Fast and Sample Near-Optimal Algorithms for Learning Multidimensional Histograms

Ilias Diakonikolas[*]
USC
diakonik@usc.edu

Jerry Li [†]
MIT
jerryzli@mit.edu

Ludwig Schmidt[‡]
MIT
ludwigs@mit.edu

June 6, 2018

Abstract

We study the problem of robustly learning multi-dimensional histograms. A d-dimensional function $h : D \to \mathbb{R}$ is called a k-histogram if there exists a partition of the domain $D \subseteq \mathbb{R}^d$ into k axis-aligned rectangles such that h is constant within each such rectangle. Let $f : D \to \mathbb{R}$ be a d-dimensional probability density function and suppose that f is OPT-close, in L_1-distance, to an unknown k-histogram (with unknown partition). Our goal is to output a hypothesis that is $O(\mathrm{OPT}) + \epsilon$ close to f, in L_1-distance. We give an algorithm for this learning problem that uses $n = \tilde{O}_d(k/\varepsilon^2)$ samples and runs in time $\tilde{O}_d(n)$. For any fixed dimension, our algorithm has optimal sample complexity, up to logarithmic factors, and runs in near-linear time. Prior to our work, the time complexity of the $d = 1$ case was well-understood, but significant gaps in our understanding remained even for $d = 2$.

1. Introduction

Density Estimation or *Distribution Learning* refers to the following unsupervised learning task: Given i.i.d. samples from an unknown target probability distribution, output a hypothesis that is a good approximation to the target distribution with high probability. Density estimation is a classical and paradigmatic statistical problem with a history of more than a century, starting with Pearson (1895) (see, e.g., Barlow et al. (1972); Devroye and Györfi (1985); Silverman (1986); Scott (1992); Devroye and Lugosi (2001) for textbook introductions). Despite this long and rich history, core computational aspects of density estimation are wide-open in a variety of settings. Starting with the pioneering work of Kearns et al. (1994), computer scientists have been working on this broad fundamental question for more than two decades.

The recent distribution learning literature usually studies *structured* settings in which the target distribution belongs to a given distribution family $\mathcal{D}$ or is well-approximated by a member of this family with respect to a global loss function. The complexity of distribution learning often depends heavily on the structure of the underlying family. The performance of a distribution learning algorithm is typically evaluated by the following criteria:

[*] Supported by NSF Award CCF-1652862 (CAREER) and a Sloan Research Fellowship.

[†] Supported by NSF grant CCF-1217921, DOE grant de-sc0008923, NSF CAREER Award CCF-145326, and a NSF Graduate Research Fellowship

[‡] Supported by a Google PhD Fellowship.

- *Sample Complexity:* For a given error tolerance, the algorithm should require a small number of samples, ideally matching the information-theoretic minimum.
- *Computational Complexity:* The algorithm should run in time polynomial (or, ideally, linear) in the number of samples provided as input.
- *Robustness:* The algorithm should provide error guarantees under model misspecification, i.e., even if the target distribution does not belong in the target family $\mathcal{D}$. The goal here is to be competitive with the best approximation of the unknown distribution by any distribution in the family $\mathcal{D}$.

There are two main strands of research in distribution learning. The first one concerns the learnability of *high-dimensional parametric* distribution families, e.g., mixtures of Gaussians. The sample complexity of learning parametric families is typically polynomial in the dimension and the goal is to design computationally efficient algorithms.

The second research strand — which is the focus of this paper — studies the learnability of *low-dimensional nonparametric* distribution families under various assumptions on the shape of the underlying density. There has been a long line of work on this strand within statistics since the 1950s and, more recently, in theoretical computer science. The reader is referred to Barlow et al. (1972) for a summary of the early work and to Groeneboom and Jongbloed (2014) for a recent book on the subject. The majority of this literature has studied the univariate (one-dimensional) setting which is by now fairly well-understood for a wide range of distributions. On the other hand, the *multivariate* setting is significantly more challenging and significant gaps in our understanding remain even for $d = 2$.

1.1. Our Results: Learning Multivariate Histograms

In this work, we study the problem of density estimation for the family of histogram distributions on d-dimensional domains. Throughout this paper, let $[m] = \{1, \ldots, m\}$ denote an ordered discrete domain of size m. A distribution on $[0, 1]^d$ or $[m]^d$ with probability density function h is a k-histogram if there exists a partition of the domain into k axis-aligned hyper-rectangles $R_1, \ldots, R_k$ such that h is constant within each of the R_i's.

Histograms constitute one of the most basic nonparametric distribution families. The algorithmic difficulty in learning such distributions lies in the fact that the location and size of these rectangles is unknown to the algorithm. Histograms have been extensively studied in statistics and computer science. Many methods have been proposed to estimate histogram distributions Scott (1979); Freedman and Diaconis (1981); Scott (1992); Lugosi and Nobel (1996); Devroye and Lugosi (2004); Willett and Nowak (2007); Klemela (2009) that are of a heuristic nature or have a strongly exponential dependence on the dimension. In the database community, histograms Jagadish et al. (1998); Chaudhuri et al. (1998); Thaper et al. (2002); Gilbert et al. (2002); Guha et al. (2006); Indyk et al. (2012); Acharya et al. (2015) constitute the most common tool for the succinct approximation of data.

The time complexity of learning *univariate* histograms is well-understood: prior work Chan et al. (2013, 2014a,b); Acharya et al. (2017) gives sample-optimal learning algorithms with near-linear running time. Perhaps surprisingly, no nearly-linear time learning algorithm is known for arbitrary histograms even in *two* dimensions. Motivated by this gap in our understanding, we study the following question:

Is there a computationally and statistically efficient algorithm to learn arbitrary histograms on $\mathbb{R}^d$, up to ℓ_1 distance ε?

Our main result answers this question in the affirmative for any constant dimension:

Theorem 1 (informal, see Theorem 17) *Fix $\varepsilon > 0$ and $k \in \mathbb{Z}_+$. Let f be an arbitrary distribution over $[m]^d$ or $[0,1]^d$. There is an algorithm which draws $n = \tilde{O}_d(k/\varepsilon^2)$ samples [1] from f, runs in time[2] $\tilde{O}_d(n)$, and outputs a hypothesis h that with high probability satisfies $\|f - h\|_1 \leq O(\mathrm{OPT}_k) + \varepsilon$, where $\mathrm{OPT}_k = \min_{h'} \|f - h'\|_1$ is the best ℓ_1-distance achievable by any k-histogram.*

It is well-known (see, e.g., Acharya et al. (2017)) that $\Omega(k/\varepsilon^2)$ samples are necessary for any histogram learning algorithm, even for $d = 1$. Hence, for any fixed dimension d, our algorithm is sample near-optimal (within logarithmic factors) and runs in sample nearly-linear time. Even for $d = 2$ and $\mathrm{OPT}_k = 0$, no non-trivial algorithm was previously known for this problem.

A few additional remarks are in order. First, we would like to stress that the focus of our work is on the case where the parameters m, k are *much larger* than the dimension d, i.e., $m, k \gg d$. For example, this condition is automatically satisfied when d is bounded from above by a fixed constant. This is arguably the most natural setting for several applications of multidimensional histograms. Second, our proof establishes that the hidden multiplicative constant in the $O(\mathrm{OPT}_k)$ of the RHS is at most 11. While we do not know the value of the optimal constant, a lower bound of 2 is known even in one dimension Chan et al. (2014b).

Third, the dependence on d in the sample complexity of our algorithm is (weakly) exponential. Such a dependence in the sample size is not necessary. Standard information-theoretic arguments (e.g. VC dimension) give that $\tilde{O}(kd/\varepsilon^2)$ samples suffice — albeit with a $(1/\varepsilon)^{\Omega(kd)}$ time learning algorithm, which is clearly unacceptable even in one dimension. Obtaining a learning algorithm with running time $\mathrm{poly}(d, k, 1/\varepsilon)$ is left as a challenging open problem. As observed in De et al. (2015), the existence of such an algorithm may be unlikely as it would imply a $\mathrm{poly}(d, k, 1/\varepsilon)$ time algorithm for PAC learning k-leaf decision trees over $\{0,1\}^d$.

As a corollary of our algorithmic techniques, we also obtain an efficient “semi-proper”[3] learning algorithm for discrete histograms with respect to the ℓ_2-distance. Specifically, we show:

Theorem 2 (informal, see Theorem 25) *Fix $\varepsilon > 0$ and $k, m, d \in \mathbb{Z}_+$. Let $f : [m]^d \to \mathbb{R}$ be an arbitrary distribution. There is an algorithm which draws $n = O(1/\varepsilon)$ samples from f, runs in $O_d(n \log^2 n)$ time, and outputs an $O_d(k \log^{d+1} 1/\varepsilon)$-histogram h so that with high probability $\|f - h\|_2^2 \leq 2 \cdot \mathrm{OPT}_k + \varepsilon$, where $\mathrm{OPT}_k = \min_{h'} \|f - h'\|_2^2$ is the best ℓ_2-squared error achievable by any k-histogram.*

It is a folklore fact (see, e.g., Acharya et al. (2015)) that $\Theta(1/\varepsilon)$ samples are necessary and sufficient for this problem and that the empirical distribution is an accurate hypothesis. Our algorithm is sample-optimal, runs in near-linear time for constant dimension d, and importantly provides a succinct “semi-proper” hypothesis distribution. Succinct data representations by multivariate histograms are well-motivated in several data analysis applications in databases, where randomness is used to sub-sample a large dataset Cormode et al. (2012).

[1] Throughout the paper we let $\tilde{O}(g) = O(g \log^{O(1)} g)$

[2] We write $O_d(g)$ to denote that the variable d is treated as a constant in the expression g.

[3] We call our algorithm semi-proper because it produces a hypothesis that is also a histogram but with more than k pieces. For our algorithm, the increase in the number of histogram pieces is a polylogarithmic factor.

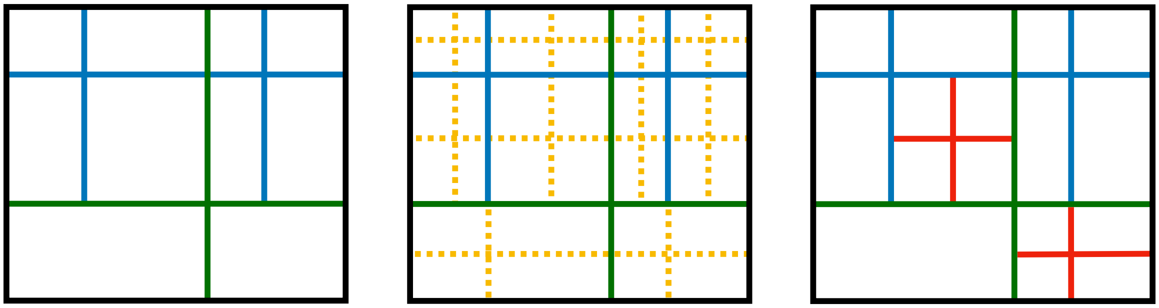
1.2. Our Techniques and Comparison to Prior Work

In this section, we provide an overview of our techniques in tandem with a comparison to prior work. Standard metric entropy arguments (see, e.g., Devroye and Lugosi (2001)) yield an inefficient method that uses $\tilde{O}(kd/\varepsilon^2)$ samples and runs in time $(1/\varepsilon)^{\Omega(kd)}$. To avoid the exponential dependence on k in the runtime, one can first partition the domain into $\mathrm{poly}(k/\varepsilon)^{\Theta(d)}$ "light" rectangles and then learn the induced probability distribution on these rectangles. This naive learning algorithm inherently incurs sample complexity and running time of $\mathrm{poly}(k/\varepsilon)^{\Theta(d)}$, which makes it unsatisfying even for 2 dimensions.

Our algorithms rely on two main ideas. The first ingredient is a *greedy splitting* scheme that enables us to approximate multi-dimensional histograms efficiently. In contrast to one-dimensional histograms, the partitions induced by multi-dimensional histograms are too complicated for a direct dynamic programming approach. Similarly, the approximate *iterative merging* strategy analyzed in Acharya et al. (2017) does not seem to generalize to the multi-dimensional setting: merging two adjacent rectangles does not necessarily yield another rectangle (as opposed to adjacent intervals). We circumvent the difficulties introduced by the complex structure of arbitrary histogram partitions by going through hierarchical histograms, which yield a more structured space of partitions that is amenable to efficient algorithms. Willett and Nowak (2007) used a related decomposition to learn smooth classes of continuous densities. First, we note that our algorithm and its analysis are significantly different from theirs. Second, Willett and Nowak (2007) do not obtain a near-linear time algorithm even in one dimension. In the univariate setting, Diakonikolas et al. (2017) used a similar algorithm to learn discrete distributions in the distributed setting with respect to the ℓ_2-norm.

Hierarchical histograms have appeared before in histogram approximation, especially in the setting of wavelet-based approaches (for instance, see Gilbert et al. (2001, 2002)) but also in approximate dynamic programs such as Muthukrishnan et al. (1999). However, these approaches do not handle the ℓ_1-setting that is standard in distribution learning. Instead, we propose a top-down splitting algorithm that expands leaf nodes in a growing hierarchical histogram according to a special error metric that we call the *$\mathcal{D}$-distance*. The $\mathcal{D}$-distance is closely related to VC theory and allows us to make good splitting decisions not only for the empirical distribution but also for the unknown distribtion we aim to recover.

The basic version of our greedy splitting scheme relies on hierarchical partitions of the distribution domain $[m]^d$, which incurs a logarithmic dependence on the domain size and does not apply to the continuous setting. The second ingredient in our paper is an *adaptive* variant of our splitting algorithm. This variant makes splitting decisions not on the dyadic boundaries of a data-independent hierarchical partition, but instead relies on the empirical distribution to build a data-dependent grid of coordinate points. By restricting our attention to the relevant coordinates, we can remove the logarithmic dependence on the domain size m and also apply our algorithm to distributions defined on $[0, 1]^d$. The adaptive approach requires a more careful analysis of our splitting algorithms and relies on the notion of a *partial* hierarchical histogram. In a partial hierarchical histogram, each partition can "shrink" to the bounding box of the samples in the partition, leaving a region on which the partition assigns value 0. Our final adaptive splitting algorithm runs in time that is nearly-linear in the number of samples with no dependence on the domain size. This is in contrast to prior wavelet-based approaches, which usually have a logarithmic dependence on the domain size and often process the entire domain $[m]^d$ as opposed to only the non-zero sample points.



(a) Current partition (b) Possible leaf refinements (c) Next partition

Figure 1: One iteration of our adaptive partitioning scheme. The left sub-figure (a) displays the hierarchical partitioning of $\mathbb{R}^2$ after two iterations of the algorithm. It is derived from two levels of splits: first the green split, then the blue splits. The location of the splits is given by locations of sample points (omitted for clarity), not by a fixed dyadic partition of the domain. The center sub-figure (b) shows the candidate leaf splits that the algorithm considers as next refinements. The algorithm chooses the splits that most reduce a certain error metric (see the right sub-figure (c)).

2. Preliminaries

We define the ℓ_p-norm of a measurable function $f : [m]^d \to \mathbb{R}$ or $f : [0,1]^d \to \mathbb{R}$ to be $\|f\|_p = \left(\sum_{x \in [m]^d} |f(x)|^p\right)^{1/p}$ or $\|f\|_p = \left(\int |f(x)|^p dx\right)^{1/p}$, for $[m]^d$ and $[0,1]^d$ respectively. For any subset $R \subseteq [m]^d \to \mathbb{R}$ (similarly for $[0,1]^d$), we let $\|f\|_{p,R}$ be the ℓ_p-norm of f restricted to R. Given $X_1, \ldots, X_n$ samples from a distribution f supported over $[m]^d$ (resp $[0,1]^d$), we let the empirical distribution induced by these samples be $\widehat{f} = \frac{1}{n}\sum_{i=1}^n \delta_{X_i}$, where δ_X is the delta distribution supported at X. Given a (measurable) set $R \subseteq [m]^d$ (resp. $[0,1]^d$), we let $|R|$ denote the counting measure of R (resp. the Lebesgue measure of R).

2.1. Histograms and Problem Definition

We first define the notion of histograms. Throughout this paper, we will assume w.l.o.g. that m is a power of 2.

Definition 3 *A distribution $h : A \to \mathbb{R}$, where A is either $[m]^d$, for $m \in \mathbb{Z}_+$, or $[0,1]^d$, is called a k-*histogram *if there exists a partition of A into k axis aligned rectangles $R_1, \ldots, R_k$ so that h is constant on R_i, for all $i = 1, \ldots, k$. We let $\mathcal{H}_k$ denote the set of k-histograms.*

We now can state the formal problem:

Problem Statement Given $0 < \varepsilon, \delta < 1$ and independent samples from some distribution $f : A \to \mathbb{R}$ where A is either $[m]^d$ or $[0,1]^d$, return $\widehat{h}$ so that with probability $1-\delta$, we have $\|\widehat{h} - f\|_1 \leq C \cdot \mathrm{OPT}_k + \varepsilon$, where C is an absolute constant and

$$\mathrm{OPT}_k = \min_{h \in \mathcal{H}_k} \|h - f\|_1 .$$

We will also crucially make use of the following definition throughout the paper:

Definition 4 *Let g be any function over $[m]^d$ (resp. $[0,1]^d$). For any set $R \subseteq [m]^d$ (resp. $[0,1]^d$), define the flattening of g over R, denoted $\overline{g}_R$, to be the constant function on R which takes on value $g(R)/|R|$ at each point in R. For any collection of disjoint sets $\mathcal{R}$, define the flattening of g over $\mathcal{R}$, denoted $\overline{g}_{\mathcal{R}}$, to be the function which is equal to the flattening of g on each set $R \in \mathcal{R}$ (and 0 otherwise).*

2.2. Hierarchical Histograms

We also require the notion of a *hierarchical* histogram, which is a histogram that respects a fixed dyadic partition. Formally:

Definition 5 *Given a grid $\mathcal{G} = P_1 \times P_2 \times \ldots \times P_d$, where each P_i is a collection of elements $x_1^{(i)} \leq x_2^{(i)} \leq \ldots \leq x_M^{(i)}$ in $[m]$ (resp. $[0,1]$) and M is a power of 2,* the level-ℓ rectangles *induced by $\mathcal{G}$, denoted $\mathcal{R}_\ell$, is defined to be*

$$\mathcal{R}_\ell = \left\{ \otimes_{i=1}^d [x^{(i)}_{2^\ell j_i + 1}, x^{(i)}_{2^\ell (j_i+1)}] : j_i \in \{0, \ldots, M/2^\ell - 1\} \right\} .$$

Moreover, the dyadic decomposition *of $\mathcal{G}$, denoted $\mathcal{D} = \mathcal{D}(\mathcal{G})$, is defined to be $\mathcal{D} = \bigcup_{\ell=1}^{\log M} \mathcal{R}_\ell$. For any $k \geq 1$, and a dyadic decomposition $\mathcal{D}$ of a grid $\mathcal{G}$ we let $\mathcal{D}_k$ denote all disjoint unions of at most k rectangles from $\mathcal{D}$.*

For instance, if the domain is $[m]^d$ and each $P_i = [m]$, then the induced dyadic decomposition is simply the set of squares R with side-length 2^ℓ for some $\ell = 1, \ldots, \log m$ and whose rightmost vertices are at a power of 2. In general, any dyadic decomposition induces a natural tree structure, which we will utilize throughout the paper. We can now define our notion of a hierarchical histogram:

Definition 6 *We say a k-histogram $f : A \to \mathbb{R}$ where A is either $[m]^d$ or $[0,1]^d$ is hierarchical with respect to a grid $\mathcal{G}$ if there exists a partition of A into rectangles $R_1, \ldots, R_k \in \mathcal{D}(\mathcal{G})$ so that f is constant on each R_i. If $\mathcal{G}$ is understood, we say f is hierarchical for short.*

We have the following simple lemma, which says that we may assume w.l.o.g. that the histogram is hierarchical, with some loss:

Lemma 7 *Fix a grid $\mathcal{G}$ with side length M. Let $f : A \to \mathbb{R}$ where A is either $[m]^d$ or $[0,1]^d$ be a k-histogram so that it is constant on $R_1, \ldots, R_k$, where every vertex of every rectangle lies on $\mathcal{G}$. Then f is a $k \log^d M$-hierarchical histogram.*

Proof For simplicity of exposition we will show this assuming $\mathcal{G} = [m]^d$, so the side length is equal to m. The same proof easily extends to general grids, and so we omit the details for conciseness. It suffices to show that any function which is supported within an axis-aligned rectangle R and which is constant within this rectangle can be represented as a $\log^d m$-hierarchical histogram. Let $R = [a_1, b_1] \times [a_2, b_2] \times \ldots \times [a_d, b_d]$. Each interval $[a_i, b_i]$ can be written as a union of at most $\log m$ disjoint dyadic intervals $\mathcal{I}_i$, so R can be decomposed as the disjoint union of all rectangles $R = \otimes_{i=1}^d I_i$ where each I_i ranges over all intervals in $\mathcal{I}_i$. By inspection, this requires $\log^d m$ pieces. ■

Thus, we lose $\log^d M$ factors going from arbitrary histograms to hierarchical histograms, where M is the side length of our grid.

3. Learning Histograms in ℓ_1-Distance

We now consider the question of histogram approximation in ℓ_1. The main difficulty in learning in ℓ_1 (as opposed to say, in ℓ_2), is that the statistical and algorithmic questions do not nicely decouple. This is because the empirical distribution is not close to the true distribution until many samples are taken. Instead, we will have to consider a different algorithmic objective inspired by VC theory.

3.1. VC theory

We now need the following classical definition of VC-dimension:

Definition 8 (VC dimension) *A collection of sets $\mathcal{A}$ is said to* shatter *a set S if for all $S' \subseteq S$, there is an $A \in \mathcal{A}$ so that $A \cap S = S'$. The VC dimension of $\mathcal{A}$, denoted $\mathrm{VC}(\mathcal{A})$, is the largest n so that there exists a S with $|S| = n$ so that $\mathcal{A}$ shatters S.*

For any collection $\mathcal{A}$ of measurable subsets of $[m]^d$ or $[0,1]^d$, define the $\mathcal{A}$-norm, denoted $\|\cdot\|_{\mathcal{A}}$, on measurable real-valued functions on $\mathbb{R}^d$ to be

$$\|f\|_{\mathcal{A}} = \sup_{A \in \mathcal{A}} |f(A)| \; .$$

For any measurable subset R of $[m]^d$ or $[0,1]$, we also define $\|\cdot\|_{\mathcal{A},R}$ to be the $\mathcal{A}$-norm of the function restricted to R. We now need the following form of the VC theorem, which follows by combining a classical form of the VC theorem along with standard uniform deviation arguments (e.g. McDiarmid's inequality):

Theorem 9 (c.f. Devroye & Lugosi Theorems 4.3 and 3.2, Chan et al. (2014a), Theorem 2.2) *Let $f : [m]^d \to \mathbb{R}$ be a distribution, and let $\widehat{f}_n$ denote the empirical distribution after n independent draws from f. Then, for all $\delta > 0$, $\Pr\left[\|f - \widehat{f}_n\|_{\mathcal{A}} \geq \sqrt{\frac{\mathrm{VC}(\mathcal{A}) + \log 1/\delta}{n}} \right] \leq \delta$.*

3.2. Computing $\mathcal{D}$-distance and fitting in $\mathcal{D}$-distance

Recall that $\mathcal{D}$ is defined to be the set of dyadic rectangles over $[m]^d$ (resp. $[0,1]^d$). By the theory developed above, this naturally induces a metric on functions from $[m]^d$ (resp. $[0,1]^d$) to $\mathbb{R}$. In this section, we show that computing and fitting with respect to $\mathcal{D}$ distance can be done in nearly input-sparsity time. Throughout this section, fix any grid $\mathcal{G}$ of side length M for $[m]^d$ or $[0,1]^d$. For any $a \in \mathbb{R}$ and $R \in \mathcal{D}$, let $\phi_{a,R} : R \to \mathbb{R}$ denote the function on R which is constantly a. We show:

Lemma 10 *Given an empirical distribution $\widehat{f}$, a rectangle $R \in \mathcal{D}$ so that $\widehat{f}$ is supported on s points in R, and a $a \in \mathbb{R}$, there is an algorithm* COMPUTED1 *that runs in time $O(2^d s \log M)$ and outputs $\|\widehat{f} - \phi_{a,R}\|_{\mathcal{D},R}$ together with a rectangle in $\mathcal{D}$ achieving this maximum.*

For conciseness we defer the proof of Lemma 10 to Appendix B. We now show that as a simple consequence of this, we can (approximately) find the constant fit to $\widehat{f}$ on any rectangle R in $\|\cdot\|_{\mathcal{D},R}$-norm in nearly linear time:

Corollary 11 *Given $\gamma > 0$, an empirical distribution $\widehat{f}$, a rectangle $R \in \mathcal{D}$ so that $\widehat{f}$ is supported on s points in R, there is an algorithm* FITD1 *which outputs an $a \in \mathbb{R}$ so that $\|\widehat{f} - \phi_{a,R}\|_{\mathcal{D},R} \leq \min_{a' \in \mathbb{R}} \|\widehat{f} - \phi_{a',R}\|_{\mathcal{D},R} + \gamma$ in time $\widetilde{O}(2^d \cdot s \log M \log 1/\gamma)$.*

The algorithm is simple: we reduce the optimization problem with binary searching over feasibility problems, then solve each feasibility problem using COMPUTED1 as a separation oracle. The details are subsumed by the calculations for Theorem 31 of Acharya et al. (2017), so we omit them.

For simplicity, we shall assume for the rest of the paper that FITD1 produces an exact fit in $\mathcal{A}_1$-distance. Because the dependence on γ in the runtime is logarithmic, it is not hard to see that by taking $\gamma = \mathrm{poly}(1/k, \varepsilon, 1/\log 1/\delta)^d$ in the remainder, we only increase the approximation errors throughout by at most additive ε factors, and this keeps the runtime unchanged, up to log factors.

3.3. The Greedy Splitting Algorithm for $\mathcal{D}_k$-Distance

In this section, we give an efficient algorithm for constructing hierarchical histograms for fitting a known empirical distribution in the norm induced by the hierarchical decomposition. Throughout this section, fix a grid $\mathcal{G}$ with side length M over either $[m]^d$ or $[0,1]^d$, and let $\mathcal{D} = \mathcal{D}(\mathcal{G})$ be the induced dyadic decomposition.

We will prove that our output, despite being a hierarchical histogram, is actually competitive with the best error achievable by a slightly more general class of functions, which we call partial hierarchical histograms. Formally:

Definition 12 *A partial k-histogram $h : [m]^d \to \mathbb{R}$ (or $h : [0,1]^d \to \mathbb{R}$) is a distribution satisfying the following: there exist k disjoint rectangles $R_1, \ldots, R_k$ such that h is supported on $\bigcup_{i=1}^k R_i$, and on each R_i, h is constant. We say that h is a partial k-hierarchical histogram with respect to a grid $\mathcal{G}$ if in addition we have $R_i \in \mathcal{D}(\mathcal{G})$ for all i.*

Our main algorithmic theorem is:

Theorem 13 *Fix $k \in \mathbb{Z}_+$, and let $\xi > 0$ be a tuning parameter. Let $\widehat{f}$ be an empirical distribution on s points. There is an algorithm* GREEDYSPLIT *which outputs a $O((1+\xi)2^d k \log M)$-hierarchical histogram h so that $\|h - \widehat{f}\|_{\mathcal{D}_k} \leq \left(3 + \frac{6}{\xi^2}\right) \cdot \widetilde{\mathrm{OPT}}_{\mathcal{D},k}$, where $\widetilde{\mathrm{OPT}}_{\mathcal{D},k} = \min_h \|h - \widehat{f}\|_{\mathcal{D}_k}$, where the minimum is taken over all partial hierarchical k-histograms h. Moreover, the algorithm runs in time $\widetilde{O}(2^d s \log^2 M)$.*

Our algorithm, given formally in Algorithm 1, is quite simple. We construct a tree of nested dyadic rectangles. Initially, this tree contains only $[m]^d$ (resp. $[0,1]^d$). Iteratively, we find the leaves of this tree with largest $\mathcal{D}$-distance error to g, and we split these into all of its children, and we repeat this for $\log M$ iterations. At the end, we return the flattening of g over all the leaves in the final tree. For conciseness, we defer the proof of Theorem 13 to Appendix B.

3.4. Warm-up: an Algorithm for Hierarchical Histograms on $[m]^d$

In this section, we will take $\mathcal{G} = [m]^d$. Assume for simplicity that m is a power of 2. Then, we may take $\mathcal{D}$ to be the dyadic partition of $[m]^d$, i.e., $\mathcal{D} = \bigcup \{\mathcal{R}_i\}_{i=1}^{\log m}$ where

$$\mathcal{R}_i = \left\{ \otimes_{i=1}^d [j_i m 2^{-i} + 1, (j_i + 1) m 2^{-i}] : j_i \in \{0, \ldots, 2^i - 1\} \right\}$$

are all rectangles on a $m2^{-i}$-spaced grid. The following are standard facts from VC theory and we defer their proof to Appendix B.

Algorithm 1 A greedy splitting algorithm for learning hierarchical histograms in $\mathcal{D}_k$-distance

```
function GreedySplit($\widehat{f}, \mathcal{D}, \xi$)
    Let $\mathcal{T}$ be a subtree of the hierarchical tree, initially containing only the root.
    for $\ell = 1, \ldots, \log M$ do
        for each leaf $R \in \mathcal{T}$ do
            Let $a_R =$ FitD1($\widehat{f}, \mathcal{D}, R$)
            Let $e_R =$ ComputeD1($\widehat{f}, R, a_R$)
        Let $\mathcal{J}$ be the set of $(1+\xi)k$ leaves $R \in \mathcal{T}$ with largest $e_R$.
        for each $R \in \mathcal{J}$ do
            if $R$ can be subdivided in $\mathcal{D}$ and $e_R > 0$ then
                Add all children of $R$ to $\mathcal{T}$
    return The function which is constantly $a_R$ for every leaf $R$ of $\mathcal{T}$
```

Corollary 14 *Let f, g be two k-hierarchical histograms. Then $2\|f - g\|_{\mathcal{D}_{2k}} = \|f - g\|_1$.*

Corollary 15 *For all $k, d \geq 1$, we have $\mathrm{VC}(\mathcal{D}_k) = O(kd)$.*

These corollaries together imply:

Corollary 16 *Fix $\varepsilon, \delta > 0$, and let $\xi > 0$. Let $f : [m]^d \to \mathbb{R}$ be an arbitrary distribution. Then, the algorithm* GreedySplit$(\widehat{f}, \mathcal{D}([m]^d), \xi)$*, given $\widehat{f} = \widehat{f}_n$ which is the empirical distribution of f after $n = \Omega\left(\frac{(1+\xi)2^d k \log^{d+1} m + \log 1/\delta}{\varepsilon^2}\right)$ samples, outputs a $(1+\xi)2^d dk \log^{d+1} m$-hierarchical histogram h so that with probability $1 - \delta$, we have $\|f - h\|_1 \leq \left(5 + \frac{6}{\xi^2}\right) \cdot \mathrm{OPT}_k + \varepsilon$. Moreover, this algorithm runs in time $O(2^d n \log^2 m)$.*

Proof The bound on the number of pieces and the runtime of the algorithm follow from Lemmata 29 and 30 immediately.[4] Thus, it suffices to argue about correctness. By Lemma 7, we know that if we let $\mathrm{OPT}'_{k \log^d m}$ be the optimal ℓ_1-error to f achievable by a hierarchical $k \log^d m$-histogram, then $\mathrm{OPT}'_{k \log^d m} \leq \mathrm{OPT}_k$. Let $h^*_{\mathcal{D}}$ be the hierarchical k-histogram which achieves the optimum.

Condition on the event that $\|f - \widehat{f}\|_{\mathcal{D}_\kappa} \leq c\varepsilon$, where $\kappa = 2(1+\xi)2^d k \log^{d+1} m$, for some universal constant c sufficiently small. By Theorem 9, this happens with probability $1 - \delta$ if we take $n = \Omega\left(\frac{(1+\xi)2^d k \log^{d+1} m + \log 1/\delta}{\varepsilon^2}\right)$ samples. Then, we have

$$\|\widehat{f} - h^*_{\mathcal{D}}\|_{\mathcal{D}_\kappa} \leq \|f - \widehat{f}\|_{\mathcal{D}_\kappa} + \|f - h^*_{\mathcal{D}}\|_{\mathcal{D}_\kappa} \leq \mathrm{OPT}'_{k \log^d m} + c\varepsilon \ .$$

Therefore, we have $\widetilde{\mathrm{OPT}}_{k \log^d m} \leq \mathrm{OPT}'_{k \log^d m} + c\varepsilon$. Combining this with the guarantee from Theorem 13 then implies

$$\|\widehat{f} - h\|_{D_k} \leq \left(3 + \frac{6}{\xi^2}\right) \mathrm{OPT}_k + c\varepsilon \ .$$

[4] While these lemmata are for ℓ_2 the proof transfers immediately over to this setting so we omit the details.

To complete proof, we now observe:

$$\begin{aligned}\|f-h\|_1 &\leq \mathrm{OPT}_k + \|h^* - h\|_{\mathcal{D}_k} \\ &\leq \mathrm{OPT}_k + \|h^* - \widehat{f}\|_{\mathcal{D}_k} + \|\widehat{f} - h\|_{\mathcal{D}_k} \\ &\leq \mathrm{OPT}_k + \|h^* - f\|_{\mathcal{D}_k} + \|\widehat{f} - f\|_{\mathcal{D}_k} + \|\widehat{f} - h\|_{\mathcal{D}_k} \\ &\leq 2\mathrm{OPT}_k + O(\varepsilon) + \left(3 + \frac{6}{\xi^2}\right)\mathrm{OPT}_k + O(\varepsilon)\;,\end{aligned}$$

which simplifies to the desired statement. ■

3.5. General Histograms via Adaptive Gridding

The framework presented above is clean, however, it has one major drawback. Namely, the conversion from arbitrary to hierarchical histograms on the grid $[m]^d$ loses $\log^d m$ factors. In particular, these factors prevent the algorithm from being useful when the support size is large or infinite. In this section, we show that a modification of the techniques presented above can remove these factors. The algorithm in this section will work even when the support size is infinite. Throughout the section, we will state our results for $[m]^d$, however, they generalize trivially to $[0,1]^d$, and we omit the details for simplicity. Our main result in this section is:

Theorem 17 *Fix $\varepsilon, \delta > 0$, and let $\xi > 0$. Let $f : [m]^d \to \mathbb{R}$ be an arbitrary distribution. There is an algorithm* ADAPTIVEGREEDYSPLIT*, which, given n independent samples from f, where $n = O\left(\frac{(1+\xi)d2^d k \log^{d+2}(k/\varepsilon) + \log 1/\delta}{\varepsilon^2}\right)$, outputs a $O((1+\xi)d2^d k \log^{d+1}(k/\varepsilon))$-hierarchical histogram h so that with probability $1-\delta$, we have $\|f-h\|_1 \leq \left(10 + \frac{12}{\xi^2}\right) \cdot \mathrm{OPT}_k + \varepsilon$. Moreover, this algorithm runs in time $O(2^d n \log^2 n)$.*

The VC dimension of (Partial) Histograms In this section, we bound the VC dimension of set systems induced by differences between k-histograms and partial histograms. We first need the following fact, which is a direct implication of the respective definitions:

Fact 18 *Given two k-partial histograms $h, g : [m]^d \to \mathbb{R}$, the set $\{x : h(x) > g(x)\}$ is of the form $\bigcup_{i=1}^{k'} A_i - \bigcup_{j=1}^{k''} B_j$, for some axis aligned rectangles A_i, B_j so that the A_i are mutually disjoint and B_j are mutually disjoint, and $k', k'' \leq k$.*

Motivated by this fact, we let

$$\mathcal{A}_k = \left\{ \bigcup_{i=1}^{k'} A_i - \bigcup_{j=1}^{k''} B_j \,:\, \{A_i\}_{i=1}^{k'}, \{B_j\}_{j=1}^{k''} \text{ are collections of disjoint rectangles, } k', k'' \leq k \right\}$$

be the set system that captures sign difference between k-partial histograms. By Fact 18, we have:

Corollary 19 *For any two k-partial histograms $h, g : [m]^d \to \mathbb{R}$ (or over $[0,1]^d$), we have $\|h - g\|_1 = 2\|h-g\|_{\mathcal{A}_k}$.*

We now require a bound on the VC dimension of $\mathcal{A}_k$, whose proof we defer to the appendix:

Lemma 20 *For all $k \geq 1$, we have $\mathrm{VC}(\mathcal{A}_k) = O(kd \log(kd))$.*

As an immediate corollary of Theorem 9 and Lemma 20, we have:

Corollary 21 *Fix $\varepsilon, \delta > 0$. Let $f : [m]^d \to \mathbb{R}$ be an arbitrary distribution. Let $\widehat{f} = \widehat{f}_n$ be the empirical distribution given n independent samples from f, where $n = O\left(\frac{kd \log(kd) + \log 1/\delta}{\varepsilon^2}\right)$. Then, with probability $1 - \delta$, we have $\|\widehat{f} - f\|_{\mathcal{A}_k} \leq \varepsilon$.*

For the rest of the section, we let f denote the unknown distribution, and we let $\widehat{f} = \widehat{f}_n$ denote the empirical distribution after n draws from f, where

$$n = C \frac{(1+\xi) d 2^d k \log^{d+2}(k/\varepsilon) + \log 1/\delta}{\varepsilon^2} ,$$

for some universal constant C sufficiently large. We let $\mathcal{X}$ denote the (multi-)set of samples, i.e., $\mathcal{X} = \mathrm{supp}(\widehat{f})$, and we will, in a slight abuse of notation, let $\mathcal{D} = \mathcal{D}(\mathrm{grid}(\mathcal{X}))$.

We will condition on the event that

$$\|\widehat{f} - f\|_{\mathcal{A}_\kappa} \leq c' \varepsilon , \tag{1}$$

for some universal constant c' sufficiently small, where $\kappa = C'(1+\xi) 2^d k \log^{d+1}(k/\varepsilon)$ for some universal constant C' sufficiently large. Observe that since $\mathcal{D}_k \subseteq \mathcal{A}_k$, this immediately implies that $\|\widehat{f} - f\|_{\mathcal{D}_k} \leq c'\varepsilon$. By Corollary 21, this holds with probability $1 - \delta$ as long as we take at least

$$n = \Omega\left(\frac{d\kappa \log(d\kappa) + \log 1/\delta}{\varepsilon^2}\right) .$$

In particular, this holds for our choice of n, for C sufficiently large.

Rounding Histograms to Partial Hierarchical Histograms Our algorithm is straightforward: we simply grid over all points where the samples land, that is, we take the grid to be $\mathcal{G} = \mathrm{grid}(\mathcal{X})$, then find the best fit hierarchical histogram with respect to this grid, and the norm it induces, using the same algorithm as above. Our algorithm will then be very similar to the algorithm presented previously, with some crucial but subtle changes, however, the analysis requires some additional steps.

In particular, now it is not a priori clear that the optimal histogram fit to the true density will have vertices on the grid, and in general, it is not too hard to show that it will not. However, we show that by only losing constant factors in the approximation ratio, we may as well assume that it does, with some important caveats. Specifically, we show that we may approximate the optimal fit k-histogram to f with a k-partial histogram with vertices on the grid. Formally:

Lemma 22 *Fix $\varepsilon > 0$, and assume that (1) holds. Then, there is a $\kappa' = O(k \log^d(k/\varepsilon))$-partial histogram h_p^* that is hierarchical with respect to $\mathcal{X}$ so that*

$$\|f - h_p^*\|_1 \leq 2 \cdot \mathrm{OPT}_k + 4c'\varepsilon . \tag{2}$$

We defer the proof of Lemma 22 to Appendix B. We also need the following lemma, which states that the $\mathcal{D}_k$-distance still captures the ℓ_1-distance between a partial hierarchical histogram and a (regular) hierarchical histogram.

Lemma 23 *Fix a grid $\mathcal{G}$, and let h be a partial hierarchical k-histogram, and let g be a hierarchical k-histogram, both with respect to $\mathcal{G}$. Then, $\{x : h(x) > g(x)\} \in \mathcal{D}_k(\mathcal{G})$. In particular, this implies that $\|h - g\|_1 = 2\|h - g\|_{\mathcal{D}_{2k}}$.*

For conciseness we defer the proof to Appendix B.

Putting Everything Together We now have the tools to prove Theorem 17. The algorithm is fairly simple: we take the grid induced by our samples, and run GREEDYSPLIT on this grid on the empirical distribution. The formal pseudocode is given in Algorithm 4 in Appendix B.

Proof [Proof of Theorem 17] The runtime guarantee and the guarantee on the number of pieces easily follow from Theorem 13. Thus, it suffices to prove correctness. Let $\mathcal{X}$ denote the set of samples, let $\mathcal{G} = \mathrm{grid}(\mathcal{X})$, and $\mathcal{D} = \mathcal{D}(\mathcal{G})$. Recall h is the output of our algorithm, and let h^* be the optimal k-histogram fit to f in ℓ_1. By Lemma 22, we know that there is some partial hierarchical $O(k \log^d(k/\varepsilon))$-histogram h_p^* so that

$$\|f - h_p^*\|_1 \le 2\mathrm{OPT}_k + 2c'\varepsilon \; . \tag{3}$$

In particular, this implies that

$$\begin{aligned} \|\widehat{f} - h_p^*\|_{\mathcal{D}_\kappa} &\le \|\widehat{f} - f\|_{\mathcal{D}_\kappa} + \|f - h_p^*\|_{\mathcal{D}_\kappa} \\ &\le \varepsilon + 2\mathrm{OPT}_k + 2c'\varepsilon \\ &= 2\mathrm{OPT}_k + 3c'\varepsilon \; , \end{aligned}$$

and hence $\widetilde{\mathrm{OPT}}_{\mathcal{D},k} \le 2\mathrm{OPT}_k + 3c'\varepsilon$. We thus have

$$\begin{aligned} \|f - h\|_1 &\le \|f - h_p^*\|_1 + \|h_p^* - h\|_1 \\ &\overset{(a)}{\le} 2 \cdot \mathrm{OPT}_k + 2c'\varepsilon + \|h_p^* - h\|_{\mathcal{D}_\kappa} \\ &\overset{(b)}{\le} 2 \cdot \mathrm{OPT}_k + 2c'\varepsilon + \|h_p^* - \widehat{f}\|_{\mathcal{D}_\kappa} + \|\widehat{f} - h\|_{\mathcal{D}_\kappa} \\ &\overset{(c)}{\le} 2 \cdot \mathrm{OPT}_k + 2c'\varepsilon + \|h_p^* - f\|_1 + \|f - \widehat{f}\|_{\mathcal{D}_\kappa} + \|\widehat{f} - h\|_{\mathcal{D}_\kappa} \\ &\overset{(d)}{\le} 4\mathrm{OPT}_k + 5c'\varepsilon + \left(3 + \frac{6}{\xi^2}\right) \widetilde{\mathrm{OPT}}_{\mathcal{D},k} \\ &\le \left(10 + \frac{12}{\xi^2}\right) \mathrm{OPT}_k + O(c'\varepsilon) \; , \end{aligned}$$

where (a) follows from a triangle inequality, (3), and Lemma 23, (b) and (c) follow from the triangle inequality, and (d) follows from (3), (1) and Theorem 13. By choosing c' sufficiently small, this completes the proof. ■

References

J. Acharya, I. Diakonikolas, C. Hegde, J. Li, and L. Schmidt. Fast and near-optimal algorithms for approximating distributions by histograms. In *Proceedings of the 34th ACM Symposium on Principles of Database Systems, PODS 2015*, pages 249–263, 2015.

J. Acharya, I. Diakonikolas, J. Li, and L. Schmidt. Sample-optimal density estimation in nearly-linear time. In *Proceedings of the Twenty-Eighth Annual ACM-SIAM Symposium on Discrete Algorithms, SODA 2017*, pages 1278–1289, 2017. Available at https://arxiv.org/abs/1506.00671.

R.E. Barlow, D.J. Bartholomew, J.M. Bremner, and H.D. Brunk. *Statistical Inference under Order Restrictions*. Wiley, New York, 1972.

S. Chan, I. Diakonikolas, R. Servedio, and X. Sun. Learning mixtures of structured distributions over discrete domains. In *SODA*, pages 1380–1394, 2013.

S. Chan, I. Diakonikolas, R. Servedio, and X. Sun. Efficient density estimation via piecewise polynomial approximation. In *STOC*, pages 604–613, 2014a.

S. Chan, I. Diakonikolas, R. Servedio, and X. Sun. Near-optimal density estimation in near-linear time using variable-width histograms. In *NIPS*, pages 1844–1852, 2014b.

S. Chaudhuri, R. Motwani, and V. R. Narasayya. Random sampling for histogram construction: How much is enough? In *SIGMOD Conference*, pages 436–447, 1998.

G. Cormode, M. Garofalakis, P. J. Haas, and C. Jermaine. Synopses for massive data: Samples, histograms, wavelets, sketches. *Found. Trends databases*, 4:1–294, 2012. ISSN 1931-7883.

A. De, I. Diakonikolas, and R. Servedio. Learning from satisfying assignments. In *Proceedings of the 26th Annual ACM-SIAM Symposium on Discrete Algorithms, SODA 2015*, pages 478–497, 2015.

L. Devroye and L. Györfi. *Nonparametric Density Estimation: The L_1 View*. John Wiley & Sons, 1985.

L. Devroye and G. Lugosi. *Combinatorial methods in density estimation*. Springer, 2001.

L. Devroye and G. Lugosi. Bin width selection in multivariate histograms by the combinatorial method. *Test*, 13(1):129–145, 2004.

I. Diakonikolas, E. Grigorescu, J. Li, A. Natarajan, K. Onak, and L. Schmidt. Communication-efficient distributed learning of discrete distributions. In *Advances in Neural Information Processing Systems 30 (NIPS 2017)*, pages 6394–6404, 2017.

D. Freedman and P. Diaconis. On the histogram as a density estimator:l2 theory. *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete*, 57(4):453–476, 1981.

A. C. Gilbert, Y. Kotidis, S. Muthukrishnan, and M. Strauss. Surfing wavelets on streams: One-pass summaries for approximate aggregate queries. In *VLDB*, 2001.

A. C. Gilbert, S. Guha, P. Indyk, Y. Kotidis, S. Muthukrishnan, and Martin Strauss. Fast, small-space algorithms for approximate histogram maintenance. In *STOC*, pages 389–398, 2002.

P. Groeneboom and G. Jongbloed. *Nonparametric Estimation under Shape Constraints: Estimators, Algorithms and Asymptotics*. Cambridge University Press, 2014.

S. Guha, N. Koudas, and K. Shim. Approximation and streaming algorithms for histogram construction problems. *ACM Trans. Database Syst.*, 31(1):396–438, 2006.

P. Indyk, R. Levi, and R. Rubinfeld. Approximating and Testing k-Histogram Distributions in Sublinear Time. In *PODS*, pages 15–22, 2012.

H. V. Jagadish, Nick Koudas, S. Muthukrishnan, Viswanath Poosala, Kenneth C. Sevcik, and Torsten Suel. Optimal histograms with quality guarantees. In *VLDB*, pages 275–286, 1998.

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

J. Klemela. Multivariate histograms with data-dependent partitions. *Statistica Sinica*, 19(1):159–176, 2009.

G. Lugosi and A. Nobel. Consistency of data-driven histogram methods for density estimation and classification. *Ann. Statist.*, 24(2):687–706, 04 1996.

S Muthukrishnan, Viswanath Poosala, and Torsten Suel. On rectangular partitionings in two dimensions: Algorithms, complexity and applications. In *ICDT*, 1999.

K. Pearson. Contributions to the mathematical theory of evolution. ii. skew variation in homogeneous material. *Philosophical Trans. of the Royal Society of London*, 186:343–414, 1895. doi: 10.1098/rsta.1895.0010.

D. W. Scott. On optimal and data-based histograms. *Biometrika*, 66(3):605–610, 1979.

D.W. Scott. *Multivariate Density Estimation: Theory, Practice and Visualization*. Wiley, New York, 1992.

B. W. Silverman. *Density Estimation*. Chapman and Hall, London, 1986.

N. Thaper, S. Guha, P. Indyk, and N. Koudas. Dynamic multidimensional histograms. In *SIGMOD Conference*, pages 428–439, 2002.

R. Willett and R. D. Nowak. Multiscale poisson intensity and density estimation. *IEEE Transactions on Information Theory*, 53(9):3171–3187, 2007.

Appendix A. Learning Histograms in ℓ_2-Distance

In this section, we consider the problem of learning the best fit k-histogram to a unknown distribution over $[m]^d$ given sample access to the distribution. Our main result is:

Theorem 24 *Fix $\varepsilon, \delta > 0$, $k \in \mathbb{Z}_+$, and let $\gamma > 0$ be a tuning parameter. Let $f : [m]^d \to R$ be an arbitrary distribution. There is an algorithm* GREEDYSPLITL2 *which takes $n = O(\log(1/\delta)/\varepsilon)$ samples and outputs a hierarchical $(1+\xi)2^d k \log^{d+1} m$-histogram h so that*

$$\|h - g\|_2^2 \leq \left(1 + \frac{1}{\xi}\right) \mathrm{OPT}_k + \varepsilon \ ,$$

where $\mathrm{OPT}_k = \min_h \|h - g\|_2^2$, where the minimum is taken over all k-histograms h. Moreover, the algorithm runs in time $O(2^d n \log^2 m)$.

While the statement of this theorem does not quite obtain the guarantees in Theorem 2, in that we have $\log m$ factors instead of $\log 1/\varepsilon$ factors, it is trivial to use the same adaptive gridding techniques as we did for ℓ_1 to replace these $\log m$ factors with $\log 1/\varepsilon$ factors. Formally:

Theorem 25 *Fix $\varepsilon, \delta > 0$, $k \in \mathbb{Z}_+$, and let $\gamma > 0$ be a tuning parameter. Let $f : [m]^d \to R$ be an arbitrary distribution. There is an algorithm* GREEDYSPLITL2 *which takes $n = O(\log(1/\delta)/\varepsilon)$ samples and outputs a hierarchical $(1+\xi)2^d k \log^{d+1} n$-histogram h so that*

$$\|h - g\|_2^2 \leq \left(1 + \frac{1}{\xi}\right) \mathrm{OPT}_k + \varepsilon \ ,$$

where $\mathrm{OPT}_k = \min_h \|h - g\|_2^2$, where the minimum is taken over all k-histograms h. Moreover, the algorithm runs in time $O(2^d n \log^2 n)$.

Since the ideas are subsumed by those described for ℓ_1, we omit these details for simplicity. We note that the runtime of this algorithm is independent of m, the domain size. This holds in any model of computation where we can read and perform arithemtic operations on samples in constant time, such as the real RAM model of computation. Such a model is standard in learning theory.

Our starting point is the following well-known statistical guarantee, which states that the empirical distribution is ε-close to the true distribution in ℓ_2-norm after roughly $O(1/\varepsilon^2)$ samples.

Fact 26 (folklore, see e.g. Acharya et al. (2015)) *Fix $\varepsilon, \delta > 0$. Let $f : [m]^d \to \mathbb{R}$ be an arbitrary distribution, and let $\widehat{f} = \widehat{f}_n$ be the empirical distribution after $n = O(\log(1/\delta)/\varepsilon)$ independent samples from f. Then, with probability $1 - \delta$, we have $\|\widehat{f} - f\|_2^2 \leq \varepsilon$.*

This fact states that the ℓ_2 learning problem is purely algorithmic: it suffices to, given $\widehat{f}$, find the best fit k-histogram approximation to $\widehat{f}$ in ℓ_2. Then by a simple application of the triangle inequality, this will be an almost optimal fit to f in ℓ_2 as well. The main challenge is to devise algorithms for this problem which exploit the sparsity of $\widehat{f}$.

We will also make crucial use of the following fact, which follows from basic calculus. Then, we have:

Fact 27 (folklore) *Let $\widehat{f}$ be an empirical distribution over $[m]^d$, and let $R \subseteq [m]^d$ be any set. Then, the best constant fit to $\widehat{f}$ in ℓ_2 on R is the flattening of $\widehat{f}$ over R.*

A.1. Greedy Splitting for Hierarchical Histograms in ℓ_2-Distance

Our main algorithmic result for the ℓ_2-norm is a greedy splitting routine which finds a nearly optimal hierarchical histogram fit to a sparse function efficiently. Throughout this section, we will let $\mathcal{D} = \mathcal{D}([m]^d)$ be the full dyadic decomposition of the domain. While it is not hard to adapt the techniques in this section to work with an adaptive grid, as we did for the ℓ_1-distance, we will not do this here for simplicity of the presentation, as this is not the main focus of our paper.

Our main theorem is:

Theorem 28 *Fix $k \in \mathbb{Z}_+$, and let $\xi > 0$ be a tuning parameter. Let $g : [m]^d \to R$ be an arbitrary function supported on at most s points. There is an algorithm* GREEDYSPLITL2 *which outputs a $(1+\xi)2^d k \log m$-hierarchical histogram h so that*

$$\|h - g\|_2^2 \leq \left(1 + \frac{1}{\xi}\right) \mathrm{OPT}_{\mathcal{D},k} ,$$

where $\mathrm{OPT}_{\mathcal{D},k} = \min_h \|h - g\|_2^2$, where the minimum is taken over all hierarchical k-histograms h. Moreover, the algorithm runs in time $O(2^d s \log^2 m)$.

Combining this with Lemma 7 and Fact 26 immediately yields Theorem 24. Thus it suffices to prove this theorem.

Our algorithm, given formally in Algorithm 2, is quite similar to Algorithm 1. We construct a tree of nested dyadic rectangles. Initially, this tree contains only $[m]^d$. Iteratively, we find the leaves of this tree with largest ℓ_2^2 error to g, and we split these into all of its children, and we repeat this for $\log m$ iterations. At the end, we return the flattening of g over all the leaves in the final tree.

Algorithm 2 Algorithm for learning a hierarchical histogram in ℓ_2

```
function GREEDYSPLITL2($g, \xi$)
    Let $\mathcal{T}$ be a subtree of the hierarchical tree, initially containing only the root.
    for $\ell = 1, \ldots, \log m$ do
        for each leaf $R \in \mathcal{T}$ do
            Let $a_R = g(R)/|R|$
            Let $e_R = \sum_{x \in R}(g(x) - a_R)^2$
        Let $\mathcal{J}$ be the set of $(1+\xi)k$ leaves $R \in \mathcal{T}$ with largest $e_R$.
        for each $R \in \mathcal{J}$ do
            if $R$ can be subdivided in $\mathcal{D}$ and $e_R > 0$ then
                Add all children of $R$ to $\mathcal{T}$
    return The flattening of $g$ for every leaf $R$ of $\mathcal{T}$
```

We will prove this theorem in three parts. First, we will prove a bound on the number of pieces of the output histogram (Lemma 29). Then, we will bound the runtime of the algorithm (Lemma 30). Finally, we will bound the error of the algorithm (Lemma 31).

We first bound the number of pieces in our output:

Lemma 29 (Number of pieces) *The output of* GREEDYSPLITL2 *has at most $(1+\xi)2^d k \log M$ pieces.*

Proof In each iteration, we split at most $(1+\xi)k$ rectangles each into 2^d pieces, so we increase the number of pieces by at most $(1+\xi)2^d k$. Since there are $\log M$ iterations, this immediately proves the bound. ■

We now prove a bound on the runtime:

Lemma 30 (Runtime) GREEDYSPLITL2 *runs in time* $O(2^d s \log^2 M)$.

Proof In each iteration, we iterate over the number of rectangles currently in the tree, and we take $O(2^d s_R \log M)$ time per rectangle R, if R contains s_R points in the support of g. Thus per iteration we do at most $O((1+\xi)2^d s \log M)$ work, and there are $\log M$ iterations. Multiplying these two terms yields the desired claim. ■

Finally, we turn our attention to correctness:

Lemma 31 $\|h-g\|_2^2 \leq \left(1+\frac{1}{\xi}\right)\mathrm{OPT}_k$.

Proof Let h^* be an optimal hierarchical k-histogram fit to g in ℓ_2 norm, and let h be the output of our algorithm. For any set $S \subseteq [m]^d$, let $\mathrm{OPT}_{\mathcal{D},k}(S) = \sum_{x\in S}(h^*(x)-g(x))^2$ be the ℓ_2-squared error incurred by h^* on S. For any collection of sets $\mathcal{S}$, let $\mathrm{OPT}_{\mathcal{D},k}(\mathcal{S}) = \mathrm{OPT}_{\mathcal{D},k}(\cup_{S\in\mathcal{S}}S)$.

Let $\mathcal{T}$ be the tree associated with h. Let $\mathcal{R}^*$ be the set of k disjoint dyadic rectangles on which h^* is supported, and let $\mathcal{R}$ be the leaves of $\mathcal{T}$. Partition $\mathcal{R}$ into three sets:

$$\begin{aligned}\mathcal{F} &= \{R\in\mathcal{R}:\ h^* \text{ is constant on } R\,\}\\ \mathcal{J}_1 &= \{R\in\mathcal{R}:\ h^* \text{ is non-constant on } R \text{ and } e_R = 0\,\}\\ \mathcal{J}_2 &= \{R\in\mathcal{R}:\ h^* \text{ is non-constant on } R \text{ and } e_R > 0\,\}\ .\end{aligned}$$

We will prove that the error is low on all three sets separately.

Error on $\mathcal{F}$ First, we will prove that the error is low in $\mathcal{F}$. In fact, we will prove a more general lemma which will be useful later:

Lemma 32 *Let g be arbitrary. Let $\mathcal{R}\in\mathcal{D}_k$ be any union of at most k disjoint rectangles in $\mathcal{D}$, and let $\overline{g}$ be the flattening of g over the rectangles in $\mathcal{R}$. Then, if h^* is constant on every rectangle in $\mathcal{R}$, we have*

$$\|g-\overline{g}\|_{2,\mathcal{R}}^2 \leq \|g-h^*\|_{2,\mathcal{R}}^2$$

Proof This follows immediately from Fact 27. ■

As an immediate corollary of this lemma, we get that

$$\|g-h\|_{2,\mathcal{F}}^2 \leq \mathrm{OPT}_{\mathcal{D},k}(\mathcal{F})\ . \tag{4}$$

Error on $\mathcal{J}_1$ By definition, we have

$$\|g-h\|_{2,\mathcal{J}_1}^2 = 0\ . \tag{5}$$

Error on $\mathcal{J}_2$ Finally, we bound the error $\mathcal{J}_2$. Fix any $R \in \mathcal{J}_2$. Observe that R cannot be an indivisible rectangle, as then otherwise $e_R = 0$ and so $R \in \mathcal{J}_1$ or $R \in \mathcal{F}$. Therefore, in some iteration, there must be some R' so that $R \subseteq R'$ so that R' was not split in this iteration. Let $A_1, \ldots, A_{(1+\xi)k}$ be the rectangles which were split in this iteration. Because the rectangles are dyadic, they are disjoint. Thus, h^* can be non-constant on at most k of them. WLOG assume that h^* is constant on $A_1, \ldots, A_{\xi k}$. Let q be the flattening of g over $\mathcal{A} = \{A_1, \ldots, A_{\xi k}\}$. We then have

$$\begin{aligned} \|g - h\|_{2,R}^2 &\overset{(a)}{\leq} \|q - \overline{g_{R'}}\|_{2,R'}^2 \\ &\leq \frac{1}{\xi k} \sum_{i=1}^{\xi k} \|g - \overline{g_{A_i}}\|_{2,A_i}^2 \\ &\overset{(b)}{\leq} \frac{1}{\xi k} \mathrm{OPT}_{\mathcal{D},k} . \end{aligned}$$

where (a) follows from the fact that h is the optimal ℓ_2 fit to g on R, and (b) follows from Lemma 32.

Summing over the elements in $\mathcal{J}_2$, we obtain that

$$\|g - h\|_{2,\mathcal{J}_2}^2 \leq \frac{1}{\xi} \mathrm{OPT}_{\mathcal{D},k} . \tag{6}$$

Combining (4), (5), and (6) and simplifying yields that

$$\|g - h\|_2^2 \leq \left(1 + \frac{1}{\xi}\right) \mathrm{OPT}_{\mathcal{D},k} ,$$

as claimed. ■

Lemmas 29, 30, and 31 together immediately imply Theorem 28.

Appendix B. Omitted Proofs from Section 3

B.1. VC Theory for Hierarchical Histograms

We first characterize exactly the structure of the difference between two hierarchical histograms that respect the dyadic decomposition on $[m]^d$:

Lemma 33 *Let f, g be two k-hierarchical histograms with respect to a grid $\mathcal{G}$. Then $f - g$ is a $2k$-hierarchical histogram.*

Proof Let $R_1, \ldots, R_k$ and $R'_1, \ldots, R'_k$ be rectangles in the hierarchical structure so that f is flat on each rectangle R_i and g is flat on each rectangle R'_i. For each pair of rectangles R_i and R'_j, we have that either $R_i \subseteq R'_j$ (or vice versa), or $R_i \cap R'_j = \varnothing$. Thus, if we choose a maximal subset $\mathcal{S}$ of $\{R_1, \ldots, R_k, R'_1, \ldots, R'_k\}$ so that (1) there do not exist $R, R' \in \mathcal{S}$ so that $R \subseteq R'$, and moreover, (2) there does not exist a $R \in \mathcal{S}$ and $R' \in \{R_1, \ldots, R_k, R'_1, \ldots, R'_k\}$ so that $R' \subset R$, then it is a partition of $[m]^d$ that consists of at most $2k$ rectangles that respect the hierarchical structure. Moreover, it is easy to see that $f - g$ is flat on every rectangle in $\mathcal{S}$. This completes the proof. ■

Lemma 33 immediately implies Corollary 14.

We also wish to instantiate these bounds for rectangles, and unions of at most k rectangles. Fortunately, the VC dimension of rectangles and unions is well-understood:

Lemma 34 (c.f. Devroye & Lugosi Lemma 4.1) *If $\mathcal{R}$ is the set of axis-aligned rectangles in $\mathbb{R}^d$, then $\mathrm{VC}(\mathcal{R}) = 2d$.*

Lemma 35 (c.f. Devroye & Lugosi Exercise 4.1) *For any two sets of sets $\mathcal{A}_1, \mathcal{A}_2$, we have $\mathrm{VC}(\mathcal{A}_1 \cup \mathcal{A}_2) \leq \mathrm{VC}(\mathcal{A}_1) + \mathrm{VC}(\mathcal{A}_2) + 1$, where*

$$\mathcal{A}_1 \cup \mathcal{A}_2 = \{A_1 \cup A_2 : A_1 \in \mathcal{A}_1, A_2 \in \mathcal{A}_2\} \ .$$

Together, these two lemmas imply Corollary 15.

Proof [Proof of Corollary 15] Let $\mathcal{R}$ denote the set of all axis aligned rectangles in $[m]^d$. The above two lemmas immediately imply that the class $\mathcal{R}_k = \bigcup_{i=1}^k \mathcal{R}$ has $\mathrm{VC}(\mathcal{R}_k) \leq 2dk + k$. Since $\mathcal{D}_k \subseteq \mathcal{R}_k$, the result follows immediately. ■

B.2. Proof of Lemma 10

Our algorithm is given in Algorithm 3. Recall that any rectangle $R \subseteq [m]^d$ (resp. $[0,1]^d$), we let $|R|$ denote its measure in $[m]^d$ (resp. $[0,1]^d$).

Algorithm 3 Approximating with histograms by splitting.

```
1: function ComputeD1($\widehat{f}, \mathcal{D}, R, a$)
2:     Let $\mathcal{T}$ be the tree of rectangles in $\mathcal{D}$ containing points in $\mathrm{supp}(\widehat{f}) \cap R$.
3:     For every rectangle $R'$ in $\mathcal{T}$, let $c(R') = |\mathrm{supp}(\widehat{f}) \cap R'|$
4:     Let $b_1 = \max_{R' \in \mathcal{T}} |c(R') - a \cdot |R'||$, and let $R_1$ be the rectangle which achieves this maxima.
5:     Let $b_2 = a \cdot |R'|$, where $R'$ is the rectangle with maximum volume not in $\mathcal{T}$, and let $R_2$ be
       the rectangle which achieves this maxima.
6:     return $\max(b_1, b_2)$ and the corresponding $R_1$ or $R_2$
```

Proof [Proof of Lemma 10] We first prove the claimed runtime bound. Observe that $\mathcal{T}$ has size at most $O(2^d s \log M)$, and by a simple recursive splitting procedure, can be generated in $O(2^d s \log M)$ time. Similarly $c(v)$ can be computed for every node in $\mathcal{T}$ in $O(s \log M)$ time overall. Therefore b_1 can be computed in $O(2^d s \log M)$ time overall. To compute b_2, it suffices to find the largest rectangle R' in $\mathcal{T}$ which does not have 2^d children in $\mathcal{T}$, and to return $\mathrm{vol}(R')/2^d$. This again can be done by iterating over the tree once, so this takes time $O(2^d s \log M)$. Therefore overall the algorithm runs in time $O(2^d s \log M)$.

We now show correctness of the algorithm. Let R' be the rectangle which achieves the maxima for the $\mathcal{D}$-distance. There are two cases. if $R' \cap \mathrm{supp}(\widehat{f}) \neq \varnothing$, then clearly it is considered in Line 3 of ComputeA1, and its contribution is considered in the distance computation. Otherwise, R' must be a rectangle with maximum volume not in $\mathcal{T}$, as otherwise we may increase the value of the maxima by taking such a rectangle. Therefore it is considered in Line 5. In either case, its contribution is considered, and thus the algorithm is correct. ■

B.3. Proof of Theorem 13

For any set R, let $\mathrm{OPT}_k(R) = \|h^* - f\|_{1,R}$ be the ℓ_1-error incurred by h^* to f on R. Similarly, let $\widetilde{\mathrm{OPT}}_{\mathcal{D},k}(R)$ be the $\mathcal{D}_k$-error incurred by the best fit hierarchical k-histogram to $\widehat{f}$ on R. For any collection of sets $\mathcal{S}$, let $\mathrm{OPT}_k(\mathcal{S}) = \mathrm{OPT}_k(\cup_{S \in \mathcal{S}} S)$ and let $\widetilde{\mathrm{OPT}}_{\mathcal{D},k}(\mathcal{S})$ be defined similarly. We now have all definitions we need for the proof.

Proof [Proof of Theorem 13] The proof of the bounds on the number of pieces and runtime are nearly identical to the proofs of Lemmas 29 and 30, so we omit them. Thus it suffices to prove correctness. This is also quite similar to the proof of correctness for ℓ_2. Let h^* be an optimal partial hierarchical k-histogram fit to $\widehat{f}$ in $\mathcal{D}_k$ norm, and let h be the output of our algorithm. Let $\mathcal{T}$ be the tree associated with h. Let $\mathcal{R}^*$ be the set of k disjoint dyadic rectangles on which h^* is supported, and let $\mathcal{R}$ be the leaves of $\mathcal{T}$. Partition $\mathcal{R}$ into three sets:

$$\begin{aligned}
\mathcal{F} &= \{R \in \mathcal{R} : \; h^* \text{ is constant on } R \;\} \\
\mathcal{J}_1 &= \{R \in \mathcal{R} : \; h^* \text{ is non-constant on } R \text{ and } e_R = 0 \;\} \\
\mathcal{J}_2 &= \{R \in \mathcal{R} : \; h^* \text{ is non-constant on } R \text{ and } e_R > 0 \;\} \; .
\end{aligned}$$

We will prove that the error is low on all three sets separately.

Error on $\mathcal{F}$ First, we will prove that the error is low in $\mathcal{F}$. In fact, we will prove a more general lemma which will be useful later:

Lemma 36 *Let $\widehat{f}$ be an empirical distribution, and let κ be arbitrary. Let $\mathcal{R} \in \mathcal{D}_k$ be any collection of at most κ disjoint rectangles in $\mathcal{D}$, and let g be the function which is, on every $R \in \mathcal{R}$, equal to the constant function $\phi_{a,R}$ which minimizes $\|\widehat{f} - \phi_{a,R}\|_{\mathcal{D},R}$. Then, if h^* is constant on every rectangle in $\mathcal{R}$, we have*

$$\|\widehat{f} - g\|_{\mathcal{D}_\kappa,\mathcal{R}} \leq 3\|\widehat{f} - h^*\|_{\mathcal{D}_\kappa,\mathcal{R}}$$

Proof By a triangle inequality, we have

$$\|\widehat{f} - g\|_{\mathcal{D}_\kappa,\mathcal{R}} \leq \|\widehat{f} - h^*\|_{\mathcal{D}_\kappa,\mathcal{R}} + \|h^* - g\|_{\mathcal{D}_\kappa,\mathcal{R}} \; .$$

Observe that on every rectangle $R \in \mathcal{R}$, both functions are constant. Hence

$$\begin{aligned}
\|h^* - g\|_{\mathcal{D}_\kappa,\mathcal{R}} &= \sum_{R \in \mathcal{R}} |h^*(R) - g(R)| \\
&= \sum_{R \in \mathcal{R}} \|h^* - g\|_{\mathcal{D},R} \\
&\leq \sum_{R \in \mathcal{R}} \|h^* - \widehat{f}\|_{\mathcal{D},R} + \sum_{R \in \mathcal{R}} \|g - \widehat{f}\|_{\mathcal{D},R} \\
&\leq 2 \sum_{R \in \mathcal{R}} \|h^* - \widehat{f}\|_{\mathcal{D},R} \\
&\leq 2\|\widehat{f} - h^*\|_{\mathcal{D}_k,\mathcal{R}} \; .
\end{aligned}$$

Putting these two inequalities together yields the desired estimate. ■

As an immediate corollary of this lemma, we get that

$$\|\widehat{f} - h\|_{\mathcal{D}_k,\mathcal{F}} \leq 3 \cdot \widetilde{\mathrm{OPT}}_{\mathcal{D},k}(\mathcal{F}) \; . \tag{7}$$

Error on $\mathcal{J}_1$ We next consider the error on $\mathcal{J}_1$. We will require the following elementary fact, which follows immediately from the definition of $\| \cdot \|_{\mathcal{D}_k}$:

Fact 37 *For any $k \geq 1$, and for any $\mathcal{R}$, we have $\|f\|_{\mathcal{D}_k, \mathcal{R}} \leq k \|f\|_{\mathcal{D}, \mathcal{R}}$.*

This immediately implies that

$$\|\widehat{f} - h\|_{\mathcal{D}_k, \mathcal{J}_1} = 0 \ . \tag{8}$$

Error on $\mathcal{J}_2$ Thus it suffices to bound the error on $\mathcal{J}_2$. By a triangle inequality, we have

$$\begin{aligned} \|\widehat{f} - h\|_{\mathcal{D}_k, \mathcal{J}_2} &\leq \widetilde{\mathrm{OPT}}_{\mathcal{D},k}(\mathcal{J}_2) + \|h^* - h\|_{\mathcal{D}_k, \mathcal{J}_2} \\ &\leq \widetilde{\mathrm{OPT}}_{\mathcal{D},k}(\mathcal{J}_2) + \|h^* - h\|_{1, \mathcal{J}_2} \\ &= \widetilde{\mathrm{OPT}}_{\mathcal{D},k}(\mathcal{J}_2) + \sum_{R \in \mathcal{J}_2} \|h^* - h\|_{1,R} \ . \end{aligned}$$

For any $R \in \mathcal{J}_2$, let $\Gamma(R) + 1$ denote the number of values that h^* takes on R. Note that

$$\begin{aligned} \|h^* - h\|_{1,R} &= \|h^* - h\|_{\mathcal{D}_{\Gamma(R)+1}, R} \\ &\leq \|h^* - \widehat{f}\|_{\mathcal{D}_{\Gamma(R)+1}, R} + \|\widehat{f} - h\|_{\mathcal{D}_{\Gamma(R)+1}, R} \\ &\leq \widetilde{\mathrm{OPT}}_{\mathcal{D},k}(R) + (\Gamma(R) + 1) \|\widehat{f} - h\|_{\mathcal{D}, R} \ . \end{aligned}$$

Observe that R cannot be an indivisible rectangle, as then otherwise $e_R = 0$ and so $R \in \mathcal{J}_1$ or $R \in \mathcal{F}$. Therefore, in some iteration, there must be some R' so that $R \subseteq R'$ so that R' was not split in this iteration. Let $\phi_{a,R'}$ be the optimal constant fit in $\mathcal{D}$ distance to $\widehat{f}$ on R'. Let $A_1, \ldots, A_{(1+\xi)k}$ be the rectangles which were split in this iteration. Since these rectangles are disjoint, this means that h^* can be non-constant on at most k of them. WLOG assume that h^* is constant on $A_1, \ldots, A_{\xi k}$. Let g be the optimal fit in $\mathcal{D}$ to $\widehat{f}$ over each rectangle in $\mathcal{A} = \{A_1, \ldots, A_{\xi k}\}$. We then have

$$\begin{aligned} \|\widehat{f} - h\|_{\mathcal{D}, R} &\overset{(a)}{\leq} \|\widehat{f} - \phi_{a,R'}\|_{\mathcal{D}, R} \\ &\leq \|\widehat{f} - \phi_{a,R'}\|_{\mathcal{D}, R'} \\ &\leq \frac{1}{\xi k} \sum_{i=1}^{\xi k} \|\widehat{f} - g\|_{\mathcal{D}, A_i} \\ &\overset{(b)}{\leq} \frac{1}{\xi k} \|\widehat{f} - g\|_{\mathcal{D}_{\xi k}, \mathcal{A}} \\ &\overset{(c)}{\leq} \frac{3}{\xi k} \|\widehat{f} - h^*\|_{\mathcal{D}_{\xi k}, \mathcal{A}} \\ &\overset{(d)}{\leq} \frac{3}{\xi^2 k} \widetilde{\mathrm{OPT}}_{\mathcal{D},k} \ . \end{aligned}$$

where (a) follows from the fact that h is the optimal $\mathcal{D}$ fit to $\widehat{f}$ on R, (b) follows from the definition of $\| \cdot \|_{\mathcal{D}_k}$, (c) follows from Lemma 36, and (d) follows from Fact 37. Hence, overall we have

$$\|h^* - h\|_{1,R} \leq \widetilde{\mathrm{OPT}}_{\mathcal{D},k}(R) + \frac{3}{\xi^2 k} \widetilde{\mathrm{OPT}}_{\mathcal{D},k} \ .$$

Summing over the elements in $\mathcal{J}_2$, we obtain that

$$\begin{aligned} \|h^* - h\|_{1,\mathcal{J}_2} &\leq \widetilde{\mathrm{OPT}}_{\mathcal{D},k}(\mathcal{J}_2) + \frac{3}{\xi^2 k}\widetilde{\mathrm{OPT}}_{\mathcal{D},k} \sum_{R \in \mathcal{J}_2} (\Gamma(R) + 1) \\ &\leq \widetilde{\mathrm{OPT}}_{\mathcal{D},k}(\mathcal{J}_2) + \frac{6}{\xi^2}\widetilde{\mathrm{OPT}}_{\mathcal{D},k} \; . \end{aligned}$$

Hence overall, we have

$$\|\widehat{f} - h\|_{\mathcal{D}_k, \mathcal{J}_2} \leq 2\widetilde{\mathrm{OPT}}_{\mathcal{D},k}(\mathcal{J}_2) + \frac{6}{\xi^2}\widetilde{\mathrm{OPT}}_{\mathcal{D},k} \; . \tag{9}$$

Combining (7), (8), and (9) and simplifying yields that

$$\|\widehat{f} - h\|_{\mathcal{D}_k} \leq \left(3 + \frac{6}{\xi^2}\right)\widetilde{\mathrm{OPT}}_{\mathcal{D},k} \; ,$$

as claimed. ■

B.4. Proof of Lemma 20

Proof [Proof of Lemma 20] Let T be a finite set of size r. If our family can shatter T, then all 2^r subsets of T must be expressible in the form

$$T \cap \left(\bigcup_{i=1}^{k'} A_i - \bigcup_{j=1}^{k''} B_j \right) = \left(\bigcup_{j=1}^{k'} A_j \cap T \right) - \left(\bigcup_{j=1}^{k''} B_j \cap T \right) \tag{10}$$

We now count the number of possible sets of the form $R \cap T$ for rectangles R. Observe that each face has a fixed normal, and for halfspaces H with a fixed normal there are clearly at most r possible sets $H \cap T$. R has $2d$ faces and so the number of possible sets of the form $R \cap T$ is at most r^{2d}. Hence, the number of sets of the form in (10) is at most r^{4dk}. This is smaller than 2^r when r is a sufficiently large multiple of $kd \log(kd)$. Thus, the VC dimension is $O(kd \log(kd))$. ■

B.5. Proof of Lemma 22

Proof [Proof of Lemma 22] Let $R_1, \ldots, R_k$ be a partition of $[m]^d$ into k disjoint rectangles so that h^* is constant on each R_i. For each i, let R'_i be the smallest rectangle so that $R'_i \subseteq R$ and so that R'_i contains every point in $\mathrm{grid}(\mathcal{X}) \cap R_i$. Let h^*_p be the k-partial histogram so that for each i, $h^*_p(x) = h^*(x)$ for all $x \in R'_i$, and $h^*_p(x) = 0$ outside of $\bigcup_{i=1}^k R'_i$. We claim this function satisfies

$$\|f - h^*_p\|_1 \leq \mathrm{OPT}_k + c'\varepsilon \; .$$

Let $R = \bigcup_{i=1}^k R'_i$. Then, we have

$$\begin{aligned} \|f - h^*_p\|_1 &= \|f - h^*_p\|_{1,R} + \|f - h^*_p\|_{1,R^c} \overset{(a)}{=} \|f - h^*\|_{1,R} + \|f\|_{1,R^c} \\ &\overset{(b)}{=} \mathrm{OPT}_k + 2\|f\|_{\mathcal{A}_k,R^c} \overset{(c)}{=} \mathrm{OPT}_k + 2\|f - \widehat{f}\|_{\mathcal{A}_k,R^c} \overset{(d)}{\leq} \mathrm{OPT}_k + 2c'\varepsilon \; , \end{aligned}$$

where (a) follows from the decomposibility of ℓ_1, (b) follows from the definition of h_p^*, (c) follows since $R \in \mathcal{A}_\kappa$ and since $\widehat{f} = 0$ on R^c, and (d) follows from (1).

The only remaining problem with this function is that it is not a distribution, namely, it does not integrate to 1. However, we know that $\left| \|h_p^*\|_1 - 1 \right| \leq \|h_p^* - f\|_1 \leq \mathrm{OPT}_k + 2c'\varepsilon$. Hence, if we renormalize h_p^* to make it integrate to 1 (say, by adding mass uniformly to one rectangle), we lose at most an additional $\mathrm{OPT}_k + 2c'\varepsilon$ factor. The claim follows then from an easy generalization of Lemma 7, since the side length of $\mathrm{grid}(\mathcal{X})$ is $\mathrm{poly}(k, 1/\varepsilon)$. ■

B.6. Proof of Lemma 23

Proof Let $Z = \{h(x) > g(x)\}$. Let $R_1, \ldots, R_k$ be k disjoint rectangles so that h is constant on every R_i, and h is supported on their union. Let $R'_1, \ldots, R'_k$ be the same for g, except that these sets form a partition of $[m]^d$. Reminiscent of the proof of Lemma 31, partition $\mathcal{R} = \{R_1, \ldots, R_k\}$ into two sets: $\mathcal{F}$, the set of $R \in \mathcal{R}$ so that g is constant on R, and $\mathcal{J}$, the set of $R \in \mathcal{R}$ so that g has a jump on R. Clearly Z is the disjoint union of $Z_1 = Z \cap \cup_{R \in \mathcal{F}} R$ and $Z_2 = Z \cap \cup_{R \in \mathcal{J}}$. Moreover, Z_1 is immediately expressible as the disjoint union of at most k dyadic rectangles: namely, the rectangles in $\mathcal{F}$ on which $h(x) > g(x)$. Thus, it suffices to show that Z_2 can be written as a disjoint union of at most k disjoint rectangles. But if $R \in \mathcal{J}$ then this means that g partitions the rectangle. Thus, Z_2 is exactly the set of R'_j so that $h(x) > g(x)$ on R'_j, and $R'_j \subset R_i$ for some $i \in [k]$. Hence Z_2 can also be written as the union of at most k rectangles, and so Z can be written as a union of at most $2k$ rectangles, which completes the proof. ■

B.7. Algorithm 4

Algorithm 4 Adaptive greedy splitting for histogram learning in ℓ_1

```
function AdaptiveGreedySplit($\widehat{f}, \xi$)
    Let $\mathcal{X}$ be the (multi-)set of points in the support of $\widehat{f}$
    return GreedySplit($\widehat{f}, \mathcal{D}(\mathrm{grid}(\mathcal{X})), \xi$)
```