Universal ε-approximators for integrals

Michael Langberg * Leonard J. Schulman †

Abstract

Let X be a space and F a family of $0,1$-valued functions on X. Vapnik and Chervonenkis showed that if F is "simple" (finite VC dimension), then for every probability measure μ on X and $\varepsilon > 0$ there is a finite set S such that for all $f \in F$, $\sum_{x \in S} f(x)/|S| = [\int f(x) d\mu(x)] \pm \varepsilon$.

Think of S as a "universal ε-approximator" for integration in F. S can actually be obtained w.h.p. just by sampling a few points from μ. This is a mainstay of computational learning theory. It was later extended by other authors to families of bounded (e.g., $[0,1]$-valued) real functions.

In this work we establish similar "universal ε-approximators" for families of unbounded nonnegative real functions — in particular, for the families over which one optimizes when performing data classification. (In this case the ε-approximation should be multiplicative.)

Specifically, let F be the family of "k-median functions" (or k-means, etc.) on $\mathbb{R}^d$ with an arbitrary norm ϱ. That is, any set $u_1, ..., u_k \in \mathbb{R}^d$ determines an f by $f(x) = (\min_i \varrho(x - u_i))^\alpha$. (Here $\alpha \geq 0$.) Then for every measure μ on $\mathbb{R}^d$ there exists a set S of cardinality $\text{poly}(k, d, 1/\varepsilon)$ and a measure ν supported on S such that for *every* $f \in F$, $\sum_{x \in S} f(x)\nu(x) \in (1 \pm \varepsilon) \cdot (\int f(x) d\mu(x))$.

1 Introduction

We study numerical integration, the problem of evaluating $\bar{f} := \int f \, d\mu$, where

(a) f is drawn from a family of nonnegative real-valued functions F on a metric space $\mathcal{X} = (X, \varrho)$; the only access we need to f is the ability to evaluate it at points of our choosing.

(b) μ is a probability measure on X. (In many cases μ has finite support of cardinality n, but this assumption plays no role in our core contributions.)

Our results pertain to families F of functions, described below, that are of importance in clustering (classification) and optimization. For these families, we prove a theorem of the following type, where ε is any small positive real; k is a parameter of complexity of the family F; and $\mathcal{X}$ is $\mathbb{R}^d$ with any norm ϱ :

For any probability measure μ there is a measure ν with $|\text{support}(\nu)| \leq \text{poly}(d, k, 1/\varepsilon)$ such that for all $f \in F$,

$$\int f \, d\nu \in (1 \pm \varepsilon) \int f \, d\mu.$$

(Here $1 \pm \varepsilon$ denotes the interval $[1 - \varepsilon, 1 + \varepsilon]$.)

The context for our work lies in strands of research in numerical analysis, statistics and computer science:

(1) *Numerical integration (or quadrature).* A typical scenario: (a) $\mathcal{X}$ is Euclidean $\mathbb{R}^d$. (b) μ is Lebesgue measure restricted to $[0,1]^d$, or to a ball, or μ is some other canonical measure such as Gaussian measure about the origin. (c) The family F is either very restricted (multivariate polynomials, trigonometric functions, etc.) or else subject only to a "tameness" condition (Lipschitz or order of continuity).

Under such conditions, various methods (familiar ones include Newton-Cotes, Gaussian or Clenshaw-Curtis quadrature in one dimension [26], but see also results in high dimension [38]) ensure the existence of measures ν with small (carefully chosen) support such that for all $f \in F$, either $\int f \, d\nu = \int f \, d\mu$ or $\int f \, d\nu \in \int f \, d\mu \pm \varepsilon$, depending on whether F is of the "restricted" or merely "tame" variety.

In this literature the support of ν is simply called the set of *evaluation points*.

(2) *Vapnik-Chervonenkis (VC) Theory* in statistics and learning theory. A typical scenario: (a) $\mathcal{X}$ can be very general but frequently is Euclidean $\mathbb{R}^d$ or Hamming $\{0,1\}^d$. (b) μ is an unknown and arbitrary probability measure on X; we (the learners) have access to μ only by sampling from it. (c) The range of the functions in F is $\{0,1\}$, or a bounded-cardinality finite set, or a bounded interval of $\mathbb{R}$; in the first case, F is simple in the sense that it has bounded *VC Dimension*, while in the latter cases, an appropriate generalization of VC dimension (there are several such) is bounded for F.

Under such conditions, various theorems [37, 32, 22, 8, 36, 31, 7, 24, 5, 6, 2, 4] (and see [33, 35] for related works) ensure that an "empirical measure" ν

*Computer Science Division, Open University of Israel, 108 Ravutski St., Raanana 43107, Israel, email: mikel@openu.ac.il. Work supported in part by The Open University of Israel's Research Fund (grant no. 46109) and Cisco Collaborative Research Initiative (CCRI).

†California Institute of Technology, MC305-16, Pasadena CA 91125. Email: schulman@caltech.edu. Work supported in part by the NSF.

obtained by selecting a small random sample from μ and fixing the uniform measure on those points, has (with high probability) the property that for all $f \in F$, $\int f\ d\nu = \int f\ d\mu \pm \varepsilon$.

In this literature the support of ν is called an *ε-net*, *ε-transversal*, *ε-sample* or *ε-approximation* [8, 28, 9].

(3) *Approximation algorithms for clustering.* A typical scenario: (a) $\mathcal{X}$ is usually Euclidean (or perhaps ℓ_1) $\mathbb{R}^d$ but may be also be an explicitly given finite metric. (b) μ is usually the uniform measure on an explicitly given finite set of n points in X. (c) F is usually one of the following two families of nonnegative real functions. In each, functions in F are parameterized by k points chosen from X:

$F_{k\text{-median}} = \{f_{c_1,\ldots,c_k}\}_{c_1,\ldots,c_k \in X}$ where $f_{c_1,\ldots,c_k}(x) = \min_{1 \le i \le k} \|x - c_i\|$

$F_{k\text{-means}} = \{f_{c_1,\ldots,c_k}\}_{c_1,\ldots,c_k \in X}$ where $f_{c_1,\ldots,c_k}(x) = \min_{1 \le i \le k} \|x - c_i\|^2$

For these families, various algorithms, some randomized [20, 19, 10] and some deterministic [14], provide a measure ν of small support such that $\int f\ d\nu \in (1 \pm \varepsilon) \int f\ d\mu$ for all $f \in F$. (This can be a crucial step in clustering, since reduction of the size of the data set allows us to run computationally intensive algorithms.)

In this literature the support of ν is usually called a *core-set*, see e.g. [1]. We elaborate on the clustering literature and its relation with our results in a later section of this Introduction.

Since the terminology for ν differs between fields, we have adopted ε-approximation as the most descriptive of the options. Let F be a family of nonnegative functions on a metric space $\mathcal{X}$.

Definition 1. *A measure ν is an ε-approximation for a measure μ with respect to $(F, \mathcal{X})$ if $\int f\ d\nu \in (1 \pm \varepsilon) \int f\ \ d\mu$ for all $f \in F$. For $g : (0, 1) \to \mathbb{N}$ nonincreasing, $(F, \mathcal{X})$ is integrable with complexity g if for every $0 < \varepsilon < 1$ and every μ there is a ν which has $|\operatorname{support}(\nu)| \le g(\varepsilon)$ and which is an ε-approximation of μ w.r.t. $(F, \mathcal{X})$. Finally, $(F, \mathcal{X})$ is finitely integrable if it is integrable with complexity g for some g.*

This is a multiplicative variant of the uniform Glivenko-Cantelli condition [13, 2].

The question we are interested in is: what classes of functions F are finitely-integrable? Or, if we consider only measures μ of finite support n, then what classes of functions are integrable with complexity independent of n? We can give a partial answer to this question.

1.1 The general approach: integration by weighted sampling. Our starting point is the observation (which has been made numerous times previously and falls in the category of "weighted" or "importance" sampling [11, 25, 3, 34]) that to any probability distribution q on X there corresponds the following unbiased estimator for $\bar{f} = \int f\ d\mu$: Sample x from q, and set

$$T = f(x)\mu(x)/q(x). \tag{1.1}$$

T is unbiased because $\int (f(x)\mu(x)/q(x))\ \ dq(x) = \int f(x)\mu(x)\ \ dx = \int f(x)\ \ d\mu(x)$. (Technically $\mathcal{X}$ needs to be finite or compact, and in the latter case, q in the denominator is a density; nothing depends on these niceties.)

One can, of course, use naïve sampling, i.e., $q = \mu$; the problem is that the standard deviation of T can be very large, arbitrarily greater than its expectation. And so the first hurdle we must cross is to reduce $\operatorname{Var} T$; this presents us with a challenging design problem for q. If we can arrange q so that $\operatorname{Var} T$ is not much larger than $\bar{f}^2$, then by collecting a small number of samples independently, we can obtain an estimator that is very likely to be within a $(1 \pm \varepsilon)$ factor of $\bar{f}$. This suggests that a plan for constructing an ε-approximator ν is to sample repeatedly, independently from a carefully chosen q, then let ν be the empirical measure (the uniform distribution on the samples), and integrate $f\mu/q$ with respect to ν.

This brings us, however, to the second and deeper challenge: our estimator needs to *simultaneously* approximate $\bar{f}$ for each of the infinitely many functions f in F. So it is not enough to ensure small probability of error for each f. This is the "uniformity" challenge that Vapnik and Chervonenkis addressed so successfully in certain (i.e., finite-VC-dimension) families of binary functions. A central part of our work will be to find a substitute argument for the case of multiplicative approximation. Without going into detail we mention that the ideas in the existing additive-approximation extensions of the VC theory do not help with the multiplicative approximation problem. Put simply, that work relies on finitely covering the range of the functions; but $(0, \infty)$ cannot be covered by finitely many intervals of the form $(y, (1 + \varepsilon)y)$.

There are essentially two things we need to do: (1) Quantify the difficulty of integrating a function family F. We do this using a new parameter which we call the *total sensitivity* $\mathfrak{S}(F)$; this parameter does not have an analogue in the additive approximation theory. (2) Show that an appropriate weighted sampling scheme addresses the "uniformity" challenge, by constructing ε-approximations whose size depends on two parameters: one is $\mathfrak{S}(F)$, the other is more combinatorial and plays the role analogous to the VC dimension, or more exactly, the shatter function.

1.2 Our results. The results of this work are threefold.

(1) Primarily, we introduce an approach to approximate integration of unbounded nonnegative functions, and show the *existence* of succinct ε-approximators for some important families of functions, while also showing that such ε-approximators do *not* exist for some other (quite simple) families of functions. To this end we introduce and show the power of the key notion of *sensitivity*; while also showing that it is logically independent of another crucial ingredient, the existence of small ε-cover-codes for the family.

We cannot yet provide VC-dimension-style characterizations of when a family of unbounded nonnegative functions is or is not finitely integrable; that is a major open problem and this paper can serve only as a starting point toward its resolution.

(2) We demonstrate the strength of our approach by showing that it yields positive results for a broad family of functions that are important to clustering. What this means is that these families have the off-line/on-line behavior familiar from numerical quadrature: once *someone* has figured out where to put the ε-approximator, *anyone* can integrate functions in the family in constant time.

(3) We show that the application in (2) has algorithmic implications: we demonstrate a generic reduction (applicable to any norm on $\mathbb{R}^d$) of the problem of near-optimal clustering of data points, to the problem of bi-criteria approximate clustering. We stress that this connection has already been demonstrated in the literature (especially in [10]); our contribution in this regard is chiefly to show how the idea can be applied generically to any norm and to any clustering exponent α. (See definition of α below.)

We now give more detail regarding the families of functions relevant to clustering. These families include and generalize the functions relevant to the well-known k-means and k-median problems. Let $\mathcal{X} = (\mathbb{R}^d, \varrho)$ for an arbitrary norm ϱ. For $\alpha > 0$ and $k \geq 1$, the "k-cluster, α-exponent" function family on $\mathcal{X}$ is defined as:

DEFINITION 2. $W(\mathcal{X}, k, \alpha) = \{f_{c_1,\ldots,c_k}\}_{c_1,\ldots,c_k \in \mathbb{R}^d}$ *where* $f_{c_1,\ldots,c_k}(x) = \min_{1\leq i\leq k} \varrho(x-c_i)^\alpha$. *(We refer to* $c_1, \ldots, c_k$ *as the* centers *of* $f_{c_1,\ldots,c_k}$*.)*

For example, setting ϱ to be the Euclidean norm, the *k-median problem* on $\mathcal{X} = (\mathbb{R}^d, \varrho)$ is: Given $x_1, \ldots, x_n \in X$ (and letting μ be the uniform measure on $x_1, \ldots, x_n$), find $\Delta(\mathcal{X}, k, 1, \mu) := \inf_{f \in W(\mathcal{X},k,1)} \bar{f}$. Likewise, we obtain the *k-means problem* with $W(\mathcal{X}, k, 2)$. More generally one seeks for general ϱ and α: $\Delta := \Delta(\mathcal{X}, k, \alpha, \mu) := \inf_{f\in W(\mathcal{X},k,\alpha)} \bar{f}$, and frequently one requires as part of the output a specific $f^* \in W(\mathcal{X}, k, \alpha)$ for which $\overline{f^*} = \Delta$. (It is easy to see that the infimum is achieved.)

As noted above, we show two major results regarding the family $W(\mathcal{X}, k, \alpha)$. First, we show that $W(\mathcal{X}, k, \alpha)$ has succinct ε-approximators (which are typically referred to as strong core-sets in the setting of clustering). Second, we describe a general reduction from finding these ε-approximators efficiently, to "bi-criteria" approximation of $W(\mathcal{X}, k, \alpha)$. Namely, we show for approximation parameters $c > 0$ and $\beta > 0$, that given a function $f^* \in W(\mathcal{X}, \beta k, \alpha)$ such that $\bar{f}^* \leq c \min_{f\in W(\mathcal{X},k,\alpha)} \bar{f}$, one can efficiently find an ε-approximator to $W(\mathcal{X}, k, \alpha)$. We note, that using standard ideas, the latter implies efficient (linear in n time) algorithms for finding near-optimal functions f^* for the family $W(\mathcal{X}, k, \alpha)$. Namely functions f^* satisfying $\overline{f^*} \leq (1+\varepsilon) \min_{f\in W(\mathcal{X},k,\alpha)} \bar{f}$. More specifically, one can exhaustively solve the clustering problem on the ε-approximator and thus obtain a solution for $\mathcal{X}$.

Our results for the function families relevant to clustering can be summarized by the following theorems. In what follows the $\tilde{O}(\cdot)$ notation neglects logarithmic terms in d, k, α and $1/\varepsilon$.

THEOREM 1.1. *Let* $\mathcal{X} = (\mathbb{R}^d, \varrho)$. $W(\mathcal{X}, k, \alpha)$ *has an* ε*-approximator of size* $\tilde{O}\left(\frac{(\alpha^4+1)d^2k^3 2^{4\alpha}}{\varepsilon^2}\right)$.

In the terms below we assume oracle access to f, ϱ, the distribution μ, and the distribution q we construct.

THEOREM 1.2. *Let* $X \subset \mathbb{R}^d$ *be of size* n. *Let* $\mathcal{X} = (X, \varrho)$. *Let* $\beta > 0$ *and* $c > 0$ *be constants. Let* $\mathcal{A}$ *be any algorithm that finds a function* $f^* \in W(\mathcal{X}, \beta k, \alpha)$ *such that* $\bar{f}^* \leq c \min_{f\in W(\mathcal{X},k,\alpha)} \bar{f}$ *in time* $T_{\beta,c}$. *Then an* ε*-approximator of size* $\tilde{O}\left(\frac{(\alpha^4+1)d^2k(c+\beta k)^2 2^{6\alpha}}{\varepsilon^2}\right)$ *can be found in time* $T_{\beta,c} + O(ndk\beta) + \tilde{O}\left(\frac{(\alpha^4+1)d^2k(c+\beta k)^2 2^{6\alpha}}{\varepsilon^2}\right)$.

1.3 Related work. As mentioned above, strong core-sets (or ε-approximators) for k-means and k-median have been extensively studied in the past. In [20], it is shown that for sets $X \subset \mathbb{R}^d$ of size n, a core-set of size $O(k\varepsilon^{-d} \log n)$ can be constructed. The dependence on n was removed in [19] to obtain core-sets of size $O(k^3\varepsilon^{-d-1})$. For high dimensional spaces, Chen [10] reduced the exponential dependence on d at the price of a logarithmic dependence in n and obtains core-sets of size $O(k^2 d\varepsilon^{-2} \log n)$. For comparison, the core-sets we obtain are of size polynomial in d, k, and $1/\varepsilon$, and independent of n. (It is possible such a result may be obtained in other ways; Feldman (private communication) has noted that a variation in the work of [15] may yield such core-sets.)

In a nutshell, the results above all use the paradigm of "bi-criteria" approximation mentioned previously. Namely, they show how to construct strong core-sets given a solution to the βk-median (or mean) of value comparable to the optimal k clustering. This then leads to efficient algorithms for the solution of k-means and k-median. The best running time to date is that of $\tilde{O}(ndk + 2^{\tilde{O}(k/\varepsilon)} + d\,\mathrm{poly}(k/\varepsilon))$ presented in [15].

In its overall structure, our algorithms for the family $W(\mathcal{X}, k, \alpha)$ closely resemble those mentioned above. Namely, given a bi-criteria approximation, a certain random process is preformed in order to obtain a small ε-approximation.

For comparison of our work with earlier works mentioned previously in VC theory establishing small ε-approximations, note that all these works provide what is essentially an additive ε-approximation. The requirement of multiplicative approximation is what makes our work very different from those referenced earlier. (However, there are interesting connections to the VC theory. The dependence on d in our argument relies upon bounding a shatter-like function.)

Finally we note that (multiplicative) ε-approximators for finite sets X and for a number of function families F defined on the line were studied in [18]. These families F include linear functions, piecewise monotone functions, and a variant of the k-median problem on the line. Most of these function families do not have finite ε-approximators when X is taken to be infinite. The work of [18] focuses on quantifying the dependence of the ε-approximator's size on the cardinality n of X.

1.4 Layout of the paper. In Section 2 we introduce the notion of total sensitivity and show how it can be used to design our sample distribution q. We then focus on the families $F = W(\mathcal{X}, k, \alpha)$: In Section 3 we analyze their total sensitivity, and in Sections 4 and 5 we use this analysis to prove Theorems 1.1 and 1.2.

2 Total sensitivity

2.1 Sensitivity and sampling. For each $x \in X$, define the *sensitivity* of x w.r.t. (F, μ) by $\sigma_{F,\mu}(x) = \sup_{f\in F} f(x)/\bar{f}$; when the identities of F and μ are either clear or immaterial we abbreviate this by $\sigma(x)$. Define the *total sensitivity* of F by $\mathfrak{S}(F) = \sup_\mu \int \sigma_{F,\mu}(x)\ d\mu(x)$. From a theoretical perspective these are the key concepts but it will be useful at a later point if the results of the current section are stated in terms of the following quantities: $s_{F,\mu}(x)$ (or $s(x)$) is any upper bound on $\sigma_{F,\mu}(x)$ (or $\sigma(x)$), and $S(F) = \sup_\mu \int s_{F,\mu}(x)\ d\mu(x)$. (Obviously $S(F) \geq \mathfrak{S}(F)$.) The inequality to keep in mind is: for any $f \in F$, any μ and any $x \in X$,

$$f(x) \leq s(x)\bar{f}. \tag{2.2}$$

The beauty of the notion of sensitivity is that it leads to a judicious weighted sampling scheme for integration. Let

$$q(x) = s(x)\mu(x)/S(F) \tag{2.3}$$

and let T be the estimator for $\bar{f}$ given in Equation 1.1. Observe that the sampling process tends to pick x's of high sensitivity. Now we'll see why total sensitivity is a crucial parameter:

THEOREM 2.1. $\operatorname{Var} T \leq (S(F) - 1)\bar{f}^2$.

Proof. Let $S := S(F)$. We analyze the value of $(1/\bar{f}^2)\operatorname{Var} T$ which is equal to

$$\begin{aligned}
& (1/\bar{f}^2)\int\left(\frac{f(x)\mu(x)}{q(x)} - \bar{f}\right)^2\ dq(x)\\
= \quad & (1/\bar{f}^2)\int \frac{\mu(x)s(x)}{S}\left(\frac{f(x)S}{s(x)} - \bar{f}\right)^2\ dx\\
= \quad & (1/\bar{f}^2)\int\left(\frac{f(x)^2\mu(x)S}{s(x)} - 2\bar{f}f(x)\mu(x) + \frac{\mu(x)s(x)\bar{f}^2}{S}\right)\ dx\\
= \quad & \left((1/\bar{f}^2)\int \frac{f(x)^2 S}{s(x)}\ d\mu(x)\right) - \left((2/\bar{f})\int f(x)\ d\mu(x)\right)\\
& \qquad + \left(\int s(x)/S\ d\mu(x)\right)\\
= \quad & (1/\bar{f}^2)\int \frac{f(x)^2 S}{s(x)}\ d\mu(x) - 2 + 1
\end{aligned}$$

Now we apply Inequality 2.2 to obtain,

$$\leq (1/\bar{f}^2)\int f(x)\bar{f}S\ d\mu(x) - 1 = (S/\bar{f})\int f(x)\ d\mu(x) - 1 = S - 1.$$

The effect of averaging over multiple samples is expressed by a simple application of Chebyshev's inequality.

LEMMA 2.1. *Let $\varepsilon > 0$. Let $f \in F$. Let R be a random sample of X of size $a \geq \frac{2(S-1)}{\varepsilon^2}$ according to the distribution q. Then* $\Pr\left[\left|\bar{f} - \frac{1}{a}\sum_{x\in R}\frac{S(F)f(x)}{s(x)}\right| \geq \varepsilon\bar{f}\right] \leq 1/2$.

Proof. Let T_f be the random variable which obtains the value of $\frac{Sf(x)}{s(x)}$ with probability $q(x)$. As we have seen, $E[T_f] = \bar{f}$ and by Theorem 2.1 $\operatorname{Var} T_f \leq (S-1)\bar{f}^2$. Now, let T be the sum of a variables i.i.d. to T_f: $T_f^{(1)}, \dots, T_f^{(a)}$. Setting $T = \sum_i T_f^{(i)}$, it follows that $E[T] = a\bar{f}$ and that $V[T] = aV[T_f] \leq a(S-1)\bar{f}^2$. Thus, $\left|\bar{f} - \frac{1}{a}\sum_{x\in R}\frac{S(F)f(x)}{s(x)}\right| \geq \varepsilon\bar{f}$ iff $|T - E[T]| > a\varepsilon\bar{f}$. Using Chebyshev's inequality, the probability of this event is at most

$$\frac{V[T]}{a^2\varepsilon^2\bar{f}^2} \leq \frac{S-1}{a\varepsilon^2} \leq \frac{1}{2}.$$

2.2 Case studies. To put our definitions in perspective we present some examples.

2.2.1 Families that do not have ε-approximators. In order to illustrate the nature of the problem we examine a few very simple families F that do not have finite ε-approximators. The simplest example is this: X is the unit interval and $F_{\text{ivl}} := \{f_{a,b}\}_{a<b}$ where $f_{a,b}(x) = 1$ if $a < x < b$ and 0 otherwise. This is a family that trivially has an ε-approximator for additive approximation: let ν be supported uniformly on the multiset $x_k = \inf\{x : \int_0^x \mu(x)\, dx \geq k\varepsilon/2\}$. The conclusion can also be seen from the more abstract consideration that F_{ivl} has VC dimension 2.

On the other hand, if μ is any measure without atoms, then for any finitely supported ν, there is a pair $a < b$ such that $\mu((a, b)) > 0$ but $(a, b) \cap \text{support}\, \nu = \emptyset$, so $(\int f_{a,b}\, d\nu)/(\int f_{a,b}\, d\mu) = 0$. It is easy to see that $\mathfrak{S}(F_{\text{ivl}}) = \infty$ because every point x has unbounded sensitivity.

For our second example we again let X be the unit interval and let $F_{\text{ray}} := \{f_a\}_a$ where $f_a(x) = 1$ if $a < x$ and 0 otherwise. For the purpose of additive approximation this is an even easier family than F_{ivl} (it has VC dimension 1), but this family too has no finite ε-approximators: again let μ be the uniform measure. If ν is finitely supported, let a be the greatest point in its support; then $(\int f_a\, d\nu)/(\int f_a\, d\mu) = 0$. To see that $\mathfrak{S}(F_{\text{ray}}) = \infty$, let μ be the uniform measure. Now $\sigma(x) \geq \sup_{a<x} 1/(1 - a) = 1/(1 - x)$, so $\mathfrak{S}(F_{\text{ray}}) \geq \int_0^1 1/(1 - x)\, dx = \infty$.

It is worth examining the "finitary" versions of each of these examples, with μ required to have support of cardinality $\leq n$. For F_{ivl}, simply take μ to be uniform on n points $x_1 < \ldots < x_n$. None of these points may be omitted in an ε-approximator. And, $\mathfrak{S}(F_{\text{ivl}})$ (under the restriction to measures of support $\leq n$) equals n. For F_{ray}, take the same points $x_1 < \ldots < x_n$ but now set $\mu(x_i) = 2^{-i}$ (except that $\mu(x_1) = 2^{-1} + 2^{-n}$). Then $\sigma(x_i) \geq 2^{i-1}$, so $\mathfrak{S}(F_{\text{ray}})$ (again under the cardinality restriction) is $\geq \sum_i \sigma(x_i)\mu(x_i) \geq \sum_i 2^{i-1}2^{-i} = n/2$.

2.2.2 Families that have ε-approximators, yet sampling from μ is futile. Our next example is the simplest and most tractable prototype of the families of functions that we shall show in this paper to have finite ε-approximators. Let $X = \mathbb{R}$ and let $F_{\text{1-means}} := \{f_a\}_{a\in\mathbb{R}}$ where $f_a(x) = (x - a)^2$. Let's see why trying to construct an ε-approximator by sampling repeatedly from μ and setting ν to be the empirical measure, fails.

For $0 < p < 1$, consider the following measure μ supported on $\{0, 1\}$: $\mu(\{0\}) = p, \mu(\{1\}) = 1 - p$. Observe that $\int f_1\, d\mu = p$ but that a sample of $\sim 1/p$ points is needed for ν to be likely to include 0 in its support; so long as this does not occur, $\int f_1\, d\nu = 0$. So there is no finite sample size at which we likely obtain an ε-approximator for every μ.

However, it is plain that a weighted sampling scheme q (as described in Sec. 1.1) can be made to work on these simple "counterexample" measures, for instance by letting q be uniform on $\{0, 1\}$. This leaves open the question of handling general measures μ. In Section 2.4 we show that indeed weighted sampling can generate ε-approximators for the "1-means" family for any μ and in any dimension, and we will actually deduce the sampling scheme q in closed form.

For the more complicated (and more important) families of functions that are important for clustering, there is almost certainly no closed form for optimal solutions. Nonetheless we show in Section 4 that it is possible to obtain an effective sampling scheme.

2.3 Properties of total sensitivity. What operations on function families preserve bounds on total sensitivity? Two are easy to see: addition and projective closure.

PROPOSITION 1. *Given families of nonnegative functions F, G on $\mathcal{X}$, let $F + G = \{f + g : f \in F, g \in G\}$. Then $\mathfrak{S}(F + G) \leq \mathfrak{S}(F) + \mathfrak{S}(G)$.*

Proof. $(f(x) + g(x))/(\bar{f} + \bar{g}) \leq f(x)/\bar{f} + g(x)/\bar{g}$ (keep in mind positivity of the quantities).

Given a family of nonnegative functions F on $\mathcal{X}$, its *closure* F^c includes any function g such that for any bounded set $A \subseteq X$ and any $\delta > 0$ there exist $f \in F$ and $c > 0$ such that for all $x \in A$, $cf(x) \leq g(x) \leq e^{\delta}cf(x)$. The *projective closure* of F, $\mathbb{P}F^c$, is formed by choosing one representative from each equivalence class in F^c w.r.t. multiplication by positive scalars. (We need this operation mainly in order to automatically include constant functions in our families.) The proof of the following proposition appears in Appendix A.

PROPOSITION 2. $\mathfrak{S}(\mathbb{P}F^c) = \mathfrak{S}(F)$.

Finally, multiplication does *not* preserve good bounds on total sensitivity. Let $FG = \{fg : f \in F, g \in G\}$. The proof of the following proposition appears in Appendix B.

PROPOSITION 3. *There exist F, G for which $\mathfrak{S}(FG) \geq e^{\Omega(\max\{\mathfrak{S}(F),\mathfrak{S}(G)\})}$.*

2.4 Vector norm functions. The key to calculating the total sensitivity of $F_{\text{1-means}}$ (mentioned above) is the observation that there is a real vector space V (the vector space of affine linear functions on $\mathbb{R}^1$) such that for every $f \in F_{1\text{-means}}$ there is a $v \in V$ satisfying $f(x) = |v(x)|^2$.

THEOREM 2.2. *Let $V \subseteq \mathbb{R}^X$ be a real vector space of dimension d. Let $F = \{|p(x)|^2 : p \in V\}$. Then $\mathfrak{S}(F) = d$.*

Proof. Let μ be any probability measure on X, and consider it as defining an inner product on V that is diagonal in the "basis" of delta-functions: the inner product of $u, v \in V$ is $\int u(x)v(x)\, d\mu(x)$. (Correspondingly, we freely consider μ as a diagonal matrix.) Let $p^1, \ldots, p^d$ be an orthonormal basis for V w.r.t. μ, and let P be the (possibly infinite) matrix having these vectors as rows. Consider $\mathbb{R}^d$ as a space of column vectors and let S^{d-1} denote the unit sphere (w.r.t. the identity inner product). Observe that

$$\sigma(x) = \sup_{u \in S^{d-1}} \frac{|(u^\dagger P)(x)|^2}{u^\dagger P \mu P^\dagger u} = \sup_{u \in S^{d-1}} |(u^\dagger P)(x)|^2$$

The last term is maximized by letting u be a scalar multiple of the column vector $P(x) = (p^1(x), \ldots, p^d(x))$. So,

$$\sigma(x) = \left| \frac{P^\dagger(x) P(x)}{\|P(x)\|} \right|^2 = \|P(x)\|^2$$

Therefore $\mathfrak{S} = \int \sigma\, d\mu = \text{Tr } P^\dagger P \mu = \text{Tr } P \mu P^\dagger$. Since the vectors p^i are orthonormal w.r.t. μ, $\mathfrak{S} = d$.

COROLLARY 2.1. *The family of squares of d-variate real polynomials of total degree $\leq t$ has total sensitivity $\binom{t+d}{d}$.*

Observe that the vector space of affine linear functions on $\mathbb{R}$, $V_{\text{aff}} = \{b + ax\}_{a,b \in \mathbb{R}}$, is of dimension 2. Let $F_{d,1\text{-means}}$ be $\{f_a\}_{a \in \mathbb{R}^d} : \mathbb{R}^d \to \mathbb{R}$ where $f_a(x) = \sum (x_i - a_i)^2$. We can see that $\mathfrak{S}(F_{d,1\text{-means}}) \leq 2d$ by applying Corollary 2.1 and Proposition 1; however, we shortly improve this bound.

Corollary 2.1 and Proposition 1 together beg the question whether one can bound the total sensitivity of the cone of nonnegative real polynomials of total degree $\leq 2t$. Proposition 1 shows that for the cone of "sum of squares" polynomials, the bound $\binom{t+d}{d}k$ holds for polynomials which are sums of k squares; we do not know whether the dependence on k is necessary. Moreover, there are nonnegative polynomials which are not sums of squares (except in the special cases $d = 1$; $t = 1$; and $(d = 2, t = 2)$) [23]).

3 Clustering functions and their total sensitivity

In this section we study the total sensitivity of the family $F = W(\mathcal{X}, k, \alpha)$. Let $\mathcal{X} = (X, \rho)$. Consider any probability measure μ on X; let $\bar{f} = \int f\, d\mu$ and $\Delta = \inf_{f \in F} \bar{f}$.

THEOREM 3.1. *(1) For $\alpha \geq 1$, $\mathfrak{S}(W(\mathcal{X}, k, \alpha)) \leq (k+1)2^{2\alpha} + 2^\alpha$. (b) For $0 \leq \alpha \leq 1$, $\mathfrak{S}(W(\mathcal{X}, k, \alpha)) \leq 2k + 3 + 2\sqrt{6k}$.*

Observe that these bounds are independent of $\mathcal{X}$.

Proof. If $\alpha = 0$ the theorem is trivial. Otherwise, let $f^* = f_{u_1^*, \ldots, u_k^*}$ be a function in $W(\mathcal{X}, k, \alpha)$ for which $\overline{f^*} = \Delta$. Let U_i be the Voronoi cell of u_i^*, and let $p_i = \mu(U_i)$. (Each p_i is positive unless the support of μ has cardinality less than k in which case the theorem is trivial.) Let $m_i = \frac{1}{p_i} \int_{U_i} \varrho(x - u_i^*)^\alpha\, d\mu(x)$, so $\Delta = \sum p_i m_i$. By a simple Markov inequality, for each i, $\mu(B(u_i^*, (2m_i)^{1/\alpha}) \cap U_i) \geq p_i/2$. (Here $B(x, r)$ denotes the closed ball of radius r about x.)

We now analyze $\mathfrak{S}$. Let $x \in U_i$, and let $f = f_{u_1, \ldots, u_k}$ be any function in $W(\mathcal{X}, k, \alpha)$. Let u denote a closest point to u_i^* in $\{u_1, \ldots, u_k\}$, and let $\varrho_i = \varrho(u - u_i^*)$. Then $\bar{f} \geq \int_{U_i} f d\mu \geq [\max(0, \varrho_i - (2m_i)^{1/\alpha})]^\alpha \frac{p_i}{2}$. Also, $\bar{f} \geq \Delta$. Thus, for a parameter $q \in [0, 1]$, $\bar{f} \geq [\max(0, \varrho_i - (2m_i)^{1/\alpha})]^\alpha \frac{qp_i}{2} + (1 - q)\Delta$. At this point the arguments for Parts (1,2) of the theorem diverge. Part (1):

$$\begin{aligned}
&\sigma(x) = \max_f f(x)/\bar{f} \leq \max_f \varrho(x - u)^\alpha / \bar{f} \\
\leq\ & \max_f \frac{(\varrho_i + \varrho(x - u_i^*))^\alpha}{[\max(0, \varrho_i - (2m_i)^{1/\alpha})]^\alpha q p_i/2 + (1 - q)\Delta} \\
\leq\ & \max_{\varrho_i \geq 0} \frac{2^{\alpha-1}(\varrho_i^\alpha + \varrho(x - u_i^*)^\alpha)}{[\max(0, \varrho_i - (2m_i)^{1/\alpha})]^\alpha q p_i/2 + (1 - q)\Delta} \\
& \qquad \text{(apply } |a + b|^\alpha \leq 2^{\alpha-1}(|a|^\alpha + |b|^\alpha)\text{)} \\
=\ & \max_{\varrho_i \geq (2m_i)^{1/\alpha}} \frac{2^{\alpha-1}(\varrho_i^\alpha + \varrho(x - u_i^*)^\alpha)}{(\varrho_i - (2m_i)^{1/\alpha})^\alpha q p_i/2 + (1 - q)\Delta} \\
\leq\ & \max_{\varrho_i^\alpha \geq 2m_i} \frac{2^{\alpha-1}(\varrho_i^\alpha + \varrho(x - u_i^*)^\alpha)}{\max\{0, 2^{1-\alpha}\varrho_i^\alpha - 2m_i\} q p_i/2 + (1 - q)\Delta} \\
& \qquad \text{(apply } |a - b|^\alpha \geq \max\{0, 2^{1-\alpha}|a|^\alpha - |b|^\alpha\}\text{)}
\end{aligned}$$

Let $G(\varrho_i^\alpha) = \frac{\varrho_i^\alpha + \varrho(x - u_i^*)^\alpha}{(2^{1-\alpha}\varrho_i^\alpha - 2m_i) q p_i/2 + (1-q)\Delta}$. Observe that sign $\left(\frac{\partial G}{\partial \varrho_i^\alpha}\right)$ is independent of ϱ_i^α and thus G is monotone as a function of ϱ_i^α. Moreover, $\frac{\varrho_i^\alpha + \varrho(x - u_i^*)^\alpha}{(1-q)\Delta}$ is increasing as a function of ϱ_i^α. Thus the bound on $\sigma(x)$ is maximized at either $\varrho_i = 2m_i^{1/\alpha}$ or $\varrho_i = \infty$. We conclude that for $x \in U_i$:

$$\begin{aligned}
\sigma(x) &\leq \max\left(\frac{2^{2\alpha-1} m_i + 2^{\alpha-1} \varrho(x - u_i^*)^\alpha}{(1 - q)\Delta}, \frac{2^{2\alpha-1}}{q p_i} \right) \\
&\leq \frac{2^{2\alpha-1} m_i + 2^{\alpha-1} \varrho(x - u_i^*)^\alpha}{(1 - q)\Delta} + \frac{2^{2\alpha-1}}{q p_i}
\end{aligned}$$

Thus $\int \sigma \, d\mu$ is equal to

$$\begin{aligned}
& \sum_i \left(\int_{U_i} \sigma \, d\mu \right) \\
\leq \;& \sum_i \left(\int_{U_i} \frac{2^{2\alpha-1}m_i + 2^{\alpha-1}\varrho(x-u_i^*)^\alpha}{(1-q)\Delta} + \frac{2^{2\alpha-1}}{qp_i} \, d\mu \right) \\
= \;& \sum_i \left(\frac{2^{2\alpha-1}p_i m_i}{(1-q)\Delta} + \frac{2^{\alpha-1}p_i m_i}{(1-q)\Delta} + \frac{2^{2\alpha-1}}{q} \right) \\
= \;& \frac{2^{2\alpha-1}+2^{\alpha-1}}{1-q} + \frac{2^{2\alpha-1}k}{q} \\
\leq \;& \left(\sqrt{2^{2\alpha-1}+2^{\alpha-1}} + \sqrt{2^{2\alpha-1}k} \right)^2 \\
& \quad \text{(the minimum of } a/(1-q)+b/q \text{ is } (\sqrt{a}+\sqrt{b})^2) \\
\leq \;& (k+1)2^{2\alpha} + 2^\alpha \qquad \text{(use } (a+b)^2 \leq 2(a^2+b^2))
\end{aligned}$$

This suffices to bound $\mathfrak{S} = \sup_\mu \int \sigma(x) \, d\mu(x)$ and to prove Part (1) of the theorem. Part (2) of the proof is similar in nature and follows.

Part (2):

$$\begin{aligned}
\sigma(x) &= \max_f f_x/\bar{f} \leq \max_f \varrho(x-u)^\alpha/\bar{f} \\
&\leq \max_f \frac{(\varrho_i + \varrho(x-u_i^*))^\alpha}{[\max(0, \varrho_i - (2m_i)^{1/\alpha})]^\alpha qp_i/2 + (1-q)\Delta} \\
&\leq \max_{\varrho_i} \frac{\varrho_i^\alpha + \varrho(x-u_i^*)^\alpha}{[\max(0, \varrho_i - (2m_i)^{1/\alpha})]^\alpha qp_i/2 + (1-q)\Delta} \\
&\leq \max_{\varrho_i \geq (2m_i)^{1/\alpha}} \frac{\varrho_i^\alpha + \varrho(x-u_i^*)^\alpha}{(\varrho_i - (2m_i)^{1/\alpha})^\alpha qp_i/2 + (1-q)\Delta} \\
&\leq \max_{\varrho_i^\alpha \geq 2m_i} \frac{\varrho_i^\alpha + \varrho(x-u_i^*)^\alpha}{(\varrho_i^\alpha - 2m_i)qp_i/2 + (1-q)\Delta}
\end{aligned}$$

In the last inequality we apply $|a+b|^\alpha \leq |a|^\alpha + |b|^\alpha$, thus $|a-b|^\alpha \geq |a|^\alpha - |b|^\alpha$. Let $G(\varrho_i^\alpha) = \frac{\varrho_i^\alpha + \varrho(x-u_i^*)^\alpha}{(\varrho_i^\alpha - 2m_i)qp_i/2 + (1-q)\Delta}$. As before, sign $\left(\frac{\partial G}{\partial \varrho_i^\alpha}\right)$ is independent of ϱ_i^α and thus G is monotone as a function of ϱ_i^α. We conclude that for $x \in U_i$ $\sigma(x)$ is at most

$$\max\left(\frac{2m_i + \varrho(x-u_i^*)^\alpha}{(1-q)\Delta}, \frac{2}{qp_i} \right) \leq \frac{2m_i + \varrho(x-u_i^*)^\alpha}{(1-q)\Delta} + \frac{2}{qp_i}$$

Thus

$$\begin{aligned}
\mathfrak{S} &= \int \sigma \, d\mu = \sum_i \int_{U_i} \sigma \, d\mu \\
&\leq \sum_i \int_{U_i} \left(\frac{2m_i + \varrho(x-u_i^*)^\alpha}{(1-q)\Delta} + \frac{2}{qp_i} \right) d\mu \\
&= \sum_i \left(\frac{2m_i p_i}{(1-q)\Delta} + \frac{m_i p_i}{(1-q)\Delta} + \frac{2}{q} \right) \\
&= \frac{3}{1-q} + \frac{2k}{q} \leq 2k + 3 + 2\sqrt{6k}
\end{aligned}$$

This completes the proof of Part (2).

4 ε-approximators, arrangements and covering codes for $W(\mathcal{X}, k, \alpha)$

Let $\mathcal{X} = (\mathbb{R}^d, \varrho)$. In this section we study the family $F = W(\mathcal{X}, k, \alpha)$ and show that it has succinct ε-approximators. Let S be a bound on the total sensitivity of F. We prove the following theorem which is a rephrased version of Theorem 1.1 stated in the Introduction.

Theorem 4.1. *Let $\mathcal{X} = (\mathbb{R}^d, \varrho)$. The family $W(\mathcal{X}, k, \alpha)$ has an ε-approximator of size $O\left(\frac{(\alpha^4+1)d^2kS^2}{\varepsilon^2}\log\left(\frac{(\alpha+1)dkS}{\varepsilon}\right)\right)$. Thus by Theorem 3.1, for $\alpha \geq 1$ the ε-approximator is of size $\tilde{O}\left(\frac{\alpha^4 d^2 k^3 2^{4\alpha}}{\varepsilon^2}\right)$ and for $\alpha \leq 1$ the ε-approximator is of size $\tilde{O}\left(\frac{d^2k^3}{\varepsilon^2}\right)$. Here $\tilde{O}(\cdot)$ neglects logarithmic terms in d, k, α and $1/\varepsilon$.*

Roughly, speaking this is done in three major steps. Primarily, recall that $\mathcal{X}$ (and thus F) is defined by a norm ϱ which in turn is defined by a centrally symmetric convex set C_ϱ. In our first step, we show that one may assume w.l.o.g. that ϱ is *well behaved*. Namely, that any arrangement (in the sense of combinatorial geometry) described by n translates of C_ϱ (i.e., the dissection of $\mathbb{R}^d$ by the boundaries of the translates) has low complexity, of approximately n^d. In our second step, we define the notion of an "ε-cover code" for the family F. Namely, a set of functions $F' \subset F$ that *approximates* the set F with respect to any finite subset of the support X. We show that F has small cover-codes F' if its underlying norm ϱ is well behaved. Finally, in our third step we show that a small ε-cover code for F implies a small ε-approximator. This will conclude the proof of Theorem 4.1. We give a detailed outline below.

4.1 Well behaved norms.

Definition 3. *Let ϱ be a norm corresponding to the centrally symmetric convex set C. Consider any collection of convex sets $\{C_1, \ldots, C_n\}$ where each set C_i is equal to $r_iC_\varrho + v_i$. Here v_i is a vector in $\mathbb{R}^d$, r_i is a positive real, and $r_iC = \{x \mid \varrho(x) \leq r_i\}$. The collection of sets C_i describes an arrangement in $\mathbb{R}^d$. We say that ϱ is Γ-well behaved if the complexity of any such arrangement is bounded by $(n\Gamma)^d$.*

Theorem 4.2. *Let ϱ be any norm. Let $\Gamma = (cd/\sqrt{\varepsilon})^d$ for a sufficiently large constant c. There exists a Γ-well behaved norm ϱ' such that any ε-approximator for ϱ' will yield an $O((1+\varepsilon)^\alpha - 1)$-approximator for the original ϱ.*

The proof of Theorem 4.2 appears in Section 6. Notice that when ε is small compared to $1/\alpha$, the quality loss in our approximator is small.

(This is the place to draw attention to a crucial distinction between Euclidean norm and general norms ϱ. Recall that the collection of positive homothetes (= translation and multiplication by positive scalars) of the unit ball in $\mathbb{R}^d$, has VC dimension $d+1$. This implies that the complexity of an arrangement of n balls is $O(n^{d+1})$. Grünbaum [16] conjectured that the same VC dimension bound held for the positive homothetes of any fixed compact convex set in $\mathbb{R}^d$. If this (or even a weaker $O(d)$ upper bound) were true, the size of our ε-approximators in Theorem 4.1 would no longer have quadratic dependence in d but rather linear dependence. However, the conjecture has been falsified. The first counterexample was due to Naiman and Wynn, who showed that the collection of translates of a box [29] has VC dimension $\lfloor 3d/2 \rfloor$. This of course is not large enough to be an obstacle to our application. But very recently, Naszódi and Taschuk showed that in dimension 3 and above there is *no* upper bound on the VC dimension of the positive homothetes of convex bodies [30]. Actually even stronger, there is a convex body whose collections of translates has infinite VC dimension.)

4.2 ε-cover codes. A central role in the VC theory of additive approximation is played by the *shatter function* of a family F of Boolean functions. (Frequently, VC dimension appears merely as a means to bound the shatter function.) The viewpoint which generalizes to our situation is this: Shatter functions measure the size of *covering codes* or *transversals* for (restrictions of) F. In the boolean case the notion of "covering" is trivial: one function covers another only if their restrictions are identical. But we require something more general. Let F be a function family and $s(x)$ a sensitivity bound (as in Section 2.1). Let $A \subseteq X$ be finite, $a = |A|$. For $g : X \to \mathbb{R}$, let $\nu_A(g) = (1/a)\sum_{x \in A} g(x)$. For $f, f' \in F$ and $x \in A$ define $\hat{f} = \nu_A(f/s)$ and

$$D_{A,x}(f,f') = \left| \frac{f(x)}{\hat{f}s(x)} - \frac{f'(x)}{\hat{f}'s(x)} \right|$$

Notice that $D_{A,x}(f,f')$ also depends on our bound $s(x)$ on the sensitivity of F at x, however, we do not write s explicitly as a parameter in D. The definition of D and that which follows are designed to fit our needs in Theorem 4.4 of Section 4.3.

Definition 4. *$F' \subseteq F$ is an ε-cover-code for (F, A, s) if for every $f \in F$ there is an $f' \in F'$ such that $\frac{\bar{f}'}{\hat{f}'} \leq \frac{\bar{f}}{\hat{f}}$ and for every $x \in A$, $D_{A,x}(f,f') \leq \frac{\varepsilon}{64S}\left(1 + \frac{f(x)}{\hat{f}s(x)} + \frac{f'(x)}{\hat{f}'s(x)}\right)$.*

In Section 7 we show that for well behaved ϱ, F has succinct cover codes.

Theorem 4.3. *Let $\mathcal{X} = (\mathbb{R}^d, \varrho)$. Let $F = W(\mathcal{X}, k, \alpha)$ where ϱ is a Γ-well behaved norm. Let $A \subset \mathbb{R}^d$ be a set of size a, F has an ε-cover-code F' of size $\left[\left(\frac{Sa}{\varepsilon}\right)^{\Theta(\alpha^2+1)}\Gamma\right]^{2dk}$.*

4.3 ε-approximators. Finally, we turn to show a general VC-type argument. Loosely speaking, we show that small cover codes imply succinct ε-approximators via random sampling. The analysis of our theorem holds for any family F with support X and total sensitivity at most S. Our proof resembles that used in [37, 22] for the proof of small ε-nets. We present our proof in Section 8.

Theorem 4.4. *Suppose that for some $a \geq 8(S-1)/\varepsilon^2$, every $A \subseteq X$ of cardinality $|A| = 2a$ possesses an ε-cover-code (w.r.t. F and sensitivity bound s) of cardinality at most $\frac{1}{8}e^{\frac{a\varepsilon^2}{100S^2}}$. Then a sample of a points from q is (with probability $\geq 1/2$) an ε-approximator for F.*

4.4 Proof of Theorem 4.1. We now are ready to prove Theorem 4.1: Let $\mathcal{X} = (\mathbb{R}^d, \varrho)$, and $F = W(\mathcal{X}, k, \alpha)$. In Theorem 4.2 we show that one may assume that ϱ is Γ-well behaved, for $\Gamma = (cd/\sqrt{\varepsilon})^d$ (here c is a sufficiently large constant). For a set A of size a, we have shown in Theorem 4.3 that for such well behaved ϱ the family F has cover codes of cardinality $\left[\left(\frac{Sa}{\varepsilon}\right)^{\Theta(\alpha^2+1)}\Gamma\right]^{2dk} = \left(\frac{cd^dSa}{\varepsilon^d}\right)^{\Theta((\alpha^2+1)dk)}$. As $\left(\frac{cd^dSa}{\varepsilon^d}\right)^{\Theta((\alpha^2+1)dk)} \leq \frac{1}{8}e^{\frac{a\varepsilon^2}{200S^2}}$ for values of a of size at least $\Theta\left(\frac{(\alpha^2+1)d^2kS^2}{\varepsilon^2}\log\left(\frac{(\alpha+1)dkS}{\varepsilon}\right)\right)$ by Theorem 4.4 (and a slight change of parameters to compensate for the loss of Theorem 4.2) we may conclude Theorem 4.1 stated in the beginning of this section.

5 Finding ε-approximators for $W(\mathcal{X}, k, \alpha)$ via bi-criteria approximation

Algorithmically, our method for finding an ε-approximator described in Section 4 is a reduction to bi-criteria approximation. In the Euclidean case, such bi-criteria approximations already exist, and our results imply that further progress on such algorithms for other norms will immediately lead to efficient clustering algorithms in those norms.

Conceptually our method is simple: First, the algorithm computes a distribution q on $X = \{x_1, \ldots, x_n\}$; then it selects independent samples from X according to q. In what follows we show how one can construct the distribution q.

A "bi-criteria" approximation to the problem is a function $f^* \in W(\mathcal{X}, \beta k, \alpha)$ such that $\bar{f^*} \leq c \min_{f \in W(\mathcal{X},k,\alpha)} \bar{f}$. Here, both β and c are parameters that will affect the running time computed below.

Assuming that finding f^* can be done in time $T_{\beta,c}$, we show that the distribution q and an ε-approximator can be found efficiently in time $\simeq T_{\beta,c} + O(|X|d\beta k)$. Namely we prove Theorem 1.2 stated in the Introduction.

In general to compute q one really only needs to compute $s(x)$ (a bound on the sensitivity) for each $x \in X$. Recall that in our setting it is necessary that the values $s(x)$ we compute be greater or equal to the ideal values $\sigma(x) = \sup_x \frac{f(x)}{\bar{f}}$ corresponding to our given family F and distribution μ. The size of our ε-approximator resulting from q will depend on the value of $S = \int_X s(x)\ d\mu$. In what follows we show how to efficiently compute such $s(x)$ for which S is at most $O(2^{2\alpha}(c + \beta k))$ (no matter what μ is). This is comparable to the (non-constructive) bounds we give on S in Section 3. Moreover, this suffices to prove Theorem 1.2. We now present our theorem for this section. We note that its proof closely resembles that of Theorem 3.1 in which we bounded the value of $s(x)$ (and thus S) for families $W(\mathcal{X}, k, \alpha)$.

THEOREM 5.1. *Let $F = W(\mathcal{X}, k, \alpha)$. Let $f^* \in W(\mathcal{X}, \beta k, \alpha)$ such that $\bar{f^*} \leq c \min_{f \in F} \bar{f}$. Given f^*, one can compute for all $x \in X$ a value $s(x) \geq \sigma(x)$ and a corresponding $q(x)$ in time $O(|X|d\beta k)$. The values of $s(x)$ computed will satisfy $S = \int_X s(x)\ d\mu \leq O(2^{2\alpha}(c + \beta k))$.*

Before we prove Theorem 5.1, a small remark is in place. If the norm ϱ corresponding to $\mathcal{X}$ is not well behaved, then naively following the flow of theorems in Section 4 one would need to *find* a well behaved approximation ϱ' to ϱ and use it in Theorem 5.1 above. However, this is not necessary. By multiplying the values of $s(x)$ that correspond to ϱ computed in Theorem 5.1 by $O((1+\varepsilon)^\alpha)$ (to compensate for the slight difference between ϱ and ϱ' stated in Theorem 4.2) we obtain values $s(x)$ (and thus a bound on S) that also correspond to ϱ'.

Proof. We concentrate on computing $s(x)$. We start with the case $\alpha \leq 1$ and then turn to the case $\alpha \geq 1$. Let $f^* = f_{u_1^*,\ldots,u_{\beta k}^*}$ be a function in $W(\mathcal{X}, \beta k, \alpha)$ for which $\bar{f^*} \leq c \min_{f \in W(\mathcal{X},k,\alpha)} \bar{f}$. Let $\overline{f^*} = \Delta$. Let U_i be the Voronoi cell of u_i^*, and let $p_i = \mu(U_i)$. Let $m_i = \frac{1}{p_i}\int_{U_i} \varrho(x - u_i^*)^\alpha d\mu(x)$, so $\Delta = \sum p_i m_i$. By a simple Markov inequality, for each i, $\mu(B(u_i^*, (2m_i)^{1/\alpha}) \cap U_i) \geq p_i/2$. (Here $B(x, r)$ denotes the closed ball of radius r about x.)

Case 1: $\alpha \leq 1$.

CLAIM 1. *Let $x \in U_i$. Setting*

$$s(x) = \frac{2c(2m_i + \varrho(x - u_i^*)^\alpha)}{\bar{f^*}} + \frac{4}{p_i}$$

satisfies $s(x) \geq \max_f \frac{f(x)}{\bar{f}}$.

Proof. Let $x \in U_i$, and let $f = f_{u_1,\ldots,u_k}$ be any function in $W(\mathcal{X}, k, \alpha)$. Let u denote a closest point to u_i^* in $\{u_1, \ldots, u_k\}$, and let $\varrho_i = \varrho(u - u_i^*)$. Then $\bar{f} \geq \int_{U_i} f d\mu \geq [\max(0, \varrho_i - (2m_i)^{1/\alpha})]^\alpha \frac{p_i}{2}$. Also, $\bar{f} \geq \Delta/c$. Thus, $\bar{f} \geq [\max(0, \varrho_i - (2m_i)^{1/\alpha})]^\alpha \frac{p_i}{4} + \bar{f^*}/2c$. We conclude that

$$\begin{aligned}
\max_f f(x)/\bar{f} &\leq \max_f \varrho(x - u)^\alpha/\bar{f} \\
&\leq \max_f \frac{(\varrho_i + \varrho(x - u_i^*))^\alpha}{[\max(0, \varrho_i - (2m_i)^{1/\alpha})]^\alpha \frac{p_i}{4} + \bar{f^*}/2c} \\
&\leq \max_{\varrho_i} \frac{\varrho_i^\alpha + \varrho(x - u_i^*)^\alpha}{[\max(0, \varrho_i - (2m_i)^{1/\alpha})]^\alpha \frac{p_i}{4} + \bar{f^*}/2c} \\
&\leq \max_{\varrho_i \geq (2m_i)^{1/\alpha}} \frac{\varrho_i^\alpha + \varrho(x - u_i^*)^\alpha}{\frac{(\varrho_i - (2m_i)^{1/\alpha})^\alpha p_i}{4} + \bar{f^*}/2c} \\
&\leq \max_{\varrho_i^\alpha \geq 2m_i} \frac{\varrho_i^\alpha + \varrho(x - u_i^*)^\alpha}{\frac{(\varrho_i^\alpha - 2m_i)p_i}{4} + \bar{f^*}/2c}
\end{aligned}$$

In the above we use the fact that (i) $|a + b|^\alpha \leq |a|^\alpha + |b|^\alpha$, and thus (ii) $|a - b|^\alpha \geq |a|^\alpha - |b|^\alpha$. Let $G(\varrho_i^\alpha) = \frac{4(\varrho_i^\alpha + \varrho(x - u_i^*)^\alpha)}{(\varrho_i^\alpha - 2m_i)p_i + 2\bar{f^*}/c}$. It is not hard to verify that $\mathrm{sign}\left(\frac{\partial G}{\partial \varrho_i^\alpha}\right)$ is independent of ϱ_i^α and thus G is monotone as a function of ϱ_i^α. We conclude that for $x \in U_i$, $\max_f \frac{f(x)}{\bar{f}}$ is at most

$$\max\left(\frac{2m_i + \varrho(x - u_i^*)^\alpha}{\bar{f^*}/2c}, \frac{4}{p_i}\right) \leq \frac{2m_i + \varrho(x - u_i^*)^\alpha}{\bar{f^*}/2c} + \frac{4}{p_i}$$

CLAIM 2. *Setting $s(x)$ according to Claim 1, it holds that $S = \int s d\mu \leq 6c + 4\beta k$.*

Proof.

$$\begin{aligned}
S &= \int s\ d\mu = \sum_i \int_{U_i} s\ d\mu \\
&\leq \sum_i \int_{U_i} \left(\frac{4cm_i + 2c\varrho(x - u_i^*)^\alpha}{\bar{f^*}} + \frac{4}{p_i}\right) d\mu \\
&= \sum_i \left(\frac{4cm_i p_i}{\bar{f^*}} + \frac{2cm_i p_i}{\bar{f^*}} + 4\right) \leq 6c + 4\beta k
\end{aligned}$$

Case 2: $\alpha \geq 1$**.** The proof below has a very similar structure to that above for $\alpha \leq 1$.

CLAIM 3. *Let $x \in U_i$. Setting*

$$s(x) = \frac{2^{2\alpha} c m_i + 2^{\alpha} c \varrho(x - u_i^*)^{\alpha}}{\bar{f}^*} + \frac{2^{2\alpha}}{p_i}$$

satisfies $s(x) \geq \max_f f(x)/\bar{f}$.

Proof. Let $x \in U_i$, and let $f = f_{u_1,\dots,u_k}$ be any function in $W(\mathcal{X}, k, \alpha)$. Let u denote a closest point to u_i^* in $\{u_1, \dots, u_k\}$, and let $\varrho_i = \varrho(u - u_i^*)$. It follows that $\bar{f} \geq \int_{U_i} f \; d\mu \geq [\max(0, \varrho_i - (2m_i)^{1/\alpha})]^{\alpha} \frac{p_i}{2}$. It also holds that $\bar{f} \geq \bar{f}^*/c$. Thus, as before, $\bar{f} \geq [\max(0, \varrho_i - (2m_i)^{1/\alpha})]^{\alpha} \frac{p_i}{4} + \bar{f}^*/2c$. We conclude that $\max_f f(x)/\bar{f}$ is at most

$$\begin{aligned}
&\max_f \varrho(x-u)^{\alpha}/\bar{f} \\
\leq\;& \max_f 2^{\alpha-1} \frac{(\varrho_i + \varrho(x - u_i^*))^{\alpha}}{\max(0, (\varrho_i - (2m_i)^{1/\alpha})^{\alpha}) \frac{p_i}{4} + \bar{f}^*/2c} \\
\leq\;& \max_{\varrho_i} 2^{\alpha-1} \frac{\varrho_i^{\alpha} + \varrho(x - u_i^*)^{\alpha}}{\max(0, (\varrho_i - (2m_i)^{1/\alpha})^{\alpha}) \frac{p_i}{4} + \bar{f}^*/2c} \\
\leq\;& \max_{\varrho_i \geq (2m_i)^{1/\alpha}} 2^{\alpha-1} \frac{\varrho_i^{\alpha} + \varrho(x - u_i^*)^{\alpha}}{\frac{(\varrho_i - (2m_i)^{1/\alpha})^{\alpha} p_i}{4} + \bar{f}^*/2c} \\
\leq\;& \max_{\varrho_i \geq (2m_i)^{1/\alpha}} 2^{\alpha-1} \frac{\varrho_i^{\alpha} + \varrho(x - u_i^*)^{\alpha}}{\frac{\max(0, (2^{1-\alpha}\varrho_i^{\alpha} - 2m_i)) p_i}{4} + \bar{f}^*/2c} \\
\leq\;& \max_{\varrho_i^{\alpha} \geq 2^{\alpha} m_i} 2^{\alpha-1} \frac{\varrho_i^{\alpha} + \varrho(x - u_i^*)^{\alpha}}{\frac{(2^{1-\alpha}\varrho_i^{\alpha} - 2m_i) p_i}{4} + \bar{f}^*/2c}
\end{aligned}$$

In the above we use the fact that (i) $|a+b|^{\alpha} \leq 2^{\alpha-1}(|a|^{\alpha} + |b|^{\alpha})$, and thus (ii) $|a-b|^{\alpha} \geq 2^{1-\alpha}|a|^{\alpha} + |b|^{\alpha}$. Let $G(\varrho_i^{\alpha}) = 2^{\alpha} \frac{\varrho_i^{\alpha} + \varrho(x - u_i^*)^{\alpha}}{(2^{-\alpha}\varrho_i^{\alpha} - m_i) p_i + \bar{f}^*/c}$. It is not hard to verify that sign $\left(\frac{\partial G}{\partial \varrho_i^{\alpha}}\right)$ is independent of ϱ_i^{α} and thus G is monotone as a function of ϱ_i^{α}. We conclude that for $x \in U_i$:

$$\begin{aligned}
\max_f f(x)/\bar{f} &\leq \max\left(\frac{2^{2\alpha} m_i + 2^{\alpha} \varrho(x - u_i^*)^{\alpha}}{\bar{f}^*/c}, \frac{2^{2\alpha}}{p_i}\right) \\
&\leq \frac{2^{2\alpha} m_i + 2^{\alpha} \varrho(x - u_i^*)^{\alpha}}{\bar{f}^*/c} + \frac{2^{2\alpha}}{p_i} = s(x)
\end{aligned}$$

CLAIM 4. *Setting $s(x)$ according to Claim 3, it holds that $S = \int s \; d\mu \leq 2^{2\alpha}(\beta k + c) + 2^{\alpha} c$*

Proof.

$$\begin{aligned}
S &= \int s \; d\mu = \sum_i \left(\int_{U_i} s \; d\mu\right) \\
&\leq \sum_i \left(\int_{U_i} \frac{2^{2\alpha} c m_i + 2^{\alpha} c \varrho(x - u_i^*)^{\alpha}}{\bar{f}^*} + \frac{2^{2\alpha}}{p_i}\right) d\mu \\
&= \sum_i \left(\frac{2^{2\alpha} c m_i p_i}{\bar{f}^*} + \frac{2^{\alpha} c m_i p_i}{\bar{f}^*} + 2^{2\alpha}\right) \\
&= 2^{2\alpha} c + 2^{\alpha} c + 2^{2\alpha} \beta k = 2^{2\alpha}(\beta k + c) + 2^{\alpha} c
\end{aligned}$$

In summary, given f^*, the computation of U_i, p_i, m_i, $s(x)$, $S = \int s \; d\mu$ and finally $q(x)$ take time $O(|X| d\beta k)$. (Recall that we assume oracle access to f, ϱ, and the distribution μ.) This concludes our proof.

6 Proof of Theorem 4.2

Let ϱ be the norm at hand. In what follows we present a proof for $\alpha = 1$. An analogous proof holds for general α. One may associate with ϱ a centrally symmetric convex set $C_{\varrho} \subset \mathbb{R}^d$ such that for each $x \in \mathbb{R}^d$ it holds that $\varrho(x) = \|x\|_{C_{\varrho}} = \inf\{r > 0 \mid \frac{x}{r} \in C_{\varrho}\}$. Let $\varepsilon > 0$. In a previous work of ours [27], certain approximations to convex sets C were studied. Namely, using a proof technique similar to that of Dudley for convex shape approximation by a polytope with few vertices [12], in Theorem 1 of [27] it is shown that every centrally symmetric convex set C has a corresponding centrally symmetric convex set $C' \subseteq C$ such that (a) C' is a polyhedral of *low* complexity, and (b) for any $x \in \mathbb{R}^d$ it holds that $(1-\varepsilon)\|x\|_{C'} \leq \|x\|_C \leq \|x\|_{C'}$. More specifically, it was shown in [27] that C' has at most $(cd/\sqrt{\varepsilon})^d$ $(d-1)$-dimensional facets. Here c is a sufficiently large universal constant.

Let ϱ' be the norm corresponding to C'. Let $f \in F$ and $x \in \mathbb{R}^d$. We denote by f_{ϱ} the function f computed with the norm ϱ and by $f_{\varrho'}$ the same function (i.e., with the same centers) computed via ϱ'. By the discussion above it holds that $(1-\varepsilon) f_{\varrho'}(x) \leq f_{\varrho}(x) \leq f_{\varrho'}(x)$. Thus, for sufficiently small $\varepsilon > 0$:

LEMMA 6.1. *Any ε-approximator for ϱ' will yield an $O(\varepsilon)$-approximator for the original ϱ.*

Proof. Let ν be an ε-approximator for a measure μ with respect to ϱ', it now holds that for every $f \in F$

$$\begin{aligned}
\int f_{\varrho} \; d\nu &\in (1 \pm \varepsilon) \int f_{\varrho'} \; d\nu \in (1 \pm 3\varepsilon) \int f_{\varrho'} \; d\mu \\
&\in (1 \pm 11\varepsilon) \int f_{\varrho} \; d\mu
\end{aligned}$$

We thus may assume w.l.o.g. that our norm ϱ is defined by a polyhedral C_{ϱ} of low complexity as described above. Consider a set $\{C_1, \dots, C_n\}$ of convex sets where each set C_i is equal to $r_i C_{\varrho} + v_i$. Here v_i is a vector in $\mathbb{R}^d$, r_i is a positive real, and $r_i C = \{x \mid \|x\|_{\varrho} \leq r_i\}$. The collection of sets C_i describes an arrangement in $\mathbb{R}^d$. In what follows we bound the complexity of this arrangement. To do so, we consider another, more complex arrangement, and present a bound on its complexity.

For each C_i, let H_i be the set of hyperplanes defining the boundary of C_i. Let $\Gamma = (cd/\sqrt{\varepsilon})^d$. It holds that $|H_i| \leq \Gamma$. Now, consider the arrangement describing the collection of sets H_i. It is known that the size of this arrangement is at most $(n\Gamma)^d$, e.g., [17]. As this arrangement is more complex than that of the C_i's we have that

LEMMA 6.2. *The arrangement of* $\{C_1, \ldots, C_n\}$ *has complexity at most* $(n\Gamma)^d = \left(n(cd/\sqrt{\varepsilon})^d\right)^d$.

Which suffices to prove our theorem.

7 Proof of Theorem 4.3

In what follows we present our proof for Theorem 4.3 when $\alpha \geq 1$. Our proof is analogous (with slight changes in parameters) for $\alpha \leq 1$. Our proof of Theorem 4.3 has two steps. In Step 1 we show the existence of a small set of functions G for which for any $f \in F$ there exists a constant c_f and a function $g \in G$ which *covers* f. Namely, for any $x \in A$:

$$\left|\frac{f(x)}{s(x)} - \frac{c_f g(x)}{s(x)}\right| \leq \frac{\varepsilon}{256S}\hat{f}. \tag{7.4}$$

We do not take G to be our ε-cover-code since G is not a subfamily of F. In Step 2 we define a mapping from G into F. We denote by f^g the function corresponding to G in this mapping. We show that any f covered by g will also be covered by f^g. The set $F' = \{f^g \mid g \in G\}$ is our final cover code. We start with the proof of Step 2.

7.1 Step 2.

Let G be a set of functions satisfying Condition 7.4. We partition the set F into $|G|$ disjoint sets, with F^g being the set of functions f covered by g. Let $f^g \in F^g$ be a function in F^g minimizing $\bar{f}/\hat{f}$. Namely, $\bar{f}^g/\hat{f}^g \leq \bar{f}/\hat{f}$ for all $f \in F^g$. If the minimum is not obtained, we may take f^g to satisfy $\bar{f}^g/\hat{f}^g \leq 2\bar{f}/\hat{f}$ for all $f \in F^g$ without changing the statement of our theorems.

LEMMA 7.1. *For any f covered by g and for any $x \in A$*

$$D_{A,x}(f, f^g) \leq \frac{\varepsilon}{64S}\left(1 + \frac{f(x)}{\hat{f}s(x)} + \frac{f^g(x)}{\hat{f}^g s(x)}\right)$$

The set $\{f^g \mid g \in G\}$ will correspond to F' in Theorem 4.3.

Proof. Let $f \in F^g$ and let c_f be as in condition 7.4. Then

$$|\hat{f} - c_f\hat{g}| \leq \nu_A\left(\left|\frac{f}{s} - \frac{c_f g}{s}\right|\right) \leq \frac{\varepsilon}{256S}\hat{f}$$

Since $256S(1 - \varepsilon/256S) \geq 128S$, it holds that

$$\left|\frac{f(x)}{s(x)} - \frac{c_f g(x)}{s(x)}\right| \leq \frac{\varepsilon}{128S}c_f\hat{g} \quad \text{for all } x \in A,$$

and thus

$$|\hat{f} - c_f\hat{g}| \leq \frac{\varepsilon}{128S}c_f\hat{g}.$$

Hence,

$$\begin{aligned}
D_{A,x}(f,g) &= \left|\frac{f(x)}{\hat{f}s(x)} - \frac{c_f g(x)}{c_f\hat{g}s(x)}\right| \\
&\leq \left|\frac{f(x)}{c_f\hat{g}s(x)} - \frac{c_f g(x)}{c_f\hat{g}s(x)}\right| + \left|\frac{f(x)}{\hat{f}s(x)} - \frac{f(x)}{c_f\hat{g}s(x)}\right| \\
&\leq \frac{\varepsilon}{128S} + \frac{f(x)}{s(x)}\left|\frac{1}{\hat{f}} - \frac{1}{c_f\hat{g}}\right| \\
&\leq \frac{\varepsilon}{128S} + \frac{f(x)}{s(x)}\left|\frac{\hat{f} - c_f\hat{g}}{c_f\hat{g}\hat{f}}\right| \\
&\leq \frac{\varepsilon}{128S} + \frac{f(x)}{s(x)}\left|\frac{\varepsilon c_f\hat{g}}{128Sc_f\hat{g}\hat{f}}\right| \\
&= \frac{\varepsilon}{128S}\left(1 + \frac{f(x)}{\hat{f}s(x)}\right)
\end{aligned}$$

As $f^g \in F^g$, the same holds for f^g, namely $D_{A,x}(f^g, g) \leq \frac{\varepsilon}{128S}\left(1 + \frac{f^g(x)}{\hat{f}^g s(x)}\right)$. Thus we conclude that

$$\begin{aligned}
D_{A,x}(f,f^g) &= \left|\frac{f(x)}{\hat{f}s(x)} - \frac{f^g(x)}{\hat{f}^g s(x)}\right| \\
&\leq \left|\frac{f(x)}{\hat{f}s(x)} - \frac{g(x)}{\hat{g}s(x)}\right| + \left|\frac{f^g(x)}{\hat{f}^g s(x)} - \frac{g(x)}{\hat{g}s(x)}\right| \\
&= D_{A,x}(f,g) + D_{A,x}(f^g,g) \\
&\leq \frac{\varepsilon}{128S}\left(1 + \frac{f(x)}{\hat{f}s(x)}\right) + \frac{\varepsilon}{128S}\left(1 + \frac{f^g(x)}{\hat{f}^g s(x)}\right) \\
&\leq \frac{\varepsilon}{64S}\left(1 + \frac{f(x)}{\hat{f}s(x)} + \frac{f^g(x)}{\hat{f}^g s(x)}\right)
\end{aligned}$$

7.2 Step 1.

Let F be $W(\mathcal{X}, k, \alpha)$. Let A be any subset of X of size a. We show there exists a set of functions G of size $\left[\left(\frac{Sa}{\varepsilon}\right)^{\Theta(\alpha^2+1)}\Gamma\right]^{2dk}$ satisfying Condition 7.4, i.e., for any $f \in F$ there is a constant $c_f > 0$ and a function $g \in G$ such that for all $x \in A$ $\left|\frac{f(x)}{s(x)} - \frac{c_f g(x)}{s(x)}\right| \leq \frac{\varepsilon}{256S}\nu_A(f/s)$.

The family G is designed by partitioning the space $\mathcal{X}$ in a certain manner. We start with some definitions. Let $Z = \sum_A 1/s(x)$, and $z(x) = 1/(s(x)Z)$. So z is a probability measure on A. Let $h \in F$ be a function satisfying $\hat{h} = \min_{f \in F}\hat{f}$. Let $v_1, \ldots, v_k$ be the centers of h, and let the Voronoi regions of these centers be $V_1, \ldots, V_k$. Namely, for $x \in V_i$, $h(x) = \varrho(x - v_i)^\alpha$.

Observe that for any $f \in F$ (in particular for h) it holds that for all $y \in A$,

$$\hat{f} = \nu_A(f(x)/s(x)) \geq \frac{f(y)}{as(y)} \tag{7.5}$$

and so $s(y) \geq \frac{f(y)}{a\hat{f}}$.

In what follows we assume that $Z(V_i) > 0$ for all i. (Otherwise, values of i for which $Z(V_i) = 0$ may be neglected in the computations to come.) Let h_i be defined as follows:

$$\begin{aligned} h_i &= \frac{1}{az(V_i)} \sum_{y \in V_i \cap A} \frac{h(y)}{s(y)} = \frac{1}{az(V_i)} \sum_{y \in V_i \cap A} \frac{\varrho(y - v_i)^\alpha}{s(y)} \\ &= \frac{Z}{a} \sum_{y \in V_i \cap A} \frac{\varrho(y - v_i)^\alpha z(y)}{z(V_i)} \end{aligned}$$

This implies that $\hat{h} = \sum_{i=1}^k z(V_i) h_i$. Observe that the average value (according to z) of h on V_i is

$$\frac{1}{z(V_i)} \sum_{y \in V_i \cap A} \varrho(y - v_i)^\alpha z(y) = \frac{ah_i}{Z}$$

Thus, at least half of $z(V_i)$ lies in the ball $B_\varrho(v_i, (2ah_i/Z)^{1/\alpha})$.

Let $s_i^{\min} = \min_{y \in V_i \cap A} s(y)$. Then

$$z(V_i) = \sum_{V_i \cap A} \frac{1}{s(x)Z} \geq \frac{1}{Z s_i^{\min}} \tag{7.6}$$

Let f be any function in F. Let the centers of f be $\{u_1, \ldots, u_k\}$. Let ϱ_i be the distance from v_i to the nearest center of f. We start by bounding $\hat{f}$ from above and below.

Lemma 7.2.

$$2^{\alpha-1} \left(\hat{h} + \sum_{i=1}^k \varrho_i^\alpha / s_i^{\min} \right) \geq \hat{f} \geq \frac{1}{4a2^{\alpha-1}} \sum_{i=1}^k \varrho_i^\alpha / s_i^{\min}$$

and thus

$$\forall i, \quad \hat{f} \geq \frac{\varrho_i^\alpha}{4a2^{\alpha-1} s_i^{\min}}$$

Proof. For a point $x \in A$ let u_x be the closest center in $\{u_1, \ldots, u_k\}$ to x. Suppose for a moment that for all i, $\varrho_i \geq (2ah_i/Z)^{1/\alpha}$. Let $B_i = V_i \cap B_\varrho(v_i, (2ah_i/Z)^{1/\alpha}) \cap A$. Then

$$\begin{aligned} \hat{f} &\geq \frac{1}{a} \sum_i \sum_{x \in B_i} \frac{f(x)}{s(x)} \\ &= \frac{1}{a} \sum_i \sum_{x \in B_i} \frac{\varrho(x - u_x)^\alpha}{s(x)} \\ &\geq \frac{1}{a} \sum_i \sum_{x \in B_i} \frac{|\varrho_i - (2ah_i/Z)^{1/\alpha}|^\alpha}{s(x)} \\ &= \frac{Z}{a} \sum_i \sum_{x \in B_i} |\varrho_i - (2ah_i/Z)^{1/\alpha}|^\alpha z(x) \\ &= \frac{Z}{a} \sum_i |\varrho_i - (2ah_i/Z)^{1/\alpha}|^\alpha \sum_{x \in B_i} z(x) \\ &\geq \frac{Z}{2a} \sum_i |\varrho_i - (2ah_i/Z)^{1/\alpha}|^\alpha z(V_i) \\ &\geq \frac{Z}{2a2^{\alpha-1}} \sum_i \varrho_i^\alpha z(V_i) - \sum_i h_i z(V_i) \\ &\geq \frac{1}{2a2^{\alpha-1}} \sum_i \varrho_i^\alpha / s_i^{\min} - \sum_i h_i z(V_i) \\ &= \frac{1}{2a2^{\alpha-1}} \sum_i \varrho_i^\alpha / s_i^{\min} - \hat{h} \end{aligned}$$

Here we used Equation 7.6. If for some i it holds that $\varrho_i \leq (2ah_i/Z)^{1/\alpha}$, then the inequalities above still hold as in this case due to the monotonicity of $|\cdot|^\alpha$ and the fact that $\alpha \geq 1$ we have that $\varrho_i^\alpha / 2^{\alpha-1} \leq \varrho_i^\alpha \leq 2ah_i/Z$, so it clearly holds that $\varrho(x - u_x)^\alpha \geq 0 \geq \varrho_i^\alpha / 2^{\alpha-1} - 2ah_i/Z$. Now, as h was chosen to minimize $\hat{f}$ for $f \in F$, it holds that $\hat{f} \geq \hat{h}$, which implies that $\hat{f} \geq \frac{1}{4a2^{\alpha-1}} \sum_i \varrho_i^\alpha / s_i^{\min}$. For the upper bound, we have by the triangle inequality that

$$\begin{aligned} \hat{f} &= \frac{1}{a} \sum_{x \in A} \frac{f(x)}{s(x)} = \frac{1}{a} \sum_{x \in A} \frac{\varrho(x - u_x)^\alpha}{s(x)} \\ &\leq \frac{1}{a} \sum_i \sum_{x \in V_i \cap A} \frac{(\varrho(x - v_i) + \varrho_i)^\alpha}{s(x)} \\ &\leq \frac{2^{\alpha-1}}{a} \sum_i \sum_{x \in V_i \cap A} \frac{\varrho(x - v_i)^\alpha}{s(x)} + \frac{2^{\alpha-1}}{a} \sum_i \sum_{x \in V_i \cap A} \frac{\varrho_i^\alpha}{s(x)} \\ &\leq 2^{\alpha-1} \hat{h} + 2^{\alpha-1} \sum_i \sum_{x \in V_i \cap A} \varrho_i^\alpha / a s_i^{\min} \\ &\leq 2^{\alpha-1} \left(\hat{h} + \sum_i \varrho_i^\alpha / s_i^{\min} \right) \end{aligned}$$

Here we use the fact that (by definition) $s_i^{\min} \leq s(x)$ for all $x \in V_i \cap A$.

We will now define a set of points in X that will act as potential centers for functions $g \in G$. In what follows we will use some parameters p_1, p_2, p_3 and p_4 to be defined at the end of the proof. For each point $x \in A$ and for $i = 1, \ldots, (ap_1)^2$, let

$$R_{x,i} = \left\{ v \in X \mid \frac{\varrho(v-x)^\alpha}{s(x)} \in \left[\frac{(i-1)\hat{h}}{ap_1}, \frac{i\hat{h}}{ap_1} \right) \right\}$$

The sets $R_{x,i}$ and their intersections form an arrangement $\mathcal{A}$ of *cells* in X. Let N be a set of points with at least one representative in each cell of $\mathcal{A}$. As we are assuming that ϱ is Γ-well behaved, we have that $|N| \leq (ap_1\Gamma)^{2d}$. (Here, to simplify our notation, the bound presented is not tight.)

CLAIM 5. *Let $f \in F$. Let U be a cell in $\mathcal{A}$, and let n be a corresponding representative of N in U, then for any $v \in U$ it holds that*

$$|\varrho(x-v)^\alpha - \varrho(x-n)^\alpha| \leq \frac{\hat{h}s(x)}{ap_1} \leq \frac{\hat{f}s(x)}{ap_1}$$

Proof. Follows directly by the definition of $\mathcal{A}$ and the family $R_{x,i}$.

We are now ready to define our function g that will approximate f. We consider two cases:

Case A: In this case we assume $\hat{f} \leq p_1\hat{h}$. We also assume w.l.o.g. that each center u_i of f is of *significance* in the sense that it is the closest center of f to some point $x \in A$. Otherwise, set the insignificant centers of f to one of the significant centers. This will not change the value of f at all—and the *new* f can be used in the analysis below.

Consider a center u_i of f and let x be a point in A for which u_i is the closest center of f to x. It holds by Equation 7.5 that

$$\frac{f(x)}{s(x)} \leq a\hat{f} \leq ap_1\hat{h}$$

We conclude that u_i is in $\cup_i R_{x,i}$, which implies that u_i is in some cell of $\mathcal{A}$. Let n_i be the representative point in the cell of u_i. We define g to be the function in F with centers $n_1, \ldots, n_k$.

We now show that g satisfies our requirements with $c_f = 1$. Namely, let $x \in A$ and consider any center u_i of f and its corresponding center n_i of g. As u_i and n_i are in the same cell of $\mathcal{A}$ it holds by Claim 5 that $|\varrho(x-u_i)^\alpha - \varrho(x-n_i)^\alpha| \leq \hat{h}s(x)/(ap_1) \leq \hat{f}s(x)/(ap_1)$. To bound $|f(x) - g(x)|$ assume that the closest center of f to x is u_i and the closest center of g to x is n_j.

It now holds that $|f(x) - g(x)|$ is equal to

$$\begin{aligned}
& |\varrho(x-u_i)^\alpha - \varrho(x-n_j)^\alpha| \\
\leq\ & |\varrho(x-u_i)^\alpha - \varrho(x-n_i)^\alpha| + |\varrho(x-n_i)^\alpha - \varrho(x-n_j)^\alpha| \\
\leq\ & \hat{f}s(x)/(ap_1) + |\varrho(x-n_i)^\alpha - \varrho(x-n_j)^\alpha|
\end{aligned}$$

Now, by Claim 5 and the fact that n_j (u_i) is the closest center of g (f) to x it holds that

$$\begin{aligned}
\varrho(x-n_i)^\alpha - 2\hat{f}s(x)/(ap_1) &\leq \varrho(x-u_i)^\alpha - \hat{f}s(x)/(ap_1) \\
&\leq \varrho(x-u_j)^\alpha - \hat{f}s(x)/(ap_1) \\
&\leq \varrho(x-n_j)^\alpha \leq \varrho(x-n_i)^\alpha
\end{aligned}$$

Thus,

$$|\varrho(x-n_i)^\alpha - \varrho(x-n_j)^\alpha| \leq 2\hat{f}s(x)/(ap_1)$$

We conclude that, $|f(x) - g(x)| \leq 3\hat{f}s(x)/(ap_1)$. Thus, setting $ap_1 \geq 768S/\varepsilon$, we obtain $|f(x) - g(x)| \leq \frac{\varepsilon}{256S}s(x)\hat{f}$. Note that the number of different functions g that we may receive in this case is at most $|N|^k = (ap_1\Gamma)^{2dk}$.

Case B: We now consider the case in which $\hat{f} \geq p_1\hat{h}$. In this case the function g we construct will be constant on each Voronoi region V_i. The values g on a given Voronoi region will be one of $O(p_2\log(p_2))$ different values to be specified shortly. This will imply that the number of different functions g specified in this case is at most this number to the power of k.

We now define g: For a parameter p_3 and i in which $\varrho_i^\alpha/s_i^{\min} \leq \hat{f}/p_3$ we define g to be 0 on V_i. We refer to such indices i as *light* indices, other indices are referred to as *heavy*. Loosely speaking, for i in which $\varrho_i^\alpha/s_i^{\min} \geq \hat{f}/p_3$ we define g to be ϱ_i^α/c_f on V_i, where c_f is defined as follows:

$$c_f = \max_{i:\varrho_i^\alpha/s_i^{\min} \geq \hat{f}/p_3} \frac{\varrho_i^\alpha}{p_2}$$

To be more precise, for i in which $\varrho_i^\alpha/s_i^{\min} \geq \hat{f}/p_3$ we define g to be the nearest value to ϱ_i^α/c_f in the set

$$\left\{ 0, \left(1 + \frac{1}{p_2}\right), \left(1 + \frac{1}{p_2}\right)^2, \ldots, p_2 \right\}$$

Notice that by definition, ϱ_i^α/c_f is at most p_2. Also notice that for heavy indices i, $c_f g$ obtains a value of approximately ϱ_i^α. More precisely the value of $c_f g$ on the heavy indices i is in the range $\left[\varrho_i^\alpha\left(1 - \frac{1}{p_2}\right), \varrho_i^\alpha\left(1 + \frac{1}{p_2}\right)\right]$. We now study $|f(x) - c_f g(x)|$ on light and heavy indices i.

Case B1: Light indices i for which $\varrho_i^\alpha/s_i^{\min} \leq \hat{f}/p_3$. Here, for $x \in V_i$, $g_x = 0$ and

$$\begin{aligned}
|f(x) - c_f g(x)| &= |f(x)| \leq (\varrho(x-v_i) + \varrho_i)^\alpha \\
&\leq 2^{\alpha-1}(\varrho(x-v_i)^\alpha + \varrho_i^\alpha) \\
&= 2^{\alpha-1}(h(x) + \varrho_i^\alpha) \\
&\leq 2^{\alpha-1}(a\hat{h}s(x) + s_i^{\min}\hat{f}/p_3) \\
&\leq 2^{\alpha-1}(as(x)\hat{f}/p_1 + s(x)\hat{f}/p_3)
\end{aligned}$$

Here, we used Equation 7.5. We will set p_1 and p_3 such that $a/p_1 + 1/p_3 \leq \frac{\varepsilon}{256S2^{\alpha-1}}$ which implies in this case $|f(x) - c_f g(x)| \leq \frac{\varepsilon}{256S}s(x)\hat{f}$.

Case B2: Heavy indices i for which $\varrho_i^\alpha / s_i^{\min} \geq \hat{f}/p_3$. Here we consider two sub-cases. First consider $x \in V_i$ for which $\varrho(x - v_i) \leq \varrho_i/p_4$ (here $p_4 \geq 1$ will be specified later). As the closest center of f to v_i is at distance ϱ_i from v_i, for such x it holds that $f(x) \in [(\varrho_i - \varrho_i/p_4)^\alpha, (\varrho_i + \varrho_i/p_4)^\alpha]$. Thus,

$$\begin{aligned}
|f(x) - c_f g(x)| &\leq |f(x) - \varrho_i^\alpha| + \frac{\varrho_i^\alpha}{p_2} \\
&\leq |(\varrho_i + \varrho_i/p_4)^\alpha - \varrho_i^\alpha| + \frac{\varrho_i^\alpha}{p_2} \\
&\leq \varrho_i^\alpha \left(\frac{2\alpha}{p_4} + \frac{1}{p_2}\right) \\
&\leq 4a2^{\alpha-1}s_i^{\min}\hat{f}\left(\frac{2\alpha}{p_4} + \frac{1}{p_2}\right)
\end{aligned}$$

Here (in the last inequality) we used the second statement in Lemma 7.2. Also, for $p_4 \geq 2\alpha$ notice that $(1 + 1/p_4)^\alpha \leq 1 + 2\alpha/p_4$. We will set p_4 and p_2 such that $\left(\frac{2\alpha}{p_4} + \frac{1}{p_2}\right) \leq \frac{\varepsilon}{1024Sa2^{\alpha-1}}$ which will imply in this case $|f(x) - c_f g(x)| \leq \frac{\varepsilon}{256S}s_i^{\min}\hat{f} \leq \frac{\varepsilon}{256S}s(x)\hat{f}$.

Now consider $x \in V_i$ for which $\varrho(x - v_i) \geq \varrho_i/p_4$. In this case, by the triangle inequality $f(x) \in [0, (\varrho_i + \varrho(x - v_i))^\alpha]$. Also, recall by Equation 7.5 that $h(x) = \varrho(x - v_i)^\alpha \leq as(x)\hat{h}$. Now,

$$\begin{aligned}
|f(x) - c_f g(x)| &\leq |f(x) - \varrho_i^\alpha| + \frac{\varrho_i^\alpha}{p_2} \\
&\leq (\varrho_i + \varrho(x - v_i))^\alpha + \varrho_i^\alpha + \frac{\varrho_i^\alpha}{p_2} \\
&\leq 2^\alpha(\varrho_i^\alpha + \varrho(x - v_i)^\alpha) + \frac{\varrho_i^\alpha}{p_2} \\
&\leq \varrho(x - v_i)^\alpha \left(\frac{p_4^\alpha}{p_2} + (2p_4)^\alpha + 2^\alpha\right) \\
&\leq a\hat{h}s(x)\left(\frac{p_4^\alpha}{p_2} + (2p_4)^\alpha + 2^\alpha\right) \\
&\leq \frac{a\hat{f}s(x)}{p_1}\left(\frac{p_4^\alpha}{p_2} + (2p_4)^\alpha + 2^\alpha\right)
\end{aligned}$$

We will set p_1, p_2 and p_4 such that $\frac{a}{p_1}\left(\frac{p_4^\alpha}{p_2} + (2p_4)^\alpha + 2^\alpha\right) \leq \frac{\varepsilon}{256S}$ which will imply in this case $|f(x) - c_f g(x)| \leq \frac{\varepsilon}{256S}s(x)\hat{f}$.

To summarize, one can set our parameters such that all the requirements stated above hold: $p_1 = \left(\frac{cSa}{\varepsilon}\right)^{\Theta(\alpha^2+1)}$; $p_2 = \frac{1024Sa2^\alpha}{\varepsilon}$; $p_3 = \frac{512S2^{\alpha-1}}{\varepsilon}$; and $p_4 = \frac{4096Sa2^{\alpha-1}\alpha}{\varepsilon}$. Here c is a sufficiently large constant.

The total size of $|G| = |F'|$ is thus

$$(cp_2 \log(p_2))^k + (ap_1\Gamma)^{2dk} = \left[\left(\frac{Sa}{\varepsilon}\right)^{\Theta(\alpha^2+1)}\Gamma\right]^{2dk}.$$

8 Proof of Theorem 4.4

Fix an integer $a \geq 8(S-1)/\varepsilon^2$. Let E_0 be the event that a sample R of size a according to the distribution q is not an ε-approximator for F. (If this occurs write $R \in E_0$.) Throughout this section R is treated as a multiset. Say that (f, R) is *ε-bad* if $\left|\bar{f} - \nu_R(Sf/s)\right| > \varepsilon\bar{f}$. (Recall, for a function g, that $\nu_R(g) = (1/|R|)\sum_{x\in R} g(x)$.) So, E_0 occurs if there is an $f \in F$ s.t. (f, R) is *ε-bad*. We shall upper-bound $\Pr[E_0]$. We start by recalling that if R chosen as above, then for any given function $f \in F$, there is only a small probability that (f, R) is ε-bad. Namely, by Lemma 2.1 we have that for $\varepsilon > 0$ and $f \in F$, if R is a random sample of X of size $a \geq \frac{2(S-1)}{\varepsilon^2}$ according to the distribution q then $\Pr[(f, R)$ is ε-bad$] \leq 1/2$.

In the manner of Vapnik and Chervonenkis, now let G be an additional multiset of size a chosen independently at random according to q. Let E_1 be the event "$\exists f \in F :$ (f, R) is ε-bad and (f, G) is $\varepsilon/2$-good" (here good is the complement of bad). In order to upper-bound $\Pr[E_0]$ we relate E_0 and E_1 as follows:

Claim 6. $\Pr[E_1] \leq \Pr[E_0] \leq 2\Pr[E_1]$.

Proof. The first inequality is trivial; for the second, we condition on R. If R has no bad functions f (namely E_0 does not happen), then $\Pr[E_1|R] = 0$. If R has bad functions, pick one and denote it f_R. Now $\Pr[E_1|R] \geq \Pr[(f_R, G)$ is $\varepsilon/2$-good$]$. By Lemma 2.1, the latter happens with probability at least $1/2$ over the set G. So, $\Pr[E_1] = \sum_{R\in E_0}\Pr[E_1|R]\Pr[R] \geq \frac{1}{2}\sum_{R\in E_0}\Pr[R] = \frac{1}{2}\Pr[E_0]$.

We now bound $\Pr[E_1]$ from above. Let A be an independent random sample of size $2a$ according to the distribution q. Let R be a random sample from A of size a (without replacement, treating all elements of the multiset as distinct) and let $G = A \setminus R$. Notice that the distribution of R and G are identical to that of R and G discussed above (namely, they both contain a independent random samples from X according to q). We now condition on the multiset $A = R \cup G$ and show that the event "$E_1 \mid A$" happens with low probability (no matter what A is). To do so we analyze an event implied by "$E_1 \mid A$" that is easier to analyze.

Specifically, it holds that $\Pr[E_1]$ is at most

$$\begin{aligned}
& \sup_A \Pr[E_1 \mid A] \\
= & \sup_A \Pr[\exists f \in F: \ (f,R) \text{ is } \varepsilon\text{-bad and} \\
& \qquad\qquad (f,G) \text{ is } \varepsilon/2\text{-good} \mid A] \\
\leq & \sup_A \Pr\left[\exists f \in F: |\nu_R(Sf/s) - \nu_G(Sf/s)| \geq \frac{\varepsilon}{2}\bar{f} \mid A\right] \\
= & \sup_A \Pr\left[\exists f \in F: |\nu_{R\cup G}(Sf/s) - \nu_G(Sf/s)| \geq \frac{\varepsilon}{4}\bar{f} \mid A\right]
\end{aligned}$$

For any specific $f \in F$, the event in brackets is unlikely:

Claim 7. *Let $\delta > 0$. Let $f \in F$ and let A be any multiset of size $2a$ in X. Let R be a random (uniform) sample from A of size a and let $G = A \setminus R$. Then*

$$\Pr[|\nu_A(Sf/s) - \nu_G(Sf/s)| \geq \delta\bar{f}] \leq 2e^{-\frac{a\delta^2}{S^2}}$$

Proof. We use the following auxiliary lemma for our proof:

Lemma 8.1. *([21]) Let $h(\cdot)$ be a function defined on a set A, such that for $x \in A$ we have $h(x) \in [0, \mathbf{max}]$. Let G be a multiset of a samples drawn independently and identically from A, and let $\delta > 0$ be a parameter. If $a \geq \mathbf{max}^2/\gamma^2 \ln(2/\varepsilon)$, then $\Pr[|\nu_A(h) - \nu_G(h)| \geq \gamma] \leq \varepsilon$*

The proof of Claim 7 now follows directly for any a by setting $h = Sf/s$, $\gamma = \delta\bar{f}$, $\varepsilon = 2e^{-\frac{a\delta^2}{S^2}}$ and noticing that $h = Sf/s \in [0, S\bar{f}]$.

At this point we need to take the key step around the union bound in

$$\Pr\left[\exists f \in F: |\nu_{R\cup G}(Sf/s) - \nu_G(Sf/s)| \geq \frac{\varepsilon}{4}\bar{f} \mid A\right].$$

In the classic VC argument for binary functions this step is trivial, but in our case it is actually what pivotally defines the definition of "ε-cover-code." The remainder of the proof is devoted to this step.

By our assumption, F contains an ε cover code F' for (F, A, s) of size at most $\frac{1}{8}e^{\frac{a\varepsilon^2}{100S^2}}$. To prove Theorem 4.4 we use Claim 7 with a union bound over F'. Namely, as F' is of size at most $\frac{1}{8}e^{\frac{a\varepsilon^2}{100S^2}}$, Claim 7 implies that with probability $1 - \frac{1}{8}e^{\frac{a\varepsilon^2}{100S^2}} \cdot 2e^{-\frac{a\varepsilon^2}{100S^2}} \geq 3/4$ for all f' in F' it holds that $|\nu_A(Sf'/s) - \nu_G(Sf'/s)| < \frac{\varepsilon}{10}\bar{f}'$.

We now show that if indeed $|\nu_A(Sf'/s) - \nu_G(Sf'/s)| < \frac{\varepsilon}{10}\bar{f}'$ for all $f' \in F'$, then $|\nu_A(Sf/s) - \nu_G(Sf/s)| < \frac{\varepsilon}{4}\bar{f}$ for all $f \in F$. This will essentially conclude the proof of the theorem. Let $f \in F$ and let $f' \in F'$ be a function covering it. Then:

$$\begin{aligned}
& |\nu_A(Sf/s) - \nu_G(Sf/s)| \\
= & \ \nu_A(f/s)\left|\nu_A\left(\frac{Sf}{\nu_A(f/s)s}\right) - \nu_G\left(\frac{Sf}{\nu_A(f/s)s}\right)\right| \\
\leq & \ \nu_A(f/s)\left|\nu_A\left(\frac{Sf}{\nu_A(f/s)s}\right) - \nu_A\left(\frac{Sf'}{\nu_A(f'/s)s}\right)\right| \\
& +\nu_A(f/s)\left|\nu_A\left(\frac{Sf'}{\nu_A(f'/s)s}\right) - \nu_G\left(\frac{Sf'}{\nu_A(f'/s)s}\right)\right| \\
& +\nu_A(f/s)\left|\nu_G\left(\frac{Sf}{\nu_A(f/s)s}\right) - \nu_G\left(\frac{Sf'}{\nu_A(f'/s)s}\right)\right| \\
\leq & \ \nu_A(Sf/s) \cdot \nu_A\left(\left|\frac{f}{\nu_A(f/s)s} - \frac{f'}{\nu_A(f'/s)s}\right|\right) \\
& +\frac{\varepsilon}{10}\frac{\bar{f}'}{\nu_A(f'/s)}\nu_A(f/s) \\
& +\nu_A(Sf/s) \cdot \nu_G\left(\left|\frac{f}{\nu_A(f/s)s} - \frac{f'}{\nu_A(f'/s)s}\right|\right) \\
\leq & \ \frac{\varepsilon}{10}\frac{\bar{f}}{\nu_A(f/s)}\nu_A(f/s) + \nu_A(Sf/s)\nu_A(D_{A,x}(f,f')) \\
& +\nu_A(Sf/s)\nu_G(D_{A,x}(f,f')) \\
\leq & \ \frac{\varepsilon}{10}\bar{f} \\
& +\frac{\varepsilon\nu_A(f/s)}{64}\left[\nu_A\left(1 + \frac{f}{\nu_A(f/s)s} + \frac{f'}{\nu_A(f'/s)s}\right)\right] \\
& +\frac{\varepsilon\nu_A(f/s)}{64}\left[\nu_G\left(1 + \frac{f}{\nu_A(f/s)s} + \frac{f'}{\nu_A(f'/s)s}\right)\right] \\
\leq & \ \frac{\varepsilon}{10}\bar{f} + \frac{\varepsilon}{32}\nu_A(f/s) \\
& +\frac{\varepsilon\nu_A(f/s)}{64}\left[\nu_A\left(\frac{f}{\nu_A(f/s)s} + \frac{f'}{\nu_A(f'/s)s}\right)\right] \\
& +\frac{\varepsilon\nu_A(f/s)}{64}\left[2\nu_A\left(\frac{f}{\nu_A(f/s)s} + \frac{f'}{\nu_A(f'/s)s}\right)\right] \\
\leq & \ \frac{\varepsilon}{10}\bar{f} + \frac{\varepsilon}{32}\nu_A(f/s) + \frac{6\varepsilon}{64}\nu_A(f/s) < \frac{\varepsilon}{4}\bar{f}
\end{aligned}$$

In the last inequality we used $\nu_A(f/s) \leq \bar{f}$ (which follows from the inequality $f(x) \leq s(x)\bar{f}$ for all $f \in F$ and $x \in X$). Thus, $\Pr[E_0]$ is at most $2\sup_A \Pr\left[\exists f \in f: |\nu_A(Sf/s) - \nu_G(Sf/s)| \geq \frac{\varepsilon}{4}\bar{f} \mid A\right] \leq 1/2$.

References

[1] Pankaj Agarwal, Sariel Har-Peled, and Kasturi R. Varadarajan. Geometric approximation via coresets. In *Current Trends in Combinatorial and Computational Geometry*. Cambridge University Press, 2005.

[2] N. Alon, S. Ben-David, N. Cesa-Bianchi, and D. Haussler. Scale-sensitive dimensions, uniform convergence, and learnability. *J. ACM*, 44(4):615–631, July 1997.

[3] N. Alon and B. Sudakov. On two segmentation problems. *J. of Algorithms*, 33:173–184, 1999.

[4] P. L. Bartlett and P. M. Long. Prediction, learning, uniform convergence, and scale-sensitive dimensions. *J. Comput. System Sci.*, 56:174–190, 1998.

[5] P. L. Bartlett, P. M. Long, and R. C. Williamson. Fat-shattering and the learnability of real-valued functions. *J. Comput. System Sci.*, 52:434–452, 1996.

[6] S. Ben-David, N. Cesa-Bianchi, D. Haussler, and P. M. Long. Characterizations of learnability for classes of $\{0, \ldots, n\}$-valued functions. *J. of Comput. Syst. Sci.*, 50(1):74–86, 1995.

[7] S. Ben-David, N. Cesa-Bianchi, and P. M. Long. Characterizations of learnability for classes of {0, ..., n}-valued functions. In *Proc. 5'th Annual ACM COLT*, pages 333–340, 1992.

[8] A. Blumer, A. Ehrenfeucht, D. Haussler, and M. K. Warmuth. Learnability and the Vapnik-Chervonenkis dimension. *J. ACM*, 36(4):929–965, October 1989.

[9] B. Chazelle. *The Discrepancy Method.* Cambridge U. Press, 2000.

[10] K. Chen. On coresets for k-median and k-means clustering in metric and Euclidean spaces and their applications. In *Proc. 17th SODA*, pages 1177–1185, 2006.

[11] W. Fernandez de la Vega and C. Kenyon. A randomized approximation scheme for metric max-cut. In *39th Ann. Symp. on Foundations of Computer Science*, pages 468–471. IEEE, 1998.

[12] R. M. Dudley. Metric entropy of some classes of sets with differentiable boundaries. *J. Approximation Theory*, 10(3):227–236, 1974.

[13] R. M. Dudley, E. Giné, and J. Zinn. Uniform and universal Glivenko-Cantelli classes. *J. Theoretical Probability*, 4(3):485–510, 1991.

[14] M. Effros and L. J. Schulman. Deterministic clustering with data nets. `http://eccc.hpi-web.de/eccc-reports/2004/TR04-050/index.html`, 2004. Technical Report TR04-085, Elec. Colloq. Comp. Complexity, 2004.

[15] D. Feldman, M. Monemizadeh, and C. Sohler. A PTAS for k-means clustering based on weak coresets. In *SOCG '07: Proceedings of the twenty-third annual symposium on Computational geometry*, pages 11–18, 2007.

[16] B. Grünbaum. Venn diagrams and independent families of sets. *Mathematics Magazine*, 48(1):12–23, 1975.

[17] D. Halperin. Handbook of Discrete and Computational Geometry, J. E. Goodman and J. O'Rourke (Editors), Ch. 24: Arrangments. *Chapman and Hall/CRC; 2nd edition*, 2004.

[18] S. Har-Peled. Coresets for Discrete Integration and Clustering. *In proceedings of FSTTCS*, pages 33–44, 2006.

[19] S. Har-Peled and A. Kushal. Smaller coresets for k-median and k-means clustering. In *SOCG '05: Proceedings of the twenty-first annual symposium on Computational geometry*, pages 126–134, New York, NY, USA, 2005. ACM.

[20] S. Har-Peled and S. Mazumdar. Coresets for k-means and k-median clustering and their applications. In *Proc. 36th STOC*, pages 291–300, 2004.

[21] D. Haussler. Decision theoretic generalizations of the pac model for neural net and other learning applications. *Inf. Comput.*, 100(1):78–150, 1992.

[22] D. Haussler and E. Welzl. Epsilon-nets and simplex range queries. *Discrete Comput. Geom.*, 2:127–151, 1987.

[23] D. Hilbert. Uber die Darstellung definiter Formen als Summe von Formenquadraten. *Math. Ann.*, 32:342–350, 1888. Ges Abh. 2:415-436, Chelsea Publishing Co., New York, 1965.

[24] M. J. Kearns and R. E. Schapire. Efficient distribution-free learning of probabilistic concepts. *J. Comput. System Sci.*, 48:464–497, 1994.

[25] J. Kleinberg, C. Papadimitriou, and P. Raghavan. Approximation algorithms for segmentation problems. In *Proc. 30th STOC*, 1998.

[26] R. Kress. *Numerical Analysis.* Springer, 1998.

[27] M. Langberg and L. J. Schulman. Contraction and expansion of convex sets. *In proceedings of CCCG*, pages 25–28, 2004.

[28] J. Matoušek. *Lectures on Discrete Geometry.* Springer, 2002.

[29] D. Q. Naiman and H. P. Wynn. Independent Collections of Translates of Boxes and a Conjecture due to Grünbaum. *Discrete and Computational Geometry*, 9:101–105, 1993.

[30] M. Naszódi and S. Taschuk. On the Transversal Number and VC-Dimension of Families of Positive Homothets of a Convex Body. *Discrete Mathematics.* To appear. Also http://arxiv.org/abs/0907.5223.

[31] B. K. Natarajan. On learning sets and functions. *Machine Learning*, 4:67–97, 1989.

[32] D. Pollard. *Convergence of stochastic processes.* Springer, 1984.

[33] N. Sauer. On the density of families of sets. *J. Combinatorial Theory Ser. A*, 13:145–147, 1972.

[34] L. J. Schulman. Clustering for edge-cost minimization. In *Proc. 32nd STOC*, pages 547–555, 2000.

[35] S. Shelah. A combinatorial problem, stability and order for models and theories in infinitary languages. *Pacific J. Math.*, 41:247–261, 1972.

[36] V. N. Vapnik. Inductive principles of the search for empirical dependences. In *Proc. 2nd Annual Workshop on Computational Learning Theory*, 1989.

[37] V. N. Vapnik and A. Y. Chervonenkis. On the uniform convergence of relative frequencies of events to their probabilities. *Theory of Probability and its Applications*, 16(2):264–280, 1971.

[38] Wikipedia. Sparse grid. `http://en.wikipedia.org/wiki/Sparse_grid`. [Online; version of April 21, 2009].

A Proof of Proposition 2

Proof. For convenience write $G = \mathbb{P}F^c$. Let $\delta > 0$ be arbitrary. If $\mathfrak{S}(G) < \infty$, let $t = e^{-\delta}\mathfrak{S}(G)$. If $\mathfrak{S}(G) = \infty$, let $t = 1/\delta$. Let μ be a measure for which

$\int \sigma_{G,\mu}(x)\ d\mu(x) \geq t$. We show there is a μ' for which $\int \sigma_{F,\mu'}(x)\ d\mu'(x) \geq e^{-3\delta}t$.

At each point x let $g^x \in G$ be a function for which $g^x(x)/\int g^x(y)\ d\mu(y) \geq e^{-\delta}\sigma_{G,\mu}(x)$. Let $A \subseteq X$ be a bounded region sufficiently large that $\int_A \sigma_{G,\mu}(x)\ d\mu(x) \geq e^{-\delta}t$. Let μ' be the probability measure defined by $\mu'(S) = \mu(S \cap A)/\mu(A)$. For each $x \in A$ let $f^x \in F$, $c^x > 0$ be such that $c^x f^x(y) \leq g^x(y) \leq e^\delta c^x f^x(y)$ for all $y \in A$. Then:

$$\begin{aligned}
\sigma_{F,\mu'}(x) &\geq \frac{f^x(x)}{\int f^x(y)\ d\mu'(y)} \geq \frac{e^{-\delta}g^x(x)/c^x}{\int g^x(y)/c^x\ d\mu'(y)} \\
&= \frac{e^{-\delta}g^x(x)}{\int g^x(y)\ d\mu'(y)} = \frac{\mu(A)e^{-\delta}g^x(x)}{\int_A g^x(y)\ d\mu(y)} \\
&\geq \frac{\mu(A)e^{-\delta}g^x(x)}{\int g^x(y)\ d\mu(y)} \geq \mu(A)e^{-2\delta}\sigma_{G,\mu}(x).
\end{aligned}$$

So: $\mathfrak{S}(F) = \int \sigma_{F,\mu'}(x)\ d\mu'(x) \geq \mu(A)e^{-2\delta}\int \sigma_{G,\mu}(x)\ d\mu'(x) = e^{-2\delta}\int_A \sigma_{G,\mu}(x)\ d\mu(x) \geq e^{-3\delta}t$. This suffices to prove our assertion as by definition $\mathfrak{S}(F) \leq \mathfrak{S}(G)$.

B Proof of Proposition 3

Let the space X be a discrete set $\{a_1, \ldots, a_n\}$. Let F be a family of n binary functions $f^1, \ldots, f^n$ each selected independently at random, by picking each $f^i(a_j) \in \{0, 1\}$ independently and uniformly at random. Let G be a family of n binary functions $g^1, \ldots, g^n$ defined as follows: for $j \leq i$, $g^i(a_j) = f^i(a_j)$; for $j > i$, $g^i(a_j) = 1 - f^i(a_j)$. Observe that G is identically distributed to F (but of course not independent of it).

Lemma B.1. *With probability* $1 - o(1)$*:* $\mathfrak{S}(F \cdot G) \in \Omega(n)$.

Proof. Let $\mu(a_j) = 2^{j-1-n}$ (except that $\mu(a_1) = 2^{1-n}$). With probability at least $1 - o(1)$ there are $\Omega(n)$ values of i for which $f^i(a_i)g^i(a_i) = 1$. For each such i, $s(a_i)\mu(a_i) \geq 1/2$.

Lemma B.2. *With probability* $1 - o(1)$*:* $\mathfrak{S}(F), \mathfrak{S}(G) \in O(\log n)$.

(This bound is optimal up to constants, because with high probability, there are $\Omega(\log n)$ points at each of which there is a function in F which is 1 there, but 0 on the others.)

Proof. It is enough to argue for F.

For functions f, f', write $(f \vee f')(a_i) = \max\{f(a_i), f'(a_i)\}$. Say that f dominates f', written $f \geq f'$, if $f(a_i) \geq f'(a_i)$ for all i. Write $f > f'$ if in addition there is an i for which $f(a_i) > f'(a_i)$.

For binary-valued families of functions, a function f optimizes $s(a_i)$ if and only if (a) $f(a_i) = 1$; (b) among functions satisfying (a), $\sum_j \mu(a_j)f(a_j)$ is minimal.

Consequently, there is a permutation π and a $1 \leq k \leq n$ such that $f^{\pi(1)}, \ldots, f^{\pi(k)} \in F$ satisfy:

(1) $f^{\pi(1)}$ is a function in F minimizing $\sum_j \mu(a_j)f(a_j)$. It is used to optimize all $s(a_i)$ for which $f^{\pi(1)}(a_i) = 1$.

(2) $f^{\pi(2)}$ is a function in F which (a) is not dominated by $f^{\pi(1)}$; (b) among functions satisfying (a), minimizes $\sum_j \mu(a_j)f(a_j)$. It is used to optimize all $s(a_i)$ for which $f^{\pi(2)}(a_i) = 1$ and which were not already optimized by $f^{\pi(1)}$.

(3) In general for $2 \leq \ell \leq k$, $f^{\pi(\ell)}$ is a function in F which (a) is not dominated by $f^{\pi(1)} \vee \ldots \vee f^{\pi(\ell-1)}$; (b) among functions satisfying (a), minimizes $\sum_j \mu(a_j)f(a_j)$. It is used to optimize all $s(a_i)$ for which $f^{\pi(2)}(a_i) = 1$ and which were not already optimized by one of $f^{\pi(1)}, \ldots, f^{\pi(\ell-1)}$.

(4) $(f^{\pi(1)} \vee \ldots \vee f^{\pi(k)})(a_j) = 1$ for all j.

For each $1 \leq \ell \leq k$, the combined contribution to the total sensitivity of all those a_j whose sensitivity is optimized by $f^{\pi(\ell)}$, is at most 1. So it is sufficient to show:

Lemma B.3. *With probability* $1 - o(1)$*: any series* $f^{\pi(1)} < (f^{\pi(1)} \vee f^{\pi(2)}) < \ldots < (f^{\pi(1)} \vee \ldots \vee f^{\pi(k)})$ *has* $k \leq 5 \log n$.

Proof. It is enough to show that with probability $1 - o(1)$, for any series of length $k = 3 \log n$, $f^{\pi(1)} \vee \ldots \vee f^{\pi(k)}$ equals 1 on all but at most $2 \log n$ points.

For any specific set T of $2 \log n$ points, and any specific set of indices $\pi(1), \ldots, \pi(3 \log n)$, $\Pr((f^{\pi(1)} \vee \ldots \vee f^{\pi(k)})(a_j) = 0 \text{ for all } j \in T) = 2^{-6 \log^2 n}$. Take a union bound over the $\binom{n}{2 \log n}$ choices of T and the $\binom{n}{3 \log n}$ choices of π (as a set, i.e., order does not matter). So the probability of "failure" is $\leq 2^{-6 \log^2 n} 2^{5 \log^2 n} \in o(1)$.