# On the Complexity of Estimating the Effective Support Size

Oded Goldreich*

October 27, 2021

**Abstract**

Loosely speaking, the effective support size of a distribution is the smallest size of the support of some 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.

In particular, for any constant $\beta > 1$ we present an algorithm that, on input $\epsilon > 0$ and oracle access to a distribution $\mathcal{D}$, uses $O(1/\epsilon^{1.001})$ samples and queries, and outputs a number $\widetilde{n}$ such that $\mathcal{D}$ is $\epsilon$-far from any distribution that has support of size $\widetilde{n}$ but is $\beta \cdot \epsilon$-close to a distribution that has support size $f \cdot \widetilde{n}$, where $f = O(\log\log\log\log\log(\widetilde{n}/\epsilon))$. (Indeed, 1.001 stands for any constant larger than 1, and $\log\log\log\log\log$ stands for any constant iterations of the logarithmic function.)

**Organization.** As is apparent from the abstract, our estimations approximate two parameters: the level of effectiveness and the support size. Hence, we start by presenting the relevant definitions (i.e., Definitions 1.1–1.3), making some initial observations (Section 1.2), and justifying our definitional choices (Section 1.3). We then state our main results (Section 1.4), which exhibit a trade-off between accuracy and sample complexity (see the various parts of Theorem 1.9), and provide overviews of their proofs (Section 1.5). The wider context is discussed in Section 1.6, whereas standard conventions and notations are presented in Section 1.7. The actual algorithms that establish the various parts of Theorem 1.9 are presented in Section 2.

## 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

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

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 $\epsilon$-*effective support size at most *$n$ if $\mathcal{D}$ is $\epsilon$-close to a distribution that has support size at most $n$, where $\mathcal{D}$ is $\epsilon$-*close *to $\mathcal{D}'$ if their total variation distance is at most $\epsilon$. The $\epsilon$-*effective support size of *$\mathcal{D}$, denoted $\mathtt{ess}_\epsilon(\mathcal{D})$, is the minimal $n$ such that $\mathcal{D}$ has $\epsilon$-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 $\delta$-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 $\epsilon$-effective support size $n$, then $\mathcal{D}$ has $\epsilon$-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 $\epsilon$-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 both 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 $\epsilon$-effective support size, for some predetermined $\epsilon > 0$, we seek a number that is upper-bounded by the $\epsilon$-effective support size and lower-bounded by the $\epsilon'$-effective support size (for some predetermined $\epsilon' > \epsilon$). Second, we may seek an approximation to the desired number rather than the number itself.

**Definition 1.2** (relaxations of the effective support size): *The natural 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 $\epsilon$-effective support size of $\mathcal{D}$; that is, $n = \mathtt{ess}_\epsilon(\mathcal{D})$. A value $\widetilde{n}$ is an $f$-*factor approximation of the $[\epsilon_1,\epsilon_2]$-effective support size of *$\mathcal{D}$ if it lies in the interval $[\mathtt{ess}_{\epsilon_2}(\mathcal{D}), f\cdot\mathtt{ess}_{\epsilon_1}(\mathcal{D})]$.*

Note that $n \in \mathbb{N}$ is an $[\epsilon_1,\epsilon_2]$-effective support size of $\mathcal{D}$ if and only if $n \in [\mathtt{ess}_{\epsilon_2}(\mathcal{D}), \mathtt{ess}_{\epsilon_1}(\mathcal{D})]$. (This is because for any $\epsilon \in [\epsilon_1,\epsilon_2]$, it holds that $\mathtt{ess}_{\epsilon_2}(\mathcal{D}) \leq \mathtt{ess}_\epsilon(\mathcal{D}) \leq \mathtt{ess}_{\epsilon_1}(\mathcal{D})$, whereas for any $n \in \mathbb{N}$ that is smaller than $\mathtt{ess}_0(\mathcal{D})$ there exists $\epsilon \in (0,1)$ such that $n = \mathtt{ess}_\epsilon(D)$.)[2]

As hinted, we are interested in algorithms that, for some $\epsilon_1,\epsilon_2$ 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] Let $\mathcal{D}'$ be the foregoing distribution that has $\eta$-effective support of size $n$. Then, the typical case is that this value of $\eta$ is not critical (w.r.t having $\eta$-effective support size $n$); that is, for some $\eta' < \eta$, the distribution $\mathcal{D}'$ has $\eta'$-effective support of size $n$. In this case, any distribution that is $(\eta-\eta')$-close to $\mathcal{D}'$ has $\eta$-effective support of size $n$. The pathological case is that $\mathcal{D}'$ has $\eta$-effective support of size $n$, but for every $\eta' < \eta$ the minimal $\eta'$-effective support size of $\mathcal{D}'$ is larger than $n$. We claim that in this case, for some $\eta' < \eta$, the distribution $\mathcal{D}'$ has $\eta'$-effective support of size $n+1$ (and it follows that any distribution that is $(\eta-\eta')$-close to $\mathcal{D}'$ has $\eta$-effective support of size $n+1$). To prove this claim, suppose that $\mathcal{D}'$ is $\eta$-close to a distribution $\mathcal{D}''$ of support size $n$, and consider the following 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 $v$ 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 $\mathcal{D}'(v) > 0$ from $\{u : \mathcal{D}''(u) > \mathcal{D}'(u)\}$ to $v$. Then, the modified distribution $\mathcal{D}'''$ has support size $n+1$ (i.e., the support of $\mathcal{D}'''$ is contained in the union of the support of $\mathcal{D}''$ and $v$) and is $(\eta-\mathcal{D}'(v))$-close to $\mathcal{D}'$. Hence, the claim follows with $\eta' = \eta - \mathcal{D}'(v)$.

[2] Needless to say, for any $r \in \mathbb{R}\setminus\mathbb{N}$ and $\epsilon \in [0,1)$, it holds that $r \neq \mathtt{ess}_\epsilon(\mathcal{D})$.

1. *What does it mean to have oracle access to a distribution?* One natural oracle associated with a distribution $\mathcal{D}$ is 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 $\mathcal{D}(e) = \Pr_{s\sim\mathcal{D}}[s=e]$; that is, $\mathtt{eval}_{\mathcal{D}}(e) = \mathcal{D}(e) = \Pr[\mathtt{samp}_{\mathcal{D}}=e]$.

   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 $\epsilon$ as a single effectiveness parameter, which we shall always keep varying. In contrast, the approximation parameters will sometimes be a function of $\epsilon$ and sometimes also depends on the distribution $\mathcal{D}$ (e.g., it may depend on the $\epsilon$-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 $\mathcal{C}$), with probability at least $2/3$, the random value $M^{\mathtt{samp}_{\mathcal{D}},\mathtt{eval}_{\mathcal{D}}}(\epsilon_1,\epsilon_2)$ is an $f$-factor approximation of the $[\epsilon_1,\epsilon_2]$-effective support size of $\mathcal{D}$; that is,*

$$\Pr\left[M^{\mathtt{samp}_{\mathcal{D}},\mathtt{eval}_{\mathcal{D}}}(\epsilon_1,\epsilon_2) \in [\mathtt{ess}_{\epsilon_2}(\mathcal{D}), f\cdot\mathtt{ess}_{\epsilon_1}(\mathcal{D})]\right] \geq 2/3.$$

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, so far, we did not restrict the complexity of the approximator; however, the complexity of such approximators is the focus of the current work. In what follows we shall consider the query complexity of the approximator as a function of $\epsilon_1,\epsilon_2$ and $f$ as well as of 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 size is obtained by omitting the lightest elements in the distribution): *If $\mathcal{D}$ has $\epsilon$-effective support size $n$, then $\mathcal{D}$ is $\epsilon$-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 $\epsilon$-close to some distribution of support size $n$).[3] Moving the probability mass of $\overline{H}$ to $H$, the claim

[3] 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 $\epsilon$, contradicting the hypothesis that $\mathtt{ess}_{\epsilon}(\mathcal{D}) = n$.

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 $\mathcal{D}$; then, $\mathtt{ess}_{\epsilon_2}(\mathcal{D}) \leq n \leq f \cdot \mathtt{ess}_{\epsilon_1}(\mathcal{D})$. Noting that $\mathcal{D}$ is $\epsilon_2$-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 $\mathcal{D}'_2$. Next, we observe that the probability mass that we moved is at most $\delta \stackrel{\text{def}}{=} (n_2 - (n_2/f)) \cdot 1/n_2 = (f-1)/f$, because the lightest element have average mass that is at most the average mass of all elements. Hence, $\mathcal{D}$ is $(\epsilon_2 + \delta)$-close to $D'_2$, 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 $\epsilon$-effective support size (i.e., Definition 1.1) by allowing two effectiveness thresholds (see Definition 1.2), because we found the former definition too restrictive. This view is substantiated by the following result.

**Proposition 1.6** (on the hardness of approximating the $\epsilon$-effective support size): *For any $\epsilon \in (0, 0.5)$ 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 $2\epsilon$-effective support size at most $n$ cannot provide an $f$-factor approximation of the $\epsilon$-effective support size of the distribution.*

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

**Proof:** Fixing $\epsilon \in (0, 0.5)$ and $n \in \mathbb{N}$, we pick a sufficiently large $N \gg n$, and consider the following two distributions:

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 $L$.
   (Indeed, $|H| \cdot \frac{1}{n} + |L| \cdot \frac{1}{N^2} = (1-\epsilon) + \epsilon$.)
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 $s \notin H' \cup L'$.
   (Indeed, $|H'| \cdot \frac{1}{n} + |L'| \cdot \frac{1}{N^2} + 1 \cdot \left(\frac{1}{n} - \frac{1}{N}\right) = \left(1 - \epsilon - \frac{1}{n}\right) + \left(\epsilon + \frac{1}{N}\right) + \left(\frac{1}{n} - \frac{1}{N}\right) = 1$.)

Note that an oracle machine that makes $o(n)$ queries cannot distinguish these two distributions (i.e., its distinguishing gap is $o(1)$).[4] On the other hand, $\mathcal{D}_1$ has $\epsilon$-effective support size $(1-\epsilon)\cdot n < n$, whereas $\mathcal{D}_2$ has $\epsilon$-effective support size $(1-\epsilon)\cdot n + N > N$, since the heaviest $(1-\epsilon)\cdot n + N$ elements in $\mathcal{D}_2$ have total weight $((1-\epsilon)\cdot n - 1)\cdot\frac{1}{n} + 1\cdot\left(\frac{1}{n}-\frac{1}{N}\right) + N\cdot\frac{1}{N^2} = 1-\epsilon$. Hence, the approximation factor provided by a $o(n)$-query machine is $\Omega(N/n)$, which cannot be bounded in terms on $\epsilon$ 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 integer multiples of $1/n$. The said result asserts that (for every $c \in (0, 1/16]$) *at least $\Omega(n^{1-3c^{1/2}})$ 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 0.04-effective support size at least $0.09n - 0.04n$ (since the lightest elements have each weight $1/n$), 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 within complexity 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 when using a number of samples that is significantly smaller than a square root of the effective support size*, since (for any $c > 0$) using $n^{0.5-c}$ samples 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, highly efficient 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 complexity of the algorithm and the approximation factor it obtains for the $[\epsilon, \beta\cdot\epsilon]$-effective support size (of distributions), where $\beta > 1$ is (typically) a constant. Specifically, at the most efficient extreme, we obtain an *algorithm that uses $O(1/\epsilon)$ samples and obtains an approximation factor that is logarithmic in the $\epsilon$-effective support size* (and almost linear in $1/\epsilon$). On the other hand, at the most accurate

[4] To streamline the argument, when considering the case that 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).

extreme, we obtain a *$\beta$-factor approximation algorithm than uses $\log^*(n/\epsilon)$ samples in expectation, where $n$ is the effective support size.*

**Theorem 1.9** (highly efficient and good approximators of the effective support size): *For any constant $\beta > 1$ and each of the following for options regarding $s$ and $f$, there exists an algorithm that, on input $\epsilon > 0$ and oracle access to $\mathcal{D}$, uses $s$ samples 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 $\epsilon$-effective support size of $\mathcal{D}$, the four options are:*

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

*In all cases, the algorithm queries only elements that have appeared in the sample. The dependence of the number of samples on $\beta$ is $\mathrm{poly}(1/(\beta-1))$. Recall that outputting an $f$-factor approximation of the $[\epsilon, \beta \cdot \epsilon]$-effective support size of $\mathcal{D}$ means that, with probability at least $2/3$, the output lies in the interval $[\mathtt{ess}_{\beta\epsilon}(\mathcal{D}), f \cdot \mathtt{ess}_\epsilon(\mathcal{D})]$.*

By Observation 1.5, using the last item, we can obtain $f = 1$ with $s = \widetilde{O}(\log^*(n/\epsilon)) \cdot \mathrm{poly}(1/\epsilon)$ samples in expectation, by letting $\beta = 1 + \epsilon$ and using the stated dependence of the sample complexity on $\beta$. Indeed, while the complexity bounds in the first three items hold for all executions, the bound in the last item refers to the expectation. This should not come as a surprise given that the fist three bounds only depend on $\epsilon$, whereas the latter bound depends also on the (*a priori*) unknown $n$.

It is not clear whether the trade-off between the sample complexity and the approximation factor exhibited by the foregoing four options is inherent. In particular, we wonder whether one can obtain $s$ and $f$ that are both functions of $\epsilon$ 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 $\mathcal{D}$, uses $s(\epsilon)$ samples and outputs an $f(\epsilon)$-factor approximation of the $[\epsilon, \beta \cdot \epsilon]$-effective support size of $\mathcal{D}$, where $s$ and $f$ are functions of $\epsilon$ only? If so, can both functions be polynomials in $1/\epsilon$? And, if so, can we have $s(\epsilon) = \mathrm{poly}(1/\epsilon)$ 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., complexity that is lower-bounded by some unbounded function of the size and is upper-bounded by a log-star in that size).

## 1.5 Techniques

The algorithms that establish the four items of Theorem 1.9 are all based on clustering the elements of the distribution's support according to their approximate probability mass (or weight). Specifically, the $i^{\text{th}}$ cluster, denoted $W_i$, contains all elements that have probability approximately $\beta^{-(i-0.5)}$; that is, $W_i \stackrel{\text{def}}{=} \{e : \beta^{-i} < \mathcal{D}(e) \leq \beta^{-(i-1)}\}$, where $\beta > 1$ is an arbitrary constant. The key observations are:

1. $\mathcal{D}(W_i)$ provides a good estimate of $|W_i|$, since $W_i \in [\beta^{i-1} \cdot \mathcal{D}(W_i), \beta^i \cdot \mathcal{D}(W_i))$.

2. The $\beta \cdot \epsilon$-effective support of $\mathcal{D}$ is contained in $\bigcup_{i \leq \ell} W_i$, where $\ell = O(\log(\mathtt{ess}_\epsilon(\mathcal{D})/\epsilon))$.

   This follows from $\sum_{i>\ell} \mathcal{D}(W_i) \leq \beta\epsilon$, which in turn follows by $\mathcal{D}(L) \leq \beta\epsilon$, where $L = \{e : \mathcal{D}(e) < (\beta-1) \cdot \epsilon/\mathtt{ess}_\epsilon(\mathcal{D})\}$. Specifically, as noted in the proof of Observation 1.4, the $n \stackrel{\text{def}}{=} \mathtt{ess}_\epsilon(\mathcal{D})$ heaviest elements of $\mathcal{D}$ are assigned probability mass at least $1-\epsilon$, and so $\mathcal{D}(L) \leq n \cdot \frac{(\beta-1)\cdot\epsilon}{n} + \epsilon = \beta\epsilon$ (see proof of Claim 2.1.1).

   Hence, if each element of $W_i$ has weight $\beta^{-(i-0.5)}$, then $\mathtt{ess}_{\beta\epsilon}(\mathcal{D})$ is determined by the values $\mathcal{D}(W_1), ..., \mathcal{D}(W_\ell)$. Using the first observation, we obtain an $[\beta^{-0.5}\epsilon, \beta^{0.5}\epsilon]$-effective support size also in general (i.e., when waiving the equal weights condition).

3. The foregoing $\ell$ can be replaced by $\widetilde{\ell}$, which can be found by sampling $O(1/\epsilon)$ elements; that is, we set $\widetilde{\ell}$ as the smallest integer such that at most $\beta^{1.1} \cdot \epsilon$ fraction of the elements in the sample reside in $\bigcup_{i>\widetilde{\ell}} W_i$.

   Note that $\sum_{i>\widetilde{\ell}} \mathcal{D}(W_i) \leq \beta^{1.2}\epsilon$, whereas $\sum_{i \geq \ell} \mathcal{D}(W_i) > \beta\epsilon$. Hence, $\widetilde{\ell} \leq \ell$.

Hence, approximating each of the $\mathcal{D}(W_i)$'s for $i \in [\ell]$ in the sense of obtaining a $\beta$-factor approximation for $\mathcal{D}(W_i) = \Omega(\epsilon/\ell)$ (and an indication that $\mathcal{D}(W_i) = o(\epsilon/\ell)$ otherwise) suffices for a $\beta^{O(1)}$-factor approximation of the $[\beta^{-O(1)}\epsilon, \beta^{O(1)} \cdot \epsilon]$-effective support size (of $\mathcal{D}$).[5] Using parameter-substitution, we get

**Theorem 1.11** (yet another approximator of the effective support size): *For any constant $\beta > 1$, there exists an algorithm that on input $\epsilon > 0$ and oracle access to $\mathcal{D}$, with probability at least $2/3$, uses $\widetilde{O}(\epsilon^{-1}\log(n/\epsilon))$ samples and outputs a $\beta$-factor approximation of the $[\epsilon, \beta \cdot \epsilon]$-effective support size of $\mathcal{D}$, where $n = \mathtt{ess}_\epsilon(\mathcal{D})$.*

Indeed, Theorem 1.11 is inferior to the last item of Theorem 1.9, but its proof is much simpler. Again, applying Observation 1.5 and relying on the fact that the sample complexity depends polynomially on $1/(\beta-1)$, we obtain an algorithm that uses $\mathrm{poly}(\epsilon^{-1} \cdot \log(n/\epsilon))$ samples and outputs an $[\epsilon, \beta \cdot \epsilon]$-effective support size of $\mathcal{D}$.

**Proving the first item of Theorem 1.9.** While the approximator underlying Theorem 1.11 does not establish any of the items of Theorem 1.9, a variant of it does establish the first item. First, observe that we can find $\widetilde{\ell}'$ such that $\sum_{i \geq \widetilde{\ell}'} \mathcal{D}(W_i) > \beta^{7/4}\epsilon$ and $\sum_{i \geq \widetilde{\ell}'+1} \mathcal{D}(W_i) < \beta^2\epsilon$. Hence, $\sum_{i \in [\widetilde{\ell}', \widetilde{\ell}]} \mathcal{D}(W_i) = \Omega(\epsilon)$, and, when $i \in [\widetilde{\ell}', \widetilde{\ell}]$ is selected with probability proportional to $\mathcal{D}(W_i)$, the value of $\mathcal{D}(W_i)$ is $\Omega(\epsilon/\ell)$ with high constant probability (e.g., at least 0.9). Note that we can select $i$ according to this distribution by sampling from $\mathcal{D}$ till the sampled element $e$ hits $\bigcup_{j \in [\widetilde{\ell}', \widetilde{\ell}]} W_j$, and set $i = \lfloor \log_\beta(1/\mathcal{D}(e)) \rfloor + 1$. Recalling that $\sum_{i \geq \widetilde{\ell}'+1} \mathcal{D}(W_i) < \beta^2\epsilon$, note that

$$\mathtt{ess}_{\beta^2\epsilon}(\mathcal{D}) \leq \sum_{j \in [\widetilde{\ell}']} |W_j| \leq \sum_{j \in [i]} |W_j| \leq \sum_{j \in [i]} \beta^j = O(\beta^i). \tag{1}$$

On the other hand, showing (see Claim 2.1.2 (2))[6] that $\mathtt{ess}_\epsilon(\mathcal{D}) = \Omega\left(\sum_{j \in [i]} |W_j|\right) = \Omega(|W_i|)$, and recalling that $\mathcal{D}(W_i) = \Omega(\epsilon/\ell)$ (with probability at least $2/3$ over the choice of $\widetilde{\ell}, \widetilde{\ell}'$ and $i$), it follows that

$$\mathtt{ess}_\epsilon(\mathcal{D}) = \Omega(|W_i|) = \Omega(\beta^{i-1} \cdot \epsilon/\ell). \tag{2}$$

Hence, with probability at least $2/3$, the value $O(\beta^i)$ constitutes an $O(\ell/\epsilon)$-factor approximation of the $[\epsilon, \beta^2 \cdot \epsilon]$-effective support size of $\mathcal{D}$. This establishes the first item of Theorem 1.9.

[5] Note that these $\ell$ approximations can be obtained by using a single sample of size $O((\ell/\epsilon) \cdot \log \ell)$.

[6] This is easy when $i < \ell$, but some difficulties arise in the case of $i = \ell$ (see proof of the main case of Part 2 of Claim 2.1.2).

**Towards proving the other items of Theorem 1.9.** Note that the foregoing procedure amounts to determining $\widetilde{\ell}$ and $\widetilde{\ell'}$, selecting $i \in [\widetilde{\ell'}, \widetilde{\ell}]$ by sampling $\mathcal{D}$ till hitting an element $e$ in $\bigcup_{j\in[\widetilde{\ell'},\widetilde{\ell}]} W_j$, and using $O(1/\mathcal{D}(e)) = O(\beta^i)$ as the approximation value. We stress that the foregoing procedure did not try to estimate $\mathcal{D}(W_i)$; it rather relied on the fact that $\sum_{j\in[i]} |W_j| = O(\beta^i)$ and that (with high constant probability) $\mathcal{D}(W_i) = \Omega(\epsilon/\ell)$, which in turn implies $\sum_{j\in[i]} |W_j| \geq |W_i| = \Omega(\beta^i \cdot \epsilon/\ell)$. The approximation factor is the ratio between these two bounds, and we seek to reduce this ratio so to obtain tighter approximators.

The first tighter approximator (i.e., the second item of Theorem 1.9) is based on the observation that if $\mathcal{D}(W_i) = o(\epsilon)$ then the lower-bound $\mathcal{D}(W_i) = \Omega(\epsilon/\ell)$ is off only by a factor of $O(\ell)$ rather than $O(\ell/\epsilon)$, whereas otherwise we can afford to approximate $\mathcal{D}(W_i)$ to within a constant factor (e.g., a $\beta$-factor). The same reasoning can be applied to each of the $W_j$ for $j \in [i - O(\log(1/\epsilon)), i-1]$, whereas $\sum_{j\in[i-O(\log(1/\epsilon))]} |W_j| = o(\epsilon \cdot \beta^i) = o(\ell \cdot |W_i|)$, provided that $\mathcal{D}(W_i) = \Omega(\epsilon/\ell)$. Hence, using $\widetilde{O}(1/\epsilon)$ samples, we can approximate each $\mathcal{D}(W_j)$ for $j \in [i - O(\log(1/\epsilon)), i]$ in the sense of obtaining a $\beta$-factor approximation for $\mathcal{D}(W_j) \geq \epsilon$ and an indication that $\mathcal{D}(W_j) < \epsilon$ otherwise. Denoting the corresponding approximations by $\widetilde{\delta_j}$'s, we essentially output $\sum_{j\in[i-O(\log(1/\epsilon),i]} \widetilde{\delta}_j \cdot \beta^j$. Recalling that $\mathtt{ess}_{\beta^2\epsilon}(\mathcal{D}) \leq \sum_{j\in[i]} |W_j|$ and $\mathtt{ess}_\epsilon(\mathcal{D}) = \Omega(|W_i|)$, we observe that our worst approximation factor (i.e., $O(\ell)$) is due to the light $W_j$'s (i.e., $\mathcal{D}(W_j) < \epsilon$) with $j \in [i-O(\log(1/\epsilon)), i]$. This suffices for establishing the second item of Theorem 1.9.

**Proving the remaining items of Theorem 1.9.** The remaining two approximators are based on a finer analysis of the interval $[\widetilde{\ell'}, \widetilde{\ell}]$, in which $i$ is selected. We start by observing that the approximation is actually off by a factor that is related to the length of the interval, $\lambda = \widetilde{\ell} - \widetilde{\ell'} + 1$; in the foregoing, we used $\ell$ merely for simplicity. Now, if $\lambda < 1/\epsilon$, then we can afford approximating all the relevant $\mathcal{D}(W_j)$'s that satisfy $\mathcal{D}(W_j) = \Omega(\epsilon/\lambda)$ up to a factor of $\beta$, since $O(\lambda/\epsilon) = O(1/\epsilon^2)$. Otherwise (i.e., $\lambda \geq 1/\epsilon$), letting $\lambda' = \Theta(\log \lambda) = \Theta(\log(\lambda/\epsilon))$ and recalling that $\sum_{j\in[\widetilde{\ell'},\widetilde{\ell'}+\lambda-1]} \mathcal{D}(W_j) = \Omega(\epsilon)$, we consider the following two cases:

- If $\sum_{j\in[\widetilde{\ell'},\widetilde{\ell'}+\lambda']} \mathcal{D}(W_j) = \Omega(\epsilon)$, then we can proceed with $\lambda'$ rather than with $\lambda$ (i.e., we recurse while resetting $\lambda \leftarrow \lambda'$).

  (In the current case, the fact that a relatively small portion of the interval $I = [\widetilde{\ell'}, \widetilde{\ell'} + \lambda - 1]$ holds much weight is used to gain a lot when proceeding to the recursive step (since $\lambda$ is replaced by $O(\log \lambda)$).)

  (In the next case, we use the fact that the rest of the interval is hit with high probability when selecting $i \in I$ with probability proportional to $\mathcal{D}(W_i)$. In that case, we can show that $\sum_{j\leq i-\lambda'} |W_j| = O(\epsilon\beta^i/\lambda)$, which means that $O(\epsilon\beta^i/\lambda)$ is a good approximator.)

- Otherwise $\sum_{i\in[\widetilde{\ell'}+\lambda',\widetilde{\ell'}+\lambda-1]} \mathcal{D}(W_i) = \Omega(\epsilon)$ holds, and we select $i \in [\widetilde{\ell'} + \lambda', \widetilde{\ell'} + \lambda - 1]$ with probability proportional to $\mathcal{D}(W_i)$. In this case, for the selected $i$ it holds that $\sum_{j\leq i-\lambda'} |W_j| = O(\epsilon\beta^i/\lambda)$, since $\beta^{-\lambda'} = O(\epsilon/\lambda)$. This implies that $\mathtt{ess}_{\beta^2\epsilon}(\mathcal{D}) = \sum_{j\leq i-\lambda'} |W_j| = O(\epsilon\beta^i/\lambda)$, since $\sum_{j>\widetilde{\ell'}} \mathcal{D}(W_j) < \beta^2\epsilon$ and $\widetilde{\ell'} < i - \lambda'$. On the other hand, with high constant probability, it holds that $|W_i| = \Omega(\epsilon\beta^i/\lambda)$, which implies $\mathtt{ess}_\epsilon(\mathcal{D}) = \Omega(\epsilon\beta^i/\lambda)$. Hence, in this case, $O(\epsilon\beta^i/\lambda)$ is a constant factor approximation (of the $[\epsilon, \beta^2 \cdot \epsilon]$-effective support size of $\mathcal{D}$).

Using recursion and a tighter analysis, we derive the fourth item of Theorem 1.9 (for the case of $k = 1$).[7] The third item (of Theorem 1.9) is established by truncating the recursion at depth $t-1$, and employing the algorithm of the second item (while recalling that here $\ell$ is replaced by $O(\log^{(t-1)} \ell)$).

[7] The case of general $k \in \mathbb{N}$ requires a finer analysis.

## 1.6 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]. (The approximators presented in [6, Sec. 2.2] are inferior to those stated in Theorem 1.9.)

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]).[8] 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).[9] 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.[10]

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 $\epsilon$ and tolerance parameter $\epsilon'$ calls for accepting distributions that have $\epsilon'$-effective support size at most $n$ (i.e., when $\mathtt{ess}_{\epsilon'}(\mathcal{D}) \leq n$) and rejecting distributions that have $\epsilon$-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 $\epsilon$ calls for accepting distributions that have support size at most $n$ and rejecting distributions that have $\epsilon$-effective support size greater than $n$. (Note that an $\epsilon'$-tolerant $\epsilon$-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 $n \in [\mathtt{ess}_{\epsilon}(\mathcal{D}), \mathtt{ess}_{\epsilon'}(\mathcal{D})]$.)[11]

## 1.7 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 $\mathcal{D} : U \to [0,1]$, 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 let $\overline{S} = U \setminus S$ and often do not specify $U$ at all.

One may assume, without loss of generality, that an approximator of the effective support size makes evaluation queries only on elements that appear in the sample. Our algorithms do satisfy this

[8] 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).

[9] The latter problem sounds related to approximating the effective support size, but is actually different from it (see next paragraph).

[10] 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 un-sampled elements are practically useless.

[11] 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, in the current work, we care about lower complexities).

assumption, and hence we only state their sample complexity, which upper-bounds their evaluation-query complexity.

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

$$\frac{1}{2} \cdot \sum_e \left|\mathcal{D}(e) - \mathcal{D}'(e)\right| \;=\; \max_S\{\mathcal{D}(S) - \mathcal{D}'(S)\}. \tag{3}$$

Otherwise, we say that $\mathcal{D}$ is $\epsilon$-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.

**Theorem 2.1** (the basic algorithm): *For any constant $\beta > 1$, there exists an algorithm that on input $\epsilon > 0$ and oracle access to $\mathcal{D}$, uses $O(1/\epsilon)$ samples and outputs an $O(\epsilon^{-1}\log(n/\epsilon))$-factor approximation of the $[\epsilon, \beta\cdot\epsilon]$-effective support size of $\mathcal{D}$, where $n = \mathtt{ess}_\epsilon(\mathcal{D})$. The dependence of the number of samples on $\beta$ is* $\mathrm{poly}(1/(\beta-1))$*. Ditto for the approximation factor.*

**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 $O(\log(n/\epsilon))$ sets (i.e., $W_i$'s), where $n = \mathtt{ess}_\epsilon(\mathcal{D})$ is the $\epsilon$-effective support size of $\mathcal{D}$.

**Claim 2.1.1** ($\mathtt{ess}_\epsilon(\mathcal{D})$ and the $\mathcal{D}(W_i)$'s): *Suppose that $\mathcal{D}$ has $\epsilon$-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 the rest of this proof (as well as the in the proofs of the subsequent theorems), we shall assume that $\ell > 1$, while noting that the case of $\ell = 1$ is easily handled (by finding all elements of $W_1$ and outputting $|W_1|$). In fact, for similar reasons, we may assume that $\ell > \log_\beta(1/\epsilon) + O(1)$. (In contrast, if $\ell \leq \log_\beta(1/\epsilon)+O(1)$, then each element of $\bigcup_{i\in[\ell]} W_i$ has weight at least $\beta^{-\ell} = \Omega(\epsilon)$, whereas we can afford finding all element of weight $\Omega(\epsilon)$.)[12]

Proof: Let $S$ be a set of size at most $n$ such that there exists a distribution that is $\epsilon$-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

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

[12] With high probability, a sample of size $O(\epsilon^{-1}\log(1/\epsilon))$, contains all elements of weight $\Omega(\epsilon)$. Hence, if $\mathtt{ess}_\epsilon(\mathcal{D}) = O(1/\epsilon)$, then $\ell \leq \log_\beta(1/\epsilon) + O(1)$, and we can determine the exact $\epsilon$-effective support size of $\mathcal{D}$ by querying $\mathtt{eval}_\mathcal{D}$ on each element in this sample.

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 $\mathcal{D}(e) \leq \beta^{-(i-1)} \leq \beta^{-k} = (\beta-1)\cdot\epsilon/n$). ■

Important thresholds ($\ell$ and $\ell'$). For $\epsilon' = \beta \cdot \epsilon$, let $\ell' \in \mathbb{N}$ be maximal such that $\sum_{i \geq \ell'} \mathcal{D}(W_i) \geq \beta \cdot \epsilon'$. Recalling that $\ell \in \mathbb{N}$ (is minimal that) satisfies $\sum_{i>\ell} \mathcal{D}(W_i) \leq \epsilon'$, we get

$$\Delta \stackrel{\text{def}}{=} \sum_{i \in [\ell', \ell]} \mathcal{D}(W_i) = \sum_{i \geq \ell'} \mathcal{D}(W_i) - \sum_{i > \ell} \mathcal{D}(W_i) \geq (\beta-1)\cdot\epsilon'. \tag{4}$$

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 $\Pr_{i \sim \mathcal{D}(W_i)}\left[i \in B | i \in [\ell', \ell]\right]$ 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

$$|W_i| \geq \mathcal{D}(W_i)/\beta^{-(i-1)} \geq \frac{(\beta-1)\cdot\epsilon'}{3\ell} \cdot \beta^{i-1} = (\beta-1)\cdot\epsilon\cdot\beta^i/3\ell. \tag{5}$$

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

$$\sum_{j \leq i} |W_j| < \frac{\beta^{i+1}}{\beta-1} \leq f \cdot \sum_{j \leq i} |W_j|. \tag{6}$$

Hence, $v \stackrel{\text{def}}{=} \beta^{i+1}/(\beta-1)$ provides an $f$-factor approximation of $\sum_{j \leq i} |W_j|$. Next (in Claim 2.1.2) we relate $\sum_{j \leq i} |W_j|$ to the $[\epsilon, \beta^2 \cdot \epsilon]$-effective support size of $\mathcal{D}$, by showing that $\mathtt{ess}_{\beta^2\epsilon}(\mathcal{D}) \leq \sum_{j \leq i} |W_j|$ and $\mathtt{ess}_\epsilon(\mathcal{D}) > (\beta-1)\cdot\beta^{-2}\cdot\sum_{j \leq i} |W_j|$ (with $\mathtt{ess}_{\beta\epsilon}(\mathcal{D}) \geq \sum_{j \leq i} |W_j|$ when $i < \ell$). It follows that $v$ constitutes an $(\beta^2/(\beta-1))\cdot f$-factor approximation of the $[\epsilon, \beta^2\cdot\epsilon]$-effective support size of $\mathcal{D}$.

**Claim 2.1.2** (effective support size vs $\sum_{j \leq i} |W_j|$): *Let $\ell$ and $\ell'$ 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. $\mathtt{ess}_\epsilon(\mathcal{D}) \geq \frac{\beta-1}{\beta^2} \cdot \sum_{j \leq i} |W_j|$. *Furthermore, if $i < \ell$, then* $\mathtt{ess}_\epsilon(\mathcal{D}) \geq \mathtt{ess}_{\beta\epsilon}(\mathcal{D}) \geq \sum_{j \leq i} |W_j|$.

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 $\bigcup_{j \leq i} W_j$. Using $\sum_{j>i} \mathcal{D}(W_j) \leq \sum_{j \geq \ell'+1} \mathcal{D}(W_j) < \beta^2\epsilon$, where the second inequality is due to the maximality of $\ell'$, it follows $\mathcal{D}'$ is $\beta^2\epsilon$-close to $\mathcal{D}$. Hence, 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, $\sum_{j>i} \mathcal{D}(W_j) \geq \sum_{j > \ell-1} \mathcal{D}(W_j) > \beta\epsilon$, where the second inequality is due to the minimality of $\ell$. Using Observation 1.4, $\sum_{j>i} \mathcal{D}(W_j) \geq \beta\epsilon$ 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., $\mathtt{ess}_{\beta\epsilon}(\mathcal{D}) \geq \sum_{j \leq i} |W_j|$).

Moving 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 $H \stackrel{\text{def}}{=} \bigcup_{j \in [\ell-1]} W_j$ and $L \stackrel{\text{def}}{=} \bigcup_{j > \ell} W_j$. Recalling that $\mathcal{D}(L \cup W_\ell) > \beta \cdot \epsilon$, we let $L'$ be a maximal set of (the lightest)[13] elements of $W_\ell$ such that $\mathcal{D}(L \cup L') \leq \epsilon$,

[13] By maximality (of size), the set $L'$ must contain the lightest elements of $W_\ell$ that satisfy the condition $\mathcal{D}(L \cup L') \leq \epsilon$. Note that $L'$ may be empty (e.g., if $\mathcal{D}(L) > \epsilon$); in this case, $\mathtt{ess}_\epsilon(\mathcal{D}) \geq |H \cup W_\ell|$ follows.

and observe that $\mathtt{ess}_\epsilon(\mathcal{D}) \geq |H \cup (W_\ell \setminus L')|$ (by the maximiality of $L'$), and that $\mathcal{D}(L) + \mathcal{D}(W_\ell) > \beta \cdot (\mathcal{D}(L) + \mathcal{D}(L'))$. Hence, $\mathcal{D}(W_\ell) > \beta \cdot \mathcal{D}(L')$, which implies $\mathcal{D}(W_\ell \setminus L') > (1-\beta^{-1}) \cdot \mathcal{D}(W_\ell)$. Noting that the elements in $W_\ell \setminus L'$ may be at most a factor of $\beta$ heavier than those in $L'$, it follows that $|W_\ell \setminus L'| > \beta^{-1} \cdot (1-\beta^{-1}) \cdot |W_\ell|$, we have

$$\mathtt{ess}_\epsilon(\mathcal{D}) \geq |H \cup (W_\ell \setminus L')| > |H| + \frac{1-\beta^{-1}}{\beta} \cdot |W_\ell|,$$

and the main claim of the second part follows. This completes the proof of the entire claim. ■

Using approximated thresholds. The foregoing presentation is idealized, since in reality we do not know $\ell'$ and $\ell$. 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 $\mathcal{D}(W_i) \geq \frac{\widetilde{\Delta}}{3\widetilde{\ell}} \geq \frac{(\beta-1)\cdot\epsilon}{6\ell}$. In this case $|W_i| \geq \frac{(\beta-1)\cdot\epsilon}{6\ell} \cdot \beta^{i-1}$ follows. Let us spell out the resulting algorithm.

**Algorithm 2.1.3** (the actual algorithm): *For fixed $\beta > 1$, on input $\epsilon > 0$ and oracle access to $\mathcal{D}$, the algorithm proceeds as follows.*

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 $\mathcal{D}(W_i)$. Recall that this can be done by rejection sampling and has complexity $O(1/\epsilon)$.*

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

Note that the operations of Algorithm 2.1.3 amount to taking samples of $\mathcal{D}$ (by invoking $\mathtt{samp}_\mathcal{D}$), evaluating their probability mass (by calling $\mathtt{eval}_\mathcal{D}$), and doing some simple manipulations.

Re-iterating the argument that was used in case $\ell$ and $\ell'$ were known, we sandwich the algorithm's output (i.e., $\beta^{i+1}/(\beta-1)$) between $\mathtt{ess}_{\beta^2\epsilon}(\mathcal{D})$ and a multiple of $\mathtt{ess}_\epsilon(\mathcal{D})$: On the one hand, as before, using Part 1 of Claim 2.1.2 (along with the l.h.s of Eq. (6)), it follows that $\mathtt{ess}_{\beta^2\epsilon}(\mathcal{D}) \leq \sum_{j \leq i} |W_j| \leq \beta^{i+1}/(\beta-1)$. On the other hand, using Part 2 of Claim 2.1.2 along with $\sum_{j \leq i} |W_j| \geq |W_i| \geq \frac{(\beta-1)\cdot\epsilon}{6\ell} \cdot \beta^{i-1}$ (which is a factor of 2 smaller than before), it follows that $\mathtt{ess}_\epsilon(\mathcal{D}) \geq \frac{\beta-1}{\beta^2} \cdot \sum_{j \leq i} |W_j| \geq \frac{(\beta-1)^2\cdot\epsilon}{6\ell} \cdot \beta^{i-3}$. Thus,

$$\mathtt{ess}_{\beta^2\epsilon}(\mathcal{D}) \leq \frac{\beta^{i+1}}{\beta-1} \leq \widetilde{f} \cdot \mathtt{ess}_\epsilon(\mathcal{D}),$$

where $\widetilde{f} = \frac{\beta^{i+1}/(\beta-1)}{(\beta-1)^2\cdot\epsilon\cdot\beta^{i-3}/6\ell} < \frac{6\cdot\beta^4}{(\beta-1)^3} \cdot \ell/\epsilon$, which means that Algorithm 2.1.3 is a $\widetilde{f}$-factor approximator of the $[\epsilon, \beta^2 \cdot \epsilon]$-effective support size of $\mathcal{D}$. Recalling that $\ell \leq \log_\beta(n/(\beta-1)\cdot\epsilon) = O((\beta-1)^{-1} \cdot \log(n/\epsilon))$, where $n = \mathtt{ess}_\epsilon(\mathcal{D})$ is the $\epsilon$-effective support size of $\mathcal{D}$, the theorem follows (by a change of parameters).[14] ■

[14] Specifically, given the parameters $\beta$ and $\epsilon$, we reset $\beta \leftarrow \beta^{1/2}$, and obtain an $O(\epsilon^{-1}\log n)$-factor approximator of the $[\epsilon, \beta\cdot\epsilon]$-effective support of $\mathcal{D}$. Note that $\beta^{1/2}-1 = (\beta-1)/(\beta^{1/2}+1) = \Omega(\beta-1)$, which implies that $\mathrm{poly}(1/(\beta^{1/2}-1)) = \mathrm{poly}(1/(\beta-1))$.

**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 $\epsilon$, 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 $\epsilon$-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 $\beta^i$.

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 $\beta$-factor, and this yields an approximation of $|W_i|$ up to a $\beta^2$-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 (i.e., $\mathcal{D}(W_i) = o(\epsilon)$) 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 $\Theta(\epsilon^{-1}\log(n/\epsilon))$ to $\Theta(\log(n/\epsilon))$, while slightly increasing the sample complexity (so to allow for obtaining $\Theta(\log(1/\epsilon))$ approximate values rather than a constant number of such values). Note that we can afford to ignore the contribution of $\sum_{j<i-\log_\beta(1/\epsilon)} |W_j|$, since it is at most $\epsilon \cdot \beta^i/(\beta-1)$.

**Theorem 2.2** (the basic algorithm, revised): *For any constant $\beta > 1$, there exists an algorithm that on input $\epsilon > 0$ and oracle access to $\mathcal{D}$, uses $\widetilde{O}(1/\epsilon)$ samples and outputs an $O(\log(n/\epsilon))$-factor approximator of the $[\epsilon, \beta\cdot\epsilon]$-effective support size of $\mathcal{D}$, where $n = \mathtt{ess}_\epsilon(\mathcal{D})$. The dependence of the number of samples on $\beta$ is* $\mathrm{poly}(1/(\beta-1))$. *Ditto for the approximation factor.*

**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 $\sum_{j\le i} \mathcal{D}(W_j)\cdot\beta^j$. (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 $1 - 0.1/\log_\beta(1/\epsilon)$)) *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.*[15] *Recall that these estimates can be obtained using a sample of size $O(\epsilon^{-1}\log\log(1/\epsilon))$.*
- *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 $\frac{\beta^{i'}}{\beta-1} + \sum_{j\in[i',i]} \widetilde{w}_j$.*

We shall show that, with high probability, the foregoing output lies in the interval $[\mathtt{ess}_{\beta^2\epsilon}(\mathcal{D}), O(\ell)\cdot \mathtt{ess}_\epsilon(\mathcal{D})]$, where $\ell = O(\log(\mathtt{ess}_\epsilon(\mathcal{D})/\epsilon))$. As in the proof of Theorem 2.1, the analysis of Algorithm 2.2.1 focus on the case that *the selected $i$ satisfies* $\mathcal{D}(W_i) \ge \frac{(\beta-1)\cdot\epsilon}{6\ell}$. But here we consider two sub-cases.

1. If $\mathcal{D}(W_i) > \epsilon/\beta^2$, then, with high probability, it holds that $\mathcal{D}(W_i) \le \widetilde{\delta}_i \le \beta\cdot\mathcal{D}(W_i)$, and $|W_i| \le \widetilde{w}_i \le \beta^2\cdot|W_i|$ follows.

[15] 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 $\epsilon/\beta^{1.5}$.

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

$$|W_i| \leq \widetilde{w}_i \leq \frac{6\ell \cdot \beta^2}{\beta-1} \cdot |W_i| \tag{7}$$

holds (where we use $6\ell \geq \beta-1$).[16] 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,

$$|W_j| < \widetilde{w}_j \leq \max(\beta^2 \cdot |W_j|, \epsilon \cdot \beta^j). \tag{8}$$

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

$$\sum_{j \leq i} |W_j| < \frac{\beta^{i'}}{\beta-1} + \sum_{j \in [i',i]} \widetilde{w}_j, \tag{9}$$

where we use $\sum_{j<i'} |W_j| \leq \sum_{j<i'} \beta^j < \beta^{i'-1} \cdot \beta/(\beta-1)$ as well as $|W_j| \leq \widetilde{w}_j$ for every $j \in [i', i]$. On the other hand, using $\mathcal{D}(W_i) \geq \frac{(\beta-1)\cdot\epsilon}{6\ell}$, which implies $|W_i| \geq \frac{(\beta-1)\cdot\epsilon}{6\ell} \cdot \beta^{i-1}$, we shall upper-bound the output value by $O(\ell) \cdot \sum_{j \leq i} |W_j|$. Intuitively, the lower bound $|W_i| = \Omega(\epsilon/\ell) \cdot \beta^i$ implies that $\sum_{j \leq i} \epsilon \cdot \beta^j = O(\ell) \cdot |W_i|$, which takes care of the $j$'s with $\mathcal{D}(W_j) < \epsilon/\beta^2$, whereas the other $W_j$'s are approximated quite well. In fact, foreseeing subsequent applications, we prove a more general statement (which refers to auxiliary parameters $\eta$ and $\epsilon'$, where we use $\eta = \frac{(\beta-1)\cdot\epsilon}{6\ell}$ and $\epsilon' = \epsilon$).

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

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

In our application $\eta = \frac{(\beta-1)\cdot\epsilon}{6\ell}$ and $\epsilon' = \epsilon$; so $i' = \max(1, i - \log_\beta(1/\epsilon))$ and $\left(1 + \frac{(\beta+1)\cdot\epsilon'}{(\beta-1)\cdot\eta}\right) \cdot \beta^2 = O(\ell)$. Note that the lower bound on $|W_i|$ holds by $|W_i| \geq \frac{(\beta-1)\cdot\epsilon}{6\ell} \cdot \beta^{i-1}$ and the r.h.s of Eq. (7), whereas the lower bound on $\widetilde{w}_j$ holds by Eq. (8).

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

$$\widetilde{w}_j \leq \beta^2 \cdot |W_j| + \frac{\epsilon' \cdot \beta^{j-i+1}}{\eta} \cdot |W_i|. \tag{10}$$

[16] In this case, $\beta^2 \leq \frac{6\ell \cdot \beta^2}{\beta-1}$. Note that $6\ell \geq \beta - 1$ holds if either $\ell = \omega(1)$ or $\beta < 7$, and each of these can be assumed without loss of generality.

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

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

Recalling that $\widetilde{w}_i \leq \frac{\beta^2\epsilon'}{\eta} \cdot |W_i|$, we get

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

The claim follow. ■

Conclusion. 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 $\sum_{j \leq i} |W_j|$ and $O(\ell) \cdot \sum_{j \leq i} |W_j|$ (see Eq. (9) and Claim 2.2.2). Proceeding as in the proof of Theorem 2.1, we infer that Algorithm 2.2.1 constitutes an $O(\ell)$-factor approximator of the $[\epsilon, \beta \cdot \epsilon]$-effective support size of $\mathcal{D}$. ■

**Reducing the factor that depends on the effective support size.** Recall that the approximation factor in Theorem 2.1 is the product of a factor of $O(1/\epsilon)$ and a factor of $O(\ell)$, where $\ell = O(\log(\mathtt{ess}_\epsilon(\mathcal{D})/\epsilon))$. In Theorem 2.2 we focused on eliminating the first factor, whereas here we shall focus on reducing the second factor (from $O(\ell)$) to $O(\log \ell)$. Needless to say, we shall actually combine both strategies, but for sake of presenting the new idea we ignore the improvement already obtained in Theorem 2.2 and the idea that underlined it. Instead, we go back to Algorithm 2.1.3 (i.e., the algorithm underlying Theorem 2.1).

Recall that the proof of Theorem 2.1 focused on the case that $\mathcal{D}(W_i) = \Omega(\epsilon/(\widetilde{\ell} - \widetilde{\ell}' + 1))$, where $i \in [\widetilde{\ell}', \widetilde{\ell}]$, and relied on the fact that $\mathtt{ess}_{\beta^2\epsilon}(\mathcal{D}) \leq \sum_{j \leq \widetilde{\ell}'} |W_j| = O(\beta^{\widetilde{\ell}'})$ and $\mathtt{ess}_\epsilon(\mathcal{D}) \geq \sum_{j \leq i} \mathcal{D}(W_j) \geq |W_i| = \Omega(\beta^i \cdot \mathcal{D}(W_i))$. Indeed, the original proof used $\lambda \stackrel{\text{def}}{=} \widetilde{\ell} - \widetilde{\ell}' + 1 \leq \ell$ and $\widetilde{\ell}' \leq i$, and yielded an approximation ratio that is upper-bounded by the ratio of the upper bound for $\mathtt{ess}_{\beta^2\epsilon}(\mathcal{D})$ over the lower bound for $\mathtt{ess}_\epsilon(\mathcal{D})$; that is, the proof uses

$$\frac{\mathtt{ess}_{\beta^2\epsilon}(\mathcal{D})}{\mathtt{ess}_\epsilon(\mathcal{D})} \leq \frac{O(\beta^{\widetilde{\ell}'})}{\Omega(\beta^i \cdot \mathcal{D}(W_i))} \leq \frac{O(\beta^i)}{\Omega(\beta^i \cdot \epsilon/\lambda)} \tag{11}$$

Note that if $\widetilde{\ell}' < i - \log_\beta(\lambda/\epsilon)$, then the middle fraction in Eq. (11) is upper-bounded by a constant, since $\mathcal{D}(W_i) = \Omega(\epsilon/\lambda)$. Otherwise (i.e., $i \leq \widetilde{\ell}' + \log_\beta(\lambda/\epsilon)$), assuming that this case is quite likely

when $i$ is chosen in proportion to $\mathcal{D}(W_i)$, we get $\sum_{j\in[\widetilde{\ell}',\widetilde{\ell}'+\log_\beta(\lambda/\epsilon)]}\mathcal{D}(W_j)=\Omega(\epsilon)$ and so $\mathcal{D}(W_i)=\Omega(\epsilon/\log_\beta(\lambda/\epsilon))$ is likely, which implies that the middle fraction in Eq. (11) can be upper-bounded by $O(\epsilon^{-1}\cdot\log(\lambda/\epsilon))=\widetilde{O}(1/\epsilon)+O(\epsilon^{-1}\log\lambda)$ rather than by $O(\epsilon^{-1}\cdot\lambda)$. Iterating this reasoning for $O(1)$ times, where in each iteration the current $\lambda$ is replaced by $\log_\beta(\lambda/\epsilon)$, we can get an upper bound of $\widetilde{O}(1/\epsilon)+O(\epsilon^{-1}\log^{(O(1))}\lambda)$. Combining this strategy with the strategy used in the proof of Theorem 2.2, while dealing differently with the case that the current $\lambda$ is smaller than $1/\epsilon$, we can get an upper bound of $O(\log^{(O(1))}\lambda)$. This suggests the following result, where the $\widetilde{O}(\epsilon^{-(k+1)/k})$ (rather than $\widetilde{O}(1/\epsilon)$) bound is due to dealing with small values of $\lambda$ (and the case of $k>1$ requires an additional twist).

**Theorem 2.3** (the iterative algorithm): *For any constants $\beta>1$ and $t,k\in\mathbb{N}$, there exists an algorithm that on input $\epsilon>0$ and oracle access to $\mathcal{D}$, uses $\widetilde{O}(t/\epsilon^{1+\frac{1}{k}})$ samples 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 $\log^{(j+1)}m=\log(\log^{(j)}m)$). *The dependence of the number of samples on $\beta$ is* $\mathrm{poly}(1/(\beta-1))$. *Ditto for the approximation factor.*

(The dependency of the sample complexity on the constant $t$ is spelled out in foreseeing Theorem 2.4.)

**Proof:** The case of $t=1$ was established in Theorem 2.2, which actually states a stronger complexity 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.

The first case is of small $\lambda\stackrel{\text{def}}{=}\widetilde{\ell}-\widetilde{\ell}'+1$ (i.e., $\lambda\leq 1/\epsilon$). In this case we cannot take the strategy of the main two cases, but we can afford dealing with it directly, while using the fact that $\ell/\epsilon=O(1/\epsilon^2)$. The main two cases differ according to the probability mass assigned to the intervals $[\widetilde{\ell}',\widetilde{\ell}'+2\log_\beta\lambda]$ and $[\widetilde{\ell}'+2\log_\beta\lambda+1,\widetilde{\ell}]$. If the first interval is assigned mass $\Omega(\epsilon)$, then we use it analogously to the use of $[\widetilde{\ell}',\widetilde{\ell}]$ in the proof of Theorem 2.2, while gaining from the fact that $\ell$ is replaced by $O(\log\ell)$. This corresponds to the following Case 3, which yields an $O(\log\ell)$-factor approximation (rather than an $O(\ell)$-factor approximation) of the $[\epsilon,\beta\cdot\epsilon]$-effective support size of $\mathcal{D}$. In contrast, Case 2 corresponds to the case that the interval $I=[\widetilde{\ell}'+2\log_\beta\lambda+1,\widetilde{\ell}]$ is assigned mass $\Omega(\epsilon)$. In this case, we select $i\in I$ with probability proportional to $\mathcal{D}(W_i)$, and output $\frac{\epsilon\cdot\beta^{i+1}}{(\beta-1)\cdot\lambda}$, which (w.h.p.) is $O(|W_i|)$, whereas $\mathtt{ess}_\epsilon=\Omega(|W_i|)$. The punchline is that

$$\mathtt{ess}_{\beta^2\epsilon}(\mathcal{D})<\sum_{j\leq i-2\log_\beta\lambda}|W_j|<\sum_{j\leq i-2\log_\beta\lambda}\beta^j<\frac{\beta^{i-2\log_\beta\lambda+1}}{(\beta-1)}\leq(\epsilon/\lambda)\cdot\frac{\beta^{i+1}}{(\beta-1)}$$

where the first inequality uses $i>\ell'+2\log_\beta\lambda$ and the last inequality uses $2\log_\beta\lambda\geq\log_\beta(\lambda/\epsilon)$, which in turn relies on Case 1 not occurring (i.e., $\lambda>1/\epsilon$).[17] Hence, the output is sandwiched between $\mathtt{ess}_{\beta^2\epsilon}(\mathcal{D})$ and $O(\mathtt{ess}_\epsilon(\mathcal{D}))$. Details follow.

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 $\widetilde{\delta}_j<\epsilon'/\beta$ to $\widetilde{\delta}_j\leftarrow\epsilon'$, and output $\frac{\beta^{i'}}{\beta-1}+\sum_{j\in[i',i]}\widetilde{w}_j$. (Recall that $\widetilde{w}_j=\widetilde{\delta}_j\cdot\beta^j$.)

[17] The gain (in Case 2) is due to the fact that $i>\widetilde{\ell}'+2\log_\beta\lambda$, which implies $i-\log_\beta(\lambda/\epsilon)>\ell'$.

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). Hence, using Claim 2.2.2 (with $\epsilon'$ as set here (i.e., $\epsilon' = \beta^{-2} \cdot \epsilon/\lambda$) and $\eta = (\beta - 1) \cdot \epsilon/6\lambda$, which implies that $\epsilon'/\eta = O(1)$)[18], we obtain an $O(1)$-factor approximation of the $[\epsilon, \beta^2 \cdot \epsilon]$-effective support size of $\mathcal{D}$.

The following two (main) cases deal with the situation in which $\lambda > 1/\epsilon$, where we want to avoid using $\widetilde{O}(\lambda/\epsilon)$ samples. In these cases, $2 \log_\beta \lambda < \lambda$.

(Recall that $\lambda = \widetilde{\ell} - \widetilde{\ell'} + 1$ and that $\widetilde{\Delta} = \sum_{j \in [\widetilde{\ell'}, \widetilde{\ell}]} \mathcal{D}(W_j) > (\beta - 1) \cdot \epsilon/2$.)

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, $\mathtt{ess}_{\beta^2 \cdot \epsilon}(\mathcal{D}) \leq \epsilon \cdot \beta^i/((\beta - 1) \cdot \lambda)$, because $i - 2 \log_\beta \lambda > \ell'$ and so $\sum_{j > i - 2 \log_\beta \lambda} \mathcal{D}(W_j) < \sum_{j > \ell'} \mathcal{D}(W_j) < \beta^2 \cdot \epsilon$, 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, $\mathtt{ess}_\epsilon(\mathcal{D}) = \Omega(|W_i|)$, whereas $|W_i| \geq \beta^{i-1} \cdot \mathcal{D}(W_i) > \frac{(\beta - 1) \cdot \epsilon}{200\lambda} \cdot \beta^{i-1}$. Hence, $\frac{\epsilon \cdot \beta^i/((\beta - 1) \cdot \lambda)}{\mathtt{ess}_\epsilon(\mathcal{D})} = O(1)$.

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, we obtain an $O(\log \ell)$-factor approximation (rather than an $O(\ell)$-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 take $\widetilde{O}(1/\epsilon^2)$ samples and obtain an $O(\log \log(\mathtt{ess}_\epsilon(\mathcal{D})/\epsilon))$-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.)

Let us recap. The key distinction is between Case 2 and Case 3. In Case 2 we select $i > \widetilde{\ell'} + 2 \log_\beta \lambda$ and are guaranteed that $\sum_{j > i - 2 \log_\beta \lambda} \mathcal{D}(W_j) < \beta^2 \cdot \epsilon$ and $\sum_{j \leq i - 2 \log_\beta \lambda} |W_j| < (\epsilon/\lambda) \cdot \beta^i/(\beta - 1)$, which (when combined) implies that $\mathtt{ess}_{\beta^2 \cdot \epsilon}(\mathcal{D}) \leq \epsilon \cdot \beta^i/((\beta - 1) \cdot \lambda)$, whereas $\mathtt{ess}_\epsilon(\mathcal{D}) = \Omega(\epsilon \cdot \beta^i/\lambda)$. In Case 3 we invoked the approximator of Algorithm 2.2.1 on an interval of length $O(\log \lambda)$ rather that length $\lambda$. When generalizing this strategy to the case of $t > 2$, Case 3 will generate a recursive call to the current procedure (which invokes Algorithm 2.2.1, if at all, only at recursion depth $t - 1$). Details follow.

More accurate approximations (i.e., $t > 2$). For $t > 2$, we proceed almost exactly in the same manner, with the following three exceptions: First, as stated above, in Case 3 we recursively invoke the current procedure with $t \leftarrow t - 1$ and $\widetilde{\Delta} \leftarrow (1 - (0.1/t)) \cdot \widetilde{\Delta}$ (rather than invoking Algorithm 2.2.1).[19] Second, the threshold for distinguishing Case 2 from Case 3 is set to equal $(1 - 0.1/t) \cdot \widetilde{\Delta}$ rather than $0.9 \cdot \widetilde{\Delta}$ (so

[18] Note that the original argument implies that $\mathcal{D}(W_i) \geq (\beta - 1) \cdot \Delta/6(\widetilde{\ell} - \widetilde{\ell'} + 1)$ (rather than $\mathcal{D}(W_i) \geq (\beta - 1) \cdot \Delta/6\ell$). (In Algorithm 2.3.1 we shall use a slightly different setting of $\epsilon'$.)

[19] Actually, when reaching the third case with $t = 2$, the recursive call (which uses $t = 1$) will actually invoke Algorithm 2.2.1.

to increase the probability mass in the last invocation).[20] 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 (i.e., Theorem 2.4).) 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-bound $\Delta'$* (initially set to $\widetilde{\Delta}$)*, and the remaining interval length $\lambda$* (initially set to $\widetilde{\ell} - \widetilde{\ell}' + 1$).[21] *If $t' = 1$, then we proceed as in Algorithm 2.2.1, and otherwise we proceed as follows, according to three cases, 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 $i' = \max(1, i - \log_\beta(6(\beta+1)\lambda/(\beta-1)^3\epsilon))$, and reset $\widetilde{\delta}_j$ to $\epsilon' \stackrel{\text{def}}{=} \frac{(\beta-1)^3\cdot\epsilon}{30(\beta+1)\lambda}$ if $\widetilde{\delta}_j < \epsilon'/\beta$. Hence, we output*

$$\frac{\beta^{i'}}{\beta-1} + \sum_{j\in[i',i]} \widetilde{w}_j \tag{12}$$

   *where $i \in [\widetilde{\ell}', \widetilde{\ell}' + \lambda - 1]$ and the $\widetilde{w}_j$'s are determined as in Algorithm 2.2.1* (i.e., $\widetilde{w}_j = \widetilde{\delta}_j \cdot \beta^j$)*, except that $\epsilon' = \Theta(\epsilon/\lambda)$* (rather than $\epsilon' = \epsilon$)

   (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, where $\epsilon' = O(\epsilon/\lambda)$ rather than $\epsilon' = \epsilon$.)[22]

   *We note 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 all 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*

$$\frac{(\beta-1)\cdot\epsilon}{300t\lambda}\cdot\beta^i \tag{13}$$

   *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-bound to $(1 - (0.1/t))\cdot\Delta'$, and the remaining interval length to $3\log_\beta\lambda$; that is, $t' \leftarrow t'-1$, $\Delta' \leftarrow (1-(0.1/t))\cdot\Delta'$, and $\lambda \leftarrow 3\log_\beta\lambda$.*

   (Note that $2\log_\beta\lambda + 1 < 3\log_\beta\lambda < \lambda$.)[23]

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

[20] This setting guarantees that, at each iteration, 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 $0.9^t = \exp(-t)$.

[21] Hence, $\sum_{j\in[\widetilde{\ell}', \widetilde{\ell}'+\lambda-1]} \mathcal{D}(W_j) \geq \Delta'$ holds initially (as well as in the recursive invocations).

[22] Recall that in case $\widetilde{\delta}_j < \epsilon'/\beta$, we reset $\widetilde{\delta}_j$ to $\epsilon'$.

[23] Both inequalities use $\lambda \geq c/\epsilon > 300/(\beta-1)^2$, 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.

**Claim 2.3.2** (analysis of Cases 1 and 2 of Algorithm 2.3.1): *Suppose that Algorithm 2.3.1 is invoked with fixed parameters $t$ and $\widetilde{\ell}'$, and uses the initial values $\widetilde{\ell}$ and $\widetilde{\Delta}$. Suppose that either Case 1 or Case 2 holds when the algorithm reached recursion depth $t - t'$ such that $t' > 1$. Then, the algorithm outputs an $\gamma_{i,\ell} \cdot \beta^4$-factor approximation of the $[\epsilon, \beta^2\epsilon]$-effective support size of $\mathcal{D}$, where $i$ is as selected in the corresponding step* (resp., case) *and $\gamma_{i,\ell} = 1$ if $i < \ell$ and $\gamma_{i,\ell} = \beta^2/(\beta-1)$ otherwise.*

Indeed, for the sake of the current proof (of Theorem 2.3), a constant upper bound on the approximation factor is more than enough, since Case 3 incurs a larger factor anyhow. However, we shall be using Claim 2.3.2 in the proof of Theorem 2.4, where the tighter bound will be useful.

Proof: In both cases, we refer to the current values of $t', \lambda$, and $\Delta' > 0.8 \cdot \widetilde{\Delta} > 0.4(\beta-1) \cdot \epsilon$. In each case, the index $i$ is selected (in proportion to $\mathcal{D}(W_i)$) in a designated interval, denoted $I$, and we assume that $\mathcal{D}(W_i)$ is at least $0.1 \cdot \sum_{j \in I} \mathcal{D}(W_j)/|I|$; indeed, this assumption holds with probability at least 0.9.

When Case 1 holds we adapt the analysis provided in the proof of Theorem 2.2, while using more accurate estimates for the $\mathcal{D}(W_j)$'s. Recall that $\epsilon' = \frac{(\beta-1)^3 \cdot \epsilon}{30(\beta+1)\cdot\lambda}$ and that (in Case 1) we obtain, for each $j \in [i', i]$, an estimate $\widetilde{\delta}_j$ of $\mathcal{D}(W_j)$ such that (with probability at least $1 - 1/10\log_\beta(1/\epsilon')$) 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, where in the latter case $\widetilde{\delta}_j$ is reset to $\epsilon'$. Hence, $|W_j| \leq \widetilde{w}_j \leq \max(\beta^2 \cdot |W_j|, \epsilon' \cdot \beta^j)$ for every $j \in [i', i]$, whereas the fact that $\widetilde{w}_j \geq |W_j|$ (for all $j \in [i', i]$) implies that

$$\mathtt{ess}_{\beta^2\cdot\epsilon}(\mathcal{D}) \leq \sum_{j \in [i'-1]} \widetilde{w}_j + \sum_{j \in [i', i]} \widetilde{w}_j < \frac{\beta^{i'}}{\beta - 1} + \sum_{j \in [i', i]} \widetilde{w}_j.$$

On the other hand, assuming that $\mathcal{D}(W_i) \geq \frac{\Delta'}{10\lambda}$, which holds with probability at least 0.9, we get $\mathcal{D}(W_i) \geq \frac{0.4(\beta-1)\epsilon}{10\lambda}$, which is lower-bounded by $\eta \stackrel{\text{def}}{=} \frac{(\beta-1)\cdot\epsilon}{30\cdot\lambda} = (\beta+1)\cdot(\beta-1)^{-2}\cdot\epsilon'$ (rather than by $(\beta-1)\cdot\epsilon/6\ell$ as in the proof of Theorem 2.2)[24]. Recalling that $i' = \max(1, i - \log_\beta(1/\epsilon'))$, we invoke Claim 2.2.2 (with $\epsilon'$ and $\eta$ as set above) and obtain

$$\begin{aligned} \frac{\beta^{i'}}{\beta-1} + \sum_{j \in [i', i]} \widetilde{w}_j &< \frac{\beta}{\beta-1} + \left(1 + \frac{(\beta+1)\cdot\epsilon'}{(\beta-1)\cdot\eta}\right) \cdot \beta^2 \cdot \sum_{j \leq i} |W_j| \\ &< \beta^4 \cdot \sum_{j \leq i} |W_j|, \end{aligned}$$

where the last inequality uses $\frac{\beta}{\beta-1} < (\beta-1)\cdot\sum_{j\leq i}|W_j|$, which may be assumed without loss of generality.[25] Recalling that $\mathtt{ess}_\epsilon(\mathcal{D}) \geq \frac{1}{\gamma_{i,\ell}} \cdot \sum_{j\leq i}|W_j|$ (by Part 2 of Claim 2.1.2), it follows that the

[24] The point is that here $\lambda$ rather than $\ell$ appears in the denominator, where the focus is on $\lambda \ll \ell$. (In the proof of Theorem 2.2, $\epsilon' = \epsilon$.)

[25] Evidently, $\frac{\beta}{\beta-1} < (\beta-1)\cdot\sum_{j\leq i}|W_j|$ 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 $\mathcal{D}$ using $O(1/\epsilon)$ samples. 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 $m'$ such the total weight of the heaviest $m'$ elements in $H$ is at most $1-\beta^2\cdot\epsilon$. Denoting the set of the heaviest $m'$ elements by $H'$, and recalling that $\mathcal{D}(H') \geq \mathcal{D}(H) > 1-\beta^3\cdot\epsilon$, it follows that $1 - \mathcal{D}(H') \in [\beta^2\cdot\epsilon, \beta^3\cdot\epsilon]$. Hence, $\mathtt{ess}_{1-\mathcal{D}(H')}(\mathcal{D}) = m'$, which implies that $m'$ is an $[\beta^2\epsilon, \beta^3\epsilon]$-effective support size of $\mathcal{D}$. Using a change of parameters, we obtained the desired approximator.

output in this case (i.e., $\frac{\beta^{i'}}{\beta-1} + \sum_{j\in[i',i]} \widetilde{w}_j$) is sandwiched between $\mathtt{ess}_{\beta^2\epsilon}(\mathcal{D})$ and $\beta^4 \cdot \gamma_{i,\ell} \cdot \mathtt{ess}_\epsilon(\mathcal{D})$. Hence, Case 1 yields a $\gamma_{i,\ell} \cdot \beta^4$-factor approximation of the $[\epsilon, \beta^2 \cdot \epsilon]$-effective support size of $\mathcal{D}$.

When Case 2 holds we use $\sum_{j\in[\widetilde{\ell'}+2\log_\beta\lambda+1,\widetilde{\ell'}+\lambda]} \mathcal{D}(W_j) > 0.09\Delta'/t$ 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 $\mathcal{D}(W_i) > \frac{0.09\Delta'/t}{10\lambda} \geq (\beta-1)\epsilon/300t\lambda$, since $\Delta' > 0.4(\beta-1)\epsilon$. Using the minimality of $\ell'$, 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 $\lambda \geq c/\epsilon$ (and $c = 300t/(\beta-1)^2$), 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 Part 2 of Claim 2.1.2, we get

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

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

The remaining cases. 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 $\Delta'$ and $\lambda$). 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 $\Delta' \geq (1-(0.11/t))^{t-1}\cdot\widetilde{\Delta} > 0.89\widetilde{\Delta}$. 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 $k = 1$.

More efficient algorithms (i.e., $k > 1$). 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 $k^2\cdot c\cdot(1/\epsilon)^{1/k}$ rather than $c/\epsilon$, 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 $\sum_{j\in[\widetilde{\ell'},\widetilde{\ell'}+2\log_\beta\lambda]} \mathcal{D}(W_j)$). Analogously, 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 $\lambda \geq k^2\cdot c\cdot(1/\epsilon)^{1/k}$ is used to argue that $(k+2)\cdot\log_\beta\lambda < \lambda$.[26] 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

[26] Letting $L \stackrel{\text{def}}{=} (\beta-1)\lambda/(k+2)$, we show that $\log_\beta((k+2)(\beta-1)^{-1}L) < L/(\beta-1)$. On the one hand,

$$\log_\beta((k+2)(\beta-1)^{-1}L) < \frac{\ln(k+2)+\ln(\beta-1)^{-1}+\ln L)}{\beta-1}$$

On the other hand, we lower-bound $L$ by $3\cdot\max(\ln L, \ln(\beta-1)^{-1}, \ln(k+2))$, while using $L \geq \frac{\beta-1}{k+2}\cdot k^2\cdot c\cdot(1/\epsilon)^{1/k}$ and $c > 300/(\beta-1)^2$ (and assuming, w.l.o.g., that $\beta \leq 2$).

of Case 2 we use $(k+1)\cdot\log_\beta\lambda \geq \log_\beta(c\lambda/\epsilon)$ (rather than $2\cdot\log_\beta\lambda \geq \log_\beta(c\lambda/\epsilon)$).[27] The theorem follows. ■

**Proving the last item of Theorem 1.9.** Intuitively, the following result is obtained by invoking Theorem 2.3, while setting $t = \log^*(\mathtt{ess}_\epsilon(\mathcal{D})/\epsilon)$. Needless to say, this is problematic because we do not know $t$, but this difficulty can be overcome. The crucial observation is that Cases 1 and 2 in Algorithm 2.3.1 provide a $\gamma_{i,\ell}\cdot\beta^4$-factor approximation of the $[\epsilon,\beta^2\cdot\epsilon]$-effective support size of $\mathcal{D}$, whereas the case of $t'=1$ can be avoided. Recalling that $\gamma_{i,\ell}=1$ if $i<\ell$ and $\gamma_{i,\ell}=\beta^2/(\beta-1)$ otherwise, we also have to deal with the latter case in order to obtain an approximation factor of $\beta^{O(1)}$.

**Theorem 2.4** (the iterative algorithm, revised): *For any constants $\beta>1$ and $k\in\mathbb{N}$, there exists an algorithm that on input $\epsilon>0$ and oracle access to $\mathcal{D}$, uses $\widetilde{O}(\log^*(n/\epsilon)/\epsilon^{1+\frac{1}{k}})$ samples in expectation and outputs an $\beta$-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 $\log_2^{(t)} m<2$. The dependence of the number of samples on $\beta$ is $\mathrm{poly}(1/(\beta-1))$.*

Unlike in the previous three theorems, the sample complexity stated in Theorem 2.4 depends on the effective support size and is bounded in expectation only.[28] These two features seem related, since an algorithm that uses a number of samples that depends on $\mathtt{ess}_\epsilon(\mathcal{D}$ must obtain some crude and necessarily randomized estimate of the effective support size of the distribution $\mathcal{D}$ in order to determine the number of sample that it asks for. On the other hand, recall that by Observation 1.5, when using $\beta=1+\epsilon'=1+\Theta(\epsilon)$, *Theorem 2.4 implies a 1-factor approximator of the $[\epsilon,\beta\cdot\epsilon+\epsilon']$-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:** As said above, we essentially 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 $\gamma_{i,\ell}\cdot\beta^4$-factor approximation of the $[\epsilon,\beta^2\cdot\epsilon]$-effective support size of $\mathcal{D}$, whereas the case of $t'=1$ can be avoided.*[29] Hence, recursing till either Cases 1 or Case 2 occurs, we essentially obtain the desired approximation factor (where the remaining slackness is addressed after describing the aforementioned ways of avoiding the case of $t'=1$)

The first way of overcoming the aforementioned difficulty is to first obtain a very crude approximation of the effective support size and the set $t$ somewhat larger than suggested in the foregoing. Specifically, invoking the basic algorithm (of Theorem 2.1), we obtain, using $O(1/\epsilon)$ samples, 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 observing that this setting of $t$ prevents the algorithm from ever reaching the case of $t'=1$ (since iterating $\lambda\leftarrow(k+2)\log_\beta\lambda$ for $t-2$ times, starting with $\lambda=O(\log_\beta(\widetilde{n}/\epsilon))$, yields a value smaller than $k^2\cdot(c/\epsilon)^{1/k}$, which means that Case 1 holds).[30]

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

[28] For any $t\in\mathbb{N}$, one can generically convert an approximator that uses $s=s(\epsilon,\mathcal{D})$ samples, where $s$ is unknown *a priori*, into an approximator that uses $O(t\cdot s)$ samples 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 $0.9t$ executions. The point is that, with probability at least 0.95, a random execution uses at most $20\cdot s$ samples. Hence, with probability at least $1-2^{-\Omega(t)}$, more than 90% of the executions will terminate while using $20\cdot s$ samples and most of them will output a correct value (i.e., an $[\epsilon,\beta\cdot\epsilon]$-effective support size).

[29] The approximation factor is due to Claim 2.3.2.

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

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).[31] Lastly, the constant $c = 300t/(\beta-1)^2$ will be replaced by a variable $v_{t''}$ that depends on the current recursion depth $t''$, where $v_{t''} = 30g(t'')/(\beta-1)^2$ if $g(m) = 20m \cdot \log_2^2(m+1)$ is used. What will happen is that we shall either stop at Case 2 or at Case 1, because if we never stop at Case 2 then at some recursion depth $t''$ we shall reach $\lambda < 600 < v_{t''}/\epsilon$ (assuming, w.l.o.g, $\beta \leq 2$). In the analysis, we use the fact that $\prod_{t'' \geq 1}\left(1 - \frac{1}{g(t'')}\right) > 1 - \sum_{t'' \geq 1}\frac{1}{g(t'')} > 0.9$.

Addressing the slackness. The foregoing description provides a $\gamma_{i,\ell} \cdot \beta^4$-factor approximation of the $[\epsilon, \beta^2 \cdot \epsilon]$-effective support size of $\mathcal{D}$, where $i \in [\widetilde{\ell'}, \widetilde{\ell'} + \lambda - 1]$ is the index selected in the terminating Case 1 (or Case 2) and $\gamma_{i,\ell} = 1$ if $i < \ell$ and $\gamma_{i,\ell} = \beta^2/(\beta-1)$ otherwise. (The approximation factor is due to Claim 2.3.2.) So the real issue is the case that $i = \ell$, which may occur only if we terminated without ever invoking Case 3.[32] In that case $\lambda = \widetilde{\ell} - \widetilde{\ell'} + 1$ and $\Delta' = \widetilde{\Delta} > (\beta-1)\epsilon/2$. (Furthermore, $i = \ell$ may only occur in Case 1, because in Case 2 the index $i$ is selected in $[\widetilde{\ell'}, \widetilde{\ell'} + 2\log_\beta \lambda]$.) More importantly, we may assume that $\sum_{j \geq i}\mathcal{D}(W_j) \geq \beta\epsilon$, since otherwise $i < \ell$ must hold. Now, if $i = \widetilde{\ell}$ and $\mathcal{D}(W_i)$ appears to be smaller than $\Delta'/2$, then we can afford to re-select $i$ in the relevant interval and proceed with $i < \widetilde{\ell} \leq \ell$ (once such $i$ is selected). Otherwise (i.e., $i = \widetilde{\ell}$ and $\mathcal{D}(W_i) > \Delta'/3$), we estimate $\mathcal{D}(W_i)$ and conduct the analysis while partitioning $W_i$ into $(W_i', W_i'')$ such that $\mathcal{D}(W_i') + \sum_{j>i}\mathcal{D}(W_j)$ appears to be $\beta\epsilon$. Using our estimate for $|W_i''|$ instead of $\widetilde{w}_i$, we obtained the desired approximation, where in the analysis we replace $W_i$ by $W_i''$ and treat $W_i'$ as if it was part of $W_{i+1}$. The theorem follows. ■

## 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.

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

[32] The fact that we get factors of $\beta^4$ and $\beta^2$ rather than a factor of $\beta$ is resolved by change of parameters (while observing that $\beta^{1/4} - 1 = \Omega(\beta - 1)$, which implies that $\mathrm{poly}(1/(\beta^{1/4}-1)) = \mathrm{poly}(1/(\beta-1))$). In contrast, $\beta^2/(\beta-1) \geq 4$ for every $\beta > 1$.

[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.