Innovations in Computer Science 2010

Space-Efficient Estimation of Robust Statistics and Distribution Testing

Steve Chien[1] Katrina Ligett[2] Andrew McGregor[3]
[1]Microsoft Research, Silicon Valley Campus
[2]Cornell University
[3]University of Massachusetts, Amherst
schien@microsoft.com katrina@cs.cornell.edu mcgregor@cs.umass.edu

Abstract: The generic problem of estimation and inference given a sequence of i.i.d. samples has been extensively studied in the statistics, property testing, and learning communities. A natural quantity of interest is the *sample complexity* of the particular learning or estimation problem being considered. While sample complexity is an important component of the computational efficiency of the task, it is also natural to consider the *space complexity*: do we need to store all the samples as they are drawn, or is it sufficient to use memory that is significantly sublinear in the sample complexity? Surprisingly, this aspect of the complexity of estimation has received significantly less attention in all but a few specific cases. While space-bounded, sequential computation is the purview of the field of data-stream computation, almost all of the literature on the algorithmic theory of data-streams considers only "empirical problems", where the goal is to compute a function of the data present in the stream rather than to infer something about the source of the stream.
Our contributions are two-fold. First, we provide results connecting space efficiency to the estimation of robust statistics from a sequence of i.i.d. samples. Robust statistics are a particularly interesting class of statistics in our setting because, by definition, they are resilient to noise or errors in the sampled data. We show that this property is enough to ensure that very space-efficient stream algorithms exist for their estimation. In contrast, the numerical value of a "non-robust" statistic can change dramatically with additional samples, and this limits the utility of any finite length sequence of samples. Second, we present a general result that captures a trade-off between sample and space complexity in the context of distributional property testing.

Keywords: data streams, property testing, robust statistics

1 Introduction

Consider a sequence of i.i.d. samples $x_1, \ldots, x_m$ drawn according to some unknown distribution.[1] [2] Estimating the parameters of the unknown distribution or testing if the distribution satisfies a given property are problems that have been extensively studied by the statistics, property testing, and learning communities. One natural quantity of interest is the *sample complexity*—that is, how large must m be such that the desired inference can be made with high probability. For example, in distributional property testing, we ask how many samples are required to determine whether the source distribution satisfies a property or is "far" (in an appropriate sense) from any distribution that satisfies the property. Unfortunately, even in the restricted case of testing discrete distributions, it can be shown that the sample complexity of many properties is polynomial in the support size of the distribution in question.

While sample complexity is an important component of the computational efficiency of the task, it is also natural to consider the *space complexity* of the task—an algorithm that can process samples sequentially using only small space might be considered more practical than one that stores the entire set of samples and then runs a complex algorithm on the entire data set. Perhaps surprisingly, this aspect of the complexity has received significantly less attention in all but a few specific cases. For example, if the distribution is promised to belong to a parametrized family of distributions, Fisher's theory of *sufficient statistics* [11] gives a framework for reasoning about the minimal statistics that should be maintained, such that no information relevant to the parameter estimation is disregarded. Some online learning algorithms are

[1]Work partly done while at Microsoft Research, Silicon Valley Campus and Carnegie Mellon University. Supported in part by an NSF Graduate Research Fellowship and a CIFellows Postdoctoral Fellowship.

[2]Supported by University of Massachusetts, Amherst startup funds. Work partly done while at Microsoft Research, Silicon Valley Campus.

either naturally space efficient in practice or can be engineered to use less space than that required by a naive implementation (e.g., Dekel et al. [8]). Kearns et al. [18] also present an intriguing example of a learning task that requires all the samples to be stored. However, our general problem is still poorly understood.

It would seem that statistical estimation and property testing are a natural fit for data-stream algorithms where an algorithm must process data sequentially given only a limited amount of memory to summarize the data. However, almost all of the literature on the algorithmic theory of data-streams to date considers "empirical problems" where the answer sought is solely determined by the m elements that appear in the stream. Yet there is no good reason to limit attention to such problems:

Isn't understanding the process generating the stream as important, if not more important, than the data in the stream itself?

This paper makes two main contributions to redressing this issue. First, we identify *robust statistics* as an important class of statistics in our setting because a) we can show that they can be approximated in small space and b) a statistic that is "not robust" is of less interest as its numerical value may change dramatically with additional samples, limiting the utility of any finite length sequence of samples. Second, following recent work by Valiant [20], we demonstrate a general result that exhibits a trade-off between sample and space complexity in the context of distributional property testing. In the rest of this section, we discuss both contributions in further detail.

1) Robust Statistics

Robust statistics is an established area of theoretical statistics that studies when statistical estimators are resilient, or "robust", to perturbation of the distribution being considered or model assumptions. It can be argued that robustness quantifies an important sense of how meaningful an estimator is, e.g., it is hard to utilize the outcome of an estimator if it is very sensitive to a slight perturbation of the distribution. The area has seen considerable growth in both understanding and importance over the past several decades. In that time, statisticians have devised a number of notions of robustness. Subsequently, various families of estimators have been identified along with the types of distributions under which they qualify as robust. We defer details and formal definitions to Section 2.

In this paper, we study the robust statistics framework from a computational perspective, in particular from the vantage point of data-stream algorithms, in which we are given a limited amount of space to process a sequence of independent samples from the distribution. Our main question is:

Does the robustness of a statistic make it easier or harder to approximate in a streaming fashion?

Our answer is a happy one. We find ourselves in the fortunate situation that the estimators that are arguably the most desirable are those that we are able to efficiently approximate. In general, robust statistics turn out to be very amenable to space-efficient streaming computation: robustness to perturbation yields a certain degree of robustness to the sampling techniques that are typical in streaming algorithms. In one sense, this is straightforward—as we describe later, it follows almost immediately from known results that robust statistics can be estimated within additive error ϵ in small space by simply computing the statistic on an $O(1/\epsilon^2)$-sized sample of the data. We show, however, that we can do better than this, and develop algorithms that achieve much better space complexity, typically $\tilde{O}(\log 1/\epsilon)$, for several broad classes of estimators, including the three major well-established classes known as location M-estimators, L-estimators, and R-estimators. Together, these results can be interpreted as a strong proof of concept, in that they cover a variety of estimators, including many of the best-known and commonly used statistics.

We are not the first to draw a connection between robust statistics and a field of computer science. For example, Dwork and Lei [10] recently demonstrated that a number of robust statistical analyses can also be computed in a differentially private fashion. Differential privacy captures the notion that the output of a computation on a database of private information should not be sensitive to a small change in the input database (the presence of absence of a particular individual), and the work of Dwork and Lei draws an parallel between the privacy literature's notion of sensitivity and the statistics literature's notion of robustness.

2) Distribution Property Testing

The study of distribution property testing is concerned with distinguishing if a distribution satisfies a certain property or is "far" from all distributions that satisfy the property. Often the property can be expressed in terms of a real-valued function defined on the distribution, e.g., is the entropy of the distribution less than some threshold? Unlike the work on robust statistics, in this setting it is standard to consider distributions over a finite domain $[n]$. Typically, proper-

ties require $n^{\Omega(1)}$ samples to test. For example, given two distributions, $\Theta(n^{2/3})$ samples are necessary and sufficient to distinguish between the case where the distributions are identical and the case where the distributions are at variation distance $1/2$ [3, 20]. In the case of entropy, for sufficiently large constants $\alpha < \beta$, $n^{\alpha/\beta - o(1)}$ samples are necessary and $n^{\alpha/\beta + o(1)}$ samples are sufficient to distinguish between the entropy being below α or above β [1, 20]. The space complexity of these problems has not been considered before. We present a technically straightforward, but very general, result showing it is possible to ensure low space complexity at the cost of increasing sample complexity; we defer further details to Section 6. For entropy, the result shows that for any $\gamma > 0$, there exists a $O(n^{\alpha/\beta + o(1) - \gamma})$ space solution if sample complexity is increased to $O(n^{\alpha/\beta + o(1) + \gamma})$. However, for $O(\text{polylog}\, n)$ space, $O(n^{1+o(1)})$ samples suffice.

3) Data Streams

The data access model we adopt assumes that the samples are processed in the order they are (independently) drawn from the distribution. This implicitly means that the m samples in the stream are in random order, i.e., conditioned on the set of samples, each of the possible $m!$ orderings of the samples are equally likely. Guha and McGregor [13] demonstrate the power of random-order data models, showing that median-finding requires exponentially more passes in the adversarial-order model. The selection algorithm for order statistics that we use as a building block in our algorithms for robust estimators is based on their algorithm. Random order streams were also considered in [5, 6, 9, 14, 21] but almost all of this work only considers empirical problems. For example, while Guha et al. [14] explored connections between the random-order model and property testing, they were interested in properties of the stream itself rather than the process that generated the stream. One exception is work on estimating the density function of a k-piece-wise linear density function from a series of samples [7, 12]. Lastly, we note that in our model we assume that samples from the distribution are stored in unit space, but that maintaining a counter requires space logarithmic in the largest value.

Notation: When discussing a probability distribution over the real line, we will always use a capital letter to denote its cumulative distribution function (cdf), and the corresponding lower case letter to denote its probability density function (pdf) when it exists. For a distribution F, F^{-1} is the well-defined function $F^{-1}(t) = \sup_{x \in \mathbb{R}}\{F(x) < t\}$. Let Δ_y refer to the cdf of the distribution where all mass is concentrated at y. We say that a sequence of values $x_1, \ldots, x_m$ defines the *empirical distribution* with cdf $F_m = \frac{1}{m}\sum_i \Delta_{x_i}$. We will use $\mathcal{D}_{[n]}$ and $\mathcal{D}_{\mathbb{R}}$ to denote the set of probability distributions over $[n] := \{1, \ldots, n\}$ and the real line respectively.

2 Robust Statistics: Background and Preliminaries

For a comprehensive overview of robust statistics, we direct the reader to an invaluable book by Hampel et al. [15].

An *estimator* is a family of real-valued functions $(T_m)_{m=1,2,\ldots}$ on collections of sample data (represented by the corresponding empirical distribution). Many estimators, including those we consider in the next three sections, are functionals (i.e. there exists some $T : \mathcal{D}_{\mathbb{R}} \to \mathbb{R}$ such that $T_m(F_m) = T(F_m)$) or converge in probability to a functional. As an example, for the mean, we have $T_m(\sum_i \Delta_{x_i}) = \frac{1}{m}\sum x_i$ and $T(F) = \int x dF(x)$. Following the statistics literature, we consider *additive* approximations of $T(F)$ where F is the source distribution: an ϵ-approximation of a value $T(F)$ is a value in the interval $[T(F) - \epsilon, T(F) + \epsilon]$.

A central concept in the study of robust statistics is the influence function:

Definition 2.1. *Given an estimator $T : \mathcal{D}_{\mathbb{R}} \to \mathbb{R}$, its* influence function *at a distribution F is*

$$\text{IF}(x; T, F) = \lim_{t \to 0} \frac{T((1-t)F + t\Delta_x) - T(F)}{t}.$$

The gross error sensitivity *is defined as* $\sup_x |\text{IF}(x; T, F)|$.

The influence function can be interpreted as a derivative, measuring the rate of change of the estimator as an infinitesimal amount of probability mass is transferred to x. Note that the influence function is parametrized by both the estimator T *and* a distribution F, and many measures of robustness are intended to describe such a (T, F) pair rather than just the estimator T. The gross error sensitivity describes the maximum change to an estimator that adding an infinitesimal amount of noise to the distribution can effect. It is almost a prerequisite that this quantity be bounded for an estimator to be intuitively thought of as robust.[3]

[3]Other measures of robustness can be derived from the influence function. These include the *local shift sensitivity*, which describes the effect of shifting a small amount of probability mass a small amount, and the *rejection point*, which, when finite, indicates the point beyond which outliers are disregarded completely.

Generally, we will say that an estimator is (σ, τ)-robust at a distribution if its "average" influence function is bounded by τ inside a σ-neighborhood of the distribution. The standard distance used in this context is the Lévy distance.

Definition 2.2 (Lévy distance)**.** *Given two distributions F and G, the* Lévy distance *between them is defined as $d(F,G) = \inf\{\epsilon : F(x-\epsilon) - \epsilon \leq G(x) \leq F(x+\epsilon) + \epsilon\}$.*

Informally, this is the side length of the largest axis-parallel square that can be inscribed between the graphs of F and G. Note that because the Lévy distance concerns cdfs, it can be quite different than other, perhaps more familiar, distances between distributions, such as variation distance between the pdfs.

Example: Our intuition suggests that the median is generally robust, while the mean is not. We can see this quantitatively in their respective influence functions. For the mean, for any distribution F, we have $\mathrm{IF}(x; T, F) = x - T(F)$, and so the influence function (and hence the gross error sensitivity) is unbounded. On the other hand, when T is the median, the influence function is $\mathrm{IF}(x; T, F) = \frac{\mathrm{sign}(x - T(F))}{2f(T(F))}$, which is bounded so long as $f(T(F)) > 0$. This is in accord with our intuition that the median should be fairly stable so long as there is some probability mass at its value; the more density at the median value, the more stable that value is. We can see that some non-robust estimators such as the mean can be impossible to approximate accurately. For example, let D_t represent the distribution $(1-t)\Delta_0 + t\Delta_{1/t^2}$; this has mean $\frac{1}{t}$. Let $\mathcal{A}$ be an approximation algorithm that (with high probability) takes at most m samples when presented with Δ_0 as its input. Then with at least constant probability, $\mathcal{A}$ will not be able to distinguish Δ_0 and $D_{1/100m}$, as it is unlikely to draw an outlier among its samples.

Our definition of robustness is based on the idea of a bounded gross error sensitivity, which intuitively limits the potential effect of a single data point as the total number of samples goes to infinity. Ideally, we would like to say that for our purposes, an estimator T is robust at a distribution F if its gross error sensitivity is bounded by a fixed constant τ. Since we cannot actually draw an infinite number of samples, however, we will use a more coarse-grained version—intuitively, we say that T is robust at F if the average of the influence function is bounded along all paths within a neighborhood of F of small constant radius σ. The intended consequence of this will serve as our formal definition:

Definition 2.3. *Given an estimator T and a distribution F, we say that T is (σ, τ)-robust at F if for all distributions F' such that $d(F, F') \leq \sigma$, $|T(F) - T(F')| \leq \tau d(F, F')$.*

Classes of Estimators As mentioned earlier, we study three types of commonly used estimators: location M-estimators, which generalize maximum likelihood estimators, L-estimators, which are linear combinations of order statistics (for example, the mean of all elements between the 25th and 75th percentiles), and R-estimators, which are based on standard statistical rank tests. All three classes contain familiar examples that are deployed in real-world practice, and will be formally defined later.

2.1 Preliminary Results

As mentioned in the introduction, any robust estimator can be approximated to within ϵ if we are willing to allow ourselves the use of space that scales with ϵ^{-2}. The essential idea is that a sample of this size induces an empirical distribution that is close to the actual distribution, and we can simply calculate the estimator on the sample. The main tool we will use is the following:

Theorem 2.4. [Dvoretzky-Kiefer-Wolfowitz inequality] *Let $x_1, \ldots, x_m$ be m samples drawn independently with respect to F, and let $F_m = \frac{1}{m}\sum_{i=1}^{m} \Delta_{x_i}$. Then $\Pr[\sup_x |F_m(x) - F(x)| > \epsilon] \leq \exp(-2m\epsilon^2)$.*

With this in mind, the next theorem follows because the Kolmogorov-Smirnov distance, i.e., $\sup_x |F_m(x) - F(x)|$, is upper bounded by the Lévy distance.

Theorem 2.5. *Let T be an estimator and F be a distribution at which T is (σ, τ)-robust, and let $\epsilon \leq \sigma\tau$. For $m = \frac{\tau^2}{2\epsilon^2} \ln \frac{1}{\delta}$, define $F_m = \sum_{i=1}^{m} \Delta_{x_i}$ where $x_1, \ldots, x_m$ are m independent samples drawn from F. With probability at least $1 - \delta$, $T(F_m)$ lies in the interval $[T(F) - \epsilon, T(F) + \epsilon]$,*

Proof. The theorem follows more or less immediately from the Dvoretzky-Kiefer-Wolfowitz inequality, which implies that $\Pr[d(F, F_m) > \frac{\epsilon}{\tau}] \leq \exp(-\frac{2m\epsilon^2}{\tau^2}) = \delta$. Since T is (σ, τ)-robust at F, we have $|T(F) - T(F_m)| \leq \tau d(F, F_m) < \tau\frac{\epsilon}{\tau} = \epsilon$, as required.

Algorithm 1, for computing order statistics, is adapted from Guha, McGregor [13]. We will use it as a sub-routine in other algorithms.

Lemma 2.6 (adapted from Guha, McGregor [13])**.** *Algorithm 1, given as input a distribution F, $t^* \in [0,1]$, and any $\epsilon, \delta > 0$, returns an element u such that*

Algorithm 1 Order statistics$(t^*; \epsilon, \delta)$ (adapted from [13])

$(a, b) = (-\infty, \infty)$.
repeat
  Sample $u \in (a, b)$ using up to $\frac{1}{\epsilon} \ln \frac{3}{\delta}$ samples (i.e. draw samples from F until finding a point in (a, b) or until reaching the given number of samples, when sampling fails.)
  Estimate $F^{-1}(u)$ using $\frac{2}{\epsilon^2} \ln \frac{3}{\delta}$ samples; call result t
  if $t - t^* < \frac{\epsilon}{2}$ **then** $a \leftarrow u$;
  else if $t - t^* > \frac{\epsilon}{2}$ **then** $b \leftarrow u$.
until sampling step fails or $|t - t^*| < \frac{\epsilon}{2}$
if $|t - t^*| < \frac{\epsilon}{2}$ **then** output u;
else [sampling step failed] output a.

$\exists t \in [t^* - \epsilon, t^* + \epsilon]$ for which $F^{-1}(t) = u$, with probability at least $1 - \delta$, while using at most $O(\log \frac{1}{\epsilon} \log\log \frac{1}{\delta})$ space and $O(\epsilon^{-2} \log \epsilon^{-1} \ln^2 \delta^{-1})$ samples.

3 M-estimators

Perhaps the most prominent class of robust estimators is that of M-estimators, so named because they generalize standard *maximum likelihood* estimators. There are two main types of M-estimators:

Definition 3.1. *Given a function $\rho : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$, a* ρ-type M-estimator *T for a given probability distribution F over $\mathbb{R}$ is defined as $T(F) = \operatorname{argmin}_\theta \left(\int \rho(x, \theta) dF(x) \right)$.*

Definition 3.2. *Given a function $\psi : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$, a* ψ-type M-estimator *T for a given probability distribution F is the value of θ (if it exists) for which $\int \psi(x, \theta) dF(x) = 0$,*

Definition 3.1 can be interpreted as minimizing the average loss as measured by ρ. Note that if ρ is differentiable, then its derivative can be used to describe an equivalent M-estimator under Definition 3.2.

For now, we consider the major class of *location* M-estimators, in which $\rho(x, \theta) = \rho(x - \theta)$ or $\psi(x, \theta) = \psi(x - \theta)$. The mean can be characterized by $\rho(y) = y^2$ or $\psi(y) = y$, while the median can be specified by $\rho(y) = |y|$ or $\psi(y) = \text{sign}(y)$ for $y \neq 0$. Among more specialized M-estimators is the Huber estimator, where for a given constant b, $\psi(x) = x$ for $|x| \leq b$ and $\psi(x) = b \cdot \text{sign}(x)$ for $|x| > b$. Each of these also has a "redescending" version in which there exists some $r > 0$ such that $\psi(x) = 0$ when $|x| \geq r$; this has the effect of removing the influence of any points that are more than a distance r from a prospective estimate θ.[4] We will find it useful to distinguish between M-estimators that are redescending and those that are not.

The influence function of an M-estimator is proportional to its ψ function, and perhaps for this reason M-estimators tend to be characterized by their ψ-functions; consequently we also generally focus on ψ-type estimators. We will make the following assumptions about ψ, which we argue are natural and satisfied by all robust location M-estimators we encountered:

- ψ is odd (i.e. $\psi(-x) = -\psi(x)$), with $\psi(x) \geq 0 \; \forall x \geq 0$.
- ψ is bounded, with $|\psi|$ denoting its supremum, and also piecewise continuous.
- There exists some neighborhood around 0 in which $\psi(x) = 0$ implies $x = 0$. The parameter ϵ will be understood to be smaller than the radius of this neighborhood.

3.1 Non-redescending M-estimators

We first present our algorithm for the case of non-redescending M-estimators, where ψ is only zero at 0. Typically, $\psi(x)$ is monotonically nondecreasing over its entire domain, and we will assume this from now on. Note that $\lim_{x \to \infty} \psi(x)$ exists, and thus there exists γ such that for all $x \geq \gamma$, $|\psi(x) - \psi(\gamma)| < \frac{\epsilon |\psi|}{2\tau}$.

For any $u \in \mathbb{R}$, define $\Psi(u) = \int \psi(x - u) dF(x)$; we wish to find the value $\theta = T(F)$ where $\Psi(\theta) = 0$. Note that because ψ is nondecreasing, Ψ must be non-increasing, with positive value at $-\infty$ and negative value at ∞.[5] The main purpose of this section is to prove the following:

Theorem 3.3. *Let T be a non-redescending location M-estimator as described above, and let F be a distribution at which T is (σ, τ)-robust. Then for any $0 < \epsilon \leq \sigma\tau$ and any $\delta > 0$, Algorithm 2 returns u within ϵ of $T(F)$ with probability $1 - \delta$, using at most $\tilde{O}(\log \frac{\tau}{\epsilon} \log\log \frac{1}{\delta})$ space and $\text{poly}(\tau, 1/\epsilon, \ln \frac{1}{\delta})$ samples.*

Before proving the theorem, we give the intuition behind the algorithm and show the key lemmas that capture the applications of robustness.

Our algorithm works in two phases, each of which is a form of binary search that uses the sign of (an estimate of) $\Psi(u)$ to determine if our current guess u is too large, too small, or close to the correct answer θ.

[4] These are sometimes referred to as the skipped mean, skipped median, and skipped Huber estimator since outlying points are skipped.

[5] In this case, we can generalize the Definition 3.2 to define $T(F) = \sup\{u : \Psi(u) > 0\}$when $\Psi(u)$ is never 0. Our algorithm works for this definition as well.

Algorithm 2 Non-redescending M-estimators

Main algorithm

$(a, b) = (-\infty, \infty)$.
repeat
  Sample $u \in (a, b)$, using up to ℓ_1 samples.
  Estimate $\Psi(u)$ by $\tilde{\Psi}(u)$, using ℓ_2 samples.
  if $\tilde{\Psi}(u) \geq \frac{\epsilon|\psi|}{2\tau}$ **then** $a \leftarrow u$;
  else if $\tilde{\Psi}(u) \leq -\frac{\epsilon|\psi|}{2\tau}$ **then** $b \leftarrow u$;
  else output u and terminate.
until sampling step fails or $b - a \leq \epsilon$ or r loop iterations finished.

if $b - a \leq \epsilon$ **then** output a and terminate;
else (sampling step failed) **call** Low probability phase (a, b).

Low probability phase (a, b)

if either $|\tilde{\Psi}(a + \gamma)| < \frac{\epsilon|\psi|}{2\tau}$ or $|\tilde{\Psi}(b - \gamma)| < \frac{\epsilon|\psi|}{2\tau}$
then output $a + \gamma$ or $b - \gamma$ as appropriate;
else [it must be that $\Psi(u)$ changes sign in $(a, a + \gamma)$ or $(b - \gamma, b)$] perform binary search in the appropriate interval.

In the first phase, we begin with the interval $(a, b) = (-\infty, \infty)$ and obtain our next guess by sampling a point inside the interval. Ideally, the probability mass in (a, b) shrinks by a constant factor in each round. The second phase begins when the probability in (a, b) becomes too small to sample and is a more traditional binary search in which the next guess is the midpoint of the interval (a, b). The search terminates when the length of the interval becomes smaller than ϵ, or we find a point u such that $|\tilde{\Psi}(u)| < \frac{\epsilon|\psi|}{2\tau}$. Robustness is critical in three ways: 1) in Lemma 3.4, it tells us that if $|\Psi(u)|$ is too small to be reliable, then u must be close to θ; 2) in Lemma 3.5, it allows to prove a concentration bound on the accuracy of our estimates of $\Psi(u)$; and 3) in Lemma 3.6, it guarantees that θ must be in a small interval when the second phase of the algorithm begins.

Lemma 3.4. *If $|\Psi(u)| < \epsilon|\psi|/\tau$, then $|u - \theta| \leq \epsilon$.*

Proof. Suppose that $u > \theta$ (the case where $u < \theta$ is handled in a similar way); hence $\Psi(u) < 0$. Starting from F, consider now a modified distribution G in which $\min\{\frac{\epsilon}{\tau}, \int_{x<u} dF(x)\}$ of probability mass is moved from the left of u to a value of at least $u + y$ to the right of u, where $\psi(y) \geq |\Psi(u)|\frac{\tau}{\epsilon}$. (Since $|\Psi(u)|\frac{\tau}{\epsilon} < |\psi|$, such a y must exist.) We will define $\Psi_G(u)$ as $\int \psi(x, u) dG(x)$ in analogy to $\Psi(u)$.

If $\int_{x<u} dF(x)$ mass is moved, then $\int_{x<u} dG(x) = 0$ and hence $\Psi_G(u) \geq 0$, implying that $T(G) \geq u$. If $\frac{\epsilon}{\tau}$ is moved, then we have $\Psi_G(u) \geq \Psi(u) + \frac{\epsilon}{\tau}\psi(y) \geq 0$, and again $T(G) \geq u$.

However, since T is assumed to be (σ, τ)-robust at F and $\epsilon < \sigma\tau$, then $T(G) - T(F) \leq \frac{\epsilon\tau}{\tau} = \epsilon$. Noticing that $T(F) \leq u \leq T(G)$ finishes the proof.

Lemma 3.5. *Let $x_1, \ldots, x_\ell$ be ℓ independent samples from F. Then* $\Pr[|\Psi(u)) - \sum_i^\ell \psi(x_i, u)| \geq \frac{\epsilon|\psi|}{2\tau}] \leq \exp(-\frac{\epsilon^2\ell}{8\tau^2})$.

Proof. This follows from a direct application of Hoeffding's inequality after observing that $\psi(x_i, u)$ is a random variable in $[-|\psi|, |\psi|]$ with expectation $\Psi(u)$.

Between this and the preceding lemma, if u is such that $|\tilde{\Psi}(u)| < \frac{\epsilon|\psi|}{2\tau}$, then $|u - \theta| \leq \epsilon$ with high probability.

Lemma 3.6. *Suppose $\int_{x\in(a,b)} dF(x) \leq \alpha$. Then if $a + \gamma \leq b - \gamma$, $\Psi(a + \gamma) - \Psi(b - \gamma) < \frac{\epsilon|\psi|}{2\tau} + 2\alpha|\psi|$.*

Proof. We will consider the contributions to $\Psi(b - \gamma)$ and $\Psi(a + \gamma)$ from the values $x \leq a$, $a < x < b$, and $x \geq b$ separately. For any $x \leq a$ and $x \geq b$, since both $|x - (a + \gamma)| \geq \gamma$ and $|x - (b - \gamma)| \geq \gamma$ then $\psi(x - (a + \gamma)) - \psi(x - (b + \gamma)) < \frac{\epsilon|\psi|}{2\tau}$ and hence $\int_{x\leq a \text{ or } x\geq b}[\psi(x-(a+\gamma))-\psi(x-(b-\gamma))]dF(x) < \frac{\epsilon|\psi|}{2\tau}$. Further, since $\int_{x\in(a,b)} dF(x) \leq \alpha$, $\int_{x\in(a,b)}[\psi(x - (a + \gamma)) - \psi(x - (b - \gamma))]dF(x) \leq 2\alpha|\psi|$. Summing the contributions gives the desired result.

This leads to the following.

Corollary 3.7. *If $\int_{x\in(a,b)} dF(x) \leq \frac{\epsilon}{4\tau}$, and $\theta \in (a + \gamma, b - \gamma)$, then at least one of $|\Psi(a + \gamma)|$ or $|\Psi(b - \gamma)|$ must be at most $\frac{\epsilon|\psi|}{2\tau}$.*

Proof. From Lemma 3.6, $\Psi(a + \gamma) - \Psi(b - \gamma) \leq \frac{\epsilon|\psi|}{\tau}$. Combining this with the fact that Ψ is zero somewhere in the interval $(a + \gamma, b - \gamma)$ finishes the proof.

Note that this proves the bracketed comment in the final else statement in the description of Algorithm 2.

We are now ready to prove Theorem 3.3.

Proof of Theorem 3.3. There are three sources of failure probability: sampling failure despite sufficiently high remaining probability in (a, b), estimation failure, and failure to reduce the range (a, b) sufficiently even after many rounds. We will set each of these failure probabilities to be at most $\frac{\delta}{3}$.

On each round when we perform a range update, with probability 1/2 we reduce $\int_{x\in(a,b)} dF(x)$ by at least a factor of 3/4. Thus, by a Chernoff bound, $r >$

$16\ln\frac{3}{\delta}+4\log_{3/4}\alpha$ iterations of the main loop of the algorithm will be sufficient to obtain $\int_{x\in(a,b)}dF(x)\leq\alpha$ with probability at least $1-\delta/3$, so long as the sampling step succeeds while $\int x\in(a,b)dF(x)>\alpha$. If $\int_{x\in(a,b)}dF(x)>\alpha$, then the probability that ℓ_1 independent samples are all outside (a,b) on any round is at most $r(1-\alpha)^{\ell_1}$. Thus if we set $\ell_1>\frac{1}{\alpha}\left(\ln\frac{3r}{\delta}\right)$, the probability that any of the sampling steps fails erroneously is at most $\delta/3$. We set $\alpha\leq\frac{\epsilon}{4\tau}$, so that by Corollary 3.7, if $\theta\in(a+\gamma,b-\gamma)$, one of $a+\gamma$ or $b-\gamma$ is within ϵ of θ, and further, detectable by estimating its $\tilde{\Psi}$ value.

The low probability phase lasts for at most $\log_2\gamma/\epsilon$ rounds, making for a total of $r+\log_2\gamma/\epsilon$ estimations of Ψ. By a union bound, all $r+\log_2\gamma/\epsilon$ estimates of $|\Psi(u)|$ are correct to within $\frac{\epsilon|\psi|}{2\tau}$ with probability at least $1-(r+\log_2\gamma/\epsilon)\exp\left(-\frac{\epsilon^2\ell_2}{2\tau^2}\right)$. Thus, $\ell_2>\frac{2\tau^2}{\epsilon^2}\left(\ln\frac{3(r+\log_2\gamma/\epsilon)}{\delta}\right)$ is sufficient.

The total number of samples needed is at most $r(\ell_1+\ell_2)+2\ell_2\log_2\gamma/\epsilon$. The space required is the logarithm of the max counter values for the ℓ_1 and ℓ_2 counters, plus a storing a constant number of samples, each of unit size.

3.2 Redescending M-estimators

Algorithm 3 Redescending M-estimators

Let $\xi_1,\ldots,\xi_{t_2}\in\mathbb{R}$ be a precomputed sequence based on ρ.
for $i=1,\ldots,t_1$ **do**
  Sample x_i from F.
  for $j=1,\ldots,t_2$ **do**
    Let $y_{ij}=x_i+\xi_j$.
    Estimate $R(y_{ij})$ using t_3 samples; if this is the minimum estimate thus far let $u=y_{ij}$.
  end for
end for
Output u.

We now turn to redescending M-estimators. Recall that for a redescending estimator T there exists some r for which $\psi(x)=0$ whenever $|x|>r$; let $r^*=\inf\{r>0:\psi(r)=0 \text{ when } |x|>r\}$. Since ψ and thus Ψ are no longer monotonic, there may not be a unique value u for which $\Psi(u)=0$. We choose one common method of remedying this by defining $T(F)$ using Definition 3.1 instead; thus $T(F)=\text{argmin}_\theta\left(\int\rho(x-\theta)dF(x)\right)$.[6]

[6]Other common methods include taking the value of u satisfying $\Psi(u)=0$ closest to the median, or using Newton's method on Ψ using the median as the starting point.

We will make the reasonable assumptions that $\rho(x)$ is bounded and continuous and will use $|\rho|$ to denote $|\rho|=\sup_x|\rho(x)|$. Let $\gamma(x)=\sup_z|\rho(z+x)-\rho(z)|$, or the maximum possible change in ρ over an interval of length x, and let $\beta(x)$ be the value of z for which this is achieved (i.e. $\rho(\beta(x)+x)-\rho(\beta(x))=\gamma(x)$.

Our main theorem here is this, which we prove after showing several useful lemmas.

Theorem 3.8. *Let T be a redescending location M-estimator, and let F be a distribution at which T is (σ,τ)-robust. Then for any $0<\epsilon\leq\sigma\tau$ and any $\delta>0$, Algorithm 3 returns u within ϵ of $T(F)$ with probability $1-\delta$, using at most $\text{poly}(\tau,|\rho|,\epsilon^{-1},(\gamma(\epsilon/2))^{-1},\ln\delta^{-1})$ samples and $\tilde{O}(\log\frac{\tau|\rho|}{\epsilon\gamma(\epsilon/2)}\log\log\frac{1}{\delta})$ space.*

The non-monotonic nature of ψ means we cannot directly apply the binary search approach described above, and so we appeal to the ρ function instead. Analogously with $\Psi(u)$, define $R(u)=\int\rho(x,u)dF(x)$; where there is the possibility of confusion, we will indicate the distribution as in $R_F(u)$. We wish to find $\theta=T(F)$, the global minimum of R over $\mathbb{R}$. On the surface, this may appear difficult, as we need to be lucky enough to sample or find a point near the global minimum of R, and successfully distinguish it from all other candidate points, some of whose own R-values may *a priori* be close to that of θ.

Fortunately, if we have a robustness guarantee, the problem becomes manageable. First, the following lemma guarantees that for any point u sufficiently far from θ, there is a gap between $R(u)$ and $R(\theta)$. Otherwise, a small change to F would cause $R(\cdot)$ to be smaller at u than at any point in a neighborhood around θ. Thus θ should be recognizable as the global minimum if we can find it.

Lemma 3.9. *Let $T(F)=\theta$. Then for all u, we have that $R(u)-R(\theta)\geq\frac{z}{2\tau}\gamma(\frac{z}{2})$, where $z=\min\{|u-\theta|,2\sigma\tau\}$.*

Proof. The result is trivial if $u=\theta$, so suppose that $u\neq\theta$ and that in fact $R(u)-R(\theta)<\frac{z}{2\tau}\gamma(\frac{z}{2})$. Intuitively, we will show that since $R(u)$ and $R(\theta)$ are relatively close, a small change to the distribution F will result in a nearby distribution F' for which $R_{F'}(\cdot)$ will be smaller at u than at any point close to θ. This will contradict the robustness of T at F.

Assume without loss of generality that $u\geq\theta$ (a similar argument holds for the other case). Specifically, let H be the subdistribution of F with $\int dH(x)\leq\frac{z}{2\tau}$ that maximizes $\int(\rho(x-\theta)-\rho(x-u)dH(x)$. Our modified distribution F' will then take the form $F'=$

$F - H + (\int dH(x))\Delta_{u+\beta(z/2)}$. Essentially, this moves the (at most) $\frac{z}{2\tau}$ of probability that most favors θ over u, and moves it to the point where it most favors u over θ. Note that $d(F, F') \leq \frac{z}{2\tau} \leq \sigma$.

We now claim that in the new distribution, $R_{F'}(u) < R_{F'}(y)$ for all $y \in [\theta - \frac{z}{2}, \theta + \frac{z}{2}]$. If $\int dH(x) < \frac{z}{2\tau}$ then H contains all of the parts of F for which $\rho(x-\theta) > \rho(x-u)$, and the claim follows. Otherwise, consider any such y in the given interval. We observe that since $\frac{z}{2} \leq \frac{|u-\theta|}{2}$, $|u-y| \geq \frac{z}{2}$. Then by definition of $\gamma(\cdot)$, we have that

$$\begin{aligned}(R(u) - R_{F'}(u)) - (R(y) - R_{F'}(y)) & \\ \geq & \frac{z}{2\tau}\gamma(\frac{z}{2}) \\ > & R(u) - R(\theta) \\ \geq & R(u) - R(y).\end{aligned}$$

Rearranging the terms yields $R_{F'}(y) - R_{F'}(u) > 0$.

Since $R_{F'}(\cdot)$ is smaller at u than at any point in $[\theta - \frac{z}{2}, \theta + \frac{z}{2}]$, we know that $T(F') \notin [\theta - \frac{z}{2}, \theta + \frac{z}{2}]$. Hence $|T(F') - T(F)| > \frac{z}{2} \geq \tau d(F, F')$, a contradiction.

Lemma 3.9 states that in particular, for any u with $|u - \theta| > \epsilon$, $R(u) - R(\theta) > \frac{\epsilon}{2\tau}\gamma(\frac{\epsilon}{2})$. Let $0 = \xi_1 < \xi_2 < \cdots < \xi_{t_2} = r^*$ be an increasing sequence of reals such that $\rho(\xi_j) - \rho(\xi_{j-1}) \leq \frac{\epsilon}{8\tau}\gamma(\frac{\epsilon}{2})$. The idea is that for any pair of points a and b with $a < b$ such that $|\rho(b) - \rho(a)| > \frac{\epsilon}{8\tau}\gamma(\frac{\epsilon}{2})$, there must exist a j for which $\xi_j \in [a, b]$. Note that this sequence can be constructed using $t_2 \leq \frac{8\tau|\rho|}{\epsilon\gamma(\frac{\epsilon}{2})} + 1$ points.

With this sequence in hand, we can lower bound the probability that sampling a point will lead us close to θ. We observe that the gap $R(u) - R(\theta)$ for all points u outside a neighborhood around θ implies a lower bound on the average derivative of $R(\cdot)$ around θ. Since ψ (the derivative of ρ, and the potential contribution to the derivative of $R(\cdot)$ from a single particle of probability) is bounded from above, many points must contribute to the relatively large average derivative of R near θ. This is the content of the following lemma.

Lemma 3.10. *Let x be a random sample from the distribution F. Then*

$$\Pr\left[\exists\ j\ s.t.\ R(x+\xi_j) - R(\theta) \leq \frac{\epsilon}{4\tau}\gamma(\frac{\epsilon}{2})\right] \geq \frac{\frac{\epsilon}{8\tau}\gamma\left(\frac{\epsilon}{2}\right)}{|\rho| - \frac{\epsilon}{8\tau}\gamma(\frac{\epsilon}{2})} \geq \frac{\epsilon\gamma(\frac{\epsilon}{2})}{8\tau|\rho|}).$$

Proof. From Lemma 3.9, we know that $R(\theta+\epsilon) - R(\theta) > \frac{\epsilon}{2\tau}\gamma(\frac{\epsilon}{2})$. Let $0 \leq z \leq \epsilon$ be the smallest value for which $R(\theta+z) - R(\theta) > \frac{\epsilon}{4\tau}\gamma(\frac{\epsilon}{2})$; thus, $\int[\rho(x - (\theta+z)) - \rho(x-\theta)]dF(x) > \frac{\epsilon}{4\tau}\gamma(\frac{\epsilon}{2})$. Given this lower bound on the average value of $\rho(x-(\theta+z)) - \rho(x-\theta)$ over all x and the upper bound of $|\rho|$ on the value of $\rho(\cdot)$, we can see that for a certain fraction of x, $\rho(x - (\theta+z)) - \rho(x-\theta) > \frac{\epsilon}{8\tau}\gamma(\frac{\epsilon}{2})$; a quick calculation shows this fraction to be $\frac{\frac{\epsilon}{8\tau}\gamma(\frac{\epsilon}{2})}{|\rho| - \frac{\epsilon}{8\tau}\gamma(\frac{\epsilon}{2})}$. By the construction of ξ_j, when we find such a value of x, there must exist some j for which $x + \xi_j \in [\theta, \theta + z]$. This finishes the proof.

Given this, our algorithm is straightforward in the extreme. It merely draws a sufficiently large number of samples, and for each sample x_i examines all points $x_i + \xi_j$ where x_i would induce a large derivative. Theorem 3.8 implies the algorithm's correctness.

Proof of Theorem 3.8. Our algorithm will succeed if (1) it samples an x_i for which there exists j such that $R(x_i + \xi_j) - R(\theta) \leq \frac{\epsilon}{4\tau}\gamma(\frac{\epsilon}{2})$ (in which case $x_i + \xi_j \in [\theta, \theta + \epsilon]$), and (2) the estimate $\tilde{R}(x_i + \xi_j)$ is smaller than $\tilde{R}(y_{ij})$ for all $y_{ij} \notin [\theta - \epsilon, \theta + \epsilon]$. This second condition will hold so long as all of the $\tilde{R}$ estimates have error at most $\frac{\epsilon}{8\tau}\gamma(\frac{\epsilon}{2})$ (as we have a gap of twice that size between $R(x_i + \xi_j)$ and $R(y_{ij})$.

By Lemma 3.10, condition (1) fails to hold with probability at most $\left(1 - \frac{\epsilon}{8\tau|\rho|}\gamma(\frac{\epsilon}{2})\right)^{t_1}$, which is less than $\frac{\delta}{2}$ when $t_1 > \frac{8\tau|\rho|}{\epsilon\gamma(\frac{\epsilon}{2})}\ln(\frac{2}{\delta})$.

Using a Hoeffding bound, when $t_3 > \frac{64\tau^2|\rho|^2}{\epsilon^2\gamma(\frac{\epsilon}{2})^2}\ln\frac{2t_1t_2}{\delta}$, then $\Pr[|R(y) - \tilde{R}(y)| > \frac{\epsilon}{8\tau}\gamma(\frac{\epsilon}{2})] < \frac{\delta}{2t_1t_2}$. Applying a union bound, all t_1t_2 estimates of $R(\cdot)$ satisfy the bound with probability at least $1 - \frac{\delta}{2}$.

The number of samples used is $t_1t_2t_3$. As observed above, $t_2 \leq \frac{8\tau|\rho|}{\epsilon\gamma(\epsilon/2)} + 1$. The space requirement is logarithmic in the maximum size of the counters t_1, t_2, t_3, plus a constant number of samples, each of unit size.

4 L-estimators

Another major class of robust estimators is that of L-estimators, or "linear combination of order statistics." In functional form, an L-estimator T is defined by a function $h : [0,1] \to \mathbb{R}$ as follows:[7]

$$T(F) = \frac{\int_0^1 F^{-1}(t)h(t)dt}{\int_0^1 h(t)dt}. \quad (1)$$

The classic example of an L-estimator is the median, defined by $h = \delta_{1/2}$, or the Dirac delta function at

[7] Another very similar definition is $T(F) = \frac{\int xh(F(x))dF(x)}{\int h(F(x))dF(x)}$. The two behave differently where F is discontinuous (i.e. has a point mass); in these cases we believe our given definition is more appropriate.

$\frac{1}{2}$. Also well-known is the α-trimmed mean for some $\alpha < \frac{1}{2}$, defined by $h(t) = 1$ for $t \in [\alpha, 1-\alpha]$ and $h(t) = 0$ elsewhere; thus the $\frac{1}{5}$-trimmed mean is the mean of the middle 60% of the distribution. A frequently used measure of spread is the interquartile distance, specified by $h(t) = \delta_{3/4} - \delta_{1/4}$.

Our algorithm for approximating L-estimators to within an additive ϵ error works essentially by discretizing the numerator in equation (1) above. To that end, we partition the interval $[0, 1]$ into $\frac{1}{w}$ subintervals, where w is chosen so that $w\tau \leq \epsilon$. We will denote these subintervals I_k, $k = 1, \ldots, 1/w$, where $I_k = [(k-1)w, kw]$.

Algorithm 4 L-estimators

for $k = 1, \ldots, 1/w$ **do**
  Compute (exactly) $\int_{I_k} h(t)dt$; denote this quantity by h_k.
  Set $t_k = (k - \frac{1}{2})w$; i.e. the midpoint of I_k.
  Estimate $x_k = F^{-1}(t_k)$, using Algorithm 1, with parameters $(w/2, w\delta)$.
  Let $X_k = h_k x_k$.
end for
Output $\sum_k X_k$.

Theorem 4.1. *Let T be an L-estimator and F be a distribution at which T is (σ, τ)-robust. Then Algorithm 4 outputs a value u within ϵ of $T(F)$ with probability at least $1-\delta$, using space $\tilde{O}(\log \frac{\tau}{\epsilon} \log\log \frac{1}{\delta})$, for any $0 < \epsilon \leq \sigma\tau$ and any $\delta > 0$.*

Proof. By (σ, τ)-robustness, it is sufficient to show that with probability at least $1 - \delta$ there exists a distribution F' such that (a) $T(F') = u$ and (b) $d(F, F') < \frac{\epsilon}{\tau}$.

We first observe that in the main loop of our algorithm, the value x_k returned by a call to Algorithm 1 is equal to $F^{-1}(t'_k)$ for some $t'_k \in [0, 1]$.[8] Hence, we have $u = \sum_{k=1}^{1/w} h_k F(t'_k)$. Further, by Lemma 2.6, for each k, there exists t'_k such that $|t_k - t'_k| \leq \frac{w}{2}$ with probability at least $1 - w\delta$. By a union bound, we obtain

$$\Pr[|t_k - t'_k| \leq \frac{w}{2} \ \forall k] \geq 1 - \delta. \tag{2}$$

In particular, when this holds, we have that $t'_k \in [(k-1)w, kw]$.

We now claim that when (2) holds, the distribution $F' = \sum_{k=1}^{1/w} w\Delta_{x_k}$ satisfies the two conditions above. Condition (a) holds by construction: note that

$$\begin{aligned} T(F') &= \int_0^1 F'^{-1}(t)h(t)dt \\ &= \sum_{k=1}^{1/w} \int_{I_k} F'^{-1}(t)h(t)dt \\ &= \sum_{k=1}^{1/w} h_k x_k = u \ . \end{aligned}$$

As for condition (b), it is sufficient to show that for all x, $|F(x) - F'(x)| \leq w$. For convenience, let x_0 and t'_0 be understood to be $-\infty$ and 0 respectively. Now consider any x and the largest value of k such that $x_k \leq x$; thus either $k = \frac{1}{w}$, an easily handled special case, or $x < x_{k+1}$ and hence $F(x) \leq F(x_{k+1})$. By the construction of F', we have $F'(x_k) = F'(x) = kw$. We also have that $t'_k \leq F(x)$, and that $F(x) < t'_{k+1}$. Now, by (2), $t'_k \geq (k-1)w$ and, if applicable, $t'_{k+1} \leq (k+1)w$, and so $|F(x) - F'(x)| \leq w$, as needed.

The space requirement is logarithmic in the counter that ranges from 1 to $1/w$, plus the space required by the call to Algorithm 1, plus space for a constant number of samples, each of unit size.

5 R-estimators

Given a distribution F and a value $u \in \mathbb{R}$, let $\bar{F}$ be the reflection of F across u; i.e. if f exists, $\bar{f}(x) = f(2u - x)$. Let G be the combined distribution $\frac{1}{2}F + \frac{1}{2}\bar{F}$, or equivalently, $G = \frac{1}{2}[1 + F(x) - F(2u-x)]$. Then for any function $a : [0, 1] \to \mathbb{R}$ that is skew-symmetric around $\frac{1}{2}$ (namely, $a(\frac{1}{2}+y) = -a(\frac{1}{2}-y)$) and for which $a(x) \geq 0$ for $x > \frac{1}{2}$), we can define the related function $S(u) = \int a_G(x)dF(x)$, where $a_G(x)$ is the average value of $a(\cdot)$ from $G(x_-)$ to $G(x_+)$, the limits of $G(x)$ from the left and right.[9] This leads to a definition for a new class of estimators called R-estimators. For convenience, we define $|a| = \int_{y > \frac{1}{2}} a(y)$, and assume without loss of generality that $|a| \geq 1$.

Definition 5.1. *Given a function $a : [0, 1] \to \mathbb{R}$, the associated R-estimator $T(F)$ is a value of θ for which $S(\theta) = 0$, if one exists.*

The median is the R-estimator with $a(t) = \text{sign}(t - \frac{1}{2})$. Other noteworthy R-estimators include the normal scores estimator ($a(t) = \Phi^{-1}(t)$, where Φ is the normal distribution), and the Hodges-Lehmann estimator ($a(t) = t - \frac{1}{2}$), derived from the well-established Wilcoxon signed-rank test.

[8] Note that there may be more than one such value of t'_k if F is discontinuous at x_k.

[9] For reasons similar to those in Section 4, we use this definition instead of the more compact $\int a(G(x))dF(x)$ to handle the case where F (and hence G) has discontinuities.

Algorithm 5 EstimateS$(u; \epsilon, \delta)$

Set $w = \frac{\epsilon}{4}$.
Define $a_k = \int_{x \in (\frac{1}{2}+(k-1)w, \frac{1}{2}+kw]} a(t)dt$.
Define $b_k = G^{-1}(u + (k-1)w)$ for $k = 0, \ldots, \frac{1}{2w}$, and $I_k = (b_{k-1}, b_k]$.
for $k = 1, \ldots, \frac{1}{2w}$ **do**
  Let $\tilde{b}_k$ be the estimate of $G^{-1}(\frac{1}{2} + kw)$ from Algorithm 1, with parameters $(\frac{w\epsilon}{16+2\epsilon}, \frac{w\delta}{2})$; let $\tilde{I}_k = (\tilde{b}_{k-1}, \tilde{b}_k]$.
  Estimate $c_k = \frac{\int_{\tilde{I}_k} dF(t)}{\int_{\tilde{I}_k} dG(t)}$ using the first t samples from G in the interval $\tilde{I}_k$. (To sample from G, take a sample x from F and return either x or $2u - x$ each with probability $\frac{1}{2}$.)
end for
Output $w \sum_k a_k(2c_k - 1)$.

Algorithm 6 R-estimators

Main
$(a, b) = (-\infty, \infty)$
repeat
  Sample $u \in (a, b)$ from F using up to ℓ_1 samples.
  Let $\tilde{S}(u) = \text{EstimateS}(u; \frac{\epsilon}{2\tau}; \frac{\delta}{3(r+r_2)})$. [Parameters r, r_2 defined in text.]
  if $|\tilde{S}(u)| > |a|\frac{\epsilon}{2\tau}$ **then**
    Update (a, b).
  end if
until sampling step fails or r loop iterations finished or $|\tilde{S}(u)| \leq |a|\frac{\epsilon}{2\tau}$ or $b - a \leq \frac{\epsilon}{2}$
if $|\tilde{S}(u)| \leq |a|\frac{\epsilon}{2\tau}$ or $b - a \leq \frac{\epsilon}{2}$ **then**
  Output u.
else {sampling step failed}
  call Low probability phase (a, b).
end if

Low probability phase (a, b)

If either a or b is infinite, then output the finite one and terminate.
repeat
  Let $u = \frac{a+b}{2}$.
  Let $\tilde{S}(u) = \text{EstimateS}(u; \frac{\epsilon}{2\tau}; \frac{r_2\delta}{3(r+r_2\log_2(b-a)/\epsilon)})$.
  if $|\tilde{S}(u)| > |a|\frac{\epsilon}{2\tau}$ **then**
    Update (a, b).
  end if
until $|\tilde{S}(u)| \leq |a|\frac{\epsilon}{2\tau}$ or $b - a \leq \frac{\epsilon}{2}$
Output u.

The formulation of R-estimators combines elements from both M-estimators and L-estimators. They resemble M-estimators in that there is an aggregate function $S(\cdot)$, analogous to $\Psi(\cdot)$ in the non-redescending case, for which we wish to find a zero. On the other hand, R-estimators and L-estimators also evaluate weighted sums of ranks, though the R-estimator variant is more complicated in that its sum depends on two different distributions (F and G) rather than one.

Nonetheless, as the following theorem states, robust R-estimators can be approximated in a space-efficient manner, and the fact that even this more complicated class can be handled gives us confidence that the classes we examine are examples of a more general principle, as opposed to a few fortunate instances.

Theorem 5.2. *Given an R-estimator T defined by $a : [0, 1] \to \mathbb{R}$ and a distribution F at which T is (σ, τ)-robust, Algorithm 6 produces a value within ϵ of $T(F)$ with probability at least $1-\delta$, for any $0 \leq \epsilon \leq \sigma\tau$ and any $\delta > 0$ using at most $\tilde{O}(\log \frac{\tau}{\epsilon} \log\log \frac{C}{\delta})$ space, where C is the maximum length of an interval (a, b) containing $T(F)$ for which $\int_a^b dF(x) \leq \frac{\epsilon}{\tau}$.*

The dependence on $\log\log C$ in the space bound is admittedly inelegant, but is significant only when F has very little density in an extremely large range around $T(F)$. We also note that C is comparable in size to sample points a and b, and thus reflects the spread of F.

The algorithm performs binary search as in Algorithm 2, using the sign of $S(u)$ for comparison. We therefore need a subroutine that gives us a useful estimate of the value of $S(u)$ for any u. This is given by Algorithm 5 and the following lemma:

Lemma 5.3. *With probability at least $1 - \frac{\delta}{2}$, the algorithm EstimateS(u) outputs a value that is the correct value of $S(u)$ for a distribution F' with $d(F, F') \leq \epsilon$.*

Informally, the EstimateS algorithm evaluates the integral $\int a_G(x)dF(x)$ by interpreting it as $\int_0^1 a(t)\frac{f(G^{-1}(t))}{g(G^{-1}(t))}dt$. In a sense, we are integrating $a(t)$ over the $[0, 1]$ interval while weighting by the fraction of $g(G^{-1}(t))$ that belongs to F. The algorithm does this by discretizing the $[0, 1]$ interval into slices of size w, while also taking advantage of the realization that because $a(\cdot)$ is skew-symmetric around $\frac{1}{2}$, we only need to consider the interval $(\frac{1}{2}, 1]$.

Proof. We first note that by a union bound, with

probability at least $1-\frac{\delta}{2}$, we can assume

$$\exists\ t_k \text{ s.t.} \quad G^{-1}(t_k) = \tilde{b}_k \quad \text{and} \quad \left| t_k - \left(\frac{1}{2} + kw\right)\right| \le \frac{w\epsilon}{16+2\epsilon}, \quad \forall k = 1,\ldots,\frac{1}{2w}. \tag{3}$$

Among other things, this implies that $\tilde{b}_{k-1} \le \tilde{b}_k$ for all k. Note that our algorithm will return the value $w\sum_{k=1}^{1/2w} a_k(2c_k-1)$; we can see that this would be the correct value of $S(u)$ for the underlying distribution $F' = w\sum_{k=1}^{1/2w} c_k\Delta_{\tilde{b}_k} + (1-c_k)\Delta_{2u-\tilde{b}_k}$. We need to show that $d(F,F') \le \epsilon$ with high probability. The following claim will be useful:

Claim 5.4. *With probability at least $1-\delta$, $|\int_{I_k} dF(x) - c_k w| \le \frac{w\epsilon}{4}$, or equivalently, $|\frac{\int_{I_k} dF(x)}{w} - c_k| \le \frac{\epsilon}{4}$, for all k.*

In words, we claim that all of our estimates of the proportion of G that belongs to F in each slice are reasonably accurate.

Proof. For each k, our estimate c_k is derived by taking t independent samples from G conditioned on their falling into the interval $\tilde{I}_k$, and observing what fraction of these belong to F. There are two possible sources of error—first, the proportions of F in the intervals $\tilde{I}_k$ and I_k will not match exactly; and second, there will be sampling error when trying to measure these proportions. We deal with both of these below.

Our algorithm will be measuring $\int_{\tilde{I}_k} dF(x)/\int_{\tilde{I}_k} dG(x)$, and we want to argue that this is close to $\int_{I_k} dF(x)/\int_{I_k} dG(x)$.[10] Recall now that we assumed that $|t_j - (\frac{1}{2}+jw)| \le \frac{w\epsilon}{16+2\epsilon}$ for all j. Hence

$$\left|\frac{\int_{\tilde{I}_k} dF(x)}{\int_{\tilde{I}_k} dG(x)} - \frac{\int_{I_k} dF(x)}{\int_{I_k} dG(x)}\right| \le \sup_{\alpha_1,\alpha_2,\beta_1,\beta_2}\left|\frac{\int_{I_k} dF(x)+\alpha_1+\alpha_2}{\int_{I_k} dG(x)+\beta_1+\beta_2} - \frac{\int_{I_k} dF(x)}{\int_{I_k} dG(x)}\right|,$$

where $|\beta_1|, |\beta_2| \le \frac{w\epsilon}{16+2\epsilon}$, $|\alpha_i| \le |\beta_i|$, and the signs of α_i and β_i are the same for $i \in \{1,2\}$. A routine calculation show that this must be bounded by

$$\max_{\alpha_1,\beta_1} \frac{2\alpha_1}{\int_{I_k} dG(x) - 2\beta_1} \le \frac{\frac{w\epsilon}{8+\epsilon}}{w - \frac{w\epsilon}{8+\epsilon}} = \frac{\epsilon}{8}. \tag{4}$$

For the sampling error, by applying a Chernoff bound, we find that as long as $t \ge \frac{32}{\epsilon^2}\ln\frac{2}{w\delta}$, then $\Pr[|c_k - \frac{\int_{\tilde{I}_k} dF(x)}{\int_{\tilde{I}_k} dG(x)}| \ge \frac{\epsilon}{8}] \le \frac{w\delta}{2}$. By a union bound, the probability that there is even one interval where this happens is at most $\frac{\delta}{2}$.

Combining this with (4), we find that $\left|\frac{\int_{I_k} dF(x)}{w} - c_k\right| \le \frac{\epsilon}{4}$ with probability at least $1-\delta$, as needed.

With the claim in hand, we can finish the proof that $d(F,F') \le \epsilon$. In the context of the above proof, for all $k \ge 1$, let $v_k = \sum_{i=1}^{k}\int_{I_k} dF(x)$ and $\tilde{v}_k = \sum_{i=1}^{k} c_k w$. Then the claim implies that for all k,

$$|v_k - \tilde{v}_k| \le \frac{\epsilon}{4}. \tag{5}$$

Further, from inequality (3), we have that $t_{k-1} \le kw \le t_{k+1}$ for all k. Now consider any $x \in \mathbb{R}$; it suffices to show that $|F(x) - F'(x)| \le \epsilon$. Let j be the largest value for which $\tilde{b}_j \le x$; thus $F'(x) = \tilde{v}_j$. If $x \le b_j$, then (5) immediately gives us what we need. Otherwise, since $x < \tilde{b}_{j+1}$, then $G(x) \le t_{j+1} \le \frac{1}{2} + (j+1)w + \frac{w\epsilon}{16}$. On the other hand, since $b_j < x$, $\frac{1}{2}+jw \le G(x)$, and so $G(x)-G(b_j) \le w+\frac{w\epsilon}{16}$ and thus $F(x)-F(b_j) \le 2(w+\frac{w\epsilon}{16})$ by (3). Expression (5) yields that $|F(b_j) - F'(x)| \le \frac{\epsilon}{4}$, and so $|F(x) - F'(x)| \le \epsilon$ for $w \le \frac{\epsilon}{4}$.

This finishes the proof of Lemma 5.3

[10] There is a technicality that needs to be addressed when F and G are discontinuous at some of the b_k and $\tilde{b}_k$. In this case, when measuring $\int_I dF(x)$ or $\int_I dG(x)$ for some interval I, we may introduce substantial error if (for example) $G(b_k) > \frac{1}{2}+kw$. In this case, the EstimateS algorithm will include all samples of b_k as belonging to I_k, as opposed to only the fraction that should be attributed to the range ending at $\frac{1}{2}+kw$.

This can be fixed by making several minor changes. First, we no longer treat all samples of a particular value as indistinguishable. Instead, each time a sample s is drawn, it is associated with a newly chosen uniform random number r in $[0,1]$ that signifies its rank among all samples with the same value. Thus $(s_1,r_1) < (s_2,r_2)$ if $s_1 < s_2$ or if $s_1 = s_2$ and $r_1 < r_2$. Algorithm 1 is then modified to take this ordering into account. Thus, when looking for the median of the distribution $\frac{1}{4}\Delta_{-1} + \frac{3}{4}\Delta_1$, Algorithm 1 will with high probability return something of the form $(1,r)$ for some r close to $\frac{1}{3}$.

Further, the expression $\int_{I_k} dF(x)$ (for example) should also be interpreted to take this into account. An interval I_k will then be of the form $((s_1,r_1),(s_2,r_2)]$, and the Monte Carlo sampling in EstimateS will use the modified comparison as well. Rather than introduce more notation and complexity to the exposition, however, we will retain the more familiar integral form above, with the implicit understanding that these changes may be necessary when F is not continuous.

Since the main R-estimator algorithm stops when it finds a small estimated value of $S(u)$, we need to show that when $S(u)$ is small, then u must be close to θ. This is shown by the following lemma.

Lemma 5.5. *Suppose* $|S(u)| \leq |a|\frac{\epsilon}{\tau}$. *Then* $|u-\theta| < \epsilon$.

Proof. For ease of exposition, we will assume that $0 < S(u) \leq |a|\frac{\epsilon}{\tau}$, with the case of $S(u) < 0$ handled analogously. Thus $T(F) > u$, or more precisely, there exists a value $\theta > u$ for which $S(\theta) = 0$.

For any x, let $\bar{x} = 2u - x$ be its reflection across u, and for a subset $I \subseteq \mathbb{R}$, let $\bar{I} = \{\bar{x} : x \in I\}$. Then note that $G(x) = 1 - G(\bar{x})$, and in particular $G(u) = \frac{1}{2}$. Along with the skew-symmetry of a, this further implies that over any subset $I \subseteq \mathbb{R}$, we have

$$\int_I a_G(x)dF(x) = -\int_{\bar{I}} a_G(x)d\bar{F}(x). \tag{6}$$

By definition, we have

$$S(u) = \int_{x>u} a_G(x)dF(x) + \int_{x<u} a_G(x)dF(x)\ .$$

Also, since u is the median of the distribution G, we have $\int_{x>u} a_G(x)dF(x) + \int_{x>u} a_G(x)d\bar{F}(x) = |a|$, and from equation (6) this becomes $\int_{x>u} a_G(x)dF(x) - \int_{x<u} a_G(x)dF(x) = |a|$. Summing this with the previous equality yields that

$$\int_{x>u} a_G(x)dF(x) = \frac{S(u)+|a|}{2}. \tag{7}$$

We now show there exists a distribution F' for which $d(F,F') < \frac{\epsilon}{\tau}$ and for which $S_{F'}(u) \stackrel{\text{def}}{=} \int a_{G'}(x))dF'(x) \leq 0$. This would imply that there exists a value for $T(F') \leq u$, and thus $|u-\theta| < \epsilon$. We can construct F' from F in a natural way, by moving probability mass from the values $x > u$ in F for which $a_G(x)$ is largest to the corresponding reflected values $\bar{x} = 2u - x$. More formally, let H be the subdistribution of F with $\int dH(x) \leq \frac{\epsilon}{\tau}$ that maximizes $\int a_G(x)dH(x)$. If $\int dH(x) < \frac{\epsilon}{\tau}$, then all density where $a_G(x) > 0$ has been removed from F and thus $S_{F'}(u) < 0$ immediately. We can therefore assume $\int dH(x) = \frac{\epsilon}{\tau}$, and therefore

$$\int a_G(x)dH(x) \geq \frac{\epsilon}{\tau}\frac{S(u)+|a|}{2}. \tag{8}$$

We then set $F' = F - H + \bar{H}$.

It remains to argue that $\int a_{G'}(x)dF'(x) \leq 0$. To see this, we observe that

$$\begin{aligned}
&\int a_{G'}(x)dF'(x)\\
&\quad= \int a_{G'}(x)dF(x)\\
&\qquad+ \int a_{G'}(x)(d\bar{H}(x) - dH(x))\\
&\quad= S(u) + \int a_G(x)(d\bar{H}(x) - dH(x))\\
&\quad= S(u) - 2\int a_G(x)dH(x)\\
&\quad\leq S(u) - \frac{\epsilon}{\tau}(S(u)+|a|).
\end{aligned}$$

The second line follows because $G = G'$, as F' was created from F by reassigning probability mass from F to $\bar{F}$ and vice versa. The third line follows from (6), and the last line from (8). To finish the proof, we need only observe that the last line is at most zero when $S(u) \leq |a|\frac{\epsilon}{\tau}$.

We are now ready to prove the main theorem for R-estimators.

Proof of Theorem 5.2. We first show the correctness of the algorithm in the absence of failures due to unlucky sampling from F, and then bound the probability of these failures happening.

We start with the following important observation: For any u, let $\tilde{S}(u)$ be the value returned from a successful call to EstimateS$(u; \epsilon^*/\tau, \delta^*)$. We know from Lemma 5.3 that $\tilde{S}(u) = S_{F'}(u)$ for some F' with $d(F,F') \leq \epsilon^*/\tau$. Then either $\text{sign}(S(u)) = \text{sign}(\tilde{S}(u))$, or $|u-\theta| \leq \epsilon^*$. For if $\text{sign}(S(u) \neq \text{sign}(\tilde{S}(u)) = \text{sign}(S_{F'}(u))$, then $T(F)$ and $T(F')$ are on opposite sides of u. From robustness, however, $|T(F) - T(F')| \leq \tau d(F,F') \leq \epsilon^*$. Since u is between $T(F)$ and $T(F')$, $|T(F) - u|$ is also bounded by ϵ^*.

In both sections of the algorithm, we claim that assuming that the EstimateS step succeeds in each iteration, it must be that $\theta \in (a - \frac{\epsilon}{2}, b + \frac{\epsilon}{2})$. The above observation tells us that we will never guess $\text{sign}(S(u))$ incorrectly whenever $|u-\theta| > \frac{\epsilon}{2}$, and thus either $\theta \in (a,b)$ or one endpoint of (a,b) is within distance $\frac{\epsilon}{2}$ of θ.

Our algorithm terminates only when $|\tilde{S}(u)| \leq |a|\frac{\epsilon}{2\tau}$ or when $b-a \leq \frac{\epsilon}{2}$, or if $b-a$ is infinite after the end of the first phase. In the first case, Lemma 5.5 implies that $|T(F') - u| \leq \frac{\epsilon}{2}$ for some F' with $d(F,F') \leq \frac{\epsilon}{2\tau}$. Since $|T(F) - T(F')| \leq \frac{\epsilon}{2}$, $|T(F) - u| \leq \epsilon$. If $b-a \leq \frac{\epsilon}{2}$, then both endpoints are within ϵ of θ, and

returning either one will suffice. Finally, if either a or b is infinite, we can simply return the one that is finite, since with parameters set as described below, with high probability only ϵ/τ total mass remains in the interval (a, b). By moving this remaining probability outside the interval, we achieve a distribution F' in which there is no probability at all in (a, b), and so $T(F') \notin (a, b)$. But now $d(F, F') \leq \frac{\epsilon}{\tau}$, and so by the robustness property, $T(F)$ must be ϵ-close to the finite endpoint.

We now bound the probability of failure. There are three sources of failure probability: sampling failure despite high remaining probability, estimation failure, and failure to reduce the range (a, b) sufficiently even after many rounds.

On each round when we perform a range update, with probability $1/2$ we reduce $\int_{x\in(a,b)} dF(x)$ by at least a factor of $3/4$. Thus, by a Chernoff bound, with probability at least $1-\delta/3$, the number of rounds r of the main loop of the algorithm to get $\int_{x\in(a,b)} dF(x) \leq \alpha$ is $r > 16 \ln \frac{3}{\delta} + 4 \log_{3/4} \alpha$. We will set $\alpha = \frac{\epsilon}{\tau}$.

The algorithm proceeds until either the sample phase fails or the EstimateS phase returns a sufficiently good value of u. Let $\ell_2 > \frac{32}{\epsilon^2} \ln \frac{6(r+r_2)}{w\delta}$ be the number of samples used in each call to EstimateS to get failure probability bounded by $\frac{\delta}{3(r+r_2)}$. If $\int_{x\in(a,b)} dF(x) > \alpha$, then the probability that ℓ_1 independent samples are all outside (a, b) on any round is at most $r(1-\alpha)^{\ell_1}$, so $\ell_1 > \frac{1}{\alpha}\left(\ln \frac{3r}{\delta}\right)$. The low probability phase involves at most r_2 repetitions, for $r_2 = \log_2(b-a)/\epsilon$.

By a union bound, all $r + r_2$ estimates of $|S(u)|$ are correct to within $\frac{\epsilon}{2\tau}$ with probability at least $1 - \delta/3$.

The total number of samples needed is at most $r(\ell_1 + \ell_2) + \ell_2 r_2$. The space requirement is logarithmic in the counters ℓ_1, ℓ_2, and t (from EstimateS); plus the space required by Algorithm 1; plus space for a constant number of samples, each of unit size.

Remark: It is possible to drop the explicit checks as to whether $|\tilde{S}(u)| \leq |a|\frac{\epsilon}{2\tau}$ in the main loop and low probability phase. Instead, we can simply run the algorithm until one of the other stopping conditions holds.

6 Property Testing

In this section, rather than estimating statistics we consider the related problem of property testing, i.e., distinguishing if the distribution satisfies a certain property or is "far" from all distributions that satisfy the property. As is standard in the property testing literature [1–4, 20], we restrict ourselves to properties defined on discrete distributions over $[n]$, denoted $\mathcal{D}_{[n]}$, and the notion of "far" considered is the variation distance between the density functions of the distributions, i.e., $L_1(p, q) = \sum_{i\in[n]} |p(i) - q(i)|$ where $p(i)$ and $q(i)$ are the probability masses at i. We consider properties defined by a real-valued function π on the density function and for a given a and b we are interested in distinguishing the case that $\pi(p) < a$ from $\pi(p) > b$. We call these two cases the "no" and "yes" cases respectively. In this setting, a "moral analogue" of robustness is the continuity of the property.

Definition 6.1 (Weakly Continuous)**.** *A property π is (ϵ, δ)-weakly-continuous if for all distributions p^+, p^- satisfying $|p^+ - p^-| \leq \delta$ we have $|\pi(p^+) - \pi(p^-)| \leq \epsilon$.*

We say π is *symmetric* if $\pi(p(1), \ldots, p(n)) = \pi(p(\sigma(1)), \ldots, p(\sigma(n))$ for any permutation σ on $[n]$. Recently, Valiant [20] generalized much of the previous work on distribution testing. In particular, he proved the following theorem:

Theorem 6.2 (Canonical Testing Theorem [20])**.** *Let π be a symmetric (ϵ, δ)-weakly-continuous property and two thresholds $a < b$. Let $\theta = 600\delta^{-2} \log n$ and consider the following "canonical" algorithm:*

1. *Draw k samples and let $s(i)$ be the count of i among the samples.*
2. *Let $\mathcal{P} \subset \mathcal{D}_{[n]}$ be the set of distributions satisfying $p(i) = s(i)/k$ if $s(i) \geq \theta$ and $p(i) < \theta/k$ otherwise.*
3. *If all $p \in \mathcal{P}$ satisfy $\pi > b$ output "yes", otherwise output "no"*

If the canonical algorithm fails to distinguish between $\pi > b + \epsilon$ and $\pi < a - \epsilon$ in k samples then no tester can distinguish between $\pi > b - \epsilon$ and $\pi < a + \epsilon$ in $k\delta/(1000 \cdot 16^{\sqrt{\log n}})$ samples.

The contrapositive of the above result states that if $f(n, a, b, \epsilon)$ is the sample complexity to distinguish between $\pi > b - \epsilon$ and $\pi < a + \epsilon$, then the canonical algorithm can distinguish $\pi > b + \epsilon$ and $\pi < a - \epsilon$ using

$$k = O(f(n, a, b, \epsilon) 16^{\sqrt{\log n}}/\delta)$$

samples. The form of the canonical algorithm is well-suited to the data-stream model as the problem reduces to finding "heavy-hitters", i.e., elements in the stream whose frequency exceeds a certain threshold. This is a well-studied problem and numerous algorithms exist. We will consider the Misra-Gries [19] algorithm, a deterministic algorithm that takes k elements as input and can be used to return all elements whose frequency exceeds a threshold $\theta/2$ while

ensuring that no elements of frequency less than $\theta/4$ are returned. It does this using $O(k\theta^{-1}\log k)$ bits of space. Using this algorithm we have an easy way to perfectly emulate the canonical testing algorithm. We proceed in two phases each of which use k samples. The second phase will be the emulation of the canonical tester. To ensure that the "heavy" elements, i.e. those whose frequency exceed θ, can be identified and counted perfectly, in the first phase we use the Misra-Gries algorithm to identify all elements that are potentially heavy in the second phase.

Phase 1: Take k samples S_1 and let $q(i)$ be the empirical distribution defined by S_1. Using Misra-Gries, find a set $\mathcal{I}$ of $O(k/\theta)$ elements i that includes all i with $q(i) > \theta/(2k)$.

Phase 2: Take another k samples S_2 and let $p(i)$ be the empirical distribution defined by S_2. For each $i \in \mathcal{I}$, compute $p(i)$ exactly.

We note that it should be possible to approximately emulate the canonical tester with only k samples by combining the two phases (i.e., using estimates for $p(i)$ based on a more accurate heavy-hitters algorithm) but this only saves constant factors. We also clarify that we are not concerned with the time and space complexity of post-processing (as is common in the data streams literature). Hence, we do not make a general claim about the space requirements of performing the third step of the canonical algorithm. However, in many cases this is not difficult. For example, for entropy, it is clear how to find the maximum and minimum entropy values of distributions from $\mathcal{P}$.

The main point we want to highlight in the emulation of the canonical tester is that it implicitly gives rise to a trade-off between sample and space complexity. Note that the sample complexity depends inversely on δ while the space complexity depends linearly on δ. Furthermore, if π is (ϵ, δ)-weakly-continuous then trivially it is also (ϵ, δ')-weakly-continuous for any $\delta' \leq \delta$. Hence, we may choose δ to be small to ensure small space complexity while increasing the sample complexity. This gives us the following trade off between space and sample complexity.

Theorem 6.3 (Trade-off Theorem)**.** *Let $f(n, a, b, \epsilon)$ be the sample complexity of the distinguishing $\pi > b-\epsilon$ from $\pi < a+\epsilon$ where π is (ϵ, δ^*) weakly continuous. Then, for any $\delta < \delta^*$ there exists a stream algorithm that distinguishes $\pi > b + \epsilon$ from $\pi < a-\epsilon$ while using $O(f(n, a, b, \epsilon)16^{\sqrt{\log n}}/\delta)$ samples and $O(f(n, a, b, \epsilon)16^{\sqrt{\log n}}\delta/\log n)$ space.*

Proof. Space use is $O(k/\theta)$ since the set $\mathcal{I}$ can be found and stored in this amount of space using, e.g., the Misra-Gries algorithm [19]. The correctness follows immediately from Theorem 6.2 since, conditioned on the event that

$$\forall i \in [n], p(i) > \theta/k \implies i \in \mathcal{I} ,$$

the set $\mathcal{P}$ constructed is the same set that would have been constructed by running the canonical algorithm on the same k samples. This event happens with probability $1 - n^{-6}$.

A natural question is whether this trade-off is optimal. This does not appear to be the case. It is not hard to show (see e.g., [12, Theorem 3]) that with $O(\delta^{-2}n\log(n))$ samples from a distribution on n points, the variation difference between the empirical distribution and source distribution is at most δ with probability 9/10. Hence, if we take this many samples and run existing data-stream algorithms that approximate the empirical value of the function with additive error δ, then the final result incurs only additive error 2δ. Many empirical problems such as estimating entropy [16] and the distance to uniformity [17] can be solved in $O(\epsilon^{-2}\log mn)$ space. The following immediate theorem states the general result.

Theorem 6.4. *Let π be $(\epsilon/2, \delta)$-weakly-continuous and suppose there exists an $s(\epsilon)$ space algorithm that returns an additive $\epsilon/2$ approximation to π evaluated on a distribution defined empirically by the stream. Then there exists a stream algorithm that processes $O(\delta^{-2}n\log(n))$ samples in $s(\epsilon)$ space and returns an ϵ additive approximation for π.*

Acknowledgments

We thank Cynthia Dwork for early conversations that helped start this line of work. We are especially indebted to Jing Lei for his time and abundant patience in helping us understand the robust statistics literature.

References

[1] T. Batu, S. Dasgupta, R. Kumar, and R. Rubinfeld. The complexity of approximating the entropy. *SIAM J. Comput.*, 35(1):132–150, 2005.

[2] T. Batu, L. Fortnow, E. Fischer, R. Kumar, R. Rubinfeld, and P. White. Testing random variables for independence and identity. In *IEEE Symposium on Foundations of Computer Science*, pages 442–451, 2001.

[3] T. Batu, L. Fortnow, R. Rubinfeld, W. D. Smith, and P. White. Testing that distributions are close. In *IEEE Symposium on Foundations of Computer Science*, pages 259–269, 2000.

[4] T. Batu, R. Kumar, and R. Rubinfeld. Sublinear algorithms for testing monotone and unimodal distributions. In *ACM Symposium on Theory of Computing*, pages 381–390, 2004.

[5] A. Chakrabarti, G. Cormode, and A. McGregor. Robust lower bounds for communication and stream computation. In *ACM Symposium on Theory of Computing*, pages 641–650, 2008.

[6] A. Chakrabarti, T. S. Jayram, and M. Pătraşcu. Tight lower bounds for selection in randomly ordered streams. In *ACM-SIAM Symposium on Discrete Algorithms*, pages 720–729, 2008.

[7] K. L. Chang and R. Kannan. Pass-efficient algorithms for learning mixtures of uniform distributions. *SIAM Journal on Computing*, 39(3):783–812, 2009.

[8] O. Dekel, S. Shalev-Shwartz, and Y. Singer. The forgetron: A kernel-based perceptron on a budget. *SIAM J. Comput.*, 37(5):1342–1372, 2008.

[9] E. D. Demaine, A. López-Ortiz, and J. I. Munro. Frequency estimation of internet packet streams with limited space. In *European Symposium on Algorithms*, pages 348–360, 2002.

[10] C. Dwork and J. Lei. Differential privacy and robust statistics. In *ACM Symposium on Theory of Computing*, pages 371–380. ACM New York, NY, USA, 2009.

[11] R. A. Fisher. On the mathematical foundations of theoretical statistics. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, 222:309–368, 1922.

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

[13] S. Guha and A. McGregor. Stream Order and Order Statistics: Quantile Estimation in Random-Order Streams. *SIAM Journal on Computing*, 38:2044, 2009.

[14] S. Guha, A. McGregor, and S. Venkatasubramanian. Streaming and sublinear approximation of entropy and information distances. In *ACM-SIAM Symposium on Discrete Algorithms*, pages 733–742. ACM New York, NY, USA, 2006.

[15] F. R. Hampel, E. M. Ronchetti, P. J. Rousseeuw, and W. A. Stahel. *Robust Statistics: The Approach Based on Influence Functions (Wiley Series in Probability and Statistics)*. Wiley-Interscience, New York, revised edition, April 2005.

[16] N. J. A. Harvey, J. Nelson, and K. Onak. Sketching and streaming entropy via approximation theory. In *IEEE Symposium on Foundations of Computer Science*, pages 489–498, 2008.

[17] P. Indyk. Stable distributions, pseudorandom generators, embeddings, and data stream computation. *J. ACM*, 53(3):307–323, 2006.

[18] M. J. Kearns, Y. Mansour, D. Ron, R. Rubinfeld, R. E. Schapire, and L. Sellie. On the learnability of discrete distributions. In *ACM Symposium on Theory of Computing*, pages 273–282, 1994.

[19] J. Misra and D. Gries. Finding repeated elements. *Sci. Comput. Program.*, 2(2):143–152, 1982.

[20] P. Valiant. Testing symmetric properties of distributions. In *ACM Symposium on Theory of Computing*, pages 383–392. ACM New York, NY, USA, 2008.

[21] D. P. Woodruff. The average-case complexity of counting distinct elements. In *ICDT*, pages 284–295, 2009.