THE LOCAL GEOMETRY OF FINITE MIXTURES

ELISABETH GASSIAT AND RAMON VAN HANDEL

Abstract. We introduce a technique to obtain local (bracketing) metric entropy bounds for subsets of a normed vector space from global entropy bounds. Using this method, we establish that for $q \geq 1$, the class of convex combinations of $q$ translates of a probability density has finite local doubling dimension under a smoothness assumption. The proof requires a detailed investigation of the local geometry of mixture classes, which is of independent interest.

1. Introduction

Let $(X, d)$ be a metric space, and consider a subset $T = \{ t_\xi : \xi \in \Xi \}$ of $X$ that is parametrized by a bounded subset $\Xi$ of $\mathbb{R}^d$. Roughly speaking, we are interested in the following question: can $T$ be viewed as a finite-dimensional subset of $X$? It is certainly tempting to think so, as the parameter set $\Xi$ is finite-dimensional, and this idea is easily made precise if the induced metric $d_T(\xi, \xi') = d(t_\xi, t_{\xi'})$ on $\Xi$ is comparable to a norm on $\mathbb{R}^d$. However, there are natural examples where control by a norm is not straightforward, or even impossible. The aim of this paper is to develop a general method to address such problems, and to study in detail a prototypical problem that arises from applications in statistics.

To set the stage for the problems that we will consider, let us recall some metric notions of dimension. For a subset $T$ of a metric space $(X, d)$, the covering number $N(T, \varepsilon)$ is the smallest cardinality of a covering of $T$ by $\varepsilon$-balls [14]:

$$N(T, \varepsilon) = \inf \left\{ n : \exists x_i \in X, i = 1, \ldots, n \text{ s.t. } T \subseteq \bigcup_{i=1}^n B(x_i, \varepsilon) \right\},$$

where $B(x, \varepsilon) = \{ x' \in X : d(x, x') \leq \varepsilon \}$. The covering number, or equivalently the metric entropy $\log N(T, \varepsilon)$, quantifies the capacity of the set $T$, and its scaling in $\varepsilon$ is closely connected to dimension. Indeed, let $|\cdot|$ be a norm on $\mathbb{R}^d$, so that $(\mathbb{R}^d, |\cdot|)$ is a finite-dimensional Banach space. A standard estimate [18, Lemma 4.14] gives

$$N(B(t, \delta), \varepsilon) \leq \left( \frac{3\delta}{\varepsilon} \right)^d$$

for any $\varepsilon \leq \delta$, where $B(t, \delta) = \{ x \in \mathbb{R}^d : |x - t| \leq \delta \}$. This estimate has two trivial consequences: first, for any bounded $T \subset (\mathbb{R}^d, |\cdot|)$, there is a constant $C_1$ so that

$$N(T, \varepsilon) \leq \left( \frac{C_1}{\varepsilon} \right)^d$$

2010 Mathematics Subject Classification. Primary 41A46; Secondary 60F15.

Key words and phrases. local metric entropy; bracketing numbers; finite mixtures.
for all $\varepsilon$ sufficiently small. On the other hand, if we fix a distinguished point $t_0 \in T$, there is a constant $C_2$ such that for all $\varepsilon/\delta$ sufficiently small

\begin{equation}
N(T \cap B(t_0, \delta), \varepsilon) \leq \left( \frac{C_2 \delta}{\varepsilon} \right)^d.
\end{equation}

Either (1.1) or (1.2) may be used as a notion of finite-dimensionality for a set $T$ in a general metric space $(X, d)$: a set satisfying the global entropy bound (1.1) has finite Kolmogorov dimension $\log N(T, \varepsilon)/\log(1/\varepsilon) \lesssim d$, while a set satisfying the local entropy bound (1.2) has finite local doubling dimension $\log N(T \cap B(t_0, 2\varepsilon), \varepsilon) \lesssim d$. Clearly (1.2) implies (1.1), but not conversely.

Now consider a parameterized set $T = \{t_ξ : ξ \in Ξ\}$ in a metric space $(X, d)$, where $Ξ$ is a bounded subset of $\mathbb{R}^d$, and let $\| \cdot \|$ be a norm on $\mathbb{R}^d$. As $(Ξ, \| \cdot \|)$ is finite-dimensional in either sense (1.1) or (1.2), these properties are inherited by $T$ provided that the metric $d$ is comparable to $\| \cdot \|$. Indeed, if we have a Hölder-type upper bound $d(t_ξ, t_ξ') \leq C|ξ - ξ'|^α$, then $T$ satisfies the global entropy bound (1.1): if we have in addition the lower bound $d(t_ξ, t_0) \geq c|ξ - ξ_0|^α$, we obtain the local entropy bound (1.2) with $t_0 = t_ξ$. The upper bound is easily obtained in many cases of interest, so that finite-dimensionality in the sense (1.1) is not too problematic. The lower bound is much more delicate, however. In its absence, finite-dimensionality in the sense (1.2) is far from obvious.

Our guiding example, which is of significant independent interest, is a problem that arises from applications in statistics. Let us fix a probability density $f_0$ on $\mathbb{R}^d$ (that is, $f_0 \geq 0$ and $\int f_0 \, dx = 1$), and consider the class

$\mathcal{M}_q = \left\{ x \mapsto \sum_{i=1}^q \pi_i f_0(x - θ_i) : \pi_i \geq 0, \sum_{i=1}^q \pi_i = 1, θ_i \in Θ \right\}$

of convex combinations of $q$ translates of $f_0$, where $Θ$ is a bounded subset of $\mathbb{R}^d$. Such densities appear in numerous applications, where they are frequently known as location mixtures. $\mathcal{M}_q$ is a subset of the space $\mathcal{M}$ of all probability densities on $\mathbb{R}^d$, endowed with a suitable metric $d$.

$\mathcal{M}_q$ is parametrized by the finite-dimensional subset $Ξ_q = Δ_{q-1} \times Θ^q$ of $\mathbb{R}^{qd+q-1}$, where $Δ_{q-1}$ is the $q$-simplex. Natural metrics $d$ satisfy a Hölder-type upper bound with respect to a norm on $Ξ_q$ (e.g., step 2 in the proof of Theorem 3.1 below). However, the corresponding lower bound can be impossible to obtain.

**Example 1.1.** We will write $f_0(x) = f_0(x - θ)$ for simplicity. Fix $θ^* \in Θ$ and let $f^* = f_0$. Then $f^* \in \mathcal{M}_2$, but $f^*$ is not uniquely represented by a parameter in $Ξ_2$:

\begin{align*}
\{ (\pi, θ) \in Ξ_2 : d(π_1 f_0 + π_2 f_0, f^*) = 0 \} = \\
\{ π \in Δ_1, θ_1 = θ_2 = θ^* \} \cup \{ π_1 = 0, θ_1 \in Θ, θ_2 = θ^* \} \cup \{ π_1 = 1, θ_1 = θ^*, θ_2 \in Θ \}.
\end{align*}

Clearly $d$ cannot be lower bounded by any norm on $Ξ_2$, as such a bound would necessarily imply that $\{ (\pi, θ) \in Ξ_2 : d(π_1 f_0 + π_2 f_0, f^*) = 0 \}$ consists of a single point. Thus the above approach to (1.2) is useless here.
THE LOCAL GEOMETRY OF FINITE MIXTURES

Figure 1. Let \( f_\theta(x) = e^{-2(x-\theta)^2} \), \( f^* = f_{0.5} \), \( M_2 = \{ pf_{\theta_1} + (1-p)f_{\theta_2} : p, \theta_1, \theta_2 \in [0,1] \} \). The plots illustrate (a) the set of parameters \( (p, \theta_1, \theta_2) \) corresponding to the Hellinger ball \( \{ f \in M_2 : h(f, f^*) \leq 0.05 \} \); and (b) the parameter set \( \{ (p, \theta_1, \theta_2) : N(p, \theta_1, \theta_2) \leq 0.05 \} \) with \( N(p, \theta_1, \theta_2) = |p(\theta_1 - 0.5) + (1-p)(\theta_2 - 0.5)| + \frac{1}{2}p(\theta_1 - 0.5)^2 + \frac{1}{2}(1-p)(\theta_2 - 0.5)^2 \). The two plots are related by the local geometry Theorem 3.9, which yields \( c^* N(p, \theta_1, \theta_2) \leq h(pf_{\theta_1} + (1-p)f_{\theta_2}, f^*) \leq C^* N(p, \theta_1, \theta_2) \).

The phenomenon illustrated in this example can be stated more generally. For \( f^* \in M_{q^*} \) such that \( q^* < q \) (note that \( f^* \in M_q \) as \( M_q \subset M_{q+1} \) for all \( q \)), the subset of parameters \( \Xi_q(\delta) \subset \Xi_q \) corresponding to the ball \( M_q(\delta) = \{ f \in M_q : d(f, f^*) \leq \delta \} \) behaves nothing at all like a ball in a finite-dimensional Banach space (see Figure 1(a)): indeed, the diameter of \( \Xi_q(\delta) \) is even bounded away from zero as \( \delta \downarrow 0 \). There is therefore no hope to deduce a local entropy bound of the form (1.2) for \( N(M_q(\delta), \varepsilon) \) directly from the corresponding bound for a finite-dimensional Banach space. This natural example provides a vivid illustration of the difficulty of establishing local entropy bounds in geometrically irregular settings. The goal of this paper is to develop an approach for the investigation of such problems.

In section 2, we develop a useful technique to obtain local entropy bounds of the form (1.2). This method is not specific to mixtures, and is developed in a very general setting. We are motivated by the fact that, as explained above, global entropy bounds of the form (1.1) are typically much easier to obtain in geometrically complex problems than local entropy bounds. The main results of this section, Theorems 2.4 and 2.6, allow to deduce a local entropy bound for a subset \( T \) of a normed vector space from a global entropy bound for a certain weighted set \( D_0 \) associated to \( T \). While the latter bound may be far from trivial to obtain, it can provide a significant simplification of the original problem.

In section 3 we obtain local entropy bounds for the mixture classes \( M_q \). For concreteness, we endow \( M_q \) with the Hellinger metric \( h(f, g) = \| \sqrt{f} - \sqrt{g} \|_{L^2} \), which is the relevant metric for statistical applications [20, ch. 7], [18] (however, our results are easily adapted to other commonly used probability metrics—the total variation metric \( d_{TV}(f, g) = \| f - g \|_{L^1} \), for example—using almost identical proofs). The main result, Theorem 3.3, provides an explicit bound of the form (1.2) for \( M_q \) under suitable smoothness assumptions on \( f_0 \).
To prove Theorem 3.3 we first reduce the local entropy bound to a global entropy bound using the technique developed in section 2. To obtain this global entropy bound, however, we must develop a precise understanding of the local geometry of mixtures, which constitutes the main effort in the proof. The key result that we prove in this direction is Theorem 3.4. One consequence of this result, for example, is as follows: given a mixture \( f^* = \sum_{i=1}^{q^*} \pi_i^* f_{\theta_i^*} \), one can choose sufficiently small neighborhoods \( A_1, \ldots, A_{q^*} \) of \( \theta_1, \ldots, \theta_{q^*} \), respectively, such that for any \( q \geq 1 \) and mixture \( f = \sum_{i=1}^{q} \pi_i f_{\theta_i} \), the Hellinger metric \( h(f, f^*) \) is of the same order as

\[
\sum_{j \in A_0} \pi_j + \sum_{i=1}^{q^*} \left( \left| \sum_{j \in A_i} \pi_j - \pi_i^* \right| + \left\| \sum_{j \in A_i} \pi_j (\theta_j - \theta_i^*) \right\| + \frac{1}{2} \sum_{j \in A_i} \pi_j \| \theta_j - \theta_i^* \|^2 \right)
\]

(here \( A_0 = \mathbb{R}^d \setminus (A_1 \cup \cdots \cup A_{q^*}) \)). This pseudodistance controls precisely the set of parameters in \( \Xi_q \) with density close to \( f^* \), see Figure 1 for an example.

Beside their intrinsic interest, the results in this paper are of direct relevance to statistical applications. Many problems in statistics and probability make use of estimates on the metric entropy of classes of densities: metric entropy controls the rate of convergence of uniform limit theorems in probability, and is therefore of central importance in the design and analysis of statistical estimators [21, 20, 18]. Such applications frequently require a slightly stronger notion of metric entropy known as bracketing entropy, which we will consider throughout this paper; see section 2. In infinite-dimensional situations, the global entropy is chiefly of interest: global entropy estimates for various classes of probability densities can be found in [21, 20, 18, 3, 9]. However, in finite-dimensional settings, global entropy bounds are known to yield sub-optimal results, and here local entropy bounds are essential to obtain optimal convergence rates of estimators [20, §7.5]. In the case of mixtures, the difficulty of obtaining local entropy bounds was noted, e.g., in [12, 19]. Applications of the results in this paper are given in [11, 10].

2. FROM GLOBAL ENTROPY TO LOCAL ENTROPY

2.1. Definitions and results. We will consider two different notions of covering in normed vector spaces. The first is the classical covering by balls.

**Definition 2.1.** Let \((X, \| \cdot \|)\) be a normed vector space. For any subset \( T \subseteq X \) and \( \varepsilon > 0 \), the covering number \( N(T, \varepsilon) \) is defined as

\[
N(T, \varepsilon) = \inf \left\{ n : \exists x_i \in X, \ i = 1, \ldots, n \text{ s.t. } T \subseteq \bigcup_{i=1}^{n} B(x_i, \varepsilon) \right\},
\]

where \( B(x, \varepsilon) = \{ x' \in X : \|x - x'\| \leq \varepsilon \} \).

The second notion that we will consider is covering by brackets (order intervals), which requires a lattice structure. We will work in the general setting of normed Riesz spaces (normed Riesz spaces, see [1] for a basic introduction).

**Definition 2.2.** Let \((X, \| \cdot \|)\) be a normed vector lattice. For any subset \( T \subseteq X \) and \( \varepsilon > 0 \), the bracketing number \( N(T, \varepsilon) \) is defined as

\[
N(T, \varepsilon) = \inf \left\{ n : \exists l_i, u_i \in X, \ |u_i - l_i| \leq \varepsilon, \ i = 1, \ldots, n \text{ s.t. } T \subseteq \bigcup_{i=1}^{n} [l_i, u_i] \right\},
\]

where \([l, u] = \{ x \in X : l \leq x \leq u \} \).
Let (\(X, \| \cdot \| \)) be a normed vector lattice, and let us fix a subset \(T \subseteq X\) and a distinguished point \(t_0 \in T\). Our aim is to obtain an estimate on the local covering (or bracketing) number \(N(T \cap B(t_0, \delta), \varepsilon)\) that is polynomial in \(\delta/\varepsilon\). As is explained in the introduction, such estimates can be obtained from a global covering estimate for the weighted class \(D_0 \subseteq X\). As global entropy estimates can be much easier to obtain than local entropy estimates, this provides a very useful approach to obtaining local entropy bounds for geometrically complex classes. We will give two versions of our main result, one for bracketing numbers (Theorem 2.4) and one for covering numbers (Theorem 2.6).

**Theorem 2.4.** Let \((X, \| \cdot \|)\) be a normed vector lattice. Fix \(T \subseteq X\) and \(t_0 \in T\), and let \(D_0\) be as above. Suppose that there exist \(q, C_0 \geq 1\) and \(\varepsilon_0 > 0\) such that

\[
N(D_0, \varepsilon) \leq \left( \frac{C_0}{\varepsilon} \right)^q \quad \text{for every } \varepsilon \leq \varepsilon_0.
\]

Choose any \(d \in X\) such that \(|d_t| \leq d\) for all \(t \in T\), \(t \neq t_0\). Then

\[
N(T \cap B(t_0, \delta), \rho) \leq \left( \frac{8C\delta}{\rho} \right)^{q+1}
\]

for all \(\delta, \rho > 0\) such that \(\rho/\delta < 4 \wedge 2\|d\|\), where \(C = C_0(1 \vee \|d\|/4\varepsilon_0)\).

**Remark 2.5.** Theorem 2.4 requires an upper bound \(d \in X\) on \(|D_0|\), that is, \(D_0\) must be order-bounded. But the assumptions of the Theorem already require that \(N(D_0, \varepsilon_0) < \infty\), which is easily seen to imply order-boundedness of \(D_0\). The latter therefore does not need to be added as a separate assumption.

**Theorem 2.6.** Let \((X, \| \cdot \|)\) be a normed vector space. Fix \(T \subseteq X\) and \(t_0 \in T\), and let \(D_0\) be as above. Suppose that there exist \(q, C_0 \geq 1\) and \(\varepsilon_0 > 0\) such that

\[
N(D_0, \varepsilon) \leq \left( \frac{C_0}{\varepsilon} \right)^q \quad \text{for every } \varepsilon \leq \varepsilon_0.
\]

Then

\[
N(T \cap B(t_0, \delta), \rho) \leq \left( \frac{3C\delta}{\rho} \right)^{q+1}
\]

for all \(\delta, \rho > 0\) such that \(\rho/\delta \leq 1\), where \(C = C_0/(1 \wedge 2\varepsilon_0)\).

**Remark 2.7.** In the above results, a global covering bound for \(D_0\) of order \((1/\varepsilon)^q\) gives a local covering bound for \(T\) of order \((\delta/\varepsilon)^{q+1}\). It is instructive to note that this polynomial scaling cannot be improved. Indeed, let \(T\) be the unit (Euclidean)
ball in $\mathbb{R}^{q+1}$, and let $t_0 = 0$. Then $D_0$ is the unit sphere in $\mathbb{R}^{q+1}$ and therefore has Kolmogorov dimension $q$, but the covering number of $B(0, \delta)$ is of order $(\delta/\varepsilon)^{q+1}$.

**Remark 2.8.** A natural question is whether a converse to the above results can be obtained. In general, however, this is not possible: the class $D_0$ can be much richer than the original class $T$, as the following simple example illustrates. Let $(X, \| \cdot \|)$ be an infinite-dimensional Hilbert space and let $(e_k)_{k \geq 1}$ be an orthonormal basis. Let $T = \{2^{-k}e_k : k \geq 1\} \cup \{0\}$ and $t_0 = 0$. Then $N(T \cap B(t_0, 2^{-r}), 2^{-k}) \leq k - r + 1$ for $k \geq r$, so $N(T \cap B(t_0, \delta), \varepsilon) \leq \log_2(8\delta/\varepsilon) \leq (8\delta/\varepsilon)^{3/2}$ for all $\varepsilon/\delta \leq 1$. But here we have $D_0 = \{e_k : k \geq 1\}$, so $N(D_0, \varepsilon) = \infty$ for all $\varepsilon > 0$ small enough.

The proofs of Theorems 2.4 and 2.6 are almost identical. We will give a complete proof for the bracketing version (Theorem 2.4) in section 2.2, and briefly sketch the changes needed for its covering counterpart (Theorem 2.6) in section 2.3.

### 2.2. Proof of Theorem 2.4

The assumption implies that

$$N(D_0, \varepsilon) \leq \left( \frac{C_0}{\varepsilon \land \varepsilon_0} \right)^q$$

for every $\varepsilon > 0$.

If $\varepsilon < \|d\|/4$, then

$$\frac{\varepsilon}{\varepsilon \land \varepsilon_0} \leq 1 \lor \frac{\|d\|}{4\varepsilon_0}.$$

We therefore have

$$N(D_0, \varepsilon) \leq \left( \frac{C}{\varepsilon} \right)^q$$

for every $\varepsilon < \|d\|/4$,

where $C$ is as defined in the Theorem. This estimate will be used below.

Fix $\varepsilon, \delta > 0$ and let $N = N(D_0, \varepsilon)$. Then there exist $l_1, u_1, \ldots, l_N, u_N \in X$ such that $\|u_i - l_i\| \leq \varepsilon$ for all $i = 1, \ldots, N$, and for every $t \in T$, $t \neq t_0$ there is an $1 \leq i \leq N$ such that $l_i \leq d_t \leq u_i$. Choose $t \in T$ such that $r^{-n}\delta \leq \|t - t_0\| \leq r^{-n+1}\delta$ (with $r > 1$ to be chosen later). Then there exists $1 \leq i \leq N$ so that

$$(r^{-n}l_i \lor r^{-n+1}l_i)\delta + t_0 \leq t \leq (r^{-n}u_i \lor r^{-n+1}u_i)\delta + t_0.$$

Note that

$$\|u_i r^{-n}\delta - l_i r^{-n}\delta\| \leq r^{-n}\delta \varepsilon,$$

$$\|u_i r^{-n+1}\delta - l_i r^{-n+1}\delta\| \leq r^{-n+1}\delta \varepsilon,$$

$$\|u_i r^{-n+1}\delta - l_i r^{-n}\delta\| \leq (r-1)r^{-n}\delta + r^{-n+1}\delta \varepsilon,$$

$$\|u_i r^{-n}\delta - l_i r^{-n+1}\delta\| \leq (r-1)r^{-n}\delta + r^{-n+1}\delta \varepsilon,$$

where the latter two estimates follow from $l_i \leq d_t \leq u_i$, $\|d_t\| = 1$, and

$$(u_i - l_i) r^{-n}\delta \leq u_i r^{-n+1}\delta - l_i r^{-n}\delta - d_t (r-1)r^{-n}\delta \leq (u_i - l_i) r^{-n+1}\delta,$$

$$(u_i - l_i) r^{-n}\delta \leq u_i r^{-n}\delta - l_i r^{-n+1}\delta + d_t (r-1)r^{-n}\delta \leq (u_i - l_i) r^{-n+1}\delta.$$

As $|a \lor b - c \land d| \leq |a - c| + |a - d| + |b - c| + |b - d|$, we can estimate

$$\|(r^{-n}u_i \lor r^{-n+1}u_i)\delta - (r^{-n}l_i \lor r^{-n+1}l_i)\delta\| \leq 2(r-1)r^{-n}\delta + 4r^{-n+1}\delta \varepsilon.$$

Therefore, we have shown that

$$N(\{t \in T : r^{-n}\delta \leq \|t - t_0\| \leq r^{-n+1}\delta\}, 2(r-1)r^{-n}\delta + 4r^{-n+1}\delta \varepsilon) \leq N(D_0, \varepsilon)$$. 


for arbitrary $\varepsilon, \delta > 0$, $r > 1$, $n \in \mathbb{N}$. In particular,

$$\mathcal{N}(\{t \in T : r^{-n}\delta \leq \|t - t_0\| \leq r^{-n+1}\delta\}, \rho) \leq \mathcal{N}(D_0, \frac{1}{2}r^{n-1}\rho/\delta - \frac{1}{2}(1 - 1/r))$$

for every $\delta > 0$, $r > 1$, $n \in \mathbb{N}$, $\rho > 2(r - 1)r^{-n}\delta$.

Choose an envelope $d \in X$ such that $|d_t| \leq d$ for all $t \in T$, $t \neq t_0$. Evidently $t_0 - r^{-n}\delta d \leq t \leq t_0 + r^{-n}\delta d$ for all $t \in T$ such that $\|t - t_0\| \leq r^{-n}\delta$. Therefore

$$\mathcal{N}(\{t \in T : \|t - t_0\| \leq r^{-[H]}\delta\}, 2r^{-H}\delta \|d\|) = 1$$

for all $\delta > 0$, $r > 1$, $H > 0$. Thus we can estimate

$$\mathcal{N}(T \cap B(t_0, \delta), 2r^{-H}\delta \|d\|) \leq 1 + \sum_{n=1}^{[H]} \mathcal{N}(D_0, \{r^{n-H-1}\|d\| - (1 - 1/r)\}/2)$$

whenever $\delta > 0$, $r > 1$, $H > 0$ such that $\|d\| > (1 - 1/r)r^H$. In particular,

$$\mathcal{N}(T \cap B(t_0, \delta), 2r^{-H}\delta \|d\|) \leq 1 + \sum_{n=1}^{[H]} \mathcal{N}(D_0, r^{n-H-1}\|d\|/4)$$

whenever $\delta > 0$, $r > 1$, $H > 0$ such that $\|d\| \geq 2(1 - 1/r)r^H$, where we have used that the bracketing number is a nonincreasing function of the bracket size.

Now recall that

$$\mathcal{N}(D_0, \varepsilon) \leq \left(\frac{C}{\varepsilon}\right)^q$$

for every $0 < \varepsilon < \|d\|/4$,

where $q, C \geq 1$. Thus

$$\mathcal{N}(T \cap B(t_0, \delta), 2r^{-H}\delta \|d\|) \leq 1 + \sum_{n=1}^{[H]} r^{-(n-1)q} \left(\frac{8C}{2r^{-H}\|d\|}\right)^q$$

whenever $\delta > 0$, $r > 1$, $H > 0$ such that $\|d\| \geq 2(1 - 1/r)r^H$. But

$$\sum_{n=1}^{[H]} r^{-(n-1)q} \leq \frac{1}{1 - 1/r^q} \leq \frac{1}{1 - 1/r} \leq \frac{\|d\|}{2(1 - 1/r)r^H} \frac{4C}{2r^{-H}\|d\|}$$

as $r > 1$ and $q, C \geq 1$. We can therefore estimate

$$\mathcal{N}(T \cap B(t_0, \delta), 2r^{-H}\delta \|d\|) \leq \frac{\|d\|}{2(1 - 1/r)r^H} \left(\frac{8C}{2r^{-H}\|d\|}\right)^{q+1}$$

whenever $\delta > 0$, $r > 1$, $H > 0$ such that $\|d\| \geq 2(1 - 1/r)r^H$.

We now fix $\delta, \rho > 0$ such that $\rho/\delta < 4 \wedge 2\|d\|$, and choose

$$r = \frac{4}{4 - \rho/\delta}, \quad H = \frac{\log(2\|d\|\delta/\rho)}{\log r}.$$
Clearly \( r > 1 \) and \( H > 0 \). Moreover, note that our choice of \( r \) and \( H \) implies that \( \|d\| = 2(1 - 1/r)^rH \) and \( \rho = 2r^{-H}\|d\| \). We have therefore shown that 

\[
N(T \cap B(t_0, \delta), \rho) \leq \left( \frac{8C\delta}{\rho} \right)^{q+1}
\]

for all \( \delta, \rho > 0 \) such that \( \rho/\delta < 4 \wedge 2\|d\| \).

2.3. **Proof of Theorem 2.6.** Fix \( \varepsilon, \delta > 0 \) and let \( N = N(D_0, \varepsilon) \). Then there exist \( x_1, \ldots, x_N \in X \) such that for every \( t \in T \), \( t \neq t_0 \) there is an \( 1 \leq i \leq N \) such that \( \|d_i - x_i\| \leq \varepsilon \). Choose \( t \in T \) such that \( r^{-n}\delta \leq \|t - t_0\| \leq r^{-n+1}\delta \) (with \( r > 1 \) to be chosen later). Then there exists \( 1 \leq i \leq N \) so that 

\[
\|t - t_0 - x_i r^{-n}\delta\| \leq \|t - t_0 - d_i r^{-n}\delta\| + r^{-n}\delta \|d_i - x_i\| \leq (r - 1)r^{-n}\delta + r^{-n}\varepsilon,
\]

where we have used that \( \|d_i\| = 1 \). Therefore, we have shown that 

\[
N(\{t \in T : r^{-n}\delta \leq \|t - t_0\| \leq r^{-n+1}\delta\}, \rho) \leq N(D_0, r^n\rho/\delta - r + 1)
\]

for every \( \delta > 0 \), \( r > 1 \), \( n \in \mathbb{N} \), \( \rho > (r - 1)r^{-n}\delta \). On the other hand, clearly 

\[
N(\{t \in T : \|t - t_0\| \leq r^{-H}\delta\}, r^{-H}\delta) = 1
\]

for all \( \delta > 0 \), \( r > 1 \), \( H > 0 \). The remainder of the proof follows along exactly the same lines as that of Theorem 2.4 and is therefore omitted.

3. **The local entropy of mixtures**

3.1. **Definitions and main results.** Let \( \mu \) be the Lebesgue measure on \( \mathbb{R}^d \). We fix a positive probability density \( f_0 \) with respect to \( \mu \) (\( f_0 > 0 \) and \( \int f_0 d\mu = 1 \)), and consider mixtures (finite convex combinations) of densities in the class 

\[
\{f_\theta : \theta \in \mathbb{R}^d\}, \quad f_\theta(x) = f_0(x - \theta) \quad \forall x \in \mathbb{R}^d.
\]

In everything that follows we fix a nondegenerate mixture \( f^* \) of the form 

\[
f^* = \sum_{i=1}^q \pi_i^* f_{\theta_i^*}.
\]

Nondegenerate means that \( \pi_i^* > 0 \) for all \( i \), and \( \theta_i^* \neq \theta_j^* \) for all \( i \neq j \).

Let \( \Theta \subset \mathbb{R}^d \) be a bounded parameter set such that \( \{\theta_i^* : i = 1, \ldots, q^*\} \subseteq \Theta \), and denote its diameter by \( 2T \) (that is, \( \Theta \) is included in some closed Euclidean ball of radius \( T \)). We consider for \( q \geq 1 \) the family of \( q \)-mixtures 

\[
\mathcal{M}_q = \left\{ \sum_{i=1}^q \pi_i f_{\theta_i} : \pi_i \geq 0, \sum_{i=1}^q \pi_i = 1, \theta_i \in \Theta \right\}.
\]

The goal of this section is to obtain a local entropy bound for \( \mathcal{M}_q \) at the point \( f^* \), where \( \mathcal{M}_q \) is endowed with the Hellinger metric 

\[
h(f, g) = \left[ \int \left( \sqrt{f} - \sqrt{g} \right)^2 d\mu \right]^{1/2}, \quad f, g \in \mathcal{M}_q.
\]

That is, we seek bounds on quantities such as \( N_h(\{f \in \mathcal{M}_q : h(f, f^*) \leq \varepsilon\}, \delta) \), where \( N_h \) denotes the covering number in the metric space \( (\mathcal{M}_q, h) \) (i.e., covering by Hellinger balls). In fact, we prove a stronger bound of bracketing type. Our choice of the Hellinger metric and the particular form of the bracketing number to be considered is directly motivated by statistical applications [20 ch. 7], [18 §7.4];
Theorem 3.1. Suppose that Assumption A holds. Then there exist constants 
Assumption A.
The following hold:
when \( f \) is sufficiently differentiable, \( 0 \leq 4 \) and \( \delta \leq 8 \).

Our aim is to obtain a polynomial bound for the bracketing number \( N(\mathcal{H}_q(\varepsilon), \delta) \).
To this end, we will apply Theorem 2.4 to the weighted class \( \mathcal{D}_q \) defined by

\[
\mathcal{D}_q = \{ d_f : f \in \mathcal{M}_q, f \neq f^* \}, \quad d_f = \frac{\sqrt{f/f^*} - 1}{\|f/f^* - 1\|_2}.
\]

The essential difficulty is now to control the global entropy of \( \mathcal{D}_q \).

The following notation will be used throughout:

\[
\begin{align*}
H_0(x) &= \sup_{\theta \in \Theta} f_0(x)/f^*(x), \\
H_1(x) &= \sup_{\theta \in \Theta} \max_{i=1,\ldots,d} |\partial f_0(x)/\partial \theta^i|/f^*(x), \\
H_2(x) &= \sup_{\theta \in \Theta} \max_{i,j=1,\ldots,d} |\partial^2 f_0(x)/\partial \theta^i \partial \theta^j|/f^*(x), \\
H_3(x) &= \sup_{\theta \in \Theta} \max_{i,j,k=1,\ldots,d} |\partial^3 f_0(x)/\partial \theta^i \partial \theta^j \partial \theta^k|/f^*(x)
\end{align*}
\]
when \( f_0 \) is sufficiently differentiable, \( \mathcal{M} = \bigcup_{q \geq 1} \mathcal{M}_q \), and \( \mathcal{D} = \bigcup_{q \geq 1} \mathcal{D}_q \).

**Assumption A.** The following hold:

1. \( f_0 \in C^3 \) and \( f_0(x), \partial f_0/\partial \theta^i(x) \) vanish as \( \|x\| \to \infty \).
2. \( H_k \in L^4 (f^* d\mu) \) for \( k = 0, 1, 2 \) and \( H_3 \in L^2 (f^* d\mu) \).

We can now state our main result, whose proof is given in section 5.

**Theorem 3.1.** Suppose that Assumption A holds. Then there exist constants \( C^{*} \) and \( \delta^{*} \), which depend on \( d, q^{*} \) and \( f^{*} \) but not on \( \Theta, q \) or \( \delta \), such that

\[
N(\mathcal{D}_q, \delta) \leq \left( \frac{C^{*}(T \vee 1)^{1/6} \|H_0\|_4 \|H_1\|_4 \|H_2\|_4 \|H_3\|_2}{\delta} \right)^{18(d+1)q}
\]

for all \( q \geq q^{*}, \delta \leq \delta^{*} \). Moreover, there is a function \( D \in L^4 (f^* d\mu) \) with

\[
\|D\|_4 \leq K^{*} \|H_0\|_4 \|H_1\|_4 \|H_2\|_4 \|H_3\|_2,
\]
where \( K^{*} \) depends only on \( d \) and \( f^{*} \), such that \( |D| \leq D \) for all \( d \in \mathcal{D} \).

**Remark 3.2.** Assumption A is essentially a smoothness assumption on \( f_0 \). Some sort of smoothness is certainly needed for a result such as Theorem 3.1 to hold: see [11, §3] for a counterexample in the non-smooth case.

The bound of Theorem 3.1 is of independent interest (such a bound was assumed, e.g., in [15, 16] without or with incorrect proof). On the other hand, combining Theorems 2.4 and 3.1 we immediately obtain a local entropy bound for \( \mathcal{M}_q \).
Theorem 3.3. Suppose that Assumption A holds. Then
\[ N(H_q(\varepsilon), \delta) \leq \left( \frac{C_\Theta \varepsilon}{\delta} \right)^{18(d+1)q+1} \]
for all \( q \geq q^* \) and \( \delta/\varepsilon \leq 1 \), where
\[ C_\Theta = L^*(T \vee 1)^{1/6} \left( \|H_0\|_4^4 \vee \|H_1\|_4^4 \vee \|H_2\|_4^4 \vee \|H_3\|_2^{5/4} \right) \]
and \( L^* \) is a constant that depends only on \( d, q^* \) and \( f^* \).

To illustrate these results, let us consider the important case of Gaussian location mixtures, which are widely used in applications (see, e.g., [12, 13, 19]).

Example 3.4 (Gaussian mixtures). Consider mixtures of standard Gaussian densities \( f_0(x) = (2\pi)^{-d/2} e^{-\|x\|^2/2} \), and let \( \Theta(T) = \{ \theta \in \mathbb{R}^d : \|\theta\| \leq T \} \). Fix a nondegenerate mixture \( f^* \), and define \( T^* = \max_{i=1, \ldots, q^*} \|\theta_i\| \). Denote by \( H_q(\varepsilon, T) \) the Hellinger ball associated to the parameter set \( \Theta(T) \). Then
\[ N(H_q(\varepsilon, T), \delta) \leq \left( \frac{C_1^* e^2 T^2 \varepsilon}{\delta} \right)^{18(d+1)q+1} \]
for all \( q \geq q^* \), \( T \geq T^* \), and \( \delta/\varepsilon \leq 1 \), where \( C_1^*, C_2^* \) are constants that depend on \( d, q^* \) and \( f^* \) only. To prove this, it evidently suffices to show that Assumption A holds and that \( \|H_k\|_4 \) for \( k = 0, 1, 2 \) and \( \|H_3\|_2 \) are of order \( e^{C T^2} \). These facts are readily verified by a straightforward computation.

Let us emphasize a key feature of Theorems 3.1 and 3.3: the dependence of the entropy bounds on the order \( q \) and on the parameter set \( \Theta \) is explicit (see, e.g., Example 3.4). In particular, we find that for every \( f^* \), the local doubling dimension of \( M_q \) at \( f^* \) is of the same order as the dimension of the natural parameter set for mixtures \( \Delta_{q-1} \times \Theta^q \), which answers the basic question posed in the introduction. Obtaining this explicit dependence, which is important in applications [11], is one of the main technical challenges of the proof. In order to show only that \( N(H_q(\varepsilon), \delta) \) is polynomial in \( \varepsilon/\delta \) without explicit control of the order, the proof could be simplified and substantially generalized—see Remark 3.5 below for some discussion. In contrast to the dependence on \( q \) and \( \Theta \), however, the proofs of Theorems 3.1 and 3.3 do not provide any control of the dependence of the constants on \( f^* \). In particular, while we can control the local doubling dimension of \( M_q \) at \( f^* \) in terms of \( q \), we do not know whether the dependence on \( f^* \) can be eliminated.

Remark 3.5. We have not optimized the constants in Theorem 3.1 and Theorem 3.3. In particular, the constant 18 in the exponent can likely be improved. On the other hand, it is unclear whether the dependence on the diameter of \( \Theta \) is optimal. Indeed, if one is only interested in global entropy \( N(H_q, \delta) \) where \( H_q = \{ \sqrt{f/f^*} : f \in M_q \} \), then it can be read off from the proof of Theorem 3.1 that the constants in the entropy bound depend on \( \|H_0\|_1 \) and \( \|H_1\|_1 \) only, which are easily seen to scale polynomially in \( T \) due to the translation invariance of the Lebesgue measure. Therefore, for example in the case of Gaussian mixtures, one can obtain a global entropy bound which scales only polynomially as a function of \( T \), whereas the above local entropy bound scales as \( e^{C T^2} \). The behavior of local entropies is much more delicate than that of global entropies, however, and we do not know whether it is possible to obtain a local entropy bound that scales polynomially in \( T \) for the Hellinger metric. On the other hand, if \( M_q \) is endowed with the total
Lemma 3.7. It is possible to choose a bounded convex neighborhood \( A \) for every \( \theta i \) such that \( \| H_i \| \) (\( i = 0, \ldots, 3 \)), and therefore scales polynomially in \( T \). In this case the scaling matches that of the global entropy, and is therefore optimal.

Proof. We first claim that one can choose linearly independent \( u, v \) for every \( \theta i \) such that \( |\{ \langle \theta i, u \rangle : i = 1, \ldots, q^* \}| = q^* \) for every \( j = 1, \ldots, d \). Indeed, note that the set \( \{ u \in \mathbb{R}^d : |\{ \langle \theta i, u \rangle : i = 1, \ldots, q^* \}| < q^* \} \) is a finite union of \((d-1)\)-dimensional hyperplanes, which has Lebesgue measure zero. Therefore, if we draw a rotation matrix \( T \) at random from the Haar measure on \( \text{SO}(d) \), and let \( u_i = T e_i \) for all \( i = 1, \ldots, d \) where \( \{ e_1, \ldots, e_d \} \) is the standard Euclidean basis in \( \mathbb{R}^d \), then the desired property will hold with unit probability. To complete the proof, it suffices to choose \( A_j = B(\theta^1_i, \varepsilon / 4) \) with \( \varepsilon = \min k \min i \min j |\langle \theta^1_i \theta^1_j, u_k \rangle| \).

We now fix once and for all a family of neighborhoods \( A_1, \ldots, A_q \), as in Lemma 3.7. The precise choice of these sets only affects the constants in the proofs below.
Then we define for each \((\eta, \beta, \rho, \tau, \nu)\) denoted by \(M\) of Definition 3.8. Let us write 
\[
\int \lambda(d\theta).
\]
Theorem 3.9. Before we turn to the proof, let us introduce a notion that is familiar in quantum mechanics. If \((\Omega, \Sigma)\) is a measurable space, call the map \(\lambda : \Sigma \to \mathbb{R}^{d \times d}\) a \textit{state} if
\[
(1)\ A \mapsto [\lambda(A)]_{ij}\ is\ a\ signed\ measure\ for\ every\ i, j = 1, \ldots, d;
\]
\[
(2)\ \lambda(A)\ is\ a\ nonnegative\ symmetric\ matrix\ for\ every\ A \in \Sigma;
\]
\[
(3)\ \text{Tr}[\lambda(\Omega)] = 1.
\]
4 Our terminology is in analogy with the notion of a state on the \(C^*\)-algebra \(C^{d \times d} \otimes C_{\mathbb{C}}(\Omega)\), where \(\Omega\) is a compact metric space and \(C_{\mathbb{C}}(\Omega)\) is the algebra of complex-valued continuous functions on \(\Omega\). Such states can be represented by the complex-valued counterpart of our definition.
It is easily seen that for any unit vector $\xi \in \mathbb{R}^d$, the map $A \mapsto \langle \xi, \lambda(A) \xi \rangle$ is a sub-probability measure. Moreover, if $\xi_1, \ldots, \xi_d \in \mathbb{R}^d$ are linearly independent, there must be at least one $\xi_i$ such that $\langle \xi_i, \lambda(\xi) \xi_i \rangle > 0$. Finally, let $B \subset \mathbb{R}^d$ be a compact set and let $(\lambda_n)_{n \geq 0}$ be a sequence of states on $B$. Then there exists a subsequence along which $\lambda_n$ converges weakly to some state $\lambda$ on $B$ in the sense that $\int \text{Tr}[M(\theta) \lambda_n(d\theta)] \to \int \text{Tr}[M(\theta) \lambda(d\theta)]$ for every continuous function $M : B \to \mathbb{R}^{d \times d}$. To see this, it suffices to note that we may extract a subsequence such that all matrix elements $|\lambda_{ij}|$ converge weakly to a signed measure by the compactness of $B$, and it is evident that the limit must again define a state.

**Proof of Theorem 3.9.** Suppose that the conclusion of the theorem does not hold. Then there must exist a sequence of coefficients $(\eta^n, \beta^n, \rho^n, \tau^n, \nu^n) \in \mathcal{D}$ with

$$\frac{\|\ell(\eta^n, \beta^n, \rho^n, \tau^n, \nu^n)\|_1}{N(\eta^n, \beta^n, \rho^n, \tau^n, \nu^n)} \not\to 0.$$ 

Let us fix such a sequence throughout the proof.

Applying Taylor’s theorem to $u \mapsto f_{\theta_i^* + u(\theta - \theta_i^*)}$, we can write for $i = 1, \ldots, q^*$

$$\eta_i^n f_{\theta_i^*} + \beta_i^n D_1 f_{\theta_i^*} + \frac{\eta_i^n}{\eta_i^n + \tau_i^n} D_2 f_{\theta_i^*} + \frac{\eta_i^n}{\eta_i^n + \tau_i^n} \int (\theta - \theta_i^*) \nu_i^n(d\theta) f_{\theta_i^*} + \frac{\eta_i^n}{\eta_i^n + \tau_i^n} \int \frac{D_2 f_{\theta_i^*} + u(\theta - \theta_i^*)}{f_{\theta_i^*}} + \frac{\eta_i^n}{\eta_i^n + \tau_i^n} \int \frac{D_2 f_{\theta_i^*} + u(\theta - \theta_i^*)}{f_{\theta_i^*}} \nu_i^n(d\theta)$$

(1) There exist $a_i \in \mathbb{R}^d$, $b_i \in \mathbb{R}^d$, $c_i \in \mathbb{M}_d$, $a_0 \geq 0$ (for $i = 1, \ldots, q^*$) with $\|a_i\| + \sum_{i=1}^{q^*} |a_i| + \|b_i\| + \text{Tr}[c_i] + \|d_i\| = 1$, such that $a_0 \to a_0$ and $a_0^n \to a_i$, $b_0^n \to b_i$, $c_0^n \to c_i$, $d_0^n \to d_i$ as $n \to \infty$ for all $i = 1, \ldots, q^*$. (2) There exists a sub-probability measure $\nu_0$ supported on $A_0$, such that $\nu_0^n$ converges vaguely to $\nu_0$ as $n \to \infty$. 

Note that

$$|a_0^n| + \sum_{i=1}^{q^*} |a_i^n| + \|b_i^n\| + \text{Tr}[c_i^n] + |d_i^n| = 1$$

for all $n$. We may therefore extract a subsequence such that:

$$\frac{\|\ell(\eta^n, \beta^n, \rho^n, \tau^n, \nu^n)\|_1}{N(\eta^n, \beta^n, \rho^n, \tau^n, \nu^n)} \not\to 0.$$ 


We may therefore apply Fubini’s theorem, giving

\[ F[\ell] = a_0 f_0 + \sum_{i=1}^{q^*} \left\{ a_i f_{\theta^*_i} + b_i^* D_1 f_{\theta^*_i} + \text{Tr}[c_i D_2 f_{\theta^*_i}] \right\} \]

\[ + d_i \int \text{Tr} \left\{ \int_0^1 D_2 f_{\theta^*_i+u(\theta-\theta^*_i)} 2(1-u) \, du \right\} \lambda_i(d\theta) \right\}. \]

But as \( \|\ell(\eta^n, \beta^n, \rho^n, \tau^n, \nu^n)/N(\eta^n, \beta^n, \rho^n, \tau^n, \nu^n) \| \to 0 \), we have \( \|h/f^*\|_1 = 0 \) by Fatou’s lemma. As \( f^* \) is strictly positive, we must have \( h \equiv 0 \).

To proceed, we need the following lemma.

**Lemma 3.10.** The Fourier transform \( F[h](s) := \int e^{i(x,s)} h(x) \, dx \) is given by

\[ F[h](s) = F[f_0](s) \left[ a_0 \int e^{i(\theta,s)} \nu_0(d\theta) + \sum_{i=1}^{q^*} \left\{ a_i e^{i(\theta^*_i,s)} + i \langle b_i, s \rangle e^{i(\theta^*_i,s)} \right\} - \langle s, c_i(s) e^{i(\theta^*_i,s)} - d_i e^{i(\theta^*_i,s)} \int \phi(i(\theta - \theta^*_i, s)) \langle s, \lambda_i(d\theta) s \rangle \right\} \right] \]

for all \( s \in \mathbb{R}^d \). Here we defined the function \( \phi(u) = 2(e^u - u - 1)/u^2 \).

**Proof.** The \( a_i, b_i, c_i \) terms are easily computed using integration by parts. It remains to compute the Fourier transform of the function

\[ \Xi_i(x)_{jk} = \int \left\{ \int_0^1 [D_2 f_{\theta^*_i+u(\theta-\theta^*_i)}(x)]_{jk} 2(1-u) \, du \right\} \lambda_i(d\theta) \].

We begin by noting that

\[ \int \int_0^1 \|D_2 f_{\theta^*_i+u(\theta-\theta^*_i)}(x)\|_{jk} 2(1-u) \, du \, dx \|\lambda_i\|_{kj} \leq \|\lambda_i\|_{TV} \int \|D_2 f_0(x)\|_{jk} \, dx < \infty. \]

We may therefore apply Fubini’s theorem, giving

\[ F[\Xi_i](s) = -F[f_0](s) s_j s_k e^{i(\theta^*_i,s)} \left\{ \int_0^1 e^{iu(\theta-\theta^*_i)s} 2(1-u)du \right\} \lambda_i(d\theta) \] \[ = -F[f_0](s) s_j s_k e^{i(\theta^*_i,s)} \int \phi(i(\theta - \theta^*_i, s)) \lambda_i(d\theta) \] \[ \text{where we have computed the inner integral using integration by parts.} \]

Let \( u_1, \ldots, u_d \in \mathbb{R}^d \) be a linearly independent family satisfying the condition of Lemma 3.7. As \( F[h](s) = 0 \) for all \( s \in \mathbb{R}^d \), we obtain

\[ \Phi^d(it) := a_0 \Phi_0^d(it) + \sum_{i=1}^{q^*} e^{it(\theta^*_i,u_i)} \left\{ a_i + it \langle b_i, u_i \rangle - t^2 \langle u_i, c_iu_i \rangle - d_i t^2 \right\} = 0 \]

for all \( \ell = 1, \ldots, d \) and \( t \in [-\ell, \ell] \subset \mathbb{R} \) for some \( \ell > 0 \), where we defined

\[ \Phi^d(it) = \int \phi(it(\theta - \theta^*_i, u_i)) \langle u_i, \lambda_i(d\theta) u_i \rangle \]
Indeed, it suffices to note that $F[f_0](0) = 1$ and that $s \mapsto F[f_0](s)$ is continuous, so that this claim follows from Lemma 3.10 and the fact that $F[f_0](s)$ is nonvanishing in a sufficiently small neighborhood of the origin.

As all $\lambda_i$ have compact support, it is easily seen that for every $i = 1, \ldots, q^*$, the function $\Phi^i(z)$ is defined for all $z \in \mathbb{C}$ by a convergent power series. The function $\Psi^i(it) := \Phi^i(it) - a_i \Phi^i_0(it)$ is therefore an entire function with $|\Psi^i(z)| \leq k_1 e^{k_2|z|}$ for some $k_1, k_2 > 0$ and all $z \in \mathbb{C}$. But as $\Phi^i(it) = 0$ for $t \in [-i, i]$, it follows from Theorem 7.2.2 that $a_i \Phi^i_0(it)$ is the Fourier transform of a finite measure with compact support. Thus we may assume without loss of generality that the law of $\langle \theta, u_t \rangle$ under the sub-probability $\nu_0$ is supported for each $\ell = 1, \ldots, d$, so by linear independence $\nu_0$ must be compactly supported. Therefore, the function $\Phi^i(z)$ is defined for all $z \in \mathbb{C}$ by a convergent power series. But as $\Phi^i(z)$ vanishes for $z \in i[-i, i]$, we must have $\Phi^i(z) = 0$ for all $z \in \mathbb{C}$, and in particular

\begin{equation}
(\text{3.1}) \quad \Phi^i(t) = a_i \Phi^i_0(t) + \sum_{i=1}^{q^*} e^{it \langle \theta^i, u_t \rangle} \left\{ a_i + t \langle b_i, u_{\ell} \rangle + t^2 \langle c_i, u_{\ell} \rangle + d_i \right\} = 0
\end{equation}

for all $t \in \mathbb{R}$ and $\ell = 1, \ldots, d$. In the remainder of the proof, we argue that (3.1) cannot hold, thus completing the proof by contradiction.

At the heart of our proof is an inductive argument. Recall that by construction, the projections $\{\langle A_i, u_t \rangle : i = 1, \ldots, q^*\}$ are disjoint open intervals in $\mathbb{R}$ for every $\ell = 1, \ldots, d$. We can therefore relabel them in increasing order: that is, define $(\ell_1), \ldots, (\ell_{q^*}) \in \{1, \ldots, q^*\}$ so that $\langle \theta_{\ell(1)}, u_\ell \rangle < \langle \theta_{\ell(2)}, u_\ell \rangle < \cdots < \langle \theta_{\ell(q^*)}, u_\ell \rangle$. The following key result provides the inductive step in our proof.

**Proposition 3.11.** Fix $\ell \in \{1, \ldots, d\}$, and define

$$
\Phi^\ell_0(t) := a_0 \Phi^\ell_0(t) + \sum_{i=1}^{q^*} a_i e^{it \langle \theta^i, u_{\ell} \rangle}.
$$

Suppose that for some $j \in \{1, \ldots, q^*\}$ we have $\Phi^{\ell,j}(t) = 0$ for all $t \in \mathbb{R}$, where

$$
\Phi^{\ell,j}(t) := \Phi^\ell_0(t) + \sum_{i=1}^{j} e^{it \langle \theta_{\ell(i)}, u_{\ell} \rangle} \left\{ t \langle b_{\ell(i)}, u_t \rangle + t^2 \langle c_{\ell(i)}, u_t \rangle + d_{\ell(i)} t^2 \Phi^\ell_{\ell(i)}(t) \right\}.
$$

Then $d_{\ell(j)}(u_{\ell}, \lambda_{\ell(j)}) (\mathbb{R}^d | u_\ell) = 0$, $u_\ell, c_{\ell(j)} u_\ell = 0$, and $b_{\ell(j)} u_\ell = 0$.

**Proof.** Let us write for simplicity $\theta^\ell_i = (\theta^i, u_{\ell})$, and denote by $\lambda_0^{\ell}$ and $\nu_0^\ell$ the finite measures on $\mathbb{R}$ defined such that $\int f(x) \lambda_0^{\ell}(dx) = \int f(\langle \theta, u_\ell \rangle)(u_{\ell}, \lambda_0(d\theta)u_{\ell})$ and $\int f(x) \nu_0^\ell(dx) = \int f(\langle \theta, u_\ell \rangle)\nu_0(d\theta)$, respectively. For notational convenience, we will assume in the following that $(\ell(i)) = i$ and $\nu_0^\ell(\{\theta^\ell_i\}) = 0$ for all $i = 1, \ldots, q^*$. This entails no loss of generality: the former can always be attained by relabeling of the points $\theta^\ell_i$, while $\Phi^\ell_0$ is unchanged if we replace $\nu_0^\ell$ and $a_i$ by $\nu_0^\ell(\cdot \cap \mathbb{R}\setminus\{\theta^\ell_1, \ldots, \theta^\ell_{q^*}\})$ and $a_i + a_i \nu_0^\ell(\{\theta^\ell_i\})$, respectively. Note that

$$
\langle A_i, u_\ell \rangle = [\theta^\ell_{i-1}, \theta^\ell_i], \quad \text{where } \theta^\ell_{i-1} < \theta^\ell_i < \theta^\ell_{i+1} < \theta^\ell_{i+1} \text{ for all } i
$$

by our assumptions ($\langle A_i, u_\ell \rangle$ must be an interval as $A_i$ is convex).
Step 1. We claim that the following hold:

\[ a_i = 0 \text{ for all } i \geq j + 1 \quad \text{and} \quad a_0 \nu_0^{\ell_j}(\theta_{j+1}, \infty) = 0. \]

Indeed, suppose this is not the case. Then it is easily seen that

\[ \lim_{t \to \infty} \frac{\Phi_0^{\ell_j}(t)}{e^{t\theta_{j+1}}} > 0, \]

where we have used that \( \nu_0^{\ell_j} \) has no mass at \( \{\theta_{j+1}^+, \theta_{j+1}^-\} \). On the other hand, as \( \phi \) is positive and increasing and as \( \lambda_i \) is supported on \( \text{cl} A_i \), we can estimate

\[
0 \leq \frac{t^2 e^{t\theta_{j+1}} \Phi_0^{\ell_j}(t)}{e^{t\theta_{j+1}}} \leq t^2 e^{-t(\theta_{j+1}^+ - \theta_{j+1}^-)} \phi(t\{\theta_{j+1}^+ - \theta_{j+1}^-\}) \lambda_j^i(\mathbb{R}) \xrightarrow{t \to \infty} 0
\]

for \( i = 1, \ldots, j \). But then we must have

\[ 0 = \lim_{t \to \infty} \frac{\Phi_0^{\ell_j}(t)}{e^{t\theta_{j+1}}} > 0, \]

which yields the desired contradiction.

Step 2. We claim that the following hold:

\[ d_j \lambda_j^i(\theta_{j+1}^+, \infty) = 0, \quad \langle u_t, c_j u_t \rangle = 0, \quad \text{and} \quad a_0 \nu_0^{\ell_j}(\theta_{j+1}^+, \infty) = 0. \]

Indeed, suppose this is not the case. As \( \nu_0^{\ell_j}(\{\theta_{j+1}^+\}) = 0 \), we can choose \( \varepsilon > 0 \) such that \( \nu_0^{\ell_j}(\theta_{j+1}^+ + \varepsilon, \infty) \geq \nu_0^{\ell_j}(\theta_{j+1}^+, \infty)/2 \). As \( a_0, d_j \geq 0 \), and using that \( \phi \) is positive and increasing with \( \phi(0) = 1 \) and that \( e^{\varepsilon t} \geq (\varepsilon t)^2/2 \) for \( t \geq 0 \), we can estimate

\[
0 = a_0 \Phi_0^{\ell_j}(t) + e^{t\theta_{j+1}} \left\{ t^2 \langle u_t, c_j u_t \rangle + d_j t^2 \Phi_0^{\ell_j}(t) \right\} \geq \frac{t^2}{4} a_0 \nu_0^{\ell_j}(\theta_{j+1}^+, \langle u_t, c_j u_t \rangle) + d_j t^2 \lambda_j^i(\theta_{j+1}^+, \infty) \geq 0
\]

for all \( t \geq 0 \). On the other hand, it is easily seen that

\[
\frac{1}{t^2} e^{t\theta_{j+1}} \left[ \sum_{i=1}^j e^{t\theta_i} \left\{ a_i + t\langle b_i, u_t \rangle \right\} + \sum_{i=1}^{j-1} e^{t\theta_i} \left\{ t^2 \langle u_t, c_i u_t \rangle + d_i t^2 \Phi_0^{\ell_j}(t) \right\} \right] \xrightarrow{t \to \infty} 0.
\]

But this would imply that

\[ 0 = \lim_{t \to \infty} \frac{\Phi_0^{\ell_j}(t)}{a_0 \Phi_0^{\ell_j}(t) + e^{t\theta_{j+1}} \left\{ t^2 \langle u_t, c_j u_t \rangle + d_j t^2 \Phi_0^{\ell_j}(t) \right\}} = 1, \]

which yields the desired contradiction.

Step 3. We claim that the following hold:

\[ d_j \lambda_j^i(\theta_{j+1}^-, \theta_{j+1}^+) = 0 \quad \text{and} \quad a_0 \nu_0^{\ell_j}(\theta_{j+1}^-, \theta_{j+1}^+) = 0. \]

Indeed, suppose this is not the case. We can compute

\[
0 = \frac{d^2}{dt^2} \left( \frac{\Phi_0^{\ell_j}(t)}{e^{t\theta_{j+1}}} \right) = d_j \int e^{t(\theta_{j+1}^+ - \theta_{j+1}^-)} \lambda_j^i(d\theta) + a_0 \int e^{t(\theta_{j+1}^+ - \theta_{j+1}^-)} (\theta_{j+1}^+ - \theta_{j+1}^-)^2 \nu_0^{\ell_j}(d\theta)
\]

\[ + \sum_{i=1}^{j-1} \frac{d^2}{dt^2} e^{-t(\theta_{j+1}^+ - \theta_{j+1}^-)} \left\{ a_i + t\langle b_i, u_t \rangle + t^2 \langle u_t, c_i u_t \rangle + d_i t^2 \Phi_0^{\ell_j}(t) \right\}, \]
where the derivative and integral may be exchanged by [22], Appendix A16. We now note that as \( t = 1 \), we can estimate for \( t \geq 0 \)

\[
d_j \int e^{i(\theta - \theta_j^i)} \lambda_j^\ell(d\theta) + a_0 \int e^{i(\theta - \theta_j^i)} (\theta - \theta_j^i) \nu_0(d\theta) \geq e^{i(\theta_j^i - \theta_j^i)} \left\{ d_j \lambda_j^\ell(\theta_j^i, \theta_j^i) + a_0 \int_{\theta_j^i - \theta_j^i} (\theta - \theta_j^i) \nu_0(d\theta) \right\} > 0.
\]

On the other hand, as \((e^x - 1)/x\) is positive and increasing, we obtain for \( t \geq 0 \)

\[
e^{-t(\theta_j^i - \theta_j^i)} \left| \frac{d^2}{dt^2} e^{-t(\theta_j^i - \theta_j^i)} \right| _2 \Phi_j^\ell(t) = e^{-t(\theta_j^i - \theta_j^i)} \times e^{-t(\theta_j^i - \theta_j^i)} \times \left| \frac{e^{i(\theta_j^i - \theta_j^i)} - 1}{\theta - \theta_j^i} \lambda_j^\ell(d\theta) + \int e^{i(\theta - \theta_j^i)} \lambda_j^\ell(d\theta) \right| \]

\[
- 2(\theta_j^i - \theta_j^i) \int e^{i(\theta - \theta_j^i)} \frac{1}{\theta - \theta_j^i} \lambda_j^\ell(d\theta) + \int e^{i(\theta - \theta_j^i)} \lambda_j^\ell(d\theta) \]

\[
\leq e^{-t(\theta_j^i - \theta_j^i)} \left\{ (\theta_j^i - \theta_j^i)^2 \phi(t) \{ \theta_j^i - \theta_j^i \} + 2(\theta_j^i - \theta_j^i)^2 \phi(\theta_j^i - \theta_j^i) \right\} \lambda_j^\ell(\mathbb{R}),
\]

which converges to zero as \( t \to \infty \) for every \( i < j \). It follows that

\[
0 = \lim_{t \to \infty} \frac{\partial^2}{\partial \theta_j^i} \left( \Phi_j^\ell(t)/e^{i\theta_j^i} \right) = 1,
\]

which yields the desired contradiction.

**Step 4.** Recall that \( \lambda_j^\ell \) is supported on \([\theta_j^{-}, \theta_j^{+}]\) by construction. We have therefore established in the previous steps that the following hold:

\[
d_j \langle u_\ell, \lambda_j(\mathbb{R}_j^d)u_\ell \rangle = \langle u_\ell, c_j u_\ell \rangle = a_0 \nu_0([\theta_j^{-}, \infty]) = 0, \quad a_i = 0 \text{ for } i > j.
\]

It is therefore easily seen that

\[
0 = \lim_{t \to \infty} \frac{\Phi_j(t)}{t e^{i\theta_j^i}} = \langle b_j, u_\ell \rangle.
\]

Thus the proof is complete. \(\square\)

We can now perform the induction by starting from (3.1) and applying Proposition 3.1 repeated. This yields \( d_j \langle u_\ell, \lambda_j(\mathbb{R}_j^d)u_\ell \rangle = \langle u_\ell, c_j u_\ell \rangle = \langle b_j, u_\ell \rangle = 0 \) for all \( j = 1, \ldots, q^* \) and \( \ell = 1, \ldots, d \). As \( u_1, \ldots, u_d \) are linearly independent and \( c_j \in M^d_j \), this implies that \( b_j = 0, c_j = 0 \) and \( d_j = 0 \) for all \( j = 1, \ldots, q^* \), so that

\[
a_0 \int e^{i(\theta, s)} \nu_0(d\theta) + \sum_{i=1}^{q^*} a_i e^{i(\theta_i, s)} = 0
\]

for all \( s \in \mathbb{R}_j^d \) (this follows as above by Lemma 3.10, \( h \equiv 0, F[f_0](s) \neq 0 \) for \( s \) in a neighborhood of the origin, and using analyticity). But by the uniqueness of Fourier transforms, this implies that the signed measure \( a_0 \nu_0 + \sum_{i=1}^{q^*} a_i \delta_{\{\theta_i\}} \) has no mass. As \( \nu_0 \) is supported on \( A_0 \), this implies that \( a_j = 0 \) for all \( j = 1, \ldots, q^* \).
We have therefore shown that $a_i, b_i, c_i, d_i = 0$ for all $i = 1, \ldots, q^*$. But recall that $|a_0| + \sum_{i=1}^{q^*} \{|a_i| + \|b_i\| + \text{Tr}[c_i] + |d_i|\} = 1$, so that evidently $a_0 = 1$.

To complete the proof, it remains to note that
\[
\int \frac{\ell(\eta^n, \beta^n, \rho^n, \tau^n, \nu^n)}{N(\eta^n, \beta^n, \rho^n, \tau^n, \nu^n)} f^* d\mu = \sum_{i=0}^{q^*} a_i^n \xrightarrow{n \to \infty} 1.
\]
But this is impossible, as
\[
\left\| \frac{\ell(\eta^n, \beta^n, \rho^n, \tau^n, \nu^n)}{N(\eta^n, \beta^n, \rho^n, \tau^n, \nu^n)} \right\|_{n \to \infty} < 0
\]
by construction. Thus we have the desired contradiction. \hfill \Box

3.3. **Proof of Theorem 3.1.** The proof of Theorem 3.1 consists of a sequence of approximations, which we develop in the form of lemmas. Throughout this section, we always presume that Assumption A holds.

We begin by establishing the existence of an envelope function.

**Lemma 3.12.** Define $S = (H_0 + H_1 + H_2) d/c^*$. Then $S \in L^4(f^* d\mu)$, and
\[
\frac{|f|}{\|f/f^* - 1\|_1} \leq S \quad \text{for all } f \in \mathcal{M}.
\]

**Proof.** That $S \in L^4(f^* d\mu)$ follows directly from Assumption A. To proceed, let $f \in \mathcal{M}_q$, so that we can write $f = \sum_{i=1}^q \pi_i f_{\theta_i}$. Then
\[
\frac{f - f^*}{f^*} = \sum_{j: \theta_j \in A_0} \pi_j f_{\theta_j} f^* + \sum_{i=1}^{q^*} \left\{ \left( \sum_{j: \theta_j \in A_i} \pi_j - \pi_i^* \right) f_{\theta_i}^* + \sum_{j: \theta_j \in A_i} \pi_j f_{\theta_j} - f_{\theta_i}^* \right\}.
\]
Taylor expansion gives
\[
f_{\theta_j}(x) - f_{\theta_i}^*(x) = (\theta_j - \theta_i^*)^* D_1 f_{\theta_i}^*(x) + \frac{1}{2} \int_0^1 (\theta_j - \theta_i^*)^* D_2 f_{\theta_i}^* + u(\theta_j - \theta_i^*) (x) (\theta_j - \theta_i^*) 2(1 - u) du.
\]
Using Assumption A, we find that
\[
\left\| \frac{f - f^*}{f^*} \right\|_1 \leq \left[ \sum_{j: \theta_j \in A_0} \pi_j + \sum_{i=1}^{q^*} \left\{ \left( \sum_{j: \theta_j \in A_i} \pi_j - \pi_i^* \right) + \left( \sum_{j: \theta_j \in A_i} \pi_j (\theta_j - \theta_i^*) \right) \right\} \right] (H_0 + H_1 + H_2) d.
\]
On the other hand, Theorem 3.9 gives
\[
\left\| \frac{f - f^*}{f^*} \right\|_1 \geq c^* \left[ \sum_{j: \theta_j \in A_0} \pi_j + \sum_{i=1}^{q^*} \left\{ \left( \sum_{j: \theta_j \in A_i} \pi_j - \pi_i^* \right) \right\} \right]
\]
\[
+ \left[ \sum_{j: \theta_j \in A_i} \pi_j (\theta_j - \theta_i^*) \right] + \frac{1}{2} \left[ \sum_{j: \theta_j \in A_i} \pi_j ||\theta_j - \theta_i^*||^2 \right].
\]
The proof follows directly. \hfill \Box

**Corollary 3.13.** $|d| \leq D$ for all $d \in \mathcal{D}$, where $D = 2S \in L^4(f^* d\mu)$. 

Proof. Using \( \|f - f^*\|_{TV} \leq 2h(f, f^*) \) and \( |\sqrt{x} - 1| \leq |x - 1| \), we find

\[
|d_f| = \left| \frac{\sqrt{f/f^*} - 1}{h(f, f^*)} \right| \leq \left| \frac{f/f^* - 1}{2\|f/f^* - 1\|_1} \right| \leq 2S,
\]

where we have used Lemma 3.12.

Next, we prove that the Hellinger normalized densities \( d_f \) can be approximated by chi-square normalized densities for small \( h(f, f^*) \).

**Lemma 3.14.** For any \( f \in \mathcal{M} \), we have

\[
\left| \frac{\sqrt{f/f^*} - 1}{h(f, f^*)} - \frac{f/f^* - 1}{\sqrt{\chi^2(f||f^*)}} \right| \leq \left\{ 4\|S\|^2_S + 2S^2 \right\} h(f, f^*),
\]

where we have defined the chi-square divergence \( \chi^2(f||f^*) = \|f/f^* - 1\|_2^2 \).

**Proof.** Let us define the function \( R \) as

\[
\sqrt{\frac{f}{f^*}} - 1 = \frac{1}{2} \left\{ \frac{f - f^*}{f^*} + R \right\}.
\]

Then we have

\[
\frac{\sqrt{f/f^*} - 1}{h(f, f^*)} \frac{f/f^* - 1}{\sqrt{\chi^2(f||f^*)}} = \frac{f/f^* - 1 + R}{\|f/f^* - 1 + R\|_2} - \frac{f/f^* - 1}{\|f/f^* - 1\|_2} = \frac{R}{\|f/f^* - 1 + R\|_2} \leq \frac{R}{\|f/f^* - 1\|_2},
\]

so that by the reverse triangle inequality and Corollary 3.13

\[
\left| \frac{\sqrt{f/f^*} - 1}{h(f, f^*)} - \frac{f/f^* - 1}{\sqrt{\chi^2(f||f^*)}} \right| \leq \frac{2\|R\|_2 S + |R|}{\|f/f^* - 1\|_2}.
\]

Now note that for all \( x \geq 1 \)

\[
-x^2 \leq -\left( \frac{\sqrt{1 + x} - 1}{2} \right)^2 = \sqrt{1 + x} - 1 - \frac{x}{2} \leq 0.
\]

Therefore, by Lemma 3.12

\[
|R| \leq \left( \frac{f - f^*}{f^*} \right)^2 \leq S^2 \left\| \frac{f - f^*}{f^*} \right\|_2^2 \leq S^2 \left\| \frac{f - f^*}{f^*} \right\|_1 \leq \left\| \frac{f - f^*}{f^*} \right\|_2.
\]

The proof is easily completed using \( \|f - f^*\|_{TV} \leq 2h(f, f^*) \).

Finally, we need one further approximation step.

**Lemma 3.15.** Let \( q \in \mathbb{N} \) and \( \alpha > 0 \). Then for every \( f \in \mathcal{M}_q \) such that \( h(f, f^*) \leq \alpha \), it is possible to choose coefficients \( \eta_i \in \mathbb{R}, \beta_i \in \mathbb{R}^d, \rho_i \in M^d_+ \) for \( i = 1, \ldots, q^* \), and \( \gamma_i \geq 0, \theta_i \in \Theta \) for \( i = 1, \ldots, q \), such that \( \sum_{i=1}^{q^*} \text{rank}[\rho_i] \leq q \wedge dq^* \),

\[
\sum_{i=1}^{q^*} |\eta_i| \leq \frac{1}{e^*} + \frac{1}{\sqrt{c^*} \alpha}, \quad \sum_{i=1}^{q^*} \|\beta_i\| \leq \frac{1}{e^*} + \frac{2T}{\sqrt{c^*} \alpha},
\]

\[
\sum_{i=1}^{q^*} \text{Tr}[\rho_i] \leq \frac{1}{e^*}, \quad \sum_{j=1}^{q} |\gamma_j| \leq \frac{1}{\sqrt{c^*} \alpha} \wedge e^*.
\]
where we have used Lemma 3.12. By Theorem 3.9, we obtain
whenever
θ
h

Therefore,
and

Proof. As
f
∈
Mq,
we can write
f = \sum_{j=1}^{q} \pi_j f_{\theta_j}. Note that by Theorem 3.9,

h(f, f^*) ≥ c^* \sum_{i=1}^{q} \sum_{j: \theta_j \in A_i} \pi_j \|\theta_j - \theta^*_i\|^2.

Therefore, \( h(f, f^*) \leq \alpha \) implies \( \pi_j \|\theta_j - \theta^*_i\|^2 \leq 4\alpha/c^* \) for \( \theta_j \in A_i \). In particular, whenever \( \theta_j \in A_i \), either \( \pi_j \leq 2\sqrt{\alpha/c^*} \) or \( \|\theta_j - \theta^*_i\|^2 \leq 2\sqrt{\alpha/c^*} \). Define

\( J = \bigcup_{i=1}^{q^*} \{ j : \theta_j \in A_i, \|\theta_j - \theta^*_i\|^2 \leq 2\sqrt{\alpha/c^*} \} \).

Taylor expansion gives

\( f_{\theta_j}(x) - f_{\theta^*_i}(x) = (\theta_j - \theta^*_i)^* D_1 f_{\theta^*_i}(x) + \frac{1}{2} (\theta_j - \theta^*_i)^* D_2 f_{\theta^*_i}(x) (\theta_j - \theta^*_i) + R_{ji}(x), \)

where \( |R_{ji}| \leq \frac{1}{6} d^{3/2} \|\theta_j - \theta^*_i\|^3 H_3 \). We can therefore write

\[
\frac{f - f^*}{f^*} = L + \sum_{i=1}^{q} \sum_{j: \theta_j \in A_i} \pi_j R_{ji},
\]

where we have defined

\[
L = \sum_{i=1}^{q^*} \left\{ \left( \sum_{j \in J: \theta_j \in A_i} \pi_j - \pi^*_i \right) \frac{f_{\theta^*_i}}{f^*} + \sum_{j \in J: \theta_j \in A_i} \pi_j (\theta_j - \theta^*_i)^* \frac{D_1 f_{\theta^*_i}}{f^*} \right. \\
+ \left. \frac{1}{2} \sum_{j \in J: \theta_j \in A_i} \pi_j (\theta_j - \theta^*_i)^* \frac{D_2 f_{\theta^*_i}}{f^*} (\theta_j - \theta^*_i) \right\} + \sum_{j \notin J} \pi_j f_{\theta_j}. 
\]

Now note that

\[
\frac{f / f^* - 1}{\sqrt{\chi^2(f\|f^*)}} - \frac{L}{\|L\|_2} \leq \frac{|f / f^* - 1|}{\|f / f^* - 1\|_2} \frac{\|f / f^* - 1 - L\|_2}{\|L\|_2} + \frac{|L / f^* - 1|}{\|L\|_2} \\
\leq \frac{|f / f^* - 1 - L\|_2 S + |f / f^* - 1 - L|}{\|L\|_2},
\]

where we have used Lemma 3.12. By Theorem 3.9, we obtain

\[
\|L\|_2 \geq \|L\| \geq c^* \sum_{i=1}^{q^*} \sum_{j: \theta_j \in A_i} \pi_j \|\theta_j - \theta^*_i\|^2.
\]

Therefore, we can estimate

\[
\frac{|f / f^* - 1 - L|}{\|L\|_2} \leq \frac{d^{3/2} H_3}{3c^*} \sum_{i=1}^{q^*} \sum_{j \in J: \theta_j \in A_i} \pi_j \|\theta_j - \theta^*_i\|^3 \leq \left( \frac{4\alpha}{c^*} \right)^{1/4} \frac{d^{3/2} H_3}{3c^*}
\]
where we have used the definition of $J$. Setting $\ell = L/\|L\|_2$, we obtain
\[
\left| \frac{f/f^* - 1}{\sqrt{\chi^2(f||f^*)}} - \ell \right| \leq \frac{d^{3/2} \sqrt{2}}{3(c^*)^{5/4}} \left( \|H_3\|_2 S + H_3 \right)^{1/4}.
\]
It remains to show that for our choice of $\ell = L/\|L\|_2$, the coefficients $\eta, \beta, \rho, \gamma$ in the statement of the lemma satisfy the desired bounds. These coefficients are
\[
\eta_i = \frac{1}{\|L\|_2} \left( \sum_{j \in J, \theta_j \in A_i} \pi_j - \pi^*_i \right), \quad \beta_i = \frac{1}{\|L\|_2} \sum_{j \in J, \theta_j \in A_i} \pi_j (\theta_j - \theta^*_i),
\]
\[
\rho_i = \frac{1}{2\|L\|_2} \sum_{j \in J, \theta_j \in A_i} \pi_j (\theta_j - \theta^*_i)(\theta_j - \theta^*_i)^*, \quad \gamma_j = \frac{\pi_j 1_{j \not\in J}}{\|L\|_2}.
\]
Clearly $\text{rank}[\rho_i] \leq \#\{j : \theta_j \in A_i\} \land d$, so $\sum_{i=1}^{q^*} \text{rank}[\rho_i] \leq q \land dq^*$. Moreover, we have
\[
\|L\|_2 \geq c^* \left[ \sum_{j \not\in J} \pi_j + \sum_{i=1}^{q^*} \left( \sum_{j : \theta_j \in A_i} \pi_j - \pi^*_i \right) \right.
\]
\[
\left. + \left( \sum_{j : \theta_j \in A_i} \pi_j (\theta_j - \theta^*_i) \right) + \frac{1}{2} \sum_{j : \theta_j \in A_i} \pi_j \|\theta_j - \theta^*_i\|^2 \right]
\]
by Theorem 5.9. It follows that $\sum_{i=1}^{q^*} \text{Tr}[\rho_i] \leq 1/c^*$. Now note that for $j \not\in J$ such that $\theta_j \in A_i$, we have $\|\theta_j - \theta^*_i\|^2 \geq 2\sqrt{\alpha/c^*}$ by construction. Therefore
\[
\|L\|_2 \geq c^* \left[ \sum_{j \not\in J} \pi_j + \sum_{i=1}^{q^*} \frac{1}{2} \sum_{j : \theta_j \in A_i} \pi_j \|\theta_j - \theta^*_i\|^2 \right] \geq \left( \sqrt{c^* c^*} \right) \sum_{j \not\in J} \pi_j.
\]
It follows that $\sum_{j=1}^{q^*} |\gamma_j| \leq 1/(\sqrt{c^* c^*})$. Next, we note that
\[
\sum_{i=1}^{q^*} \sum_{j : \theta_j \in A_i} \pi_j - \pi^*_i \leq \sum_{i=1}^{q^*} \sum_{j : \theta_j \in A_i} \pi_j - \pi^*_i + \sum_{j \not\in J} \pi_j.
\]
Therefore $\sum_{i=1}^{q^*} \|\eta_i\| \leq 1/c^* + 1/\sqrt{c^* c^*}$. Finally, note that
\[
\sum_{i=1}^{q^*} \sum_{j : \theta_j \in A_i} \pi_j (\theta_j - \theta^*_i) \leq \sum_{i=1}^{q^*} \sum_{j : \theta_j \in A_i} \pi_j (\theta_j - \theta^*_i) + 2T \sum_{j \not\in J, \theta_j \in A_0} \pi_j.
\]
Therefore $\sum_{i=1}^{q^*} \|\beta_i\| \leq 1/c^* + 2T/\sqrt{c^* c^*}$. The proof is complete.

We can now complete the proof of Theorem 5.11.

**Proof of Theorem 5.11** Let $\alpha > 0$ be a constant to be chosen later on, and
\[
\mathcal{D}_{q, \alpha} = \{d_f : f \in M_q, f \neq f^*, h(f, f^*) \leq \alpha\}.
\]
Then clearly
\[
\mathcal{N}(\mathcal{D}_q, \delta) \leq \mathcal{N}(\mathcal{D}_{q, \alpha}, \delta) + \mathcal{N}(\mathcal{D}_q \setminus \mathcal{D}_{q, \alpha}, \delta).
\]
We will estimate each term separately.

**Step 1 (the first term).** Define
\[
M_q = \{(m_1, \ldots, m_{q^*}) \in \mathbb{Z}^q_+ : m_1 + \cdots + m_{q^*} = q \land dq^*\}.
\]
For every \( m \in \mathbb{N}_q \), we define the family of functions

\[
\mathcal{L}_{q,m,\alpha} = \left\{ \sum_{i=1}^{q^*} \left\{ \eta_i f_{\gamma_i *}, \beta_i f_{\gamma_i *}, \sum_{j=1}^{m_i} \rho_{ij} D_2 f_{\gamma_i *}, \rho_{ij} \right\} + \sum_{j=1}^{q} \gamma_j f_{\gamma_i *} : (\eta, \beta, \gamma, \theta) \in \mathcal{I}_{q,m,\alpha} \right\},
\]

where

\[
\mathcal{I}_{q,m,\alpha} = \left\{ (\eta, \beta, \rho, \gamma, \theta) \in \mathbb{R}^q \times (\mathbb{R}^d)^{q^*} \times \mathbb{R}^{q^*} \times \cdots \times (\mathbb{R}^d)^{m_1} \times \mathbb{R}^q \times \Theta^q : \right.
\]

\[
\sum_{i=1}^{q^*} |\eta_i| \leq \frac{1}{c^*} + \frac{1}{\sqrt{c^*}}, \quad \sum_{i=1}^{q^*} \|\beta_i\| \leq \frac{1}{c^*} + \frac{2T}{\sqrt{c^*}},
\]

\[
\sum_{i=1}^{q^*} \sum_{j=1}^{m_i} \|\rho_{ij}\|^2 \leq \frac{1}{c^*}, \quad \sum_{j=1}^{q} |\gamma_j| \leq \frac{1}{\sqrt{c^*} \alpha} \wedge c^* \right\}.
\]

Define the family of functions

\[
\mathcal{L}_{q,\alpha} = \bigcup_{m \in \mathbb{N}_q} \mathcal{L}_{q,m,\alpha}
\]

From Lemmas 3.14 and 3.15, we find that for any function \( d \in \mathcal{D}_{q,\alpha} \), there exists a function \( \ell \in \mathcal{L}_{q,\alpha} \) such that (here we use that \( h(f, f^*) \leq \sqrt{2} \) for any \( f \))

\[
|d - \ell| \leq \{4\|S\|^2 S + 2S^2 \} (\alpha \wedge \sqrt{2}) + \frac{d^{3/2} \sqrt{2}}{3(c^*)^{5/4}} \{\|H_3\|_2 S + H_3\} \alpha^{1/4}.
\]

Using \( \alpha \wedge \sqrt{2} \leq 2^{3/8} \alpha^{1/4} \) for all \( \alpha > 0 \), we can estimate

\[
|d - \ell| \leq \alpha^{1/4} U, \quad U = \left( \frac{1 + \|H_3\|_2}{(c^*)^{5/4}} + 8\|S\|^2_4 + 4 \right) d^{3/2} \{S + S^2 + H_3\},
\]

where \( U \in L^2(f^* d\mu) \) by Assumption A. Now note that if \( m_1 \leq \ell \leq m_2 \) for some functions \( m_1, m_2 \) with \( \|m_2 - m_1\|_2 \leq \varepsilon \), then \( m_1 - \alpha^{1/4} U \leq d \leq m_2 + \alpha^{1/4} U \) with \( \|(m_2 + \alpha^{1/4} U) - (m_1 - \alpha^{1/4} U)\|_2 \leq \varepsilon + 2\alpha^{1/4} \|U\|_2 \). Therefore

\[
\mathcal{N}(\mathcal{D}_{q,\alpha}, \varepsilon + 2\alpha^{1/4} \|U\|_2) \leq \mathcal{N}(\mathcal{L}_{q,\alpha}, \varepsilon) \leq \sum_{m \in \mathbb{N}_q} \mathcal{N}(\mathcal{L}_{q,m,\alpha}, \varepsilon) \quad \text{for } \varepsilon > 0.
\]

Of course, we will ultimately choose \( \varepsilon, \alpha \) such that \( \varepsilon + 2\alpha^{1/4} \|U\|_2 = \delta \).

We proceed to estimate the bracketing number \( \mathcal{N}(\mathcal{L}_{q,m,\alpha}, \varepsilon) \). To this end, let \( \ell, \ell' \in \mathcal{L}_{q,m,\alpha} \), where \( \ell \) is defined by the parameters \( (\eta, \beta, \rho, \gamma, \theta) \in \mathcal{I}_{q,m,\alpha} \) and \( \ell' \) is defined by the parameters \( (\eta', \beta', \rho', \gamma', \theta') \in \mathcal{I}_{q,m,\alpha} \). Note that

\[
\sum_{i=1}^{q^*} \sum_{j=1}^{m_i} \left| \rho_{ij} D_2 f_{\gamma_i *}, \rho_{ij} - (\rho'_{ij})^* D_2 f_{\gamma_i *}, \rho_{ij} \right| \leq \frac{2d}{\sqrt{c^*} \alpha} H_2 \sum_{i=1}^{q} \sum_{j=1}^{m_i} \|\rho_{ij} - \rho'_{ij}\|.
\]
We can therefore estimate

\[ |\ell - \ell'| \leq H_0 \sum_{i=1}^{q^*} |\eta_i - \eta_i'| + H_1 \sqrt{d} \sum_{i=1}^{q^*} \|\beta_i - \beta_i'|| + H_0 \sum_{j=1}^{q} |\gamma_j - \gamma_j'| + \frac{\sqrt{d}}{V^\alpha (c^*)^\epsilon}\]

where we have used that \( |f_\theta - f_{\theta'}|/f^* \leq \|\theta - \theta'|| H_1\sqrt{d} \) by Taylor expansion.

Therefore, writing \( V = (H_0 + H_1 + H_2) d\sqrt{d\epsilon} \), we have

\[ |\ell - \ell'| \leq V \|((\eta, \beta, \rho, \gamma, \theta) - (\eta', \beta', \rho', \gamma', \theta'))\|_{q,m,\alpha} \]

where \( \|\cdot\|_{q,m,\alpha} \) is the norm on \( \mathbb{R}^{(1+d)q^*+d(q^*dq^*)+(1+d)q} \) defined by

\[ \|((\eta, \beta, \rho, \gamma, \theta))\|_{q,m,\alpha} = \sum_{i=1}^{q^*} |\eta_i| + \sum_{i=1}^{q^*} \|\beta_i\| + \sum_{j=1}^{q} |\gamma_j| + \frac{1}{\sqrt{c^*} (c^*)^\epsilon}\max_{j=1,\ldots,q^*} \|\theta_j\| + \frac{2}{\sqrt{c^*}} \left[ \sum_{j=1}^{q^*} \sum_{i=1}^{m_i} \|\rho_{ij}\|^2 \right]^{1/2}. \]

Note that if \( \|((\eta, \beta, \rho, \gamma, \theta) - (\eta', \beta', \rho', \gamma', \theta'))\|_{q,m,\alpha} \leq \epsilon' \), then we obtain a bracket \( \ell' - \epsilon' V \leq \ell \leq \ell' + \epsilon' V \) of size \( \|\ell' + \epsilon' V\|_2 = 2\epsilon' \|V\|_2 \). Therefore, if we denote by \( N(\mathcal{L}_{q,m,\alpha}, \|\cdot\|_{q,m,\alpha}, \epsilon') \) the cardinality of the largest packing of \( \mathcal{L}_{q,m,\alpha} \) by \( \epsilon' \)-separated points with respect to the \( \|\cdot\|_{q,m,\alpha} \)-norm, then

\[ N(\mathcal{L}_{q,m,\alpha}, \epsilon) \leq N(\mathcal{L}_{q,m,\alpha}, \|\cdot\|_{q,m,\alpha}, \epsilon/2\|V\|_2) \quad \text{for } \epsilon > 0. \]

But note that, by construction, \( \mathcal{L}_{q,m,\alpha} \) is included in a \( \|\cdot\|_{q,m,\alpha} \)-ball of radius not exceeding \((6+3T)/\sqrt{c^* (c^*)^\epsilon}\). Therefore, using the standard fact that the packing number of the \( r \)-ball \( B(r) = \{x \in B : \|x\| \leq r\} \) in any \( n \)-dimensional normed space \( (B, \|\cdot\|) \) satisfies \( N(B(r), \|\cdot\|, \epsilon) \leq \left( \frac{2r}{\epsilon} \right)^n \), we can estimate

\[ N(\mathcal{L}_{q,m,\alpha}, \epsilon) \leq \left( \frac{4\|V\|_2(6+3T)/\sqrt{c^* (c^*)^\epsilon} + \epsilon}{\epsilon} \right)^{(1+d)q^*+d(q^*dq^*)+(1+d)q}. \]

In particular, if \( \epsilon \leq 1 \) and \( \alpha \leq c^* \), then

\[ N(\mathcal{L}_{q,m,\alpha}, \epsilon) \leq \left( \frac{24 + 12T}{\epsilon\sqrt{c^*} + \sqrt{c^*}} \right)^{3(1+d)q}. \]

Finally, note that the cardinality of \( \mathbb{M}_q \) can be estimated as

\[ \#\mathbb{M}_q = \frac{q^* + q \wedge dq^* - 1}{q \wedge dq^*} \leq e^{q^*dq^*} \left( \frac{q^* + q \wedge dq^* - 1}{q \wedge dq^*} \right)^{q^*dq^*} \leq 2^{4q}, \]

where we have used that \( q \geq q^* \). We therefore obtain

\[ N(\mathcal{D}_{q,\alpha}, \delta) \leq \sum_{m \in \mathbb{M}_q} N(\mathcal{L}_{q,m,\alpha}, \delta - 2\alpha^{1/4}\|U\|_2) \leq \left( \frac{24(2 + T)/\|V\|_2/\sqrt{c^*} + \sqrt{c^*}}{\delta - 2\alpha^{1/4}\|U\|_2/\sqrt{c^*}} \right)^{3(1+d)q}, \]

whenever \( \delta \leq 1 \) and \( \alpha \leq (\delta/2\|U\|_2)^4 \wedge c^* \).
Step 2 (the second term). For \( f, f' \in \mathcal{M}_q \) with \( h(f, f^*) > \alpha \) and \( h(f', f^*) > \alpha \),
\[
|d_f - d'_f| = \frac{|(\sqrt{f/f^*} - 1)\|\sqrt{f'/f'} - 1\|_2 - (\sqrt{f'/f'} - 1)\|\sqrt{f/f^*} - 1\|_2|}{h(f, f^*) h(f', f^*)} \\
\leq \frac{\|\sqrt{f'/f'} - \sqrt{f/f^*}\|_2 (\sqrt{f'/f'} - 1) + \sqrt{2} (\sqrt{f/f^*} - \sqrt{f'/f'})}{\alpha^2},
\]
where we have used that \( h(f, f^*) \leq \sqrt{2} \) for any \( f \). Now note that
\[
|\sqrt{a} - \sqrt{b}| \leq |\sqrt{a} - \sqrt{b}| (\sqrt{a} + \sqrt{b}) = |a - b|
\]
for any \( a, b \geq 0 \). We can therefore estimate
\[
|d_f - d'_f| \leq \frac{\|(f - f')/f^*\|_1^{1/2} (\sqrt{H_0} + 1) + \sqrt{2} |(f - f')/f^*|^{1/2}}{\alpha^2},
\]
where we have used that \( |\sqrt{f/f^*} - 1| \leq \sqrt{H_0} + 1 \) for any \( f \in \mathcal{M} \). Now note that if we write \( f = \sum_q \pi_i f_q \) and \( f' = \sum_q \pi'_i f'_q \), then we can estimate
\[
\left| \frac{f - f'}{f^*} \right| \leq H_0 \sum_{i=1}^q |\pi_i - \pi'_i| + H_1 \sqrt{\alpha} \max_{i=1,...,q} \|\theta_i - \theta'_i\|.
\]
Defining
\[
W = (\sqrt{H_0} + 1)\|H_0 + H_1\sqrt{\alpha}\|_1^{1/2} + \sqrt{2} (H_0 + H_1\sqrt{\alpha})^{1/2},
\]
we obtain
\[
|d_f - d'_f| \leq \frac{W}{\alpha^2} \|(\pi, \theta) - (\pi', \theta')\|_q^{1/2}, \quad \|(\pi, \theta)\|_q = \sum_{i=1}^q |\pi_i| + \max_{i=1,...,q} \|\theta_i\|
\]
(clearly \( \| \cdot \|_q \) defines a norm on \( \mathbb{R}^{(d+1)q} \)). Now note that if \( \|(\pi, \theta) - (\pi', \theta')\|_q \leq \epsilon \),
then we obtain a bracket \( d_f' - \epsilon^{1/2} W/\alpha^2 \leq d_f \leq d_f' + \epsilon^{1/2} W/\alpha^2 \) of size \( \|(d_f' + \epsilon^{1/2} W/\alpha^2) - (d_f' - \epsilon^{1/2} W/\alpha^2)\|_2 = 2\epsilon^{1/2} W_2/\alpha^2 \). Therefore
\[
\mathcal{N}(D_q \setminus D_{q, \alpha, \delta}) \leq \mathcal{N}(\Delta_q \times \Theta_q, \| \cdot \|_q, \alpha^4 \delta^2/4 W_2^{1/2}),
\]
where we have defined the simplex \( \Delta_q = \{ \pi \in \mathbb{R}_+^q : \sum_{i=1}^q \pi_i = 1 \} \). We can now estimate the quantity on the right hand side of this expression as before, giving
\[
\mathcal{N}(D_q \setminus D_{q, \alpha, \delta}) \leq \left( \frac{8(1 + T) W_2^{1/2} + (c^*)^4}{\alpha^4 \delta^2} \right)^{(d+1)q}
\]
for \( \delta \leq 1 \) and \( \alpha \leq c^* \).

End of proof. Choose \( \alpha = (\delta/4 \|U\|_2^4) \). Collecting the various estimates above,
we find that for \( \delta \leq 1 \wedge 4(c^*)^{1/4} \) (as \( \|U\|_2 \geq \|S\|_1 \geq 1 \) by Lemma 3.12)
\[
\mathcal{N}(D_q, \delta) \leq \left( \frac{768(2 + T)\|U\|_2^3 / \sqrt{c^*} + 32\|U\|_2 \sqrt{c^*}}{\delta^3} \right)^3 (d+1)^q \\
+ \left( \frac{4^{18}(1 + T)\|U\|_2^6 \|W\|_2^2 + 4^{16}\|U\|_2^6 (c^*)^4}{\delta^{18}} \right)^{(d+1)q} \\
\leq \left( \frac{c^0 (T \vee 1)^{1/6} (\|U\|_2 \vee \|V\|_2 \vee \|W\|_2)}{\delta} \right)^{18(d+1)^q}
\]
where \( c_0^* = 12(c^*)^{-1/12} + 2(c^*)^{1/12} + 4(c^*)^{4/18} + 8 \). It follows that

\[
N(D_q, \delta) \leq \left( \frac{C^*(T \lor 1)^{1/6}}{\delta} \left( \|H_0\|_4^4 \lor \|H_1\|_4^4 \lor \|H_2\|_4^4 \lor \|H_3\|_2^2 \right) \right) \frac{18(q+1)d}{d+1} \]

for all \( \delta \leq \delta^* \), where \( C^* \) and \( \delta^* \) are constants that depend only on \( c^*, d, \) and \( q^* \). This establishes the estimate given in the statement of the Theorem. The proof of the second half of the Theorem follows from Corollary 3.13 and \( \|H_0\|_4 \geq 1 \). □

Acknowledgment. The authors would like to thank Jean Bretagnolle for providing an enlightening counterexample that guided some of our proofs.

References

1. C. D. Aliprantis and K. C. Border, *Infinite dimensional analysis*, third ed., Springer, 2006.
2. P. Assouad, “Plongements lipschitziens dans \( \mathbb{R}^n \),” Bull. Soc. Math. France 111 (1983), 429–448.
3. R. Bléi, F. Gao, and W. V. Li, “Metric entropy of high dimensional distributions,” Proc. Amer. Math. Soc. 135 (2007), 4009–4018 (electronic).
4. B. Carl, I. Kyrezi, and A. Pajor, “Metric entropy of convex hulls in Banach spaces,” J. London Math. Soc. (2) 60 (1999), 871–896.
5. G. Ciuperca, “Likelihood ratio statistic for exponential mixtures,” Ann. Inst. Statist. Math. 54 (2002), 585–594.
6. F. Gao, “Metric entropy of convex hulls,” Israel J. Math. 123 (2001), 359–364.
7. F. Gao, W. V. Li, and J. A. Wellner, “How many Laplace transforms of probability measures are there?” Proc. Amer. Math. Soc. 138 (2010), 4331–4344.
8. E. Gassiat and J. Rousseau, “On the asymptotic behaviour of the posterior distribution in hidden Markov models,” 2012, preprint.
9. E. Gassiat and R. van Handel, “Consistent order estimation and minimal penalties,” 2012, preprint.
10. C. R. Genovese and L. Wasserman, “Rates of convergence for the Gaussian mixture sieve,” Ann. Statist. 28 (2000), 1105–1127.
11. S. Ghosal and A. W. van der Vaart, “Entropies and rates of convergence for maximum likelihood and Bayes estimation for mixtures of normal densities,” Ann. Statist. 29 (2001), 1233–1263.
12. J. Heinonen, “Lectures on analysis on metric spaces,” Springer, 2001.
13. A. N. Kolmogorov and V. M. Tihomirov, “\( \varepsilon \)-entropy and \( \varepsilon \)-capacity of sets in function spaces,” Uspehi Mat. Nauk 14 (1959), 3–86.
14. X. Liu and Y. Shao, “Asymptotics for likelihood ratio tests under loss of identifiability,” Ann. Statist. 31 (2003), 807–832.
15. A. W. van der Vaart and J. A. Wellner, “Weak convergence and empirical processes,” Springer Series in Statistics, Springer-Verlag, New York, 1996.
16. D. Williams, “Probability with martingales,” Cambridge University Press, 1991.