_A + \frac{1}{\beta}\ln\left(\frac{A}{h}\right) \tag{49}$$

$$x_2 = x_B + \frac{1}{\beta}\ln\left(\frac{B}{h}\right) \tag{50}$$

$$= \frac{-1}{\beta}\ln\left(\frac{A}{B}\right) + x_1 - x_A + x_B \tag{51}$$

$$\Rightarrow x_2 - x_1 = \frac{\beta - 1}{\beta}(x_B - x_A)\,. \tag{52}$$

We use $F_{\exp}$ and F_H to refer to the CDF of $f_{\exp}$ and f_H as well. We get that the CDF is given by the following:

$$F_{\mathrm{hard}}(x) = \begin{cases} F_{\exp}(x_A) + \frac{A}{\beta}\left(1 - e^{-\beta(x-x_A)}\right) & x \in [x_A, x_1) \\ F_{\exp}(x_A) + \frac{A}{\beta}\left(1 - e^{-\beta(x_1-x_A)}\right) + h(x - x_1) & x \in [x_1, x_2) \\ F_{\exp}(x_A) + \frac{A}{\beta}\left(1 - e^{-\beta(x_1-x_A)}\right) + h(x_2 - x_1) + \frac{B}{\beta}(e^{-\beta(x-x_B)} - 1) & x \in [x_2, x_B) \end{cases} \tag{53}$$

Knowing that $F_{\mathrm{hard}}(x_B) - F_{\mathrm{hard}}(x_A) = C$, we enforce the area constraint to solve for h:

$$\frac{A}{\beta}\left(1 - e^{-\beta(x_1-x_A)}\right) + h(x_2 - x_1) + \frac{B}{\beta}\left(e^{-\beta(x_2-x_B)} - 1\right) = C \tag{54}$$

$$\frac{A}{\beta}\left(1 - \frac{h}{A}\right) + h \cdot \frac{\beta - 1}{\beta}(x_B - x_A) + \frac{B}{\beta}\left(\frac{h}{B} - 1\right) = C \tag{55}$$

$$\frac{h}{\beta}\left(-1 + (\beta - 1)(x_B - x_A) + 1\right) = C - \frac{A}{\beta} + \frac{B}{\beta} \tag{56}$$

$$h = \frac{\beta C - A + B}{(\beta - 1)(x_B - x_A)} = \frac{C}{(x_B - x_A)} \tag{57}$$

Analysis of $\overline{f}_H$: Recall that we simply set $\overline{f}_H(x) = 2f_{\exp}(x) - f_H(x)$. We need to show that on the region of interest, $\overline{f}_H(x)$ is never below zero, and it is monotone decreasing. Given that, it is clear that the probability mass of a chunk when we replace $f_{\exp}$ by $\overline{f}_H$ remains equal to C as required.

Existence: To show that such an $\overline{f}_H$ exists, we must show that $2e^{-x} \geq f_{\mathrm{hard}}(x)$. In the first region, this is clearly true, as $2e^{-x} \geq e^{-x} = Ae^{-(x-x_A)} \geq Ae^{-\beta(x-x_A)}$ when $\beta \geq 1$. In the uniform region, we compare $2e^{-x}$ to a constant function, so it suffices to compare the smallest value of $2e^{-x}$ to the constant height:

$$2e^{-x} \geq h \Leftrightarrow \ln\left(\frac{2}{h}\right) \geq x \tag{58}$$

$$\ln\left(\frac{2}{h}\right) \geq x_2 = x_B + \frac{1}{\beta}\ln\frac{B}{h} = \ln\left(\frac{1}{B^{1-1/\beta}h^{1/\beta}}\right) \tag{59}$$

$$\Leftrightarrow 2B^{1-1/\beta} \geq h^{1-1/\beta} \tag{60}$$

We return to this condition in a moment after considering the second exponential region, where we want to determine when: $2e^{-x} \geq Be^{-\beta(x-x_B)}\,, x \in [x_2, x_B)$. This is equivalent to checking where $2B^{\beta-1} \geq e^{-x(\beta-1)}$ holds. For this to hold, it is sufficient for $2B^{\beta-1} \geq 1$, meaning that $B \geq (1/2)^{1/(\beta-1)}$. Note that if this holds, then Eqn. 60 also holds. When $B \leq e^{-\ln(4)}\,, \beta = 1.5$, this condition indeed holds, which means Eqn. 60 also holds.

Monotonicity: To show montonicity, we must show that $\bar{f}_H(x) = 2e^{-x} - f_H(x)$ is monotone decreasing as well in the regions in which we are interested. We calculate the derivative:

$$f'_H(x) = \begin{cases} -\beta A e^{-\beta(x-x_A)} & x \in [x_A, x_1) \\ 0 & x \in (x_1, x_2) \\ -\beta B e^{-\beta(x-x_B)} & x \in (x_2, x_B] \end{cases} \tag{61}$$

$$\Rightarrow \text{ we must check } \bar{f}'_H(x) = -2e^{-x} - f'_H(x) = -2e^{-x} + \beta A e^{-\beta(x-x_A)} \leq 0\,, \tag{62}$$

$$\bar{f}'_H(x) = -2e^{-x} - 0 \leq 0\,, \tag{63}$$

$$\bar{f}'_H(x) = -2e^{-x} - f'_H(x) = -2e^{-x} + \beta B e^{-\beta(x-x_B)} \leq 0\,, \tag{64}$$

where the second condition clearly holds, and we can summarize the first and third as checking when $-2e^{-x} + \beta D e^{-\beta(x-x_D)} \leq 0$ holds, where $D = A, x_D = x_A$ or $D = B, x_D = x_B$. Then:

$$2e^{-x} \geq \beta D e^{-\beta(x-x_D)} \tag{65}$$

$$\Leftrightarrow e^{-\beta(x-x_D)} \leq \frac{2}{\beta} e^{-(x-x_D)} \Leftrightarrow e^{-(\beta-1)(x-x_D)} \leq \frac{2}{\beta} \Leftrightarrow -(\beta-1)(x-x_D) \leq \ln\left(\frac{2}{\beta}\right) \tag{66}$$

$$\Leftarrow \frac{\ln\frac{2}{\beta}}{\beta-1} \geq |x - x_D| \Leftarrow \frac{\ln\frac{2}{\beta}}{\beta-1} \geq (x_B - x_A)\,. \tag{67}$$

Since we have that $x_B \leq \ln(4)$, and $\beta = 1.5$, still consider chunks in a constant fraction of the mass of the distribution, so we have that $x_B - x_A \leq \ln(1+4C)$. Thus, we are interested in the region where the left hand side of Eqn. 67 is greater than $1 + 4C$. We have full freedom to pick β subject to $2 > \beta > 1$. If we set $\beta = 1.5$, then this condition holds when $C \leq 0.194$, (i.e., $s \geq 5.2$) which is satisfied. Thus, we can conclude the existence and monotonicity of $\bar{\bar{f}}_H$.

Drawing samples from classes: Here, we explain a slightly modified process, that is called *poissonization method*, to draw roughly m samples from the two classes of distribution. Suppose we generate samples from $\mathcal{C} \in \{\mathcal{C}_L, \mathcal{C}_H\}$. Let p be a random distribution in $\mathcal{C}$. That is, instead of drawing m samples from p, we draw $\mathit{Poi}(m')$ samples from p where $m' = \Theta(m)$. In this way, the number of samples we see in every chunk is independent from the rest which simplifies the proof greatly. Let X denotes the sample set we obtain according to this process. We use $\mathcal{D}_L$ and $\mathcal{D}_H$ to denote the distributions of X when $\mathcal{C} = \mathcal{C}_L$ and $\mathcal{C} = \mathcal{C}_H$ respectively.

Proof of indistinguishability: Suppose we set $\mathcal{C}$ to be $\mathcal{C}_L$ and $\mathcal{C}_H$ each with probability a half. Then we draw a sample from X from the class $\mathcal{C}$ as explained above. We define event $\mathcal{E}$ to be the probabilistic event when X contains at least m samples and p is a light-tailed distribution, or an (α, ρ)-scattered-heavy tailed one. It is worth noting that condition on $\mathcal{E}$, then algorithm $\mathcal{A}$ distinguishes whether $\mathcal{C} = \mathcal{C}_L$ or $\mathcal{C} = \mathcal{C}_H$ with probability at least $0.5 + \delta$. In the following lemma, we claim that event $\mathcal{E}$ holds with high probability.

Lemma 22 *Suppose X is generated from class $\mathcal{C} = \mathcal{C}_H$ (similarly $\mathcal{C} = \mathcal{C}_L$). Then, with probability at least $1-\delta/2$, p is an (α, ρ)-scattered-heavy-tailed (similarly light-tailed) distribution, and X contains at least m samples.*

We prove this lemma in Section F.1. Using the above lemma, we can show that if X is generated from a random $\mathcal{C}$, then $\mathcal{A}$ still distinguishes whether $\mathcal{C} = \mathcal{C}_L$ or $\mathcal{C} = \mathcal{C}_H$ with some reasonable probability while $\mathcal{E}$ does not necessary holds. In particular, we have:

$$\begin{aligned}\mathbf{Pr}[\mathcal{A}(X) = \mathcal{C}] &\geq \mathbf{Pr}[\mathcal{A}(X) = \mathcal{C} \mid \mathcal{E}] \cdot \mathbf{Pr}[\mathcal{E}] \\ &\geq \left(\frac{1}{2} + \delta\right) \cdot \left(1 - \frac{\delta}{2}\right) > \frac{1}{2} + \frac{\delta}{4}\end{aligned} \tag{68}$$

Now, by Le Cam's lemma, the probability of distinction is bounded by the total variation distance between the input distributions, i.e., $\mathcal{D}_L$ and $\mathcal{D}_H$. More formally, we have:

$$\mathbf{Pr}[\mathcal{A}(X) = \mathcal{C}] \leq \frac{1}{2} + \frac{\|\mathcal{D}_L - \mathcal{D}_H\|_{tv}}{2}\,. \tag{69}$$

Now, putting Equation (69) and Equation (68) together, we obtain:

$$\|\mathcal{D}_L - \mathcal{D}_H\|_{tv} > \frac{\delta}{2}\,. \tag{70}$$

However, we have the following upper bound on the total variation distance between $\mathcal{D}_H$ and $\mathcal{D}_X$.

Lemma 23 *For $s = \Omega(m^2)$, the total variation distance between the distribution over the sample sets, $\mathcal{D}_H$ and $\mathcal{D}_L$ is at most $\delta/2$.*

For the full proof of this lemma, see Section F.2. Thus, by having the contradicting bounds for the total variation distance, we conclude that our original hypothesis regarding the existence of $\mathcal{A}$ was false. ■

F.1. Proof of Lemma 22

Lemma 24 *Suppose X is generated from class $\mathcal{C} = \mathcal{C}_H$ (similarly $\mathcal{C} = \mathcal{C}_L$). Then, with probability at least $1 - \delta/2$, p is an (α, ρ)-scattered-heavy-tailed (similarly light-tailed) distribution, and X contains at least m samples.*

Proof We start off by showing that the sample set X contains at least m samples. Recall that the number of samples, $|X|$ is a Poisson random variable with mean (and variance) m'. We set $m' = m + 2/\delta + 2\sqrt{\delta\, m + 1}/\delta$ which satisfies $m' - m = 2\sqrt{m/\delta}$. Note that since we assumed $1/\delta$ is a constant, m is $\Theta(m)$. Now, by Chebyshev's inequality, we have:

$$\mathbf{Pr}[\,|X| < m] \leq \mathbf{Pr}\left[m' - |X| > m' - m = \frac{2\sqrt{m'}}{\sqrt{\delta}}\right] \leq \frac{\delta}{4}\,.$$

Clearly, when $\mathcal{C} = \mathcal{C}_L$, p is a light-tailed distribution. Now, we need show that if p is drawn randomly from $\mathcal{C}_H$, then it is (α, ρ)-scattered-heavy-tailed. Our main claim concerns the mass in the fooling region:

Lemma 25 *For a sufficiently large $s = \Omega(\log \delta^{-1}/\rho)$, if we alter chunk i for $i \leq 3s/4$, then the probability mass in the* fooling region *is at least $2/(3s)$ and the hazard rate of the distribution in the fooling region is decreasing by rate at least $\alpha = 0.0043$.*

We prove this lemma in Section F.3. Now, let $s' > s/2$ denote the number of chunks such that $i \leq 3s/4$. Let s_a denote the number of chunks we alter and replace f by f_H among these s' intervals. It is not hard to see that $\mathbf{E}[\,s_a] = \rho' \cdot s'/2$. Given the above lemma, if $s_a \geq 3\,s\,\rho/2$, it is clear that at least on ρ-fraction of the domain, the hazard rate decreases by rate at lest α. Therefore, p will be an

(α, ρ)-scattered-heavy-tailed. We set $\rho' = 12 \cdot \rho$. Now, by the Chernoff bound, we show that s_a is large enough with high probability:

$$\mathbf{Pr}[s_a < 3\,s\,\rho/2] = \mathbf{Pr}\left[\frac{s_a}{s'} < \frac{\rho'}{2} \cdot \left(1 - \frac{1}{2}\right)\right] \leq \exp\left(-3\,s\,\rho/8\right) \leq \delta/4.$$

where the last inequality holds for $s = \Omega(\log \delta^{-1}/\rho)$. Now, by the union bound, with probability $\delta/2$, we have at least m samples, and a random $p \in C$ is either a light tailed distribution, or an (α, ρ)-scattered-heavy-tailed one. ■

F.2. Proof of Lemma 23

Lemma 26 *For $s = \Omega(m^2)$, the total variation distance between the distribution over the sample sets, $\mathcal{D}_H$ and $\mathcal{D}_L$ is at most $\delta/2$.*

Proof To prove the lemma, we propose a new process for drawing samples from a random $\mathcal{C} \in \{\mathcal{C}_L, \mathcal{C}_H\}$ based on piossonization method. In this new process, instead of drawing sample from the distribution directly, we first determine the number of samples in a chunk, and then draw samples from each chunk separately.

Let p be a random distribution in $\mathcal{C}$. We use m_i to denote the number of sample in chunk i when we draw $Poi(m')$ samples from p. Given the properties of the poissonization method, m_i is a random variable drawn from $Poi(m'/s)$. We claim if $m_i = 1$, regardless of $\mathcal{C}$ being equal to $\mathcal{C}_L$ or $\mathcal{C}_H$, the sample from chunk i, x_i, is drawn from the exponential distribution f_{exp} over chunk i. The claim is trivial when $\mathcal{C} = \mathcal{C}_L$ or when we have not alter chunk i in p. Now, assume $\mathcal{C} = \mathcal{C}_H$, and the alteration happens. In this case, since p is randomly selected, the conditional distribution in chunk i is either f_H or $\overline{f}_H$ each with probability half. Recall that we define $\overline{f}_H$ to be in such a way that the mixture of f_H and $\overline{f}_H$ is exactly f_{exp}. Thus, for a random $p \in \mathcal{C}_H$, when $m_i = 1$, sample x comes from exactly f_{exp}.

The above property implies that a sample set X for which each chunk has one sample will be generated with the same probability regardless of the choice of $\mathcal{C}$. Therefore, the total variation distance between $\mathcal{D}_H$ and $\mathcal{D}_L$ is bounded by the probability of at seeing at least two sample from a chunk. By properties of the Poisson distribution, we show it is very unlikely to see two samples from the same chunk:

$$\mathbf{Pr}[m_i \geq 2] = 1 - \mathbf{Pr}[m_i = 0] - \mathbf{Pr}[m_i = 1] = 1 - e^{-m'/s} - m'\,e^{-m'/s}/s \leq \frac{m'^2}{2\,s^2}\,.$$

Using the union bound, the probability of having a two samples from the same chunk is

$$\|\mathcal{D}_L - \mathcal{D}_H\|_{tv} \leq \mathbf{Pr}[\exists i : m_i \geq 2] \leq \frac{m'^2}{2\,s} \leq \frac{\delta}{2}\,.$$

where the last inequality is true when $s \geq m'^2/\delta = \Theta(m^2)$. ■

F.3. Proof of Lemma 25

Lemma 27 *For a sufficiently large $s = \Omega(\log \delta^{-1}/\rho)$, if we alter chunk i for $i \leq 3s/4$, then the probability mass in the* fooling region *is at least $2/(3s)$ and the hazard rate of the distribution in the fooling region is decreasing by rate at least $\alpha = 0.0043$.*

Proof We prove this lemma in two parts. First, we show that the derivative of the hazard rate is bounded above in the regions in which it is decreasing. Then, we show that in those decreasing regions, the probability mass is a constant fraction of the total mass in that chunk. We previously set $\beta = 1.5$ which we will use here, as well.

Bounded Derivative of Hazard Rate First, we calculate the hazard rate for a chunk:

$$HR_{\text{hard}}(x) = \begin{cases} \frac{Ae^{-\beta(x-x_A)}}{1-\left(F_{\exp}(x_A)+\frac{A}{\beta}\left(1-e^{-\beta(x-x_A)}\right)\right)} & x \in [x_A, x_1) \\ \frac{h}{1-\left(F_{\exp}(x_A)+\frac{A}{\beta}\left(1-e^{-\beta(x_1-x_A)}\right)+h(x-x_1)\right)} & x \in [x_1, x_2) \\ \frac{Be^{-\beta(x-x_B)}}{1-\left(F_{\exp}(x_A)+\frac{A}{\beta}\left(1-e^{-\beta(x_1-x_A)}\right)+h(x_2-x_1)+\frac{B}{\beta}(e^{-\beta(x-x_B)}-1)\right)} & x \in [x_2, x_B) \end{cases} \tag{71}$$

$$= \begin{cases} \frac{A}{\left(A-\frac{A}{\beta}\right)e^{\beta(x-x_A)}+\frac{A}{\beta}} & x \in [x_A, x_1) \\ \frac{h}{A-\frac{A}{\beta}\left(1-e^{-\beta(x_1-x_A)}\right)-h(x-x_1)} & x \in [x_1, x_2) \\ \frac{B}{\left(A-\frac{A}{\beta}\left(1-e^{-\beta(x_1-x_A)}\right)-h(x_2-x_1)+\frac{B}{\beta}\right)e^{\beta(x-x_B)}+\frac{B}{\beta}} & x \in [x_2, x_B) \end{cases} \tag{72}$$

We consider the derivative of the hazard rate to determine the fooling regions and show that the derivative is below $-\alpha$:

$$\frac{d}{dx}HR(x) = \begin{cases} \frac{-(\beta-1)e^{\beta(x-x_A)}}{\left(\left(1-\frac{1}{\beta}\right)e^{\beta(x-x_A)}+\frac{1}{\beta}\right)^2} & x \in [x_A, x_1) \\ \frac{h^2}{(C_1-hx)^2} & x \in (x_1, x_2) \\ \frac{-B\beta C_2 e^{\beta(x-x_B)}}{\left(C_2 e^{\beta(x-x_B)}+\frac{B}{\beta}\right)^2} & x \in (x_2, x_B) \end{cases} \tag{73}$$

where $C_1 = \left(A - \frac{A}{\beta}\left(1 - e^{-\beta(x_1-x_A)}\right) + hx_1\right)$ and $C_2 = \left(A - \frac{A}{\beta}\left(1 - e^{-\beta(x_1-x_A)}\right) - h(x_2 - x_1) + \frac{B}{\beta}\right)$. We first observe that because $\beta > 1$, $C_1, C_2 > 0$, the derivative of the hazard rate is positive in the uniform region and negative in the two exponential regions. Thus, these are the fooling regions. We now show that for a constant α, the derivatives of the hazard rate in the fooling regions are at most $-\alpha$. First, we lower bound the absolute value of this quantity in the first exponential region:

$$\frac{(\beta-1)e^{\beta(x-x_A)}}{\left(\left(1-\frac{1}{\beta}\right)e^{\beta(x-x_A)}+\frac{1}{\beta}\right)^2} \geq \frac{\beta-1}{\left(\left(1-\frac{1}{\beta}\right)e^{\beta(x-x_A)}+\frac{1}{\beta}\right)^2} \tag{74}$$

$$\geq \frac{\beta-1}{\left(\left(1-\frac{1}{\beta}\right)\frac{A}{h}+\frac{1}{\beta}\right)^2} = \frac{\beta^2(\beta-1)}{\left((\beta-1)\frac{A}{h}+1\right)^2} \tag{75}$$

$$\geq \frac{\beta^2(\beta-1)}{\left(\frac{\beta-1}{h}+1\right)^2} \tag{76}$$

Next, we lower bound the absolute value of the derivative of the hazard rate in the second exponential region:

$$\frac{B\beta C_2 e^{\beta(x-x_B)}}{\left(C_2 e^{\beta(x-x_B)}+\frac{B}{\beta}\right)^2} \geq \frac{B\beta C_2 B/h}{\left(C_2+\frac{B}{\beta}\right)^2} = \frac{\beta C_2 B^2}{h\left(C_2+\frac{B}{\beta}\right)^2} \tag{77}$$

$$\geq \frac{\beta B^3}{h\left(A+\frac{B}{\beta}+\frac{B}{\beta}\right)^2} \geq \frac{\beta B^3}{\left(1+\frac{2}{\beta}\right)^2} . \tag{78}$$

The transition from Eqn. 77 to Eqn. 78 results from bounding C_2 as follows:

$$C = A - B = \frac{A}{\beta}\left(1 - e^{-\beta(x_1-x_A)}\right) + h(x_2 - x_1) + \frac{B}{\beta}\left(e^{-\beta(x_2-x_B)} - 1\right) \tag{79}$$

$$\Rightarrow A - \frac{A}{\beta}\left(1 - e^{-\beta(x_1-x_A)}\right) - h(x_2 - x_1) + \frac{B}{\beta} = B + \frac{B}{\beta}e^{-\beta(x_2-x_B)} = C_2 . \tag{80}$$

Thus, $A + \frac{B}{\beta} > C_2 > B$.

Finally, it suffices to bound these quantities for intervals that lie in a certain region of the distribution that comprises a constant fraction of the mass of the distribution. Thus, since we have that $i \leq 3s/4$, we know that $x_B \leq -\ln(1/4) = \ln(4)$, implying that $1/4 \leq A, B, h$.This allows us to bound Eqn. 78 by:

$$\frac{\beta B^3}{\left(1+\frac{2}{\beta}\right)^2} \geq \frac{\beta (1/4)^3}{\left(1+\frac{2}{\beta}\right)^2} = \frac{\beta}{4^3\left(1+\frac{2}{\beta}\right)^2}$$

We also require a lower bound on h ; by definition, we have that $h \geq B = e^{-x_B} \geq 1/4$. Finally, we plug this back into the bounds from Eqn. 76 and Eqn. 78 to determine α :

$$\begin{aligned} \alpha &= \min\left(\frac{\beta^2(\beta-1)}{\left(\frac{\beta-1}{1/4}+1\right)^2}, \frac{\beta}{4^3\left(1+\frac{2}{\beta}\right)^2}\right) \\ &= \min\left(\frac{\beta^2(\beta-1)}{(4\beta+3)^2}, \frac{\beta^3}{4^3(\beta+2)^2}\right) \end{aligned}$$

When $\beta = 1.5$, $\alpha \approx 0.0043$.

Area Under Fooling Region Finally, we must show that the area under the fooling region is large. To do this, we evaluate $F_{\text{hard}}(x_B) - F_{\text{hard}}(x_2) + F_{\text{hard}}(x_1) - F_{\text{hard}}(x_A)$ and show that it is a large fraction of C for the whole domain.

Evaluating this area, we get:

$$F_{\text{hard}}(x_B) - F_{\text{hard}}(x_2) + F_{\text{hard}}(x_1) - F_{\text{hard}}(x_A) = \frac{A}{\beta}\left(1 - e^{-\beta(1/\beta \ln(A/h))}\right) + \frac{B}{\beta}\left(e^{-\beta(1/\beta \ln(B/h))} - 1\right) \tag{81}$$

$$= \frac{A}{\beta} - \frac{h}{\beta} + \frac{h}{\beta} - \frac{B}{\beta} = \frac{C}{\beta} \tag{82}$$

Over the whole domain of the distribution, this is a constant fraction of C, the total area in a chunk. When $\beta = 1.5$, as before, $1/1.5 = 2/3$ of the area in the chunk comes from the fooling region.

Thus, we have shown that in a chunk, a constant fraction of the area lies in a region with hazard rate $\leq -\alpha$, for a constant α .

■

Appendix G. Discussion of Experiments

Here, we include a version of the algorithm with weaker theoretical guarantees that we used in our experiments. We also discuss some preliminary experiments run on real-world data. Finally, we present the details of the experiments run to compare our algorithm to a naive one.

G.1. TPC-H Job Distribution

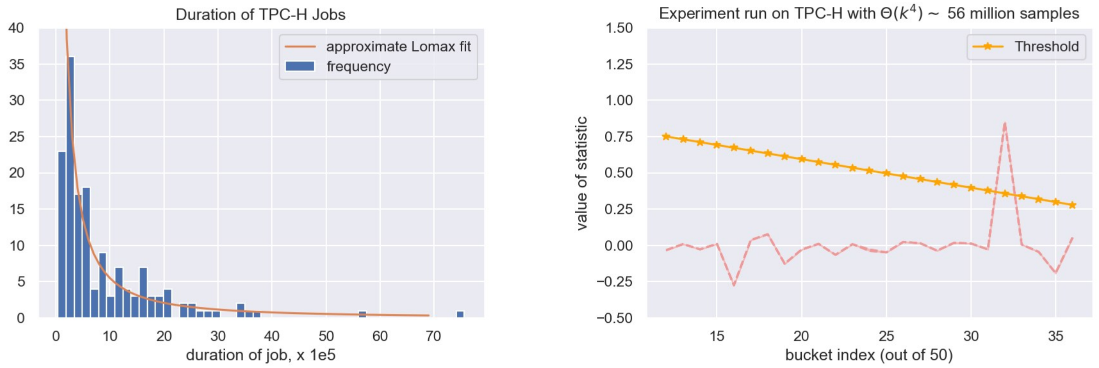
Data and Methods We run experiments on a 2.8 GHz Intel i7 core using the algorithm presented in Appendix G. The dataset considered is sampled from TPC-H queries (TPC). We use the dataset curated by Mao et al. (2019), in which they sample jobs from different input sizes from 22 different TPC-H queries. Our experiments verify their claim that the distribution of jobs over work is heavy-tailed, which they justify by noting that about 20% of the jobs contain about 80% of the work, a common heuristic for judging heavy-tailedness. In Figure 6, we plot the distribution of job durations

Algorithm 2 Weak (α, ρ)-Heavy-Tailed Test

```
$\mathcal{S}_1, \mathcal{S}_2, \mathcal{S}_3, \mathcal{S}_4 \leftarrow$ each $n$ samples from the distribution
Sort $\mathcal{S}_1, \mathcal{S}_2, \mathcal{S}_3, \mathcal{S}_4$
Split into $k$ equal weight buckets and determine interval endpoints $I$
Calculate $L[i] = I[i+1] - I[i]$ and $dL[i] = L[i+1] - L[i]$
Calculate $S[i] = \frac{L[i]}{dL[i]}$ for $i \in \{1, 2, ..., k-2\}$
if $S[i] < 1 - \frac{i}{k} - \frac{1}{2}gap(\alpha)$ for any $i \in \{c_1 \cdot k, ..., c_2 \cdot k\}$ then
    PASS
else
    FAIL
end
```

from the dataset and overlay an approximate Lomax fit. Notably, the distribution is not monotone, which means it does not entirely obey our assumptions. Further, due to a limited number of jobs, the distribution does not lie on an infinite domain. However, we show that the algorithm still picks up on heavy-tailedness. We sample $n = \Theta(k^4) \approx 56$ million samples for $k = 50$.

Results and Discussion We find that the test statistic considers the distribution heavy-tailed, since not all of the indices considered "look" light-tailed in our statistic. In particular, the red dashed line represents the test statistic calculated on samples from the distribution of jobs over work in TPC-H. Thus, we are able to verify the claim of the authors of Mao et al. (2019) that this distribution is heavy-tailed. It is worth noting that we have applied our algorithm to a practical setting where several of the assumptions of our work do not hold, and the algorithm still shows some signal when consider a heavy-tailed distribution. In order for this validation to be complete, however, we would need to show that a light-tailed distribution that doesn't match our assumptions exactly does not behave unexpectedly.



(a)

(b)

Figure 6: The plot on the left shows the distribution of job duration with an approximate Lomax fit. On the right, in (b), we plot the calculated value of the test statistic in comparison to the threshold for considering a distribution heavy-tailed. Since the dashed red line is below the orange at at least one point, the algorithm would consider this distribution heavy-tailed, in line with the assessment of the authors of Mao et al. (2019). We note that we don't see a significant spread over many runs, since the underlying distribution is finite. For the same reason, increasing the number of samples would not add any benefit.

G.2. Comparison Against a Naive Algorithm

In this section, we describe experiments we ran to compare our algorithm to a naive algorithm, showing that our algorithm has an edge in distinguishing difficult cases. We first explain the experimental setup, including some context for why we chose the examples we did and the naive algorithm. We then describe the results we saw, and finally posit an explanation as to why this is the case.

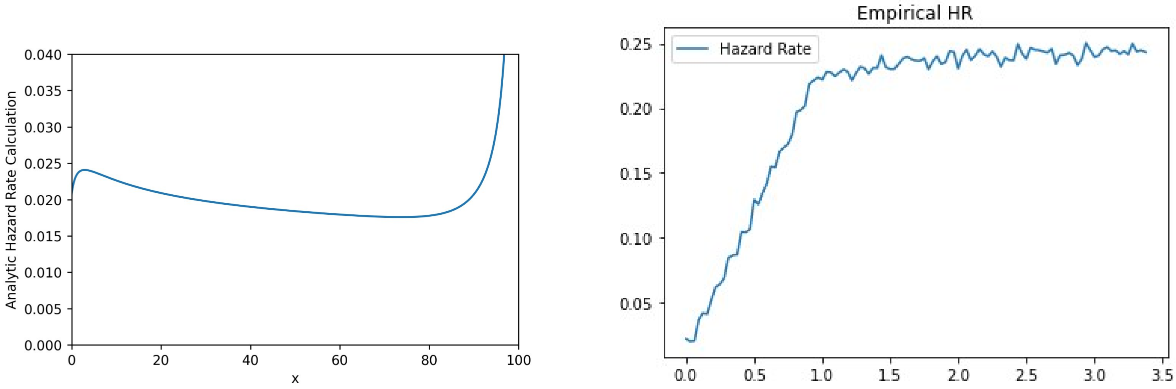
Motivation and Experimental Setup Our algorithm, to our knowledge, is the first finite-sample algorithm for this problem over a continuous domain. As simple as our algorithm is, a natural question is that of how a very naive algorithm might perform. In particular, we could attempt to learn the CDF from samples, and then directly calculate the derivative of the hazard rate from that. Based on the Dvoretzky–Kiefer–Wolfowitz (DKW) inequality, we know that with enough samples, we can get a good approximation of the CDF. Thus, a simple idea would be to approximate the PDF and derivative of the PDF from the empirical CDF and from there compute an approximation to the derivative of the hazard rate. Since the DKW inequality provides a guarantee in terms of the supremum of the difference between the true and empirical CDFs, a hard case for an algorithm that learns the CDF would be one where the CDFs of two distributions are very close to one another in CDF but are classified differently, one as heavy-tailed and one as light-tailed. To this end, we choose the likelihood $f(x) = \exp(-x/\ln(x))$[9] and the PDF for an exponential, $f(x) = \exp(-x)$.

As discussed in Appendix G, we implemented a weaker version of our algorithm that didn't involve the finer-grained buckets. We sought to distinguish the aforementioned distributions based on samples drawn from the likelihood Fixing the number of samples, over the course of 50 repetitions, we set half of them to be exponential and half to be the hard case, sampled from the chosen one, and ran both our tester and the naive tester. We used appropriate settings for the constants β, B_1, B_2 for

[9] the normalization constant is hard to calculate analytically so we did so numerically.

the threshold of our algorithm. We then computed what percentage of the time each algorithm had been correct, and we conducted this process for various sample sizes.

Results and Discussion From analytically calculating the hazard rate, we see that the distribution has decreasing hazard rate over most of its support. However, from the plot of the hazard rate from samples, the naive algorithm considers the distribution light-tailed early on, and almost exponential later (constant hazard rate) (Figure 7). Our algorithm, too, confuses the distribution for exponential late in the distribution. However, with the same number of samples, crucially, it identifies the heavy-tailed-ness early on (Figure 8).



(a)

(b)

Figure 7: Analytic Hazard Rate and Hazard Rate calculated by the naive algorithm for a hard distribution, $f(x) = \exp(-x/\log(x))$

Finally, we consider the success rates of our algorithm vs the naive one. In Figure 9, we note that our algorithm starts to do better than random guessing with around 10^7 samples, where the naive algorithm still does not do better than random guessing.

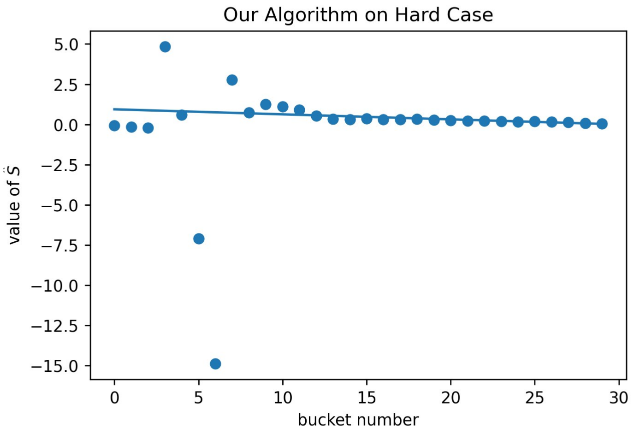


Figure 8: The value of the statistic for our algorithm (averaged over two runs) also confuses the distribution with an exponential toward the end of the domain (as evidenced by the coincidence of the statistic with the threshold) but determines the heavy-tailed-ness from early buckets.

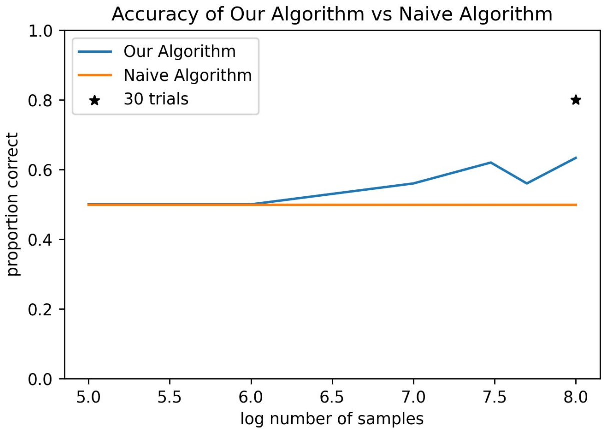


Figure 9: Performance of Each Algorithm. Note that 0.5 is random guessing, and so only our algorithm does better than that at any point.

Appendix H. Low-sample Analysis

We analyze the performance of our algorithm when it is given fewer samples than what our theoretical analysis requires. In specific applications, where the provable guarantees we give may not be required, we may be able to get good enough results without as many buckets and samples.

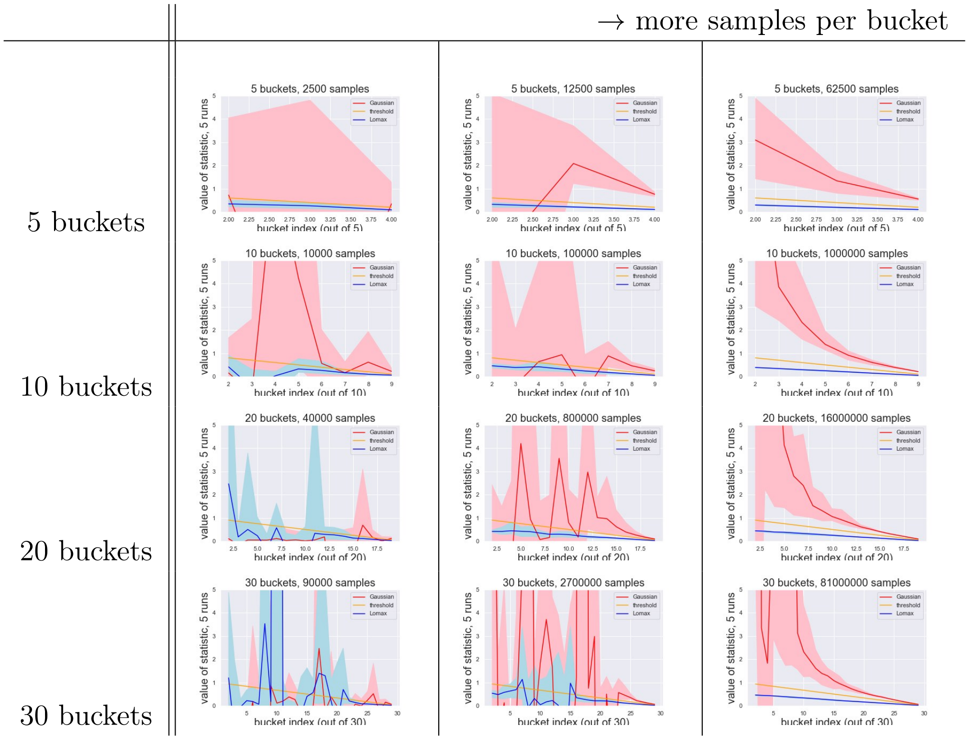


Table 1: In this table, we consider (half) Gaussian (light-tailed) and Lomax (heavy-tailed) distributions. We use smaller sample sizes (2500 to 20 million) than the experiments in the main paper (Except for the bottom-rightmost plot, which has 81 million samples). Observe that with fewer samples per bucket (further to left), there is much more variation (pink area and light blue area) in the computed statistics, and they do not behave as we would wish for our statistic to behave. In the rightmost column, however, we start to see the value of our statistic gets very close to what we theoretically expect, even with a small value for k. In particular, it concentrates well in the tail. Indeed, in the rightmost column we see that for smaller values for k (62,500 samples for $k = 5$) the test clearly distinguishes between the (half) Gaussian and Lomax distributions. Despite promising behavior of the test even from such few buckets and samples, it is essential to note that large k is theoretically necessary to capture the behavior of an *arbitrary* underlying (well-behaved) distribution. However, in specific applications, these provable guarantees may not be required, and good enough results may be seen without as many buckets and samples in these cases.