1

Learning Structured Distributions

CONTENTS

1.1 Introduction

Discovering hidden structure in data is one of the cornerstones of modern data analysis. Due to the diversity and complexity of modern data sets, this is a very challenging task and the role of efficient algorithms is of paramount importance in this context. The majority of available data sets are in raw and unstructured form, consisting of example points without corresponding labels. A large class of unlabeled datasets can be modeled as samples from a probability distribution over a very large domain. An important goal in the exploration of these datasets is understanding the underlying distributions.

Estimating distributions from samples is a paradigmatic and fundamental unsupervised learning problem that has been studied in statistics since the late nineteenth century, starting with the pioneering work of Karl Pearson [55]. During the past couple of decades, there has been a large body of work in computer science on this topic with a focus on *computational efficiency.*

The area of distribution estimation is well-motivated in its own right, and has seen a recent surge of research activity, in part due to the ubiquity of structured distributions in the natural and social sciences. Such structural properties of distributions are sometimes direct consequences of the underlying application problem, or they are a plausible explanation of the model under investigation.

In this chapter, we give a survey of both classical and modern techniques for distribution estimation, with a focus on recent algorithmic ideas developed in theoretical computer science. These ideas have led to computationally and statistically efficient algorithms for learning broad families of models. For the sake of concreteness, we illustrate these ideas with specific examples. Finally, we highlight outstanding challenges and research directions for future work.

1.2 Historical Background

The construction of an estimate of an unknown probability density function based on observed data is a classical problem in statistics with a rich history and extensive literature (see e.g., [4, 26, 60, 59, 27]). A number of generic methods have been proposed in the mathematical statistics literature, including histograms, kernels, nearest neighbor estimators, orthogonal series estimators, maximum likelihood, and more. The reader is referred to [44] for a survey of these techniques.

The oldest and most natural estimator is the histogram, first introduced by Karl Pearson [55]. Given a number of samples (observations) from a probability density function, the method partitions the domain into a number of bins, and outputs the empirical density which is constant within each bin. It should be emphasized that the number of bins to be used and the width and location of each bin are unspecified by the method. The problem of finding the optimal number and location of the bins to minimize the error is an inherently algorithmic question, since the ultimate goal is to obtain learning algorithms that are computationally efficient.

Suppose that we are given a number of samples from a density that we believe is from (or very close to) a given family $\mathcal{C}$, e.g., it is a mixture of a small number of Gaussian distributions. Our goal is to estimate the target distribution in a precise, well-defined way. There are three different goals in this context:

1. In *non-proper* learning (density estimation) the goal is to output an approximation to the target density without any constraints on its representation. That is, the output distribution is not necessarily a member of the family $\mathcal{C}$.
2. In *proper* learning the goal is to output a density in $\mathcal{C}$ that is a good approximation to the target density.
3. In *parameter* learning the goal is to identify the parameters of the target distribution, e.g., the mixing weights and the parameters of

the components up to a desired accuracy. (The notion of parameter learning is well-defined for parametric classes $\mathcal{C}$.)

Note that non-proper learning and proper learning are equivalent in terms of sample size: given any (non-proper) hypothesis we can do a brute-force search to find its closest density in $\mathcal{C}$. However, it is not clear whether this computation can be performed efficiently.

We remark that the task of parameter learning is possible only under certain separation assumptions on the components. Even under such assumptions, it can be a more demanding task than proper learning. In particular, it is possible that two distinct distributions in $\mathcal{C}$ whose parameters are far from each other give rise to densities that are close to each other. Moreover, parameter learning strongly relies on the assumption that there is no noise in the data, and hence it may not be meaningful in many realistic settings. These facts motivate the study of proper learning algorithms in the noisy setting.

The focus of this chapter is on general techniques and efficient algorithms for density estimation and proper learning. Due to space constraints, we do not elaborate on algorithmic methods used for the problem of parameter learning.

The structure of this chapter is as follows: After some basic definitions (Section 1.3), in Section 1.4 we give a classification of the types of distribution families studied in the literature. In Section 1.5 we describe a classical method from statistics to efficiently select from a given a set of candidate hypothesis distributions. Section 1.6 describes recent algorithmic ideas from theoretical computer science to learn structured univariate densities. Section 1.7 discusses the challenging case of high-dimensional distributions. We conclude with some future directions in Section 1.8.

1.3 Definitions and Preliminaries

We consider a standard notion of learning an unknown probability distribution from samples [46], which is a natural analogue of Valiant's well-known PAC model for learning Boolean functions [67] to the unsupervised setting of learning an unknown probability distribution. We remark that our definition is essentially equivalent to the notion of minimax rate of convergence in statistics [27].

Given access to independent draws from an unknown probability density function (pdf) p, the goal is to approximate p in a certain well-defined sense. More specifically, the goal of a learning algorithm is to output a hypothesis distribution h that is "close" to the target distribution p. One can choose various metrics to measure the distance between distributions. Throughout this chapter, we measure the closeness between distributions using the *statistical distance* or total variation distance. The *statistical distance* between two densities $p, q : \Omega \to \mathbb{R}_+$ is defined as:

$$d_{\mathrm{TV}}(p, q) = \frac{1}{2}\|p - q\|_1 = \frac{1}{2}\int_\Omega |p(x) - q(x)|dx.$$

(When Ω is discrete the above integral is replaced by a sum.)

A distribution learning problem is defined by a class $\mathcal{C}$ of probability distributions over a domain Ω. The domain Ω may be discrete, e.g., $\Omega = [n] := \{1, \ldots, n\}$, or continuous, e.g., $\Omega = \mathbb{R}$, one-dimensional or high-dimensional. In the "noiseless" setting, we are promised that $p \in \mathcal{C}$ and the goal is to construct a hypothesis h such that with probability at least 9/10[1] the total variation distance $d_{\mathrm{TV}}(h, p)$ between h and p is at most ϵ, where $\epsilon > 0$ is the accuracy parameter.

The "noisy" or *agnostic* model captures the situation of having adversarial noise in the data. In this setting, we do not make any assumptions about the target density p and the goal is to find a hypothesis h that is almost as accurate as the "best" approximation of p by any distribution in $\mathcal{C}$. Formally, given $\epsilon > 0$ and sample access to a target distribution p, the goal of an *agnostic learning algorithm for* $\mathcal{C}$ is to compute a hypothesis distribution h such that, with probability at least 9/10, it holds

$$d_{\mathrm{TV}}(h, p) \le \alpha \cdot \mathrm{opt}_{\mathcal{C}}(p) + \epsilon,$$

where $\mathrm{opt}_{\mathcal{C}}(p) := \inf_{q \in \mathcal{C}} d_{\mathrm{TV}}(q, p)$, i.e., $\mathrm{opt}_{\mathcal{C}}(p)$ is the statistical distance between p and the closest distribution to it in $\mathcal{C}$, and $\alpha \ge 1$ is a universal constant.

We will use the following two standard metrics to measure the performance of a learning algorithm: (i) the *sample complexity*, i.e., the number of samples drawn by the algorithm, and (ii) the *computational complexity*, i.e., the worst-case running time of the algorithm. An algorithm is statistically efficient if its sample complexity is information–theoretically optimal, and it is computationally efficient if its computational complexity is polynomial in its sample complexity. The "gold standard" is a statistically efficient algorithm whose computational complexity is linear in its sample size.

As mentioned in the introduction, proper and non-proper learning of any class $\mathcal{C}$ are equivalent in terms of sample complexity, but not necessarily equivalent in terms of computational complexity. We also remark that, for broad classes of distributions $\mathcal{C}$, agnostic learning and noiseless learning are equivalent in terms of sample complexity. However, designing computationally efficient agnostic learning algorithms is in general a much more challenging task.

[1] We note that, using standard techniques, the confidence probability can be boosted to $1 - \delta$, for any $\delta > 0$, with a multiplicative overhead of $O(\log(1/\delta))$ in the sample size.

1.4 Types of Structured Distributions

In this section we provide a broad categorization of the most common types of structured distributions that have been considered in the statistics and computer science literatures. We also briefly summarize a few standard methods to learn such distributions in statistics.

In the following sections, we will describe a set of algorithmic techniques that lead to provably efficient learning algorithms for most of these distribution families.

Shape Constrained Distributions. For distributions over $\mathbb{R}^d$ (or a discrete d-dimensional subset, e.g., $[n]^d$), a very natural type of structure to consider is some sort of "shape constraint" on the probability density function (pdf) defining the distribution.

Statistical research in this area started in the 1950's, and the reader is referred to the book [4] for a summary of the early work. Most of the literature has focused on one-dimensional distributions, with a few exceptions during the past decade. Various structural restrictions have been studied over the years, starting from monotonicity, unimodality, convexity, and concavity [38, 9, 56, 71, 41, 39, 6, 7, 35, 12, 45], and more recently focusing on structural restrictions such as log-concavity and k-monotonicity [2, 32, 1, 37, 3, 49, 69]. The reader is referred to [40] for a recent book on the subject.

The most common method used in statistics to address shape constrained inference problems is the Maximum Likelihood Estimator (MLE) and its variants. The challenge is to analyze the performance of the MLE in this context. It turns out that for several univariate learning problems of this sort the MLE performs quite well in terms of statistical efficiency. While the MLE is very popular and quite natural, there exist natural inference problems (see, e.g., [10]) where it performs poorly in terms of statistical and computational efficiency, as well as noise tolerance.

A related line of work in mathematical statistics [47, 29, 48, 30, 28] uses non-linear estimators based on wavelet techniques to learn continuous distributions whose densities satisfy various smoothness constraints, such as Triebel and Besov-type smoothness. We remark that the focus of these works is on the statistical efficiency of the proposed estimators and not on computational complexity.

Aggregation of Structured Distributions. Aggregations of structured random variables are very popular as they can model many rich phenomena. Two prominent examples of this sort are mixture models and sums of simple random variables. Mixtures of structured distributions have received much attention in statistics [51, 57, 64] and, more recently, in theoretical computer science [19, 54].

We remark that early statistical work on mixture models focuses on parameter learning. In practice, this problem is typically handled with non-convex heuristics such as the Expectation–Maximization (EM) algorithm. Recent algorithmic techniques rely on the moment problem and tensor decomposition. However, such algorithms lead to sample complexities that are inherently exponential in the number of components.

Learning sums of simple random variables has received recent attention in the computer science literature [23, 20]. Such distributions have various applications in areas such as survey sampling, case-control studies, and survival analysis (see e.g., [16] for the case of sums of indicators).

1.5 The Cover Method and Sample Bounds

The first fundamental question that arises in the context of learning an unknown probability distribution is information-theoretic:

What is the minimum sample size that is necessary and sufficient to learn an unknown $p \in \mathcal{C}$ up to total variation distance ϵ?

While this question has been extensively investigated in statistics, information theory, and, more recently, computer science, the information–theoretically optimal sample size is not yet understood, even for some relatively simple families of distributions. It turns out that the optimal sample complexity depends on the structure of the underlying density class in a subtle way.

In this section we describe a general powerful method that yields nearly tight upper bounds on the sample complexity of learning. The method, which we term the *cover method*, is classical in statistics and information theory, and has its roots in early work of A. N. Kolmogorov. The high-level idea is to analyze the structure of the metric space $\mathcal{C}$ under total variation distance. The method postulates that the structure of this metric space characterizes the sample complexity of learning. To describe the method in detail we introduce some basic terminology.

Let $(\mathcal{X}, d)$ be a metric space. Given $\delta > 0$, a subset $\mathcal{Y} \subseteq \mathcal{X}$ is said to be a *δ-cover of $\mathcal{X}$* with respect to the metric $d : \mathcal{X}^2 \to \mathbb{R}_+$ if for every $\mathbf{x} \in \mathcal{X}$ there exists some $\mathbf{y} \in \mathcal{Y}$ such that $d(\mathbf{x}, \mathbf{y}) \leq \delta$. There may exist many δ-covers of $\mathcal{X}$, but one is typically interested in those with minimum cardinality. The *δ-covering number* of $(\mathcal{X}, d)$ is the minimum cardinality of any δ-cover of $\mathcal{X}$. Intuitively, the covering number captures the "size" of the metric space.

Covering numbers – and their logarithms, known as *metric entropy* numbers – were first defined by Kolmogorov in the 1950's and have since played a central role in a number of areas, including approximation theory, geometric functional analysis (see, e.g., [31, 53, 8] and the books [50, 52, 11, 33]), information theory, statistics, and machine learning (see, e.g., [74, 5, 42, 43, 73] and the books [68, 27, 65]).

In the context of distribution learning, the cover method is summarized in the following theorem:

Theorem 1.5.1. *Let $\mathcal{C}$ be an arbitrary family of distributions and $\epsilon > 0$. Let $\mathcal{C}_\epsilon \subseteq \mathcal{C}$ be an ϵ-cover of $\mathcal{C}$ of cardinality N. Then there is an algorithm that uses $O(\epsilon^{-2} \log N)$ samples from an unknown distribution $p \in \mathcal{C}$ and, with probability at least $9/10$, outputs a distribution $h \in \mathcal{C}_\epsilon$ that satisfies $d_{\mathrm{TV}}(h, p) \leq 6\epsilon$.*

An equivalent version of Theorem 1.5.1 (with a slightly different terminology) was given by Yatracos [74] (see also Chapter 7 of [27] for a detailed discussion). The above statement appears as Lemma C.1 in [21].

As we explain in detail below, the algorithm implicit in the above theorem is *not* computationally efficient in general. Indeed, even assuming that we have an explicit construction of a minimal size ϵ-cover, the algorithm takes time at least $\Omega(N/\epsilon^2)$ – that is, *exponential* in its sample size.

We point out that the cover method can serve as a very useful tool in the design of computationally efficient learning algorithms. Indeed, many algorithms in the literature work by constructing a *small* set S of candidate hypotheses with the guarantee that at least one of them is close to the target distribution. The cover method can be used as a post-processing step to efficiently select an appropriate candidate in the set S. This simple idea has been used in the design of fast proper learning algorithms for various natural classes of distributions, including sums of independent integer random variables [23, 20], Gaussian mixtures [24, 63], and other high-dimensional distributions [25].

We now provide a brief intuitive explanation of the argument in [21] establishing Theorem 1.5.1. (The corresponding proof of [74, 27] is quite similar.) Given a description of the cover $\mathcal{C}_\epsilon$, the algorithm performs a tournament between the distributions in $\mathcal{C}_\epsilon$, by running a hypothesis testing routine for every pair of distributions in $\mathcal{C}_\epsilon$. The obvious implementation of this tournament takes time $\Omega(N^2/\epsilon^2)$. Recent algorithmic work [24, 63] has improved this to nearly-linear in N, namely $O(N \log N/\epsilon^2)$. However, this running time bound is still exponential in the sample complexity of the algorithm.

The hypothesis testing routine can be viewed as a simple "competition" between two candidate hypothesis distributions. If at least one of the two candidate hypotheses is close to the target distribution p, then with high probability over the samples drawn from p the hypothesis testing routine selects as winner a candidate that is close to p. The algorithm outputs a distribution in the cover $\mathcal{C}_\epsilon$ that was never a loser (i.e., won or tied against all other distributions in the cover). We remark that the analysis of the algorithm is elementary, relying only on the Chernoff bound and the union bound.

Another important property of the cover method is its noise tolerance. It generalizes naturally yielding an agnostic learning algorithm with the same sample complexity. More specifically, for an arbitrary target distribution p with $\mathrm{opt}_{\mathcal{C}}(p) = \inf_{q \in \mathcal{C}} d_{\mathrm{TV}}(q, p)$, the tournament–based algorithm makes $O(\epsilon^{-2} \log N)$ i.i.d. draws from p and outputs a hypothesis h in $\mathcal{C}_\epsilon$ satisfying

$d_{\mathrm{TV}}(h,p) \leq O(\mathrm{opt}_{\mathcal{C}}(p)+\epsilon)$. The reader is referred to Chapter 7.3 of [27] for an explicit proof of this fact.

The sample upper bound of $O(\epsilon^{-2}\log N)$ cannot be improved in general, in the sense that there exist distribution families where it is information–theoretically optimal up to constant factors. In fact, Yang and Barron [73] showed that for many smooth nonparametric classes the metric entropy number characterizes the sample complexity of learning. We note, however, that metric entropy does not provide a characterization in general: there exist distribution families where the $O(\epsilon^{-2}\log N)$ sample upper bound is sub-optimal.

As a simple example consider the set of all "singleton" distributions over $[n]$, i.e., the class contains n distinct distributions each supported on a single point of the domain. It is easy to see that Theorem 1.5.1 gives a sample upper bound of $O(\epsilon^{-2}\log n)$ for this case, while one sample suffices to uniquely specify the target distribution. For a more natural example, consider the class of Poisson Binomial Distributions (PBDs), i.e., sums $\sum_{i=1}^{n} X_i$ of n mutually independent Bernoulli random variables $X_1,\ldots,X_n$. It is not difficult to show that the covering number of the set of PBDs is $\Omega(n/\epsilon)$. Hence, Theorem 1.5.1 cannot give an upper bound better than $\widetilde{\Omega}(\epsilon^{-2})\cdot\log n$. On the other hand, a sample upper bound of $\widetilde{O}(\epsilon^{-2})$ was recently obtained in [23]. These examples raise the following natural question:

Open Problem 1.5.1. *Is there a "complexity measure" of a distribution class $\mathcal{C}$ that* characterizes *the sample complexity of learning $\mathcal{C}$?*

We recall that the Vapnik–Chervonenkis dimension of a class of Boolean functions plays such a role in Valiant's PAC model [66], i.e., it tightly characterizes the number of examples that are required to PAC learn an arbitrary function from the class.

1.6 Learning Univariate Structured Distributions

In this section we consider the problem of non-proper learning of an unknown univariate probability distribution, i.e., a distribution with a density function $p:\Omega\to\mathbb{R}_+$, where the sample space Ω is a subset of the real line. We focus on two basic cases: (i) $\Omega=[n]$ where the set $[n]$ is viewed as an ordered set, and (ii) $\Omega=[a,b]$ with $a\leq b\in\mathbb{R}$. Given a family $\mathcal{C}$ of univariate distributions, can we design a sample optimal and computationally efficient learning algorithm for $\mathcal{C}$? Can we achieve this goal in the more challenging agnostic setting? It turns out that the answer to both questions turns out to be "yes" for broad classes of structured families $\mathcal{C}$.

If the target distribution is arbitrary, the learning problem is well-understood. More specifically, suppose that the class $\mathcal{C}$ of target distributions is the set of all distributions over $[n]$. It is a folklore fact that $\Theta(n/\epsilon^2)$ samples are necessary and sufficient for learning within total variation distance ϵ in this case. The underlying algorithm is also straightforward: output the empirical distribution. For distributions over very large domains, a linear dependence on n is of course impractical, both from running time and sample complexity perspective.

For continuous distributions the learning problem is not solvable without any assumptions. Indeed, learning an arbitrary distribution over $[0, 1]$ to any constant accuracy $\epsilon < 1$ requires infinitely many samples. This follows, for example, from the aforementioned discrete lower bound for $n \to \infty$. Hence, it is important to focus our attention on structured distribution families.

In the main part of this section we describe recent work from theoretical computer science that yields sample–optimal and computationally efficient algorithms for learning broad classes of structured distributions. The main idea of the approach is that the *existence* of good piecewise polynomial approximations for a family $\mathcal{C}$ can be leveraged for the design of efficient learning algorithms for $\mathcal{C}$. The approach is inspired and motivated by classical results in statistics, and combines a variety of techniques from algorithms, probability, and approximation theory.

Piecewise polynomials (splines) have been extensively used in statistics as tools for inference tasks, including density estimation, see, e.g., [70, 72, 61, 62]. We remark that splines in statistics have been used in the context of the MLE, which is very different than the aforementioned approach. Moreover, the degree of the splines used in statistical literature is typically bounded by a small constant.

In Section 1.6.1 we describe classical work in statistics on learning monotone densities that served as an inspiration for the piecewise polynomial approach. In Section 1.6.2 we describe how to use piecewise constant approximations for learning and argue why it is insufficient for some cases. Finally, in Section 1.6.3 we describe the general approach in detail.

1.6.1 Learning Monotone Distributions

Monotonicity is arguably one of the simplest shape constraints. Learning a monotone density was one of the first problems studied in this context by Grenander [38]. We present a result by Birgé [6, 7] who gave a sample–optimal and computationally efficient algorithm for this problem. More specifically, Birgé showed the following:

Theorem 1.6.1 ([6, 7]). *Fix $L, H > 0$, let $\mathcal{M}$ be the set of non-increasing densities $p : [0, L] \to [0, H]$. There is a computationally efficient algorithm that given $m = O((1/\epsilon^3) \log(1 + H \cdot L))$ samples from an arbitrary $p \in \mathcal{M}$ outputs a hypothesis h satisfying $d_{\mathrm{TV}}(h, p) \le \epsilon$ with probability at least $9/10$. Moreover, $\Omega((1/\epsilon^3) \log(1 + H \cdot L))$ samples are information-theoretically necessary for this problem.*

An adaptation of the above theorem holds for monotone distributions over $[n]$, yielding an efficient algorithm with optimal sample complexity of $O((1/\epsilon^3)\log n)$ for the discrete setting as well.

To sketch the proof of this theorem, we will need a few definitions. Given m independent samples $s_1, \ldots, s_m$, drawn from a density $p : \Omega \to \mathbb{R}_+$ the *empirical distribution* $\widehat{p}_m$ is the discrete distribution supported on $\{s_1, \ldots, s_m\}$ defined as follows: for all $z \in \Omega$, $\widehat{p}_m(z) = |\{j \in [m] \mid s_j = z\}|/m$.

For a measurable function $f : I \to \mathbb{R}_+$ and $A \subseteq I$ we will denote $f(A) = \int_A f(x)dx$.

Definition 1.6.2. A function $f : I \to \mathbb{R}$ is called a *t-histogram* if it is piecewise constant with at most t interval pieces. For a function $f : I \to \mathbb{R}$ and an interval partition $\{I_1, \ldots, I_t\}$ of the domain, the *flattened version* $\bar{f}$ of f is the t-histogram defined by $\bar{f}(x) = f(I_j)/|I_j|$ for all $x \in I_j$.

Birgé's algorithm works as follows [7]: it partitions the domain into a set of intervals and outputs the flattened empirical distribution on those intervals. Its correctness relies on an approximation lemma that he proves:

Lemma 1.6.3. *([7]) Fix $L, H > 0$. There exists a partition of $[0, H]$ into $t = O((1/\epsilon)\log(1+H\cdot L))$ intervals such that for any $p \in \mathcal{M}$ it holds $d_{\mathrm{TV}}(\bar{p}, p) \leq \epsilon$.*

An analogue of the lemma holds for discrete monotone distributions over $[n]$ establishing a bound of $t = O((1/\epsilon)\log n)$ on the number of intervals.

Note that the interval decomposition of the lemma is *oblivious*, in the sense that it does not depend on the underlying monotone density. This is a very strong guarantee that facilitates the learning algorithm. Indeed, given the guarantee of the lemma, the algorithm is straightforward. The monotone learning problem is *reduced* to the problem of learning a distribution over a known finite support of cardinality $t = O((1/\epsilon)\log(1 + H \cdot L))$.

In summary, one can break Birgé's approach in two conceptual steps:

- Prove that any monotone distribution is ϵ-close in total variation distance to a t-histogram distribution, where the parameter t is small.
- Agnostically learn the target distribution using the class of t-histogram distributions as a hypothesis class.

This scheme is quite general and can be applied to any structured distribution class as long as there exists a good piecewise constant approximation. In general, such a histogram approximation may not be fixed for all distributions in the family. Indeed, this is the case for most natural families of distributions. To handle this case, we need an agnostic learning algorithm for t-histogram distributions with an *unknown* partition.

1.6.2 Agnostically Learning Histograms

In this section, we study the problem of agnostically learning t-histogram distributions with an unknown partition. Formally, given a bound t on the number of intervals, we want to design a computationally efficient algorithm that uses an optimal sample size and approximates the target distribution nearly as accurately as the best t-histogram. As sketched in the previous section, such an algorithm would have several applications in learning classes of shape restricted densities.

Denote by $\mathcal{H}_t$ the family of t-histogram distributions over $[0, 1]$. [2] The first step is to determine the optimal sample complexity of the learning problem. It is easy to see that $\Omega(t/\epsilon^2)$ is a lower bound and simple arguments can be used to get an upper bound of $\tilde{O}(t/\epsilon^2)$ using the cover method described in Section 1.5.

The problem of agnostically learning t-histogram distributions with $\tilde{O}(t/\epsilon^2)$ samples and $\mathrm{poly}(t/\epsilon)$ time[3] is algorithmically non-trivial. If one is willing to relax the sample size to $O(t/\epsilon^3)$, it is easy to obtain a computationally efficient algorithm [22, 13]. The first efficient algorithm with near-optimal sample complexity was obtained in [14] and is based on dynamic programming.

To sketch the algorithm in [14] we will need a more general metric between distributions that generalizes the total variation distance. Fix a family of subsets $\mathcal{A}$ over $[0, 1]$. We define the *$\mathcal{A}$–distance* between p and q by $\|p-q\|_{\mathcal{A}} := \max_{A\in\mathcal{A}} |p(A)-q(A)|$. (Note that if $\mathcal{A}$ is the set of all measurable subsets of the domain, the $\mathcal{A}$–distance is identical to the total variation distance.) The *VC–dimension* of $\mathcal{A}$ is the maximum size of a subset $X \subseteq [0, 1]$ that is shattered by $\mathcal{A}$ (a set X is shattered by $\mathcal{A}$ if for every $Y \subseteq X$ some $A \in \mathcal{A}$ satisfies $A \cap X = Y$).

The VC inequality. Fix a family of subsets $\mathcal{A}$ over $[n]$ of VC-dimension d. The *VC inequality* is the following result from empirical process theory:

Theorem 1.6.4 ([27, p.31])**.** *Let $\widehat{p}_m$ be an empirical distribution of m samples from p. Let $\mathcal{A}$ be a family of subsets of VC–dimension d. Then*

$$\mathbb{E}\left[\|p - \widehat{p}_m\|_{\mathcal{A}}\right] \leq O(\sqrt{d/m}).$$

In other words, for $m = \Omega(d/\epsilon^2)$, with probability 9/10 the empirical distribution $\widehat{p}_m$ will be ϵ-close to p in $\mathcal{A}$-distance. We remark that this sample bound is asymptotically optimal (up to a constant factor) for all values of d and ϵ.

Let $\mathcal{A}_k$ be the collection of all subsets of the domain that can be expressed as unions of at most k (disjoint) intervals. The intuition is that the collection $\mathcal{A}_{2t}$ characterizes t-histograms in a precise way. Consider the following algorithm for agnostically learning a distribution p:

(i) Draw $m = \Theta(t/\epsilon^2)$ samples from p;

[2] We choose the domain to be $[0, 1]$ for simplicity. All the results that we will describe extend straightforwardly to distributions over any interval or over a discrete set.

[3] We use the notation $\mathrm{poly}(x)$, $x \in \mathbb{R}_+$, to denote a function that is bounded from above by a fixed degree polynomial in x.

(ii) Output the distribution $h \in \mathcal{H}_t$ that minimizes the quantity $\|h - \widehat{p}_m\|_{\mathcal{A}_k}$ (up to an additive error $\gamma = O(\epsilon)$).

It is not difficult to show that this is an agnostic learning algorithm for $\mathcal{H}_t$. The main observation needed for the proof is that the $\mathcal{A}_{2t}$ distance between two t-histograms is identical to their total variation distance.

The algorithm in [14] uses a dynamic programming approach to efficiently perform step (ii) above, and its analysis relies on the VC inequality. More recently, a near-linear time algorithm, i.e., an algorithm with running time $\tilde{O}(t/\epsilon^2)$, was developed in [15].

Applications to Learning Structured Distributions. The aforementioned agnostic learning algorithm has been used as the key algorithmic ingredient to learn various classes of structured distributions. An additional ingredient needed is a structural approximation result stating that for the underlying distribution family $\mathcal{C}$ there exists an ϵ-approximation by t-histograms for an appropriately small value of the parameter t. For example, by using the structural approximation results of [13], one obtains near-sample optimal and near-linear time estimators for various well-studied classes including multimodal densities, monotone hazard rate (MHR) distributions, and others.

However, there exist distribution families where the approach of approximating by histograms *provably* leads to suboptimal sample complexity. A prominent such example is the class of log-concave distributions. This motivates the more general approach of approximating by piecewise polynomials.

1.6.3 Agnostically Learning Piecewise Polynomials

We say that a distribution q over $[0,1]$ is a *t-piecewise degree-d distribution* if there is a partition of $[0,1]$ into t disjoint intervals $I_1, \ldots, I_t$ such that $q(x) = q_j(x)$ for all $x \in I_j$, where each of $q_1, \ldots, q_t$ is a univariate polynomial of degree at most d. Let $\mathcal{P}_{t,d}$ denote the class of all t-piecewise degree-d probability density functions over $[0,1]$. We have the following theorem:

Theorem 1.6.5 ([14])**.** *Let p be any pdf over $[0,1]$. There is an algorithm that, given t, d, ϵ and $\tilde{O}(t(d+1)/\epsilon^2)$ samples from p, runs in time $\mathrm{poly}(t, d+1, 1/\epsilon)$ and with high probability outputs an $O(t)$-piecewise degree-d hypothesis h such that $d_{\mathrm{TV}}(p,h) \leq O(\mathrm{opt}_{t,d}) + \epsilon$, where $\mathrm{opt}_{t,d} := \inf_{r \in \mathcal{P}_{t,d}} d_{\mathrm{TV}}(p,r)$ is the error of the best t-piecewise degree-d distribution for p.*

It is shown in [14] that the number of samples used by the aforementioned algorithm is information–theoretically optimal in all three parameters up to logarithmic factors.

The high-level approach to prove this theorem is similar to the one described in the previous paragraph for the case of histograms. Let $\mathcal{A}_k$ be the collection of all subsets of the domain that can be expressed as unions of at most $k = 2t(d+1)$ intervals. The intuition is that the collection $\mathcal{A}_k$ characterizes piecewise polynomials with t pieces and degree d. Similarly, the following is an agnostic learning algorithm for p:

(i) Draw $m = \Theta(t(d+1)/\epsilon^2)$ samples from p;

(ii) Output $h \in \mathcal{P}_{t,d}$ that minimizes the quantity $\|h - \widehat{p}_m\|_{\mathcal{A}_k}$ (up to an additive error $\gamma = O(\epsilon)$).

We remark that the optimization problem in Step (ii) is non-convex. However, it has sufficient structure so that (an appropriately relaxed version of) it can be solved in polynomial time by a combination of convex programming and dynamic programming.

Applications to Learning Structured Distributions. Theorem 1.6.5 yields near-sample optimal and computationally efficient estimators for a very broad class of structured distribution families, including arbitrary mixtures of natural distribution families, such as multi-modal, concave, convex, log-concave, monotone hazard rate, sums of indicators, and others. Given a class $\mathcal{C}$ that we want to learn, we have the following general approach:

- Prove that any distribution in $\mathcal{C}$ is ϵ-close in total variation distance to a t-piecewise degree-d distribution, for appropriate values of t and d.
- Agnostically learn the target distribution using the class of t-piecewise degree-d distributions.

We emphasize that there are many combinations of (t, d) that guarantee an ϵ-approximation. To minimize the sample complexity of the learning algorithm in the second step, one would like to use the values that minimize the product $t(d+1)$. This is, of course, an approximation theory problem that depends on the structure of the family $\mathcal{C}$.

For example, if $\mathcal{C}$ is the class of log-concave distributions, the optimal t-histogram ϵ-approximation requires $\tilde{\Theta}(1/\epsilon)$ intervals. This leads to an algorithm with sample complexity $\tilde{\Theta}(1/\epsilon^3)$. On the other hand, it can be shown that any log-concave distribution has a piecewise *linear* ϵ-approximation with $\tilde{\Theta}(1/\epsilon^{1/2})$ intervals, which gives us a $\tilde{\Theta}(1/\epsilon^{5/2})$ sample algorithm. Perhaps surprisingly, this cannot be improved using higher degrees as one can show a sample lower bound of $\Omega(1/\epsilon^{5/2})$.

As a second example, let $\mathcal{C}$ be the class of k-mixtures of Gaussians in one dimension. By approximating these functions by piecewise polynomials of degree $O(\log(1/\epsilon))$, we obtain an efficient agnostic algorithm using $\tilde{O}(k/\epsilon^2)$ samples. This sample bound is optimal up to logarithmic factors. It should be noted that this is the first computationally efficient and sample near-optimal algorithm for this problem.

It should be emphasized that algorithm of [14] is theoretically efficient (polynomial time), but it may be relatively slow for real applications with large data sets. This prompts the following question: Is the full algorithmic power of convex programming and dynamic programming necessary to achieve this level of sample efficiency? Ideally one would like a simple combinatorial algorithm for these estimation tasks that runs in near-linear time. This is an interesting open problem of significant practical interest:

Open Problem 1.6.1. *Is there a sample optimal and linear time algorithm for agnostically learning piecewise polynomial distributions?*

Note that the aforementioned approach leads to *non-proper* learning algorithms. In many settings, e.g., for latent variable models, obtaining proper learning algorithms is important for the underlying application. In particular, we pose the following concrete open problem:

Open Problem 1.6.2. *Is there a* $\mathrm{poly}(k, 1/\epsilon)$ *time algorithm for* properly *learning k-mixtures of simple parametric classes?*

1.7 Learning Multivariate Structured Distributions

The problem of learning an unknown structured density over $\mathbb{R}^d$, $d > 1$, has been studied in statistics and machine learning in many settings. We refer the reader to a relatively recent survey on multi-dimensional density estimation [58] with a focus on sample complexity.

Despite intense research efforts, our understanding of the high-dimensional setting is still quite limited. There are two regimes of interest: (i) the dimension d is small, i.e., a fixed constant independent of the problem size, and (ii) the dimension d is large, i.e., part of the input.

For low-dimensional settings, one may be able to handle learning problems whose sample complexity (and hence, running time) is exponential in d. For some natural distribution families such an exponential dependence is inherent, e.g., for high-dimensional arbitrary log-concave densities. Recent statistical research on the topic attempts to determine tight upper and lower bounds on the minimax rate of convergence [17, 18] in the context of the MLE. From a computer science perspective, the goal for these settings is the same: design an algorithm with information-theoretic optimal sample size and polynomial running time.

For high-dimensional settings, problems that inherently require sample complexity exponentially in d are considered intractable. Interestingly enough, a wide variety of natural and important high-dimensional estimation problems have sample complexity polynomial (or even linear) in the dimension. The bottleneck for such problems is to design computationally efficient algorithms. Circumventing the curse of dimensionality is one of the most challenging research directions in distribution learning.

During the past couple of decades, several natural high-dimensional learning problems have been studied in the theoretical computer science literature. In a few prominent cases, theoretically efficient algorithms have been discovered. Two examples include the development of computationally efficient algorithms for learning mixtures of a constant number of high-dimensional

Gaussian distributions [54], and a constant number of discrete product distributions [36, 34]. We remark that both of these algorithms are based on the method of moments and are in fact proper. These algorithms represent important progress in our theoretical understanding of these challenging and important problems. However, while they run in polynomial time, the exponents in their running time are quite high. Both algorithms [54, 34] run in time $(d/\epsilon)^{f(k)}$, where d is the dimension and k is the number of components in the mixture. Hence, there is still a lot of ground to be covered in our understanding of these questions.

At this point, we would like to highlight a fundamental high-dimensional problem that has received significant attention in statistics, but no non-trivial algorithm is known to date. A *t-piece d-dimensional histogram* is a probability density function p over the domain $[0,1]^d$ of the following sort: the domain $[0,1]^d$ is partitioned into t axis-aligned hyper-rectangles $R_1, \ldots, R_t$, and the distribution p is piecewise constant over each rectangle R_i. It follows from Theorem 1.6.4 that $O(td/\epsilon^2)$ samples information–theoretically suffice to learn such distributions (even agnostically). However, no algorithm with subexponential running time is known. A major goal is to answer the following question:

Open Problem 1.7.1. *Is there a* $\mathrm{poly}(d, t, 1/\epsilon)$ *time algorithm for learning t-piece d-dimensional histograms?*

Another fundamental gap in our understanding concerns mixtures of high-dimensional Gaussians. Recall that there exists a learning algorithm for k-mixtures of d-dimensional Gaussians that runs in time $(d/\epsilon)^{f(k)}$ [54]. The learning algorithm follows from the corresponding parameter learning algorithm. For the parameter learning setting, however, the exponential dependence on k is inherent in the sample complexity of the problem (hence, also in the running time) even for $d = 1$. However, no such information–theoretic barrier exists for the problem of density estimation. This motivates the following problem:

Open Problem 1.7.2. *Is there a* $\mathrm{poly}(d, k, 1/\epsilon)$ *time algorithm for learning a mixture of k d-dimensional Gaussians?*

Analogous questions can be asked for various mixture models and more general latent variable models.

1.8 Conclusions and Future Directions

In this chapter, we gave a biased survey of the distribution learning literature from a computer science perspective, i.e., with an explicit focus on the computational efficiency of our estimators. We presented recent work in theoretical computer science on the design of sample optimal and computationally efficient estimators. Many important questions remain open, and the interplay between algorithms and statistics is crucial to their resolution. We conclude this chapter with two important research directions.

One of the most important challenges in statistical learning is handling data that are corrupted by noise. In most cases, the difficulty is not information–theoretic but rather computational. Many popular algorithms (e.g., MLE) are not tolerant to even a small amount of noise in the data. A fundamental gap in our understanding concerns high-dimensional problems.

Research Direction 1.8.1. *Develop computationally efficient* agnostic *learning algorithms for high-dimensional distribution learning problems.*

A concrete problem is that of learning a binary product distribution with adversarial noise. This problem is, of course, straightforward in the noiseless setting; however, it becomes very challenging in the presence of even a small constant fraction of noisy observations. A more challenging open problem is agnostically learning mixture models, e.g., for two high-dimensional Gaussians or even binary product distributions.

Overall, the body of work on statistical estimation has focused on worst-case instances both in terms of algorithms and lower bounds. A natural goal is to go beyond worst-case analysis and design algorithms that provably perform near optimally on *every* input.

Research Direction 1.8.2. *Develop "instance-by-instance optimal" algorithms for distribution learning problems.*

We believe that progress in this direction will lead to efficient algorithms that perform very well in practice.

Bibliography

[1] F. Balabdaoui, K. Rufibach, and J. A. Wellner. Limit distribution theory for maximum likelihood estimation of a log-concave density. *The Annals of Statistics*, 37(3):pp. 1299–1331, 2009.

[2] F. Balabdaoui and J. A. Wellner. Estimation of a k-monotone density: Limit distribution theory and the spline connection. *The Annals of Statistics*, 35(6):pp. 2536–2564, 2007.

[3] F. Balabdaoui and J. A. Wellner. Estimation of a k-monotone density: characterizations, consistency and minimax lower bounds. *Statistica Neerlandica*, 64(1):45–70, 2010.

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

[5] L. Birgé. On estimating a density using Hellinger distance and some other strange facts. *Probab. Theory Relat. Fields*, 71(2):271–291, 1986.

[6] L. Birgé. Estimating a density under order restrictions: Nonasymptotic minimax risk. *Annals of Statistics*, 15(3):995–1012, 1987.

[7] L. Birgé. On the risk of histograms for estimating decreasing densities. *Annals of Statistics*, 15(3):1013–1022, 1987.

[8] R. Blei, F. Gao, and W. V. Li. Metric entropy of high dimensional distributions. *Proceedings of the American Mathematical Society (AMS)*, 135(12):4009 – 4018, 2007.

[9] H. D. Brunk. On the estimation of parameters restricted by inequalities. *The Annals of Mathematical Statistics*, 29(2):pp. 437–454, 1958.

[10] L. Le Cam. Maximum likelihood: An introduction. *Intl. Stat. Rev*, 58:153–171, 1990.

[11] B. Carl and I. Stephani. *Entropy, compactness and the approximation of operators*, volume 98 of *Cambridge Tracts in Mathematics.* Cambridge University Press, Cambridge, 1990.

[12] K.S. Chan and H. Tong. Testing for multimodality with dependent data. *Biometrika*, 91(1):113–123, 2004.

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

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

[15] 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, 2014.

[16] S.X. Chen and J.S. Liu. Statistical applications of the Poisson-Binomial and Conditional Bernoulli Distributions. *Statistica Sinica*, 7:875–892, 1997.

[17] M. Cule and R. Samworth. Maximum likelihood estimation of a multi-dimensional log-concave density. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 72:545607, 2010.

[18] M. Cule and R. Samworth. Theoretical properties of the log-concave maximum likelihood estimator of a multidimensional density. *Electron. J. Statist.*, 4:254–270, 2010.

[19] S. Dasgupta. Learning mixtures of Gaussians. In *Proceedings of the 40th Annual Symposium on Foundations of Computer Science*, pages 634–644, 1999.

[20] C. Daskalakis, I. Diakonikolas, R. O'Donnell, R.A. Servedio, and L. Tan. Learning Sums of Independent Integer Random Variables. In *FOCS*, pages 217–226, 2013.

[21] C. Daskalakis, I. Diakonikolas, and R. A. Servedio. Learning k-modal distributions via testing. *Theory of Computing*, 10(20):535–570, 2014.

[22] C. Daskalakis, I. Diakonikolas, and R.A. Servedio. Learning k-modal distributions via testing. In *SODA*, pages 1371–1385, 2012.

[23] C. Daskalakis, I. Diakonikolas, and R.A. Servedio. Learning Poisson Binomial Distributions. In *STOC*, pages 709–728, 2012.

[24] C. Daskalakis and G. Kamath. Faster and sample near-optimal algorithms for proper learning mixtures of gaussians. In *Proceedings of The 27th Conference on Learning Theory, COLT 2014*, pages 1183–1213, 2014.

[25] 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.

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

[27] L. Devroye and G. Lugosi. *Combinatorial methods in density estimation.* Springer Series in Statistics, Springer, 2001.

[28] D. L. Donoho and I. M. Johnstone. Minimax estimation via wavelet shrinkage. *Ann. Statist.*, 26(3):879–921, 1998.

[29] D. L. Donoho, I. M. Johnstone, G. Kerkyacharian, and D. Picard. Wavelet shrinkage: asymptopia. *Journal of the Royal Statistical Society, Ser. B*, pages 371–394, 1995.

[30] D. L. Donoho, I. M. Johnstone, G. Kerkyacharian, and D. Picard. Density estimation by wavelet thresholding. *Ann. Statist.*, 24(2):508–539, 1996.

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

[32] L. Dumbgen and K. Rufibach. Maximum likelihood estimation of a log-concave density and its distribution function: Basic properties and uniform consistency. *Bernoulli*, 15(1):40–68, 2009.

[33] D. E. Edmunds and H. Triebel. *Function spaces, entropy numbers, differential operators*, volume 120 of *Cambridge Tracts in Mathematics.* Cambridge University Press, Cambridge, 1996.

[34] J. Feldman, R. O'Donnell, and R. A. Servedio. Learning mixtures of product distributions over discrete domains. *SIAM J. Comput.*, 37(5):1536–1564, 2008.

[35] A.-L. Fougères. Estimation de densités unimodales. *Canadian Journal of Statistics*, 25:375–387, 1997.

[36] Y. Freund and Y. Mansour. Estimating a mixture of two product distributions. In *Proceedings of the Twelfth Annual Conference on Computational Learning Theory*, pages 183–192, 1999.

[37] F. Gao and J. A. Wellner. On the rate of convergence of the maximum likelihood estimator of a k-monotone density. *Science in China Series A: Mathematics*, 52:1525–1538, 2009.

[38] U. Grenander. On the theory of mortality measurement. *Skand. Aktuarietidskr.*, 39:125–153, 1956.

[39] P. Groeneboom. Estimating a monotone density. In *Proc. of the Berkeley Conference in Honor of Jerzy Neyman and Jack Kiefer*, pages 539–555, 1985.

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

[41] D. L. Hanson and G. Pledger. Consistency in concave regression. *The Annals of Statistics*, 4(6):pp. 1038–1050, 1976.

[42] R. Hasminskii and I. Ibragimov. On density estimation in the view of kolmogorov's ideas in approximation theory. *Ann. Statist.*, 18(3):999–1010, 1990.

[43] D. Haussler and M. Opper. Mutual information, metric entropy and cumulative relative entropy risk. *Ann. Statist.*, 25(6):2451–2492, 1997.

[44] A. J. Izenman. Recent developments in nonparametric density estimation. *Journal of the American Statistical Association*, 86(413):205–224, 1991.

[45] H. K. Jankowski and J. A. Wellner. Estimation of a discrete monotone density. *Electronic Journal of Statistics*, 3:1567–1605, 2009.

[46] 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.

[47] G. Kerkyacharian and D. Picard. Density estimation in Besov spaces. *Statistics & Probability Letters*, 13(1):15–24, 1992.

[48] G. Kerkyacharian, D. Picard, and K. Tribouley. Lp adaptive density estimation. *Bernoulli*, 2(3):pp. 229–247, 1996.

[49] R. Koenker and I. Mizera. Quasi-concave density estimation. *Ann. Statist.*, 38(5):2998–3027, 2010.

[50] A. N. Kolmogorov and V. M. Tihomirov. ε-entropy and ε-capacity of sets in function spaces. *Uspehi Mat. Nauk*, 14:3–86, 1959.

[51] B. Lindsay. *Mixture models: theory, geometry and applications.* Institute for Mathematical Statistics, 1995.

[52] G. G. Lorentz. Metric entropy and approximation. *Bull. Amer. Math. Soc.*, 72:903–937, 1966.

[53] Y. Makovoz. On the kolmogorov complexity of functions of finite smoothness. *Journal of Complexity*, 2(2):121 – 130, 1986.

[54] A. Moitra and G. Valiant. Settling the polynomial learnability of mixtures of Gaussians. In *FOCS*, pages 93–102, 2010.

[55] 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.

[56] B.L.S. Prakasa Rao. Estimation of a unimodal density. *Sankhya Ser. A*, 31:23–36, 1969.

[57] R. A. Redner and H. F. Walker. Mixture densities, maximum likelihood and the EM algorithm. *SIAM Review*, 26:195–202, 1984.

[58] D. W. Scott and S. R. Sain. Multidimensional density estimation. volume 24 of *Handbook of Statistics*, pages 229 – 261. 2005.

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

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

[61] C. J. Stone. The use of polynomial splines and their tensor products in multivariate function estimation. *The Annals of Statistics*, 22(1):pp. 118–171, 1994.

[62] C. J. Stone, M. H. Hansen, C. Kooperberg, and Y. K. Truong. Polynomial splines and their tensor products in extended linear modeling: 1994 wald memorial lecture. *Ann. Statist.*, 25(4):1371–1470, 1997.

[63] A. T. Suresh, A. Orlitsky, J. Acharya, and A. Jafarpour. Near-optimal-sample estimators for spherical gaussian mixtures. In *Advances in Neural Information Processing Systems (NIPS)*, pages 1395–1403, 2014.

[64] D.M. Titterington, A.F.M. Smith, and U.E. Makov. *Statistical analysis of finite mixture distributions.* Wiley & Sons, 1985.

[65] A. B. Tsybakov. *Introduction to Nonparametric Estimation.* Springer Publishing Company, Incorporated, 2008.

[66] L. Valiant. A theory of the learnable. *Communications of the ACM*, 27(11):1134–1142, 1984.

[67] L. G. Valiant. A theory of the learnable. In *Proc. 16th Annual ACM Symposium on Theory of Computing (STOC)*, pages 436–445. ACM Press, 1984.

[68] A. W. van der Vaart and J. A. Wellner. *Weak convergence and empirical processes.* Springer Series in Statistics. Springer-Verlag, New York, 1996. With applications to statistics.

[69] G. Walther. Inference and modeling with log-concave distributions. *Statistical Science*, 24(3):319–327, 2009.

[70] E. J. Wegman and I. W. Wright. Splines in statistics. *Journal of the American Statistical Association*, 78(382):pp. 351–365, 1983.

[71] E.J. Wegman. Maximum likelihood estimation of a unimodal density. I. and II. *Ann. Math. Statist.*, 41:457–471, 2169–2174, 1970.

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

[73] Y. Yang and A. Barron. Information-theoretic determination of minimax rates of convergence. *Ann. Statist.*, 27(5):1564–1599, 1999.

[74] Y. G. Yatracos. Rates of convergence of minimum distance estimators and Kolmogorov's entropy. *Annals of Statistics*, 13:768–774, 1985.