On the Complexity of Estimating the Effective Support Size*

Oded Goldreich†

June 16, 2019

Abstract

Loosely speaking, the effective support size of a distribution is the size of the support of a distribution that is close to it (in totally variation distance). We study the complexity of estimating the effective support size of an unknown distribution when given samples of the distributions as well as an evaluation oracle (which returns the probability that the queried element appears in the distribution). In this context, we present several algorithms that exhibit a trade-off between the quality of the approximation and the complexity of obtaining it, and leave open the question of their optimality.

Stating the actual results requires some definitions, which are provided in Section 1.1, based on Definition 1.1. Once Definitions 1.1–1.3 are internalized, one can find the main result stated in Theorem 1.9 (in Section 1.4).

Keywords: Distribution Testing, Effective Support Size, Evaluation oracle.

Contents

*Preliminary version; comments are most welcome.

†Faculty of Mathematics and Computer Science, Weizmann Institute of Science, Rehovot, Israel. Email: oded.goldreich@weizmann.ac.il.

1 Introduction

The support size of a (discrete) probability distribution is a natural parameter of a distribution: Defined as the number of elements that appear with positive probability (in the distribution), the support size measures the "scope" of the distribution; that is, the number of different elements that may occur when sampling from this distribution. Unfortunately, this parameter is highly sensitive to insignificant changes in the distribution; for example, any distribution is infinitesimally close to having an arbitrary large support size.

A much more robust notion, which maintains the intuitive appeal of the support size, is the "effective support size" of a distribution (cf., [2]). Loosely speaking, the "effective support size" of a distribution is the number of elements that remain in the support after discarding from it a set of elements that has a "small" total probability mass. Alternatively, the "effective support size" of a distribution $\mathcal{D}$ is the minimum support size of distributions that are "close" to $\mathcal{D}$. Hence, $\mathcal{D}$ has effective support size at most n if it is "close" to some distribution that has support size (at most) n. Needless to say, the actual definition should specify what is considered "close".

Definition 1.1 (effective support size): *We say that the distribution $\mathcal{D}$ has ϵ-*effective support size at most *n if $\mathcal{D}$ is ϵ-close to a distribution that has support size at most n, where $\mathcal{D}$ is ϵ-*close *to $\mathcal{D}'$ if their total variation distance is at most ϵ. The ϵ-*effective support size of *$\mathcal{D}$, denoted $\mathtt{ess}_\epsilon(\mathcal{D})$, is the minimal n such that $\mathcal{D}$ has ϵ-effective support size at most n.*

Note that the 0-effective support size of a distribution equals its support size, whereas its 1-effective support size equals 1. (Actually, for any distribution $\mathcal{D}$, there exists a number $\delta \in [0,1)$ such that the δ-effective support size of $\mathcal{D}$ equals 1.)

The notion of effective support size is much more robust than the notion of the support size; in particular, if $\mathcal{D}$ is infinitesimally close to a distribution that has ϵ-effective support size n, then $\mathcal{D}$ has ϵ-effective support size at most $n+1$ (where the additional unit is needed only in pathological cases).[1] In general, if $\mathcal{D}$ is $o(\epsilon)$-close to a distribution that has ϵ-effective support size at most n, then $\mathcal{D}$ that has $(1+o(1))\cdot\epsilon$-effective support size at most n.

1.1 Beyond the straightforward definition

The foregoing discussion hints at two aspects of slackness that may be applied to the effective support size. Actually, one better apply these slackness aspects (or notions of approximation) if wishing to *actually* find the effective support size of unknown distributions. First, rather than fixing the effectiveness parameter, one may want to allow it to vary within a fixed interval; that is, rather than seeking the ϵ-effective support size, for some predetermined $\epsilon > 0$, we seek a number that is upper-bounded by the ϵ-effective support size and lower-bounded by the ϵ'-effective support size (for some predetermined $\epsilon' > \epsilon$). Second, we may seek an approximation to the desired number rather than the number itself.

[1] Let $\mathcal{D}'$ be the foregoing distribution that has ϵ-effective support size n. Then, the typical case is that, for some $\epsilon' < \epsilon$, the distribution $\mathcal{D}'$ has ϵ'-effective support size n. In this case, any distribution that is $(\epsilon-\epsilon')$-close to $\mathcal{D}'$ has ϵ-effective support size n. The pathological case is that $\mathcal{D}'$ has ϵ-effective support size n, but for every $\epsilon' < \epsilon$ the ϵ'-effective support size of $\mathcal{D}'$ is larger than n. We claim that, in this case, for some $\epsilon' < \epsilon$, the distribution $\mathcal{D}'$ has ϵ'-effective support size at most $n+1$ (and it follows that any distribution that is $(\epsilon-\epsilon')$-close to $\mathcal{D}'$ has ϵ-effective support size at most $n+1$). To prove this claim, suppose that $\mathcal{D}'$ is ϵ-close to a distribution $\mathcal{D}''$ that has support size n. We prove the claim by considering two cases.

1. If the support of $\mathcal{D}'$ is contained in the support of $\mathcal{D}''$, then the claim is trivial (since then $\mathcal{D}'$ has support size n).
2. Otherwise, let e be in the support of $\mathcal{D}'$ but not in the support of $\mathcal{D}''$, and consider modifying $\mathcal{D}''$ by moving a probability mass of $\mathcal{D}'(e) > 0$ from $\{u : \mathcal{D}''(u) > \mathcal{D}'(u)\}$ to e. Then, the modified distribution $\mathcal{D}'''$ has support size $n+1$ and is $(\epsilon - \mathcal{D}'(e))$-close to $\mathcal{D}'$, and so the claim follows with $\epsilon' = \epsilon - \mathcal{D}'(e)$.

Definition 1.2 (relaxations of the effective support size): *The number n is an $[\epsilon_1, \epsilon_2]$-*effective support size of *$\mathcal{D}$ if there exists $\epsilon \in [\epsilon_1, \epsilon_2]$ such that n is the ϵ-effective support size of $\mathcal{D}$. A random variable X is an f-*factor approximation of the $[\epsilon_1, \epsilon_2]$-effective support size of *$\mathcal{D}$ if $\Pr[n \leq X \leq f \cdot n] \geq 2/3$ for n that is an $[\epsilon_1, \epsilon_2]$-effective support size of $\mathcal{D}$.*

Note that if n is an f-factor approximation of the $[\epsilon_1, \epsilon_2]$-effective support size of $\mathcal{D}$, then it lies in the interval $[\mathtt{ess}_{\epsilon_2}(\mathcal{D}), f \cdot \mathtt{ess}_{\epsilon_1}(\mathcal{D})]$. (This is because of any $\epsilon \in [\epsilon_1, \epsilon_2]$, it holds that $\mathtt{ess}_{\epsilon_2}(\mathcal{D}) \leq \mathtt{ess}_{\epsilon}(\mathcal{D}) \leq \mathtt{ess}_{\epsilon_1}(\mathcal{D})$.)

As hinted, we are interested in algorithms that, for some ϵ_1, ϵ_2 and f, when given oracle access to an arbitrary distribution $\mathcal{D}$, output an f-factor approximation of the $[\epsilon_1, \epsilon_2]$-effective support size of $\mathcal{D}$. Two questions arise:

1. *What does it mean to have oracle access to a distribution?* One natural oracle associated with a distribution $\mathcal{D}$ is actually a sampling device, denoted $\mathtt{samp}_{\mathcal{D}}$, that on each invocation returns a sample of $\mathcal{D}$ (i.e., an element drawn according to the distribution $\mathcal{D}$). Another natural oracle is an evaluation oracle, denoted $\mathtt{eval}_{\mathcal{D}}$, that answers each query e with $\mathcal{D}(e) = \Pr_{s \sim \mathcal{D}}[s = e]$, which equals $\Pr[\mathtt{samp}_{\mathcal{D}} = e]$.

   We shall focus on oracle machines that are given oracle access to both oracles, but will also discuss the case of machines that only get access to a sampling device. Actually, we shall consider the latter setting as a special case.

2. *What parameters $\epsilon_1 < \epsilon_2$ and f can we handle and at what cost?* Wishing to reduce the number of parameters, we fix an arbitrary small constant $\beta > 1$, and consider the setting $\epsilon_1 = \epsilon$ and $\epsilon_2 = \beta \cdot \epsilon$; that is, we keep ϵ as a single effectiveness parameter, which we shall always keep varying. In contrast, the approximation parameters will sometimes be a function of ϵ and sometimes also depends on the distribution $\mathcal{D}$ (e.g., it may depend on the ϵ-effective support size of $\mathcal{D}$).

With these preliminaries in place, our main definition is the following.

Definition 1.3 (approximating the effective support size): *We say that a* (two oracle) *machine M is an f-*factor approximator of the $[\epsilon_1, \epsilon_2]$-effective support size of distributions *(in a class $\mathcal{C}$) if, for every distribution $\mathcal{D}$ (in $\mathcal{C}$), it holds that $M^{\mathtt{samp}_{\mathcal{D}}, \mathtt{eval}_{\mathcal{D}}}(\epsilon_1, \epsilon_2)$ is an f-factor approximation of the $[\epsilon_1, \epsilon_2]$-effective support size of $\mathcal{D}$.*

Algorithms that have no access to an evaluation oracle may be viewed as a special case in which the oracle machine makes no queries to $\mathtt{eval}_{\mathcal{D}}$. Note that, in general, we did not restrict the complexity of the approximator so far. Indeed, in what follows we shall consider the query complexity of the approximator as a function of ϵ_1, ϵ_2 and f as well as on the distribution $\mathcal{D}$ itself (e.g., the complexity may depend on the output estimate of the effective support size of $\mathcal{D}$).

1.2 Initial observations

We start with two simple but clarifying observations:

Observation 1.4 (The effective support is obtained by omitting the lightest elements in the distribution): *If $\mathcal{D}$ has ϵ-effective support size n, then $\mathcal{D}$ is ϵ-close to a distribution that has support that consists of the n heaviest elements in $\mathcal{D}$, with ties broken arbitrarily.*

Proof: Assuming that $n = \mathtt{ess}_{\epsilon}(\mathcal{D})$, let H denote the set of n heaviest elements in $\mathcal{D}$, where ties are broken arbitrarily; that is, $|H| = n$ and for every $e \notin H$ it holds that $\mathcal{D}(e) \leq \min_{h \in H}\{\mathcal{D}(h)\}$. Then, $\mathcal{D}(H) \stackrel{\text{def}}{=} \sum_{h \in H} \mathcal{D}(h) \geq 1 - \epsilon$, because otherwise we derive a contradiction (to the hypothesis that $\mathcal{D}$

is ϵ-close to some distribution of support size n).[2] Moving the probability mass of $\overline{H}$ to H, the claim follows (e.g., fixing any $h \in H$, we may let $\mathcal{D}'(h) \stackrel{\text{def}}{=} \mathcal{D}(h) + \sum_{e \notin H} \mathcal{D}(e)$ and $\mathcal{D}'(e) \stackrel{\text{def}}{=} \mathcal{D}(e)$ for every $e \in H \setminus \{h\}$). ■

Observation 1.5 (Small approximation factors can be eliminated by moderately increasing the larger effectiveness threshold): *If a random variable X is an f-factor approximation of the $[\epsilon_1, \epsilon_2]$-effective support size of $\mathcal{D}$, then X/f is an $[\epsilon_1, \epsilon_2 + (f-1)/f]$-effective support size of $\mathcal{D}$.* (In particular, for $f = 1 + \epsilon > 1$, we have $(f-1)/f < \epsilon$.)

Proof: Suppose that n is an f-factor approximation of the $[\epsilon_1, \epsilon_2]$-effective support size of $\mathcal{D}$; that is, $\mathtt{ess}_{\epsilon_2}(\mathcal{D}) \leq n \leq f \cdot \mathtt{ess}_{\epsilon_1}(\mathcal{D})$. Observing that $\mathcal{D}$ is ϵ_2-close to a distribution $\mathcal{D}_2$ that has support size $n_2 \stackrel{\text{def}}{=} \mathtt{ess}_{\epsilon_2}(\mathcal{D})$, we move the probability mass of the $n_2 - (n_2/f)$ lightest elements of $\mathcal{D}_2$ to its n_2/f heaviest elements, obtaining a distribution of support size n_2/f, denoted $\mathcal{D}'_2$. Then, $\mathcal{D}$ is $(\epsilon_2 + \delta)$-close to D'_2, where $\delta = (n_2 - (n_2/f)) \cdot 1/n_2 = (f-1)/f$, and it follows that $\mathtt{ess}_{\epsilon_2+\delta}(\mathcal{D}) \leq n_2/f \leq n/f$. On the other hand, using $n \leq f \cdot \mathtt{ess}_{\epsilon_1}(\mathcal{D})$, we infer that $n/f \leq \mathtt{ess}_{\epsilon_1}(\mathcal{D})$. Hence, n/f is an $[\epsilon_1, \epsilon_2 + (f-1)/f]$-effective support size of $\mathcal{D}$. ■

1.3 Justifying the general framework of Section 1.1

Next, we show that only poor approximations can be obtained when not using the general framework outlined above (i.e., not using an effectiveness interval and an evaluation oracle).

Justifying the use of an effectiveness interval. As hinted upfront, we chose to relax the definition of ϵ-effective support size (i.e., Definition 1.1) by allowing two effectiveness thresholds (see Definition 1.2), because we found the former too restrictive. This view is substantiated in the following result.

Proposition 1.6 (on the hardness of approximating the ϵ-effective support size): *For any $\epsilon \in (0,1)$ and $n, f \in \mathbb{N}$, an algorithm that makes $o(n)$ queries to* (the sampling and evaluation oracles of) *an arbitrary distribution that has 2ϵ-effective support size at most n cannot provide an f-factor approximation of the ϵ-effective support size of the distribution.*

Note that the approximation factor (i.e., f) may depend arbitrarily on ϵ and n, but not on other parameters of the distribution (like its actual ϵ-effective support size). Indeed, n is merely an upper bound on the 2ϵ-effective support size of the distribution, whereas its actual ϵ-effective support size may be unrelated to its 2ϵ-effective support size. In fact, the proof capitalizes on two extreme cases: For $\epsilon \in (0, 0.5)$, in one case the ϵ-effective support size is quite close to the 2ϵ-effective support size, whereas in the other case the ϵ-effective support size is arbitrary larger than its 2ϵ-effective support size. (In both cases, the 2ϵ-effective support size is $(1-2\epsilon) \cdot n$.)

Proof: Fixing $\epsilon \in (0,1)$ and $n \in \mathbb{N}$, we pick a sufficiently large $N \gg n$, and consider the following two distributions:

[2] That is, supposed towards the contradiction that $\mathcal{D}(H) < 1 - \epsilon$. Then, for every distribution $\mathcal{D}'$ having a support S' such that $|S'| = n$, it holds that the total variation distance between $\mathcal{D}$ and $\mathcal{D}'$ equals

$$\begin{aligned} \max_S\{\mathcal{D}'(S) - \mathcal{D}(S)\} &\geq \mathcal{D}'(S') - \mathcal{D}(S') \\ &= 1 - \mathcal{D}(S') \\ &\geq 1 - \max_{S:|S|=n}\{\mathcal{D}(S)\} \\ &= 1 - \mathcal{D}(H), \end{aligned}$$

which (by the contradiction hypothesis) is greater than $1 - (1-\epsilon)$. Hence, the total variation distance between $\mathcal{D}$ and an arbitrary distribution of support size n is greater than ϵ, contradicting the hypothesis that $\mathtt{ess}_\epsilon(\mathcal{D}) = n$.

1. For arbitrary sets H and L such that $|H| = (1-\epsilon)\cdot n$ and $|L| = \epsilon \cdot N^2$, the distribution $\mathcal{D}_1$ assigns probability $1/n$ to each of element in H, and probability $1/N^2$ to each element in L.

2. For arbitrary sets H' and L' such that $|H'| = (1-\epsilon)\cdot n - 1$ and $|L'| = \epsilon \cdot N^2 + N$, the distribution $\mathcal{D}_2$ assigns probability $1/n$ to each of element in H', probability $1/N^2$ to each element in L', and probability $(1/n) - (1/N)$ to a single element $s \notin H' \cup L'$.

Note that an oracle machine that makes $o(n)$ queries cannot distinguish these two distributions (i.e., its distinguishing gap is $o(1)$).[3] On the other hand, $\mathcal{D}_1$ has ϵ-effective support size $(1-\epsilon)\cdot n < n$, whereas $\mathcal{D}_2$ has ϵ-effective support size $(1-\epsilon)\cdot n + N > N$. Hence, the approximation factor provided by a $o(n)$-query machine is $\Omega(N/n)$, which cannot be bounded in terms on ϵ and n. ■

Justifying the use of an evaluation oracle. In the rest of this paper, we shall focus on algorithms that use both a sampling device and an evaluation oracle, because algorithms that use only a sampling device perform quite poorly. This fact is an immediate corollary of a result of Raskhodnikova, Ron, Shpilka, and Smith [9].

Corollary 1.7 (on the hardness of approximating the effective support size when using a sampling device only): *For any constant $c \in (0, 0.06]$, an $0.04n^c$-factor approximator of the $[0, 0.04]$-effective support size of distributions over $[n]$ that makes no evaluation queries, must take $\Omega(n^{1-3c^{1/2}})$ samples.*

Proof: Restating the first part of [9, Cor. 2.2], we consider n-grained distributions over $[n]$, where a distribution is n-grained if all probabilities are multiples of $1/n$. The said result asserts that (for every $c \in (0, 1/16]$) $\Omega(n^{1-3c^{1/2}})$ *samples are needed in order to distinguish an n-grained distribution of support size at least $n/11 > 0.09n$ from an n-grained distribution with support size at most n^{1-c}*. Note that the first distribution has 0.05-effective support size at least $0.09n - 0.04n$, whereas the second distribution has 0-effective support size at most n^{1-c}. Lastly, note that $0.05n/n^{1-c}$ is greater than the desired approximation factor (i.e., $0.04n^c$). ■

We stress that Corollary 1.7 does not rule out the possibility of obtaining more crude approximations of the effective support size in time that is significantly smaller than linear in the effective support size. On the other hand, note that, in this setting (i.e., without using an evaluation oracle), *nothing significant can be done in time that is significantly smaller than a square root of the effective support size*, since (for any $c > 0$) in $n^{0.5-c}$ time one cannot distinguish a uniform distribution on n elements from a uniform distribution on $n^{1-2c-o(1)}$ elements. So the real questions are of the following type.

Open Problem 1.8 (obtaining crude approximation of the effective support size when using a sampling device only): *Fixing any positive $\epsilon_1 < \epsilon_2 < 0.5$, for which values of $c, c' \in (0, 0.5)$ does there exist an n^c-factor approximator of the $[\epsilon_1, \epsilon_2]$-effective support of distributions over $[n]$ that uses $n^{0.5+c'}$ samples, when making no evaluation queries at all?*

1.4 Our main results

In contrast to Corollary 1.7, fast and good approximations of the effective support size (of distributions) can be obtained when using both types of queries (i.e., a sampling device as well as an evaluation oracle). In fact, we obtain several different algorithms that exhibit a trade-off between the running time and the approximation factor of the $[\epsilon, \beta\cdot\epsilon]$-effective support size (of distributions), for any constant $\beta > 1$.

[3] To streamline the argument, when the machine queries $\mathcal{D}_1$, let s be an arbitrary element in H. Then, the distinguishing gap is mainly due to the case that the machine obtained s as a sample, where we neglect the different collision probabilities for L and L' (since it is extremely small).

Specifically, at the fastest extreme, we obtain an $O(1/\epsilon)$-*time algorithm with an approximation factor that is logarithmic in the* ϵ-*effective support size* (and almost linear in $1/\epsilon$). On the other hand, at the most accurate extreme, we obtain a β-*factor approximation algorithm than runs in* $\log^*(n/\epsilon)$-*time, where* n *is the effective support size.*

Theorem 1.9 (fast and good approximators of the effective support size): *For every constant* $\beta > 1$ *and each of the following for options regarding* T *and* f*, there exists an algorithm that, on input* $\epsilon > 0$ *and oracle access to* $\mathcal{D}$*, runs in time* T *and outputs an* f*-factor approximation of the* $[\epsilon, \beta \cdot \epsilon]$*-effective support size of* $\mathcal{D}$*. Letting* $n = \mathtt{ess}_\epsilon(\mathcal{D})$ *denote the* ϵ*-effective support size of* $\mathcal{D}$*, the four options are:*

1. $T = O(1/\epsilon)$ *and* $f = O(\epsilon^{-1} \log(n/\epsilon))$.
2. $T = \widetilde{O}(1/\epsilon)$ *and* $f = O(\log(n/\epsilon))$.
3. *For any constants* $t, k \in \mathbb{N}$*, it holds that* $T = \widetilde{O}(t/\epsilon^{1+\frac{1}{k}})$ *and* $f = \widetilde{O}(\log^{(t)}(n/\epsilon))$*, where* $\log^{(t)}$ *denotes* t *iterated logarithms.*
4. *For any constant* $k \in \mathbb{N}$*, it holds that* $T = \widetilde{O}(\log^*(n/\epsilon)/\epsilon^{1+\frac{1}{k}})$ *in expectation and* $f = \beta$.

The running time dependence on β *is* $\mathrm{poly}(1/(\beta - 1))$.

By Observation 1.5, in the last item, we can obtain $f = 1$ (with $T = \widetilde{O}(\epsilon^{-1} \log^*(n/\epsilon))$ in expectation). It is not clear whether the trade-off between the running time and the approximation factor exhibited by the foregoing four options is inherent. In particular, we wonder whether one can obtain T and f that are both functions of ϵ only.

Open Problem 1.10 (approximators of the effective support size with performance guarantees that are oblivious of the distribution): *For a constant* $\beta > 1$*, does there exist an algorithm that, on input* $\epsilon > 0$ *and oracle access to* $\mathcal{D}$*, runs in time* $T(\epsilon)$ *and outputs an* $f(\epsilon)$*-factor approximation of the* $[\epsilon, \beta \cdot \epsilon]$*-effective support size of* $\mathcal{D}$*, where* T *and* f *are functions of* ϵ *only? If so, can both functions be polynomials in* $1/\epsilon$*? And, if so, can we have* $T(\epsilon) = \mathrm{poly}(1/\epsilon)$ *and* $f = 1$*?*

A negative answer would join the small collection of natural computational problems having computational complexity that depends extremely mildly on the object's size (i.e., the complexity is lower-bounded by some unbounded function of the size and is upper-bounded by a log-star in that size).

1.5 Wider context

Our original motivation for the current study arose in the context of "vertex-distribution-free" (VDF) models for testing properties of graphs [5]. Loosely speaking, in these models the tester is provided with a sampling device to an arbitrary distribution, $\mathcal{D}$, over the vertex set (as well as with query access to the graph itself). Our focus in [5] was on strong testers; that is, tester whose complexity depends only on the proximity parameter. Nevertheless, in [5, Sec. 5.2], we suggested to consider also testers of complexity that depends on (label-invariant) parameters of the vertex distribution such as its effective support size. This immediately raises the problem of approximating these parameters. Indeed, an initial study of this problem was provided by us in [6, Sec. 2.2], and it was used in the construction of a Bipartite tester (in (a variant of) the bounded-degree VDF model), which is the actual focus of [6]. (We reproduce the relevant parts of [6, Sec. 2.2], which represent a somewhat different algorithmic approach, in the appendix.)

Access to an evaluation oracle may not be very natural in the context of the "vertex-distribution-free" testing model (yet, it was postulated, motivated, and relied upon in [6]). In contrast, an evaluation oracle is quite natural in the context of studying computational problems regarding distributions (see,

e.g., [1, 3, 8]).[4] In particular, prior works [1, 3, 8] considered a variety of computational problems such as approximating the distance to a known distribution, approximating the entropy of a distribution, and approximating the size of the support of distributions (when given a lower bound on the probability of the lightest element in the support, and allowed an additive approximation error that is inversely proportional to that bound).[5] We comment that the different models of [1, 3] and [8] coincide in our setting, where the domain of the distributions is not *a priori* known.[6]

Approximating the effective support size is somewhat related to (tolerantly) testing the support size of distributions, a task that has been studied extensively (see [4, Sec. 11.4] and the references therein). Specifically, tolerantly testing that $\mathcal{D}$ has support size n under proximity parameter ϵ and tolerance parameter ϵ' calls for accepting distributions that have ϵ'-effective support size at most n (i.e., when $\mathtt{ess}_{\epsilon'}(\mathcal{D}) \leq n$) and rejecting distributions that have ϵ-effective support size greater than n (i.e., when $\mathtt{ess}_{\epsilon}(\mathcal{D}) > n$). In particular, testing that $\mathcal{D}$ has support size n under proximity parameter ϵ calls for accepting distributions that have support size at most n and rejecting distributions that have ϵ-effective support size greater than n. (Note that an ϵ'-tolerant ϵ-tester is (given n and) allowed arbitrary behaviour in case $n \in [\mathtt{ess}_{\epsilon}(\mathcal{D}), \mathtt{ess}_{\epsilon'}(\mathcal{D}))$, whereas a 1-factor approximator of the $[\epsilon', \epsilon]$-effective support size is required to find $n \in [\mathtt{ess}_{\epsilon}(\mathcal{D}), \mathtt{ess}_{\epsilon'}(\mathcal{D})]$.)[7]

1.6 Conventions and notations

Throughout this work we refer to discrete probability distributions, which may be thought of as ranging either over binary strings or over natural numbers. For such a distribution $\mathcal{D}$, we denote by $\mathcal{D}(e)$ the probability (or weight or mass) that $\mathcal{D}$ assigns e; that is, $\mathcal{D}(e) = \Pr_{s\sim\mathcal{D}}[s = e]$. For a set S, we define $\mathcal{D}(S) \stackrel{\text{def}}{=} \sum_{e\in S} \mathcal{D}(e)$.

We say that $\mathcal{D}$ is ϵ-close to $\mathcal{D}'$ if the total variation distance between them is at most ϵ, where the total variation distance between $\mathcal{D}$ and $\mathcal{D}'$ equals

$$\frac{1}{2} \cdot \sum_{e} \left|\mathcal{D}(e) - \mathcal{D}'(e)\right| \;=\; \max_{S}\{\mathcal{D}(S) - \mathcal{D}'(S)\}. \tag{1}$$

Otherwise, we say that $\mathcal{D}$ is ϵ-far from $\mathcal{D}'$.

2 Algorithms

In this section we establish the four items of Theorem 1.9 by proving four corresponding theorems. Our starting point is an algorithm that is based on clustering the elements of the distribution's support according to their approximate probability mass (or weight). The key observation is that the number of relevant clusters (i.e., clusters having noticeable weight) is logarithmically related to the effective support size. Furthermore, the effective support size can be related to the size of a random relevant cluster (i.e., a relevant cluster selected with probability that is proportional to its total mass). The resulting approximation factor is linearly related to the number of relevant clusters (which is logarithmic in the effective support size) and is inversely related to the effectiveness threshold.

[4] Prior works (see, e.g., [7]) have also considered the problem of learning the evaluation function of a distribution (rather than learning to generate the distribution).

[5] The latter problem sounds related to approximating the effective support size, but is actually different from it.

[6] In general, in [1, 3] the algorithm is allowed arbitrary evaluation queries, whereas [8] provide it only with the probability mass of each sampled element. But in setting in which the domain of the distribution is arbitrary, evaluation queries to unsampled elements is practically useless.

[7] Hence, it is unclear how to convert an approximator into a tester. As for the opposite direction, we face the generic problem of converting a decision procedure into a search procedure, and note that we cannot afford a logarithmic factor overhead (since we care about lower complexities).

Theorem 2.1 (the basic algorithm): *For every constant $\beta > 1$, there exists an algorithm that on input $\epsilon > 0$ and oracle access to $\mathcal{D}$, runs in time $O(1/\epsilon)$ and outputs an $O(\epsilon^{-1}\log(n/\epsilon))$-factor approximation of the $[\epsilon, \beta\cdot\epsilon]$-effective support size of $\mathcal{D}$, where $n = \mathtt{ess}_\epsilon(\mathcal{D})$.*

Proof: Fixing $\beta > 1$ and $\mathcal{D}$, for every $i \in \mathbb{N}$, we consider the set of elements having probability approximately $\beta^{-(i-0.5)}$; that is, we let $W_i \stackrel{\text{def}}{=} \{e : \beta^{-i} < \mathcal{D}(e) \leq \beta^{-(i-1)}\}$. We first observe that almost all of the probability mass of $\mathcal{D}$ is assigned to the first $O(\epsilon^{-1}\cdot\log n)$ sets (i.e., W_i's), where $n = \mathtt{ess}_\epsilon(\mathcal{D})$ is the ϵ-effective support size of $\mathcal{D}$.

Claim 2.1.1 *Suppose that $\mathcal{D}$ has ϵ-effective support size at most n, and let $\ell \in \mathbb{N}$ be minimal such that $\sum_{i>\ell}\mathcal{D}(W_i) \leq \beta\cdot\epsilon$. Then, $\ell \leq \log_\beta(n/(\beta-1)\cdot\epsilon)$.*

Throughout this proof (as well as the subsequent proofs), we shall assume that $\ell > 1$, while noting that the case of $\ell = 1$ is easily handled (by outputting $|W_1|$).[8] In fact, for similar reasons, we may assume that $\ell > \log_\beta(1/\epsilon) + O(1)$.

Proof: Let S be a set of size at most n such that there exists a distribution that is ϵ-close to $\mathcal{D}$ and has support S. Then, letting $L \stackrel{\text{def}}{=} \{e : \mathcal{D}(e) \leq (\beta-1)\cdot\epsilon/n\}$, we have

$$\begin{aligned}\mathcal{D}(L) &= \mathcal{D}(L\cap S) + \mathcal{D}(L\setminus S)\\ &\leq |S|\cdot\max_{e\in L}\{\mathcal{D}(e)\} + \mathcal{D}(\overline{S})\\ &\leq n\cdot\frac{(\beta-1)\cdot\epsilon}{n} + \epsilon\end{aligned}$$

which equals $\beta\cdot\epsilon$. The claim follows, because, for every $i > k \stackrel{\text{def}}{=} \log_\beta(n/(\beta-1)\cdot\epsilon)$, it holds that $W_i \subseteq L$ (since $e \in W_i$ implies $\mathcal{D}(e) \leq \beta^{-(i-1)} \leq \beta^{-k} = (\beta-1)\cdot\epsilon/n$). ■

For $\epsilon' = \beta\cdot\epsilon$, let $\ell' \in \mathbb{N}$ be maximal such that $\sum_{i\geq\ell'}\mathcal{D}(W_i) \geq \beta\cdot\epsilon'$. Hence, $\Delta \stackrel{\text{def}}{=} \sum_{i\in[\ell',\ell]}\mathcal{D}(W_i) \geq (\beta-1)\cdot\epsilon'$. Suppose that we select $i \in [\ell',\ell]$ with probability proportional to $\mathcal{D}(W_i)$; this can be done by "rejection sampling" (and has complexity $O(1/\epsilon)$). The key observation is that, with probability at least 2/3, it holds that the selected i satisfies $\mathcal{D}(W_i) \geq \frac{\Delta}{3\ell} \geq \frac{(\beta-1)\epsilon'}{3\ell}$, because for $B \stackrel{\text{def}}{=} \{j \in [\ell',\ell] : \mathcal{D}(W_j) < \Delta/3\ell\}$ it holds that $\Pr_{i\sim\mathcal{D}}[i\in B | i\in[\ell',\ell]]$ equals $\sum_{j\in B}\mathcal{D}(W_j)/\Delta < |B|/3\ell \leq 1/3$. Hence, with probability at least 2/3, it holds that

$$|W_i| \geq \mathcal{D}(W_i)/\beta^{-(i-1)} \geq \frac{(\beta-1)\cdot\epsilon'}{3\ell}\cdot\beta^{i-1} = (\beta-1)\cdot\epsilon\cdot\beta^i/3\ell. \tag{2}$$

On the other hand, $\sum_{j\leq i}|W_j| < \sum_{j\leq i}\beta^j < \beta^{i+1}/(\beta-1)$. Now, letting $f = \frac{\beta^{i+1}/(\beta-1)}{(\beta-1)\cdot\epsilon\cdot\beta^i/3\ell} = \frac{3\cdot\beta}{(\beta-1)^2}\cdot\epsilon^{-1}\cdot\ell$ and combining the foregoing bounds, we get

$$\sum_{j\leq i}|W_j| < \frac{\beta^{i+1}}{\beta-1} \leq f\cdot\sum_{j\leq i}|W_j|. \tag{3}$$

Using $\sum_{j>i}\mathcal{D}(W_j) \leq \sum_{j\geq\ell'+1}\mathcal{D}(W_j) < \beta^2\cdot\epsilon$ and $\sum_{j\geq i}\mathcal{D}(W_j) \geq \sum_{j>\ell-1}\mathcal{D}(W_j) \geq \beta\cdot\epsilon$, we infer that $v \stackrel{\text{def}}{=} \beta^{i+1}/(\beta-1)$ constitutes an $f\cdot\beta^2$-factor approximation of the $[\epsilon,\beta^2\cdot\epsilon]$-effective support size of $\mathcal{D}$, since $\mathtt{ess}_{\beta^2\epsilon}(\mathcal{D}) \leq \sum_{j\leq i}|W_j| < v$ and $\mathtt{ess}_\epsilon(\mathcal{D}) \geq \beta^{-2}\cdot\sum_{j\leq i}|W_j| \geq v/\beta^2 f$, where $\mathtt{ess}_{\beta^2\epsilon}(\mathcal{D}) \leq \sum_{j\leq i}|W_j|$ and $\mathtt{ess}_\epsilon(\mathcal{D}) \geq \sum_{j\leq i}|W_j|$ for $i < \ell$ are quite obvious (see below). Let us spell out the latter two facts as well as their proofs.

[8] In this case (i.e., $\ell = 1$), it holds that $|W_1| < 1/\beta$ and all elements of W_1 can be found in constant time. Furthermore, this case can be detected in constant time (since $\mathcal{D}(W_1)$ can be computed in constant time).

Claim 2.1.2 *Let ℓ and ℓ' be as define above. Then, for every $i \in [\ell', \ell]$ it holds that:*

1. $\mathtt{ess}_{\beta^2\epsilon}(\mathcal{D}) \leq \sum_{j \leq i} |W_j|$.
2. $\mathtt{ess}_{\epsilon}(\mathcal{D}) \geq \beta^{-1} \cdot \sum_{j \leq i} |W_j| - 1$. *Furthermore, if $i < \ell$, then* $\mathtt{ess}_{\beta\epsilon}(\mathcal{D}) \geq \sum_{j \leq i} |W_j|$.

As done above, in the sequel we shall use a simpler (but more wasteful) form of Part 2, which asserts that $\mathtt{ess}_{\epsilon}(\mathcal{D}) \geq \beta^{-2} \cdot \sum_{j \leq i} |W_j|$. This is justified (for $i = \ell$) by recalling that we may assume that $\ell > \log_\beta(\beta^3/(\beta-1)\epsilon)$ (see discussion following Claim 2.1.1). On the other hand, we may assume that $\mathcal{D}(W_\ell) \geq (\beta-1) \cdot \epsilon$, since otherwise $\sum_{j>\ell} \mathcal{D}(W_j) \geq \epsilon$ and $\mathtt{ess}_\epsilon(\mathcal{D}) \geq \sum_{j \leq \ell} |W_j|$ follows. Hence, $|W_\ell| \geq \epsilon \cdot \beta^{\ell-1} > \beta^2/(\beta-1)$ and $\beta^{-1} \cdot \sum_{j \leq \ell} |W_j| - 1 > \beta^{-2} \cdot \sum_{j \leq \ell} |W_j|$ follows.

Proof: To see the first part, consider a distribution $\mathcal{D}'$ in which the probability mass of $\bigcup_{j>i} W_j$ is moved to $\bigcup_{j \leq i} W_j$. By maximality of ℓ', it holds that $\sum_{j>i} \mathcal{D}(W_j) \leq \sum_{j \geq \ell'+1} \mathcal{D}(W_j) < \beta^2\epsilon$. Hence, $\mathcal{D}'$ is $\beta^2\epsilon$-close to $\mathcal{D}$, which implies that there exists a distribution that is $\beta^2\epsilon$-close to $\mathcal{D}$ and has support of size $\sum_{j \leq i} |W_j|$ (i.e., $\mathtt{ess}_{\beta^2\epsilon}(\mathcal{D}) \leq \sum_{j \leq i} |W_j|$).

Turning to the second part, we start with the furthermore case (i.e., $i < \ell$). In this case, using the minimality of ℓ, it holds that $\sum_{j>i} \mathcal{D}(W_j) \geq \sum_{j>\ell-1} \mathcal{D}(W_j) > \beta\epsilon$. Using Observation 1.5, $\sum_{j>i} \mathcal{D}(W_j) \geq \beta\epsilon$ implies that any distribution that is $\beta\epsilon$-close to $\mathcal{D}$ must have support size at least $\sum_{j \leq i} |W_j|$ (i.e., $\mathtt{ess}_{\beta\epsilon}(\mathcal{D}) \geq \sum_{j \leq i} |W_j|$).

Turning to the main claim of the second part and focusing on the case of $i = \ell$ (since a stronger claim was already established for $i < \ell$), we let $\delta \stackrel{\text{def}}{=} \sum_{j \geq \ell} \mathcal{D}(W_j)$ and observe that $\delta > \beta\epsilon$ (by the minimality of ℓ). Below, we introduce a distribution $\mathcal{D}'$ that is $(1-\beta^{-1}) \cdot \delta$-close to $\mathcal{D}$ and satisfies $\sum_{j \leq \ell-1} |W'_j| \geq \beta^{-1} \cdot \sum_{j \leq \ell} |W_j| - 1$ and $\sum_{j>\ell} \mathcal{D}'(W'_j) = \sum_{j>\ell-1} \mathcal{D}(W_j)$, where $W'_j = \{e : \beta^{-j} < \mathcal{D}'(e) \leq \beta^{-(j-1)}\}$. Using this distribution, we observe that

$$\begin{aligned}
\mathtt{ess}_\epsilon(\mathcal{D}) &\geq \mathtt{ess}_{\delta/\beta}(\mathcal{D}) \\
&\geq \mathtt{ess}_\delta(\mathcal{D}') \\
&\geq \sum_{j \leq \ell-1} |W'_j| \\
&\geq \beta^{-1} \cdot \sum_{j \leq \ell} |W_j| - 1,
\end{aligned}$$

where the first inequality is due to $\epsilon \leq \delta/\beta$, the second inequality is due to fact that a distribution that is δ/β-close to $\mathcal{D}$ must be δ-close to $\mathcal{D}'$ (which implies that $\mathtt{ess}_\delta(\mathcal{D}') \leq \mathtt{ess}_{\delta/\beta}(\mathcal{D})$),[9] and the third inequality follows by using $\sum_{j>\ell} \mathcal{D}'(W'_j) = \sum_{j>\ell-1} \mathcal{D}(W_j) = \delta$ (which implies $\mathtt{ess}_\delta(\mathcal{D}') \geq \sum_{j \leq \ell-1} |W'_j|$).[10] So it is left to demonstrate the existence of the foregoing distribution $\mathcal{D}'$. This is done by shifting the probability mass of W_ℓ to $W'_{\ell-1}$; specifically, letting L_ℓ denote a maximal set of the lightest elements of W_ℓ such that $\mathcal{D}(L_\ell) \leq \beta^{-1} \cdot \mathcal{D}(W_\ell)$, we increase the probability mass of each $e \in L_\ell$ by a factor of β, and assign each element in $W_\ell \setminus L_\ell$ weight 0 (while leaving the rest of $\mathcal{D}$ intact).[11] Note that $|L_\ell| \geq \lfloor \beta^{-1} \cdot |W_\ell| \rfloor$. Hence, the resulting distribution $\mathcal{D}'$ satisfies all the foregoing requirements; specifically:

- $\mathcal{D}'$ is at distance $(1-\beta^{-1}) \cdot \mathcal{D}(W_\ell)$ from $\mathcal{D}$ (i.e., $\mathcal{D}'$ is $(1-\beta^{-1}) \cdot \delta$-close to $\mathcal{D}$).

[9] Let $\mathcal{D}''$ be a distribution that is δ/β-close to $\mathcal{D}$ and has support size $\mathtt{ess}_{\delta/\beta}(\mathcal{D})$. Then, $\mathcal{D}''$ is $((\delta/\beta)+(1-\beta^{-1})\cdot\delta)$-close to $\mathcal{D}'$, since $\mathcal{D}$ is $(1-\beta^{-1}) \cdot \delta$-close to $\mathcal{D}'$. It follows that $\mathtt{ess}_\delta(\mathcal{D}')$ is upper-bounded by the support size of $\mathcal{D}''$.

[10] Indeed, as in the argument used in the furthermore claim, this implication is due to Observation 1.5, which implies that $\mathtt{ess}_\delta(\mathcal{D}')$ is minimized by moving a probability mass of δ from the lighter elements to the heavier ones.

[11] Actually, the above description is slightly inaccurate, since $\mathcal{D}(W_\ell) - \beta \cdot \mathcal{D}(L_\ell) \in [0, \beta^{-(i-1)})$. In case the difference is positive, we move the access probability mass (from one of the elements of $W_\ell \setminus L_\ell$) to arbitrary elements in $\bigcup_{j \leq \ell-1} W'_j$.

- $|W'_{\ell-1}| \geq |W_{\ell-1}| + \lfloor \beta^{-1} \cdot |W_\ell| \rfloor$, which implies $\sum_{j \leq \ell-1} |W'_j| \geq \beta^{-1} \cdot \sum_{j \leq \ell} |W_j| - 1$.
- $\mathcal{D}'(W'_{\ell-1}) = \mathcal{D}(W_{\ell-1}) + \mathcal{D}(W_\ell)$ and $\mathcal{D}'(W'_j) = \mathcal{D}(W_j)$ for every $j \notin \{\ell-1, \ell\}$, which implies $\sum_{j>\ell} \mathcal{D}'(W'_j) = \sum_{j>\ell-1} \mathcal{D}(W_j)$.

This completes the proof of the claim. ■

The foregoing presentation is idealized, since in reality we do not know ℓ' and ℓ. Yet, we can find "good enough" approximations for them. Specifically, taking a sample S of $\mathrm{poly}(1/(\beta-1)) \cdot \epsilon^{-1}$ elements of $\mathcal{D}$, we set $\widetilde{\ell}$ to be minimal such that $|\{e \in S : e \in \bigcup_{j>\widetilde{\ell}} W_j\}| < \beta^{1.1} \cdot \epsilon \cdot |S|$, while noting that with high probability $\widetilde{\ell} \leq \ell$. Likewise, we set $\widetilde{\ell'}$ to be maximal such that $|\{e \in S : e \in \bigcup_{j \geq \widetilde{\ell'}} W_j\}| > \beta^{1.9} \cdot \epsilon \cdot |S|$, while noting that with high probability $\widetilde{\ell'} \geq \ell'$. On the other hand, $\sum_{j>\widetilde{\ell}} \mathcal{D}(W_j) \leq \beta^{5/4} \cdot \epsilon$ and $\sum_{j \geq \widetilde{\ell'}} \mathcal{D}(W_j) \geq \beta^{7/4} \cdot \epsilon$. Hence, $\widetilde{\Delta} \stackrel{\text{def}}{=} \sum_{i \in [\widetilde{\ell'}, \widetilde{\ell}]} \mathcal{D}(W_i) \geq \beta^{7/4}\epsilon - \beta^{5/4}\epsilon$, which is lower-bounded by $(\beta^{0.5} - 1) \cdot \beta\epsilon > (\beta - 1) \cdot \epsilon/2$. Hence, selecting $i \in [\widetilde{\ell'}, \widetilde{\ell}]$ with probability proportional to $\mathcal{D}(W_i)$, with probability at least $2/3$ it holds that $\mathcal{D}(W_i) \geq \frac{\widetilde{\Delta}}{3\widetilde{\ell}} \geq \frac{(\beta-1)\cdot\epsilon}{6\ell}$ (and $|W_i| \geq \frac{(\beta-1)\cdot\epsilon}{6\ell} \cdot \beta^{i-1}$ follows). Let us spell out the resulting algorithm.

Algorithm 2.1.3 (For fixed $\beta > 1$, on input $\epsilon > 0$ and oracle access to $\mathcal{D}$:)

1. *Using a sample of size $O(1/\epsilon)$, determine $\widetilde{\ell}$ and $\widetilde{\ell'}$ as outlined above.*
2. *Select $i \in [\widetilde{\ell'}, \widetilde{\ell}]$ with probability proportional to $\mathcal{D}(W_i)$.*

Output $\beta^{i+1}/(\beta-1)$.

Applying the simplified form of Claim 2.1.2 , it follows that $\mathtt{ess}_{\beta^2\epsilon}(\mathcal{D}) \leq \sum_{j \leq i} |W_j| < \beta^{i+1}/(\beta-1)$ and $\mathtt{ess}_\epsilon(\mathcal{D}) \geq \beta^{-2} \cdot \sum_{j \leq i} |W_j| > \frac{(\beta-1)\cdot\epsilon}{6\ell} \cdot \beta^{i-3}$. It follows that Algorithm 2.1.3 is a f'-factor approximator of the $[\epsilon, \beta^2 \cdot \epsilon]$-effective support size of $\mathcal{D}$, where $f' = \frac{\beta^{i+1}/(\beta-1)}{(\beta-1)\cdot\epsilon\cdot\beta^{i-3}/6\ell} < \frac{6\cdot\beta^4}{(\beta-1)^2} \cdot \epsilon^{-1} \cdot \ell$. Recalling that $\ell \leq \log_\beta(n/(\beta-1)\cdot\epsilon)$, where $n = \mathtt{ess}_\epsilon(\mathcal{D})$ is the ϵ-effective support size of $\mathcal{D}$, the claim follows (by a change of parameters). ■

Improving over Theorem 2.1. The approximation factor provided by Theorem 2.1 is essentially the multiple of two factors: The first factor is the reciprocal of the effectiveness parameter ϵ, and the second factor is essentially the logarithm of the effective support size; actually, the second factor is $O(\ell) = O(\log(n/\epsilon))$, where n is the ϵ-effective support size of the distribution. Both factors are an artifact of using $\Theta(\epsilon/\log(n/\epsilon)) \cdot \beta^{i-1}$ as a lower bound on the size of $|W_i|$, whereas $|W_i|$ could be as large as β^i.

An immediate improvement follows from the observation that we can afford to identify the case that $|W_i| = \Omega(\epsilon \cdot \beta^i)$, since in this case $\mathcal{D}(W_i) = \Omega(\epsilon)$, and output a much better estimate in this case. Specifically, when $\mathcal{D}(W_i) = \Omega(\epsilon)$, we can afford to approximate $\mathcal{D}(W_i)$ up to a β factor, and this yields an approximation of $|W_i|$ up to a β^2 factor. On the other hand, we can easily detect the case that $\mathcal{D}(W_i) = o(\epsilon)$ (or even distinguish $\mathcal{D}(W_i) < \epsilon/100$ from $\mathcal{D}(W_i) > \epsilon/99$), and in this case using $\Theta(\epsilon/\log(n/\epsilon)) \cdot \beta^{i-1}$ as an estimate of $|W_i|$ is only a factor of $O(\log(n/\epsilon))$ off. The foregoing considerations ignore the contribution of $\sum_{j<i} |W_j|$ to the effective support size, but employing the same considerations to W_j for each $j \in [i - \log_\beta(1/\epsilon), i-1]$, we reduce the approximation factor from $\Theta(\epsilon/\log(n/\epsilon))$ to $\Theta(\log(n/\epsilon))$, while slightly increasing the running time (so to allow for obtaining $\Theta(\log(1/\epsilon))$ approximate values rather than a constant number of such values).

Theorem 2.2 (the basic algorithm, revised): *For every constant $\beta > 1$, there exists an algorithm that on input $\epsilon > 0$ and oracle access to $\mathcal{D}$, runs in time $\widetilde{O}(1/\epsilon)$ and outputs an $O(\log(n/\epsilon))$-factor approximator of the $[\epsilon, \beta \cdot \epsilon]$-effective support size of $\mathcal{D}$, where $n = \mathtt{ess}_\epsilon(\mathcal{D})$.*

Proof: The algorithm starts by determining $\widetilde{\ell}$ and $\widetilde{\ell}'$ and selecting $i \in [\widetilde{\ell}', \widetilde{\ell}]$ as in Algorithm 2.1.3. Next, rather than outputting $\beta^{i+1}/(\beta-1)$, the algorithm uses $O(\epsilon^{-1}\log\log(1/\epsilon))$ samples in order to estimate $\mathcal{D}(W_j)$ for each $j \in [i', i]$, where $i' = \max(1, i - \log_\beta(1/\epsilon))$, and (essentially) outputs the corresponding estimate of $\sum_{j\leq i}\mathcal{D}(W_j)\cdot\beta^j$. (The upper bound of $O(\log(1/\epsilon))$ on the length of the interval $[i', i]$ is used when employing a union bound on the probability that some of these estimates are wrong.)

Algorithm 2.2.1 (refining Algorithm 2.1.3): *After setting $\widetilde{\ell}, \widetilde{\ell}'$ and i as in Algorithm 2.1.3, the algorithm proceeds as follows* (where $i' = \max(1, i - \log_\beta(1/\epsilon))$)*:*

- *For each $j \in [i', i]$, the algorithm first obtains an estimate $\widetilde{\delta}_j$ of $\mathcal{D}(W_j)$ such that* (with probability at least $1 - 1/10\log_\beta(1/\epsilon)$)) *it holds that $\widetilde{\delta}_j \in [\mathcal{D}(W_j), \beta\cdot\mathcal{D}(W_j)]$ if $\mathcal{D}(W_j) > \epsilon/\beta^2$ and $\widetilde{\delta}_j < \epsilon/\beta$ otherwise.*[12]
- *Next, for each $j \in [i', i]$, if $\widetilde{\delta}_j < \epsilon/\beta$, then the algorithm resets $\widetilde{\delta}_j \leftarrow \epsilon$.*
- *Finally, for each $j \in [i', i]$, the algorithm sets $\widetilde{w}_j \leftarrow \widetilde{\delta}_j \cdot \beta^j$, and outputs $\frac{\beta^{i'}}{\beta-1} + \sum_{j\in[i',i]}\widetilde{w}_j$ as its estimate of the effective support size.*

As in the proof of Theorem 2.1, the analysis of Algorithm 2.2.1 focus on the case that *the selected i satisfies* $\mathcal{D}(W_i) \geq \frac{(\beta-1)\cdot\epsilon}{6\ell}$. But here we consider two sub-cases.

1. If $\mathcal{D}(W_i) > \epsilon/\beta^2$, then, with high probability, it holds that $\mathcal{D}(W_i) \leq \widetilde{\delta}_i \leq \beta\cdot\mathcal{D}(W_i)$, and $|W_i| \leq \widetilde{w}_i \leq \beta^2\cdot|W_i|$ follows.
2. Otherwise (i.e., $\mathcal{D}(W_i) \leq \epsilon/\beta^2$), with high probability, the algorithm reset $\widetilde{\delta}_i \leftarrow \epsilon$. In this case, relying on the foregoing hypotheses, we have $\mathcal{D}(W_i) < \widetilde{\delta}_i = \epsilon \leq 6\ell\cdot(\beta-1)^{-1}\cdot\mathcal{D}(W_i)$, and $|W_i| \leq \widetilde{w}_i \leq \frac{6\ell}{\beta-1}\cdot\beta\cdot|W_i|$ follows.

Hence, in both cases

$$|W_i| \leq \widetilde{w}_i \leq \frac{6\ell\cdot\beta^2}{\beta-1}\cdot|W_i| \tag{4}$$

holds (where we either use $\ell = \omega(1)$ or assume $\beta < 7$). We stress that here the estimate for $|W_i|$ is sandwiched more tightly than in the proof of Theorem 2.1; that, is the ratio between the upper and lower bounds is $\frac{6\beta^2}{\beta-1}\cdot\ell$ (rather than is $\frac{6\beta^2}{\beta-1}\cdot\ell/\epsilon$).

A similar (but slightly different) analysis applies to each $j \in [i', i-1]$. Specifically, with high probability, it holds (for each $j \in [i', i-1]$) that if $\mathcal{D}(W_j) > \epsilon/\beta^2$ then $|W_j| \leq \widetilde{w}_j \leq \beta^2\cdot|W_j|$, whereas if $\mathcal{D}(W_j) \leq \epsilon/\beta^2$ then $\widetilde{\delta}_j = \epsilon$ and $|W_j| < \mathcal{D}(W_j)\cdot\beta^j \leq \epsilon\beta^{j-2} < \widetilde{\delta}_j\cdot\beta^j = \widetilde{w}_j$ follows. Hence, $|W_j| < \widetilde{w}_j \leq \max(\beta^2\cdot|W_j|, \epsilon\cdot\beta^j)$.

Using the forgoing bounds we sandwich the output value (i.e., $(\beta-1)^{-1}\cdot\beta^{i'} + \sum_{j\in[i',i]}\widetilde{w}_j$) between $\sum_{j\leq i}|W_j|$ and $O(\ell)\cdot\sum_{j\leq i}|W_j|$. First, we observe that, with high probability, the output is lower-bounded by $\sum_{j\leq i}|W_j|$, since

$$\begin{aligned}\sum_{j\leq i}|W_j| &= \sum_{j<i'}|W_j| + \sum_{j\in[i',i]}|W_j| \\ &< \frac{\beta^{i'}}{\beta-1} + \sum_{j\in[i',i]}\widetilde{w}_j,\end{aligned}$$

[12] This estimate, $\widetilde{\delta}_j$, is merely $\sqrt{\beta}$ times the fraction of the number of occurrences of elements in W_j in the foregoing sample. Note that if $\mathcal{D}(W_j) > \epsilon/\beta^2$, then (w.h.p.) the empirical measure resides in $[\beta^{-0.5}\cdot\mathcal{D}(W_j), \beta^{0.5}\cdot\mathcal{D}(W_j)]$, and otherwise the empirical count is smaller than $\epsilon/\beta^{1.5}$.

where we use $\sum_{j<i'}\beta^j < \beta^{i'-1}\cdot\beta/(\beta-1)$ as well as the fact that $|W_j| \le \beta^j$ for every j. Next, using $\mathcal{D}(W_i) \ge \frac{(\beta-1)\cdot\epsilon}{6\ell}$, which implies $|W_i| \ge \frac{(\beta-1)\cdot\epsilon}{6\ell}\cdot\beta^{i-1}$, we upper-bound the output value by $O(\ell)\cdot\sum_{j\le i}|W_j|$. In fact, foreseeing subsequent applications, we prove a more general statement (which refers to auxiliary parameters η and ϵ').

Claim 2.2.2 *Suppose that $|W_i| \ge \max(\eta\cdot\beta^{i-1}, \widetilde{w}_i/\beta^2) \ge 1$ and $\widetilde{w}_j \le \max(\beta^2\cdot|W_j|, \epsilon'\cdot\beta^j)$ for every $j\in[i',i-1]$. Then, for $i' = \max(1, i-\log_\beta(1/\epsilon'))$, it holds that*

$$\frac{\beta^{i'}}{\beta-1} + \sum_{j\in[i',i]}\widetilde{w}_j < \frac{\beta}{\beta-1} + \left(1+\frac{2\epsilon'}{(\beta-1)\cdot\eta}\right)\cdot\beta^2\cdot\sum_{j\le i}|W_j|.$$

In our application $\eta = \frac{(\beta-1)\cdot\epsilon}{6\ell}$ and $\epsilon' = \epsilon$; so $i' = \max(1, i-\log_\beta(1/\epsilon))$ and $\left(1+\frac{2\epsilon'}{(\beta-1)\cdot\eta}\right)\cdot\beta^2 = O(\ell)$.

Proof: For each $j\in[i',i-1]$, combining $\widetilde{w}_j \le \max(\beta^2\cdot|W_j|, \epsilon'\cdot\beta^j)$ and $|W_i| \ge \eta\cdot\beta^{i-1}$, we get

$$\widetilde{w}_j \le \beta^2\cdot|W_j| + \frac{\epsilon'\cdot\beta^{j-i+1}}{\eta}\cdot|W_i|. \tag{5}$$

We also use $\beta^{i'} \le \max(\beta, \epsilon'\cdot\beta^i) \le \beta + \frac{\beta\cdot\epsilon'}{\eta}\cdot|W_i|$, where the first inequality is due to the definition of i' and the second inequality is due to $|W_i| \ge \eta\cdot\beta^{i-1}$. Hence,

$$\begin{aligned}
\frac{\beta^{i'}}{\beta-1} + \sum_{j\in[i',i]}\widetilde{w}_j &= \frac{\beta^{i'}}{\beta-1} + \widetilde{w}_i + \sum_{j\in[i',i-1]}\widetilde{w}_j \\
&\le \frac{\beta}{\beta-1} + \frac{\beta\cdot\epsilon'}{(\beta-1)\cdot\eta}\cdot|W_i| + \beta^2\cdot|W_i| + \sum_{j\in[i',i-1]}\left(\beta^2\cdot|W_j| + \frac{\epsilon'\cdot\beta^{j-i+1}}{\eta}\cdot|W_i|\right) \\
&= \frac{\beta}{\beta-1} + \left(\frac{\beta\cdot\epsilon'}{(\beta-1)\cdot\eta} + \frac{\beta\cdot\epsilon'}{\eta}\cdot\sum_{j\in[i',i-1]}\beta^{j-i}\right)\cdot|W_i| + \beta^2\cdot\sum_{j\in[i',i]}|W_j| \\
&< \frac{\beta}{\beta-1} + \frac{2\beta\cdot\epsilon'}{(\beta-1)\cdot\eta}\cdot|W_i| + \beta^2\cdot\sum_{j\in[i',i]}|W_j| \\
&< \frac{\beta}{\beta-1} + \left(1+\frac{2\epsilon'}{(\beta-1)\cdot\eta}\right)\cdot\beta^2\cdot\sum_{j\in[i',i]}|W_j|.
\end{aligned}$$

The claim follow. ■

Recall that we have sandwiched the output of Algorithm 2.2.1 (i.e., $(\beta-1)^{-1}\cdot\beta^{i'} + \sum_{j\in[i',i]}\widetilde{w}_j$) between $\sum_{j\le i}|W_j|$ and $O(\ell)\cdot\sum_{j\le i}|W_j|$. Proceeding as in the proof of Theorem 2.1, we infer that Algorithm 2.2.1 constitutes a $O(\ell)$-factor approximator of the $[\epsilon, \beta\cdot\epsilon]$-effective support size of $\mathcal{D}$. ■

Reducing the factor that depends on the effective support size. The (approximation) factor of $O(\ell) = O(\log(n/\epsilon))$, which remains in Theorem 2.2, is due to the fact that we use $\mathcal{D}(W_i) = \Omega(\epsilon/\ell)$ for the selected $i\in[\ell',\widetilde{\ell}]$, and did not try to farther capitalize on the fact that $\sum_{\in[\widetilde{\ell}',\widetilde{\ell}]}\mathcal{D}(W_j) = \Omega(\epsilon)$.

The key observation is that *if Algorithm 2.1.3 happens to select $i > \widetilde{\ell}' + \log_\beta(\widetilde{\ell}/\epsilon)$ such that $\mathcal{D}(W_i) = \Omega(\epsilon/\widetilde{\ell})$, then we can output $\Theta(\epsilon/\widetilde{\ell})\cdot\beta^i$ and use a better analysis of approximation factor.* Loosely speaking, in this case, we can show the existence of a distribution $\mathcal{D}'$ that $\beta^2\cdot\epsilon$-close to $\mathcal{D}$ such that $\sum_{j\in[i-\log_\beta(\widetilde{\ell}/\epsilon),i]}\mathcal{D}'(W'_j) = \mathcal{D}'(W'_i) = \Theta(\epsilon/\widetilde{\ell})$ and $\sum_{j>i}\mathcal{D}'(W'_j) = 0$, where $W'_j = \{e : \beta^{-j} < \mathcal{D}'(e) \le \beta^{-(j-1)}\}$, while noting that $\sum_{j<i-\log_\beta(\widetilde{\ell}/\epsilon)}|W'_j| = O(\epsilon/\widetilde{\ell})\cdot\beta^i$. Hence, in this case, there exists a constant

$c > 0$ such that $\mathtt{ess}_{\beta^2\cdot\epsilon}(\mathcal{D}) \leq c \cdot (\epsilon/\widetilde{\ell}) \cdot \beta^i \leq O(1) \cdot \mathtt{ess}_\epsilon(\mathcal{D})$. On the other hand, we show that if the foregoing event (i.e., selecting $i > \widetilde{\ell}' + \log_\beta(\widetilde{\ell}/\epsilon)$ s.t. $\mathcal{D}(W_i) = \Omega(\epsilon/\widetilde{\ell})$) is unlikely, then this is due to $\sum_{j\in[\widetilde{\ell}',\widetilde{\ell}'+\log_\beta(\widetilde{\ell}/\epsilon)]} \mathcal{D}(W_j) = \Omega(\epsilon)$, and in this case we can use $\widetilde{\ell}' + \log_\beta(\widetilde{\ell}/\epsilon)$ instead of $\widetilde{\ell}$, which means that we lose a factor of $O(\log(\widetilde{\ell}/\epsilon))$ rather than a factor of $\widetilde{\ell}$. Iterating this reasoning for $t-1$ times, we get –

Theorem 2.3 (the iterative algorithm): *For every constants $\beta > 1$ and $t, k \in \mathbb{N}$, there exists an algorithm that on input $\epsilon > 0$ and oracle access to $\mathcal{D}$, runs in time $\widetilde{O}(t/\epsilon^{1+\frac{1}{k}})$ and outputs an $O(\log^{(t)}(n/\epsilon))$-factor approximator of the $[\epsilon, \beta\cdot\epsilon]$-effective support size of $\mathcal{D}$, where $n = \mathtt{ess}_\epsilon(\mathcal{D})$ and $\log^{(t)}$ denotes t iterated logarithms* (i.e., $\log^{(1)} m = \log m$ and $\log^{(j+1)} m = \log(\log^{(j)} m)$).

Proof: The case of $t = 1$ was established in Theorem 2.2, which actually states a stronger running-time bound. Hence, we start by considering the case of $t = 2$ (and $k = 1$), where all notations are as in the proofs of Theorems 2.1 and 2.2. We consider three cases, where the main cases are the last two.

1. If $\lambda \stackrel{\text{def}}{=} \widetilde{\ell} - \widetilde{\ell}' + 1 \leq 1/\epsilon$, then we can proceed as in the proof of Theorem 2.2, except that we use a sample of size $\widetilde{O}(\lambda/\epsilon) = \widetilde{O}(1/\epsilon^2)$ in order to obtain more accurate estimates of the $\mathcal{D}(W_j)$'s. Specifically, with such a sample, setting $i' = \max(1, i - \log_\beta(\lambda/\epsilon))$, we can obtain, for each $j \in [i', i]$, an estimate $\widetilde{\delta}_j$ of $\mathcal{D}(W_j)$ such that (with probability at least $1 - 1/10\log_\beta(\lambda/\epsilon)$) it holds that $\widetilde{\delta}_j \in [\mathcal{D}(W_j), \beta\cdot\mathcal{D}(W_j)]$ if $\mathcal{D}(W_j) > \beta^{-3}\cdot\epsilon/\lambda$ and $\widetilde{\delta}_j < \epsilon' \stackrel{\text{def}}{=} \beta^{-2}\cdot\epsilon/\lambda$ otherwise. We then proceed as in Algorithm 2.2.1, while resetting $\widetilde{\delta}_j < \epsilon'$ to $\widetilde{\delta}_j \leftarrow \epsilon'$, and output $\frac{\beta^{i'}}{\beta-1} + \sum_{j\in[i',i]} \widetilde{w}_j$.

   The crucial fact is that with such better estimates, for each $j \in [i', i]$, it holds that $|W_j| < \widetilde{w}_j \leq \max(\beta^2 \cdot |W_j|, O(\beta^{j-i}) \cdot |W_i|)$ (rather than $|W_j| < \widetilde{w}_j \leq \max(\beta^2 \cdot |W_j|, O(\beta^{j-i} \cdot \ell) \cdot |W_i|)$ as in the proof of Theorem 2.2).[13] Hence, we obtain a $O(1)$-factor approximation of the $[\epsilon, \beta^2 \cdot \epsilon]$-effective support size of $\mathcal{D}$.

   The main two cases deal with the situation in which $\lambda > 1/\epsilon$, where we want to avoid running in time $\widetilde{O}(\lambda/\epsilon)$, which we cannot afford when λ is much larger than $1/\epsilon$. (Recall that $\lambda = \widetilde{\ell} - \widetilde{\ell}' + 1$.)

2. If $\lambda > 1/\epsilon$ and $\sum_{j\in[\widetilde{\ell}',\widetilde{\ell}'+2\log_\beta\lambda]} \mathcal{D}(W_j) < 0.9\widetilde{\Delta}$, then, by repeatedly selecting i with probability proportional to $\mathcal{D}(W_i)$, we obtain $i \in [\widetilde{\ell}' + 2\log_\beta\lambda + 1, \widetilde{\ell}]$ after $O(1/\epsilon)$ trials. (Here we use $\sum_{j\in[\widetilde{\ell}'+2\log_\beta\lambda+1,\widetilde{\ell}]} \mathcal{D}(W_j) > 0.1\widetilde{\Delta}$, and in the analysis (which follows) we shall also use $\lambda > 1/\epsilon$.) Furthermore, with probability at least 0.9, it holds that $\mathcal{D}(W_i) > \widetilde{\Delta}/100\lambda > (\beta-1)\cdot\epsilon/200\lambda$. In this case, we output $\frac{\epsilon\cdot\beta^i}{(\beta-1)\cdot\lambda}$ as the estimated size of the effective support size, and show that this yields an $O(1)$-factor approximation of the $[\epsilon, \beta^2 \cdot \epsilon]$-effective support size of $\mathcal{D}$.

   The crux of the analysis is showing that the output (i.e., $\epsilon \cdot \beta^i/(\beta-1)\cdot\lambda$) is sandwiched between $\mathtt{ess}_{\beta^2\cdot\epsilon}(\mathcal{D})$ and $O(\mathtt{ess}_\epsilon(\mathcal{D}))$. On the one hand, $\mathtt{ess}_{\beta^2\cdot\epsilon}(\mathcal{D}) \leq (\beta-1)^{-1} \cdot \epsilon \cdot \beta^i/\lambda$, because $i - 2\log_\beta\lambda > \ell'$ and so $\sum_{j>i-2\log_\beta\lambda} \mathcal{D}(W_j) < \beta^2 \cdot \epsilon$, whereas $\sum_{j\leq i-2\log_\beta\lambda} |W_j| < \sum_{j\leq i-2\log_\beta\lambda} \beta^j < (\epsilon/\lambda)\cdot\beta^i/(\beta-1)$ (using $2\log_\beta\lambda \geq \log_\beta(\lambda/\epsilon)$). On the other hand, $\mathtt{ess}_\epsilon(\mathcal{D}) \geq \beta^{-2}\cdot|W_i|$, which is lower-bounded by $\mathcal{D}(W_i)\cdot\beta^{i-3} > \frac{(\beta-1)\cdot\epsilon}{200\lambda}\cdot\beta^{i-3}$. Hence, $\frac{\epsilon\cdot\beta^i/((\beta-1)\cdot\lambda)}{\mathtt{ess}_\epsilon(\mathcal{D})} < 200\cdot(\beta-1)^{-2}\cdot\beta^3 = O(1)$.

3. If $\lambda > 1/\epsilon$ and $\sum_{j\in[\widetilde{\ell}',\widetilde{\ell}'+2\log_\beta\lambda]} \mathcal{D}(W_j) \geq 0.9\widetilde{\Delta}$, then we can proceed as in the proof of Theorem 2.2 except that we use $\widetilde{\ell}' + 2\log_\beta\lambda = \widetilde{\ell}' + O(\log\ell)$ instead of $\widetilde{\ell}$, and $0.9\widetilde{\Delta}$ instead of $\widetilde{\Delta}$. In this case,

[13] That is, the analysis uses Claim 2.2.2 with ϵ' as set here (i.e., $\epsilon' = \beta^{-2}\cdot\epsilon/\lambda$) and $\eta = (\beta-1)\cdot\epsilon/6\lambda$, while noting that the original argument implies that $\mathcal{D}(W_i) \geq (\beta-1)\cdot\Delta/6\cdot(\widetilde{\ell} - \widetilde{\ell}' + 1)$ (rather than $\mathcal{D}(W_i) \geq (\beta-1)\cdot\Delta/6\cdot\ell$). (In Algorithm 2.3.1 we shall use a slightly different setting.)

we obtain a $O(\log \ell)$-factor approximation of the $[\epsilon, \beta \cdot \epsilon]$-effective support size of $\mathcal{D}$. (Note that $O(\log \ell) = O(\log\log(\mathtt{ess}_\epsilon(\mathcal{D})/\epsilon))$.)

Hence, in each case we make $\widetilde{O}(1/\epsilon^2)$ steps and obtain a $O(\log\log(\mathtt{ess}_\epsilon(\mathcal{D})/\epsilon))$-factor approximation of the $[\epsilon, \beta \cdot \epsilon]$-effective support size of $\mathcal{D}$. This establishes the claim for $t = 2$ and $k = 1$. We shall extend this result to general $k \in \mathbb{N}$ at the end of this proof.

For $t > 2$, we proceed almost exactly in the same manner, with the following three exceptions: First, the threshold for the main case analysis is set to equal $1 - 0.1/t$ rather than 0.9 (so to increase the probability mass in the last case).[14] Second, in the third case we continue as in the current proof with $t \leftarrow t - 1$ and $\widetilde{\Delta} \leftarrow (1 - (0.1/t)) \cdot \widetilde{\Delta}$ (rather than as in the proof of Theorem 2.2).[15] Last, we slightly modify the threshold distinguishing Case 1 from Cases 2–3 and the setting of i'. (The latter modification as well as the tightening of the analysis are performed in preparation for the proof of the next theorem.) For sake of clarity, we detail the recursive procedure next.

Algorithm 2.3.1 (recursive procedure with fixed parameters t and $\widetilde{\ell'}$): *The varying parameters are the remaining recursion-depth t'* (initially set to t)*, the remaining probability mass Δ'* (initially set to $\widetilde{\Delta}$)*, and the remaining interval length λ* (initially set to $\widetilde{\ell} - \widetilde{\ell'} + 1$).[16] *If $t' = 1$, then we proceed as in the proof of Theorem 2.2, and otherwise we proceed as follows, when setting $c = 300t/(\beta - 1)^2$.*

1. *If $\lambda < c/\epsilon$, then we proceed as in the proof of Theorem 2.2, except that we use a sample of size $\widetilde{O}(\lambda/\epsilon) = \widetilde{O}(t/\epsilon^2)$, set $i' = \max(1, i - \log_\beta(6\lambda/(\beta-1)^3\epsilon))$, and output $\frac{\beta^{i'}}{\beta-1} + \sum_{j \in [i', i]} \widetilde{w}_j$, where $i \in [\widetilde{\ell'}, \widetilde{\ell'} + \lambda - 1]$ and the $\widetilde{w}_j$'s are determined as in Algorithm 2.2.1, except that $\widetilde{\delta}_j$ is reset to $\epsilon' \stackrel{\text{def}}{=} (\beta-1)^3 \cdot \epsilon/6\lambda$ if $\widetilde{\delta}_j < \epsilon'$.*

   (Recall that our estimates of the $\mathcal{D}(W_j)$'s are better than in the proof of Theorem 2.2, since we use a larger sample. Specifically, for each $j \in [i', i]$, with high probability, $\widetilde{\delta}_j \in [\mathcal{D}(W_j), \beta \cdot \mathcal{D}(W_j)]$ if $\mathcal{D}(W_j) > \epsilon'/\beta^2$ and $\widetilde{\delta}_j < \epsilon'/\beta$ otherwise.)

   We warn that approximately distinguishing between the following two cases requires approximating the value of $\sum_{j \in [\widetilde{\ell'} + 2\log_\beta \lambda + 1, \widetilde{\ell'} + \lambda - 1]} \mathcal{D}(W_j)$ in the sense of distinguishing a value above $0.11\Delta'/t$ from a value below $0.09\Delta'/t$. This can be done using $O(t/\Delta') = O(t/\epsilon)$ samples. Using the same sample in all $t - 1$ recursion levels, it suffices to use a single sample of size $\widetilde{O}(t)/\epsilon$ for these approximations.

2. *If $\lambda \geq c/\epsilon$ and $\sum_{j \in [\widetilde{\ell'}, \widetilde{\ell'} + 2\log_\beta \lambda]} \mathcal{D}(W_j) < \left(1 - \frac{0.1}{t}\right) \cdot \Delta'$, then, by repeatedly selecting i with probability proportional to $\mathcal{D}(W_i)$, we obtain $i \in [\widetilde{\ell'} + 2\log_\beta \lambda + 1, \widetilde{\ell'} + \lambda - 1]$ after $O(t/\epsilon)$ trials. In this case, we output $(\beta - 1) \cdot \epsilon \cdot \beta^i/300t\lambda$ as the estimated size of the effective support size.*

3. *If $\lambda \geq c/\epsilon$ and $\sum_{j \in [\widetilde{\ell'}, \widetilde{\ell'} + 2\log_\beta \lambda]} \mathcal{D}(W_j) \geq \left(1 - \frac{0.1}{t}\right) \cdot \Delta'$, then we invoke this very procedure while setting the remaining recursion-depth to $t' - 1$, the remaining probability mass to $(1 - (0.1/t)) \cdot \Delta'$, and the remaining interval length to $3\log_\beta \lambda$.*

   (Note that $2\log_\beta \lambda + 1 < 3\log_\beta \lambda < \lambda$.)[17]

Hence, Cases 1 and 2 produce output by themselves, whereas Case 3 initiates a recursive call.

[14] This setting guarantees that the residual probability mass is reduced by a factor of $1 - 0.1/t$ rather than by a constant factor (of 0.9). The point is that $(1 - 0.1/t)^t > 0.9$, whereas $0.9^t = \exp(-t)$.

[15] Actually, when reaching the third case with $t = 2$, the recursive call will actually invoke Algorithm 2.2.1.

[16] Hence, $\sum_{j \in [\widetilde{\ell'}, \widetilde{\ell'} + \lambda - 1]} \mathcal{D}(W_j) \geq \Delta'$ holds.

[17] Both inequalities use $\lambda \geq c/\epsilon > 300/(\beta - 1)$, while assuming (w.l.o.g.) that $\beta \leq 2$.

The total complexity of the invocation of Algorithm 2.3.1 (with $t' = t$) is $\widetilde{O}(t/\epsilon^2)$, which fits our aim for $k = 1$. Before modifying the algorithm for general $k \in \mathbb{N}$, let us analyze its performance.

If Case 1 holds, then the analysis provided in the proof of Theorem 2.2 holds, when adapted to using more accurate estimates for the $\mathcal{D}(W_j)$'s. Recall that in this case we obtain, for each $j \in [i', i]$, an estimate $\widetilde{\delta}_j$ of $\mathcal{D}(W_j)$ such that (with probability at least $1 - 1/10\log_\beta(6\lambda/(\beta-1)^3\epsilon)$) it holds that $\widetilde{\delta}_j \in [\mathcal{D}(W_j), \beta \cdot \mathcal{D}(W_j)]$ if $\mathcal{D}(W_j) > \epsilon'/\beta^2$ and $\widetilde{\delta}_j < \epsilon'/\beta$ otherwise, where in the latter case $\widetilde{\delta}_j$ is reset to $\epsilon' = (\beta-1)^3 \cdot \epsilon/6\lambda$. Hence, $|W_j| \leq \widetilde{w}_j \leq \max(\beta^2 \cdot |W_j|, \epsilon' \cdot \epsilon^j)$ for every $j \in [i', i]$, whereas the fact that $\widetilde{w}_j \geq |W_j|$ (for all $j \in [i', i]$) implies that $\mathtt{ess}_{\beta^2 \cdot \epsilon}(\mathcal{D}) \leq \frac{\beta^{i'}}{\beta-1} + \sum_{j \in [i', i]} \widetilde{w}_j$. Observing that $\mathcal{D}(W_i) \geq \eta \stackrel{\text{def}}{=} (\beta-1) \cdot \epsilon/6\lambda = \epsilon'/(\beta-1)^2$ (rather than merely $\mathcal{D}(W_i) \geq (\beta-1) \cdot \epsilon/6\ell$ as stated in the proof of Theorem 2.2) and recalling that $i' = \max(1, i - \log_\beta(1/\epsilon'))$, we invoke Claim 2.2.2 and obtain

$$\begin{aligned}
\frac{\beta^{i'}}{\beta-1} + \sum_{j \in [i', i]} \widetilde{w}_j &< \frac{\beta}{\beta-1} + \left(1 + \frac{2\epsilon'}{(\beta-1) \cdot \eta}\right) \cdot \beta^2 \cdot \sum_{j \leq i} |W_j| \\
&= \frac{\beta}{\beta-1} + (1 + 2 \cdot (\beta-1)) \cdot \beta^2 \cdot \sum_{j \leq i} |W_j| \\
&< \beta^5 \cdot \sum_{j \leq i} |W_j|,
\end{aligned}$$

where the last inequality may be assumed without loss of generalty.[18] Recalling that $\mathtt{ess}_\epsilon(\mathcal{D}) \geq \beta^{-2} \cdot \sum_{j \leq i} |W_j|$ (by the simplified form of Part 2 of Claim 2.1.2), it follows that the output in this case (i.e., $\frac{\beta^{i'}}{\beta-1} + \sum_{j \in [i', i]} \widetilde{w}_j$) is a β^7-factor approximation of the $[\epsilon, \beta^2 \cdot \epsilon]$-effective support size of $\mathcal{D}$.

When Case 2 holds we use $\sum_{j \in [\widetilde{\ell}' + 2\log_\beta \lambda + 1, \widetilde{\ell}' + \lambda]} \mathcal{D}(W_j) > 0.1\Delta'/t$ in order to infer that an adequate i (i.e., $i \in [\widetilde{\ell}' + 2\log_\beta \lambda + 1, \widetilde{\ell}' + \lambda]$) is indeed selected (w.h.p.) after $O(t/\epsilon)$ trials. Furthermore, with probability at least 0.9, it holds that $\mathcal{D}(W_i) > \Delta'/100t\lambda \geq (\beta-1)\epsilon/300t\lambda$. Using the minimality of ℓ', which implies $\sum_{j \leq \widetilde{\ell}'+1} \mathcal{D}(W_j) \leq \beta^2 \cdot \epsilon$, and $i > \widetilde{\ell}' + 2\log_\beta \lambda$ (equiv., $i - 2\log_\beta \lambda > \widetilde{\ell}'$), we upper-bound $\mathtt{ess}_{\beta^2 \cdot \epsilon}(\mathcal{D})$ by $\sum_{j \leq i - 2\log_\beta \lambda - 1} |W_j|$. Hence, using $i > \widetilde{\ell}' + 2\log_\beta \lambda \geq \widetilde{\ell}' + \log_\beta(300t\lambda/(\beta-1)^2\epsilon)$, where the last inequality is due to $\lambda \geq c/\epsilon$, we get

$$\begin{aligned}
\mathtt{ess}_{\beta^2 \cdot \epsilon}(\mathcal{D}) &\leq \sum_{j \leq i - 2\log_\beta \lambda - 1} \beta^j \\
&< \frac{\beta^{i - \log_\beta(300t\lambda/(\beta-1)^2\epsilon)}}{\beta-1} \\
&= \frac{\epsilon \cdot (\beta-1)}{300t\lambda} \cdot \beta^i
\end{aligned}$$

which implies that the output (in this case) is at least $\mathtt{ess}_{\beta^2 \cdot \epsilon}(\mathcal{D})$. On the other hand, $\mathcal{D}(W_i) > (\beta-1)\epsilon/300t\lambda$ implies that $|W_i| > (\beta-1) \cdot \epsilon \cdot \beta^{i-1}/300t\lambda$, and applying (the simplified form of) Part 2 of Claim 2.1.2, we get

$$\mathtt{ess}_\epsilon(\mathcal{D}) \geq \beta^{-2} \cdot \frac{(\beta-1) \cdot \epsilon}{300t\lambda} \cdot \beta^{i-1}.$$

[18] Specifically, $\frac{\beta}{\beta-1} < (\beta-1) \cdot \sum_{j \leq i} |W_j|$ follows from $\sum_{j \leq i} |W_j| = \omega(1)$, which can be justified by an alternative approximation procedure that holds in case $m \stackrel{\text{def}}{=} \sum_{j \leq i} |W_j| = O(1)$. Recalling that $\sum_{j > i} \mathcal{D}(W_j) < \beta^2 \cdot \epsilon$, we show how to find an $[\beta^2\epsilon, \beta^3\epsilon]$-effective support size of $\mathcal{D}$ in $O(1/\epsilon)$ time. Specifically, letting $W = \bigcup_{j \leq i} W_j$ and $H = \{e \in W : \mathcal{D}(e) \geq (\beta^3 - \beta^2) \cdot \epsilon/m\}$, observe that $\mathcal{D}(H) \geq \mathcal{D}(W) - (\beta^3 - \beta^2) \cdot \epsilon > 1 - \beta^3 \cdot \epsilon$. The suggested procedure finds all elements in H using $O(1/\epsilon)$ samples, and outputs the largest v such the total weight of the heaviest v elements in H is at most $1 - \beta^2 \cdot \epsilon$. Denoting the set of the heaviest v elements by H', it follows that $1 - \mathcal{D}(H') \in [\beta^2 \cdot \epsilon, \beta^3 \cdot \epsilon]$ and $\mathtt{ess}_{1-\mathcal{D}(H')}(\mathcal{D}) = v$. Hence, v is an $[\beta^2\epsilon, \beta^3\epsilon]$-effective support size of $\mathcal{D}$.

Hence, the output (i.e., $(\beta-1)\cdot\epsilon\cdot\beta^i/300t\lambda$) is at most $\beta^3\cdot\mathtt{ess}_\epsilon(\mathcal{D})$. Combining both bounds, we infer that (in this case) the output is an β^3-factor approximation of the $[\epsilon,\beta^2\cdot\epsilon]$-effective support size of $\mathcal{D}$.

We are left with two cases: The case of $t'=1$ (handled in the preamble of Algorithm 2.3.1) and Case 3 (in which $t'>1$). In the latter case (i.e., for $t'>1$), we recurse, and otherwise (i.e., for $t'=1$), we invoke Algorithm 2.2.1 (with the current Δ' and λ). The key observation is that, at this time (i.e., when $t'=1$), it holds that $\lambda=O(\log^{(t-1)}(O(\log(n/\epsilon))))$ and $\Delta'\geq(1-(0.1/t))^{t-1}\cdot\widetilde{\Delta}>0.9\widetilde{\Delta}$. Hence, this invocation produces an $O(\lambda)$-factor approximation of the $[\epsilon,\beta\cdot\epsilon]$-effective support size of $\mathcal{D}$. This establishes the theorem for $k=1$.

Turning to general $k\in\mathbb{N}$, we modify Algorithm 2.3.1 by merely replacing the thresholds that govern the choice of cases. Specifically, for distinguishing Case 1 from Cases 2–3, we use a threshold of $k^2\cdot(c/\epsilon)^{1/k}$ rather than c/ϵ, whereas distinguishing between Case 2 and Case 3 is done based on the value of $\sum_{j\in[\widetilde{\ell}',\widetilde{\ell}'+(k+1)\cdot\log_\beta\lambda]}\mathcal{D}(W_j)$ (rather than $\sum_{j\in[\widetilde{\ell}',\widetilde{\ell}'+2\log_\beta\lambda]}\mathcal{D}(W_j)$). Similarly, at the end of Case 3, the `remaining interval length` is set to $(k+2)\cdot\log_\beta\lambda$ (rather than to $3\cdot\log_\beta\lambda$), and $\lambda\geq k^2\cdot(c/\epsilon)^{1/k}$ is used to argue that $(k+2)\cdot\log_\beta\lambda<\lambda$. Hence, the complexity of Case 1 is $\widetilde{O}(\lambda/\epsilon)=\widetilde{O}(t/\epsilon^{1+\frac{1}{k}})$ (rather than $\widetilde{O}(t/\epsilon^2)$), whereas in the analysis of Case 2 we use $(k+1)\cdot\log_\beta\lambda\geq\log_\beta(c\lambda/\epsilon)$ (rather than $2\cdot\log_\beta\lambda\geq\log_\beta(c\lambda/\epsilon)$).[19] The theorem follows. ■

Theorem 2.4 (the iterative algorithm, revised): *For every constants $\beta>1$ and $k\in\mathbb{N}$, there exists an algorithm that on input $\epsilon>0$ and oracle access to $\mathcal{D}$, runs in expected time $\widetilde{O}(\log^*(n/\epsilon)/\epsilon^{1+\frac{1}{k}})$ and outputs an β-factor approximator of the $[\epsilon,\beta\cdot\epsilon]$-effective support size of $\mathcal{D}$, where $n=\mathtt{ess}_\epsilon(\mathcal{D})$ and $\log^*m$ is the minimal $t\in\mathbb{N}$ satisfying $\log_2^{(t)}m<2$.*

Unlike in the previous three theorems, the running time stated in Theorem 2.4 depends on the effective support size and is bounded in expectation only.[20] These two features seem related, since such an algorithm should obtain some crude and necessarily randomized estimate of the effective support size of the distribution in order to determine its own running time. On the other hand, recall that by Observation 1.5, *Theorem 2.4 implies a 1-factor approximator of the $[\epsilon,(\beta+(\beta-1))\cdot\epsilon]$-effective support size of $\mathcal{D}$*, and by change of parameters we infer that the output is an $[\epsilon,\beta\cdot\epsilon]$-effective support size of $\mathcal{D}$.

Proof: Intuitively, we invoke Theorem 2.3, while setting $t=\log^*(n/\epsilon)$. However, in this case, t is not a constant, and we do not know it. Still, we can overcome these difficulties in one of two ways, where the more elegant way (presented first) was suggested to us by Clement Canonne. The crucial observation, used in both ways, is that *Cases 1 and 2 in Algorithm 2.3.1 provide a β-factor approximation, whereas the case of $t'=1$ can be avoided*. Hence, recursing till either Cases 1 or Case 2 occurs, we obtain the desired approximation factor.

The first way to do so is to first obtain a very crude approximation of the effective support size. Specifically, invoking the basic algorithm (of Theorem 2.1), we obtain in $O(1/\epsilon)$-time an $O(\epsilon^{-1}\log(n/\epsilon))$-factor approximation of the $[\epsilon,\beta\cdot\epsilon]$-effective support size of $\mathcal{D}$, where $n=\mathtt{ess}_\epsilon(\mathcal{D})$. Denoting this value by $\widetilde{n}$, we set $t=4k\cdot\log^*_\beta(\widetilde{n}/\epsilon))=O(\log^*(n/\epsilon))$, and invoke Algorithm 2.3.1, while noting that this setting of t prevents the algorithm from ever reaching the case of $t'=1$ (since iterating $\lambda\leftarrow(k+2)\log_\beta\lambda$

[19] Indeed, here we use $\lambda\geq(c/\epsilon)^{1/k}$. The fact that Cases 2 and 3 actually presumes $\lambda\geq k^2\cdot(c/\epsilon)^{1/k}$ is used only when verifying that $(k+2)\log_\beta\lambda<\lambda$.

[20] For any $t\in\mathbb{N}$, one can generically convert an approximator that runs for expected time $T=T(\epsilon,\mathcal{D})$, where T is unknown *a priori*, into an approximator that runs for $O(t\cdot T)$-time with probability at least $1-2^{-\Omega(t)}$. To do so, we invoke the algorithm t times in parallel, suspends all executions as soon as 90% of the them terminate, and output the median value obtained in these $0.9t$ invocations. The point is that, with probability at least 0.95, a random execution runs for at most $20\cdot T$ steps. Hence, with probability at least $1-2^{-\Omega(t)}$, more than 90% of the executions will terminate within $20\cdot T$ steps and most of them will output a correct value (i.e., an $[\epsilon,\beta\cdot\epsilon]$-effective support size).

for $t-2$ times, starting with $\lambda = O(\log_\beta(\widetilde{n}/\epsilon))$, yields a value smaller than $k^2 \cdot (c/\epsilon)^{1/k}$).[21] The theorem follows.

The alternative way is to adapt Algorithm 2.3.1 so that it does not use t at all. Specifically, first, we replace the varying parameter t', which represent the *remaining* recursion-depth, by a varying parameter that represents the *current* recursion-depth, and remove the stopping rule that refers to the case that $t' = 1$. Second, we change the threshold that distinguishes the two main cases (i.e., Cases 2 and 3) from $\left(1 - \frac{0.1}{t}\right) \cdot \Delta'$ to $\left(1 - \frac{1}{g(t'')}\right) \cdot \Delta'$, where t'' represents the current recursion depth and $g(m) = \widetilde{O}(m)$ satisfies $\sum_{m \geq 1}(1/g(m)) < 0.1$ (e.g., $g(m) = 20m \cdot \log_2^2(m+1)$ will do).[22] Lastly, the constant $c = 300t/(\beta-1)^2$, will be replaced by a variable c that will be set to $c = 30g(t'')/(\beta-1)^2$ (assuming $g(m) = 20m \cdot \log_2^2(m+1)$ is used). What will happen is that we shall either stop at Case 2 or at Case 1, since if we never stop at Case 2 then at some point we shall reach $\lambda < 600 < c/\epsilon$. In the analysis, we use the fact that $\prod_{t'' \geq 1}\left(1 - \frac{1}{g(t'')}\right) > 1 - \sum_{t'' \geq 1}\frac{1}{g(t'')} > 0.9$. ■

[21] In fact, for sufficiently large k, this $(t-2)$-step iterative process yields a value smaller than k^2.

[22] Note that $\sum_{m \geq 1}\frac{1}{m \cdot \log_2^2(m+1)} < \sum_{m \geq 1}\frac{1}{m \cdot \lfloor \log_2(m+1) \rfloor^2} < \sum_{i \geq 1} 2^i \cdot \frac{1}{2^i \cdot i^2} < 2$.

Acknowledgments

I am grateful to Clement Canonne for numerous comments and suggestions regarding a prior version of this write-up.

This project was partially supported by the Israel Science Foundation (grant No. 1146/18), and has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No. 819702).

References

[1] T. Batu, S. Dasgupta, R. Kumar, and R. Rubinfeld. The complexity of approximating the entropy. *SICOMP*, Vol. 35 (1), pages 132–150, 2005.

[2] E. Blais, C.L. Canonne, and T. Gur. Distribution Testing Lower Bounds via Reductions from Communication Complexity. In *32nd Computational Complexity Conference*, pages 28:1–28:40, 2017.

[3] C.L. Canonne and R. Rubinfeld. Testing Probability Distributions Underlying Aggregated Data. In *41st ICALP*, pages 283–295, 2014.

[4] O. Goldreich. *Introduction to Property Testing*. Cambridge University Press, 2017.

[5] O. Goldreich. Testing Graphs in Vertex-Distribution-Free Models. *ECCC*, TR18-171, 2018. (See Revision Nr 1, March 2019.)

[6] O. Goldreich. Testing Bipartitness in an Augmented VDF Bounded-Degree Graph Model. *arxiv*, 1905.03070 [cs.DS], 2019.

[7] M.J. Kearns, Y. Mansour, D. Ron, R. Rubinfeld, R.E. Schapire, and L. Sellie. On the learnability of discrete distributions. In *26th STOC*, pages 273–282, 1994.

[8] K. Onak and X. Sun. Probability-Revealing Samples. In *21st AISTATS*, pages 2018–2026, 2018.

[9] S. Raskhodnikova, D. Ron, A. Shpilka, and A. Smith. Strong Lower Bounds for Approximating Distribution Support Size and the Distinct Elements Problem. *SICOMP*, Vol. 39 (3), pages 813–842, 2009.

Appendix A: Reproducing Algorithms from [6]

This appendix reproduces the algorithms presented by us in [6, Sec. 2.2], while adapting the notation to the one used in the current write-up. Both algorithms are different from all algorithms presented in Section 2, but they obtain inferior results.

A simple approximation of the effective support size. We first present a simple algorithm for obtaining a rather rough (but sufficiently good for our purposes) approximation. Given an "effectiveness" parameter ϵ, we proceed in iterations such that in the i^{th} iteration we *take a sample of* $m = O(\epsilon^{-1} \log i)$ *elements of* $\mathcal{D}$*, and halt outputting* $2^i/\epsilon$ *if the number of samples that have* $\mathcal{D}$*-value below* $\epsilon/2^i$ *is at most* $3\epsilon \cdot m$.

Observer that (in iteration i), with probability at least $1 - 0.01/i^2$, the sample approximates the total weight of the "light elements" (i.e., elements having $\mathcal{D}$-value below $\epsilon/2^i$) in the sense that if $\Pr_{e \leftarrow \mathcal{D}}[\mathcal{D}(e) < \epsilon/2^i] < 2\epsilon$, (resp., if $\Pr_{e \leftarrow \mathcal{D}}[\mathcal{D}(e) < \epsilon/2^i] \geq 4\epsilon$), then the number of samples that have $\mathcal{D}$-value below $\epsilon/2^i$ is at most $3\epsilon \cdot m$ (resp., is greater than $3\epsilon \cdot m$).

Now, letting n be an upper bound on the ϵ-effective support size of $\mathcal{D}$ (and assuming for simplicity that n is a power of two), observe that, with high (constant) probability, we halt by iteration $i^* = \log_2 n$, because $\Pr_{e \leftarrow \mathcal{D}}[\mathcal{D}(e) < \epsilon/2^{i^*}] < n \cdot \epsilon/2^{i^*} + \epsilon = 2\epsilon$, where the first (resp., second) term is due to elements that are (resp., are not) in the effective support of $\mathcal{D}$. Hence, with high (constant) probability, the algorithm outputs a value that is at most n/ϵ. (Actually, if we reach iteration i^*, we halt in it with probability at least $1 - 0.01/(i^*)^2$; it follows that with probability at $1 - 0.01/\log^2 n$, the algorithm outputs a value that is at most n/ϵ.)

On the other hand, assuming that the 4ϵ-effective support size of $\mathcal{D}$ is at least n' (and assuming that n' is also a power of two), with high (constant) probability, we do not halt before iteration $i^+ = \log_2(\epsilon \cdot n')$, because otherwise $\Pr_{e \leftarrow \mathcal{D}}[\mathcal{D}(e) < \epsilon/2^{i^+-1}] < 4\epsilon$, which implies that $\mathcal{D}$ has 4ϵ-effective support size at most $n'/2$ (since $\Pr_{e \leftarrow \mathcal{D}}[\mathcal{D}(e) < 2/n'] < 4\epsilon$ implies that $\mathcal{D}$ is 4ϵ-close to a distribution of support size at most $n'/2$). Hence, with high (constant) probability (i.e., with probability at least $1 - \sum_{i<i^+} 0.01/i^2 > 0.98$), the algorithm outputs a value that is at least n'.

It follows that, with high (constant) probability, the algorithm outputs a number that lies between n' and n/ϵ (i.e., between the half the 4ϵ-effective support size of $\mathcal{D}$ and $2/\epsilon$ times its ϵ-effective support size). Furthermore, with high (constant) probability, this algorithm runs for $\sum_{i \leq i^*} O(\epsilon^{-1} \cdot \log i) = \widetilde{O}(\log n)/\epsilon$ steps (and its expected number of steps can be similarly bounded). Hence, we got

Theorem A.1 (obtaining a crude approximation of the effective support size): *There exist an* $(4/\epsilon)$*-factor approximator of the* $[\epsilon, 4\epsilon]$*-effective support of distributions that runs in expected time* $\widetilde{O}(\epsilon^{-1} \log n)$*, where* n *denote the approximator's output.*

Building on any such crude approximation of the effective support size, one can obtain a better approximation is shown next.

Obtaining better approximations of the effective support size. For starters, we present a tighter analysis of (a minor variant of) the foregoing algorithm. Specifically, for any constant $\beta > 1$, in the i^{th} iteration, we *take a sample of* $m = O(\epsilon^{-1} \log i)$ *elements of* $\mathcal{D}$*, and halt outputting* $2^i/\epsilon$ *if the number of samples that have* $\mathcal{D}$*-value below* $(\beta - 1) \cdot \epsilon/2^i$ *is at most* $\beta^2 \cdot \epsilon \cdot m$. In analyzing the probability that this algorithm halts by iteration $i^* = \log_2 n$, we use the fact that $\Pr_{e \leftarrow \mathcal{D}}[\mathcal{D}(e) < (\beta - 1) \cdot \epsilon/2^{i^*}] < n \cdot (\beta - 1) \cdot \epsilon/2^{i^*} + \epsilon = \beta \cdot \epsilon$, whereas in analyzing the probability that the algorithm does halts before iteration $i^+ = \log_2(\epsilon \cdot n')$ implies $\Pr_{e \leftarrow \mathcal{D}}[\mathcal{D}(e) < (\beta - 1) \cdot \epsilon/2^{i^+-1}] < \beta^3 \cdot \epsilon$ (where here n' is the $\beta^3\epsilon$-effective support size).[23] It follows that, with high (constant) probability, the algorithm

[23] Here we use the fact that (in iteration i), with probability at least $1 - 0.01/i^2$, the sample approximates the total weight of the "light elements" (i.e., elements having $\mathcal{D}$-value below $\epsilon/2^i$) in the sense that if $\Pr_{e \leftarrow \mathcal{D}}[\mathcal{D}(e) < (\beta - 1) \cdot \epsilon/2^i] < \beta \cdot \epsilon$,

outputs a number, denoted $\widetilde{n}$, that lies between half the $\beta^3 \cdot \epsilon$-effective support size of $\mathcal{D}$ and $2/\epsilon$ times its ϵ-effective support size. That is, we obtain a $(4/\epsilon)$-factor approximator of the $[\epsilon, \beta^3\epsilon]$-effective support of distributions.

To obtain an even better approximation of the effective support size, we use the rough estimate $\widetilde{n}$ obtained above in order to approximate the number of elements that have $\mathcal{D}$-value approximately $\beta^{i-0.5}$ for every $i \in [O(\log \widetilde{n}/\eta)]$. Indeed, our first step is ignoring elements having $\mathcal{D}$-value below $\epsilon/\widetilde{n}$ or so. Specifically, setting $\epsilon' = \beta^3 \cdot \epsilon$, recall that if $\mathcal{D}$ has an ϵ'-effective support of size $\widetilde{n}$, then $\mathcal{D}(H) \geq 1-\beta\epsilon'$ for $H \stackrel{\text{def}}{=} \{v : \mathcal{D}(v) \geq (\beta-1)\cdot\epsilon'/\widetilde{n}\}$ (see prior paragraph, while replacing i^* by $\log_2 \widetilde{n}$).[24] Hence, assuming that $\mathcal{D}(H) \leq 1-\epsilon$ and letting $\epsilon'' = \beta \cdot \epsilon' = \beta^4 \cdot \epsilon$, it holds that $|H|$ lies between the minimal ϵ''-effective support size of $\mathcal{D}$ and its minimal ϵ-effective support size, and so providing a good approximation of the "effective size" of H will do. (In the case that $\mathcal{D}(H) > 1-\epsilon$ additional steps will be needed.)

To (effectively) approximate $|H|$, we let $W_i = \{e : \beta^{-i} < \mathcal{D}(e) \leq \beta^{-(i-1)}\}$, and observe that it suffice the approximate $\mathcal{D}(W_i)$ for $i = 1, ..., \ell$, where $\ell \stackrel{\text{def}}{=} \log_\beta(\widetilde{n}/(\beta-1)\cdot\epsilon') = O((\beta-1)^{-1} \cdot \log(\widetilde{n}/\epsilon))$. Actually, letting $I = \{i \in [\ell] : \mathcal{D}(W_i) \geq (\beta-1)\epsilon''/\ell\}$, it suffices to approximate $\mathcal{D}(W_i)$ for every $i \in I$, which yields approximations of the corresponding $|W_i|$'s (using $|W_i| \approx \mathcal{D}(W_i)/\beta^{-(i-0.5)}$). That is, we do not actaully approximate $|H|$ but rather approximate the size of $H' \stackrel{\text{def}}{=} \bigcup_{i\in I} W_i \subseteq H$, while capitalizing on $\mathcal{D}(H \setminus H') \leq (\beta-1)\cdot\epsilon''$. Hence, we will approximate the minimal ϵ'''-support size for some $\epsilon \leq \epsilon''' \leq \beta\epsilon'' = \beta^5\epsilon$. Specifically, for each $i \in [\ell]$, using a sample of $O(t\ell/(\beta-1)^2 \cdot \epsilon')$ samples, we obtain (with probability $1-2^{-t}$) a β-factor approximation of $\mathcal{D}(W_i)$ for each $i \in I$, which yields a β^2-factor approximation of $|W_i|$ (since $|W_i| \in [\beta^{i-1}\mathcal{D}(W_i), \beta^i \cdot \mathcal{D}(W_i))$). Note that the rough estimate of the effective support size of $\mathcal{D}$ (i.e., $\widetilde{n}$) is only used in order to determine ℓ.

Recall that we have assumed that $\mathcal{D}(H) \leq 1-\epsilon$, whereas this is not necessarily the case. Needless to say, we can easily estimate $1-\mathcal{D}(H)$ up to any desired constant factor (using $O(1/\epsilon)$ samples of $\mathcal{D}$). In case we are quite sure that $\mathcal{D}(H) > 1-\epsilon$ (which will happen if $\mathcal{D}(H) > 1-\beta\epsilon$ but not if $\mathcal{D}(H) < 1-\beta^{-1}\epsilon$), we can reduce the estimate obtained for $|H|$ by disposing an adequate weight of H; that is, we dispose as many as the the lightest elements as possible till reaching a subset of H' that has $\mathcal{D}$-value that is most likely to be below $1-\epsilon$. Note that the foregoing process is performed without making any samples or queries; it is solely based on the estimated values of $\mathcal{D}(W_i)$ for $i \in I$ already obtained. Hence, we get.

Theorem A.2 (obtaining a good approximation of the effective support size): *For every constant $\beta > 1$, there exist a β^2-factor approximator of the $[\beta^{-1}\cdot\epsilon, \beta^5\cdot\epsilon]$-effective support of distributions that makes a number of queries that is almost logarithmic in the quantity it outputs. Specifically, when given oracle access to $\mathcal{D}$, the expected number of steps performed by the approximator when outputting the value n is $\widetilde{O}(\epsilon^{-1}\log n)$.*

Needless to say, by a change in parameters we can make n lie between the $\beta\cdot\eta$-effective support size of $\mathcal{D}$ and β times its η-effective support size. Hence, we obtain a β-factor approximator of the $[\epsilon, \beta\cdot\epsilon]$-effective support of distributions.

Appendix B: Another Inferior Algorithm

This appendix presents an algorithm that is similar to the one stated in Theorem A.2. Achieving the same result as Theorem A.2, it is inferior to Theorem 2.4. (This algorithm was devised by us before

(resp., if $\Pr_{e\leftarrow\mathcal{D}}[\mathcal{D}(e) < (\beta-1)\cdot\epsilon/2^i] \geq \beta^3\cdot\epsilon$), then the number of samples that have $\mathcal{D}$-value below $(\beta-1)\cdot\epsilon/2^i$ is at most $\beta^2\cdot\epsilon\cdot m$ (resp., is greater than $\beta^2\cdot\epsilon\cdot m$).

[24] In other words, observe that $\Pr_{v\leftarrow\mathcal{D}}[\mathcal{D}(v) < (\beta-1)\cdot\epsilon'/\widetilde{n}] < \widetilde{n}\cdot(\beta-1)\cdot\epsilon'/\widetilde{n} + \epsilon' = \beta\cdot\epsilon'$, where the first (resp., second) term is due to elements that are (resp., are not) in the effective support of $\mathcal{D}$.

realizing that the analysis of Algorithm 2.3.1 can be tighted so that Theorem 2.4 can yield a β-factor approximation rather than an $O(1)$-factor approximation.)

Using any crude approximation of the effective support size, one can obtain a better approximation factor by estimating the values of the $\mathcal{D}(W_j)$'s for all $j \in [\widetilde{\ell}]$. The point is that the crude approximation is used only to determine a good upper bound on $\ell = O(\log(n/\epsilon))$, where $n = \mathtt{ess}_\epsilon(\mathcal{D})$. The best result of this type is obtained by using Algorithm 2.2.1 as a starting point, or rather by using a variant of the proof of Theorem 2.2.

Theorem B.1 (obtaining a β-factor approximation): *For every constant $\beta > 1$, there exists an algorithm that on input $\epsilon > 0$ and oracle access to $\mathcal{D}$, runs in expected time $\widetilde{O}(\epsilon^{-1} \cdot \log(\mathtt{ess}_\epsilon(\mathcal{D})))$ and outputs an β-factor approximator of the $[\epsilon, \beta \cdot \epsilon]$-effective support size of $\mathcal{D}$.*

Recall that by Observation 1.5, Theorem B.1 implies a 1-factor approximator of the $[\epsilon, (\beta + (\beta - 1)) \cdot \epsilon]$-effective support size of $\mathcal{D}$, and by change of parameters we infer that the output is an $[\epsilon, \beta \cdot \epsilon]$-effective support size of $\mathcal{D}$.

Proof: We adapt the proof of Theorem 2.2 by obtaining more accurate approximations of the relevant $\mathcal{D}(W_j)$'s and tightening the analysis. Specifically, rather than obtaining a constant factor approximation for each $\mathcal{D}(W_j) = \Omega(\epsilon)$, we obtain such an approximation for each $\mathcal{D}(W_j) = \Omega(\epsilon/\ell)$. Of course, this requires increasing the sample complexity by a factor of $O(\ell)$, which we can afford per the claim of the current theorem. (Indeed, the adaptation is identical to the one used in Case 1 of Algorithm 2.3.1, except that here we use ϵ/ℓ instaed of ϵ/λ.) Specifically, our algorithm proceeds as follows.

- As in Algorithm 2.2.1, the algorithm starts by determining $\widetilde{\ell}$ and $\widetilde{\ell'}$ and selecting $i \in [\widetilde{\ell'}, \widetilde{\ell}]$ as in Algorithm 2.1.3. (Actually, we could have afforded to select $i \in [\widetilde{\ell'}, \ell]$ as having the highest $\widetilde{\delta}_i$ value among those in $[\widetilde{\ell'}, \widetilde{\ell}]$, rather than selecting $i \in [\widetilde{\ell'}, \widetilde{\ell}]$ with proportional to $\mathcal{D}(W_i)$.)
- Next, letting $i' = \max(1, i - \log_\beta(6\beta\ell/((\beta-1)^3 \cdot \epsilon)))$ (rather than $i' = \max(1, i - \log_\beta(1/((\beta-1)\epsilon)))$), and using a sample of size $\widetilde{O}(\ell/\epsilon)$, the algorithm obtains, for each $j \in [i', i]$, an estimate $\widetilde{\delta}_j$ of $\mathcal{D}(W_j)$ such that (with probability at least $1 - 1/10\log_\beta(\ell/\epsilon)$) it holds that $\widetilde{\delta}_j \in [\mathcal{D}(W_j), \beta \cdot \mathcal{D}(W_j)]$ if $\mathcal{D}(W_j) \geq \eta \stackrel{\text{def}}{=} \frac{(\beta-1)^3 \cdot \epsilon}{6\beta\ell}$ and $\widetilde{\delta}_j < \beta \cdot \eta$ otherwise.
- As in Algorithm 2.2.1, for each $j \in [i', i]$, if $\widetilde{\delta}_j < \beta \cdot \eta$, then the algorithm resets $\widetilde{\delta}_j \leftarrow \eta$.
- Finally, for each $j \in [i', i]$, the algorithm sets $\widetilde{w}_j \leftarrow \widetilde{\delta}_j \cdot \beta^j$, it outputs $\frac{\beta^{i'}}{\beta-1} + \sum_{j \in [i', i]} \widetilde{w}_j$ as its estimate of the effective support size.

As in the proofs of Theorems 2.1 and 2.2, the analysis of the foregoing algorithm focus on the case that the selected i satisfies $\mathcal{D}(W_i) \geq \frac{(\beta-1) \cdot \epsilon}{6\ell} = \frac{\beta \cdot \eta}{(\beta-1)^2} > \eta$ (assuming, w.l.o.g, that $\beta < 2.618$). It follows that $\widetilde{\delta}_i \in [\mathcal{D}(W_i), \beta \cdot \mathcal{D}(W_i)]$ (w.h.p.). Hence, $|W_i| \leq \widetilde{w}_i \leq \beta^2 \cdot |W_i|$ holds. Likewise, with high probability, for each $j \in [i', i-1]$, if $\mathcal{D}(W_j) \geq \eta$ then $|W_j| \leq \widetilde{w}_j \leq \beta^2 \cdot |W_j|$, and otherwise $\widetilde{\delta}_j = \eta$ and $|W_j| < \mathcal{D}(W_j) \cdot \beta^j < \eta \cdot \beta^j = \widetilde{w}_j$. Hence, with high probability, the output (i.e., $(\beta-1)^{-1} \cdot \beta^{i'} + \sum_{j \in [i', i]} \widetilde{w}_j$) is lower-bounded by $\sum_{j \leq i} |W_j|$, since

$$\begin{aligned} \sum_{j \leq i} |W_j| &= \sum_{j < i'} |W_j| + \sum_{j \in [i', i]} |W_j| \\ &< \frac{\beta^{i'}}{\beta - 1} + \sum_{j \in [i', i]} \widetilde{w}_j. \end{aligned}$$

Next, using $|W_i| \geq \beta^{i-1} \cdot \mathcal{D}(W_i) \geq \frac{\eta}{(\beta-1)^2} \cdot \beta^i$, we upper-bound the output value by $\beta \cdot \sum_{j \leq i} |W_j|$. We first recall that for each $j \in [i', i]$, either $\widetilde{w}_j \leq \beta^2 \cdot |W_j|$ or $\widetilde{w}_j = \eta \cdot \beta^{j-i} \cdot \beta^i \leq \beta^{j-i} \cdot (\beta-1)^2 \cdot |W_i|$, which implies

$$\widetilde{w}_j \leq \beta^2 \cdot |W_j| + (\beta-1)^2 \cdot \beta^{j-i} \cdot |W_i|. \tag{6}$$

We also use $\beta^{i'} \leq \max(\beta, \frac{(\beta-1)^3 \cdot \epsilon}{6\beta\ell} \cdot \beta^i) = \max(\beta, \eta \cdot \beta^i) \leq (\beta-1)^2 \cdot |W_i|$, where the first inequality is due to the definition of i' and the second inequality is due to $|W_i| \geq \beta^i \cdot \eta/(\beta-1)^2$. Hence,

$$\begin{aligned}
\frac{\beta^{i'}}{\beta-1} + \sum_{j \in [i',i]} \widetilde{w}_j &= \frac{\beta^{i'}}{\beta-1} + \widetilde{w}_i + \sum_{j \in [i',i-1]} \widetilde{w}_j \\
&\leq (\beta-1) \cdot |W_i| + \beta^2 \cdot |W_i| + \sum_{j \in [i',i-1]} \left(\beta^2 \cdot |W_j| + (\beta-1)^2 \cdot \beta^{j-i} \cdot |W_i|\right) \\
&= \left((\beta-1) + (\beta-1)^2 \cdot \sum_{j \in [i',i-1]} \beta^{j-i}\right) \cdot |W_i| + \beta^2 \cdot \sum_{j \in [i',i]} |W_j| \\
&< \left((\beta-1) + (\beta-1)^2 \cdot \frac{1}{\beta-1}\right) \cdot |W_i| + \beta^2 \cdot \sum_{j \in [i',i]} |W_j| \\
&< (\beta^2 + 2 \cdot (\beta-1)) \cdot \sum_{j \in [i',i]} |W_j|.
\end{aligned}$$

Noting that $\beta^2 + 2(\beta-1) < \beta^4$, we have established that the algorithm outputs a β^6-factor approximation of the $[\epsilon, \beta \cdot \epsilon]$-effective support size of $\mathcal{D}$ (where the extra factor of β^2 is due to the use of the simplified form of Part 2 of Claim 2.1.2). Replacing β^6 by β, the proof is completed. ■