Appendix: Beyond Smoothness: Incorporating Low-Rank Analysis into Nonparametric Density Estimation

Robert A. Vandermeulen
Machine Learning Group
Technische Universiät Berlin
Berlin, Germany
vandermeulen@tu-berlin.de

Antoine Ledent
Machine Learning Group
Technische Universität Kaiserslautern
Kaiserslautern, Germany
ledent@cs.uni-kl.de

Appendix Overview

Section A contains the basic theoretical tools we need for the rest of the appendix. We will also be using some notation from the paper without introduction. In Section B we have all the proofs for the results relating to asymptotic parameter rates of the histogram spaces in our estimators; this section corresponds to Section 2.2.1 in the main text. Section C contains the proofs relating to finite-sample rates, and several expanded or additional results, corresponding to Section 2.2.2 in the main text. Section 2.2.2 in the main text contains some results that are simplified for readability and this appendix has some results which are tighter, but less readable. Additionally Section C contains all the finite-sample analysis for Tucker histograms which were omitted in the main text. Experimental details, specifically the equivalency of L^2 minimization and nonnegative tensor factorization can be found in Section D. Section E is a bit different from the other sections and contains some discussion relevant to the future of this direction of research. There we show that not all densities can be written as a summation of separable densities with nonnegative coefficients.

A Theoretical Basics

A.1 Notation

In this section we include some notation that was not contained in the main text but will be necessary for the rest of the appendix. In particular we will we need to introduce a fair amount of tensor notation which will be used to represent the various histograms. A histogram on the unit cube can naturally be represented as a tensor, which will be useful for deriving many of the results in this paper.

From the main text recall that $\mathcal{T}_{d,b}$ is the set of probability tensors in $\mathbb{R}^{b^{\times d}}$, i.e. their entries are nonnegative and sum to one.

Let $\mathcal{T}_{d,b}^k$ be the set of tensors that are a convex combination of k separable probability tensors (which are analogous to multi-view models) i.e.

$$\mathcal{T}_{d,b}^k \triangleq \left\{ \sum_{i=1}^k w_i \prod_{j=1}^d p_{i,j} \middle| w \in \Delta_k, p_{i,j} \in \Delta_b \right\}.$$

35th Conference on Neural Information Processing Systems (NeurIPS 2021).

The following is the set of probability tensors constructed via a nonnegative Tucker factorization ($[k]^d$ represents a multi-index)

$$\widetilde{\mathcal{T}}_{d,b}^{k} \triangleq \left\{ \sum_{S\in[k]^d} W_S \prod_{i=1}^{d} p_{i,S_i} \middle| W \in \mathcal{T}_{d,k}, p_{i,j} \in \Delta_b \right\}.$$

For a multi-index $A \in [b]^d$ we define $\mathbf{e}_{d,b,A}$ as the element of $\mathcal{T}_{d,b}$ where the $(A_1, \ldots, A_d)$-th entry is one and is zero elsewhere.

Note that there exists a $\ell^1 \to L^1$ linear isometry $U_{d,b} : \mathcal{T}_{d,b} \to \mathcal{H}_{d,b}$ with $U_{d,b}$ defined as

$$U_{d,b}(\mathbf{e}_{d,b,A}) = h_{d,b,A}.$$

The inverse function, $U_{d,b}^{-1}$, simply transforms a histogram to the tensor representing its bin weights and $U_{d,b}$ performs the reverse transformation. Note that $U_{d,b}$ is also a bijection between $\mathcal{T}_{d,b}^k \to \mathcal{H}_{d,b}^k$ and $\widetilde{\mathcal{T}}_{d,b}^k \to \widetilde{\mathcal{H}}_{d,b}^k$. Much of our analysis on histograms will be performed on the space of probability tensors with the analysis being translated to histograms via this operator.

For a set of vectors $\mathcal{V}$ we define $k\text{-mix}(\mathcal{V}) \triangleq \left\{\sum_{i=1}^k w_i v_i \middle| w \in \Delta_k, v_i \in \mathcal{V}\right\}$, i.e. the set of convex combinations of collections of k vectors from $\mathcal{V}$. We define $N(\mathcal{V}, \varepsilon)$ to be the minimum cardinality for a subset of of $\mathcal{V}$ which ε-covers $\mathcal{V}$ (with closed balls) with respect to the $\|\cdot\|_1$ metric. It will be clear from context whether $\|\cdot\|_1$ represents the ℓ^1, L^1, or total variation norm.

A.2 Preliminary Results

The following lemmas will be useful for all the theoretical results.

Lemma A.1. *For all* $0 < \varepsilon \leq 1$ *we have that* $N(\Delta_b, \varepsilon) \leq \left(\frac{2b}{\varepsilon}\right)^b$.

Lemma A.2. *For all* $0 < \varepsilon \leq 1$ *we have that* $N\left(\mathcal{T}_{d,b}^1, \varepsilon\right) \leq \left(\frac{2bd}{\varepsilon}\right)^{bd}$.

Lemma A.3. *Let* $\mathcal{P}$ *be a set of probability measures, then*

$$N(k\text{-mix}(\mathcal{P}), \varepsilon + \delta) \leq N(\mathcal{P}, \varepsilon)^k N(\Delta_k, \delta).$$

Lemma A.4. *For all* $0 < \varepsilon \leq 1$ *the following holds* $N\left(\mathcal{T}_{d,b}^k, \varepsilon\right) \leq \left(\frac{4bd}{\varepsilon}\right)^{bdk} \left(\frac{4k}{\varepsilon}\right)^k$.

Through application of the $U_{d,b}$ operator we now have a characterization of the complexity of the space $\mathcal{H}_{d,b}^k$.

Corollary A.1. *For all* $0 < \varepsilon \leq 1$ *following holds* $N\left(\mathcal{H}_{d,b}^k, \varepsilon\right) \leq \left(\frac{4bd}{\varepsilon}\right)^{bdk} \left(\frac{4k}{\varepsilon}\right)^k$.

The following are analogous results for Tucker histograms.

Lemma A.5. *For all* $0 < \varepsilon \leq 1$ *the following holds* $N\left(\widetilde{\mathcal{T}}_{d,b}^k, \varepsilon\right) \leq \left(\frac{4bd}{\varepsilon}\right)^{bdk} \left(\frac{4k^d}{\varepsilon}\right)^{k^d}$.

Corollary A.2. *For all* $0 < \varepsilon \leq 1$ *following holds* $N\left(\widetilde{\mathcal{H}}_{d,b}^k, \varepsilon\right) \leq \left(\frac{4bd}{\varepsilon}\right)^{bdk} \left(\frac{4k^d}{\varepsilon}\right)^{k^d}$.

The following lemma from [1] provides us with a way to choose good estimators from finite collections of densities. It can be proven by applying a Chernoff bound to [3], Theorem 6.3.

Lemma A.6 (Thm 3.4 page 7 of [1], Thm 3.6 page 54 of [3])**.** *There exists a deterministic algorithm that, given a collection of distributions* $p_1, \ldots, p_M$*, a parameter* $\varepsilon > 0$ *and at least* $\frac{\log(3M^2/\delta)}{2\varepsilon^2}$ *iid samples from an unknown distribution* p*, outputs an index* $j \in [M]$ *such that*

$$\|p_j - p\|_1 \leq 3 \min_{i\in[M]} \|p_i - p\|_1 + 4\varepsilon$$

with probability at least $1 - \frac{\delta}{3}$.

We present the following asymptotic version of the previous lemma. We highlight the use of finding sufficiently slow rates on parameters in order to establish asymptotic results, a technique which we will use in later proofs.

Lemma A.7. *Let $(\mathcal{P}_n)_{n\in\mathbb{N}}$ be a sequence of finite collections of densities in $\mathcal{D}_d$ where $|\mathcal{P}_n| \to \infty$ with $n/\log(|\mathcal{P}_n|) \to \infty$. Then there exists a sequence of estimators $V_n \in \mathcal{P}_n$ such that, for all $\gamma > 0$,*

$$\sup_{p\in\mathcal{D}_d} P\left(\|V_n - p\|_1 > 3 \min_{q\in\mathcal{P}_n} \|p-q\|_1 + \gamma\right) \to 0,$$

where V_n is a function of $X_1, \ldots, X_n \overset{iid}{\sim} p$.

Proof of Lemma A.7. Let $M = M(n) = |\mathcal{P}_n|$. Since $n/\log(M) \to \infty$ we have that for all $c > 0$ there exists a N_c such that, for all $n \geq N_c$ we have $n/\log(M) \geq c$ or equivalently $n \geq c\log(M)$. Because of this there exists sequence of positive values $C = C(n)$ such that $C \to \infty$ and $n \geq C\log(M)$.

We will be making use of the algorithm in Lemma A.6 as well as its notation. If we can show that there exist sequences of positive values $\varepsilon(n) \to 0, \delta(n) \to 0$ such that, for sufficiently large n, the following holds

$$\frac{\log\left(3M^2/\delta\right)}{2\varepsilon^2} \leq n,$$

then can simply set V_n equal to be the estimator from Lemma A.6 for sufficiently large n and, because the lemma holds independent of choice of p, the theorem statement follows.

Let $\varepsilon = (2/C)^{1/4}$ and $\delta = 3/\left(\exp\left(2\sqrt{\frac{C}{2}}\right)\right)$. Note that these are both positive sequences which converge to zero. Now we have

$$\begin{aligned}
\frac{\log\left(3M^2/\delta\right)}{2\varepsilon^2} &= \frac{\log\left(M^2\right)+\log(3/\delta)}{2\varepsilon^2} \\
&= \frac{2\log(M)+\log(3/\delta)}{2\varepsilon^2} = \frac{\log(M)+\frac{1}{2}\log(3/\delta)}{\varepsilon^2} \\
&= \varepsilon^{-2}\left(\log(M) + \frac{1}{2}\log(3/\delta)\right) \\
&= \left((2/C)^{1/4}\right)^{-2}\left(\log(M) + \frac{1}{2}\log\left(\exp\left(2\sqrt{\frac{C}{2}}\right)\right)\right) \\
&= \sqrt{\frac{C}{2}}\left(\log(M) + \sqrt{\frac{C}{2}}\right) = \sqrt{\frac{C}{2}}\log(M) + \frac{C}{2}. \qquad (1)
\end{aligned}$$

For sufficiently large C and M we have that the RHS of (1) is less than or equal to

$$\begin{aligned}
\frac{C}{2}\log(M) + \frac{C}{2} &\leq \frac{C}{2}\log(M) + \frac{C}{2}\log(M) \\
&= C\log(M) \leq n.
\end{aligned}$$

which completes our proof. □

A.3 Theoretical Basics Proofs

All norms are either the ℓ^1, L^1, or total variation norm, which are equivalent with respect to our analysis and the proper norm will be clear from context.

Proof of Lemma A.1. In Section 7.4 from [3], the authors show that for any collection of measures $\mu_1, \ldots, \mu_b$, for all $\varepsilon > 0$, that

$$N\left(\mathrm{Conv}\left(\{\mu_1, \ldots, \mu_b\}\right), \varepsilon\right) \leq \left(b + \frac{b}{\varepsilon}\right)^b.$$

With the additional assumption that $\varepsilon \le 1$ we have that $b + \frac{b}{\varepsilon} \le \frac{b}{\varepsilon} + \frac{b}{\varepsilon} = \frac{2b}{\varepsilon}$ and thus

$$N\left(\mathrm{Conv}\left(\{\mu_1, \ldots, \mu_b\}\right)\right) \le \left(\frac{2b}{\varepsilon}\right)^b .$$

If we let $\mu_i = \mathbf{e}_i$, the indicator vector at index i, then the lemma follows. □

Proof of Lemma A.2. From Lemma A.1 we know there exists a finite collection of probability vectors $\widetilde{\mathcal{P}}$ such that $\widetilde{\mathcal{P}}$ is an ε/d-covering of Δ_b and $\left|\widetilde{\mathcal{P}}\right| \le \left(\frac{2bd}{\varepsilon}\right)^b$. Note that the set $\left\{\widetilde{p}_1 \otimes \cdots \otimes \widetilde{p}_d \,|\, \widetilde{p}_i \in \widetilde{\mathcal{P}}\right\}$ contains at most $\left(\left(\frac{2bd}{\varepsilon}\right)^b\right)^d = \left(\frac{2bd}{\varepsilon}\right)^{bd}$ elements. We will now show that this set is an ε-cover of $\mathcal{T}^1_{d,b}$. Let $p_1 \otimes \cdots \otimes p_d \in \mathcal{T}^1_{d,b}$ be arbitrary. From our construction of $\widetilde{\mathcal{P}}$ there exist elements $\widetilde{p}_1, \ldots, \widetilde{p}_d \in \widetilde{\mathcal{P}}$ such that $\|p_i - \widetilde{p}_i\|_1 \le \frac{\varepsilon}{d}$.

We will now make use of Lemma 3.3.7 in [10], which states that, for any collection of probability vectors $q_1, \ldots, q_d$ and $\widetilde{q}_1, \ldots, \widetilde{q}_d$, the following holds

$$\left\|\prod_{i=1}^d q_i - \prod_{j=1}^d \widetilde{q}_j\right\|_1 \le \sum_{i=1}^d \|q_i - \widetilde{q}_i\|_1 .$$

From this it follows that

$$\left\|\prod_{i=1}^d p_i - \prod_{j=1}^d \widetilde{p}_j\right\|_1 \le \sum_{i=1}^d \|p_i - \widetilde{p}_i\|_1 \le d\frac{\varepsilon}{d} = \varepsilon$$

thus completing our proof. □

Proof of Lemma A.3. Let $\widetilde{\mathcal{P}}$ be the finite collection of probability measures with $|\widetilde{\mathcal{P}}| = N(\mathcal{P}, \varepsilon)$ which ε-covers $\mathcal{P}$. Similarly let $W \subset \Delta_k$ with $|W| = N(\Delta_k, \delta)$ such that W is a δ-cover of Δ_k. Consider the set

$$\Omega = \left\{\sum_{i=1}^k \widetilde{w}_i \widetilde{p}_i \middle| \widetilde{w} \in W, \widetilde{p}_i \in \widetilde{\mathcal{P}}\right\}.$$

Note that this set contains at most $N(\mathcal{P}, \varepsilon)^k N(\Delta_k, \delta)$ elements. We will now show that it $(\delta + \varepsilon)$-covers $k\text{-mix}(\mathcal{P})$, which completes the proof. Let $\sum_{i=1}^k p_i w_i \in k\text{-mix}(\mathcal{P})$. We know there exists elements $\widetilde{p}_1, \ldots, \widetilde{p}_k \in \widetilde{\mathcal{P}}$ such that $\|\widetilde{p}_i - p_i\|_1 \le \varepsilon$ and $\widetilde{w} \in W$ such that $\|w - \widetilde{w}\|_1 \le \delta$ and thus $\sum_{i=1}^k \widetilde{p}_i \widetilde{w}_i \in \Omega$. Now observe that

$$\begin{aligned}
\left\|\sum_{i=1}^k \widetilde{p}_i \widetilde{w}_i - \sum_{j=1}^k p_j w_j\right\|_1 &= \left\|\sum_{i=1}^k \widetilde{p}_i \widetilde{w}_i - \sum_{j=1}^k p_j \widetilde{w}_j + \sum_{l=1}^k p_l \widetilde{w}_l - \sum_{r=1}^k p_r w_r\right\|_1 \\
&\le \left\|\sum_{i=1}^k \left(\widetilde{p}_i - p_i\right) \widetilde{w}_i\right\|_1 + \left\|\sum_{i=1}^k p_i \left(\widetilde{w}_i - w_i\right)\right\|_1 \\
&\le \sum_{i=1}^k \widetilde{w}_i \|\widetilde{p}_i - p_i\|_1 + \sum_{i=1}^k |\widetilde{w}_i - w_i| \\
&\le \sum_{i=1}^k \widetilde{w}_i \varepsilon + \|\widetilde{w} - w\|_1 \\
&\le \varepsilon + \delta.
\end{aligned}$$

□

Proof of Lemma A.4. Note that $\mathcal{T}_{d,b}^k = k\text{-mix}\left(\mathcal{T}_{d,b}^1\right)$. Applying Lemma A.3 followed by Lemmas A.1 and A.2 we have that

$$N\left(\mathcal{T}_{d,b}^k, \varepsilon\right) \le N\left(\mathcal{T}_{d,b}^1, \varepsilon/2\right)^k N\left(\Delta_k, \varepsilon/2\right) \le \left(\frac{4bd}{\varepsilon}\right)^{bdk} \left(\frac{4k}{\varepsilon}\right)^k.$$

□

Proof of Lemma A.5. Fix k, d, b and $0 < \varepsilon \le 1$. We are going to construct an ε-cover of $\widetilde{\mathcal{T}}_{d,b}^k$. From Lemma A.1 we know that there exists a set $\mathcal{B} \subset \Delta_b$ which $\left(\frac{\varepsilon}{2d}\right)$-covers of Δ_b and contains no more than $\left(\frac{4bd}{\varepsilon}\right)^b$ elements. Let $\mathcal{P}$ be the collection of all $d \times k$ arrays whose entries are elements from $\mathcal{B}$. So we have that

$$|\mathcal{P}| = |\mathcal{B}|^{dk} \le \left(\frac{4bd}{\varepsilon}\right)^{bdk}.$$

From Lemma A.1 there exists $\mathcal{W}$ which is an $\varepsilon/2$-cover of $\mathcal{T}_{d,k}$ and contains no more than $\left(4k^d/\varepsilon\right)^{\left(k^d\right)}$ elements. Now let

$$\mathcal{L}_{d,b}^k = \left\{ \sum_{S\in[k]^d} \widetilde{W}_S \prod_{i=1}^d \widetilde{p}_{i,S_i} \middle| \widetilde{W} \in \mathcal{W}, \widetilde{p} \in \mathcal{P} \right\}.$$

Note that

$$\left|\mathcal{L}_{d,b}^k\right| \le |\mathcal{W}|\,|\mathcal{P}| \le \left(\frac{4k^d}{\varepsilon}\right)^{k^d} \left(\frac{4bd}{\varepsilon}\right)^{bdk}.$$

We will now show that $\mathcal{L}_{d,b}^k$ is an ε-cover of $\widetilde{\mathcal{T}}_{d,b}^k$. To this end let $\sum_{S\in[k]^d} W_S \prod_{i=1}^d p_{i,S_i} \in \widetilde{\mathcal{T}}_{d,b}^k$ be arbitrary, where $W \in \mathcal{T}_{d,k}$ and $p_{i,j} \in \Delta_b$. From our construction of $\mathcal{W}$, there exists $\widetilde{W} \in \mathcal{W}$ such that $\left\|W - \widetilde{W}\right\|_1 \le \varepsilon/2$. There also exists $\widetilde{p} \in \mathcal{P}$ such that $\left\|\widetilde{p}_{i,j} - p_{i,j}\right\|_1 \le \varepsilon/2$ for all i, j. Therefore we have that

$$\sum_{S\in[k]^d} \widetilde{W}_S \prod_{i=1}^d \widetilde{p}_{i,S_i} \in \mathcal{L}_{d,b}^k.$$

So finally

$$\begin{aligned}
&\left\| \sum_{S\in[k]^d} W_S \prod_{i=1}^d p_{i,S_i} - \sum_{R\in[k]^d} \widetilde{W}_R \prod_{j=1}^d \widetilde{p}_{j,R_j} \right\|_1 \\
&\le \left\| \sum_{S\in[k]^d} W_S \prod_{i=1}^d p_{i,S_i} - \sum_{R\in[k]^d} W_R \prod_{j=1}^d \widetilde{p}_{j,R_j} \right\|_1 + \left\| \sum_{S\in[k]^d} W_S \prod_{i=1}^d \widetilde{p}_{i,S_i} - \sum_{R\in[k]^d} \widetilde{W}_R \prod_{j=1}^d \widetilde{p}_{j,R_j} \right\|_1 \\
&\le \sum_{S\in[k]^d} W_S \left\| \prod_{i=1}^d p_{i,S_i} - \prod_{j=1}^d \widetilde{p}_{j,S_j} \right\|_1 + \sum_{R\in[k]^d} \left|W_R - \widetilde{W}_R\right| \left\| \prod_{j=1}^d \widetilde{p}_{j,R_j} \right\|_1 \\
&\le \sum_{S\in[k]^d} W_S \sum_{i=1}^d \left\|p_{i,S_i} - \widetilde{p}_{i,S_i}\right\|_1 + \sum_{R\in[k]^d} \left|W_R - \widetilde{W}_R\right| \left\| \prod_{j=1}^d \widetilde{p}_{j,R_j} \right\|_1 \\
&\le \sum_{S\in[k]^d} W_S \frac{\varepsilon}{2} + \left\|W - \widetilde{W}\right\|_1 \\
&\le \varepsilon/2 + \varepsilon/2 = \varepsilon.
\end{aligned}$$

□

B Asymptotic Theoretical Results

This section contains results related to the asymptotic results related to the growth of model parameters with respect to the number of training samples. It corresponds to Section 2.2.1 in the main text.

Proof of Theorem 2.1. We will be applying the estimator from Lemma A.7 to a series of δ-covers of $\mathcal{H}_{d,b}^k$. We begin by constructing a series of δ-covers whose cardinality doesn't grow too quickly. Corollary A.1 states that, for all $0 < \delta \le 1$, that $N\left(\mathcal{H}_{d,b}^k, \delta\right) \le \left(\frac{4bd}{\delta}\right)^{bdk} \left(\frac{4k}{\delta}\right)^k$. *For sufficiently large* b and k and *sufficiently small* δ, the following holds

$$\begin{aligned}
\log\left(\left(\frac{4bd}{\delta}\right)^{bdk}\left(\frac{4k}{\delta}\right)^{k}\right) &= bdk\log\left(\frac{4bd}{\delta}\right) + k\log\left(\frac{4k}{\delta}\right)\\
&= bdk\left[\log\left(b\right) + \log\left(\frac{4d}{\delta}\right)\right] + k\left[\log\left(k\right) + \log\left(\frac{4}{\delta}\right)\right]\\
&\le bdk\left[\log\left(b\right) + \log\left(b\right)\log\left(\frac{4d}{\delta}\right)\right] + dk\left[\log\left(k\right) + \log\left(k\right)\log\left(\frac{4d}{\delta}\right)\right]\\
&= \left(bk\log\left(b\right) + k\log\left(k\right)\right)d\left(1 + \log\left(\frac{4d}{\delta}\right)\right). \qquad (2)
\end{aligned}$$

Using the argument from the proof of Lemma A.7 we have that, because $n/(bk\log(b) + k\log(k)) \to \infty$ there exists a sequence of positive values $C = C(n)$ such that $C \to \infty$ and $n > C\left[bk\log(b) + k\log(k)\right]$. If we let $\delta = \frac{4d}{\exp\left(\frac{C}{d}-1\right)}$ we have that $\delta \to 0$ and

$$\left(bk\log\left(b\right) + k\log\left(k\right)\right)d\left(1 + \log\left(\frac{4d}{\delta}\right)\right) \le n.$$

Because of this we can construct collections of densities $\widetilde{\mathcal{P}}_n \subset \mathcal{H}_{d,b}^k$ such that $\widetilde{\mathcal{P}}_n$ is a δ-covering of $\mathcal{H}_{d,b}^k$ with $\left|\widetilde{\mathcal{P}}\right| \to \infty$, $n/\log\left|\widetilde{\mathcal{P}}_n\right| \to \infty$ and $\delta \to 0$. Let V_n be the estimator from Lemma A.7 applied to the sequence $\widetilde{\mathcal{P}}_n$.

Let $\varepsilon > 0$ be arbitrary. Due to the way that we have constructed the sequence $\widetilde{\mathcal{P}}_n$, for sufficiently large n, we have that $3\sup_{q\in\mathcal{H}_{d,b}^k}\min_{\widetilde{q}\in\widetilde{\mathcal{P}}_n}\|q-\widetilde{q}\|_1 \le \varepsilon/2$. It therefore follows that, for sufficiently large n, the following holds for all $p \in \mathcal{D}_d$

$$\begin{aligned}
3\min_{q\in\mathcal{H}_{d,b}^k}\|p-q\|_1 + \varepsilon &\ge 3\min_{q\in\mathcal{H}_{d,b}^k}\|p-q\|_1 + 3\sup_{q\in\mathcal{H}_b^k}\min_{\widetilde{q}\in\widetilde{\mathcal{P}}_n}\|q-\widetilde{q}\|_1 + \varepsilon/2\\
&\ge 3\min_{q\in\mathcal{H}_{d,b}^k}\left[\|p-q\|_1 + \min_{\widetilde{q}\in\widetilde{\mathcal{P}}_n}\|q-\widetilde{q}\|_1\right] + \varepsilon/2\\
&= 3\min_{q\in\mathcal{H}_{d,b}^k}\min_{\widetilde{q}\in\widetilde{\mathcal{P}}_n}\|p-q\|_1 + \|q-\widetilde{q}\|_1 + \varepsilon/2\\
&\ge 3\min_{\widetilde{q}\in\widetilde{\mathcal{P}}_n}\|p-\widetilde{q}\|_1 + \varepsilon/2.
\end{aligned}$$

From this we have that, for sufficiently large n

$$\sup_{p\in\mathcal{D}_d} P\left(\|V_i - p\|_1 > 3\min_{q\in\mathcal{H}_{d,b}^k}\|p-q\|_1 + \varepsilon\right) \le \sup_{p\in\mathcal{D}_d} P\left(\|V_i - p\|_1 > 3\min_{\widetilde{q}\in\widetilde{\mathcal{P}}_n}\|p-\widetilde{q}\|_1 + \varepsilon/2\right)$$

and the right side goes to zero due to Lemma A.7, thus completing the proof. $\square$

Proof of Theorem 2.2. This proof is very similar to the proof of Theorem 2.1. We will be applying the estimator from Lemma A.7 to a series of δ-covers of $\widetilde{\mathcal{H}}_{d,b}^k$. We begin by constructing a series of δ-covers whose cardinality doesn't grow too quickly. Corollary A.2 states that, for all $0 < \delta \le 1$, that $N\left(\widetilde{\mathcal{H}}_{d,b}^k,\delta\right) \le \left(\frac{4bd}{\delta}\right)^{bdk}\left(\frac{4k^d}{\delta}\right)^{k^d}$. *For sufficiently large* b and k and *sufficiently small* δ, the following holds

$$
\begin{aligned}
\log\left(\left(\frac{4bd}{\delta}\right)^{bdk}\left(\frac{4k^d}{\delta}\right)^{k^d}\right) &= bdk\log\left(\frac{4bd}{\delta}\right) + k^d\log\left(\frac{4k^d}{\delta}\right)\\
&\le d\left(bk\log\left(\frac{4bd}{\delta}\right) + k^d\log\left(\frac{4k^d}{\delta}\right)\right)\\
&= d\left(bk\left(\log(b)+\log\left(\frac{4d}{\delta}\right)\right) + k^d\left(\log\left(k^d\right)+\log\left(\frac{4}{\delta}\right)\right)\right)\\
&\le d\left(bk\left(\log(b)+\log\left(\frac{4d}{\delta}\right)\right) + k^d\left(\log\left(k^d\right)+\log\left(\frac{4d}{\delta}\right)\right)\right)\\
&= \left(bk\log(b)+k^d\log\left(k^d\right)\right)d\left(1+\log\left(\frac{4d}{\delta}\right)\right).
\end{aligned}
$$

Note that replacing $bk\log(b)+k\log(k)$ with $bk\log(b)+k^d\log\left(k^d\right)$ in the last line is exactly (2) in our proof of Theorem 2.1 . From here we can proceed exactly as in the proof of Theorem 2.1 by replacing $\mathcal{H}_{d,b}^k$ with $\widetilde{\mathcal{H}}_{d,b}^k$ and $bk\log(b)+k\log(k)$ with $bk\log(b)+k^d\log\left(k^d\right)$. $\square$

Proof of Corollary 2.1. This follows directly from Theorems 2.1 and 2.2 and selecting an appropriately slow rate for $k\to\infty$. $\square$

Proof of Lemma 2.1. Let $\varepsilon>0$. Theorem 5 in Chapter 2 of [6][1] states that, for any $p\in\mathcal{D}_d$, that $\min_{h\in\mathcal{H}_{d,b}}\|p-h\|_1\to 0$ as $b\to\infty$, i.e. the bias of a histogram estimator goes to zero as the number of bins per dimension goes to infinity. Thus there exists a sufficiently large B such that there exists a histogram $h\in\mathcal{H}_{d,B}$ which is a good approximation of p, $\|p-h\|_1<\varepsilon/2$. In this proof we we will argue that once $k\ge B^d$ and b is sufficiently large, we can find an element of $\mathcal{H}_{d,b}^k$ where the multi-view components can approximate the B^d bins of h.

We have that, for some $w\in\mathcal{T}_{d,B}$

$$
h=\sum_{A\in[B]^d} w_A h_{d,B,A}.
$$

From the same theorem in [6] there exists a_0 such that, for all $a\ge a_0$, for all i, there exists $\widetilde{h}_{1,a,i}\in\mathcal{H}_{1,a}$ such that $\left\|h_{1,B,i}-\widetilde{h}_{1,a,i}\right\|_1<\varepsilon/(2d)$ for all $i\in[B]$. For any multi-index $A\in[B]^d$, we define

$$
\widetilde{h}_{d,a,A}=\prod_{j=1}^d \widetilde{h}_{1,a,A_j}.
$$

Now we have that, for all $a\ge a_0$ and $A\in[B]^d$,

$$
\begin{aligned}
\left\|h_{d,B,A}-\widetilde{h}_{d,a,A}\right\|_1 &= \left\|\prod_{i=1}^d h_{1,B,A_i}-\prod_{j=1}^d \widetilde{h}_{1,a,A_j}\right\|_1\\
&\le \sum_{i=1}^d \left\|h_{1,B,A_i}-\widetilde{h}_{1,a,A_i}\right\|_1 \qquad (3)\\
&\le d\frac{\varepsilon}{2d}\\
&=\varepsilon/2,
\end{aligned}
$$

[1]See p. 20 in this text for the application to histograms.

where we use the previously mentioned product measure inequality for (3). As soon as $k \geq B^d$ and $a \geq a_0$ the set $\mathcal{H}_{d,a}^k$ contains the element,

$$\widetilde{h} \triangleq \sum_{A \in [B]^d} w_A \widetilde{h}_{d,a,A}.$$

Now we have that, for all $a \geq a_0$.

$$\begin{aligned}
\left\| h - \widetilde{h} \right\|_1 &= \left\| \sum_{A \in [B]^d} w_A h_{d,B,A} - \sum_{Q \in [B]^d} w_Q \widetilde{h}_{d,a,Q} \right\|_1 \\
&\leq \sum_{A \in [B]^d} w_A \left\| h_{d,B,A} - \widetilde{h}_{d,a,A} \right\|_1 \\
&\leq \varepsilon/2.
\end{aligned}$$

From the triangle inequality we have that

$$\left\| p - \widetilde{h} \right\|_1 \leq \|p - h\|_1 + \left\| h - \widetilde{h} \right\|_1 \leq \varepsilon.$$

So we have that, for sufficiently large b and k

$$\min_{q \in \mathcal{H}_{d,b}^k} \|p - q\|_1 \leq \varepsilon$$

which completes our proof. $\square$

Proof of Lemma 2.2. We will show that $\mathcal{H}_{d,b}^k \subset \widetilde{\mathcal{H}}_{d,b}^k$ and the lemma clearly follows due to Lemma 2.1. Any element of $\mathcal{H}_{d,b}^k$ will have the following representation

$$\sum_{i=1}^k w_i \prod_{j=1}^d f_{i,j} : w \in \Delta_k, f_{i,j} \in \mathcal{H}_{1,b}. \tag{4}$$

Letting $W \in \mathcal{T}_{d,k}$ with $W_{i,\ldots,i} = w_i$ for all i, the rest of the entries of W be zero, and $\widetilde{f}_{j,i} = f_{i,j}$ for all i, j we have that

$$\begin{aligned}
\sum_{S \in [k]^d} W_S \prod_{j=1}^d \widetilde{f}_{j,S_j} &= \sum_{i=1}^k W_{i,\ldots,i} \prod_{j=1}^d \widetilde{f}_{j,i} \\
&= \sum_{i=1}^k w_i \prod_{j=1}^d f_{i,j}
\end{aligned}$$

so we have that (4) is an element of $\widetilde{\mathcal{H}}_{d,b}^k$ and we are done. $\square$

Proof of Theorem 2.3. We will proceed by contradiction. Suppose V_n is an estimator violating the theorem statement, i.e. there exist sequences $b \to \infty$ and $k \to \infty$ with $n/(bk) \to 0$ and $b \geq k$ such that, for all $\varepsilon > 0$,

$$\sup_{p \in \mathcal{D}_d} P\left(\|V_n - p\|_1 > 3 \min_{q \in \mathcal{H}_{d,b}^k} \|p - q\|_1 + \varepsilon \right) \to 0.$$

Let $(p_n)_{n=1}^\infty$ be a sequence of probability vectors $p_n \in \Delta_{b(n) \times k(n)}$ which represent distributions over $[b(n)] \times [k(n)]$. Let $\mathcal{X}_n \triangleq (X_{n,1}, \ldots, X_{n,n})$ with $X_{n,1}, \ldots, X_{n,n} \overset{iid}{\sim} p_n$.

We will now construct a series of estimators for p_n using V_n. Let $\widetilde{\mathcal{X}}_n = \left(\widetilde{X}_{n,1}, \ldots, \widetilde{X}_{n,n} \right)$ which are independent random variables with $\widetilde{X}_{n,i} \sim h_{d,b,(X_{n,i},1,\ldots,1)}$, so $\widetilde{X}_{n,i}$ is uniformly distributed over the bin designated by $X_{n,i}, 1, \ldots, 1$. For this proof we will assume $d > 2$ but the proof can be simplified in a straightforward manner to the $d=2$ case by ignoring the indices and modes beyond the second. Note that that $X_{n,i}$ contains two indices. Now we have the following for the densities of $\widetilde{X}_{n,i}$

$$\begin{aligned}
p_{\widetilde{X}_{n,i}} &= \sum_{(j,\ell)\in[b]\times[k]} p_{\widetilde{X}_{n,i}|X_{n,i}=(j,\ell)} P(X_{n,i}=(j,\ell)) \\
&= \sum_{(j,\ell)\in[b]\times[k]} h_{d,b,(j,\ell,1,\ldots,1)} p_n(j,\ell) \\
&= \sum_{\ell\in[k]}\sum_{j\in[b]} h_{d,b,(j,\ell,1,\ldots,1)} p_n(j,\ell) \\
&= \sum_{\ell\in[k]}\sum_{j\in[b]} p_n(j,\ell)\, h_{1,b,j}\otimes h_{1,b,\ell}\otimes \prod_{a\in[d-2]} h_{1,b,1} \\
&= \sum_{\ell\in[k]}\left(\sum_{j\in[b]} p_n(j,\ell)\, h_{1,b,j}\right)\otimes h_{1,b,\ell}\otimes \prod_{a\in[d-2]} h_{1,b,1} \qquad (5)\\
&= \sum_{\ell\in[k]}\left(\sum_{q\in[b]} p_n(q,\ell)\right)\left(\sum_{j\in[b]} \frac{p_n(j,\ell)}{\sum_{q\in[b]} p_n(q,\ell)} h_{1,b,j}\right)\otimes h_{1,b,\ell}\otimes \prod_{a\in[d-2]} h_{1,b,1}. \qquad (6)
\end{aligned}$$

For (6) we let $0/0$ be equal to zero as is common for discrete conditioning. This last line is in the form of (4) in the main text and is thus an element of $\mathcal{H}^k_{d,b}$. To see this we will show the correspondence between the terms in (6) from here and the terms in (4) in the main text:

$$\begin{aligned}
w_\ell &:= \left(\sum_{q\in[b]} p_n(q,\ell)\right)\\
f_{\ell,1} &:= \left(\sum_{j\in[b]} \frac{p_n(j,\ell)}{\sum_{q\in[b]} p_n(q,\ell)} h_{1,b,j}\right)\\
f_{\ell,2} &:= h_{1,b,\ell}\\
f_{i,j} &:= h_{1,b,1}, \forall j>2, \forall i.
\end{aligned}$$

Let V_n estimate $\widetilde{P}_n \triangleq p_{\widetilde{X}_{n,i}}$ so $\widetilde{X}_{n,1},\ldots,\widetilde{X}_{n,n} \overset{iid}{\sim} \widetilde{P}_n$. We will use V_n to construct an estimator v_n for p_n.

Because $\widetilde{P}_n \in \mathcal{H}^k_{d,b}$ [2] for all n and our contradiction hypothesis we have that $\left\|V_n - \widetilde{P}_n\right\|_1 \xrightarrow{p} 0$. From this it follows that $\left\|U^{-1}_{d,b}(V_n) - U^{-1}_{d,b}(\widetilde{P}_n)\right\|_1 \xrightarrow{p} 0$. Note that $\left[U^{-1}_{d,b}(\widetilde{P}_n)\right]_{j,\ell,A} = p_n(j,\ell)$ when $A=(1,\ldots,1)$ and zero otherwise (see (5)). We define the linear operator $B_n : \mathcal{T}_{d,b} \to \Delta_{b\times k}$ as

$$[B_n(T)]_{j,\ell} \triangleq \sum_{A\in[b]^{d-2}} T_{j,\ell,A}$$

i.e. the linear operator which sums out all modes except for the first two. We have that $B_n(U^{-1}_{d,b}(\widetilde{P}_n)) = p_n$. Now let $v_n = B_n(U^{-1}_{d,b}(V_n))$ be the estimator for p_n. Now we have that

$$\|v_n - p_n\|_1 = \left\|B_n(U^{-1}_{d,b}(\widetilde{P}_n)) - B_n(U^{-1}_{d,b}(V_n))\right\|_1 = \left\|B_n(U^{-1}_{d,b}(\widetilde{P}_n - V_n))\right\|_1.$$

We have that B_n is an ℓ^1-nonexpansive operator due to the triangle inequality,

$$\|B_n(T)\|_1 = \sum_{j,l}\left|\sum_{A\in[b]^{d-2}} T_{j,\ell,A}\right| \le \sum_{j,l}\sum_{A\in[b]^{d-2}} |T_{j,\ell,A}| = \|T\|_1,$$

[2] We will use this portion of the proof again for our proof of Theorem 2.4

so the operator norm of B_n is less than or equal to one. We also know that $U_{d,b}^{-1}$ an isometry and $\left\|\widetilde{P}_n - V_n\right\|_1 \xrightarrow{p} 0$, so it follows that $\|v_n - p_n\|_1 \xrightarrow{p} 0$ for any sequence of $p_n \in \Delta_{[b(n)]\times[k(n)]}$. We will now use following theorem from [7] to show that no such estimator v_n can exist.

Theorem B.1 ([7] Theorem 2.)**.** *For any $\zeta \in (0,1]$, we have*

$$\inf_{\hat{p}} \sup_{p\in\Delta_a} \mathbb{E}_p \|\hat{p} - p\|_1 \geq \frac{1}{8}\sqrt{\frac{ea}{(1+\zeta)\,n}}\mathbb{1}\left(\frac{(1+\zeta)\,n}{a} > \frac{e}{16}\right)$$
$$+ \exp\left(-\frac{2\,(1+\zeta)\,n}{a}\right)\mathbb{1}\left(\frac{(1+\zeta)\,n}{a} \leq \frac{e}{16}\right) - \exp\left(-\frac{\zeta^2 n}{24}\right) - 12\exp\left(-\frac{\zeta^2 a}{32\,(\log a)^2}\right)$$

where the infimum is over all estimators.

Our estimator is equivalent to estimating a categorical distribution with $a = bk$ categories. Letting $\zeta = 1$, $bk \to \infty$, and $n \to \infty$, with $n/(bk) \to 0$, we get that for sufficiently large n

$$\inf_{\hat{p}} \sup_{p\in\Delta_{bk}} \mathbb{E}_p \|\hat{p} - p\|_1 \geq \exp\left(-\frac{4n}{bk}\right) - \exp\left(-\frac{n}{24}\right) - 12\exp\left(-\frac{bk}{32\,(\log bk)^2}\right)$$

whose right hand side converges to 1. From this we get that

$$\liminf_{n\to\infty} \sup_{p_n\in\Delta_{bk}} \mathbb{E}_{p_n} \|v_n - p_n\|_1 > \frac{1}{2}$$

which contradicts $\|v_n - p_n\|_1 \xrightarrow{p} 0$ for arbitrary sequences p_n. □

Proof of Theorem 2.4. We will proceed by contradiction. Suppose V_n is an estimator violating the theorem statement, i.e. there exist sequences $b \to \infty$ and $k \to \infty$ with $n/\left(bk + k^d\right) \to 0$ and $b \geq k$ such that, for all $\varepsilon > 0$,

$$\sup_{p\in\mathcal{D}_d} P\left(\|V_n - p\|_1 > 3 \min_{q\in\widetilde{\mathcal{H}}_{d,b}^k} \|p - q\|_1 + \varepsilon\right) \to 0.$$

Since $n/(bk + k^d) \to 0$ we have that $(bk + k^d)/n \to \infty$ so there is a subsequence n_i such that $b(n_i)k(n_i)/n_i \to \infty$ or $k(n_i)^d/n_i \to \infty$, or equivalently $n_i/(b(n_i)k(n_i)) \to 0$ or $n_i/k(n_i)^d \to 0$. We will show that both cases lead to a contradiction. We will let b and k be functions of n_i implicitly when defining limits.

Case $n_i/(bk) \to 0$**:** We proceed similarly to the proof of Theorem 2.3. Let $(p_n)_{n=1}^\infty$, $\widetilde{P}_n$, and $\mathcal{X}_n$ be defined as in the proof of Theorem 2.3. Note that $\mathcal{H}_{d,b}^k \subset \widetilde{\mathcal{H}}_{d,b}^k$ (see proof of Lemma 2.2) and thus $\widetilde{P}_n \in \widetilde{\mathcal{H}}_{d,b}^k$. We can proceed exactly as in our proof of Theorem 2.3 at footnote 2, by simply replacing $\mathcal{H}_{d,b}^k$ with $\widetilde{\mathcal{H}}_{d,b}^k$ and n with n_i which finishes this case.

Case $n_i/k^d \to 0$**:** Let $(p_n)_{n=1}^\infty$ be a sequence of elements in $\mathcal{T}_{d,k}$ which represents distributions over $[k]^d$. Let $\mathcal{X}_n \triangleq (X_{n,1},\ldots,X_{n,n})$ with $X_{n,1},\ldots,X_{n,n} \overset{iid}{\sim} p_n$. Let $\widetilde{\mathcal{X}}_n = \left(\widetilde{X}_{n,1},\ldots,\widetilde{X}_{n,n}\right)$ which are independent random variables with $\widetilde{X}_{n,i} \sim h_{d,b,X_{n,i}}$. Let $\widetilde{P}_n$ be the density for $\widetilde{X}_{n,i}$. Note that $k \leq b$. So we have that

$$\begin{aligned}
\widetilde{P}_n &= \sum_{S\in[k]^d} p_{\widetilde{X}_{n,i}|X_{n,i}=S} P(X_{n,i} = S) \\
&= \sum_{S\in[k]^d} h_{d,b,S}\, p_n(S) \\
&= \sum_{S\in[k]^d} p_n(S) \prod_{i=1}^d h_{1,b,S_i}
\end{aligned}$$

and thus $\widetilde{P}_n \in \widetilde{\mathcal{H}}_{d,b}^k$. We proceed as in Theorem 2.3 to find an estimator for elements of $\mathcal{T}_{d,k}$ which is equivalent to estimating elements of Δ^{k^d} which is impossible since $n_i/k^d \to 0$. □

C Finite Sample Theoretical Results

In this section we cover the proofs for the finite-sample results in our paper, Section 2.2.2 in the main text. This includes proofs related to estimator bias on Lipschitz continuous functions. We note that all projections in this paper are in their respective L^2 space. As noted in the main text we will be using *set projection*, $\operatorname{Proj}_S x \triangleq \arg\min_{s \in S} \|x - s\|_2$ [2]. When S is a (closed) linear subspace this is equivalent to a linear projection. Every instance of set projection in this work yields a unique minimizer. As in the main text let Lip_L be the set of L-Lipschitz continuous densities on $[0, 1]$ and let $mL = \sup_{f \in \mathrm{Lip}_L} \|f\|_2$.

C.1 Bias

Theorem 2.5 is in the main text but is not used as is in any of our other proofs and is meant simply to be illustrative of the behavior of the bias as in the main text. We will get the proof of this theorem out of the way before moving on to the core results of this portion of the proofs. Note that λ is the standard Lebesgue measure.

Proof of Theorem 2.5. From Hölder's Inequality we have that for any function $f : [0, 1]^d \to \mathbb{R}$ that

$$\|f\|_1 = \|f \cdot \mathbb{1}\|_1 \leq \|f\|_2 \|\mathbb{1}\|_2 = \|f\|_2 . \tag{7}$$

Applying this directly to the inequality from Theorem C.1 we have

$$\left\| \prod_{i=1}^d f_i - \operatorname{Proj}_{\mathcal{H}^1_{d,b}} \prod_{i=1}^d f_i \right\|_2^2 \leq m_L^{2d} - \left(m_L^2 - \frac{L^2}{12b^2} \right)^d .$$

□

We now developing the core results needed for the paper. Because of the consequence of Hölder's Inequality (7) we will focus mainly on integrated squared distance between functions.

Lemma C.1. *Let $f : [a, b] \to \mathbb{R}$ be an L-Lipschitz continuous function. Then*

$$\min_{\alpha \in \mathbb{R}} \int_a^b (\alpha - f(x))^2 \, dx \leq \frac{L^2 (b-a)^3}{12}.$$

Proof of Lemma C.1. Let $\alpha = f\left(\frac{a+b}{2}\right)$. Then we have that

$$\begin{aligned}
\int_a^b (\alpha - f(x))^2 \, dx &\leq \int_a^b \left(L \left| x - \frac{a+b}{2} \right| \right)^2 dx \\
&= L^2 \int_a^b \left(x - \frac{a+b}{2} \right)^2 dx \\
&= L^2 \int_{a - \frac{a+b}{2}}^{b - \frac{a+b}{2}} x^2 dx \\
&= L^2 \int_{\frac{a-b}{2}}^{\frac{b-a}{2}} x^2 dx \\
&= \frac{L^2}{3} \left[x^3 \right]_{\frac{a-b}{2}}^{\frac{b-a}{2}} \\
&= \frac{L^2}{3} 2 \left(\frac{b-a}{2} \right)^3 \\
&= \frac{L^2 (b-a)^3}{12}.
\end{aligned}$$

□

Lemma C.2. *Let f be an L-Lipschitz function on $[0,1]$, then*

$$\left\| f - \operatorname{Proj}_{\operatorname{span}(\mathcal{H}_{1,b})} f \right\|_2^2 \leq \frac{L^2}{12b^2}.$$

Proof of Lemma C.2. Applying Lemma C.1 we have that

$$\begin{aligned}
\left\| f - \operatorname{Proj}_{\operatorname{span}(\mathcal{H}_{1,b})} f \right\|_2^2 &= \min_{w \in \mathbb{R}^b} \left\| f - \sum_{i=1}^{b} w_i h_{1,b,i} \right\|_2^2 \\
&= \min_{w \in \mathbb{R}^b} \left\| f - \sum_{i=1}^{b} w_i b \mathbb{1}_{\left[\frac{i-1}{b}, \frac{i}{b}\right)} \right\|_2^2 \\
&= \min_{w \in \mathbb{R}^b} \int_{[0,1]} \left(f(x) - \sum_{i=1}^{b} w_i \mathbb{1}\left(\frac{i-1}{b} \leq x < \frac{i}{b} \right) \right)^2 dx \\
&= \min_{w \in \mathbb{R}^b} \int_{[0,1]} \left(\sum_{i=1}^{b} (f(x) - w_i)\, \mathbb{1}\left(\frac{i-1}{b} \leq x < \frac{i}{b} \right) \right)^2 dx \qquad (8) \\
&= \min_{w \in \mathbb{R}^b} \int_{[0,1]} \sum_{i=1}^{b} \left((f(x) - w_i)\, \mathbb{1}\left(\frac{i-1}{b} \leq x < \frac{i}{b} \right) \right)^2 dx \qquad (9) \\
&= \min_{w \in \mathbb{R}^b} \sum_{i=1}^{b} \int_{\frac{i-1}{b}}^{\frac{i}{b}} (f(x) - w_i)^2\, dx \\
&\leq b \frac{L^2}{12b^3} = \frac{L^2}{12b^2},
\end{aligned}$$

where (8) to (9) is justified since, when distributing the square, the cross terms of the form $\mathbb{1}\left(\frac{i-1}{b} \leq x < \frac{i}{b}\right) \mathbb{1}\left(\frac{j-1}{b} \leq x < \frac{j}{b}\right)$ are equal to zero when $i \neq j$. □

Lemma C.3. *Let $f_1, \ldots, f_d \in L^2([0,1])$. Then*

$$\operatorname{Proj}_{\operatorname{span}(\mathcal{H}_{d,b})} \prod_{i=1}^{d} f_i = \prod_{i=1}^{d} \operatorname{Proj}_{\operatorname{span}(\mathcal{H}_{1,b})} f_i.$$

Proof of Lemma C.3. We will be using a multi-index $A = (i_1, \ldots, i_d) \in [b]^d$ so that $\prod_{j=1}^{d} h_{1,b,i_j} = h_{d,b,A}$ and note that $h_{d,b,A} \perp h_{d,b,A'}$ for $A \neq A'$, so $h_{d,b,A}$ is an *orthogonal* basis for $\mathcal{H}_{d,b}$. We have the following

$$\begin{aligned}
\prod_{i=1}^{d} \operatorname{Proj}_{\operatorname{span}(\mathcal{H}_{1,b})} f_i &= \prod_{i=1}^{d} \sum_{j=1}^{b} \frac{h_{1,b,j}}{\|h_{1,b,j}\|_2^2} \langle h_{1,b,j}, f_i \rangle \\
&= \left(\frac{h_{1,b,1}}{\|h_{1,b,1}\|_2^2} \langle h_{1,b,1}, f_1 \rangle + \cdots + \frac{h_{1,b,b}}{\|h_{1,b,b}\|_2^2} \langle h_{1,b,b}, f_1 \rangle \right) \otimes \\
&\quad \cdots \otimes \left(\frac{h_{1,b,1}}{\|h_{1,b,1}\|_2^2} \langle h_{1,b,1}, f_d \rangle + \cdots + \frac{h_{1,b,b}}{\|h_{1,b,b}\|_2^2} \langle h_{1,b,b}, f_d \rangle \right). \qquad (10)
\end{aligned}$$

Now, distributing the terms in (10) and consolidating the subscripts into A we get that (10) is equal to

$$\begin{aligned}\sum_{A\in[b]^d}\prod_{i=1}^d \frac{h_{1,b,A_i}}{\|h_{1,b,A_i}\|_2^2}\left\langle h_{1,b,A_i}, f_i\right\rangle &= \sum_{A\in[b]^d}\frac{\prod_{i=1}^d h_{1,b,A_i}}{\prod_{i=1}^d \|h_{1,b,A_i}\|_2^2}\prod_{i=1}^d\left\langle h_{1,b,A_i}, f_i\right\rangle\\ &= \sum_{A\in[b]^d}\frac{h_{d,b,A}}{\|h_{d,b,A}\|_2^2}\left\langle h_{d,b,A}, \prod_{i=1}^d f_i\right\rangle\\ &= \operatorname{Proj}_{\operatorname{span}(\mathcal{H}_{d,b})}\prod_{i=1}^d f_i.\end{aligned}$$

□

Lemma C.4. *Let $f \in L^2\left([0,1]^d\right)$ be a probability density. Then*

$$\operatorname{Proj}_{\operatorname{span}(\mathcal{H}_{d,b})} f = \operatorname{Proj}_{\mathcal{H}_{d,b}} f.$$

Proof of Lemma C.4. Since $h_{d,b,A}\perp h_{d,b,A'}$ for $A\neq A'$ we have

$$\operatorname{Proj}_{\operatorname{span}(\mathcal{H}_{d,b})} f = \sum_{A\in[b]^d}\frac{h_{d,b,A}}{\|h_{d,b,A}\|_2^2}\left\langle h_{d,b,A}, f\right\rangle$$

and

$$\operatorname{Proj}_{\operatorname{span}(\mathcal{H}_{d,b})} f = \sum_{A\in[b]^d} w_A h_{d,b,A}$$

where

$$w_A = \frac{\left\langle h_{d,b,A}, f\right\rangle}{\|h_{d,b,A}\|_2^2}.$$

Clearly $w_A \ge 0$ so to finish we need only show that $\sum_{A\in[b]^d} w_A = 1$. To this end we have

$$\begin{aligned}\sum_{A\in[b]^d} w_A &= \sum_{A\in[b]^d}\frac{\left\langle h_{d,b,A}, f\right\rangle}{\|h_{d,b,A}\|_2^2}\\ &= \sum_{A\in[b]^d}\int_{[0,1]^d} b^d \mathbb{1}\left(x\in\Lambda_{d,b,A}\right) f(x)dx/b^d\\ &= \int_{[0,1]^d}\sum_{A\in[b]^d}\mathbb{1}\left(x\in\Lambda_{d,b,A}\right) f(x)dx\\ &= \int_{[0,1]^d} f(x)dx\\ &= 1.\end{aligned}$$

□

Corollary C.1. *Let $f_1,\dots,f_d \in L^2([0,1])$ be probability densities, then*

$$\operatorname{Proj}_{\mathcal{H}_{d,b}}\prod_{i=1}^d f_i = \prod_{i=1}^d \operatorname{Proj}_{\mathcal{H}_{1,b}} f_i = \operatorname{Proj}_{\mathcal{H}_{d,b}^1}\prod_{i=1}^d f_i.$$

Proof of Corollary C.1. The first equality follows from Lemmas C.3 and C.4. To second equality follows from the first equality, the observation that $\mathcal{H}_{d,b}^1 \subset \mathcal{H}_{d,b}$, and that Lemma C.4 implies $\prod_{i=1}^d \operatorname{Proj}_{\mathcal{H}_{1,b}} f_i \in \mathcal{H}_{d,b}^1$. □

Theorem C.1. *Let $m_L = \sup_{f \in \mathrm{Lip}_L} \|f\|_2$ and let $b^2 \geq L^2/12$ then, for any $f_1, \ldots, f_d \in \mathrm{Lip}_L$, we have that*

$$\left\| \prod_{i=1}^d f_i - \mathrm{Proj}_{\mathcal{H}^1_{d,b}} \prod_{i=1}^d f_i \right\|_2^2 \leq m_L^{2d} - \left(m_L^2 - \frac{L^2}{12b^2} \right)^d .$$

Proof of Theorem C.1. We have

$$\begin{aligned}
\left\| \prod_{i=1}^d f_i - \mathrm{Proj}_{\mathcal{H}^1_{d,b}} \prod_{i=1}^d f_i \right\|_2^2 &= \left\| \prod_{i=1}^d f_i - \mathrm{Proj}_{\mathcal{H}_{d,b}} \prod_{i=1}^d f_i \right\|_2^2 && \text{Corollary C.1} \\
&= \left\| \prod_{i=1}^d f_i \right\|_2^2 - \left\| \mathrm{Proj}_{\mathcal{H}_{d,b}} \prod_{i=1}^d f_i \right\|_2^2 && \text{(Lemma C.4 with linear projection} \quad (11) \\
&= \left\| \prod_{i=1}^d f_i \right\|_2^2 - \left\| \prod_{i=1}^d \mathrm{Proj}_{\mathcal{H}_{1,b}} f_i \right\|_2^2 && \text{Corollary C.1} \\
&= \prod_{i=1}^d \|f_i\|_2^2 - \prod_{i=1}^d \left\| \mathrm{Proj}_{\mathcal{H}_{1,b}} f_i \right\|_2^2 && \text{Def'n of tensor product norm.}
\end{aligned}$$

Noting that $\left\| \mathrm{Proj}_{\mathcal{H}_{1,b}} f_i \right\|_2^2 = \|f_i\|_2^2 - \left\| f_i - \mathrm{Proj}_{\mathcal{H}_{1,b}} f_i \right\|_2^2$ (rearrangement of the property used in (11)) we have that

$$\left\| \prod_{i=1}^d f_i - \prod_{i=1}^d \mathrm{Proj}_{\mathcal{H}_{d,b}} f_i \right\|_2^2 = \prod_{i=1}^d \|f_i\|_2^2 - \prod_{i=1}^d \left(\|f_i\|_2^2 - \left\| f_i - \mathrm{Proj}_{\mathcal{H}_{1,b}} f_i \right\|_2^2 \right). \tag{12}$$

We will now turn our attention to the subtrahend on the right hand side of the previous equation. Note that the product terms $\|f_i\|_2^2 - \left\| f_i - \mathrm{Proj}_{\mathcal{H}_{1,b}} f_i \right\|_2^2 = \left\| \mathrm{Proj}_{\mathcal{H}_{1,b}} f_i \right\|_2^2 > 0$ so $\prod_{i=1}^d \left(\|f_i\|_2^2 - \left\| f_i - \mathrm{Proj}_{\mathcal{H}_{1,b}} f_i \right\|_2^2 \right)$ is a product of positive values. From Lemmas C.2 and C.4 we have that $\left\| f_i - \mathrm{Proj}_{\mathcal{H}_{1,b}} f_i \right\|_2^2 \leq \frac{L^2}{12b^2}$. So we have

$$\|f_i\|_2^2 - \left\| f_i - \mathrm{Proj}_{\mathcal{H}_{1,b}} f_i \right\|_2^2 \geq \|f_i\|_2^2 - \frac{L^2}{12b^2}.$$

From Hölder's Inequality we have that

$$\|f \cdot \mathbb{1}\|_1 \leq \|f\|_2 \|\mathbb{1}\|_2 \Rightarrow \|f\|_2 \geq 1.$$

Combining this with the hypothesis $\frac{L^2}{12b^2} \leq 1$ we have that

$$\|f_i\|_2^2 - \left\| f_i - \mathrm{Proj}_{\mathcal{H}_{1,b}} f_i \right\|_2^2 \geq \|f_i\|_2^2 - \frac{L^2}{12b^2} \geq 0.$$

For a pair of tuples $b_i \geq a_i \geq 0$ it follows that $\prod b_i \geq \prod a_i$ and thus we have that

$$\prod_{i=1}^d \left(\|f_i\|_2^2 - \left\| f_i - \mathrm{Proj}_{\mathcal{H}_{1,b}} f_i \right\|_2^2 \right) \geq \prod_{i=1}^d \left(\|f_i\|_2^2 - \frac{L^2}{12b^2} \right).$$

Plugging this back into the RHS of (12) we get

$$\prod_{i=1}^d \|f_i\|_2^2 - \prod_{i=1}^d \left(\|f_i\|_2^2 - \left\| f_i - \mathrm{Proj}_{\mathcal{H}_{1,b}} f_i \right\|_2^2 \right) \leq \prod_{i=1}^d \|f_i\|_2^2 - \prod_{i=1}^d \left(\|f_i\|_2^2 - \frac{L^2}{12b^2} \right). \tag{13}$$

For some arbitrary j we perform the following derivative

$$\frac{d}{d\left\|f_j\right\|_2^2}\left(\prod_{i=1}^d\|f_i\|_2^2-\prod_{i=1}^d\left(\|f_i\|_2^2-\frac{L^2}{12b^2}\right)\right)=\prod_{i\neq j}\|f_i\|_2^2-\prod_{i\neq j}\left(\|f_i\|_2^2-\frac{L^2}{12b^2}\right)$$
$$\geq 0.$$

Thus we can find an upper bound to the RHS of (13) by maximizing $\|f_i\|_2^2$ over f_i, thus yielding

$$\left\|\prod_{i=1}^d f_i-\mathrm{Proj}_{\mathcal{H}_{d,b}}\prod_{i=1}^d f_i\right\|_2^2\leq m_L^{2d}-\left(m_L^2-\frac{L^2}{12b^2}\right)^d.$$

□

Calculating the value of M_L is quite involved and is done in the next subsection. However we will first present the following result which gives a more tractable bound on bias.

Proposition C.1. *Let $L\geq 2$, $b^2\geq L^2/12$, and let $f_1,\ldots,f_d$ be elements of Lip_L. Then*

$$\left\|\prod_{i=1}^d f_i-\mathrm{Proj}_{\mathcal{H}^1_{d,b}}\prod_{i=1}^d f_i\right\|_2^2\leq\frac{dL^{\frac{d+3}{2}}}{12b^2}\left[\frac{\sqrt{8}}{3}\right]^{d-1}.$$

If $0\leq L\leq 2$ instead of $L\geq 2$, then

$$\left\|\prod_{i=1}^d f_i-\mathrm{Proj}_{\mathcal{H}^1_{d,b}}\prod_{i=1}^d f_i\right\|_2^2\leq d\frac{L^2}{12b^2}\exp\left(\frac{(d-1)L^2}{12}\right).$$

Proof of Proposition C.1. For the $L\geq 2$ case, applying Theorem C.1 and Proposition C.3 gives us

$$\left\|\prod_{i=1}^d f_i-\mathrm{Proj}_{\mathcal{H}^1_{d,b}}\prod_{i=1}^d f_i\right\|_2^2\leq\left[\frac{2\sqrt{2L}}{3}\right]^d-\left[\frac{2\sqrt{2L}}{3}-\frac{L^2}{12b^2}\right]^d.$$

We will use the identity $a^n-b^n=(a-b)\sum_{i=0}^{n-1}a^ib^{n-1-i}$ with $a=\frac{2\sqrt{2L}}{3}$ and $b=\frac{2\sqrt{2L}}{3}-\frac{L^2}{12b^2}$. Note that $a>b>0$. From $a^d-b^d=(a-b)\sum_{i=0}^{d-1}a^ib^{d-1-i}$ we have

$$a^d-b^d=(a-b)\sum_{i=0}^{d-1}a^ib^{d-1-i}<(a-b)\sum_{i=0}^{d-1}a^ia^{d-1-i}=d(a-b)a^{d-1}$$

and thus

$$\left[\frac{2\sqrt{2L}}{3}\right]^d-\left[\frac{2\sqrt{2L}}{3}-\frac{L^2}{12b^2}\right]^d\leq d\frac{L^2}{12b^2}\left[\frac{2\sqrt{2L}}{3}\right]^{d-1}$$
$$=\frac{dL^{\frac{d+3}{2}}}{12b^2}\left[\frac{\sqrt{8}}{3}\right]^{d-1}.$$

The other case proceeds analogously. If $L=0$ the proposition statement is trivial, so we can assume $L>0$ and thus $a>b>0$ as before. Now we can finish with

$$\left\|\prod_{i=1}^d f_i-\mathrm{Proj}_{\mathcal{H}^1_{d,b}}\prod_{i=1}^d f_i\right\|_2^2\leq\left[\frac{L^2}{12}+1\right]^d-\left[\frac{L^2}{12}+1-\frac{L^2}{12b^2}\right]^d$$
$$\leq d\frac{L^2}{12b^2}\left[\frac{L^2}{12}+1\right]^{d-1}$$
$$\leq d\frac{L^2}{12b^2}\exp\left(\frac{(d-1)L^2}{12}\right).$$

□

We find that the rate of the convergence of this bias term doesn't depend on the Lipschitz constants nor dimension and always has order $O\left(b^{-2}\right)$. Applying a square root and applying Hölder's Inequality as we did in (7) we find that the L^1 bias convergence rate is $O(b^{-1})$, regardless of dimension.

C.1.1 Maximizing the L^2 norm of a Lipschitz density

In this section, we compute the value of m_L, the largest possible L^2 norm of an L-Lipschitz density function on $[0,1]$. We will show that if $L \leq 2$, $m_L^2 = L^2/12 + 1$, and if $L \geq 2$, $m_L^2 = 2\sqrt{2L}/3$ (i.e. Proposition 2.2). This will follow directly from the more general Proposition C.3 below.

We will need several lemmas, starting with a discretization of the problem.

For the following results we are going to use a notion of *monotonic rearrangement*. For a finite sequence of real numbers $(y_1, \ldots, y_n)$ its rearrangement $R(y) = \widetilde{y} = (\widetilde{y}_1, \ldots, \widetilde{y}_m)$ which is a reordering of y so that it is increasing, it essentially sorts y to be ascending. Interestingly this concept can also be applied to functions [8]. It is usually defined so that the function is decreasing for the functional case. Interestingly if one monotonically rearranges a Lipschitz continuous function, $f : \mathbb{R}^+ \to \mathbb{R}$ for example, then its rearrangement will also be Lipschitz continuous with a Lipschitz constant no larger than that for f (and potentially smaller). The following lemma says that the discrete equivalent of the "Lipschitz constant" of a sequence is always higher than that of its monotone reordering.

Lemma C.5. *Let $y = (y_1, \ldots, y_n)$ be a sequence of real numbers, and let $\widetilde{y} = (\widetilde{y}_1, \ldots, \widetilde{y}_n)$ be its monotone reordering so that $\widetilde{y}_1 \leq \widetilde{y}_2 \leq \cdots \leq \widetilde{y}_n$. We have*

$$\max_{i \leq n-1} |y_{i+1} - y_i| \geq \max_{i \leq n-1} \widetilde{y}_{i+1} - \widetilde{y}_i.$$

Proof of Lemma C.5. If y is constant or $n = 1$ then we are done, so we will assume y is nonconstant. Let i^* be such that $\widetilde{y}_{i^*+1} - \widetilde{y}_{i^*} = \max_{i \leq n-1} \widetilde{y}_{i+1} - \widetilde{y}_i$. Define the sets

$$\begin{aligned} A &:= \{i \in \{1, 2, \ldots, n\} : y_i \leq \widetilde{y}_{i^*}\} \\ B &:= \{i \in \{1, 2, \ldots, n\} : y_i \geq \widetilde{y}_{i^*+1}\}. \end{aligned}$$

These sets partition $[n]$ so $A \sqcup B = \{1, 2, \ldots, n\}$ and A and B are non empty. Thus there must exist $j \in [n]$ such that $j \in A$ and $j + 1 \in B$ or vice versa. For the first case we have that

$$y_{j+1} - y_j \geq \widetilde{y}_{i^*+1} - \widetilde{y}_{i^*} = \max_{i \leq n-1} \widetilde{y}_{i+1} - \widetilde{y}_i$$

and for the vice versa case we have

$$|y_{j+1} - y_j| = y_j - y_{j+1} \geq \widetilde{y}_{i^*+1} - \widetilde{y}_{i^*} = \max_{i \leq n-1} \widetilde{y}_{i+1} - \widetilde{y}_i.$$

So we have demonstrated that there exits a j such that $|y_{j+1} - y_j| \geq \max_{i \leq n-1} \widetilde{y}_{i+1} - \widetilde{y}_i$. $\square$

We are now in a position to prove the following discrete version of Proposition C.3 below.

Proposition C.2. *Let $n \in \mathbb{N}_{\geq 2}$, $U > 0, L > 0$, and let S_n denote the set of sequences $y = (y_1, \ldots, y_n)$ such that the following conditions are satisfied:*

$$y_i \geq 0 \qquad (\forall i) \tag{14}$$

$$|y_{i+1} - y_i| \leq L \qquad (\forall 1 \leq i \leq n-1) \tag{15}$$

$$\sum_{i=1}^{n} y_i = U. \tag{16}$$

A sequence $y^* = (y_1, \ldots, y_n) \in S_n$, which maximizes the quantity $\sum_{i=1}^n y_i^2$ subject to the conditions above is given by the following formulae.

If $L \leq \frac{2U}{n(n-1)}$, for all i,

$$y_i^* = (i-1)L + [U/n - (n-1)L/2]. \tag{17}$$

If $L \geq \frac{2U}{n(n-1)}$, for all i,

$$y_i^* = ((i - n + m)L + U/m - L(m+1)/2)_+, \tag{18}$$

where x_+ denotes the positive part of x and m is the smallest integer such that $m(m+1)L \geq 2U$.

Proof of Proposition C.2. First note that S_n is compact since the intersection of the closed sets denoted by (14) and (16) is compact and the set denoted by (15) is closed and thus further intersecting with that set gives us compact set. Because $y \mapsto \sum_{i=1}^n y_i^2 = \|y\|_2^2$ is continuous it must therefore attain a maximum on S_n. Let $\widetilde{\zeta} \in S_n$ be some sequence which attains the maximum. Define the sequence ζ as a monotone increasing (non-decreasing) rearrangement of $\widetilde{\zeta}$. By Lemma C.5, we have $\zeta \in S_n$ and clearly $\|\zeta\|_2^2 = \left\|\widetilde{\zeta}\right\|_2^2$ so ζ is also a maximizer.

Consider the following class of sequences.

$\mathfrak{L}$-*sequences:* A sequence $x = (x_1, \ldots, x_n)$ is an $\mathfrak{L}$-sequence if

1. $x \in S_n$.
2. x is non-decreasing.
3. For all $i < n$, $x_i = 0$ or $x_{i+1} - x_i = L$.

We are first going to show that y^* is the only element in this class.

Consider some arbitrary $\mathfrak{L}$-sequence y be and let $k = \min(i : y_i \neq 0)$. We have $y_i - y_k = L(i - k)$ for all $i \geq k$ and therefore

$$\begin{aligned} U &= \sum_{i=1}^{n} y_i \\ &= \sum_{i=k}^{n} y_k + L \sum_{i=k+1}^{n} (i-k) && (19) \\ \Rightarrow U &= (n-k+1)y_k + \frac{L(n-k)(n-k+1)}{2}. && (20) \end{aligned}$$

One can check that (20) holds when $k = n$, even though the summation in (19) is ill-posed. We will now split into two cases, corresponding to (17) and (18), to show that there is only one $\mathfrak{L}$-sequence for fixed U, L, n and that this sequence is given by y^*.

Case 1: $\frac{Ln(n-1)}{2} < U$.
If $\frac{Ln(n-1)}{2} < U$ then (20) cannot hold unless $k = 1$ because if $k \geq 2$ then $y_k \leq L$ (since $y_{k-1} = 0$) and

$$\begin{aligned} U &= (n-k+1)y_k + \frac{L(n-k)(n-k+1)}{2} \\ &\leq (n-1)L + \frac{L(n-2)(n-1)}{2} \\ &= \frac{Ln(n-1)}{2} < U, \end{aligned}$$

a contradiction.

Solving for y_k in (20) with $k = 1$ we get

$$y_k = y_1 = \frac{U}{n} - \frac{L(n-1)}{2},$$

so k and y_k are both unique, thus S_n contains only one element. We see that this coincides with the expression for y^* in (17) by noting that the derived y_1 is equal to y_1^*. The situation where $L = \frac{2U}{n(n-1)}$ in (17) will be addressed at the end of Case 2.

Case 2: $\frac{Ln(n-1)}{2} \geq U$.
For this case we need $k \geq 2$ since $y_k > 0$ and letting $k = 1$ in (20) yields

$$\begin{aligned} U &= (n-k+1)y_k + \frac{L(n-k)(n-k+1)}{2} \\ &> \frac{Ln(n-1)}{2} \geq U \end{aligned}$$

a contradiction. This implies $y_{k-1} = 0$, so $y_k \le L$. Define $m \triangleq n - k + 1$. From (20) we have

$$\begin{aligned} U &= (n-k+1)y_k + \frac{L(n-k)(n-k+1)}{2} \\ &= my_k + \frac{L(m-1)m}{2} \\ &= m\left(y_k + \frac{L(m-1)}{2}\right). \end{aligned} \tag{21}$$

Since $y_k > 0$ we have $U \ge \frac{Lm(m-1)}{2}$. Additionally since $y_k \le L$,

$$\begin{aligned} U &\le m\left(L + \frac{L(m-1)}{2}\right) \\ &= \frac{Lm(m+1)}{2}. \end{aligned}$$

It now follows that m must be the smallest integer such that $\frac{Lm(m+1)}{2} \ge U$.
Solving for y_k in (21) we have

$$y_k = \frac{U}{m} - \frac{L(m-1)}{2},$$

so both k and y_k are unique so there is only one element in S_n. Further note that

$$\begin{aligned} y_k &= \frac{U}{m} - \frac{L(m-1)}{2} \\ &= \frac{U}{m} - \frac{L(m+1)}{2} + L. \end{aligned}$$

From the definition of S_n it follows that

$$\begin{aligned} y_i &= (L(i-k) + y_k)_+ \\ &= \left(L(i-n+m-1) + \frac{U}{m} - \frac{L(m+1)}{2} + L\right)_+ \\ &= \left(L(i-n+m) + \frac{U}{m} - \frac{L(m+1)}{2}\right)_+ \end{aligned}$$

which coincides with y^* from (18).

To finish this case we will show that when $\frac{Ln(n-1)}{2} = U$ then (17) equals (18). To see this first note that $\frac{Ln(n-1)}{2} = U$ implies $m = n - 1$. Therefore (18) equals

$$\begin{aligned} y_i^* &= \left((i-n+m)L + \frac{U}{m} - \frac{L(m+1)}{2}\right)_+ \\ &= \left((i-1)L + \frac{U}{n-1} - \frac{Ln}{2}\right)_+ \\ &= \left((i-1)L + \frac{Ln}{2} - \frac{Ln}{2}\right)_+ \\ &= (i-1)L \end{aligned}$$

and accordingly (17) equals

$$\begin{aligned} y_i^* &= (i-1)L + \left[\frac{U}{n} - \frac{(n-1)L}{2}\right] \\ &= (i-1)L + \left[\frac{(n-1)L}{2} - \frac{(n-1)L}{2}\right] \\ &= (i-1)L. \end{aligned}$$

This finishes Case 2.

We now have that y^* satisfies property 1 of $\mathfrak{L}$-sequences (in particular (16) was nontrivial) as well as properties 2 and 3. This concludes the proof that y^* is the only $\mathfrak{L}$-sequence.

Now we are going to show that ζ is also a an $\mathfrak{L}$-sequence which will complete this proof, since y^* is the only $\mathfrak{L}$-sequence. For sake of contradiction assume that ζ is not an $\mathfrak{L}$-sequence. From this there must exist $i_* \in \{1, 2, \ldots, n-1\}$ such that $\zeta_{i_*+1} - \zeta_{i_*} < L$ and $\zeta_{i_*} \neq 0$. We then define a sequence t by:

$$t_i = \begin{cases} \zeta_i & \text{if} \quad i \neq i_*, i_* + 1 \\ \zeta_{i_*} - \delta & \text{if} \quad i = i_* \\ \zeta_{i_*+1} + \delta & \text{if} \quad i = i_* + 1, \end{cases}$$

where $\delta = \min\left(\frac{L - \zeta_{i_*+1} + \zeta_{i_*}}{2}, \zeta_{i_*}\right)$. Note that $0 < \delta \leq L/2$, a fact which we will be using extensively.

We will now show that $t \in S_n$.

t **satisfies** (14)**:** Note that by construction, since $\delta \leq \zeta_{i_*}$ and $t_{i_*} \geq 0$ (and clearly, $t_i \geq 0$ for $i \neq i_*$).

t **satisfies** (15)**:** The sequences t and ζ differ on only two indices so (15) holds trivially for any pair of indices not involving i_* or $i_* + 1$. We will address the remaining cases:

Case 1 $|t_{i_*+1} - t_{i_*}|$**:** We now have that

$$\begin{aligned} t_{i_*+1} - t_{i_*} &= \zeta_{i_*+1} - \zeta_{i_*} + 2\delta \\ &\leq \zeta_{i_*+1} - \zeta_{i_*} + L - \zeta_{i_*+1} + \zeta_{i_*} \\ &= L. \end{aligned}$$

Since ζ is monotonic and $\delta > 0$ we further have that $\zeta_{i_*+1} - \zeta_{i_*} + 2\delta > 0$ thereby finishing this case.

Case 2 $|t_{i_*+2} - t_{i_*+1}|$**:** We have that

$$\begin{aligned} t_{i_*+2} - t_{i_*+1} &= \zeta_{i_*+2} - \zeta_{i_*+1} - \delta \\ &\geq -\delta \\ &\geq -L/2 \end{aligned}$$

and

$$t_{i_*+2} - t_{i_*+1} = \left(\zeta_{i_*+2} - \zeta_{i_*+1}\right) - \delta \leq L$$

thereby completing this case.

Case 3: $|t_{i_*} - t_{i_*-1}|$**:** This is virtually identical to the previous case so we omit it.

Thus, we have that t satisfies condition (15).

t **satisfies** (16)**:** finally, it is clear that $t_{i_*} + t_{i_*+1} = \zeta_{i_*} + \zeta_{i_*+1}$ and thus $\sum_i t_i = \sum_i \zeta_i = U$. Hence, t satisfies (16) and we have proved the claim that $t \in S_n$.

Now, writing θ for $\frac{\zeta_{i_*+1} + \zeta_{i_*}}{2}$ and Δ for $\frac{\zeta_{i_*+1} - \zeta_{i_*}}{2}$:

$$\begin{aligned} \sum_{i=1}^{n} (t_i^2 - \zeta_i^2) &= t_{i_*+1}^2 + t_{i_*}^2 - \zeta_{i_*+1}^2 - \zeta_{i_*}^2 \\ &= (\zeta_{i_*+1} + \delta)^2 + (\zeta_{i_*} - \delta)^2 - \zeta_{i_*+1}^2 - \zeta_{i_*}^2 \\ &= (\theta + \Delta + \delta)^2 + (\theta - \Delta - \delta)^2 - (\theta + \Delta)^2 - (\theta - \Delta)^2 \\ &= 2\,\theta^2 + 2(\Delta + \delta)^2 - [2\,\theta^2 + 2\Delta^2] = 4\Delta\delta + 2\delta^2 \geq 2\delta^2 > 0, \end{aligned}$$

where at the last line, we have used the fact that since by construction, since $\zeta \neq y^*$, we have $\delta > 0$. This shows that:

$$\sum_{i=1}^{n} t_i^2 > \sum_{i=1}^{n} \zeta_i^2,$$

a contradiction. Hence we conclude that ζ must equal y^*. □

We can now proceed with the statement and proof of the continuous case.

Proposition C.3. *Let $L, U > 0$ be given and let S be the set of functions $f : [0,1] \to \mathbb{R}^+$ such that the following conditions are satisfied:*

1. *f is Lipschitz continuous with Lipschitz constant L.*
2. *$\int_0^1 f(x)dx = U$.*

Let $f_*^{U,L}$ be defined as follows

$$f_*^{U,L}(x) = \left(x - \frac{1}{2}\right) L + U \qquad \text{if } L \leq 2U \tag{22}$$

$$f_*^{U,L}(x) = L\left(x - 1 + \sqrt{\frac{2U}{L}}\right)_+ \qquad \text{if } L > 2U. \tag{23}$$

Then $f_*^{U,L} \in \arg\max_{f \in S} \int_0^1 f^2(x)dx$ and

$$\max_{f \in S} \int_0^1 f^2(x)dx = M(U,L) \triangleq \frac{W^3 L^2}{12} + \frac{U^2}{W},$$

where $W = 1$ if $L \leq 2U$ and $W = \sqrt{\frac{2U}{L}}$ if $L > 2U$. Equivalently, $M = \frac{L^2}{12} + U^2$ if $L \leq 2U$ and $M = \frac{2\sqrt{2}}{3}\sqrt{L}U^{3/2}$ if $L \geq 2U$.

Remark (Proposition C.3 applied to probability densities)**.** The highly general formation of Proposition C.3 is useful for its proof. However in density estimation only the case where f_* is a pdf and $U = 1$ is particularly meaningful. This gives

$$f_*(x) = \left(x - \frac{1}{2}\right) L \qquad \|f_*\|_2^2 = \frac{L^2}{12} + 1 \qquad \text{if } L \leq 2$$

$$f_*(x) = L\left(x - 1 + \sqrt{\frac{2}{L}}\right)_+ \qquad \|f_*\|_2^2 = \frac{2\sqrt{2}}{3}\sqrt{L} \qquad \text{if } L > 2.$$

Proof of Proposition C.3. We first introduce some notation. For $n \in \mathbb{N}$ and $U', L' \geq 0$ we will denote by $\mathcal{G}^n(U', L')$ the nonnegative sequence $(y_1, y_2, \ldots, y_n)$ maximizing $\sum_{i=1}^n y_i^2$ subject to $\sum_{i=1}^n y_i \frac{1}{n} = U'$ and $|y_{i+1} - y_i| \leq \frac{L'}{n}$ (i.e. $\mathcal{G}^n(U', L') = y^*$ from Proposition C.2 with $U \leftarrow U'n$ and $L \leftarrow \frac{L'}{n}$). We similarly write $M_n(U', L')$ for the optimal value, i.e., $M_n(U', L') = \sum_{i=1}^n (\mathcal{G}^n(U', L')_i)^2 \frac{1}{n}$. For ease of notation we set $\mathcal{G}^n \triangleq \mathcal{G}^n(U', L')$, $M_n \triangleq M_n(U', L')$, and $f_* = f_*^{U',L'}$. Later we will set $U' \leftarrow U$ and $L' \leftarrow L$ to prove the proposition statement, but it will be useful to establish some results for general U' and L'. We use the prime symbol to help avoid confusion between proving the final proposition statement and proving supporting results. Unless otherwise specified, all limits are taken as $n \to \infty$.

Let $\widetilde{\mathcal{G}}^n$ be the piecewise linear function from $[0,1]$ to $\mathbb{R}^+$ with $\widetilde{\mathcal{G}}^n(i/n) = \mathcal{G}_i^n$ for $i \in [n]$ and $\widetilde{\mathcal{G}}^n(0) = \mathcal{G}_1^n$.

Lemma C.6. *The following limits hold,*

$$\int_0^1 \widetilde{\mathcal{G}}^n(x)dx \to U'$$

$$\int_0^1 \widetilde{\mathcal{G}}^n(x)^2 dx - M_n \to 0.$$

Proof of Lemma C.6. Note that the function $\widetilde{\mathcal{G}}^n$ is L'-Lipschitz. The Lipschitz continuity also implies the following bounds on the difference between the Riemann sums below and their corresponding integrals:

$$
\begin{aligned}
\left|\int_0^1 \widetilde{\mathcal{G}}^n(x)dx - U'\right| &= \left|\int_0^1 \widetilde{\mathcal{G}}^n(x)dx - \sum_{i=1}^n \mathcal{G}_i^n \frac{1}{n}\right| \\
&\leq \sum_{i=1}^n \left|\int_{(i-1)/n}^{i/n} \widetilde{\mathcal{G}}^n(x)dx - \mathcal{G}_i^n \frac{1}{n}\right| \\
&= \sum_{i=1}^n \left|\int_{(i-1)/n}^{i/n} \widetilde{\mathcal{G}}^n(x) - \mathcal{G}_i^n dx\right| \\
&\leq \sum_{i=1}^n \int_{(i-1)/n}^{i/n} \left|\widetilde{\mathcal{G}}^n(x) - \mathcal{G}_i^n\right| dx \\
&\leq \sum_{i=1}^n \int_{(i-1)/n}^{i/n} \frac{L'}{n} dx \\
&= n\frac{L'}{n^2} \to 0,
\end{aligned}
$$

and also

$$
\begin{aligned}
\left|\int_0^1 \widetilde{\mathcal{G}}^n(x)^2dx - M_n\right| &= \left|\int_0^1 \widetilde{\mathcal{G}}^n(x)^2dx - \sum_{i=1}^n (\mathcal{G}_i^n)^2 \frac{1}{n}\right| \\
&= \left|\sum_{i=1}^n \left[\int_{(i-1)/n}^{i/n} \widetilde{\mathcal{G}}^n(x)^2 - (\mathcal{G}_i^n)^2 dx\right]\right| \\
&= \left|\sum_{i=1}^n \left[\int_{(i-1)/n}^{i/n} \left[\widetilde{\mathcal{G}}^n(x) - (\mathcal{G}_i^n)\right]\left[\widetilde{\mathcal{G}}^n(x) + (\mathcal{G}_i^n)\right] dx\right]\right| \\
&\leq n\frac{1}{n}\frac{L'}{n} \max_i \left[2\mathcal{G}_i^n\right] \\
&\leq \frac{L'}{n}\left[2\left(U' + L'\right)\right] \to 0,
\end{aligned}
$$

where at the last line we have used the fact that $\mathcal{G}_i^n \leq U' + L'$ for all i. This is because $\sum_{i=1}^n \mathcal{G}_i^n = U'n$ implies there exists an i_* such that $\mathcal{G}_{i_*}^n \leq U'$ and the Lipschitz condition further implies (for all i):

$$
\mathcal{G}_i^n \leq \mathcal{G}_{i_*}^n + \frac{L'}{n}|i - i_*| \leq U' + L'. \tag{24}
$$

$\square$

Lemma C.7. *The sequence of functions $\widetilde{\mathcal{G}}^n \to f_*$ pointwise.*

Proof of Lemma C.7. Given a continuous function $f : [0,1] \to \mathbb{R}$, that is also differentiable on $(0,1)$, for all $x \in [0,1]$ the fundamental theorem of calculus implies

$$
\begin{aligned}
&\int_0^x \partial_y f(y)dy = f(x) - f(0) \\
\Rightarrow &f(x) = \int_0^x \partial_y f(y)dy + f(0) \\
\Rightarrow &|f(x)| \leq \int_0^x |\partial_y f(y)|\, dy + |f(0)| \leq \int_0^1 |\partial_y f(y)|\, dy + |f(0)|\,.
\end{aligned}
$$

Suppose f was not differentiable, but was still continuous, at some $c \in (0, x)$. We would still have

$$\begin{aligned} f(x) &= f(x) - f(c) + f(c) \\ &= \int_c^x \partial_y f(y) dy + \int_0^c \partial_y f(y) dy + f(0) \\ \Rightarrow |f(x)| &\leq \int_0^1 |\partial_y f(y)| \, dy + |f(0)| \, . \end{aligned}$$

Both f_* and $\widetilde{\mathcal{G}}^n$ are piecewise linear. They are therefore continuous and their derivatives exist on all but a finite set of points, so for all $x \leq 1$,

$$|f_*(x) - \widetilde{\mathcal{G}}^n(x)| \leq \left| f_*(0) - \widetilde{\mathcal{G}}^n(0) \right| + \int_0^1 |\partial_y f_*(y) - \partial_y \widetilde{\mathcal{G}}^n(y)| dy. \tag{25}$$

We now go on with the proof of Lemma C.7 in two cases.

Case 1: $L' \leq 2U'$

Note that if $L' \leq 2U'$, then for all $n \geq 2$ we have that $\frac{L'}{n} \leq \frac{2U'n}{n(n-1)}$ and thus $\mathcal{G}^n$ is defined by (17). Then for any n we have $\widetilde{\mathcal{G}}^n(i/n) - \widetilde{\mathcal{G}}^n((i-1)/n) = L'/n$ for all $i \geq 2$, and $\widetilde{\mathcal{G}}^n(1/n) - \widetilde{\mathcal{G}}^n(0/n) = 0$. In particular, we have under these conditions that $\partial_x \widetilde{\mathcal{G}}^n(x) = L'$ for all $x > 1/n$ and $\partial_x \widetilde{\mathcal{G}}^n(x) = 0$ for $x < 1/n$. Hence by (25) we have for all $x \leq 1$,

$$\begin{aligned} |f_*(x) - \widetilde{\mathcal{G}}^n(x)| &\leq \left| f_*(0) - \widetilde{\mathcal{G}}^n(0) \right| + \int_0^1 |\partial_y f_*(y) - \partial_y \widetilde{\mathcal{G}}^n(y)| dy \\ &\leq \left| U' - \frac{1}{2} L' - \left[U' - \frac{n-1}{n} \frac{L'}{2} \right] \right| + \int_0^{1/n} |\partial_y f_*(y) - \partial_y \widetilde{\mathcal{G}}^n(y)| dy \\ &\leq \frac{L'}{2n} + \frac{L'}{n} \to 0. \end{aligned}$$

Case 2: $L' > 2U'$

Since $L' > 2U'$, for sufficiently large n we have that $\frac{L'}{n} > \frac{2U'n}{n(n-1)}$ and $\mathcal{G}^n$ is defined using (18). Since we are interested the limit as $n \to \infty$ we will proceed using (18). As in the proof of Proposition C.2 we let k be the smallest natural number with $\mathcal{G}_k^n > 0$, noting also that $k \geq 2$ (see Case 2 in that proof).

For all $i \leq k-1$ we have $\widetilde{\mathcal{G}}^n(i/n) - \widetilde{\mathcal{G}}^n((i-1)/n) = 0$ so $\partial_x \widetilde{\mathcal{G}}^n(x) = 0$ for all $x < (k-1)/n$. Similarly for all $i \geq k+1$ we have $\widetilde{\mathcal{G}}^n(i/n) - \widetilde{\mathcal{G}}^n((i-1)/n) = L'/n$ so $\partial_x \widetilde{\mathcal{G}}^n(x) = L'$ for all $x > (k+1)/n$. We also have that $\widetilde{\mathcal{G}}^n(0) = 0$.

By definition (23) we have $\partial_x f_*(x) = 0$ for all $x < \left[1 - \sqrt{\frac{2U'}{L'}}\right]$ and $\partial_x f_*(x) = L'$ for all $x > 1 - \sqrt{\frac{2U'}{L'}}$. We also have $f_*(0) = 0$.

Again from the proof of Proposition C.2 we know $k = n - m + 1$, where m is the smallest integer such that $m(m+1)L'/n \geq 2U'n$. It follows that $m(m+1)/n^2 \to 2U'/L'$. Since $m \to \infty$ and m^2/n^2 doesn't diverge we have that $m/n^2 \to 0$ and thus $m/n \to \sqrt{2U'/L'}$. Substituting in $k = n - m + 1$ gives $k/n \to 1 - \sqrt{2U'/L'}$.

Let $\varepsilon > 0$. For sufficiently large n both $(k+1)/n$ and $(k-1)/n$ lie in $\left[1 - \sqrt{2U'/L'} - \varepsilon, 1 - \sqrt{2U'/L'} + \varepsilon\right]$. Letting $x \leq 1$, for large enough n the following holds by

(25)

$$
\begin{aligned}
|f_*(x) - \widetilde{\mathcal{G}}^n(x)| &\le \left|f_*(0) - \widetilde{\mathcal{G}}^n(0)\right| + \int_0^1 |\partial_y f_*(y) - \partial_y \widetilde{\mathcal{G}}^n(y)| dy \\
&= \left|f_*(0) - \widetilde{\mathcal{G}}^n(0)\right| + \int_{1-\sqrt{2U'/L'}-\varepsilon}^{1-\sqrt{2U'/L'}+\varepsilon} |\partial_y f_*(y) - \partial_y \widetilde{\mathcal{G}}^n(y)| dy \\
&\le 2\varepsilon \left(\max_y |\partial_y f_*(y)| + |\partial_y \widetilde{\mathcal{G}}^n(y)| \right) \\
&\le 2\varepsilon \left(L' + L' + U' \right) \qquad \text{using (24).}
\end{aligned}
$$

Since this holds for all $\varepsilon > 0$, $|f_*(x) - \widetilde{\mathcal{G}}^n(x)|$ goes to zero, and we have pointwise convergence for Case 2. ☐

Lemma C.8. *We have that $M_n \to \int_0^1 f_*(x)^2 dx$.*

Proof of Lemma C.8. From (24) we can bound $\max_x \widetilde{\mathcal{G}}^n(x)^2 \le (U' + L')^2$. This bound, along with the pointwise convergence of Lemma C.7, allows us to apply the dominated convergence theorem, thus giving $\int_0^1 \widetilde{\mathcal{G}}^n(x)^2 dx \to \int_0^1 f_*(x)^2 dx$. Simply applying Lemma C.6 finishes our proof. ☐

We can now proceed with the proof of the proposition. First we will calculate the integral $\int_0^1 f_*(x)^2 dx$, thus establishing that $M(U', L') = \int_0^1 f_*(x)^2 dx$. To avoid confusion and to be fully precise: M is defined by the $\triangleq$ symbol in the proposition statement; the left equality in that equation being proven.

We define W' from U' and L' analogously to the way W is defined from U and L in the proposition statement. Note that f_* is linear on the interval $[1 - W', 1]$ and zero elsewhere. From Lemmas C.6 and C.7 and using the dominated convergence theorem as before we have that $\int_0^1 f_*(x) dx = U'$. Thus we can write

$$
\begin{aligned}
M(U', L') &= \int_0^1 [f_*(x)]^2 dx \\
&= \int_{1-W'}^1 [f_*(x)]^2 dx \\
&= \int_{1-W'}^1 \left[f_*(x) - \frac{U'}{W'} + \frac{U'}{W'} \right]^2 dx \\
&= \int_{1-W'}^1 \left[f_*(x) - \frac{U'}{W'} \right]^2 dx + \int_{1-W'}^1 \left[\frac{U'}{W'} \right]^2 dx + 2 \int_{1-W'}^1 \left[f_*(x) - \frac{U'}{W'} \right] \left[\frac{U'}{W'} \right] dx \\
&= \int_{1-W'}^1 \left[f_*(x) - \frac{U'}{W'} \right]^2 dx + \int_{1-W'}^1 \left[\frac{U'}{W'} \right]^2 dx,
\end{aligned}
$$

where at the last line we have used the fact that $\int_{1-W'}^1 f_*(x) - \frac{U'}{W'} dx = 0$.

Clearly $\int_{1-W'}^1 [\frac{U'}{W'}]^2 dx = \frac{U'^2}{W'}$. Further, $f_*|_{[1-W',1]} - \frac{U'}{W'}$ is linear with slope L' and $f_*(1 - W'/2) - \frac{U'}{W'} = 0$, so $f_*|_{[1-W',1]}(x) - \frac{U'}{W'}$ is antisymmetric about $x = 1 - \frac{W'}{2}$. Using these facts we can continue,

$$\begin{aligned}
\int_0^1 [f_*(x)]^2 dx &= \int_{1-W'}^1 [f_*(x)]^2 dx \\
&= \frac{U'^2}{W'} + 2\int_{1-\frac{W'}{2}}^1 \left[f_*(x) - \frac{U'}{W'}\right]^2 dx \\
&= \frac{U'^2}{W'} + 2\int_{1-\frac{W'}{2}}^1 L'^2 \left(x - 1 + \frac{W'}{2}\right)^2 dx \\
&= \frac{U'^2}{W'} + 2L'^2 \int_0^{\frac{W'}{2}} x^2 dx \\
&= \frac{U'^2}{W'} + 2L'^2 \frac{(W'/2)^3}{3} = \frac{W'^3 L'^2}{12} + \frac{U'^2}{W'},
\end{aligned}$$

as expected.

We now fix $U' \leftarrow U$ and $L' \leftarrow L$. We have shown $\int_0^1 f_*(x)dx = U$ which, along direct inspection of the definition of f_*, establishes that f_* satisfies the properties of S. Let $g \in S$ be arbitrary. We will show that $\int_0^1 g(x)^2 dx \leq \int_0^1 f_*(x)^2 dx$, which demonstrates that $f_* \in \arg\max_{f \in S} \int_0^1 f^2(x)dx$ and finishes our proof.

For all $n \geq 2$, define the sequence $g^n = (g_1^n, \ldots, g_n^n)$ with $g_i^n = g(\frac{i}{n})$. From the Lipschitz continuity of g we have as $n \to \infty$,

$$\sum_{i=1}^n g_i^n \frac{1}{n} \to \int_0^1 g(x)dx = U, \tag{26}$$

and

$$\sum_{i=1}^n (g_i^n)^2 \frac{1}{n} \to \int_0^1 g(x)^2 dx. \tag{27}$$

Let $\varepsilon > 0$ be arbitrary. By (26) we can set N such that $\sum_{i=1}^n g_i^n \frac{1}{n} \leq U + \varepsilon$ for all $n \geq N$. Note that $\left|g_{i+i}^n - g_i^n\right| \leq L/n$ and $g_i^n \geq 0$ for all i. Let $Q_n \triangleq \frac{1}{n}\left(n\varepsilon + nU - \sum_{i=1}^n g_i^n\right)$. For $n \geq N$ we have that $Q_n \geq 0$ and the following hold,

$$\begin{aligned}
g_i^n + Q_n \geq 0 \qquad & \text{for all } i \\
\left|(g_i^n + Q_n) - \left(g_{i+1}^n + Q_n\right)\right| \leq L/n \qquad & \text{for all } i \\
\frac{1}{n}\left(\sum_{i=1}^n g_i^n + Q_n\right) = U + \varepsilon. &
\end{aligned}$$

It follows that

$$\begin{aligned}
\int_0^1 g(x)^2 dx &= \lim_{n\to\infty} \sum_{i=1}^n (g_i^n)^2 \frac{1}{n} \\
&\leq \limsup_{n\to\infty} \sum_{i=1}^n \left(g_i^n + Q_n\right)^2 \frac{1}{n} \\
&\leq \lim_{n\to\infty} M_n\left(U + \varepsilon, L\right) \\
&= M\left(U + \varepsilon, L\right). \qquad \text{Lemma C.8}
\end{aligned}$$

Since $M(\cdot, L)$ is continuous and ε was arbitrary we have that $\int_0^1 g(x)^2 dx \leq M(U, L)$. □

C.1.2 Lower bound for Lipschitz densities in L^1.

Lemma C.9. *Let $f : [a, b] \to \mathbb{R}$ be the function $f(x) = Lx + c$ for some $a, b, c, L \in \mathbb{R}$ with $a < b$. Then*

$$\min_{\alpha\in\mathbb{R}} \int_a^b |\alpha - f(x)|\, dx = \frac{L\,(b-a)^2}{4}, \tag{28}$$

and

$$\arg\min_{\alpha\in\mathbb{R}} \int_a^b |\alpha - f(x)|\, dx = L\frac{a+b}{2} + c = \frac{1}{b-a}\int_a^b f(x)dx. \tag{29}$$

Proof of Lemma C.9. When $L = 0$ the result follows from just setting $\alpha = c$. We will assume that $L \neq 0$.

Suppose that $\alpha > \max(f(a), f(b))$, then we have

$$\begin{aligned}
\int_a^b |\alpha - f(x)|\, dx &= \int_a^b \alpha - f(x)dx \\
&= \int_a^b \alpha - \max\left(f(a), f(b)\right) + \max(f(a), f(b)) - f(x)dx \\
&= \int_a^b |\alpha - \max(f(a), f(b))| + |\max(f(a), f(b)) - f(x)|\, dx \\
&> \int_a^b |\max(f(a), f(b)) - f(x)|\, dx.
\end{aligned}$$

Therefore $\alpha > \max(f(a), f(b))$ cannot be the minimizer since we can simply let $\alpha = \max(f(a), f(b))$ and we have a better minimizer. So we have that $\alpha \leq \max(f(a), f(b))$ and a similar argument gives us that $\alpha \geq \min(f(a), f(b))$. Now we have that $\alpha \in \left[\min_{x\in[a,b]} f(x), \max_{x\in[a,b]} f(x)\right]$. From this and the continuity of f there exists $r \in [a, b]$ such that $f(r) = \alpha$, specifically $\alpha = Lr + c$. We now assume that $L > 0$ as the other case ($L < 0$) is analogous.

Continuing with this r we get

$$\begin{aligned}
\int_a^b |\alpha - f(x)|\, dx &= \int_a^b |Lr + c - (Lx + c)|\, dx \\
&= \int_a^b L\,|r - x|\, dx \\
&= L\int_a^r (r - x)dx + L\int_r^b (x - r)dx \\
&= L(r - a)^2/2 + L(b - r)^2/2 \\
&= L(r - a)^2/2 + L(r - b)^2/2 \\
&= \frac{L(\delta - A)^2}{2} + \frac{L(\delta + A)^2}{2} \qquad \text{letting } \delta := r - \frac{a+b}{2} \text{ and } A := \frac{a-b}{2} \\
&= L\left(A^2 + \delta^2\right) \\
&= \frac{L\,(b-a)^2}{4} + L\delta^2.
\end{aligned}$$

Upon noting that the last line is minimized for $\delta = 0$ we arrive at (28) and the first equality in (29). The second equality in (29) follows from noting

$$\int_a^b f(x)dx = \left[\frac{1}{2}Lx^2 + cx\right]_{x=a}^{x=b} = \frac{1}{2}L\left(b^2 - a^2\right) + c(b-a) = \left(L\frac{a+b}{2} + c\right)(b-a).$$

□

The following lemma addresses the point in the main text "[w]e show in the appendix that this decays at a rate of $O\left(b^{-1}\right)$ and that *this rate is tight*." We use it later to show a lower bound for the rate of convergence of the standard histogram.

Lemma C.10. *Let $b \in \mathbb{N}$ and $f_L : [0,1] \to \mathbb{R}$ be a collection of pdfs indexed by $L \in [0,\infty)$ via*

$$f_L(x) = \begin{cases} 1 + \left(x - \frac{1}{2}\right)L & L \leq 2 \\ \left(xL - L + \sqrt{2L}\right)_+ & L \geq 2. \end{cases}$$

We have that

(a) If $L \le 2$ then $\min_{g\in\mathcal{H}_{1,b}} \|f_L - g\|_1 = \frac{L}{4b}$.

(b) If $L \ge 2$ and b is a multiple of $\sqrt{L/2}$, we have $\min_{g\in\mathcal{H}_{1,b}} \|f_L - g\|_1 = \frac{\sqrt{2L}}{4b}$.

Proof of Lemma C.10. We will begin by optimizing over $\mathrm{span}\,(\mathcal{H}_{1,b})$ and will then show that the optimum lies in $\mathcal{H}_{1,b}$. To begin

$$\min_{g\in\mathrm{span}(\mathcal{H}_{1,b})} \|f_L - g\|_1 = \min_{w} \int_0^1 \left| f_L(x) - \sum_{i=1}^{b} w_i b \mathbb{1}\left((i-1)/b \le x < i/b\right)\right| dx \tag{30}$$

$$= \min_{\widetilde{w}} \sum_{i=1}^{b} \int_{(i-1)/b}^{i/b} |f_L(x) - \widetilde{w}_i|\, dx.$$

For case (b) there is a breakpoint at x satisfying $xL - L + \sqrt{2L} = 0 \iff x = 1 - \sqrt{\frac{2}{L}}$. Since b is a multiple of $\sqrt{L/2}$, there exists an integer z such that

$$\begin{aligned} b &= z\sqrt{L/2} \\ \iff z/b &= \sqrt{2/L} \\ \iff (b-z)/b &= 1 - \sqrt{2/L} \end{aligned} \tag{31}$$

and we have for both cases (a) and (b) that f_L is linear on the bins $[(i-1)/b, < i/b)$.

Applying Lemma C.9 we have the following for (a),

$$\min_{\widetilde{w}} \sum_{i=1}^{b} \int_{(i-1)/b}^{i/b} |f_L(x) - \widetilde{w}_i|\, dx = \sum_{i=1}^{b} \frac{L\left(i/b - (i-1)/b\right)^2}{4} = \frac{L}{4b}.$$

For (b) we split the summation between the breakpoint

$$\begin{aligned} &\min_{\widetilde{w}} \sum_{i=1}^{b} \int_{(i-1)/b}^{i/b} |f_L(x) - \widetilde{w}_i|\, dx \\ &= \min_{\widetilde{w}} \sum_{i\in[b]:\frac{i}{b}\le 1-\sqrt{\frac{2}{L}}} \int_{(i-1)/b}^{i/b} |f_L(x) - \widetilde{w}_i|\, dx + \sum_{i\in[b]:\frac{i}{b}> 1-\sqrt{\frac{2}{L}}} \int_{(i-1)/b}^{i/b} |f_L(x) - \widetilde{w}_i|\, dx \\ &= \sum_{i\in[b]:\frac{i}{b}\le 1-\sqrt{\frac{2}{L}}} 0 + \sum_{i\in[b]:\frac{i}{b}> 1-\sqrt{\frac{2}{L}}} \frac{L}{4b^2} \\ &= \left|\left\{ i \in [b] : \frac{i}{b} > 1 - \sqrt{\frac{2}{L}} \right\}\right| \frac{L}{4b^2}. \end{aligned} \tag{32}$$

From (31) and because $1 - \sqrt{2/L} \ge 0$ it follows that $b\left(1 - \sqrt{\frac{2}{L}}\right)$ is a nonnegative integer so

$$\begin{aligned} \left|\left\{ i \in [b] : \frac{i}{b} > 1 - \sqrt{\frac{2}{L}} \right\}\right| &= b - \left|\left\{ i \in [b] : \frac{i}{b} \le 1 - \sqrt{\frac{2}{L}} \right\}\right| \\ &= b - \left|\left\{ i \in [b] : i \le b\left(1 - \sqrt{\frac{2}{L}}\right) \right\}\right| \\ &= b - \left(b\left(1 - \sqrt{\frac{2}{L}}\right)\right) \\ &= b\sqrt{\frac{2}{L}}. \end{aligned}$$

Now (32) is equal to

$$b\sqrt{\frac{2}{L}}\frac{L}{4b^2} = \frac{\sqrt{2L}}{4b}.$$

We will show that the argument for the minimum w (with $bw = \widetilde{w}$) from (30) lies in Δ_b to finish the proof. From the second equality in (29) we clearly get that $\widetilde{w}_i \geq 0$ for all i. Again using the second equality in (29) we have

$$\begin{aligned}
\sum_{i=1}^{b} w_i = \frac{1}{b} \sum_{i=1}^{b} \widetilde{w}_i &= \frac{1}{b} \sum_{i=1}^{b} \frac{1}{i/b - (i-1)/b} \int_{(i-1)/b}^{i/b} f_L(x) dx \\
&= \sum_{i=1}^{b} \int_{(i-1)/b}^{i/b} f_L(x) dx \\
&= \int_0^1 f_L(x) dx = 1.
\end{aligned}$$

□

C.2 Finite Sample Rate: Multi-view

This section contains the finite-sample bounds from Section 2.2.2 in the main text. The results here are a bit stronger than those from the main text at the cost of being a bit less concise. In the main text we weakened the results by assuming that Lipschitz constants were greater than or equal to 2. The next section contains corresponding results for Tucker models. First we will prove the following finite-sample bound.

Proposition C.4 (Proposition 2.1 in main text)**.** *Let $d, b, k, n \in \mathbb{N}$ and $0 < \delta \leq 1$. There exists an estimator $V_n \in \mathcal{H}_{d,b}^k$ such that for all densities $p \in \mathcal{D}_d$ the following holds with probability at least $1 - \delta$*

$$\|p - V_n\|_1 \leq \min_{q \in \mathcal{H}_{d,b}^k} 3\|p - q\|_1 + 7\sqrt{\frac{2bdk \log(4bdkn)}{n}} + 7\sqrt{\frac{\log(\frac{3}{\delta})}{2n}} \tag{33}$$

where V_n is a function of $X_1, \ldots, X_n \overset{iid}{\sim} p$.

Proof of Proposition C.4. We begin by showing that (33), and therefore the proposition, holds if $n = 1$, and $d, b, k \geq 1$. Bounding the second term in the next summation with $b, d, k \geq 1$ and $n = 1$ yields the following inequality

$$\begin{aligned}
3 \min_{q \in \mathcal{H}_{d,b}^k} \|p - q\|_1 + 7\sqrt{\frac{2bdk \log(4bdkn)}{n}} + 7\sqrt{\frac{\log(\frac{3}{\delta})}{2n}} &\geq 7\sqrt{\frac{2bdk \log(4bdkn)}{n}} \\
&\geq 7\sqrt{2\log(4)} \\
&\geq 7.
\end{aligned}$$

The triangle inequality gives us $\|p - V_n\|_1 \leq 2 \leq 7$. Therefore Proposition C.5 holds for $n = 1$. We will now proceed assuming that $n \geq 2$.

Let $1 \geq \varepsilon > 0$ be arbitrary. By Corollary A.1 there exists $p_1, \ldots, p_M \in \mathcal{H}_{d,b}^k$ such that $M \leq \left(\frac{4bd}{\varepsilon}\right)^{bdk} \left(\frac{4k}{\varepsilon}\right)^k$ and for all $q \in \mathcal{H}_{d,b}^k$ there exists $i \leq M$ with $\|p_i - q\|_1 \leq \varepsilon$. Applying Lemma A.6 with the same ε gives a deterministic algorithm V_n that, given at least $\frac{\log(3M^2/\delta)}{2\varepsilon^2}$ samples from a density p, outputs an index $j \in [M]$ where, with probability at least $1 - \delta$, the following holds

$$\begin{aligned}
\left\|p_j - p\right\|_1 \leq& 3 \min_{i \in [M]} \left\|p_i - p\right\|_1 + 4\varepsilon \\
\leq& 3 \min_{q \in \mathcal{H}_{d,b}^k} \min_{i \in [M]} \left(\left\|p_i - q\right\|_1 + \left\|q - p\right\|_1\right) + 4\varepsilon \\
=& \min_{q \in \mathcal{H}_{d,b}^k} 3 \left(\min_{i \in [M]} \left\|p_i - q\right\|_1\right) + 3 \left\|q - p\right\|_1 + 4\varepsilon \\
\leq& 7\varepsilon + 3 \min_{q \in \mathcal{H}_{d,b}^k} \left\|p - q\right\|_1
\end{aligned}$$

(we are loosing $\delta/3$ from Lemma A.6 to δ for convenience).

Note that

$$\frac{\log(3M^2/\delta)}{2\varepsilon^2} = \frac{\log(M)}{\varepsilon^2} + \frac{\log(3/\delta)}{2\varepsilon^2}. \tag{34}$$

We will now bound $\log(M)$ which, because ε is positive and $\log$ is strictly increasing, will give us an upper bound on the previous term. The following follows from the fact that b, d and k are all greater than or equal to 1 and $1 \geq \varepsilon > 0$,

$$\begin{aligned} M &\leq \left(\frac{4bd}{\varepsilon}\right)^{bdk} \left(\frac{4k}{\varepsilon}\right)^{k} \\ &\leq \left(\frac{4bdk}{\varepsilon}\right)^{bdk} \left(\frac{4bdk}{\varepsilon}\right)^{bdk} \\ &= \left(\frac{4bdk}{\varepsilon}\right)^{2bdk}. \end{aligned}$$

Applying this to (34) we have

$$\frac{\log(3M^2/\delta)}{2\varepsilon^2} \leq \frac{2bdk\log(\frac{4bdk}{\varepsilon})}{\varepsilon^2} + \frac{\log(\frac{3}{\delta})}{2\varepsilon^2}. \tag{35}$$

The rest of the proof will be primarily concerned with choosing $\varepsilon \in (0,1]$ so that the RHS of (35) is less than or equal to n so the hypotheses of Lemma A.6 are satisfied. We begin by eliminating some settings where selecting V_n trivial; we will then apply Lemma A.6 for the remaining settings.

Observe that if $n < 4bdk\log(4bdkn)$, then

$$7\sqrt{\frac{2bdk\log(4bdkn)}{n}} > 2,$$

and inequality (33) holds trivially. Similarly, if $n < \log\left(\frac{3}{\delta}\right)$, then

$$7\sqrt{\frac{\log(\frac{3}{\delta})}{2n}} \geq \frac{7}{\sqrt{2}} > 2,$$

and again inequality (33) holds trivially.

Thus, we can proceed with the setting

$$\begin{aligned} n &\geq 2\max\left(2bdk\log(4bdkn), \frac{\log(\frac{3}{\delta})}{2}\right) \\ &\geq 2bdk\log(4bdkn) + \frac{\log(\frac{3}{\delta})}{2}. \end{aligned} \tag{36}$$

Defining the function $\rho(\varepsilon) := \frac{2bdk\log(\frac{4bdk}{\varepsilon})}{\varepsilon^2} + \frac{\log(\frac{3}{\delta})}{2\varepsilon^2}$, inequality (36) implies that

$$\rho(1) = \frac{2bdk\log(\frac{4bdk}{1})}{1} + \frac{\log(\frac{3}{\delta})}{2} \leq n. \tag{37}$$

Furthermore,

$$\lim_{x\to 0}\rho(x) = \infty. \tag{38}$$

Together with the mean value theorem, (37) and (38) imply that we can now pick $1 \geq \varepsilon > 0$ such that

$$\rho(\varepsilon) = \frac{2bdk\log(\frac{4bdk}{\varepsilon})}{\varepsilon^2} + \frac{\log(\frac{3}{\delta})}{2\varepsilon^2} = n. \tag{39}$$

We can now apply the estimator from Lemma A.6 to select the estimator V_n. As we have shown before the estimator in Lemma A.6 outputs a density in $\mathcal{H}_{d,b}^k$ such that

$$\|V_n - p\|_1 \leq 7\varepsilon + 3\min_{q\in\mathcal{H}_{d,b}^k}\|p-q\|_1. \tag{40}$$

By (39), we have

$$\varepsilon = \sqrt{\frac{2bdk\log(\frac{4bdk}{\varepsilon})}{n} + \frac{\log(\frac{3}{\delta})}{2n}} \geq \sqrt{\frac{1}{2n}} \geq \frac{1}{n},$$

since $0 < \delta, \varepsilon \leq 1$ and $n \geq 2$. Using this in (39), we obtain

$$\begin{aligned}\varepsilon &= \sqrt{\frac{2bdk\log(\frac{4bdk}{\varepsilon})}{n} + \frac{\log(\frac{3}{\delta})}{2n}}\\ &\leq \sqrt{\frac{2bdk\log(4bdkn)}{n} + \frac{\log(\frac{3}{\delta})}{2n}}\\ &\leq \sqrt{\frac{2bdk\log(4bdkn)}{n}} + \sqrt{\frac{\log(\frac{3}{\delta})}{2n}}.\end{aligned}$$

The result follows upon plugging this back into inequality (40). □

Now we can prove the key result from the paper.

Proposition C.5 (Theorem 2.6 in main text)**.** *Let $L \geq 2$, $0 < \delta \leq 1$ and $k, n \in \mathbb{N}$. Then there exists b and an estimator $V_n \in \mathcal{H}_{d,b}^k$ such that for any density $p \triangleq \sum_{i=1}^k w_i \prod_{j=1}^d p_{i,j}$ where $p_{i,j} \in \mathrm{Lip}_L$ and w is in the probability simplex, the following holds with probability at least $1-\delta$,*

$$\|V_n - p\|_1 \leq \frac{21dk^{1/3}L^{\frac{d+3}{12}}}{n^{\frac{1}{3}}}\sqrt{\log(3Ldkn)} + 7\sqrt{\frac{\log(\frac{3}{\delta})}{2n}} \tag{41}$$

where V_n is a function of $X_1, \ldots, X_n \overset{iid}{\sim} p$.

This also holds with "$L \leq 2$" replacing "$L \geq 2$" and the following inequality replacing (41)

$$\|V_n - p\|_1 \leq \sqrt{d}\frac{k^{1/3}}{n^{1/3}}\left[L^{1/3}\exp\left(\frac{L^2(d-1)}{24}\right) + 20\sqrt{\log(7dnk)}\right] + 7\sqrt{\frac{\log(\frac{3}{\delta})}{2n}}. \tag{42}$$

Proof of Proposition C.5. We begin with the $L \geq 2$ case and the other case will follow with some minor adjustments. From Hölder's Inequality followed by Proposition C.1, if $b^2 \geq L^2/12$ and $L \geq 2$, then for any collection $f_1, \ldots, f_d$ in Lip_L we have that

$$\left\|\prod_{j=1}^d f_j - \mathrm{Proj}_{\mathcal{H}_{d,b}^1}\prod_{j=1}^d f_j\right\|_1^2 \leq \left\|\prod_{j=1}^d f_j - \mathrm{Proj}_{\mathcal{H}_{d,b}^1}\prod_{j=1}^d f_j\right\|_2^2 \leq \frac{dL^{\frac{d+3}{2}}}{12b^2}\left[\frac{\sqrt{8}}{3}\right]^{d-1} \leq \frac{dL^{\frac{d+3}{2}}}{9b^2}.$$

Taking the square root, we have

$$\left\|\prod_{j=1}^d f_j - \mathrm{Proj}_{\mathcal{H}_{d,b}^1}\prod_{j=1}^d f_j\right\|_1 \leq \frac{\sqrt{d}L^{\frac{d+3}{4}}}{3b}.$$

Consider some density p like that from the theorem statement

$$p = \sum_{i=1}^k w_i \prod_{j=1}^d p_{i,j}.$$

Since $\sum_{i=1}^k w_i \prod_{j=1}^d \mathrm{Proj}_{\mathcal{H}_{1,b}} p_{i,j}$ is an element of $\mathcal{H}_{d,b}^k$, using Corollary C.1 gives us

$$\begin{aligned}
\min_{q\in\mathcal{H}_{d,b}^k} \|p-q\|_1 &\le \left\| \sum_{i=1}^k w_i \prod_{j=1}^d p_{i,j} - \sum_{i=1}^k w_i \prod_{j=1}^d \mathrm{Proj}_{\mathcal{H}_{1,b}} p_{i,j} \right\|_1 \\
&= \left\| \sum_{i=1}^k w_i \prod_{j=1}^d p_{i,j} - \sum_{i=1}^k w_i \mathrm{Proj}_{\mathcal{H}_{d,b}^1} \prod_{j=1}^d p_{i,j} \right\|_1 \\
&\le \sum_{i=1}^k w_i \left\| \prod_{j=1}^d p_{i,j} - \mathrm{Proj}_{\mathcal{H}_{d,b}^1} \prod_{j=1}^d p_{i,j} \right\|_1 \\
&\le \sum_{i=1}^k w_i \frac{\sqrt{d}L^{\frac{d+3}{4}}}{3b} = \frac{\sqrt{d}L^{\frac{d+3}{4}}}{3b}.
\end{aligned}$$

Invoking the estimator V_n from Proposition C.4 and using the previous bound it follows that, for any choice of k and b such that $b^2 \ge L^2/12$, there exists an estimator $V_n \in \mathcal{H}_{d,b}^k$ where for all densities p from our theorem statement, with probability at least $1-\delta$, the following holds

$$\|p - V_n\|_1 \le \frac{\sqrt{d}L^{\frac{d+3}{4}}}{b} + 7\sqrt{\frac{2bdk\log(4bdkn)}{n}} + 7\sqrt{\frac{\log(\frac{3}{\delta})}{2n}}. \tag{43}$$

We now set $b = \left\lceil n^{\frac{1}{3}} L^{\frac{d+3}{6}} k^{-1/3} \right\rceil$. Note that if $n < kL$ then (41) is trivially satisfied since $L \ge 2$ and

$$\frac{21dk^{1/3}L^{\frac{d+3}{12}}}{n^{\frac{1}{3}}} \ge \frac{21dk^{1/3}L^{\frac{1+3}{12}}}{n^{\frac{1}{3}}} \ge \frac{21dk^{1/3}L^{\frac{1}{3}}}{n^{\frac{1}{3}}} \ge 21.$$

Thus we can proceed with the assumption that $n \ge kL$ for the $L \ge 2$ case of this proof. Under this assumption, noting again that $L \ge 2$, we have the following bound on b,

$$b \ge n^{\frac{1}{3}} L^{\frac{d+3}{6}} k^{-1/3} \ge L^{1/3} L^{\frac{1+3}{6}} = L,$$

which implies $b^2 \ge \frac{L^2}{12}$. So for the remainder of the $L \ge 2$ case of this proof we will proceed with the assumption $b^2 \ge \frac{L^2}{12}$ and use the estimator from (43).

Since $n \ge kL \ge k$ and $L \ge 2$, we also have $1 \le n^{\frac{1}{3}} L^{\frac{d+3}{6}} k^{-1/3} \le b \le 2n^{\frac{1}{3}} L^{\frac{d+3}{6}} k^{-1/3}$. Thus

$$\begin{aligned}
7\sqrt{\frac{2bdk\log(4bdkn)}{n}} &\le 7\sqrt{\frac{4dk^{2/3}n^{\frac{1}{3}}L^{\frac{d+3}{6}} \log\left(8L^{\frac{d+3}{6}}dk^{2/3}n^{4/3}\right)}{n}} \\
&= 14\frac{\sqrt{d}k^{1/3}L^{\frac{d+3}{12}}}{n^{1/3}}\sqrt{\log\left(8L^{\frac{d+3}{6}}dk^{2/3}n^{4/3}\right)} \\
&\le 14\frac{\sqrt{d}k^{1/3}L^{\frac{d+3}{12}}}{n^{1/3}}\sqrt{\log\left(3^{2d}L^{2d}d^{2d}k^{2d}n^{2d}\right)} \\
&= 14\frac{\sqrt{d}k^{1/3}L^{\frac{d+3}{12}}}{n^{1/3}}\sqrt{2d\log(3Ldkn)} \\
&\le 20\frac{dk^{1/3}L^{\frac{d+3}{12}}}{n^{1/3}}\sqrt{\log(3Ldkn)}.
\end{aligned}$$

Using this with (43) and applying $n^{\frac{1}{3}} L^{\frac{d+3}{6}} k^{-1/3} \le b$ to the first summand in (43) we have

$$\begin{aligned}
\|p - V_n\|_1 &\le \frac{\sqrt{d}k^{1/3}L^{\frac{d+3}{12}}}{n^{\frac{1}{3}}} + \frac{20dk^{1/3}L^{\frac{d+3}{12}}}{n^{\frac{1}{3}}}\sqrt{\log(3Ldkn)} + 7\sqrt{\frac{\log(\frac{3}{\delta})}{2n}} \\
&\le \frac{21dk^{1/3}L^{\frac{d+3}{12}}}{n^{\frac{1}{3}}}\sqrt{\log(3Ldkn)} + 7\sqrt{\frac{\log(\frac{3}{\delta})}{2n}},
\end{aligned}$$

as expected. This finishes the $L \geq 2$ case.

For the $L \leq 2$ case, we obtain (using Proposition C.1 and taking the square root along with some simple manipulations) instead,

$$\min_{q \in \mathcal{H}_{d,b}^k} \|p - q\|_1 \leq \sqrt{d} \frac{L}{3b} \exp\left(\frac{L^2(d-1)}{24}\right),$$

and instead of (43),

$$\|p - V_n\|_1 \leq \sqrt{d} \frac{L}{b} \exp\left(\frac{L^2(d-1)}{24}\right) + 7\sqrt{\frac{2bdk \log(4bdkn)}{n}} + 7\sqrt{\frac{\log(\frac{3}{\delta})}{2n}}. \tag{44}$$

Since $L \leq 2$, we clearly have $\frac{L^2}{12} < b^2$ so we can use Proposition C.1 without issue. Set $b = \lceil n^{1/3} k^{-1/3} L^{2/3} \rceil$ (we can assume $L > 0$ and thus $b \geq 1$ since the $L = 0$ case can be solved by simply setting V_n to output the uniform distribution). If $k > n$ then (42) is trivially satisfied so we can proceed with the assumption that $k \leq n$.

Since $n \geq k$ and $n^{1/3}k^{-1/3} \geq 1$ it follows that $b \geq n^{1/3}k^{-1/3}L^{2/3}$ and $b \leq \lceil n^{1/3}k^{-1/3}2^{2/3} \rceil \leq \lceil n^{1/3}k^{-1/3}2 \rceil \leq 3n^{1/3}k^{-1/3}$. Letting $C = 7\sqrt{\frac{\log(\frac{3}{\delta})}{2n}}$ we can bound (44) as follows

$$\begin{aligned}
\|p - V_n\|_1 &\leq \sqrt{d} \frac{L}{b} \exp\left(\frac{L^2(d-1)}{24}\right) + 7\sqrt{\frac{2bdk \log(4bdkn)}{n}} + C \\
&\leq \sqrt{d} \frac{L}{b} \exp\left(\frac{L^2(d-1)}{24}\right) + 7\sqrt{\frac{2\left(3n^{1/3}k^{-1/3}\right)k}{n}} \sqrt{d \log\left(4\left(3n^{1/3}k^{-1/3}\right)dkn\right)} + C \\
&\leq \sqrt{d} \frac{L^{1/3}k^{1/3}}{n^{1/3}} \exp\left(\frac{L^2(d-1)}{24}\right) + \frac{7k^{1/3}}{n^{1/3}} \sqrt{6d \log\left(12dn^{4/3}k^{2/3}\right)} + C \\
&\leq \sqrt{d} \frac{L^{1/3}k^{1/3}}{n^{1/3}} \exp\left(\frac{L^2(d-1)}{24}\right) + \frac{7k^{1/3}}{n^{1/3}} \sqrt{6d \log\left(7^{4/3}d^{4/3}n^{4/3}k^{4/3}\right)} + C \\
&\leq \sqrt{d} \frac{L^{1/3}k^{1/3}}{n^{1/3}} \exp\left(\frac{L^2(d-1)}{24}\right) + \frac{7k^{1/3}}{n^{1/3}} \sqrt{8d \log(7dnk)} + C \\
&\leq \sqrt{d} \frac{L^{1/3}k^{1/3}}{n^{1/3}} \exp\left(\frac{L^2(d-1)}{24}\right) + \frac{20k^{1/3}\sqrt{d}}{n^{1/3}} \sqrt{\log(7dnk)} + C \\
&= \sqrt{d} \frac{k^{1/3}}{n^{1/3}} \left[L^{1/3} \exp\left(\frac{L^2(d-1)}{24}\right) + 20\sqrt{\log(7dnk)}\right] + 7\sqrt{\frac{\log(\frac{3}{\delta})}{2n}},
\end{aligned}$$

which finishes the proof for the $L \leq 2$ case. $\square$

C.3 Finite Sample Rate: Tucker

Here we present results for the Tucker decomposition that are analogous to those in the last section. The results here were omitted from the main text "[f]or brevity, and because the results are virtually direct analogues of their multi-view histogram counterparts..." We begin again with a proof of a finite-sample bound, which will then be used to prove a distribution-free bound.

Proposition C.6. *Let $d, b, k, n \in \mathbb{N}$ and $0 < \delta \leq 1$. There exists an estimator $V_n \in \widetilde{\mathcal{H}}_{d,b}^k$ such that for all densities $p \in \mathcal{D}_d$, the following holds with probability at least $1 - \delta$*

$$\|p - V_n\|_1 \leq 3 \min_{q \in \widetilde{\mathcal{H}}_{d,b}^k} \|p - q\|_1 + 7\sqrt{\frac{2bdk \log(4bdn)}{n}} + 7\sqrt{\frac{2k^d \log(4k^d n)}{n}} + 7\sqrt{\frac{\log(\frac{3}{\delta})}{2n}} \tag{45}$$

where V_n is a function of $X_1, \ldots, X_n \overset{iid}{\sim} p$.

Proof of Proposition C.6. This proof is very similar to the proof of Proposition C.4 and we will provide fewer intermediate steps when they are virtually identical to those that proof.

Similarly to the proof of Proposition C.4, we can assume $n \geq 2$. Let $1 \geq \varepsilon > 0$ be arbitrary. By Corollary A.2 there exists $p_1, \ldots, p_M \in \widetilde{\mathcal{H}}^k_{d,b}$ such that $M \leq \left(\frac{4bd}{\varepsilon}\right)^{bdk} \left(\frac{4k^d}{\varepsilon}\right)^{k^d}$ and for all $q \in \widetilde{\mathcal{H}}^k_{d,b}$ there exists $i \leq M$ with $\|p_i - q\|_1 \leq \varepsilon$.

Now, by applying Lemma A.6 with the same ε, there exists a deterministic algorithm which, for all densities p, can output an index $j \in [M]$ such that

$$\|p_j - p\|_1 \leq 7\varepsilon + 3 \min_{q \in \widetilde{\mathcal{H}}^k_{d,b}} \|p - q\|_1,$$

with probability at least $1 - \delta$ given at least N samples from the distribution, where

$$N \geq \frac{\log(3M^2/\delta)}{2\varepsilon^2} = \frac{\log(M)}{\varepsilon^2} + \frac{\log(3/\delta)}{2\varepsilon^2}.$$

Now we can bound

$$\frac{\log(3M^2/\delta)}{2\varepsilon^2} \leq \frac{2bdk \log(\frac{4bd}{\varepsilon})}{\varepsilon^2} + \frac{2k^d \log(\frac{4k^d}{\varepsilon})}{\varepsilon^2} + \frac{\log(\frac{3}{\delta})}{2\varepsilon^2}.$$

Note that:

- If $n \leq 6bdk \log(4bdn)$, then $7\sqrt{\frac{2bdk \log(4bdn)}{n}} \geq 7/\sqrt{3} \geq 2$, making (45) trivial.
- If $n \leq 6k^d \log(4k^d n)$ then $7\sqrt{\frac{2k^d \log(4k^d n)}{n}} \geq 7/\sqrt{3} \geq 2$, also making (45) trivial.
- Similarly, if $n \leq 3\log(\frac{3}{\delta})$, then $7\sqrt{\frac{\log(\frac{3}{\delta})}{2n}} \geq 2$, making (45) trivial yet again.

Thus we can assume that

$$\begin{aligned} n &\geq 3 \max\left(2bdk \log(4bdn), 2k^d \log(4k^d n), \log\left(\frac{3}{\delta}\right)\right) \\ &\geq 2bdk \log(4bdn) + 2k^d \log(4k^d n) + \log\left(\frac{3}{\delta}\right). \end{aligned} \tag{46}$$

We now define $\rho(\varepsilon)$ as

$$\rho(\varepsilon) := \frac{2bdk \log(\frac{4bd}{\varepsilon})}{\varepsilon^2} + \frac{2k^d \log(\frac{4k^d}{\varepsilon})}{\varepsilon^2} + \frac{\log(\frac{3}{\delta})}{2\varepsilon^2}. \tag{47}$$

So (46) gives us $\rho(1) \leq n$ and, as before, $\lim_{x \to 0} \rho(x) = \infty$. Together with the mean value theorem it follows we can now pick $1 \geq \varepsilon > 0$ such that

$$\rho(\varepsilon) = \frac{2bdk \log(\frac{4bd}{\varepsilon})}{\varepsilon^2} + \frac{2k^d \log(\frac{4k^d}{\varepsilon})}{\varepsilon^2} + \frac{\log(\frac{3}{\delta})}{2\varepsilon^2} = n. \tag{48}$$

We can now apply the estimator from Lemma A.6 to select the estimator V_n. As we have shown before the estimator in Lemma A.6 outputs a density in $\mathcal{H}^k_{d,b}$ such that

$$\|V_n - p\|_1 \leq 7\varepsilon + 3 \min_{q \in \widetilde{\mathcal{H}}^k_{d,b}} \|p - q\|_1. \tag{49}$$

Now, note that by (48), we have

$$\varepsilon = \sqrt{\frac{2bdk \log(\frac{4bd}{\varepsilon})}{n} + \frac{2k^d \log(\frac{4k^d}{\varepsilon})}{n} + \frac{\log(\frac{3}{\delta})}{2n}} \geq \sqrt{\frac{1}{2n}} \geq \frac{1}{n},$$

since $0 < \delta, \varepsilon \leq 1$ and $n \geq 2$. Plugging this back into (48), we obtain

$$\begin{aligned} \varepsilon &\leq \sqrt{\frac{2bdk \log(4bdn)}{n} + \frac{2dk^d \log(4k^d n)}{n} + \frac{\log(\frac{3}{\delta})}{2n}} \\ &\leq \sqrt{\frac{2bdk \log(4bdn)}{n}} + \sqrt{\frac{2k^d \log(4k^d n)}{n}} + \sqrt{\frac{\log(\frac{3}{\delta})}{2n}}. \end{aligned}$$

The result follows upon plugging this back into inequality (49). $\square$

Now we can prove our distribution-free bound.

Proposition C.7. *Let $L \geq 2$, $0 < \delta \leq 1$ and $k, n \in \mathbb{N}$. Then there exists $b \in \mathbb{N}$ and an estimator $V_n \in \widetilde{\mathcal{H}}^k_{d,b}$ such that for any density $p \triangleq \sum_{S\in[k]^d} W_S \prod_{i=1}^d p_{i,S_i}$ where $p_{i,j} \in \mathrm{Lip}_L$ and W is a probability tensor, the following holds with probability at least $1-\delta$,*

$$\|p - V_n\|_1 \leq \frac{21dk^{1/3}L^{\frac{d+3}{12}}}{n^{\frac{1}{3}}}\sqrt{\log(3Ldkn)} + 7\sqrt{\frac{2k^d\log(4k^dn)}{n}} + 7\sqrt{\frac{\log(\frac{3}{\delta})}{2n}} \tag{50}$$

where V_n is a function of $X_1, \ldots, X_n \overset{iid}{\sim} p$.

This also holds with "$L \leq 2$" replacing "$L \geq 2$" and the following inequality replacing (50)

$$\|p-V_n\|_1 \leq \sqrt{d}\frac{k^{1/3}}{n^{1/3}}\left[L^{1/3}\exp\left(\frac{L^2(d-1)}{24}\right) + 20\sqrt{\log(7dnk)}\right] + 7\sqrt{\frac{2k^d\log(4k^dn)}{n}} + 7\sqrt{\frac{\log(\frac{3}{\delta})}{2n}}.$$

Proof of Proposition C.7. This proof is very similar to the proof of Proposition C.5 and we will provide fewer intermediate steps when they are virtually identical to those that proof. We begin with the $L \geq 2$ case. Consider some density p like that from the theorem statement

$$p = \sum_{S\in[k]^d} W_S \prod_{i=1}^d p_{i,S_i}.$$

Since $\sum_{S\in[k]^d} W_S \prod_{i=1}^d \mathrm{Proj}_{\mathcal{H}_{1,b}} p_{i,S_i}$ is an element of $\widetilde{\mathcal{H}}^k_{d,b}$, using Corollary C.1, Hölder's Inequality, and Proposition C.1 yields the following when $b^2 \geq L^2/12$

$$\begin{aligned}
\min_{q\in\widetilde{\mathcal{H}}^k_{d,b}} \|p-q\|_1 &\leq \left\| \sum_{S\in[k]^d} W_S \prod_{i=1}^d p_{i,S_i} - \sum_{S\in[k]^d} W_S \prod_{i=1}^d \mathrm{Proj}_{\mathcal{H}_{1,b}} p_{i,S_i} \right\|_1 \\
&= \left\| \sum_{S\in[k]^d} W_S \prod_{i=1}^d p_{i,S_i} - \sum_{S\in[k]^d} W_S \mathrm{Proj}_{\mathcal{H}^1_{d,b}} \prod_{i=1}^d p_{i,S_i} \right\|_1 \\
&\leq \sum_{S\in[k]^d} W_S \left\| \prod_{i=1}^d p_{i,S_i} - \mathrm{Proj}_{\mathcal{H}^1_{d,b}} \prod_{i=1}^d p_{i,S_i} \right\|_1 \\
&\leq \sum_{S\in[k]^d} W_S \frac{\sqrt{d}L^{\frac{d+3}{4}}}{3b} = \frac{\sqrt{d}L^{\frac{d+3}{4}}}{3b}.
\end{aligned} \tag{51}$$

Combining this with the estimator from Proposition C.6 we get that, for any b such that $b^2 \geq L^2/12$, we have

$$\|p - V_n\|_1 \leq \frac{\sqrt{d}L^{\frac{d+3}{4}}}{b} + 7\sqrt{\frac{2bdk\log(4bdn)}{n}} + 7\sqrt{\frac{2k^d\log(4k^dn)}{n}} + 7\sqrt{\frac{\log(\frac{3}{\delta})}{2n}}.$$

If $n < kL$, the RHS of (50) is greater than 2, which means that (50) holds trivially. Thus we assume $n \geq kL$. Since b doesn't appear in the third summand in the previous inequality, and the first, second, and fourth summand are exactly the same as those from (43) in the proof Proposition C.5, we can set $b = \left\lceil n^{\frac{1}{3}}L^{\frac{d+3}{6}}k^{-1/3}\right\rceil$ and proceed exactly as we did in that proof (again $b^2 \geq L^2/12$ so Proposition C.1 holds). We then obtain

$$\|p - V_n\|_1 \leq \frac{21dk^{1/3}L^{\frac{d+3}{12}}}{n^{\frac{1}{3}}}\sqrt{\log(3Ldkn)} + 7\sqrt{\frac{2k^d\log(4k^dn)}{n}} + 7\sqrt{\frac{\log(\frac{3}{\delta})}{2n}},$$

as expected.

For the $L \leq 2$ case, (51) becomes

$$\min_{q\in\widetilde{\mathcal{H}}^k_{d,b}} \|p-q\|_1 \leq \sqrt{d}\frac{L}{3b}\exp\left(\frac{L^2(d-1)}{24}\right),$$

from which we get

$$\|p - V_n\|_1 \leq \sqrt{d}\frac{L}{b}\exp\left(\frac{L^2(d-1)}{24}\right) + 7\sqrt{\frac{2bdk\log(4bdn)}{n}} + 7\sqrt{\frac{2k^d\log(4k^d n)}{n}} + 7\sqrt{\frac{\log(\frac{3}{\delta})}{2n}}.$$

Since b doesn't appear in the third summand in the previous inequality, and the rest of the inequality is exactly the same as in the proof of Proposition C.5, we can again let $b = \lceil n^{1/3}k^{-1/3}L^{2/3}\rceil$, assume $n \geq k$ and, proceed identically as we did in the proof of Proposition C.5. This yields

$$\|p - V_n\|_1 \leq \sqrt{d}\frac{k^{1/3}}{n^{1/3}}\left[L^{1/3}\exp\left(\frac{L^2(d-1)}{24}\right) + 20\sqrt{\log(7dnk)}\right] + 7\sqrt{\frac{2k^d\log(4k^d n)}{n}} + 7\sqrt{\frac{\log(\frac{3}{\delta})}{2n}},$$

as expected. $\square$

C.4 Lower Bound: Standard Histogram

Proof of Proposition 2.3. Let $p \in \mathcal{D}_d$ with $p = \prod_{i=1}^d p_i$ where p_1 is the density from Lemma C.10 for $L = 2$, i.e. $p_1(x) = 2x$, and $p_i(x) \equiv 1$ for $i > 1$. Let $(Y_1, \ldots, Y_d) \sim V_n$ and $(X_1, \ldots, X_d) \sim p$. Because total variation distance is never increased through mappings of the random variables (see Theorem 5.2 in [3]) we have that $\|V_n - p\|_1 \geq \|f - p_1\|_1$ where f is the probability density associated with Y_1 . We will now show that f is an element of $\mathcal{H}_{1,b}$. Let $S \subset [0,1]$ be an arbitrary (Borel) measurable set and note that V_n has the form $\sum_{A \in [b]^d} \hat{w}_A h_{d,b,A}$. Then we have that

$$\begin{aligned}
P(Y \in S) &= \int_S f d\lambda \\
&= \int_{S \times [0,1] \times \cdots \times [0,1]} V_n d\lambda \\
&= \int_{S \times [0,1] \times \cdots \times [0,1]} \sum_{A \in [b]^d} \hat{w}_A h_{d,b,A} d\lambda \\
&= \sum_{A \in [b]^d} \hat{w}_A \int_{S \times [0,1] \times \cdots \times [0,1]} h_{d,b,A} d\lambda \\
&= \sum_{A \in [b]^d} \hat{w}_A \int_{S \times [0,1] \times \cdots \times [0,1]} \prod_{i=1}^d h_{1,b,A_i} d\lambda \\
&= \sum_{A \in [b]^d} \hat{w}_A \left(\int_S h_{1,b,A_1} d\lambda \right) \left(\int_{[0,1]} h_{1,b,A_2} d\lambda \right) \cdots \left(\int_{[0,1]} h_{1,b,A_d} d\lambda \right) \\
&= \sum_{A \in [b]^d} \hat{w}_A \int_S h_{1,b,A_1} d\lambda \\
&= \int_S \sum_{A \in [b]^d} \hat{w}_A h_{1,b,A_1} d\lambda.
\end{aligned}$$

note that $\sum_{A \in [b]^d} \hat{w}_A h_{1,b,A_1}$ is a histogram and thus the density associated with f is a histogram. Using this fact with the earlier mentioned inequality we have that

$$\begin{aligned}
\|V_n - p\|_1 &\geq \|f - p_1\|_1 \\
&\geq \min_{h \in \mathcal{H}_{1,b}} \|h - p_1\|_1 .
\end{aligned}$$

From Lemma C.10 it then follows that

$$\|V_n - p\|_1 \geq \frac{1}{2b}.$$

Let $D > 0$ be arbitrary. From our assumption that $n/b^d \to \infty$ it follows that, for sufficiently large n, that $n/b^d \geq (2D)^d$ and furthermore

$$n/b^d \geq (2D)^d \iff \sqrt[d]{n}/b \geq 2D \iff 1/(2b) \geq D/\sqrt[d]{n} \Rightarrow \|V_n - p\|_1 \geq D/\sqrt[d]{n}$$

so $\|V_n - p\|_1 \in \omega(1/\sqrt[d]{n})$ by definition. □

D Experimental Setting

Consider the problem of finding some density estimator $\hat{p}$ with minimal L_2 distance to an unknown density p (p is the various projections of MNIST and Diabetes from the main text). This is equivalent to minimizing the squared L^2 loss:

$$\begin{aligned}
&\int_{[0,1]^d} (p(x) - \hat{p}(x))^2 \, dx \\
&= \int_{[0,1]^d} \hat{p}(x)^2 \, dx - 2 \int_{[0,1]^d} p(y)\hat{p}(y) dy + \int_{[0,1]^d} p(z)^2 dz. \qquad (52)
\end{aligned}$$

Because the right term in (52) does not depend on $\hat{p}$ it can be ignored when finding optimal $\hat{p}$. The left term in (52) is known. The middle term in (52) can be estimated with the following approximation

$$\int_{[0,1]^d} p(x)\hat{p}(x) dx = \mathbb{E}_{X \sim p}[\hat{p}(X)] \approx \frac{1}{n} \sum_{i=1}^{n} \hat{p}(X_i)$$

where $\mathcal{X} = X_1, \ldots, X_n \overset{iid}{\sim} p$. We can use this to find a good estimate $\hat{H} \in \mathcal{R}^k_{d,b}$ for p which represents $\mathcal{H}^k_{d,b}$ or $\widetilde{\mathcal{H}}^k_{d,b}$:

$$\begin{aligned}
&\arg\min_{\hat{H} \in \mathcal{R}^k_{d,b}} \int_{[0,1]^d} \left(\hat{H}(x) - \hat{p}(x)\right)^2 dx = \arg\min_{\hat{H} \in \mathcal{R}^k_{d,b}} \left\langle \hat{H}, \hat{H} \right\rangle - 2 \int_{[0,1]^d} \hat{H}(x) p(x) dx \\
&\approx \arg\min_{\hat{H} \in \mathcal{R}^k_{d,b}} \left\langle \hat{H}, \hat{H} \right\rangle - 2 \frac{1}{n} \sum_{i=1}^{n} \hat{H}(X_i). \qquad (53)
\end{aligned}$$

Recall that the standard histogram estimator is $H = H_{d,b}(\mathcal{X}) = \frac{1}{n} \sum_{i=1}^{n} \sum_{A \in [b]^d} h_{d,b,A} \mathbb{1}(X_i \in \Lambda_{d,b,A})$ and let $\hat{H} = \sum_{A \in [b]^d} \hat{w}_A h_{d,b,A}$. We have the following

$$\begin{aligned}
\left\langle \hat{H}, H \right\rangle &= \left\langle \sum_{A \in [b]^d} \hat{w}_A h_{d,b,A}, \frac{1}{n} \sum_{i=1}^{n} \sum_{B \in [b]^d} h_{d,b,B} \mathbb{1}(X_i \in \Lambda_{d,b,B}) \right\rangle \\
&= \frac{1}{n} \sum_{i=1}^{n} \sum_{A \in [b]^d} \hat{w}_A \mathbb{1}(X_i \in \Lambda_B) b^d = \frac{1}{n} \sum_{i=1}^{n} \hat{H}(X_i).
\end{aligned}$$

As a consequence (53) is equal to

$$\begin{aligned}
\arg\min_{\hat{H} \in \mathcal{R}^k_{d,b}} \left\langle \hat{H}, \hat{H} \right\rangle - 2 \left\langle H, \hat{H} \right\rangle &= \arg\min_{\hat{H} \in \mathcal{R}^k_{d,b}} \left\langle \hat{H}, \hat{H} \right\rangle - 2 \left\langle H, \hat{H} \right\rangle + \langle H, H \rangle \\
&= \arg\min_{\hat{H} \in \mathcal{R}^k_{d,b}} \left\| H - \hat{H} \right\|_2^2.
\end{aligned}$$

Using the $U_{d,b}$ operator we can reformulate this into a tensor factorization problem

$$\min_{\hat{T} \in \mathcal{Q}^k_{d,b}} \left\| H - U_{d,b}(\hat{T}) \right\|_2^2 = \min_{\hat{T} \in \mathcal{Q}^k_{d,b}} b^d \left\| U^{-1}_{d,b}(H) - \hat{T} \right\|_2^2$$

where $\mathcal{Q}^k_{d,b}$ could be either $\mathcal{T}^k_{d,b}$ or $\widetilde{\mathcal{T}}^k_{d,b}$. Because of this equivalence, to find estimates in $\mathcal{H}^k_{d,b}$ or $\widetilde{\mathcal{H}}^k_{d,b}$ we can simply use nonnegative tensor decomposition algorithms, which minimize ℓ^2 loss, to find NNTF histograms that approximate H.

E Nonexistence of Infinite Tensor Decomposition

Let $p : \mathbb{R}^2 \to \mathbb{R}$ be a probability density function. Consider the possibility of decomposing p as follows

$$p(x, y) = \sum_{i=1}^{\infty} w_i f_i(x) g_i(y) \tag{54}$$

where, for all i, $w_i \geq 0$ and f_i and g_i are probability densities. We are going to show that this is not always possible, which we will do by contradiction. Let λ be the Lebesgue measure (dimensionality will be left implicit). We are going to use the following proposition which we will prove later.

Proposition E.1. *There exists a set $E \subset [0,1] \times [0,1]$ such that $\lambda(E) > 0$ and for all non-null measurable sets $A, B \subset [0,1]$ we have that $\lambda(E \cap A \times B) < \lambda(A \times B)$.*

Let E be a set satisfying the property in Proposition E.1. Let $\mathbb{1}_S$ be the indicator function a set S. We will let $p = \mathbb{1}_E$ and assume that p has a decomposition as in (54).

We will assume that $w_1 > 0$. Clearly we have that $p - w_1 f_1 g_1$ is an almost everywhere (a.e.) nonnegative function (all products of functions here are outer products). Let $\varepsilon > 0$ such that $A \triangleq f_1^{-1}([\varepsilon, \infty))$ and $B \triangleq g_1^{-1}([\varepsilon, \infty))$ have positive measure. Such an ε must exist otherwise f_1 and g_1 are 0 a.e.. Now we have that $\varepsilon^2 \mathbb{1}_A \mathbb{1}_B \leq f_1 g_1$. And thus $p - w_1 \varepsilon^2 \mathbb{1}_A \mathbb{1}_B \geq 0$ a.e or equivalently $\lambda(E)^{-1} \mathbb{1}_E - w_1 \varepsilon^2 \mathbb{1}_{A \times B} \geq 0$ a.e.. From our definition of E we know that $\lambda(A \times B \setminus E) = \lambda(A \times B) - \lambda(E \cap A \times B) > 0$ so $\lambda(E)^{-1} \mathbb{1}_E - w_1 \varepsilon^2 \mathbb{1}_{A \times B}$ is negative on a set of positive measure, a contradiction.

We will now address the existence of the set E. The most direct statement of the existence of such an E can be found in [9], the following is the exact statement from the text.

Theorem E.1 ([9] Theorem 2.1)**.** *There exist Borel measurable subsets $E \subset [0,1]^2$ of positive measure which are rectangle free, so that if $A \times B \subseteq E$ then* $\operatorname{area}(A \times B) = 0$.

That paper builds the set E via a random construction and contains an image which showing an example that approximates a randomly sampled E. Their construction seems to imply that the condition "$A \times B \subseteq E$" was intended to be interpreted measure theoretically, i.e. "$\operatorname{area}(A \times B \setminus E) = 0$"; it is not particularly difficult to construct a measurable subset of $[0,1] \times [0,1]$ which contains all but a null set of $[0,1] \times [0,1]$ and is "rectangle free" as described in the theorem statement (see [4] and references therein). If the measure theoretic strengthening is true it would imply the existence of the set E from Proposition E.1 above. Since we are not *totally* certain that this strengthening is possible we include a proof of the existence of E above.

For a topological space (Ω, τ) equipped with a Borel measure μ, a set $S \subseteq \Omega$ is called *essentially dense* if, for any nonempty open set I, $\mu(I \cap S) > 0$. For any measurable set in $\mathbb{R}^d$ we will equip it with the standard subspace topology and measure induced by the standard Lebesgue measure. There exists a measurable set $D \subset \mathbb{R}$ such that D and D^C are essentially dense (see [5] 134J(a)). The following is a simplification of Theorem 1 in [4] that we will use to construct E.

Theorem E.2 ([4] Theorem 1)**.** *For a measurable set of the form $E = \{(x, y) : x - y \in D\}$ the following two conditions are equivalent*

1. *D is essentially dense on $\mathbb{R}$.*
2. *$\lambda(E \cap A \times B) > 0$ for all A, B such that $\lambda(A)\lambda(B) > 0$.*

From this we have that $E \triangleq \{(x, y) : x - y \in D\} \cap [0,1]^2$ and $E^C = \left\{(x, y) : x - y \in D^C\right\} \cap [0,1]^2$ (we let E live in $[0,1]^2$) have that property that for non null sets $A, B \subset [0,1]$, $\lambda(E \cap A \times B) > 0$, $\lambda(E^C \cap A \times B) > 0$. Note that $\lambda(E \cap A \times B) + \lambda(E^C \cap A \times B) = \lambda(A \times B)$ and thus $\lambda(E \cap A \times B) < \lambda(A \times B)$ so we have constructed E.

We mention that the E we have constructed contains a non-null rectangle when rotated by 45 degrees. Thusly rotating p to give $\widetilde{p}$ allows us to find f, g and $w > 0$ such that $\widetilde{p} - wfg$ is a.e. nonnegative. In [4] the authors discuss the existence of sets E where, for all non null A, B, we have that $\lambda(f(E) \cap A \times B) > 0$, for all f in certain classes of transforms. These results hint towards research directions of finding transforms so that our data is better approximated by nice NNTF model.

References

[1] Hassan Ashtiani, Shai Ben-David, Nicholas Harvey, Christopher Liaw, Abbas Mehrabian, and Yaniv Plan. Nearly tight sample complexity bounds for learning mixtures of gaussians via sample compression schemes. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems 31*, pages 3412–3421. Curran Associates, Inc., 2018.

[2] H.H. Bauschke and P.L. Combettes. *Convex analysis and monotone operator theory in Hilbert spaces*. CMS Books in Mathematics, Ouvrages de mathématiques de la SMC. Springer New York, 2011.

[3] L. Devroye and G. Lugosi. *Combinatorial Methods in Density Estimation*. Springer, New York, 2001.

[4] P. Erdős and John C. Oxtoby. Partitions of the plane into sets having positive measure in every non-null measurable product set. *Transactions of the American Mathematical Society*, 79(1):91–102, 1955.

[5] D.H. Fremlin. *Measure Theory*. Number v. 1 in Measure theory. Torres Fremlin, 2000.

[6] Làszlò Györfi, Luc Devroye, and Laszlo Gyorfi. *Nonparametric density estimation: the L1 view*. John Wiley & Sons, New York; Chichester, 1985.

[7] Yanjun Han, Jiantao Jiao, and Tsachy Weissman. Minimax estimation of discrete distributions under ℓ_1 loss. *CoRR*, abs/1411.1467, 2014.

[8] Bernhard Kawohl. *Rearrangements*, pages 7–99. Springer Berlin Heidelberg, Berlin, Heidelberg, 1985.

[9] Wilfrid S. Kendall and Giovanni Montana. Small sets and markov transition densities. *Stochastic Processes and their Applications*, 99(2):177 – 194, 2002.

[10] R.D. Reiss. *Approximate distributions of order statistics: with applications to nonparametric statistics*. Springer series in statistics. Springer, 1989.