Discrete minimax estimation with trees

Luc Devroye* **and Tommy Reddad†**

School of Computer Science
McGill University
3480 University Street
Montréal, Québec, Canada
H3A 2A7
e-mail: lucdevroye@gmail.com; tommy.reddad@gmail.com

Abstract: We propose a simple recursive data-based partitioning scheme which produces piecewise-constant or piecewise-linear density estimates on intervals, and show how this scheme can determine the optimal L_1 minimax rate for some discrete nonparametric classes.

MSC 2010 subject classifications: 60G07.
Keywords and phrases: density estimation, minimax theory, discrete probability distribution, Vapnik-Chervonenkis dimension, monotone density, convex density, histogram.

Contents

1. Introduction

Density estimation or *distribution learning* refers to the problem of estimating the unknown probability density function of a common source of independent sample observations. In any interesting case, we know that the unknown source density may come from a known class. In the *parametric* case, each density in this class can be specified using a bounded number of real parameters, e.g., the class of all Gaussian densities with any mean and any variance. The remaining cases are called *nonparametric*. Examples of nonparametric classes include bounded monotone densities on $[0, 1]$, L-Lipschitz densities for a given constant $L > 0$, and log-concave densities, to name a few. By *minimax estimation*, we mean density estimation in the minimax sense, i.e., we are interested in the existence of a density estimate which minimizes its approximation error, even in the worst case.

There is a long line of work in the statistics literature about density estimation, and a growing interest coming from the theoretical computer science and machine learning communities; for a selection of new and old books on this topic, see [13, 14, 16, 22, 32, 33]. The study of nonparametric density estimation began as early as in the 1950's, when Grenander [20] described and studied properties of the maximum likelihood estimate of an unknown density taken from the

*Supported by NSERC Grant A3456.
†Supported by NSERC PGS D scholarship 396164433.

class of bounded monotone densities on $[0, 1]$. Grenander's estimator and this class received much further treatment over the years, in particular by Prakasa Rao [29], Groeneboom [21], and Birgé [7, 8, 9], who identified the optimal L_1-error minimax rate up to a constant factor, and also gave an efficient adaptive estimator which worked even when the boundedness parameter was unknown. Since then, countless more nonparametric classes have been studied, and many different all-purpose methods have been developed to obtain minimax results about these classes: for the construction of density estimates, see e.g., the maximum likelihood estimate, skeleton estimates, kernel estimates, and wavelet estimates, to name a few; and for minimax rate lower bounds, see e.g., the methods of Assouad, Fano, and Le Cam [13, 14, 16, 35]. See [5, 10, 18, 24] for recent related works in nonparametric shape-constrained regression.

One very popular style of density estimate is the *histogram*, in which the support of the random data is partitioned into bins, where each bin receives a weight proportional to the number of data points contained within, and such that the estimate is constant with the given weight along each bin. Then, the selection of the bins themselves becomes critical in the construction of a good histogram estimate. Birgé [8] showed how histograms with carefully chosen exponentially increasing bin sizes will have L_1-error within a constant factor of the optimal minimax rate for the class of bounded non-increasing densities on $[0, 1]$. In general, the right choice of an underlying partition for a histogram estimate is not obvious.

In this work, we devise a recursive data-based approach for determining the partition of the support for a histogram estimate of discrete non-increasing densities. We also use a similar approach to build a piecewise-linear estimator for discrete non-increasing convex densities—see Anevski [1], Jongbloed [26], and Groeneboom, Jongbloed, and Wellner [23] for works concerning the maximum likelihood and minimax estimation of continuous non-increasing convex densities. Both of our estimators are *minimax-optimal*, i.e., their minimax L_1-error is within a constant factor of the optimal rate. Recursive data-based partitioning schemes have been extremely popular in density estimation since the 1970's with Gessaman [19], Chen and Zhao [11], Lugosi and Nobel [28], and countless others, with great interest coming from the machine learning and pattern recognition communities [15]. Still, it seems that most of the literature involving recursive data-based partitions are not especially concerned with the rate of convergence of density estimates, but rather other properties such as consistency under different recursive schemes. Moreover, most of the density estimation literature is concerned with the estimation of continuous probability distributions. In discrete density estimation, not all of the constructions or methods used to develop arguments for analogous continuous classes will neatly apply, and in some cases, there are discrete phenomena that call for a different approach. See Jankowski and Wellner [25] for a recent treatment on the properties of a variety of estimators of discrete non-increasing densities.

2. Preliminaries and summary

Let $\mathcal{F}$ be a given class of probability densities with respect to a base measure μ on the measurable space $(\mathcal{X}, \Sigma)$, and let $f \in \mathcal{F}$. If X is a random variable taking values in $(\mathcal{X}, \Sigma)$, we write $X \sim f$ to mean that

$$\mathbf{P}\{X \in A\} = \int_A f \, \mathrm{d}\mu, \qquad \text{for each } A \in \Sigma.$$

The notation $X_1, \ldots, X_n \overset{i.i.d.}{\sim} f$ means that $X_i \sim f$ for each $1 \leq i \leq n$, and that $X_1, \ldots, X_n$ are mutually independent.

Typically in density estimation, either $\mathcal{X} = \mathbb{R}^d$, Σ is the Borel σ-algebra, and μ is the Lebesgue measure, or $\mathcal{X}$ is countable, $\Sigma = \mathcal{P}(\mathcal{X})$, and μ is the counting measure. The former case is referred to as the *continuous setting*, and the latter case as the *discrete setting*, where f is more often called a *probability mass function* in the literature. Throughout this paper, we will only be concerned with the discrete setting, and even so, we still refer to $\mathcal{F}$ as a class of densities, and f as a density itself. Plainly, in this case, $X \sim f$ signifies that

$$\mathbf{P}\{X \in A\} = \sum_{x \in A} f(x), \qquad \text{for each } A \in \mathcal{P}(\mathcal{X}).$$

Let $f \in \mathcal{F}$ be unknown. Given the n samples $X_1, \ldots, X_n \overset{i.i.d.}{\sim} f$, our goal is to create a *density estimate*

$$\hat{f}_n \colon \mathcal{X}^n \to \mathbb{R}^{\mathcal{X}},$$

such that the probability measures corresponding to f and $\hat{f}_n(X_1, \ldots, X_n)$ are close in *total variation (TV) distance*, where for any probability measures μ, ν, their TV-distance is defined as

$$\mathrm{TV}(\mu, \nu) = \sup_{A \in \Sigma} |\mu(A) - \nu(A)|. \tag{1}$$

The TV-distance has several equivalent definitions; importantly, if μ and ν are probability measures with corresponding densities f and g, then

$$\mathrm{TV}(\mu, \nu) = \|f - g\|_1 / 2, \tag{2}$$

$$= \inf_{(X,Y) \colon X \sim f, Y \sim g} \mathbf{P}\{X \neq Y\}, \tag{3}$$

where for any function $h \colon \mathcal{X} \to \mathbb{R}$, we define the L_1-norm of h as

$$\|h\|_1 = \sum_{x \in \mathcal{X}} |h(x)|.$$

(In the continuous case, this sum is simply replaced with an integral.) In view of the relation between TV-distance and L_1-norm in (

There are various possible measures of dissimilarity between probability distributions which can be considered in density estimation, e.g., the Hellinger

distance, Wasserstein distance, L_p-distance, χ^2-divergence, Kullback-Leibler divergence, or any number of other divergences; see Sason and Verdú [31] for a survey on many such functions and the relations between them. Here, we focus on the TV-distance due to its several appealing properties, such as being a metric, enjoying the natural probabilistic interpretation of (

If $\hat{f}_n$ is a density estimate, we define the *risk* of the estimator $\hat{f}_n$ with respect to the class $\mathcal{F}$ as

$$\mathcal{R}_n(\hat{f}_n, \mathcal{F}) = \sup_{f \in \mathcal{F}} \mathbf{E}\big\{\mathrm{TV}(\hat{f}_n(X_1, \ldots, X_n), f)\big\},$$

where the expectation is over the n i.i.d. samples from f, and possible randomization of the estimator. From now on we will omit the dependence of $\hat{f}_n$ on $X_1, \ldots, X_n$ unless it is not obvious. The *minimax risk* or *minimax rate* for $\mathcal{F}$ is the smallest risk over all possible density estimates,

$$\mathcal{R}_n(\mathcal{F}) = \inf_{\hat{f}_n} \mathcal{R}_n(\hat{f}_n, \mathcal{F}).$$

We can now state our results precisely. Let $k \in \mathbb{N}$ and let $\mathcal{F}_k$ be the class of non-increasing densities on $\{1, \ldots, k\}$, i.e., set of of all probability vectors $f\colon \{1, \ldots, k\} \to \mathbb{R}$ for which

$$f(x+1) \le f(x), \qquad \text{for all } x \in \{1, \ldots, k-1\}. \tag{4}$$

Theorem 2.1. *Let $f\colon \mathbb{N} \times \mathbb{N} \to \mathbb{R}$ be*

$$f(n,k) = \begin{cases} \sqrt{k/n} & \text{if } 2 \le k < 2n^{1/3}, \\ \left(\frac{\log_2(k/n^{1/3})}{n}\right)^{1/3} & \text{if } 2n^{1/3} \le k < n^{1/3}2^n, \\ 1 & \text{if } n^{1/3}2^n \le k. \end{cases}$$

There is a universal constant $C \ge 1$ such that, for sufficiently large n not depending on k,

$$\frac{1}{C} \le \frac{\mathcal{R}_n(\mathcal{F}_k)}{f(n,k)} \le C.$$

Let $\mathcal{G}_k$ be the class of all non-increasing convex densities on $\{1, \ldots, k\}$, so each $f \in \mathcal{G}_k$ satisfies (

Our upper bounds will crucially rely on the next results, which allow us to relate the minimax rate of a class to an old and well-studied combinatorial quantity called the *Vapnik-Chervonenkis (VC) dimension* [34]: For $\mathcal{A} \subseteq \mathcal{P}(\mathcal{X})$ a family of subsets of $\mathcal{X}$, the VC-dimension of $\mathcal{A}$, denoted by $\mathrm{VC}(\mathcal{A})$, is the size of the largest set $X \subseteq \mathcal{X}$ such that for every $Y \subseteq X$, there exists $B \in \mathcal{A}$ such that $X \cap B = Y$. See, e.g., the book of Devroye and Lugosi [16] for examples and applications of the VC-dimension in the study of density estimation.

Theorem 2.2 (Devroye, Lugosi [16])**.** *Let $\mathcal{F}$ be a class of densities supported on $\mathcal{X}$, and let $\mathcal{F}_\Theta = \{f_{n,\theta}\colon \theta \in \Theta\}$ be a class of density estimates satisfying $\sum_{x \in \mathcal{X}} f_{n,\theta}(x) = 1$ for every $\theta \in \Theta$. Let $\mathcal{A}_\Theta$ be the* Yatracos class *of $\mathcal{F}_\Theta$,*

$$\mathcal{A}_\Theta = \Big\{\{x \in \mathcal{X}\colon f_{n,\theta}(x) > f_{n,\theta'}(x)\}\colon \theta \ne \theta' \in \Theta\Big\}.$$

For $f \in \mathcal{F}$, let μ be the probability measure corresponding to f. Let also μ_n be the empirical measure based on $X_1, \ldots, X_n \overset{i.i.d.}{\sim} f$, where

$$\mu_n(A) = \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}\{X_i \in A\}, \qquad \text{for } A \in \mathcal{A}_\Theta.$$

Then, there is an estimate ψ_n for which

$$\mathrm{TV}(\psi_n, f) \leq 3 \inf_{\theta \in \Theta} \mathrm{TV}(f_{n,\theta}, f) + 2 \sup_{A \in \mathcal{A}_\Theta} |\mu_n(A) - \mu(A)| + \frac{3}{2n}.$$

The estimate ψ_n in Theorem

Theorem 2.3 (Devroye, Lugosi [16])**.** *Let $\mathcal{F}, \mathcal{X}, f, \mu, \mu_n$ be as in Theorem*

Remark 2.4. The quantity $\sup_{A \in \mathcal{A}} |\mu(A) - \nu(A)|$ in Theorem

Corollary 2.5. *Let $\mathcal{F}, \mathcal{A}_\Theta, f_{n,\theta}, f, \mu, \mu_n$ be as in Theorem*

3. Non-increasing densities

This section is devoted to presenting a proof of the upper bound of Theorem

3.1. A greedy tree-based estimator

Suppose that k is a power of two. This assumption can only, at worst, smudge some constant factors in the final minimax rate. Using the samples $X_1, \ldots, X_n \overset{i.i.d.}{\sim} f \in \mathcal{F}_k$, we recursively construct a rooted ordered binary tree $\widehat{T}$ which determines a partition of the interval $\{1, \ldots, k\}$, from which we can build a histogram estimate $\hat{f}_n$ for f. Specifically, let ρ be the root of $\widehat{T}$, where $I_\rho = \{1, \ldots, k\}$. We say that ρ covers the interval I_ρ. Then, for every node u in $\widehat{T}$ covering the interval

$$I_u = \{a_u, a_u + 1, \ldots, a_u + |I_u| - 1\},$$

we first check if $|I_u| = 1$, and if so we make u a leaf in $\widehat{T}$. Otherwise, if

$$\begin{aligned} I_v &= \{a_u, a_u + 1, \ldots, a_u + |I_u|/2 - 1\}, \\ I_w &= I_u \setminus I_v \end{aligned}$$

are the first and second halves of I_u, we verify the condition

$$|N_v - N_w| > \sqrt{N_v + N_w}, \tag{5}$$

where N_v, N_w are the number of samples which fall into the intervals I_v, I_w, i.e.,

$$N_z = \sum_{i=1}^{n} \mathbf{1}\{X_i \in I_z\}, \qquad \text{for } z \in \{v, w\}.$$

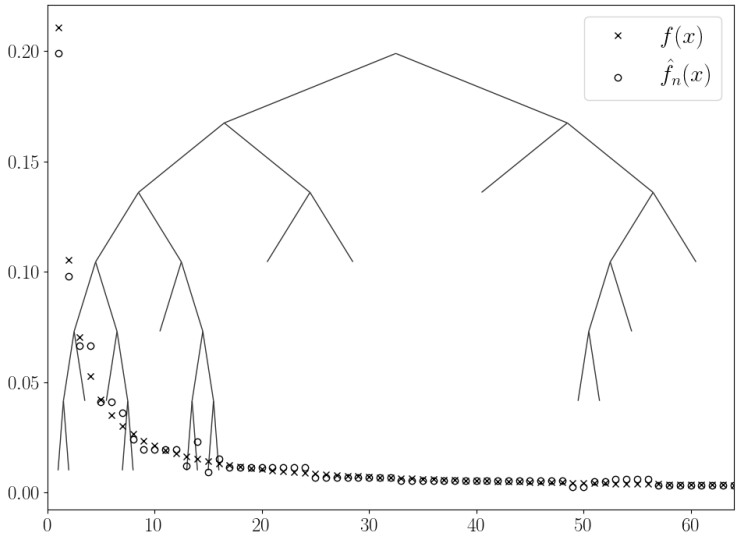

FIG 1. *One sample of $\hat{f}_n$ for $n = 1000$ and $k = 64$, where $f(x) = 1/(xH_k)$ for H_k the k-th Harmonic number. The tree $\widehat{T}$ is overlayed.*

The inequality (

After applying this procedure, one obtains a (random) tree $\widehat{T}$ with leaves $\widehat{L}$, and the set $\{I_u : u \in \widehat{L}\}$ forms a partition of the support $\{1, \ldots, k\}$. Let $\hat{f}_n$ be the histogram estimate based on this partition, i.e.,

$$\hat{f}_n(x) = \frac{N_u}{n|I_u|}, \qquad \text{if } x \in I_u, u \in \widehat{L}.$$

The density estimate $\hat{f}_n$ is called the greedy tree-based estimator*. See Figure for a typical plot of $\hat{f}_n$, and a visualization of the tree $\widehat{T}$.*

Remark 3.1. Intuitively, we justify the rule (

Remark 3.2. One could argue that any good estimate of a non-increasing density should itself be non-increasing, and the estimate $\hat{f}_n$ does not have this property. This can be rectified using a method of Birgé [8], who described a transformation of piecewise-constant density estimates which does not increase risk with respect to non-increasing densities. Specifically, suppose that the estimate $\hat{f}_n$ is not non-increasing. Then, there are consecutive intervals I_v, I_w such that $\hat{f}_n$ has constant value y_v on I_v and y_w on I_w, and $y_v < y_w$. Let the transformed estimate be constant on $I_v \cup I_w$, with value

$$\frac{y_v|I_v| + y_w|I_w|}{|I_v| + |I_w|},$$

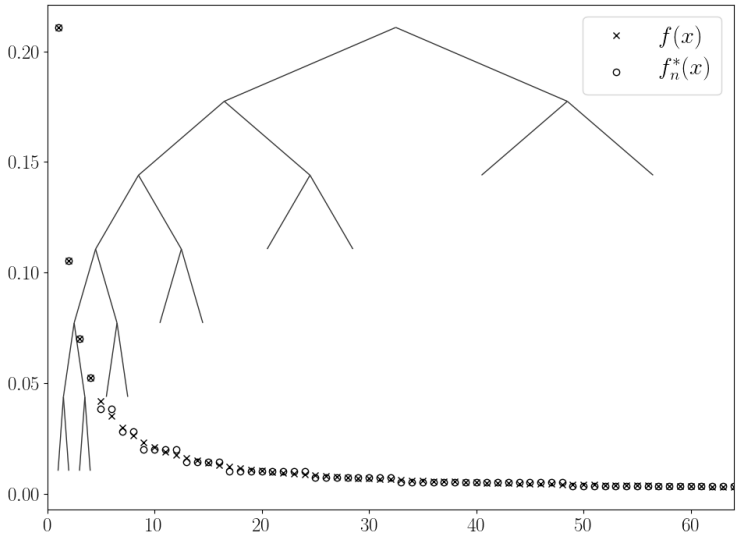

FIG 2. *A plot of f_n^* for $n = 1000$ and $k = 64$, where $f(x) = 1/(xH_k)$. The tree T^* is overlayed.*

i.e., the average value of $\hat{f}_n$ on $I_v \cup I_w$. Iterate the above transformation until a non-increasing estimate is obtained. It can be proven that this results in a unique estimate $\hat{f}'_n$, regardless of the order of merged intervals, and that

$$\mathrm{TV}(\hat{f}'_n, f) \leq \mathrm{TV}(\hat{f}_n, f).$$

3.2. An idealized tree-based estimator

Instead of analyzing the greedy tree-based estimator $\hat{f}_n$ of the preceding section, we fully analyze an idealized version. Indeed, in (

Of course, T^* and f_n^* both depend intimately upon knowledge of the density f; in practice, we only have access to the samples $X_1, \ldots, X_n \overset{i.i.d.}{\sim} f$, and the density f itself is unknown. In particular, we cannot practically use f_n^* as an estimate for unknown f. Importantly, as we will soon show, we can still use f_n^* along with Corollary

Proposition 3.3.

$$\mathrm{TV}(f_n^*, f) \leq \frac{5}{2}\sqrt{\frac{|\{u \in L^* \colon |I_u| > 1\}|}{n}}.$$

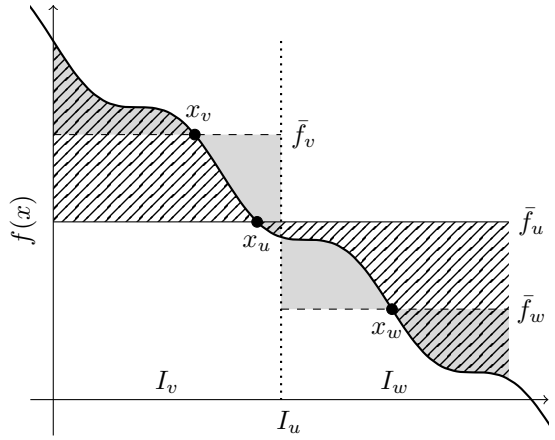

FIG 3. *A visualization of the L_1 distance between f_n^* and f on I_u.*

Proof. Writing out the TV-distance explicitly, we have

$$\begin{aligned}\mathrm{TV}(f_n^*, f) &= \frac{1}{2}\sum_{x\in\{1,\dots,k\}} |f_n^*(x) - f(x)| \\ &= \frac{1}{2}\sum_{u\in L^*}\sum_{x\in I_u} |f_n^*(x) - f(x)| \\ &= \frac{1}{2}\sum_{u\in L^*}\sum_{x\in I_u} |\bar{f}_u - f(x)|.\end{aligned}$$

Let $u \in L^*$, and define $A_u = \sum_{x\in I_u} |\bar{f}_u - f(x)|$. If $|I_u| = 1$, then $A_u = 0$, so assume that $|I_u| > 1$. In this case, let I_v and I_w be the left and right halves of the interval I_u, and let $\bar{f}_v$ and $\bar{f}_w$ be the average value of f on I_v and I_w respectively. Write also

$$B_v = \sum_{x\in I_v} |\bar{f}_v - f(x)|, \qquad B_w = \sum_{x\in I_w} |\bar{f}_w - f(x)|.$$

Refer to Figure . We view A_u as the positive area between the curve f and the line $\bar{f}_u$; in the figure, this is the patterned area. Then, B_v is the positive area between f and $\bar{f}_v$ on I_v, which is represented as the gray area on I_v in Figure , and B_w is the positive area between f and $\bar{f}_w$ on I_w, the gray area on I_w in Figure . For $z \in \{v, u, w\}$, let x_z be the largest point in I_z for which $f(x_z) \geq \bar{f}_z$. By the triangle inequality,

$$\begin{aligned}A_u &\leq (\bar{f}_v - \bar{f}_u)|I_v| + (\bar{f}_u - \bar{f}_w)|I_w| + B_v + B_w \\ &= (f_v - f_w) + B_v + B_w.\end{aligned}$$

Furthermore,

$$\begin{aligned} B_v &= \sum_{x \in I_v,\, x \le x_v} (f(x) - \bar{f}_v) + \sum_{x \in I_v,\, x > x_v} (\bar{f}_v - f(x)) \\ &= 2 \sum_{x \in I_v,\, x > x_v} (\bar{f}_v - f(x)) \\ &\le 2|I_v|(\bar{f}_v - \bar{f}_w) \\ &= 2(f_v - f_w), \end{aligned}$$

where the second equality follows by the choice of x_v. A similar relation holds for B_w, whence

$$A_u \le 5(f_v - f_w) \le 5\sqrt{f_u/n},$$

where this last inequality follows from the splitting rule (

Proposition 3.4. *If $n \ge 64$ and $2n^{1/3} \le k < n^{1/3}2^n$, then*

$$|L^*| \le 12n^{1/3}\Big(\log_2(k/n^{1/3})\Big)^{2/3}.$$

Proof. Note that T^* has height at most $\log_2 k$. Let U_j be the set of nodes at depth $j-1$ in T^* which have at least one leaf as a child, for $1 \le j \le \log_2 k$, and label the children of the nodes in U_j in order of appeareance from right to left in T^* as $u_1, u_2, \ldots, u_{2|U_j|}$. Since none of the nodes in U_j are themselves leaves, then by (

Recall that $\mathcal{G}_k$ is the class of non-increasing convex densities supported on $\{1, \ldots, k\}$. Then, $\mathcal{G}_k$ forms a subclass of $\mathcal{F}_k$, which we considered in Section . This section is devoted to extending the techniques of Section in order to obtain a minimax rate upper bound on $\mathcal{G}_k$. Again, the lower bound is proved using standard techniques in Appendix . In this section, we assume that k is a power of three. In order to prove the upper bound of Theorem

$$\mathrm{TV}(f_n^\dagger, f) \le \frac{41}{48}\sqrt{\frac{|\{u \in L^\dagger : |I_u| > 1\}|}{n}}.$$

Before proving this, we first note that by convexity of f, the slope of the line passing through $(m_w, \bar{f}_w)$ and $(m_r, \bar{f}_r)$ is at least the slope of the line passing through $(m_v, \bar{f}_v)$ and $(m_w, \bar{f}_w)$. Equivalently,

$$f_r - f_w \ge f_w - f_v \iff f_v - 2f_w + f_r \ge 0.$$

$$B_w \le \int_{I_w(f_n^\dagger - g_{vw})} (f_n^\dagger - g_{vw}) + \int_{I_w(f_n^\dagger - g_{wr})} (f_n^\dagger - g_{wr}) = 3(f_v - 2f_w + f_r)/8.$$

Proof of the upper bound in Theorem The case $k \ge n^{1/3}2^n$ is trivial, and follows simply because the TV-distar

It remains to bound B_v and B_r. Let $x_v \in I_v$ be the point where the line passing through $(m_v, \bar{f}_v)$ and $(m_r, \bar{f}_r)$ intersects f. As before, this points exists, and since f is non-increasing, $x_v \leq m_v$. Futhermore,

$$\begin{aligned} B_v &= \int_{I_v \cap (-\infty, x_v]} (f - f_n^\dagger) + \int_{I_v \cap (x_v, \infty)} (f_n^\dagger - f) \\ &= 2 \int_{I_v \cap (x_v, \infty)} (f_n^\dagger - f) \\ &\leq 2 \int_{I_v} (f_n^\dagger - g_{wr}) \\ &= 2(f_v - 2f_w + f_r)/3, \end{aligned}$$

where the inequality follows from convexity and earlier remarks. A similar argument follows for B_r.

In total,

$$A_u = B_v + B_w + B_r \leq 41(f_v - 2f_w + f_r)/24.$$

The result then follows from the splitting rule (

Proposition 4.2. *If $n \geq 3^{10}$ and $3n^{1/5} \leq k < n^{1/5}3^n$, then*

$$|L^\dagger| \leq 34n^{1/5}\Big(\log_3(k/n^{1/5})\Big)^{4/5}.$$

Proof. The tree $T^\dagger$ has height at most $\log_3 k$. Let U_j be the set of nodes at depth $j-1$ in $T^\dagger$ with at least one leaf as a child, for $1 \leq j \leq \log_3 k$, labelled in order of appearance from right to left in $T^\dagger$ as $u_1, u_2, \ldots, u_{3|U_j|}$. By the convex splitting rule (

It seems likely, given our results on the idealized tree-based estimators from Section and Section , that the greedy tree-based estimators also behave well. In particular, we suspect that our greedy tree-based estimators are minimax-optimal within logarithmic factors. We leave this open to future work. It is also often desirable for nonparametric estimators to be adaptive, in the sense that they attain the optimal minimax rate without depending on some of the important features of the nonparametric class in question. In some cases, an adaptive density estimate can be constructed by first estimating these features, and then building a density estimate assuming the estimated features. For example, in [8], an adapative estimate for non-increasing densities is developed by first estimating the size of the support, and plugging this estimated support size into a non-adaptive estimate. We expect that in this manner, our method can be made adaptive. The techniques of this paper seem to naturally extend to higher dimensions. Take, for instance, the class of block-decreasing densities, whose minimax rate was identified by Biau and Devroye [6]. This is the class of densities supported on $[0,1]^d$ bounded by some constant $B \geq 0$, such that each density is non-increasing in each coordinate if all other coordinates are held fixed. The

discrete version of this class has each density supported on $\{1,\dots,k\}^d$, with the monotonicity constraint. In order to estimate such a density, one could devise an oriented binary splitting rule analogous to (Furthermore, we expect that there are many other classes of one-dimensional densities whose optimal minimax rate could be identified using our approach, like the class of ℓ-monotone densities on $\{1,\dots,k\}$, where a function f is called *ℓ-monotone* if it is non-negative and if $(-1)^j f^{(j)}$ is non-increasing and convex for all $j \in \{0,\dots,\ell-2\}$ if $\ell \geq 2$, and where f is non-negative and non-increasing if $\ell = 1$. This paper tackles the cases of $\ell = 1$ and $\ell = 2$. Write $\mathcal{F}_{k,\ell}$ for the class of ℓ-monotone densities on $\{1,\dots,k\}$. See Balabdaoui and Wellner [3, 4] for texts concerning the density estimation of ℓ-monotone densities. It seems likely that our method could be applied to prove the following conjecture. *Let $f\colon \mathbb{N}\times\mathbb{N}\times\mathbb{N}\times\mathbb{R}\to\mathbb{R}$ be*

$$f(n,k,\ell,\alpha) = \begin{cases} \sqrt{k/n} & \text{if } 2 \leq k \leq \alpha n^{\frac{1}{2\ell+1}}, \\ \left(\frac{\log_\alpha(k/n^{\frac{1}{2\ell+1}})}{n}\right)^{\frac{\ell}{2\ell+1}} & \text{if } \alpha n^{\frac{1}{2\ell+1}} \leq k \leq n^{\frac{1}{2\ell+1}}\alpha^n, \\ 1 & \text{if } n^{\frac{1}{2\ell+1}}\alpha^n \leq k. \end{cases}$$

Let $\ell \geq 1$ be fixed. There are constants $\alpha, C, n_0 \geq 1$ depending only on ℓ such that, for $n \geq n_0$,

$$\frac{1}{C} \leq \frac{\mathcal{R}_n(\mathcal{F}_{k,\ell})}{f(n,k,\ell,\alpha)} \leq C.$$

The main obstacle in proving the above would be the development of good local estimates for ℓ-monotone densities, in the same flavor as Proposition

Our approach also likely can be applied to the class of all log-concave discrete distributions, where we recall that $f : \mathbb{N} \to [0,1]$ is called *log-concave* if

$$f(x)f(x+2) \leq f(x+1)^2, \quad \text{for all } x \geq 1.$$

See [17, 27, 30] for a small selection of works on the density estimation of d-dimensional log-concave continuous densities. The optimal Hellinger distance minimax rate (within logarithmic factors) for this class was recently obtained by Dagan and Kur [12], who showed that it is attained by the maximum-likelihood estimate. There remains a small gap between the best known upper and lower bounds in the TV-distance minimax rate as of the time of writing.

Acknowledgments

We would like to thank the three reviewers and an associate editor for their helpful comments and suggestions.

Appendix A: Lower bounds

Lemma A.1 (Assouad's Lemma [2, 16])**.** *Let $\mathcal{F}$ be a class of densities supported on the set $\mathcal{X}$. Let $A_0, A_1, \dots, A_r$ be a partition of $\mathcal{X}$, and $g_{ij}\colon A_i \to \mathbb{R}$ for $0 \leq i \leq$*

r and $j \in \{0,1\}$ be some collection of functions. For $\theta = (\theta_1, \ldots, \theta_r) \in \{0,1\}^r$, define the function $f_\theta \colon \mathcal{X} \to \mathbb{R}$ by

$$f_\theta(x) = \begin{cases} g_{00}(x) & \text{if } x \in A_0, \\ g_{i\theta_i}(x) & \text{if } x \in A_i, \end{cases}$$

such that each f_θ is a density on $\mathcal{X}$. Let $\zeta_i \in \{0,1\}^n$ agree with θ on all bits except for the i-th bit. Then, suppose that

$$0 < \beta \le \inf_\theta \inf_{1 \le i \le r} \int \sqrt{f_\theta f_{\zeta_i}},$$

and

$$0 < \alpha \le \inf_\theta \inf_{1 \le i \le r} \int_{A_i} |f_\theta - f_{\zeta_i}|.$$

Let $\mathcal{H}$ be the hypercube of densities

$$\mathcal{H} = \{f_\theta \colon \theta \in \{0,1\}^r\}.$$

If $\mathcal{H} \subseteq \mathcal{F}$, then

$$\mathcal{R}_n(\mathcal{F}) \ge \frac{r\alpha}{4}\Big(1 - \sqrt{2n(1-\beta)}\Big).$$

A.1. Proof of the lower bound in Theorem

Suppose first that $e^8 n^{1/3} \le k \le n^{1/3} e^n$. Let $A_1, \ldots, A_r$ be consecutive intervals of even cardinality, starting from the leftmost atom 1. Split each A_i in two equal parts, A_i' and A_i''. Let $\varepsilon \in (0, 1/\sqrt{2})$, and set

$$g_{i0}(x) = \begin{cases} \frac{1+\varepsilon}{r|A_i|} & \text{if } x \in A_i', \\ \frac{1-\varepsilon}{r|A_i|} & \text{if } x \in A_i'', \end{cases}$$

$$g_{i1}(x) = \frac{1}{r|A_i|}.$$

It is clear that each f_θ is a density. In order for each f_θ to be monotone, we require that

$$\frac{1-\varepsilon}{|A_i|} \ge \frac{1+\varepsilon}{|A_{i+1}|},$$

and in particular

$$|A_i| \ge |A_1| \left(\frac{1+\varepsilon}{1-\varepsilon}\right)^{i-1}.$$

Pick $|A_1| = 2$. Since $\log(1+\varepsilon) - \log(1-\varepsilon) \le 4\varepsilon$ for $\varepsilon \in (0, 1/\sqrt{2})$, it suffices to take

$$|A_i| \ge a_i = 2e^{4\varepsilon(i-1)}.$$

Let $|A_i|$ be the smallest even integer at least equal to a_i, so that $a_i \le |A_i| \le a_i+2$, and thus

$$\sum_{i=1}^{r} |A_i| \le 2r + \frac{2e^{4\varepsilon r}}{e^{4\varepsilon}-1} \le 2r + \frac{e^{4\varepsilon r}}{2\varepsilon}.$$

Since the support of our densities is $\{1,\dots,k\}$, then we ask that this last upper bound not exceed k. We can guarantee this in particular with a choice of r and ε for which

$$2r \le \frac{k}{2}, \quad \text{and} \quad \frac{e^{4\varepsilon r}}{2\varepsilon} \le \frac{k}{2}. \tag{6}$$

Fix $1 \le i \le r$. Then,

$$\begin{aligned}\int \Big(\sqrt{f_\theta} - \sqrt{f_{\zeta_i}}\Big)^2 &= \sum_{x\in A_i} \Big(\sqrt{f_\theta(x)} - \sqrt{f_{\zeta_i}(x)}\Big)^2 \\ &= \frac{2}{r} - \frac{1}{r}\big(\sqrt{1+\varepsilon} + \sqrt{1-\varepsilon}\big) \\ &\le \frac{\varepsilon^2}{r},\end{aligned}$$

so

$$\begin{aligned}\int \sqrt{f_\theta f_{\zeta_i}} &= 1 - \frac{1}{2}\int \Big(\sqrt{f_\theta} - \sqrt{f_{\zeta_i}}\Big)^2 \\ &\ge 1 - \frac{\varepsilon^2}{2r}.\end{aligned}$$

On the other hand,

$$\int_{A_i} |f_\theta - f_{\zeta_i}| = \sum_{x\in A_i} |f_\theta(x) - f_{\zeta_i}(x)| = \frac{\varepsilon}{r}.$$

Now pick

$$\varepsilon = \frac{1}{4}\left(\frac{\log(k/n^{1/3})}{n}\right)^{1/3},$$

and r for which

$$\sqrt{\frac{n\varepsilon^2}{r}} \le \frac{1}{2},$$

or equivalently,

$$r \ge \frac{1}{4}\Big(n\log^2(k/n^{1/3})\Big)^{1/3}.$$

Note that $k \le n^{1/3}e^n$ now implies that $\varepsilon \in (0, 1/\sqrt{2})$. With this choice, Lemma

When $k \ge n^{1/3}e^n$, we argue by inclusion that

$$\mathcal{R}_n(\mathcal{F}_k) \ge \inf_{k\ge n^{1/3}e^n} \mathcal{R}_n(\mathcal{F}_k) \ge \frac{1}{32}.$$

The only remaining case is $k \leq e^8 n^{1/3}$. In this case, we offer a different construction. Now, each A_i will have size 2 for $1 \leq i \leq r$, where $r = \lfloor k/2 \rfloor$. Fix $a, b \in \mathbb{R}$ to be specified later, and set

$$g_{i0}(x) = \begin{cases} a - b(2i-1) & \text{if } x \in A_i', \\ a - b(2i+1) & \text{if } x \in A_i'', \end{cases}$$
$$g_{i1}(x) = a - 2bi,$$

We insist that

$$a - b(2r+1) = \frac{1-\varepsilon}{2r}$$

for some $0 \leq \varepsilon \leq 1$. Since each f_θ must be a density, we need that

$$\sum_{i=1}^{r} 2(a - 2bi) = 1.$$

Both of these conditions will be satisfied if we pick

$$b = \frac{\varepsilon}{2r^2}, \quad \text{and} \quad a = b + \frac{1+\varepsilon}{2r},$$

Furthermore, the largest probability of an atom here is

$$a - b = \frac{1+\varepsilon}{2r} \leq 1,$$

for $k \geq 2$. Then, for $1 \leq i \leq r$, we can compute

$$\begin{aligned} \int \left(\sqrt{f_\theta} - \sqrt{f_{\zeta_i}}\right)^2 &\leq \frac{2b^2}{a - 2bi} \\ &\leq \frac{\varepsilon^2}{r^3(1-\varepsilon)}, \end{aligned}$$

so

$$\int \sqrt{f_\theta f_{\zeta_i}} \geq 1 - \frac{\varepsilon^2}{2r^3(1-\varepsilon)}.$$

and

$$\int_{A_i} |f_\theta - f_{\zeta_i}| = 2b = \frac{\varepsilon}{r^2}.$$

Pick $\varepsilon = e^{-12} r\sqrt{k/n}$. Then, since $2 \leq k \leq e^8 n^{1/3}$ and $r = \lfloor k/2 \rfloor \geq k/3$, then $\varepsilon \leq 1/2$, and

$$\sqrt{\frac{n\varepsilon^2}{r^3(1-\varepsilon)}} \leq \frac{1}{2},$$

so that

$$\mathcal{R}_n(\mathcal{F}_k) \geq \frac{\varepsilon}{4r}\left(1 - \sqrt{\frac{n\varepsilon^2}{r^3(1-\varepsilon)}}\right) \geq \frac{1}{8e^{12}}\sqrt{k/n}. \qquad \square$$

A.2. Proof of the lower bound in Theorem

Let $A_1, \ldots, A_r$ be the partition in Lemma

When $k \geq n^{1/5}e^n$, we argue by inclusion that

$$\mathcal{R}_n(\mathcal{G}_k) \geq \inf_{k \geq n^{1/5}e^n} \mathcal{R}_n(\mathcal{G}_k) \geq \frac{1}{1152}.$$

It remains to prove the case $k \leq e^{40}n^{1/5}$. Observe that $\mathcal{G}_2 = \mathcal{F}_2$, so the lower bound for $k = 2$ follows from Appendix , so we assume that $k \geq 3$. Now, each A_i will have size 3 for $1 \leq i \leq r$, where $r = \lfloor k/3 \rfloor$. Fix $a, b \in \mathbb{R}$ to be specified later, and set

$$\begin{aligned} g_{i0}(j_i) &= \beta_i \\ g_{i0}(j_i + 1) &= \frac{2\beta_i + \beta_{i+1}}{3} - \Delta_i \\ g_{i0}(j_i + 2) &= \frac{\beta_i + 2\beta_{i+1}}{3} - \frac{\Delta_i}{2} \end{aligned}$$

and

$$\begin{aligned} g_{i1}(j_i) &= \beta_i \\ g_{i1}(j_i + 1) &= \frac{2\beta_i + \beta_{i+1}}{3} - \frac{\Delta_i}{2} \\ g_{i1}(j_i + 2) &= \frac{\beta_i + 2\beta_{i+1}}{3} - \Delta_i. \end{aligned}$$

Each f_θ will be non-increasing as long as $\beta_i \geq \beta_{i+1}$, and

$$\Delta_i \leq \frac{\beta_i - \beta_{i+1}}{3},$$

for each $1 \leq i \leq r$. Convexity will follow if

$$\beta_{i+1} - \left(\frac{\beta_i + 2\beta_{i+1}}{3} - \Delta_i\right) \leq \left(\frac{2\beta_{i+1} + \beta_{i+2}}{3} - \Delta_{i+1}\right) - \beta_{i+1},$$

or equivalently,

$$\frac{\beta_i - \beta_{i+1}}{3} - \Delta_i \geq \frac{\beta_{i+1} - \beta_{i+2}}{3} + \Delta_{i+1}.$$

We need also that $\beta_1 \leq 1$, $\beta_{r+1} \geq 0$, and

$$\sum_{i=1}^{r} \left(2\beta_i + \beta_{i+1} - \frac{3\Delta_i}{2}\right) = 1.$$

Take $\beta_{i+1} = \beta_i - 3\Delta_i - \alpha(r - i)$ for $\alpha \geq 0$ to be specified.. Monotonicity follows, and convexity will follow if

$$\begin{aligned} &\frac{\alpha(r-i)}{3} \geq 2\Delta_{i+1} + \frac{\alpha(r-i-1)}{3} \\ \iff &\Delta_{i+1} \leq \frac{\alpha}{6}. \end{aligned}$$

So take each $\Delta_i = \alpha/6$. Then,

$$\beta_i = \beta_1 - \frac{\alpha(i-1)}{2} - \alpha\sum_{j=1}^{i-1}(r-i),$$

and in particular,

$$\beta_{r+1} = \beta_1 - \frac{\alpha r}{2} - \frac{\alpha(r-1)r}{2} = \beta_1 - \frac{\alpha r^2}{2}.$$

Take $\alpha = \varepsilon/r^3$ for some $0 \le \varepsilon \le 1$, whence

$$\beta_{r+1} = \beta_1 - \frac{\varepsilon}{2r}.$$

By monotonicity,

$$1 \le \sum_{i=1}^{r} 3\beta_i \le 3r\beta_1,$$

and

$$1 \ge \sum_{i=1}^{r} 3\beta_{i+1} \ge 3r\Big(\beta_1 - \frac{\varepsilon}{2r}\Big) \ge 3r\beta_1 - \frac{3\varepsilon}{2},$$

so that the right choice of β_1 satisfies

$$\frac{1}{3r} \le \beta_1 \le \frac{5}{6r}.$$

Fix $1 \le i \le r$. Then,

$$\int_{A_i} |f_\theta - f_{\zeta_i}| = \Delta_i = \frac{\varepsilon}{6r^3},$$

and if $\varepsilon \le 1/2$,

$$\int_{A_i} \Big(\sqrt{f_\theta} - \sqrt{f_{\zeta_i}}\Big)^2 \le \frac{\Delta_i^2}{8\beta_{i+1}} \le \frac{\varepsilon^2}{24r^5}.$$

Finally, pick $\varepsilon = e^{-100}r^2\sqrt{k/n}$. Since $k \le e^{40}n^{1/5}$ and $r = \lfloor k/3 \rfloor \ge k/6$, then $\varepsilon \le 1/2$, and

$$\sqrt{\frac{n\varepsilon^2}{24r^5}} \le \frac{1}{2},$$

so by Lemma

References

[1] Anevski, D. (2003). Estimating the derivative of a convex density. *Statist. Neerlandica* **57** 245–257. MR2028914

[2] Assouad, P. (1983). Deux remarques sur l'estimation. *C. R. Acad. Sci. Paris Sér. I Math.* **296** 1021–1024. MR777600

[3] Balabdaoui, F. and Wellner, J. A. (2007). Estimation of a k-monotone density: limit distribution theory and the spline connection. *Ann. Statist.* **35** 2536–2564. MR2382657

[4] Balabdaoui, F. and Wellner, J. A. (2010). Estimation of a k-monotone density: characterizations, consistency and minimax lower bounds. *Stat. Neerl.* **64** 45–70. MR2830965

[5] Bellec, P. C. (2018). Sharp oracle inequalities for least squares estimators in shape restricted regression. *Ann. Statist.* **46** 745–780. MR3782383

[6] Biau, G. and Devroye, L. (2003). On the risk of estimates for block decreasing densities. *J. Multivariate Anal.* **86** 143–165. MR1994726

[7] Birgé, L. (1987). Estimating a density under order restrictions: nonasymptotic minimax risk. *Ann. Statist.* **15** 995–1012. MR902241

[8] Birgé, L. (1987). On the risk of histograms for estimating decreasing densities. *Ann. Statist.* **15** 1013–1022. MR902242

[9] Birgé, L. (1989). The Grenander estimator: a nonasymptotic approach. *Ann. Statist.* **17** 1532–1549. MR1026298

[10] Chatterjee, S., Guntuboyina, A. and Sen, B. (2015). On risk bounds in isotonic and other shape restricted regression problems. *Ann. Statist.* **43** 1774–1800. MR3357878

[11] Chen, X. R. and Zhao, L. C. (1987). Almost sure L_1-norm convergence for data-based histogram density estimates. *J. Multivariate Anal.* **21** 179–188. MR877850

[12] Dagan, Y. and Kur, G. (2019). The log-concave maximum likelihood estimator is optimal in high dimensions. *arXiv e-prints* **abs/1903.05315**.

[13] Devroye, L. (1987). *A Course in Density Estimation. Progress in Probability and Statistics* **14**. Birkhäuser Boston, Inc., Boston, MA. MR891874

[14] Devroye, L. and Györfi, L. (1985). *Nonparametric Density Estimation: The L_1 View. Wiley Series in Probability and Mathematical Statistics: Tracts on Probability and Statistics*. John Wiley & Sons, Inc., New York. MR780746

[15] Devroye, L., Györfi, L. and Lugosi, G. (1996). *A Probabilistic Theory of Pattern Recognition. Applications of Mathematics (New York)* **31**. Springer-Verlag, New York. MR1383093

[16] Devroye, L. and Lugosi, G. (2001). *Combinatorial Methods in Density Estimation. Springer Series in Statistics*. Springer-Verlag, New York. MR1843146

[17] Diakonikolas, I., Kane, D. M. and Stewart, A. (2017). Learning multivariate log-concave distributions. In *Proceedings of Machine Learning Research. COLT '17* **65** 1–17.

[18] Gao, C., Han, F. and Zhang, C.-H. (2017). Minimax risk bounds for piecewise constant models. *Ann. Statist.*

[19] Gessaman, M. P. (1970). A consistent nonparametric multivariate density estimator based on statistically equivalent blocks. *Ann. Math. Statist.* **41** 1344–1346.

[20] Grenander, U. (1956). On the theory of mortality measurement. I, II.

Skand. Aktuarietidskr. **39** 70–96, 125–153.

[21] Groeneboom, P. (1985). Estimating a monotone density. In *Proceedings of the Berkeley conference in honor of Jerzy Neyman and Jack Kiefer, Vol. II (Berkeley, Calif., 1983). Wadsworth Statist./Probab. Ser.* 539–555. Wadsworth, Belmont, CA. MR822052

[22] Groeneboom, P. and Jongbloed, G. (2014). *Nonparametric Estimation Under Shape Constraints: Estimators, Algorithms, and Asymptotics. Cambridge Series in Statistical and Probabilistic Mathematics* **38**. Cambridge University Press, New York. MR3445293

[23] Groeneboom, P., Jongbloed, G. and Wellner, J. A. (2001). Estimation of a convex function: characterizations and asymptotic theory. *Ann. Statist.* **29** 1653–1698. MR1891742

[24] Guntuboyina, A. and Sen, B. (2015). Global risk bounds and adaptation in univariate convex regression. *Probab. Theory Related Fields* **163** 379–411. MR3405621

[25] Jankowski, H. K. and Wellner, J. A. (2009). Estimation of a discrete monotone distribution. *Electron. J. Stat.* **3** 1567–1605. MR2578839

[26] Jongbloed, G. (1995). Three Statistical Inverse Problems, PhD thesis, Delft University of Technology.

[27] Kim, A. K. H. and Samworth, R. J. (2016). Global rates of convergence in log-concave density estimation. *Ann. Statist.* **44** 2756–2779. MR3576560

[28] Lugosi, G. and Nobel, A. (1996). Consistency of data-driven histogram methods for density estimation and classification. *Ann. Statist.* **24** 687–706. MR1394983

[29] Prakasa Rao, B. L. S. (1969). Estimation of a unimodal density. *Sankhyā Ser. A* **31** 23–36. MR0267677

[30] Samworth, R. J. (2017). Recent progress in log-concave density estimation. *arXiv e-prints* **abs/1709.03154**.

[31] Sason, I. and Verdú, S. (2016). f-divergence inequalities. *IEEE Trans. Inform. Theory* **62** 5973–6006. MR3565096

[32] Scott, D. W. (2015). *Multivariate Density Estimation: Theory, Practice, and Visualization*, second ed. *Wiley Series in Probability and Statistics*. John Wiley & Sons, Inc., Hoboken, NJ. MR3329609

[33] Silverman, B. W. (1986). *Density Estimation for Statistics and Data Analysis. Monographs on Statistics and Applied Probability*. Chapman & Hall, London. MR848134

[34] Vapnik, V. N. and Červonenkis, A. J. (1971). The uniform convergence of frequencies of the appearance of events to their probabilities. *Teor. Verojatnost. i Primenen.* **16** 264–279. MR0288823

[35] Yu, B. (1997). Assouad, Fano, and Le Cam. In *Festschrift for Lucien Le Cam* 423–435. Springer, New York. MR1462963