

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	1
2	Preliminaries and summary	3
3	Non-increasing densities	6
3.1	A greedy tree-based estimator	6
3.2	An idealized tree-based estimator	8
4	Non-increasing convex densities	12
5	Discussion	15
A	Lower bounds	16
A.1	Proof of the lower bound in Theorem 2.1.	16
A.2	Proof of the lower bound in Theorem 2.2.	19
	References	24

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

^{*}Supported by NSERC Grant A3456.

[†]Supported by NSERC PGS D scholarship 396164433.

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 [10, 11, 13, 18, 26, 27]. The study of nonparametric density estimation began as early as in the 1950's, when Grenander [16] described and studied properties of the maximum likelihood estimate of an unknown density taken from the class of bounded monotone densities on $[0, 1]$. This estimator and this class received much further treatment over the years, in particular by Prakasa Rao [23], Groeneboom [17], and Birgé [6, 7, 8], who identified the optimal L_1 -error minimax rate, up to a constant factor. 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 [10, 11, 13, 29].

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 [7] 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 [20], and Groeneboom, Jongbloed, and Wellner [19] 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 [15], Chen and Zhao [9], Lugosi and Nobel [22], and countless others, with great interest coming from the machine learning and pattern recognition communities [12]. 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.

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}, Σ) , and let $f \in \mathcal{F}$. If X is a random variable taking values in (\mathcal{X}, Σ) , we write $X \sim f$ to mean that

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

The notation $X_1, \dots, X_n \stackrel{i.i.d.}{\sim} f$ means that $X_i \sim f$ for each $1 \leq i \leq n$, and that X_1, \dots, 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

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

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

$$\hat{f}_n: \mathcal{X}^n \rightarrow \mathbb{R}^{\mathcal{X}},$$

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

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

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

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

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

where for any function $h: \mathcal{X} \rightarrow \mathbb{R}$, 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 (2), we will abuse notation and write

$$\text{TV}(f, g) = \|f - g\|_1/2.$$

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 [25] 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 (1), and having the coupling characterization (3).

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

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

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, \dots, 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 $\{1, \dots, k\}$, i.e., set of all probability vectors $f: \{1, \dots, k\} \rightarrow \mathbb{R}$ for which

$$f(x+1) \leq f(x), \quad \text{for all } x \in \{1, \dots, k-1\}. \quad (4)$$

Theorem 2.1. *For sufficiently large n ,*

$$\mathcal{R}_n(\mathcal{F}_k) \asymp \begin{cases} \sqrt{k/n} & \text{if } 2 \leq k < 2n^{1/3}, \\ \left(\frac{\log_2(k/n^{1/3})}{n}\right)^{1/3} & \text{if } 2n^{1/3} \leq k < n^{1/3}2^n, \\ 1 & \text{if } n^{1/3}2^n \leq k. \end{cases}$$

Let \mathcal{G}_k be the class of all non-increasing convex densities on $\{1, \dots, k\}$, so each $f \in \mathcal{G}_k$ satisfies (4) and

$$f(x) - 2f(x+1) + f(x+2) \geq 0, \quad \text{for all } x \in \{1, \dots, k-2\}.$$

Theorem 2.2. *For sufficiently large n ,*

$$\mathcal{R}_n(\mathcal{G}_k) \asymp \begin{cases} \sqrt{k/n} & \text{if } 2 \leq k < 3n^{1/5}, \\ \left(\frac{\log_3(k/n^{1/5})}{n}\right)^{2/5} & \text{if } 3n^{1/5} \leq k < n^{1/5}3^n, \\ 1 & \text{if } n^{1/5}3^n \leq k. \end{cases}$$

¹For sequences $(a_n: n \in \mathbb{N})$, $(b_n: n \in \mathbb{N})$, we write $a_n \asymp b_n$ if there is a constant $c \geq 1$ for which $(1/c)b_n \leq a_n \leq cb_n$ for all $n \in \mathbb{N}$.

We emphasize here that the above results give upper and lower bounds on the minimax rates $\mathcal{R}_n(\mathcal{F}_k)$ and $\mathcal{R}_n(\mathcal{G}_k)$ which are within universal constant factors of one another, for the entire range of k and n .

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* [28]: For $\mathcal{A} \subseteq \mathcal{P}(\mathcal{X})$ a family of subsets of \mathcal{X} , the VC-dimension of \mathcal{A} , denoted by $\text{VC}(\mathcal{A})$, 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 [13] for examples and applications of the VC-dimension in the study of density estimation.

Theorem 2.3 (Devroye, Lugosi [13]). *Let \mathcal{F} be a class of densities supported on \mathcal{X} , and let $\mathcal{F}_\Theta = \{f_{n,\theta} : \theta \in \Theta\}$ 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}_Θ be the Yatracos class of \mathcal{F}_Θ ,*

$$\mathcal{A}_\Theta = \left\{ \{x \in \mathcal{X} : f_{n,\theta}(x) > f_{n,\theta'}(x)\} : \theta \neq \theta' \in \Theta \right\}.$$

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

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

Then, there is an estimate ψ_n for which

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

The estimate ψ_n in Theorem 2.3 is called the *minimum distance estimate* in [13]—we omit the details of its construction, though we emphasize that if computing $\int_A f_{n,\theta}$ takes one unit of computation for any θ and A , then selecting ψ_n takes time polynomial in the size of \mathcal{A} , which is often exponential in the quantities of interest.

Theorem 2.4 (Devroye, Lugosi [13]). *Let $\mathcal{F}, \mathcal{X}, f, \mu, \mu_n$ be as in Theorem 2.3, and let $\mathcal{A} \subseteq \mathcal{P}(\mathcal{X})$. Then, there is a universal constant $c > 0$ for which*

$$\sup_{f \in \mathcal{F}} \mathbf{E} \left\{ \sup_{A \in \mathcal{A}} |\mu_n(A) - \mu(A)| \right\} \leq c \sqrt{\frac{\text{VC}(\mathcal{A})}{n}}.$$

Remark 2.5. The quantity $\sup_{A \in \mathcal{A}} |\mu(A) - \nu(A)|$ in Theorem 2.4 is precisely equal to $\text{TV}(\mu, \nu)$ if \mathcal{A} is the Borel σ -algebra on \mathcal{X} .

Corollary 2.6. *Let $\mathcal{F}, \mathcal{A}_\Theta, f_{n,\theta}, f, \mu, \mu_n$ be as in Theorem 2.3. Then, there is a universal constant $c > 0$ for which*

$$\mathcal{R}_n(\mathcal{F}) \leq 3 \sup_{f \in \mathcal{F}} \inf_{\theta \in \Theta} \text{TV}(f_{n,\theta}, f) + c \sqrt{\frac{\text{VC}(\mathcal{A}_\Theta)}{n}} + \frac{3}{2n}.$$

3. Non-increasing densities

This section is devoted to presenting a proof of the upper bound of Theorem 2.1. The lower bound is proved in Appendix A.1 using a careful but standard application of Assouad's Lemma [2]. Part of our analysis in proving Theorem 2.1 will involve the development of an explicit efficient estimator for a density in \mathcal{F}_k .

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 $X_1, \dots, X_n \stackrel{i.i.d.}{\sim} f \in \mathcal{F}_k$, we will recursively construct a rooted ordered binary tree \hat{T} which determines a partition of the interval $\{1, \dots, k\}$, from which we can build a histogram estimate \hat{f}_n for f . Specifically, let ρ be the root of \hat{T} , where $I_\rho = \{1, \dots, k\}$. We say that ρ covers the interval I_ρ . Then, for every node u in \hat{T} covering the interval

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

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

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

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

$$|N_v - N_w| > \sqrt{N_v + N_w}, \quad (5)$$

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

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

The inequality (5) is referred to as the *greedy splitting rule*. If (5) is satisfied, then create nodes v, w covering I_v and I_w respectively, and add them to \hat{T} as left and right children of u . If not, make u a leaf in \hat{T} .

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

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

The density estimate \hat{f}_n will be called the *greedy tree-based estimator*. See Figure 1 for a typical plot of \hat{f}_n , and a visualization of the tree \hat{T} .

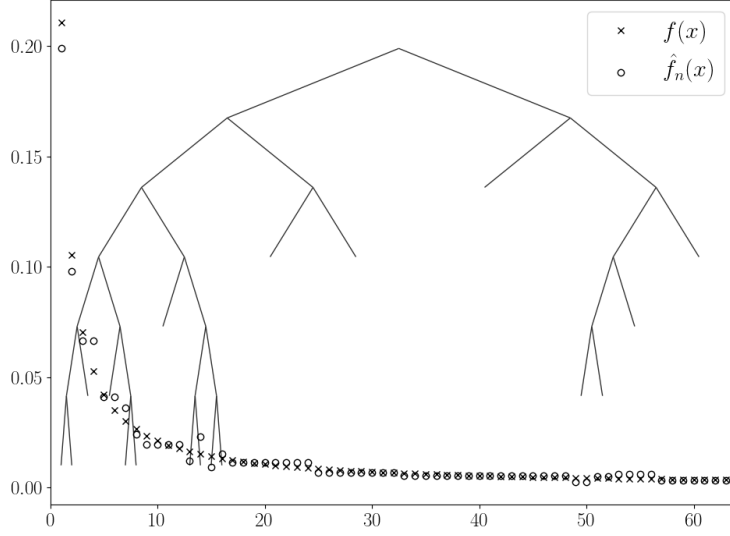


FIG 1. One sample of \hat{f}_n for $n = 1000$ and $k = 64$, where $f(x) = 1/(xH_k)$ for H_k the k -th Harmonic number. The tree \hat{T} is overlayed.

Remark 3.1. Intuitively, we justify the rule (5) as follows: We expect that N_v is at least as large as N_w by monotonicity of the density f , and the larger the difference $|N_v - N_w|$, the finer a partition around I_v and I_w should be to minimize the error of a piecewise constant estimate of f . However, even if N_v and N_w were equal in expectation, we expect with positive probability that N_v may deviate from N_w on the order of a standard deviation, i.e., on the order of $\sqrt{N_v + N_w}$, and this determines the threshold for splitting.

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 [7], 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

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

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 \tilde{f}_n , regardless of the order of merged intervals, and that

$$\text{TV}(\tilde{f}_n, f) \leq \text{TV}(\hat{f}_n, f).$$

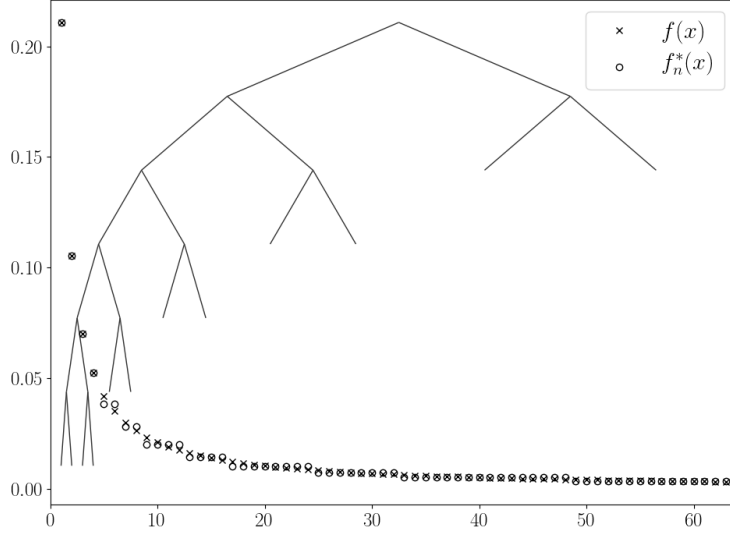


FIG 2. A plot of f_n^* for $n = 1000$ and $k = 64$, where $f(x) = 1/(xH_k)$. The tree T^* is overlaid.

3.2. An idealized tree-based estimator

Instead of analyzing the greedy tree-based estimator \hat{f}_n of the preceding section, we will fully analyze an idealized version. Indeed, in (5), the quantities N_z are distributed as $\text{Binomial}(n, f_z)$ for $z \in \{v, w\}$, where we define

$$f_z = \sum_{x \in I_z} f(x).$$

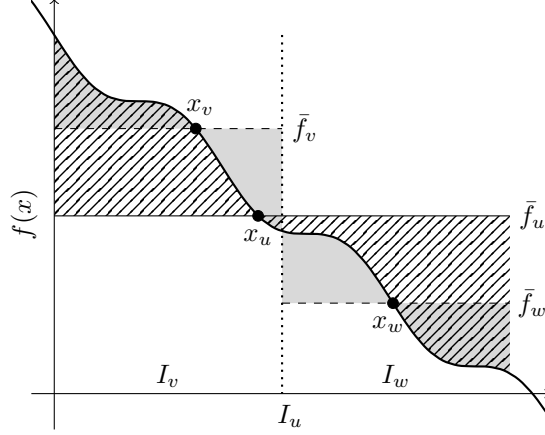
If we replace the quantities in (5) with their expectations, we obtain the *idealized splitting rule*

$$f_v - f_w > \sqrt{\frac{f_v + f_w}{n}}, \quad (6)$$

where we note that $f_v \geq f_w$, since f is non-increasing. Using the same procedure as in the preceding section, replacing the splitting rule with (6), we obtain a deterministic tree $T^* = T^*(f)$ with leaves L^* , and we set

$$f_n^*(x) = \bar{f}_u = \frac{f_u}{|I_u|}, \quad \text{if } x \in I_u, u \in L^*,$$

i.e., f_n^* is constant and equal to the average value of f on each interval I_u for $u \in L^*$. We call f_n^* the *idealized tree-based estimate*. See Figure 2 for a visualization of f_n^* and T^* . It may be instructive to compare Figure 2 to Figure 1.

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

Of course, T^* and f_n^* both depend intimately upon knowledge of the density f ; in practice, we only have access to the samples $X_1, \dots, X_n \stackrel{i.i.d.}{\sim} f$, 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 2.6 to get a minimax rate upper bound for \mathcal{F}_k .

Proposition 3.3.

$$\text{TV}(f_n^*, f) \leq \frac{3}{2} \sqrt{\frac{|\{u \in L^* : |I_u| > 1\}|}{n}}.$$

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

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

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 . In order to compute A_u , we refer to Figure 3. We view A_u as the positive area between the curve f and the line \bar{f}_u ; in Figure 3, this is the patterned area. Let \bar{f}_v and \bar{f}_w be the average value of f on I_v and I_w respectively. 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 3, and B_w is the positive area between f and \bar{f}_w

on I_w , the gray area on I_w in Figure 3. The points x_v, x_u, x_w are the unique points for which

$$f(x_z) = \bar{f}_z, \quad \text{for } z \in \{v, u, w\}.$$

Since the gray area on I_v to the left of x_v is equal to the gray area on I_v to the right of x_v , and the gray area on I_w to the left of x_w is equal to the gray area on I_w to the right of x_w , then

$$B_v + B_w \leq (\bar{f}_v - \bar{f}_w)|I_u|.$$

Then

$$\begin{aligned} A_u &\leq B_v + B_w + (\bar{f}_v - \bar{f}_w)|I_u|/2 \\ &\leq 3(\bar{f}_v - \bar{f}_w)|I_u|/2 \\ &= 3(f_v - f_w) \\ &\leq 3\sqrt{f_u/n}, \end{aligned}$$

where this last line follows from the splitting rule (6), since $u \in L^*$ and $|I_u| > 1$. So,

$$\begin{aligned} \text{TV}(f_n^*, f) &= \frac{1}{2} \sum_{u \in L^*} A_u \\ &\leq \frac{3}{2\sqrt{n}} \sum_{u \in L^*: |I_u| > 1} \sqrt{f_u} \\ &\leq \frac{3}{2} \sqrt{\frac{|\{u \in L^*: |I_u| > 1\}|}{n}}, \end{aligned}$$

where the last line follows from the Cauchy-Schwarz inequality. \square

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

$$|\{u \in L^*: |I_u| > 1\}| \leq 10n^{1/3} \left(\log_2(k/n^{1/3}) \right)^{2/3}.$$

Proof. Note that T^* has height at most $\log_2 k$. Let U_j be the set of nodes at depth j in T^* which have at least one leaf as a child, for $0 \leq j \leq \log_2 k$, and label the children of the nodes in U_j in order of appearance from right to left in T^* as $u_1, u_2, \dots, u_{2|U_j|}$. Since none of the nodes in U_j are themselves leaves, then by (6),

$$f_{u_2} - f_{u_1} > \sqrt{\frac{f_{u_1} + f_{u_2}}{n}},$$

and in particular since $f_{u_1} \geq 0$, then $f_{u_2} > \sqrt{f_{u_2}/n}$, so that $f_{u_2} > 1/n$. In general,

$$f_{u_{2i}} > f_{u_{2i-2}} + \sqrt{\frac{2f_{u_{2i-2}}}{n}},$$

and this recurrence relation can be solved to obtain that

$$f_{u_{2i}} \geq \frac{i^2}{n}. \quad (7)$$

Let L_j be the set of leaves at level j in T^* . The leaves at level $j+1$ form a subsequence $v_1, v_2, \dots, v_{|L_{j+1}|}$ of $u_1, u_2, \dots, u_{2|U_j|}$. Write q_j for the total probability mass of f held in the leaves L_j , i.e.,

$$q_j = \sum_{u \in L_j} f_u = \sum_{i=1}^{|L_j|} f_{v_i}.$$

By (7),

$$\sum_{i=1}^{|L_j|} f_{v_i} \geq \sum_{i=1}^{\lfloor |L_j|/2 \rfloor} f_{u_{2i}} \geq \sum_{i=1}^{\lfloor |L_j|/2 \rfloor} \frac{i^2}{n} \geq \frac{(\lfloor |L_j|/2 \rfloor)^3}{3n},$$

so that

$$|L_j| \leq 2 + 2(3nq_j)^{1/3} \leq 2 + 3(nq_j)^{1/3}.$$

Summing over all leaves except on the last level, and using the facts that $n \geq 64$ and $2n^{1/3} \leq k < n^{1/3}2^n$,

$$\begin{aligned} |\{u \in L^*: |I_u| > 1\}| &= \sum_{j=0}^{\lfloor (1/3)\log_2 n \rfloor - 1} |L_j| + \sum_{j=\lfloor (1/3)\log_2 n \rfloor}^{\log_2 k - 1} |L_j| \\ &\leq n^{1/3} + \sum_{j=\lfloor (1/3)\log_2 n \rfloor}^{\log_2 k - 1} (2 + 3(nq_j)^{1/3}) \\ &\leq n^{1/3} + 4\log_2(k/n^{1/3}) + 3n^{1/3} \sum_{j=\lfloor (1/3)\log_2 n \rfloor}^{\log_2 k - 1} q_j^{1/3}. \end{aligned}$$

By Hölder's inequality,

$$\begin{aligned} \sum_{j=\lfloor (1/3)\log_2 n \rfloor}^{\log_2 k - 1} q_j^{1/3} &\leq \left(\sum_{j=1}^{\log_2 k} q_j \right)^{1/3} \left(\sum_{j=\lfloor (1/3)\log_2 n \rfloor}^{\log_2 k - 1} 1 \right)^{2/3} \\ &\leq \left(2\log_2(k/n^{1/3}) \right)^{2/3}, \end{aligned}$$

so finally

$$\begin{aligned} |L^*| &\leq n^{1/3} + 4\log_2(k/n^{1/3}) + 5n^{1/3} \left(\log_2(k/n^{1/3}) \right)^{2/3} \\ &\leq 10n^{1/3} \left(\log_2(k/n^{1/3}) \right)^{2/3}. \end{aligned} \quad \square$$

Proof of the upper bound in Theorem 2.1. The case $k \geq n^{1/3}2^n$ is trivial, and follows simply because the TV-distance is always upper bounded by 1.

Suppose next that $2n^{1/3} > k$. In this regime, we can use a histogram estimator for f with bins of size 1 for each element of $\{1, \dots, k\}$. It is well known that risk of this estimator is on the order of $\sqrt{k/n}$ [13].

Finally, suppose that $2n^{1/3} \leq k < n^{1/3}2^n$. Let \mathcal{F}_Θ be the class of all piecewise-constant probability densities on $\{1, \dots, k\}$ which have $\ell = |\{u \in L^*: |I_u| > 1\}|$ parts; in particular, $f_n^* \in \mathcal{F}_\Theta$. Let \mathcal{A}_Θ be the Yatracos class of \mathcal{F}_Θ ,

$$\mathcal{A}_\Theta = \left\{ \{x: f(x) > g(x)\}: f \neq g \in \mathcal{F}_\Theta \right\}.$$

Then, $\mathcal{A}_\Theta \subseteq \mathcal{A}$, where \mathcal{A} is the class of all unions of 2ℓ intervals in \mathbb{N} . It is well known that $\text{VC}(\mathcal{A}) = 4\ell$, so $\text{VC}(\mathcal{A}_\Theta) \leq 4\ell$. By Corollary 2.6 and Proposition 3.3, there are universal constants $c_1, c_2 > 0$ for which

$$\begin{aligned} \mathcal{R}_n(\mathcal{F}_k) &\leq 3 \sup_{f \in \mathcal{F}_k} \inf_{\theta \in \Theta} \text{TV}(f_{n,\theta}, f) + c_1 \sqrt{\ell/n} \\ &\leq 3 \sup_{f \in \mathcal{F}_k} \text{TV}(f_n^*, f) + c_1 \sqrt{\ell/n} \\ &\leq c_2 \sqrt{\ell/n}. \end{aligned}$$

By Proposition 3.4, we see that for sufficiently large n , there is a universal constant $c_3 > 0$ such that

$$\mathcal{R}_n(\mathcal{F}_k) \leq c_3 \left(\frac{\log(k/n^{1/3})}{n} \right)^{1/3}. \quad \square$$

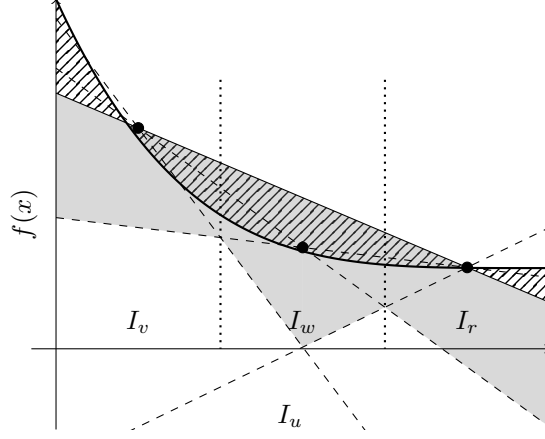
4. Non-increasing convex densities

Recall that \mathcal{G}_k is the class of non-increasing convex densities supported on $\{1, \dots, k\}$. Then, \mathcal{G}_k forms a subclass of \mathcal{F}_k , which we considered in Section 3. This section is devoted to extending the techniques of Section 3 in order to obtain a minimax rate upper bound on \mathcal{G}_k . Again, the lower bound is proved using standard techniques in Appendix A.2.

In this section, we assume that k is a power of three. In order to prove the upper bound of Theorem 2.2, we construct a ternary tree just as in Section 3, now with a ternary splitting rule, where if a node u has children v, w, r in order from left to right, we split and recurse if

$$f_v - 2f_w + f_r > \sqrt{\frac{f_v + f_w + f_r}{n}}. \quad (8)$$

Here we obtain a tree $T^\dagger = T^\dagger(f)$ with leaves L^\dagger . If $u \in L^\dagger$ has children v, w, r from left to right, let m_z be the midpoint of I_z for $z \in \{v, w, r\}$. Let the estimate f_n^* on I_u to be composed of the secant line passing through the points (m_v, \bar{f}_v) and (m_r, \bar{f}_r) . The estimate f_n^\dagger is again called the *idealized tree-based estimate* for f .

FIG 4. A visualization of the L_1 distance between f_n^\dagger and f on I_u .**Proposition 4.1.**

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

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

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

Proof of Proposition 4.1. As in the proof of Proposition 3.3, we have

$$\text{TV}(f_n^\dagger, f) = \frac{1}{2} \sum_{u \in L^\dagger : |I_u| > 1} A_u,$$

for $A_u = \sum_{x \in I_u} |f_n^\dagger(x) - f(x)|$. We refer to Figure 4 for a visualization of the quantity A_u , which is depicted as the patterned area.

Let B_v , B_w , and B_r denote the gray area on I_v , I_w , and I_r respectively. It can be shown that

$$A_u \leq 2B_v + B_w + 2B_r,$$

and

$$B_v = B_r = (f_v - 2f_w + f_r)/3, \quad B_w = 3(f_v - 2f_w + f_r)/8,$$

whence

$$A_u \leq 2(f_v - 2f_w + f_r).$$

The result then follows from the splitting rule (8) and the Cauchy-Schwarz inequality. \square

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

$$|\{u \in L^\dagger : |I_u| > 1\}| \leq 27n^{1/5} \left(\log_3(k/n^{1/5}) \right)^{4/5}.$$

Proof. The tree T^\dagger has height at most $\log_3 k$. Let U_j be the set of nodes at depth j in T^\dagger with at least one leaf as a child, for $0 \leq j \leq \log_3 k$, labelled in order of appearance from right to left in T^\dagger as $u_1, u_2, \dots, u_{3|U_j|}$. By the convex splitting rule (8), and since f is non-increasing,

$$f_{u_3} - f_{u_2} > f_{u_2} - f_{u_1} + \sqrt{\frac{f_{u_1} + f_{u_2} + f_{u_3}}{n}} \geq \sqrt{\frac{f_{u_3} - f_{u_2}}{n}},$$

so in particular, $f_{u_3} - f_{u_2} > 1/n$, and $f_{u_3} > 1/n$. In general,

$$\begin{aligned} f_{u_{3i}} - f_{u_{3i-1}} &> f_{u_{3(i-1)}} - f_{u_{3(i-1)-1}} + \sqrt{\frac{3f_{u_{3(i-1)}}}{n}} \\ &\geq f_{u_{3(i-1)}} - f_{u_{3(i-1)-1}} + \sqrt{\frac{3 \sum_{j=1}^{i-1} (f_{u_{3j}} - f_{u_{3j-1}})}{n}}. \end{aligned} \quad (9)$$

We claim now that $f_{u_{3i}} - f_{u_{3i-1}} > \frac{i^3}{27n}$, which we prove by induction; the base case is shown above, and by the induction hypothesis,

$$\begin{aligned} (9) &\geq \frac{(i-1)^3}{27n} + \sqrt{\frac{3 \sum_{j=1}^{i-1} (j^3/27n)}{n}} \\ &\geq \frac{(i-1)^3}{27n} + \frac{(i-1)^2}{6n} \\ &\geq \frac{i^3}{27n}, \end{aligned}$$

for all $i \geq 4$, while the cases $i = 2, 3$ can be manually verified. Then, by monotonicity of f ,

$$\begin{aligned} f_{u_{3i}} &> \frac{i^3}{27n} + f_{u_{3i-1}} \\ &\geq \frac{i^3}{27n} + f_{u_{3(i-1)}} \\ &\geq \sum_{j=1}^i \frac{j^3}{27n} \\ &\geq \frac{i^4}{108n}. \end{aligned} \quad (10)$$

Let now L_j be the set of leaves at level j in T^\dagger . The leaves at level $j+1$ form a subsequence $v_1, \dots, v_{|L_{j+1}|}$ of $u_1, \dots, u_{3|U_j|}$. Let q_j be the total probability mass

of f held in the leaves L_j . By (10),

$$q_j \geq \sum_{i=1}^{\lfloor |L_j|/3 \rfloor} f_{u_{3i}} \geq \sum_{i=1}^{\lfloor |L_j|/3 \rfloor} \frac{i^4}{108n} \geq \frac{(\lfloor |L_j|/3 \rfloor)^5}{540n},$$

so that

$$|L_j| \leq 3 + 3(540nq_j)^{1/5} \leq 3 + 11(nq_j)^{1/5}.$$

Summing over all leaves except on the last level,

$$\begin{aligned} |\{u \in L^\dagger : |I_u| > 1\}| &\leq n^{1/5} + \sum_{j=\lfloor (1/5) \log_3 n \rfloor}^{\log_3 k - 1} |L_j| \\ &\leq n^{1/5} + 6 \log_3(k/n^{1/5}) + 11n^{1/5} \sum_{j=\lfloor (1/5) \log_3 n \rfloor}^{\log_3 k - 1} q_j^{1/5}. \end{aligned}$$

By Hölder's inequality,

$$\begin{aligned} \sum_{j=\lfloor (1/5) \log_3 n \rfloor}^{\log_3 k - 1} q_j^{1/5} &\leq \left(\sum_{j=0}^{\log_3 k} q_j \right)^{1/5} \left(\sum_{j=\lfloor (1/5) \log_3 k \rfloor}^{\log_3 k - 1} 1 \right)^{4/5} \\ &\leq \left(2 \log_3(k/n^{1/5}) \right)^{4/5}, \end{aligned}$$

so finally

$$\begin{aligned} |\{u \in L^\dagger : |I_u| > 1\}| &\leq n^{1/5} + 6 \log_3(k/n^{1/5}) + 20n^{1/5} \left(\log_3(k/n^{1/5}) \right)^{4/5} \\ &\leq 27n^{1/5} \left(\log_3(k/n^{1/5}) \right)^{4/5}. \quad \square \end{aligned}$$

Proof of the upper bound in Theorem 2.2. The proof is similar to that of Theorem 2.1. \square

5. Discussion

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 [5]. 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. In order to estimate such a density, one could devise an oriented binary splitting rule analogous to (6) and carry out a similar analysis as performed in Section 3.2. 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 $\ell = 2$. See Balabdaoui and Wellner [3, 4] for texts concerning the density estimation of ℓ -monotone densities. Our approach also likely can be applied to the class of all log-concave discrete distributions, where we recall that $f : \mathbb{N} \rightarrow [0, 1]$ is called *log-concave* if

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

We note here that the optimal minimax rate for the class of d -dimensional log-concave continuous densities is unknown as of the time of writing [14, 21, 24].

Appendix A: Lower bounds

Lemma A.1 (Assouad's Lemma [2, 13]). *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 $g_{ij} : A_i \rightarrow \mathbb{R}$ for $0 \leq i \leq r$ and $j \in \{0, 1\}$ be some collection of functions. For $\theta = (\theta_1, \dots, \theta_r) \in \{0, 1\}^r$, define the function $f_\theta : \mathcal{X} \rightarrow \mathbb{R}$ 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

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

and

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

Let \mathcal{H} be the hypercube of densities

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

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

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

A.1. Proof of the lower bound in Theorem 2.1.

Suppose first that $e^8 n^{1/3} \leq k \leq n^{1/3} e^n$. Let A_1, \dots, 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

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

and in particular

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

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

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

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

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

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

$$2r \leq \frac{k}{2}, \quad \text{and} \quad \frac{e^{4\varepsilon r}}{2\varepsilon} \leq \frac{k}{2}. \quad (11)$$

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

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

so

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

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

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

or equivalently,

$$r \geq \frac{1}{4} \left(n \log^2(k/n^{1/3}) \right)^{1/3}.$$

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

$$\mathcal{R}_n(\mathcal{F}_k) \geq \frac{\varepsilon}{4} \left(1 - \sqrt{\frac{n\varepsilon^2}{r}} \right) \geq \frac{1}{32} \left(\frac{\log(k/n^{1/3})}{n} \right)^{1/3}. \quad (12)$$

So we need only verify that these choices of ε and r are compatible with (11). Since $k \geq e^8 n^{1/3}$, then there is an integer choice of r in the range

$$\frac{1}{4} \left(n \log^2(k/n^{1/3}) \right)^{1/3} \leq r \leq \frac{1}{2} \left(n \log^2(k/n^{1/3}) \right)^{1/3}.$$

In particular, we can verify that

$$2r \leq \left(n \log^2(k/n^{1/3}) \right)^{1/3} \leq \frac{k}{2},$$

since $2 \log^{2/3} x \leq x$ for all $x \geq 0$. Moreover, since $k \geq e^8 n^{1/3}$, then $\varepsilon \geq \frac{1}{2n^{1/3}}$, so that

$$\begin{aligned} \frac{e^{4\varepsilon r}}{2\varepsilon} &\leq n^{1/3} e^{4\varepsilon r} \\ &\leq n^{1/3} e^{\frac{1}{2} \log(k/n^{1/3})} \\ &= n^{1/3} \sqrt{\frac{k}{n^{1/3}}} \\ &\leq \frac{k}{2}, \end{aligned}$$

where this last inequality holds since $\sqrt{x} \leq x/2$ for all $x \geq e^8$, so that (12) is proved.

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

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

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

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

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

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

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

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

A.2. Proof of the lower bound in Theorem 2.2.

Let A_1, \dots, A_r be the partition in Lemma A.1, for an integer $r \geq 1$ to be specified. Let j_i be the smallest element of A_i , and suppose that each $|A_i|$ is chosen to be a positive multiple of 3. We will define the functions f_θ based on parameters $\beta_i, \Delta_i \in \mathbb{R}$, to be specified. Let g_{i0} linearly interpolate between the points

$$(j_i, \beta_i), \quad \text{and} \quad \left(j_i + \frac{|A_i|}{3}, \frac{\beta_{i+1} - \beta_i}{3} + \beta_i - \Delta_i \right),$$

on $\{j_i, j_i + 1, \dots, j_i + |A_i|/3 - 1\}$, and between the points

$$\left(j_i + \frac{|A_i|}{3}, \frac{\beta_{i+1} - \beta_i}{3} + \beta_i - \Delta_i\right), \quad \text{and} \quad (j_i + |A_i|, \beta_{i+1})$$

on $\{j_i + |A_i|/3, \dots, j_i + |A_i| - 1\}$. Let g_{i1} linearly interpolate between the points

$$(j_i, \beta_i), \quad \text{and} \quad \left(j_i + \frac{2|A_i|}{3}, \frac{2(\beta_{i+1} - \beta_i)}{3} + \beta_i - \Delta_i\right)$$

on $\{j_i, j_i + 1, \dots, j_i + 2|A_i|/3 - 1\}$, and between the points

$$\left(j_i + \frac{2|A_i|}{3}, \frac{2(\beta_{i+1} - \beta_i)}{3} + \beta_i - \Delta_i\right), \quad \text{and} \quad (j_i + |A_i|, \beta_{i+1})$$

on $\{j_i + 2|A_i|/3, \dots, j_i + |A_i| - 1\}$. Then, each f_θ will be nonincreasing as long as $\beta_i \geq \beta_{i+1}$ for each $1 \leq i \leq r$, and

$$\frac{2(\beta_{i+1} - \beta_i)}{3} + \beta_i - \Delta_i \geq \beta_{i+1} \iff \frac{\beta_i - \beta_{i+1}}{3} \geq \Delta_i. \quad (13)$$

Each f_θ will be convex as long as the largest slope on A_i is at most the smallest slope on A_{i+1} for each $1 \leq i \leq r$. Equivalently,

$$\frac{|A_{i+1}|}{|A_i|} \geq \frac{\frac{\beta_{i+1} - \beta_{i+2}}{3} + \Delta_{i+1}}{\frac{\beta_i - \beta_{i+1}}{3} - \Delta_i}. \quad (14)$$

Now, pick $\beta_i = (1 - \varepsilon)^{i-1} \beta$ for some $\beta \in \mathbb{R}$, $\varepsilon \in (0, 1)$ to be specified, and

$$\Delta_i = \frac{(\beta_i - \beta_{i+1})\varepsilon}{3} = \frac{\beta\varepsilon^2(1 - \varepsilon)^{i-1}}{6},$$

for which (13) is immediately satisfied. The condition (14) is then equivalent to

$$\frac{|A_{i+1}|}{|A_i|} \geq 1 + \varepsilon.$$

Pick $|A_1| = 3$. It is sufficient to make the choice

$$|A_i| \geq a_i = \frac{3}{(1 - \varepsilon)^{i-1}}.$$

Let $|A_i|$ be the smallest integer multiple of 3 at least as large as a_i , so that $a_i \leq |A_i| \leq a_i + 3$. If $\varepsilon \leq 1/2$, then

$$\sum_{i=1}^r |A_i| \leq 3r + \frac{3e^{2\varepsilon r}}{2\varepsilon}.$$

Since the support of our densities is $\{1, \dots, k\}$, this upper bound must not exceed k , so we impose that

$$3r \leq \frac{k}{2}, \quad \text{and} \quad \frac{3e^{2\varepsilon r}}{2\varepsilon} \leq \frac{k}{2}. \quad (15)$$

We must tune β in order for each f_θ to be a density. By monotonicity, we must have

$$1 \leq \sum_{i=1}^r \beta_i |A_i| \leq 6\beta r,$$

and

$$1 \geq \sum_{i=1}^r \beta_{i+1} |A_i| \geq 3(1-\varepsilon)\beta r,$$

so there is a choice of β where

$$\frac{1}{6r} \leq \beta \leq \frac{2}{3r},$$

as long as $\varepsilon \leq 1/2$. Now, fix $1 \leq i \leq r$. Then,

$$\int_{A_i} |f_\theta - f_{\zeta_i}| \geq \frac{|A_i| \Delta_i}{12} \geq \frac{\varepsilon^2 \beta}{12} \geq \frac{\varepsilon^2}{72r},$$

and

$$\begin{aligned} \int (\sqrt{f_\theta} - \sqrt{f_{\zeta_i}})^2 &= \int_{A_i} \frac{(f_\theta - f_{\zeta_i})^2}{(\sqrt{f_\theta} + \sqrt{f_{\zeta_i}})^2} \\ &\leq \frac{1}{4\beta_{i+1}} \int_{A_i} (f_\theta - f_{\zeta_i})^2 \\ &\leq \frac{|A_i| \Delta_i^2}{4\beta_{i+1}} \\ &\leq \frac{\left(3 + \frac{3}{(1-\varepsilon)^{i-1}}\right) \left(\frac{\beta \varepsilon^2 (1-\varepsilon)^{i-1}}{3}\right)^2}{4\beta(1-\varepsilon)^i} \\ &\leq \frac{2\varepsilon^4}{9r}, \end{aligned}$$

as long as $\varepsilon \leq 1/2$, whence

$$\int \sqrt{f_\theta f_{\zeta_i}} \geq 1 - \frac{\varepsilon^4}{9r}.$$

Now, pick

$$\varepsilon = \frac{1}{2} \left(\frac{\log(k/n^{1/5})}{n} \right)^{1/5},$$

and r for which

$$\sqrt{\frac{2n\varepsilon^4}{9r}} \leq \frac{1}{2},$$

or equivalently,

$$r \geq \frac{1}{18} n^{1/5} \log^{4/5}(k/n^{1/5}).$$

Note that $k \leq n^{1/5}e^n$ now implies that $\varepsilon \leq 1/2$. With this choice, Lemma A.1 has

$$\mathcal{R}_n(\mathcal{G}_k) \geq \frac{\varepsilon^2}{288} \left(1 - \sqrt{\frac{2n\varepsilon^4}{9r}}\right) \geq \frac{1}{1152} \left(\frac{\log(k/n^{1/5})}{n}\right)^{2/5}, \quad (16)$$

so it remains to verify that our choices of ε and r are compatible with (15). Since $k \geq e^{40}n^{1/5}$, then there is an integer choice of r in the range

$$\frac{1}{18}n^{1/5} \log^{4/5}(k/n^{1/5}) \leq r \leq \frac{1}{9}n^{1/5} \log^{4/5}(k/n^{1/5}).$$

In particular, we can verify that

$$3r \leq \frac{1}{3}n^{1/5} \log^{4/5}(k/n^{1/5}) \leq \frac{k}{2},$$

since $(2/3) \log^{4/5} x \leq x$ for all $x \geq 0$. Moreover, since $k \geq e^{40}n^{1/5}$, then $\varepsilon \geq 1/n^{1/5}$, so that

$$\begin{aligned} \frac{3e^{2\varepsilon r}}{2\varepsilon} &\leq \frac{3n^{1/5}e^{2\varepsilon r}}{2} \\ &\leq \frac{3n^{1/5}e^{\frac{1}{9}\log(k/n^{1/5})}}{2} \\ &= \frac{3n^{1/5}}{2} \left(\frac{k}{n^{1/5}}\right)^{1/9} \\ &\leq \frac{k}{2}, \end{aligned}$$

since $x^{1/9} \leq x/3$ for all $x \geq e^{40}$, so that (16) is proved.

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 A.1, 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,

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

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 $0 \leq \varepsilon \leq 1$, whence

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

By monotonicity,

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

and

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

so that the right choice of β_1 satisfies

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

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

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

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

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

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

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

so by Lemma A.1,

$$\mathcal{R}_n(\mathcal{G}_k) \geq \frac{\varepsilon}{24r^2} \left(1 - \sqrt{\frac{n\varepsilon^2}{24r^5}} \right) \geq \frac{1}{48e^{100}} \sqrt{k/n}. \quad \square$$

References

- [1] ANEVSKI, D. (2003). Estimating the derivative of a convex density. *Statist. Neerlandica* **57** 245–257. MR2028914
- [2] ASSOUD, 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] BIAU, G. and DEVROYE, L. (2003). On the risk of estimates for block decreasing densities. *J. Multivariate Anal.* **86** 143–165. MR1994726
- [6] BIRGÉ, L. (1987). Estimating a density under order restrictions: nonasymptotic minimax risk. *Ann. Statist.* **15** 995–1012. MR902241
- [7] BIRGÉ, L. (1987). On the risk of histograms for estimating decreasing densities. *Ann. Statist.* **15** 1013–1022. MR902242
- [8] BIRGÉ, L. (1989). The Grenander estimator: a nonasymptotic approach. *Ann. Statist.* **17** 1532–1549. MR1026298
- [9] 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

- [10] DEVROYE, L. (1987). *A Course in Density Estimation. Progress in Probability and Statistics* **14**. Birkhäuser Boston, Inc., Boston, MA. MR891874
- [11] 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
- [12] 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
- [13] DEVROYE, L. and LUGOSI, G. (2001). *Combinatorial Methods in Density Estimation. Springer Series in Statistics*. Springer-Verlag, New York. MR1843146
- [14] 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.
- [15] GESSAMAN, M. P. (1970). A consistent nonparametric multivariate density estimator based on statistically equivalent blocks. *Ann. Math. Statist.* **41** 1344–1346.
- [16] GRENANDER, U. (1956). On the theory of mortality measurement. I, II. *Skand. Aktuarietidskr.* **39** 70–96, 125–153.
- [17] 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
- [18] 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
- [19] GROENEBOOM, P., JONGBLOED, G. and WELLNER, J. A. (2001). Estimation of a convex function: characterizations and asymptotic theory. *Ann. Statist.* **29** 1653–1698. MR1891742
- [20] JONGBLOED, G. (1995). Three Statistical Inverse Problems, PhD thesis, Delft University of Technology.
- [21] KIM, A. K. H. and SAMWORTH, R. J. (2016). Global rates of convergence in log-concave density estimation. *Ann. Statist.* **44** 2756–2779. MR3576560
- [22] LUGOSI, G. and NOBEL, A. (1996). Consistency of data-driven histogram methods for density estimation and classification. *Ann. Statist.* **24** 687–706. MR1394983
- [23] PRAKASA RAO, B. L. S. (1969). Estimation of a unimodal density. *Sankhyā Ser. A* **31** 23–36. MR0267677
- [24] SAMWORTH, R. J. (2017). Recent progress in log-concave density estimation. *CoRR* **abs/1709.03154**.
- [25] SASON, I. and VERDÚ, S. (2016). f -divergence inequalities. *IEEE Trans. Inform. Theory* **62** 5973–6006. MR3565096
- [26] 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

- [27] SILVERMAN, B. W. (1986). *Density Estimation for Statistics and Data Analysis. Monographs on Statistics and Applied Probability*. Chapman & Hall, London. MR848134
- [28] 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
- [29] YU, B. (1997). Assouad, Fano, and Le Cam. In *Festschrift for Lucien Le Cam* 423–435. Springer, New York. MR1462963