Pathwise fluctuations of likelihood ratios and consistent order estimation

Elisabeth Gassiat, Ramon van Handel

PATHWISE FLUCTUATIONS OF LIKELIHOOD RATIOS AND CONSISTENT ORDER ESTIMATION

BY ELISABETH GASSIAT AND RAMON VAN HANDEL

Université Paris-Sud and Princeton University

Consider an i.i.d. sequence of random variables whose distribution $f^\star$ lies in one of a nested family of models $\mathcal{M}_q$, $q \geq 1$. We obtain a sharp characterization of the pathwise fluctuations of the generalized likelihood ratio statistic under entropy assumptions on the model classes $\mathcal{M}_q$. Moreover, we develop a technique to obtain local entropy bounds from global entropy computations, so that these results can be applied in models with non-regular geometric structure. Finally, the results are applied to prove strong consistency and to identify minimal penalties for penalized likelihood order estimators in the absence of prior upper bounds on the model order and the underlying parameter set. Location mixture models, which possess a notoriously complicated geometric structure, are used as a case study throughout the paper, and the requisite geometric analysis is of independent interest.

CONTENTS

AMS 2000 subject classifications: 62G20, 60F15, 60F10, 41A46

Keywords and phrases: uniform law of iterated logarithm, local entropy with bracketing, entropy of finite mixtures, penalized likelihood order estimation, strong consistency

1. Introduction. Let $(X_k)_{k\geq 1}$ be a sequence of random variables whose distribution $f^\star$ lies in one of a nested family of models $(\mathcal{M}_q)_{q\geq 1}$, indexed (and ordered) by the integers. We define the model order as the smallest index $q^\star$ such that the true distribution $f^\star$ lies in the corresponding model class. The model order typically determines the most parsimonious representation of the true distribution of the underlying model (for example, it might determine the parametrization of the model which has the smallest possible dimension). On the other hand, the model order often has a concrete interpretation in terms of the modelling of the underlying phenomenon (for example, the estimation of the number of clusters in a data set, or the number of regimes in an economic time series). Therefore, the problem of estimating the model order from observed data is of significant practical, as well as theoretical, interest.

Of course, a satisfactory solution to this problem must provide an estimation method that does not assume prior knowledge on the underlying unknown distribution $f^\star$. In particular, prior bounds on model order and on parameter sets should be avoided. Yet, in this light, even one of the most widely used model selection criteria—the Bayesian Information Criterion (BIC) of Schwarz—is poorly understood. The chief motivation for the use of BIC (as opposed to other model selection criteria, such as Akaike's Information Criterion) is that it is expected to yield a strongly consistent estimator of the model order. However, as is pointed out by Csiszár and Shields [9], almost all existing consistency proofs assume a prior upper bound on the order as well as compactness of the underlying parameter sets. This is hardly satisfactory from the theoretical point of view, and provides little confidence in the basic motivation for this method. More delicate questions, such as the minimal penalty that yields a strongly consistent order estimator in the absence of a prior bound on the order, also remain open (the problem of identifying the minimal penalty, which minimizes the probability of underestimating the model order, was also raised in [9]).

Characterizing strong consistency of penalized likelihood order estimators hinges on a precise understanding of the pathwise fluctuations of the likelihood ratio statistic

$$\sup_{f\in\mathcal{M}_q} \ell_n(f) - \sup_{f\in\mathcal{M}_{q^\star}} \ell_n(f),$$

as $n\to\infty$, *uniformly* in the model order $q > q^\star$ (here $\ell_n(f)$ is the likelihood of $(X_k)_{k\leq n}$ under the distribution $f\in\mathcal{M}_q$). When there is a known upper bound on the order $q^\star \leq q_{\max} < \infty$ and the model classes $\mathcal{M}_q$ are parametrized by a compact subset of Euclidean space, an upper bound on the pathwise fluctuations can be obtained by classical parametric methods: Taylor expansion of the likelihood and an application of a law of iterated logarithm. This approach forms the basis for most consistency proofs for penalized likelihood order estimators in the litera-

ture, for example [14, 22, 10, 16, 6]. However, such techniques fail in the absence of a prior upper bound: even though each model class $\mathcal{M}_q$ is finite dimensional, the full model $\mathcal{M} = \bigcup_q \mathcal{M}_q$ is infinite dimensional and, as such, the problem in the absence of a prior upper bound is inherently nonparametric.[1] When the model classes $\mathcal{M}_q$ are noncompact, one must introduce sieves $\mathcal{M}_q^n \subset \mathcal{M}_q^{n+1} \subset \cdots \subset \mathcal{M}_q$ which complicate the problem further (in this case even the parametric theory remains poorly understood, see [15, 3, 19]). An entirely different approach based on universal coding theory [10, 12, 5, 7] yields pathwise upper bounds on the likelihood ratio statistic that do not require prior bounds on the order or compactness of the models. However, these bounds are far from tight and cannot even establish consistency of BIC, let alone smaller penalties (this appears to be a fundamental limitation of this approach due to Rissanen's theorem, see [24, 2]).

To our knowledge, the only setting in which the pathwise fluctuations of the likelihood ratio statistic has been studied in the absence of a prior bound on the order is that of higher-order Markov chains, where Csiszár and Shields [9, 8] proved consistency of BIC. The proofs in [9, 8] use delicate estimates specific to Markov chains, and do not yield minimal penalties. However, it was shown in [27] that a sharp bound can be obtained in the Markov chain case using techniques from empirical process theory, the main difficulty being the dependence structure of Markov chains.

The aim of this paper is to obtain generally applicable upper and lower bounds on the pathwise fluctuations of the likelihood ratio statistic uniformly in the model order $q > q^\star$, in the case of i.i.d. observations $(X_k)_{k\geq 1}$, without a prior bound on the model order and in possibly noncompact parameter spaces. We use empirical process methods as in [27], but the difficulties to be surmounted in the present setting are of a different nature. Though the Markov chain models in [9, 8, 27] suffer from a lack of independence, geometrically these models are exceedingly simple: the family of qth-order Markov chains endowed with the Hellinger distance is simply a Euclidean ball when viewed in the appropriate parametrization. In contrast, in general order estimation problems, one is often faced with model classes that are geometrically very complex. An important case study that will be considered in this paper are finite mixture models (widely used in practice for clustering), which possess a notoriously complicated non-regular geometry. To obtain sharp bounds in such models, we will develop tools that can be used to obtain local and weighted

[1] One of the key issues in this setting is to understand the dependence of the fluctuations of the likelihood ratio statistic on the dimension of the model classes $\mathcal{M}_q$. However, one of the main results of this paper shows that for regular parametric models, the fluctuations of the likelihood ratio statistic uniformly in $q > q^\star$ are dimension independent when a prior upper bound is assumed (cf. Remark 2.7), which is certainly not the case in the absence of a prior upper bound. Therefore, we find that the pathwise fluctuations of the likelihood ratio statistic with and without a prior upper bound are qualitatively different.

entropy bounds, required for our pathwise fluctuation theorems, in models with non-regular geometric structure. These results are of independent interest: we are not aware of any existing local entropy results for models that possess a nontrivial geometric structure (the difficulty of obtaining local entropy bounds for mixture models is noted, for example, in [13, 21]). Finally, we will apply our results to establish strong consistency of BIC and to identify minimal penalties for order estimation for general model classes, in the absence of prior bounds on the order and the underlying parameter set.

The remainder of this paper is organized as follows. Section 2 introduces the general model under consideration, and states our results on the pathwise fluctuations of the likelihood ratio statistic. Section 3 states our general results on local and weighted entropies, and considers also the special case of mixture models. Section 4 derives the consequences for order estimation. Proofs are given in section 5.

2. Pathwise fluctuations of the likelihood ratio statistic.

2.1. *Basic setting and notation.* Let $(E, \mathcal{E}, \mu)$ be a measure space. For each $q, n \geq 1$, let $\mathcal{M}_q^n$ be a given family of strictly positive probability densities with respect to μ (that is, we assume that $\int f d\mu = 1$ and that $f > 0$ μ-a.e. for every $f \in \mathcal{M}_q^n$). Moreover, we assume that $(\mathcal{M}_q^n)_{q,n\geq 1}$ is a nested family of models in the sense that $\mathcal{M}_q^n \subseteq \mathcal{M}_{q+1}^n$ and $\mathcal{M}_q^n \subseteq \mathcal{M}_q^{n+1}$ for all $q, n \geq 1$. Let $\mathcal{M}_q = \bigcup_n \mathcal{M}_q^n$, $\mathcal{M}^n = \bigcup_q \mathcal{M}_q^n$, $\mathcal{M} = \bigcup_{q,n} \mathcal{M}_q^n$.

Consider an i.i.d. sequence of E-valued random variables $(X_k)_{k\geq 1}$ whose common distribution under the measure $\mathbf{P}^\star$ is $f^\star d\mu$, where $f^\star \in \mathcal{M}_{q^\star} \setminus \operatorname{cl} \mathcal{M}_{q^\star-1}$ for some $q^\star \geq 1$ (here $\operatorname{cl} \mathcal{M}_q$ denotes the $L^1(d\mu)$-closure of $\mathcal{M}_q$). The index $q^\star$ is called the *model order*. Let us define the log-likelihood function

$$\ell_n(f) = \sum_{i=1}^n \log f(X_i), \qquad f \in \mathcal{M}.$$

Evidently $\ell_n(f)$ is the log-likelihood of the i.i.d. sequence $(X_k)_{k\leq n}$ when $X_k \sim f d\mu$. Our aim is to study the pathwise fluctuations of the likelihood ratio statistic

$$\sup_{f\in\mathcal{M}_q^n} \ell_n(f) - \sup_{f\in\mathcal{M}_{q^\star}^n} \ell_n(f)$$

as $n \to \infty$, *uniformly* over the order parameter $q \geq q^\star$. Pathwise upper and lower bounds on the likelihood ratio statistic are the key ingredient in the study of strong consistency of penalized likelihood order estimators (see section 4).

EXAMPLE 2.1 (Location mixtures). The guiding example for our theory, the case of location mixtures, will be studied in detail in sections 3.2 and 4.2 below. We presently introduce this example in order to clarify our basic setup.

Let $E = \mathbb{R}^d$ (with its Borel σ-field $\mathcal{E}$) and let μ be the Lebesgue measure on $\mathbb{R}^d$. We fix a strictly positive probability density f_0 with respect to μ, and define $f_\theta(x) = f_0(x-\theta)$ for $x, \theta \in \mathbb{R}^d$. Fix a sequence $T(n) \uparrow \infty$ and define

$$\mathcal{M}_q^n = \left\{ \sum_{i=1}^q \pi_i f_{\theta_i} : \pi_i \geq 0,\ \sum_{i=1}^q \pi_i = 1,\ \|\theta_i\| \leq T(n) \right\}.$$

Then $\mathcal{M}_q$ is the family of all q-component mixtures of translates of the density f_0, while $\mathcal{M}_q^n$ is the subset of the mixtures $\mathcal{M}_q$ whose translation parameters $(\theta_i)_{i=1,\ldots,q}$ are restricted to a ball of radius $T(n)$. The number of components $q^\star$ of the true mixture $f^\star \in \mathcal{M}$ can be estimated from observations using the order estimator

$$\hat{q}_n = \operatorname*{argmax}_{q \geq 1} \left\{ \sup_{f \in \mathcal{M}_q^n} \ell_n(f) - \mathrm{pen}(n,q) \right\}.$$

Pathwise control of the likelihood ratio statistic allows us to identify what penalties $\mathrm{pen}(n,q)$ and cutoff sequences $T(n)$ yield strong consistency of $\hat{q}_n$ (cf. section 4.2).

REMARK 2.2. To avoid measurability problems and other technical complications, we employ throughout this paper the simplifying convention that all uncountable suprema (such as $\sup_{f \in \mathcal{M}_q^n} \ell_n(f)$) are interpreted as essential suprema with respect to the measure $\mathbf{P}^\star$. In the majority of applications the model classes $\mathcal{M}_q^n$ will be separable, in which case the supremum and essential supremum coincide.

In the sequel, we will denote by $\|\cdot\|_p$ the $L^p(f^\star d\mu)$-norm, that is, $\|g\|_p^p = \int |g(x)|^p f^\star(x)\mu(dx)$, and we denote by $\langle f, g\rangle = \int f(x)g(x)f^\star(x)\mu(dx)$ the Hilbert space inner product in $L^2(f^\star d\mu)$. Define the Hellinger distance

$$h(f,g)^2 = \int (\sqrt{f} - \sqrt{g})^2 d\mu, \qquad f, g \in \mathcal{M}.$$

It is easily seen that $h(f, f^\star) = \|\sqrt{f/f^\star} - 1\|_2$. Finally, we will denote by $\mathcal{N}(\mathcal{Q}, \delta)$ for any class of functions $\mathcal{Q}$ and $\delta > 0$ the minimal number of brackets of $L^2(f^\star d\mu)$-width δ needed to cover $\mathcal{Q}$: that is, $\mathcal{N}(\mathcal{Q}, \delta)$ is the smallest cardinality N of a collection of pairs of functions $\{g_i^L, g_i^U\}_{i=1,\ldots,N}$ such that $\max_{i \leq N} \|g_i^U - g_i^L\|_2 \leq \delta$ and for every $g \in \mathcal{Q}$ we have $g_i^L \leq g \leq g_i^U$ pointwise for some $i \leq N$.

2.2. *Upper bound.* We aim to obtain a pathwise upper bound on the likelihood ratio statistic that holds *uniformly* in $q > q^\star$. To this end, define for $q, n \geq 1$ and $\varepsilon > 0$ the Hellinger ball

$$\mathcal{H}_q^n(\varepsilon) = \{\sqrt{f/f^\star} : f \in \mathcal{M}_q^n,\ h(f, f^\star) \leq \varepsilon\}.$$

Note that the definition of $\mathcal{H}_q^n(\varepsilon)$ depends on $f^\star$ (which is fixed throughout the paper). The following result shows that the geometry of the Hellinger balls $\mathcal{H}_q^n(\varepsilon)$ controls the pathwise fluctuations of the likelihood ratio statistic.

THEOREM 2.3. *Suppose that for all n sufficiently large, we have*

$$\mathcal{N}(\mathcal{H}_q^n(\varepsilon),\delta) \le \left(\frac{K(n)\varepsilon}{\delta}\right)^{\eta(q)}$$

for all $q \ge q^\star$ and $\delta \le \varepsilon$, with $K(n) \ge 1$ and $\eta(q) \ge q$ increasing functions. Then

$$\limsup_{n\to\infty} \frac{1}{\log K(2n) \vee \log\log n} \sup_{q\ge q^\star} \frac{1}{\eta(q)} \left\{ \sup_{f\in\mathcal{M}_q^n} \ell_n(f) - \sup_{f\in\mathcal{M}_{q^\star}^n} \ell_n(f) \right\} \le C$$

$\mathbf{P}^\star$*-a.s., where $C > 0$ is a universal constant.*

The proof of Theorem 2.3 is given in section 5.1 below.

The assumption of Theorem 2.3 on the entropy of the Hellinger balls $\mathcal{H}_q^n(\varepsilon)$ states, roughly speaking, that the class of densities $\mathcal{M}_q^n$ endowed with the Hellinger distance has the same metric structure as a Euclidean ball of dimension $\eta(q)$ and radius of order $K(n)$, at least locally in a neighborhood of the true density $f^\star$. The effective dimension $\eta(q)$ controls the fluctuations of the likelihood ratio statistic as a function of the model order, while the effective radius $K(n)$ controls the fluctuations as a function of time up to a minimal rate of order $\log\log n$. In the following section we will see that the minimal $\log\log n$ rate is indeed optimal.

Let us note that the geometric structure required by Theorem 2.3 is far from obvious in many cases of practical interest. For example, in the case of finite mixtures, the geometry of the parameter sets corresponding to Hellinger balls is notoriously complex and highly non-regular, but we will nonetheless verify the assumption of Theorem 2.3 (see section 3.2). In order to apply Theorem 2.3 in such cases, we therefore need to develop tools to establish local entropy bounds in models that possess nontrivial geometric structure. Section 3 below is devoted to this problem.

2.3. *Lower bound.* Throughout this section, we specialize to the case that $\mathcal{M}_q^n = \mathcal{M}_q$ does not depend on n (this implies essentially that $\mathcal{M}_q$ is compact). In this setting, Theorem 2.3 yields an upper bound of order $\log\log n$ on the pathwise fluctuations of the likelihood ratio statistic. The aim of this section is to obtain a matching lower bound of order $\log\log n$, which shows that the minimal rate in Theorem 2.3 is essentially optimal. For the purposes of a lower bound, the uniformity in q is irrelevant, so that it suffices to restrict attention to some fixed $q > q^\star$. We will in fact

obtain a much stronger result in this case, which completely characterizes the precise pathwise asymptotics of the likelihood ratio statistic for fixed q in sufficiently smooth families.

The geometric structure required in the present section is somewhat different than that of Theorem 2.3. Instead of Hellinger balls, we consider the classes of weighted densities $\mathcal{D}_q = \{d_f : f \in \mathcal{M}_q,\ f \neq f^\star\}$ and $\mathcal{D} = \bigcup_q \mathcal{D}_q$, where

$$d_f = \frac{\sqrt{f/f^\star} - 1}{h(f, f^\star)}, \qquad f \in \mathcal{M}, \quad f \neq f^\star.$$

In addition, we define for $\varepsilon > 0$ and $q \geq 1$ the local weighted classes

$$\mathcal{D}_q(\varepsilon) = \{d_f : f \in \mathcal{M}_q,\ 0 < h(f, f^\star) \leq \varepsilon\}, \qquad \bar{\mathcal{D}}_q = \bigcap_{\varepsilon > 0} \operatorname{cl} \mathcal{D}_q(\varepsilon),$$

where the closure $\operatorname{cl} \mathcal{D}_q(\varepsilon)$ is in $L^2(f^\star d\mu)$. Evidently $\bar{\mathcal{D}}_q$ is the set of all possible limit points of d_f as $h(f, f^\star) \to 0$ in $\mathcal{M}_q$. If the neighborhoods of $\bar{\mathcal{D}}_q$ are sufficiently rich, such limits can be taken along a continuous path in the following sense.

DEFINITION 2.4. A point $d \in \bar{\mathcal{D}}_q$ is called *continuously accessible* if there is a path $(f_t)_{t \in]0,1]} \subset \mathcal{M}_q \backslash \{f^\star\}$ such that the map $t \mapsto h(f_t, f^\star)$ is continuous, $h(f_t, f^\star) \to 0$ as $t \to 0$, and $d_{f_t} \to d$ in $L^2(f^\star d\mu)$ as $t \to 0$. The subset of all continuously accessible points in $\bar{\mathcal{D}}_q$ will be denoted as $\bar{\mathcal{D}}_q^c$.

We can now formulate the main result of this section.

THEOREM 2.5. *Let $q^\star \leq p < q$. Assume that*

$$\int_0^1 \sqrt{\log \mathcal{N}(\mathcal{D}_q, u)}\, du < \infty,$$

and that $|d| \leq D$ for all $d \in \mathcal{D}_q$ with $D \in L^{2+\alpha}(f^\star d\mu)$ for some $\alpha > 0$. Then

$$\limsup_{n \to \infty} \frac{1}{\log \log n} \left\{ \sup_{f \in \mathcal{M}_q} \ell_n(f) - \sup_{f \in \mathcal{M}_p} \ell_n(f) \right\} \geq \sup_{g \in L_0^2(f^\star d\mu)} \left\{ \sup_{f \in \bar{\mathcal{D}}_q^c} (\langle f, g \rangle)_+^2 - \sup_{f \in \bar{\mathcal{D}}_p} (\langle f, g \rangle)_+^2 \right\} \quad \mathbf{P}^\star\text{-a.s.},$$

as well as

$$\limsup_{n\to\infty} \frac{1}{\log\log n} \left\{ \sup_{f\in\mathcal{M}_q} \ell_n(f) - \sup_{f\in\mathcal{M}_p} \ell_n(f) \right\} \le \sup_{g\in L_0^2(f^\star d\mu)} \left\{ \sup_{f\in\bar{\mathcal{D}}_q} (\langle f,g\rangle)_+^2 - \sup_{f\in\bar{\mathcal{D}}_p^c} (\langle f,g\rangle)_+^2 \right\} \quad \mathbf{P}^\star\text{-a.s.},$$

where $L_0^2(f^\star d\mu) = \{g \in L^2(f^\star d\mu) : \|g\|_2 \le 1,\ \langle 1, g\rangle = 0\}$.

Only the first (lower bound) part of the theorem is needed to conclude optimality of the minimal $\log\log n$ rate in Theorem 2.3. Indeed, we will obtain as a corollary the following lower bound counterpart to Theorem 2.3.

COROLLARY 2.6. *Suppose there exists $q > q^\star$ such that the following hold.*

1. *There is an envelope function $D : E \to \mathbb{R}$ such that $|d| \le D$ for all $d \in \mathcal{D}_q$ and $D \in L^{2+\alpha}(f^\star d\mu)$ for some $\alpha > 0$. Moreover, $\int_0^1 \sqrt{\log \mathcal{N}(\mathcal{D}_q, u)}\, du < \infty$.*
2. *$\bar{\mathcal{D}}_q^c \backslash \bar{\mathcal{D}}_{q^\star}$ is nonempty.*

Let $\eta(q) > 0$ be an arbitrary positive function. Then

$$\limsup_{n\to\infty} \frac{1}{\log\log n} \sup_{q\ge q^\star} \frac{1}{\eta(q)} \left\{ \sup_{f\in\mathcal{M}_q} \ell_n(f) - \sup_{f\in\mathcal{M}_{q^\star}} \ell_n(f) \right\} \ge C_0$$

$\mathbf{P}^\star$*-a.s., where $C_0 > 0$ is nonrandom but may depend on $f^\star$ and η.*

The proofs of Theorem 2.5 and Corollary 2.6 are given in section 5.2 below.

The fact that the geometric assumptions in Theorem 2.5 and Corollary 2.6 are expressed in terms of weighted classes is not surprising, as the sharp asymptotic expression provided by Theorem 2.5 for the pathwise fluctuations of the likelihood ratio statistic are expressed in terms of a variational problem on the weighted classes. Nonetheless, we are naturally led to ask whether there is any relation between the geometric assumptions imposed in the upper bound Theorem 2.3 and the lower bound Theorem 2.5, which appear to be quite different at first sight. In section 3, we will show that the global entropy of the weighted class is closely related to local entropy, so that the geometric assumptions for the upper and lower bounds are not too far apart. Beside the fundamental value of this observation, the relation between global and local entropies will prove to be an essential tool in order to verify these geometric assumptions in models with a complicated geometry, such as finite mixture models.

REMARK 2.7. When $\bar{\mathcal{D}}_q$ and $\bar{\mathcal{D}}_p$ each contain an $L^2(f^\star d\mu)$-dense subset of continuously accessible points (which is typically the case in sufficiently smooth models), then Theorem 2.5 provides the exact characterization

$$\limsup_{n\to\infty} \frac{1}{\log\log n} \left\{ \sup_{f\in\mathcal{M}_q} \ell_n(f) - \sup_{f\in\mathcal{M}_p} \ell_n(f) \right\} = \sup_{g\in L_0^2(f^\star d\mu)} \left\{ \sup_{f\in\bar{\mathcal{D}}_q} (\langle f,g\rangle)_+^2 - \sup_{f\in\bar{\mathcal{D}}_p} (\langle f,g\rangle)_+^2 \right\} \quad \mathbf{P}^\star\text{-a.s.}$$

Beside its intrinsic interest, this result has a surprising consequence. In the case that $\mathcal{M}_q$ and $\mathcal{M}_p$ are regular parametric models with $\dim(\mathcal{M}_q) > \dim(\mathcal{M}_p)$, one can choose $g \in \bar{\mathcal{D}}_q$ which is orthogonal to $\bar{\mathcal{D}}_p$. As $\bar{\mathcal{D}}_q, \bar{\mathcal{D}}_p \subseteq L_0^2(f^\star d\mu)$ (see the proof of Corollary 2.6), it follows easily that in this case the right-hand side of the previous equation display is precisely equal to 1. In particular, we obtain the curious conclusion that in regular parametric models, the magnitude of the fluctuations of the likelihood ratio statistic does not depend on the dimensions $\dim(\mathcal{M}_q)$ and $\dim(\mathcal{M}_p)$. In contrast, it is well known that in regular parametric models, the likelihood ratio statistic itself converges weakly to a chi-square distribution with $\dim(\mathcal{M}_q) - \dim(\mathcal{M}_p)$. degrees of freedom, so the tails of the distribution of the likelihood ratio statistic do in fact depend strongly on the dimensions $\dim(\mathcal{M}_q)$ and $\dim(\mathcal{M}_p)$. Of course, the dimension independence of the pathwise fluctuations will also cease to hold if we are interested in a result that is uniform in the order q, as in Theorem 2.3.

3. Entropy bounds. In section 2, we obtained pathwise bounds on the fluctuations of the likelihood ratio statistic in terms of the geometry of the underlying model classes. However, we have required two distinct types of geometric conditions: local entropy bounds for classes of densities, and global entropy bounds for classes of weighted densities. In this section, we will show that the latter implies the former under appropriate conditions, so that a suitable global entropy bound for weighted densities suffices for all the results in section 2. We will subsequently show how the requisite entropy bounds can be obtained for the case of location mixtures (cf. Example 2.1). The latter is significant both as an important application, and as a nontrivial case study in obtaining local entropy bounds in models with a complicated geometry.

3.1. *From global entropy to local entropy.* We are going to establish that local entropy estimates for a class of densities $\mathcal{M}$ can be obtained from global entropy estimates on the associated weighted class $\mathcal{D}$. To this end, let us consider for the purposes of this section a general class of positive probability densities $\mathcal{M}$ with

respect to some reference measure μ, a fixed $f^\star \in \mathcal{M}$, and define the class of weighted densities $\mathcal{D} = \{d_f : f \in \mathcal{M},\ f \neq f^\star\}$. In addition, we define for $\delta > 0$ the Hellinger ball $\mathcal{H}(\delta) = \{\sqrt{f/f^\star} : h(f, f^\star) \leq \delta\}$. We obtain the following result, whose proof is given in section 5.3.

THEOREM 3.1. *Suppose that there exist $q, C_0 \geq 1$ and $\varepsilon_0 > 0$ such that*

$$\mathcal{N}(\mathcal{D}, \varepsilon) \leq \left(\frac{C_0}{\varepsilon}\right)^q \quad \text{for every } \varepsilon \leq \varepsilon_0.$$

Let $R \geq \sup_f |d_f|$ be an envelope function such that $\|R\|_2 < \infty$. Then

$$\mathcal{N}(\mathcal{H}(\delta), \rho) \leq \left(\frac{C_1 \delta}{\rho}\right)^{q+1}$$

for all $\delta, \rho > 0$ such that $\rho/\delta < 4 \wedge 2\|R\|_2$, where $C_1 = 8C_0(1 \vee \|R\|_2/4\varepsilon_0)$.

Of course, in the setting of section 2, we would apply this result to $\mathcal{M}_q^n$, $\mathcal{D}_q^n$, $\mathcal{H}_q^n(\varepsilon)$ for given n, q instead of to $\mathcal{M}$, $\mathcal{D}$, $\mathcal{H}(\varepsilon)$.

3.2. *The entropy of mixtures.* We now develop the requisite entropy bounds in the case of mixtures (Example 2.1). In this section, let μ be the Lebesgue measure on $\mathbb{R}^d$. We fix a strictly positive probability density f_0 with respect to μ, and consider mixtures of densities in the class

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

In everything that follows we fix a nondegenerate mixture $f^\star$ of the form

$$f^\star = \sum_{i=1}^{q^\star} \pi_i^\star f_{\theta_i^\star}.$$

Nondegenerate means that $\pi_i^\star > 0$ for all i, and $\theta_i^\star \neq \theta_j^\star$ for all $i \neq j$.

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

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

and define the class of weighted densities as $\mathcal{D}_q = \{d_f : f \in \mathcal{M}_q,\ f \neq f^\star\}$. Let

$$\begin{aligned}
H_0(x) &= \sup_{\theta\in\Theta} f_\theta(x)/f^\star(x),\\
H_1(x) &= \sup_{\theta\in\Theta} \max_{i=1,\dots,d} |\partial f_\theta(x)/\partial\theta^i|/f^\star(x),\\
H_2(x) &= \sup_{\theta\in\Theta} \max_{i,j=1,\dots,d} |\partial^2 f_\theta(x)/\partial\theta^i\partial\theta^j|/f^\star(x),\\
H_3(x) &= \sup_{\theta\in\Theta} \max_{i,j,k=1,\dots,d} |\partial^3 f_\theta(x)/\partial\theta^i\partial\theta^j\partial\theta^k|/f^\star(x)
\end{aligned}$$

when f_0 is sufficiently differentiable, and let $\mathcal{M} = \bigcup_{q\geq 1}\mathcal{M}_q$ and $\mathcal{D} = \bigcup_{q\geq 1}\mathcal{D}_q$.

REMARK 3.2. In the setting of Example 2.1, the parameter set $\Theta = \Theta(n)$ depends on n, and we then write $\mathcal{M}_q^n$ instead of $\mathcal{M}_q$, etc. However, as the dependence on n is irrelevant for the entropy computation, we consider a fixed parameter set Θ in this section, and drop the dependence on n in our notation for simplicity.

We can now state the result of this section, whose proof is given in section 5.4.

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. $H_k \in L^4(f^\star d\mu)$ for $k = 0, 1, 2$ and $H_3 \in L^2(f^\star d\mu)$.

THEOREM 3.3. *Suppose that Assumption A holds. Then there exist constants $C^\star$ and $\delta^\star$, which depend on d, $q^\star$ and $f^\star$ but not on Θ, q or δ, such that*

$$\mathcal{N}(\mathcal{D}_q,\delta) \leq \left(\frac{C^\star(T\vee 1)^{1/6}(\|H_0\|_4^4 \vee \|H_1\|_4^4 \vee \|H_2\|_4^4 \vee \|H_3\|_2^2)}{\delta}\right)^{18(d+1)q}$$

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

$$\|D\|_4 \leq K^\star(\|H_0\|_4 \vee \|H_1\|_4 \vee \|H_2\|_4),$$

where $K^\star$ depends only on d and $f^\star$, such that $|d| \leq D$ for all $d \in \mathcal{D}$.

Let us note that a key aspect of this result is that the dependence of the entropy bound on the order q and on the parameter set Θ is essentially explicit (see Example 3.5 below, for example). However, even for fixed q and Θ, the existence of a polynomial bound on the bracketing number of $\mathcal{D}_q$ is far from obvious (previous claims [16, 6, 1] that such bracketing numbers are polynomial were stated without proof).

Define the Hellinger ball $\mathcal{H}_q(\varepsilon) = \{\sqrt{f/f^\star} : f \in \mathcal{M}_q,\ h(f,f^\star) \leq \varepsilon\}$. Using Theorem 3.1, we immediately obtain the following result on the local entropy of $\mathcal{M}_q$.

COROLLARY 3.4. *Suppose that Assumption A holds. Then*

$$\mathcal{N}(\mathcal{H}_q(\varepsilon),\delta) \le \left(\frac{C_\Theta\,\varepsilon}{\delta}\right)^{18(d+1)q+1}$$

for all $q \ge q^\star$ and $\delta/\varepsilon \le 1$, where

$$C_\Theta = L^\star\,(T\vee 1)^{1/6}\,(\|H_0\|_4^4 \vee \|H_1\|_4^4 \vee \|H_2\|_4^4 \vee \|H_3\|_2^2)^{5/4}$$

and $L^\star$ is a constant that depends only on d, $q^\star$ and $f^\star$.

EXAMPLE 3.5 (Gaussian mixtures). Consider mixtures of standard Gaussian densities $f_0(x) = (2\pi)^{-d/2}e^{-\|x\|^2/2}$, and let $\Theta(T) = \{\theta \in \mathbb{R}^d : \|\theta\| \le T\}$. Fix a nondegenerate mixture $f^\star$, and define $T^\star = \max_{i=1,\ldots,q^\star}\|\theta_i^\star\|$. Denote by $\mathcal{H}_q(\varepsilon,T)$ the Hellinger ball associated to the parameter set $\Theta(T)$. Then

$$\mathcal{N}(\mathcal{H}_q(\varepsilon,T),\delta) \le \left(\frac{C_1^\star e^{C_2^\star T^2}\varepsilon}{\delta}\right)^{18(d+1)q+1}$$

for all $q \ge q^\star$, $T \ge T^\star$, and $\delta/\varepsilon \le 1$, where $C_1^\star, C_2^\star$ are constants that depend on d, $q^\star$ and $f^\star$ 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^{CT^2}. These facts are readily verified by a straightforward but tedious computation.

REMARK 3.6. We have not optimized the constants in Theorem 3.3 and Corollary 3.4. 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 Θ is optimal. Indeed, if one is only interested in global entropy $\mathcal{N}(\mathcal{H}_q,\delta)$ where $\mathcal{H}_q = \{\sqrt{f/f^\star} : f \in \mathcal{M}_q\}$, then it can be read off from the proof of Theorem 3.3 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^{CT^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.

The proof of Theorem 3.3 is long and rather technical. Nonetheless, there are some key ideas underlying the proof, which we aim to briefly explain here.

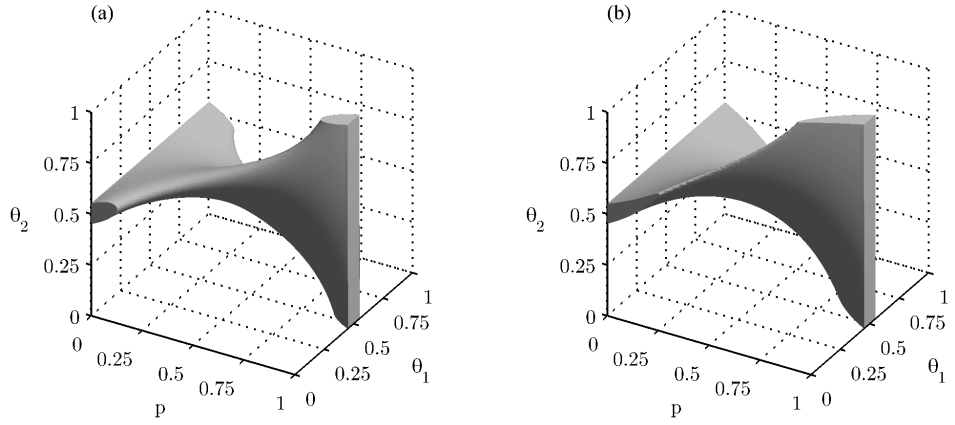

FIG 1. *Let* $f_\theta(x) = \sqrt{2/\pi}\, e^{-2(x-\theta)^2}$ *and* $f^\star = f_{0.5}$*, and consider the mixture family* $\mathcal{M}_2 = \{pf_{\theta_1} + (1-p)f_{\theta_2} : p, \theta_1, \theta_2 \in [0,1]\}$*. The plots illustrate (a) the set of parameters* (p, θ_1, θ_2) *corresponding to the Hellinger ball* $\{f \in \mathcal{M}_2 : h(f, f^\star) \le 0.05\}$*; and (b) the set of parameters* $\{(p, \theta_1, \theta_2) : N(p, \theta_1, \theta_2) \le 0.05\}$ *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 5.11, which yields* $c^\star N(p, \theta_1, \theta_2) \le h(pf_{\theta_1} + (1-p)f_{\theta_2}, f^\star) \le C^\star N(p, \theta_1, \theta_2)$ *for all* $p, \theta_1, \theta_2 \in [0,1]$*.*

The classical approach to controlling local entropies of a parametric class $\mathcal{G} = \{g_\xi : \xi \in \Xi\}$ with $\Xi \subset \mathbb{R}^d$ is as follows (cf. [26], Example 19.7). Suppose that the square root densities $h_\xi = \sqrt{g_\xi / g_{\xi^\star}}$ satisfy the pointwise Lipschitz condition

$$|h_\xi(x) - h_{\xi'}(x)| \le H(x)\, |||\xi - \xi'|||, \qquad \xi, \xi' \in \Xi,$$

where H is a function in L^2 and $|||\cdot|||$ is a norm on Ξ. Suppose, moreover, that

$$h(g_\xi, g_{\xi^\star}) \ge c\, |||\xi - \xi^\star|||, \qquad \xi \in \Xi.$$

Define $\Xi(\varepsilon) = \{\xi \in \Xi : |||\xi - \xi^\star||| \le \varepsilon\}$ and $\mathcal{H}(\varepsilon) = \{h_\xi : \xi \in \Xi,\ h(g_\xi, g_{\xi^\star}) \le \varepsilon\}$. If $|||\xi - \xi'||| \le \delta$, then $h_{\xi'} - \delta H \le h_\xi \le h_{\xi'} + \delta H$. Therefore, we can control the local bracketing entropy by $\mathcal{N}(\mathcal{H}(c\varepsilon), 2\delta\|H\|_2) \le N(\Xi(\varepsilon), \delta)$, where $N(\Xi(\varepsilon), \delta)$ denotes metric entropy. But the metric entropy of a ball can be controlled by a standard volume comparison argument, yielding $N(\Xi(\varepsilon), \delta) \le ((2\varepsilon + \delta)/\delta)^d$.

Clearly the above properties require $c|||\xi - \xi^\star||| \le h(g_\xi, g_{\xi^\star}) \le \|H\|_2 |||\xi - \xi^\star|||$ for all $\xi \in \Xi$. Therefore, such an approach can only work when the class $\mathcal{G}$ endowed with the Hellinger distance has a regular geometry (i.e., equivalent to a subset of a finite dimensional Banach space), at least in a neighborhood of the true parameter. This fails miserably in the case of mixture classes $\mathcal{M}_q$, which possess a highly non-regular geometry in a neighborhood of $f^\star$ when $q > q^\star$. In fact, it is easily seen that

$h(f,f^\star)=0$ does not even select a unique set of parameters $(\pi_i,\theta_i)_{i=1,\ldots,q}$, as mixture models are non-identifiable, and consequently the Hellinger balls $\mathcal{H}_q(\varepsilon)$ look nothing like norm-balls when viewed as a subset of the parameters $(\pi_i,\theta_i)_{i=1,\ldots,q}$ (cf. Figure 1). Thus we are faced with two basic difficulties:

1. How does one control the subset of parameters $(\pi_i,\theta_i)_{i=1,\ldots,q}$ corresponding to the Hellinger balls $\mathcal{H}_q(\varepsilon)$?
2. How does one control the metric entropy of these sets?

The resolution of the first problem requires us to develop a precise understanding of the local geometry of mixture classes, which is done in Theorem 5.11 below. One key consequence of this result, for example, is as follows: one can choose sufficiently small neighborhoods $A_1,\ldots,A_{q^\star}$ of $\theta_1,\ldots,\theta_{q^\star}$, respectively, such that the Hellinger distance $h(f,f^\star)$ is bounded above and below up to a constant by the pseudodistance

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

(here $f=\sum_{i=1}^q\pi_i f_{\theta_i}$ and $A_0=\mathbb{R}^d\backslash(A_1\cup\cdots\cup A_{q^\star})$). This pseudodistance quantifies precisely (and rather intuitively) the set of parameters with density close to $f^\star$, see Figure 1 for an illustration in the simplest possible case.

As for the second problem, we avoid it entirely by exploiting Theorem 3.1 instead of computing directly the local entropy. Using the local geometry Theorem 5.11 and Taylor expansion, we can approximate the weighted densities d_f by linear combinations of their first and second derivatives with coefficients in a Euclidean ball. The entropy of the latter is easily estimated by the Lipschitz argument indicated above. However, the details are somewhat intricate: Taylor expansion should only be applied to parameters θ_j that lie close to some $\theta_i^\star$, which requires careful bookkeeping.

The local geometry Theorem 5.11 and the relation between global entropy of weighted densities and local entropy developed in Theorem 3.1 are key ideas that allow us to obtain local entropy estimates in a geometrically nontrivial model. Let us note that the restriction to location mixtures is only used in the proof of Theorem 5.11. We believe that the same technique is applicable to other classes of mixtures (for example, Poisson mixtures or mixtures of densities in an exponential family) provided that the proof of Theorem 5.11 can be adapted to this setting.

4. Strongly consistent order estimation. The goal of this section is to apply the results of sections 2 and 3 to identify what penalties and cutoffs yield strong consistency of penalized likelihood order estimators. We first develop some general

consistency and inconsistency results, and then consider specifically the problem of mixture order estimation.

4.1. *Consistency and minimal penalties.* In this section we consider the general setting introduced in section 2.1. We now suppose, however, that the true model order $q^\star$ (as well as the true density $f^\star$) is not known, so that we must aim to estimate $q^\star$ from an observation sequence $(X_k)_{k\geq 1}$. To this end, define the *penalized likelihood order estimator* as

$$\hat{q}_n = \operatorname*{argmax}_{q\geq 1} \left\{ \sup_{f\in\mathcal{M}_q^n} \ell_n(f) - \mathrm{pen}(n,q) \right\},$$

where $\mathrm{pen}(n,q)$ is a penalty function. Our goal is to show that the penalized likelihood order estimator is strongly consistent, that is, $\hat{q}_n \to q^\star$ as $n\to\infty$ $\mathbf{P}^\star$-a.s., for a suitable choice of the penalty (that does not depend on $q^\star$ or $f^\star$). Let us emphasize that the maximum in the definition of $\hat{q}_n$ is taken over *all* model orders $q\geq 1$, that is, we do not assume that an a priori upper bound on the order is available, in contrast to most previous work on this topic. We obtain the following general result.

THEOREM 4.1. *Suppose that for all n sufficiently large, we have*

$$\mathcal{N}(\mathcal{H}_q^n(\varepsilon),\delta) \leq \left(\frac{K(n)\varepsilon}{\delta}\right)^{\eta(q)}$$

for all $q\geq q^\star$ and $\delta\leq\varepsilon$, where $K(n)\geq 1$ and $\eta(q)\geq q$ are increasing functions and we assume that $\log K(n) = o(n)$. Let $\mathrm{pen}(n,q)$ be a penalty such that

$$\lim_{n\to\infty}\sup_{q>q^\star}\frac{\eta(q)\{\log K(2n)\vee\log\log n\}}{\mathrm{pen}(n,q)-\mathrm{pen}(n,q^\star)} = 0, \qquad \lim_{n\to\infty}\max_{q<q^\star}\frac{\mathrm{pen}(n,q)}{n} = 0,$$

and $\mathrm{pen}(n,q)$ is increasing in q. Then $\hat{q}_n\to q^\star$ as $n\to\infty$ $\mathbf{P}^\star$-a.s.

Theorem 4.1 is proved in section 5.5 below.

Let us now specialize to the case that $\mathcal{M}_q^n = \mathcal{M}_q$ does not depend on n, as in section 2.3. In this case, Theorem 4.1 immediately yields the following corollary.

COROLLARY 4.2. *Suppose that for all $q\geq q^\star$ and $\delta\leq\varepsilon$*

$$\mathcal{N}(\mathcal{H}_q(\varepsilon),\delta) \leq \left(\frac{K\varepsilon}{\delta}\right)^{\eta(q)},$$

where $K \geq 1$ and $\eta(q) \geq q$ is a strictly increasing function. Define the penalty

$$\mathrm{pen}(n, q) = \eta(q)\, \varpi(n),$$

where $\varpi(n)$ is any function such that

$$\lim_{n\to\infty} \frac{\log\log n}{\varpi(n)} = 0, \qquad \lim_{n\to\infty} \frac{\varpi(n)}{n} = 0.$$

Then $\hat{q}_n \to q^\star$ as $n \to \infty$ $\mathbf{P}^\star$-a.s.

Corollary 4.2 states that, when $\mathcal{M}_q^n = \mathcal{M}_q$ does not depend on n, the penalized likelihood order estimator is strongly consistent provided the penalty grows faster than $\log\log n$ and slower than n. Clearly the $\log\log n$ rate is the minimal one attainable by applying Theorem 4.1. This raises the question whether the $\log\log n$ rate is indeed minimal, in the sense that smaller penalties yield inconsistent estimators. The following result (which follows easily from Theorem 2.5) shows that this is indeed the case, so that the result of Corollary 4.2 is essentially optimal.

COROLLARY 4.3. *Suppose there exists $q > q^\star$ such that the following hold.*

1. *There is an envelope function $D : E \to \mathbb{R}$ such that $|d| \leq D$ for all $d \in \mathcal{D}_q$ and $D \in L^{2+\alpha}(f^\star d\mu)$ for some $\alpha > 0$. Moreover, $\int_0^1 \sqrt{\log \mathcal{N}(\mathcal{D}_q, u)}\, du < \infty$.*
2. *$\bar{\mathcal{D}}_q^c \backslash \bar{\mathcal{D}}_{q^\star}$ is nonempty.*

Let $\eta(q) > 0$ be any strictly increasing function, and define the penalty

$$\mathrm{pen}(n, q) = C\, \eta(q)\, \log\log n.$$

If the constant $C > 0$ is chosen sufficiently small, then $\hat{q}_n \neq q^\star$ infinitely often $\mathbf{P}^\star$-a.s.

The proof of Corollary 4.3 is given in section 5.5. Let us note that the proof of Corollary 4.3 actually shows that $\sup_{f\in\mathcal{M}_q} \ell_n(f) - \mathrm{pen}(n, q) > \sup_{f\in\mathcal{M}_{q^\star}} \ell_n(f) - \mathrm{pen}(n, q^\star)$ infinitely often $\mathbf{P}^\star$-a.s., so the conclusion of Corollary 4.3 is not altered even if we were to impose a prior upper bound on the order.

In conclusion, we have shown that when $\mathcal{M}_q^n = \mathcal{M}_q$ does not depend on n, penalties growing faster than $\log\log n$ are consistent while the penalty $C\,\eta(q)\log\log n$ is inconsistent when the constant C is sufficiently small. From the proof of Theorem 4.1, we can also see that the penalty $C\,\eta(q)\log\log n$ is consistent when C is sufficiently large. However, the critical value of C may depend on the unknown parameter $f^\star$, so that this *minimal* penalty may not be implementable. On the other

hand, assuming that $\eta(q)$ does not depend on $f^\star$ (as is typically the case), penalties satisfying the assumptions of Theorem 4.1 obviously do not depend on the unknown parameter $f^\star$ and therefore define admissible estimators. When $\mathcal{M}_q^n$ depends on n, larger penalties may be required to ensure consistency, depending on the growth rate of $K(n)$.

4.2. *Mixture order estimation.* We finally apply the results in the previous section to mixture order estimation. Throughout this section, let $E = \mathbb{R}^d$ and let μ be the Lebesgue measure on $\mathbb{R}^d$. Fix a strictly positive probability density f_0 with respect to μ, and define

$$\mathcal{M}_q^n = \left\{ \sum_{i=1}^q \pi_i f_{\theta_i} : \pi_i \geq 0,\ \sum_{i=1}^q \pi_i = 1,\ \theta_i \in \Theta(n) \right\},$$

where $f_\theta(x) = f_0(x-\theta)$ and $\cdots \subseteq \Theta(n) \subseteq \Theta(n+1) \subseteq \cdots \subset \mathbb{R}^d$ is an increasing family of bounded subsets of $\mathbb{R}^d$. We fix $f^\star \in \mathcal{M}$ throughout this section. In the following, we consider two separate cases. The first case is that of a compact parameter set, where $\Theta(n) = \Theta$ does not depend on n. In this setting, we obtain a general result. Then, we consider the noncompact case in the setting of Gaussian mixtures, and illustrate how Theorem 4.1 can be used to obtain consistency results in this case.

Let us first consider the case of a compact parameter set. Then we obtain a general consistency result under Assumption A (cf. section 3.2).

PROPOSITION 4.4. *Suppose that the parameter set $\Theta(n) = \Theta$ is a bounded subset of $\mathbb{R}^d$ independent of n, and that Assumption A holds. If we choose a penalty of the form*

$$\mathrm{pen}(n,q) = q\,\omega(n), \qquad \lim_{n\to\infty} \frac{\log\log n}{\omega(n)} = \lim_{n\to\infty} \frac{\omega(n)}{n} = 0,$$

then $\hat{q}_n \to q^\star$ as $n \to \infty$ $\mathbf{P}^\star$-a.s. On the other hand, if we choose the penalty

$$\mathrm{pen}(n,q) = C\,q\,\log\log n$$

with a sufficiently small constant $C > 0$, then $\hat{q}_n \neq q^\star$ infinitely often $\mathbf{P}^\star$-a.s.

We therefore find that in the setting of location mixtures with a compact parameter set, the minimal penalty is of order $\log\log n$. Moreover, the popular BIC penalty

$$\mathrm{pen}(n,q) = \frac{dq+q-1}{2}\,\log n \tag{4.1}$$

yields a strongly consistent mixture order estimator in this setting, without a prior upper bound on the order. The requisite Assumption A is a very mild one, which highlights the broad applicability of this result. However, the assumption of a compact parameter space can be quite restrictive in practice.

Let us therefore consider a case where the parameter space is noncompact. For simplicity we restrict our attention to Gaussian mixtures, that is, we choose $f_0(x) = (2\pi)^{-d/2} e^{-\|x\|^2/2}$, and we choose the restricted parameter sets $\Theta(n) = \{\theta \in \mathbb{R}^d : \|\theta\| \le T(n)\}$ for some sequence $T(n) \uparrow \infty$. Our aim is to choose the penalty $\mathrm{pen}(n,q)$ and cutoff $T(n)$ so that the penalized likelihood order estimator is strongly consistent. In this setting, we obtain the following result.

PROPOSITION 4.5. *Let $f_0(x) = (2\pi)^{-d/2} e^{-\|x\|^2/2}$ and $\Theta(n) = \{\theta \in \mathbb{R}^d : \|\theta\| \le T(n)\}$, and choose a penalty of the form $\mathrm{pen}(n,q) = q\,\omega(n)$. If*

$$\lim_{n\to\infty} \frac{\log\log n}{\omega(n)} = \lim_{n\to\infty} \frac{\omega(n)}{n} = 0, \qquad T(n) = O(\sqrt{\log\log n}),$$

then $\hat{q}_n \to q^\star$ as $n \to \infty$ $\mathbf{P}^\star$-a.s. On the other hand, the BIC penalty (4.1) yields a strongly consistent order estimator if $T(n) = o(\sqrt{\log n})$.

This result illustrates that our theory can establish consistency of the penalized likelihood mixture order estimator without any prior upper bounds on the model order or the magnitude of the true parameters. Let us note that there is nothing particularly special about the Gaussian case: a similar result can be obtained, in principle, for any mixture distribution, as long as one can obtain suitable estimates on the quantities $\|H_i\|_4$ that appear in Corollary 3.4 (see Example 3.5 for the Gaussian case).

The proofs of Propositions 4.4 and 4.5 appear in section 5.6 below.

5. Proofs.

5.1. *Proof of Theorem 2.3.* The proof of Theorem 2.3 is based on the following deviation bound for the log-likelihood ratio. This bound is essentially from [25], Corollary 7.5, but the additional maximum inside the probability is essential for our purposes.

THEOREM 5.1. *Let $\mathcal{M}$ be a family of strictly positive probability densities with respect to a reference measure μ, fix some $f^\star \in \mathcal{M}$, and define the Hellinger ball $\mathcal{H}(\varepsilon) = \{\sqrt{f/f^\star} : f \in \mathcal{M},\ h(f,f^\star) \le \varepsilon\}$ where $h(f,g)^2 = \int(\sqrt{f} - \sqrt{g})^2 d\mu$. Suppose that for some constants $K \ge 1$, $p \ge 1$ and all $\delta \le \varepsilon$*

$$\mathcal{N}(\mathcal{H}(\varepsilon), \delta) \le \left(\frac{K\varepsilon}{\delta}\right)^p,$$

where $\mathcal{N}(\mathcal{H}(\varepsilon),\delta)$ is the minimal number of brackets of $L^2(f^\star d\mu)$-width δ needed to cover $\mathcal{H}(\varepsilon)$. Let $(X_i)_{i\in\mathbb{N}}$ be i.i.d. with distribution $f^\star d\mu$. Then

$$\mathbf{P}\left[\max_{n\le k\le 2n}\sup_{f\in\mathcal{M}}\sum_{j=1}^{k}\log\left(\frac{f(X_j)}{f^\star(X_j)}\right)\ge\alpha\right]\le C\,e^{-\alpha/C}$$

for all $\alpha\ge Cp(1+\log K)$ and $n\ge 1$, where C is a universal constant.

PROOF. Define $\bar f=(f+f^\star)/2$ for any $f\in\mathcal{M}$, and define the empirical process $\nu_n(g)=n^{-1/2}\sum_{k=1}^{n}\{g(X_k)-\mathbf{E}[g(X_k)]\}$. Using concavity of $\log x$ we have

$$\sum_{j=1}^{k}\log\left(\frac{f(X_j)}{f^\star(X_j)}\right)\le 2k^{1/2}\nu_k(\log(\bar f/f^\star))-2kD(f^\star||\bar f),$$

where $D(f^\star||f)=\int\log(f^\star/f)f^\star d\mu$ is relative entropy. As $D(f^\star||f)\ge h(f,f^\star)^2$

$$\begin{aligned}
&\mathbf{P}\left[\max_{n\le k\le 2n}\sup_{f\in\mathcal{M}}\sum_{j=1}^{k}\log\left(\frac{f(X_j)}{f^\star(X_j)}\right)\ge\alpha\right]\\
&\le\mathbf{P}\left[\max_{n\le k\le 2n}\sup_{f\in\mathcal{M}}\{2k^{1/2}\nu_k(\log(\bar f/f^\star))-2kh(\bar f,f^\star)^2\}\ge\alpha\right]\\
&\le\sum_{s=0}^{S}\mathbf{P}\left[\max_{n\le k\le 2n}\sup_{f\in\mathcal{M}:nh(\bar f,f^\star)^2\le\alpha 2^s}|k^{1/2}\nu_k(\log(\bar f/f^\star))|\ge\alpha 2^{s-1}\right]\\
&\le 3\sum_{s=0}^{S}\max_{n\le k\le 2n}\mathbf{P}\left[\sup_{f\in\mathcal{M}:h(\bar f,f^\star)^2\le\alpha 2^s n^{-1}}|\nu_k(\log(\{\bar f/f^\star\}^{1/2}))|\ge\alpha 2^{s-5}n^{-1/2}\right],
\end{aligned}$$

where $S=\min\{s:\alpha 2^s n^{-1}>2\}$, and we have used Lemma 5.2 below for the last inequality. The remainder of the proof is identical to that of [25], Theorem 7.4 provided we show that for $\bar{\mathcal{H}}(\varepsilon)=\{\sqrt{\bar f/f^\star}:f\in\mathcal{M},\ h(\bar f,f^\star)\le\varepsilon\}$

$$\mathcal{N}(\bar{\mathcal{H}}(\varepsilon),\delta)\le\left(\frac{2\sqrt{2}K\varepsilon}{\delta}\right)^p.$$

To this end, fix $\delta\le\varepsilon$, and note that $h(f,f^\star)\le 4h(\bar f,f^\star)$ by [25], Lemma 4.2 so that $\{f\in\mathcal{M}:h(\bar f,f^\star)\le\varepsilon\}\subseteq\{f\in\mathcal{M}:h(f,f^\star)\le 4\varepsilon\}$. By assumption, there exist $N\le(2\sqrt{2}K\varepsilon/\delta)^p$ and functions $g_1,\dots,g_N,h_1,\dots,h_N$ such that $\|h_i-g_i\|_2\le\delta\sqrt{2}$ for every i, and for every $u\in\mathcal{H}(4\varepsilon)$ there is an i such that $g_i\le u\le h_i$. But for every $f\in\mathcal{M}$ such that $h(\bar f,f^\star)\le\varepsilon$, we then have for some i

$$2^{-1/2}\sqrt{g_i^2+1}\le\sqrt{\bar f/f^\star}\le 2^{-1/2}\sqrt{h_i^2+1}.$$

Moreover, using $|\sqrt{a+c}-\sqrt{b+c}| \le |\sqrt{a}-\sqrt{b}|$ for $a,b,c \ge 0$ we obtain

$$\left\| 2^{-1/2}\sqrt{h_i^2+1} - 2^{-1/2}\sqrt{g_i^2+1} \right\|_2 \le 2^{-1/2}\|h_i - g_i\|_2 \le \delta.$$

The result now follows directly. □

The following variant of Etemadi's inequality was used in the proof. The proof follows closely that of the classical Etemadi inequality, see [4], Appendix M19.

LEMMA 5.2. *Let $\mathcal{Q}$ be a family of measurable functions $f : E \to \mathbb{R}$. Then we have for every $\alpha > 0$ and $m, n \in \mathbb{N}$, $m \le n$*

$$\mathbf{P}^\star\left[\max_{k=m,\ldots,n}\sup_{f\in\mathcal{Q}}|S_k(f)| \ge 3\alpha\right] \le 3\max_{k=m,\ldots,n}\mathbf{P}^\star\left[\sup_{f\in\mathcal{Q}}|S_k(f)| \ge \alpha\right],$$

where $S_n(f) = n^{1/2}\nu_n(f)$.

PROOF. Define the stopping time $\tau = \inf\{k \ge m : \sup_{f\in\mathcal{Q}}|S_k(f)| \ge 3\alpha\}$. Then

$$\begin{aligned}\mathbf{P}^\star\left[\max_{k=m,\ldots,n}\sup_{f\in\mathcal{Q}}|S_k(f)| \ge 3\alpha\right] &= \mathbf{P}^\star[\tau \le n] \\ &\le \mathbf{P}^\star\left[\sup_{f\in\mathcal{Q}}|S_n(f)| \ge \alpha\right] + \sum_{k=m}^{n}\mathbf{P}^\star\left[\tau = k \text{ and } \sup_{f\in\mathcal{Q}}|S_n(f)| < \alpha\right].\end{aligned}$$

But on the event $\{\tau = k \text{ and } \sup_{f\in\mathcal{Q}}|S_n(f)| < \alpha\}$, we clearly have

$$2\alpha \le \sup_{f\in\mathcal{Q}}|S_k(f)| - \sup_{f\in\mathcal{Q}}|S_n(f)| \le \sup_{f\in\mathcal{Q}}|S_k(f) - S_n(f)|.$$

Therefore, we can estimate

$$\begin{aligned}&\mathbf{P}^\star\left[\max_{k=m,\ldots,n}\sup_{f\in\mathcal{Q}}|S_k(f)| \ge 3\alpha\right] \\ &\quad\le \mathbf{P}^\star\left[\sup_{f\in\mathcal{Q}}|S_n(f)| \ge \alpha\right] + \sum_{k=m}^{n}\mathbf{P}^\star\left[\tau = k \text{ and } \sup_{f\in\mathcal{Q}}|S_n(f) - S_k(f)| \ge 2\alpha\right] \\ &\quad\le \mathbf{P}^\star\left[\sup_{f\in\mathcal{Q}}|S_n(f)| \ge \alpha\right] + \max_{k=m,\ldots,n}\mathbf{P}^\star\left[\sup_{f\in\mathcal{Q}}|S_n(f) - S_k(f)| \ge 2\alpha\right],\end{aligned}$$

where we have used that $\sup_{f\in\mathcal{Q}}|S_n(f) - S_k(f)|$ and $\{\tau = k\}$ are independent to obtain the last inequality. The proof is now easily completed. □

We can now complete the proof of Theorem 2.3.

PROOF OF THEOREM 2.3. By assumption, we have $f^\star \in \mathcal{M}_q^n$ for all $q \geq q^\star$ when n is sufficiently large. Then by Theorem 5.1, we have for n sufficiently large

$$\mathbf{P}^\star\left[\max_{n\leq k\leq 2n}\sup_{f\in\mathcal{M}_q^{2n}}\{\ell_k(f)-\ell_k(f^\star)\}\geq\alpha\right]\leq C\,e^{-\alpha/C}$$

for all $\alpha \geq C\eta(q)(1+\log K(2n))$ and $q \geq q^\star$. Using that $\mathcal{M}_q^k \subseteq \mathcal{M}_q^{2n}$ for $n \leq k \leq 2n$ and $\ell_k(f^\star) \leq \sup_{f\in\mathcal{M}_{q^\star}^k}\ell_k(f)$, we have for n sufficiently large

$$\mathbf{P}^\star\left[\max_{n\leq k\leq 2n}\sup_{q\geq q^\star}\frac{1}{\eta(q)}\left\{\sup_{f\in\mathcal{M}_q^k}\ell_k(f)-\sup_{f\in\mathcal{M}_{q^\star}^k}\ell_k(f)\right\}\geq\alpha\right]\leq\sum_{q=q^\star}^{\infty}C\,e^{-\alpha\eta(q)/C}$$

for all $\alpha \geq C(1+\log K(2n))$. Let $\beta(n)$ be an increasing function. Then

$$\mathbf{P}^\star\left[\max_{2^n\leq k\leq 2^{n+1}}\frac{1}{\beta(k)}\sup_{q\geq q^\star}\frac{1}{\eta(q)}\left\{\sup_{f\in\mathcal{M}_q^k}\ell_k(f)-\sup_{f\in\mathcal{M}_{q^\star}^k}\ell_k(f)\right\}\geq 2C\right]\leq\frac{2C}{n^2}$$

for all n sufficiently large, provided that $\beta(2^n) \geq \log K(2^{n+1}) \vee \log\log 2^n$. The proof is now easily completed using the Borel-Cantelli lemma. □

5.2. *Proof of Theorem 2.5.* The proof of Theorem 2.5 is based on a sequence of auxiliary results. First, we will need a compact law of iterated logarithm for the Strassen functional

$$I_n(g)=\frac{1}{\sqrt{2n\log\log n}}\sum_{i=1}^{n}\{g(X_i)-\mathbf{E}^\star(g(X_1))\}\,.$$

We state the requisite result for future reference.

THEOREM 5.3. *Let $\mathcal{Q}$ be a family of measurable functions $f : E \to \mathbb{R}$ with*

$$\int_0^1\sqrt{\log\mathcal{N}(\mathcal{Q},u)}\,du<\infty.$$

Then, $\mathbf{P}^\star$-a.s., the sequence $(I_n)_{n\geq 0}$ is relatively compact in $\ell_\infty(\mathcal{Q})$, and its set of cluster points coincides precisely with the set $\mathcal{K}=\{f\mapsto\langle f,g\rangle : g\in L_0^2(f^\star d\mu)\}$.

Proofs of this result can be found in [23], Theorem 4.2 or in [17], Theorem 9.

We will also need the following simple result, whose proof is omitted.

LEMMA 5.4. *Let $(X_i)_{i\geq 1}$ be an i.i.d. sequence of random variables, and suppose $\mathbf{E}[|X_1|^p] < \infty$. Then $n^{-1/p} \max_{i=1,\ldots,n} |X_i| \to 0$ a.s. as $n \to \infty$.*

Finally, we will need the following likelihood inequality that relates the log-likelihood ratio $\ell_n(f) - \ell_n(f^\star)$ to the empirical process. Related inequalities appear in [11, 18, 6], but the following form is perhaps the most natural.

LEMMA 5.5. *For any strictly positive probability density $f \neq f^\star$, we have*

$$\ell_n(f) - \ell_n(f^\star) \leq |\nu_n(d_f)|^2,$$

where $\nu_n(g) = n^{-1/2} \sum_{k=1}^n \{g(X_k) - \mathbf{E}^\star[g(X_k)]\}$ denotes the empirical process.

PROOF. Note that

$$h(f,f^\star)^2 = 2 - \int 2\sqrt{f f^\star}\, d\mu = -2\, h(f,f^\star)\, \mathbf{E}^\star(d_f(X_1)).$$

Using $\log(1+x) \leq x$, we can estimate

$$\begin{aligned}\ell_n(f) - \ell_n(f^\star) &= \sum_{i=1}^n 2\, \log(1 + h(f,f^\star)\, d_f(X_i)) \leq \sum_{i=1}^n 2\, h(f,f^\star)\, d_f(X_i) \\ &= 2\, \nu_n(d_f)\, h(f,f^\star)\, \sqrt{n} - h(f,f^\star)^2\, n \leq \sup_{p\in\mathbb{R}} \left\{2\, \nu_n(d_f)\, p - p^2\right\}.\end{aligned}$$

The proof is easily completed. □

We can now obtain the following asymptotic expansion of the log-likelihood, which provides a pathwise counterpart to the weak convergence theory in [11, 18].

PROPOSITION 5.6. *Let $q \geq q^\star$. Assume that*

$$\int_0^1 \sqrt{\log \mathcal{N}(\mathcal{D}_q, u)}\, du < \infty,$$

and that $|d| \leq D$ for all $d \in \mathcal{D}_q$ with $D \in L^{2+\alpha}(f^\star d\mu)$ for some $\alpha > 0$. Then

$$\begin{aligned}\sup_{f\in\mathcal{M}_q(4\sqrt{\log\log n/n})} &\left\{2\, I_n(d_f)\, h(f,f^\star)\, \sqrt{\frac{2n}{\log\log n}} - h(f,f^\star)^2\, \frac{2n}{\log\log n}\right\} \\ &- \frac{1}{\log\log n} \left\{\sup_{f\in\mathcal{M}_q} \ell_n(f) - \ell_n(f^\star)\right\} \xrightarrow{n\to\infty} 0 \quad \mathbf{P}^\star\text{-a.s.},\end{aligned}$$

where we have defined $\mathcal{M}_q(\varepsilon) = \{f \in \mathcal{M}_q : h(f,f^\star) \leq \varepsilon\}$.

PROOF. We proceed in several steps.

Step 1 (*localization*). As $q \geq q^\star$ (hence $f^\star \in \mathcal{M}_q$), clearly

$$\sup_{f\in\mathcal{M}_q} \ell_n(f) - \ell_n(f^\star) = \sup_{f\in\mathcal{M}_q:\ell_n(f)-\ell_n(f^\star)\geq 0} \{\ell_n(f) - \ell_n(f^\star)\}.$$

Now note that, as in the proof of Lemma 5.5,

$$\ell_n(f) - \ell_n(f^\star) \leq 2\,\nu_n(d_f)\,h(f,f^\star)\,\sqrt{n} - h(f,f^\star)^2\,n.$$

Therefore, we can estimate

$$\begin{aligned}
&\sup_{f\in\mathcal{M}_q:\ell_n(f)-\ell_n(f^\star)\geq 0} h(f,f^\star) \\
&\quad\leq \sup_{f\in\mathcal{M}_q:\ell_n(f)-\ell_n(f^\star)\geq 0} \left\{ h(f,f^\star) + \frac{\ell_n(f)-\ell_n(f^\star)}{n\,h(f,f^\star)} \right\} \\
&\quad\leq \frac{2}{\sqrt{n}} \sup_{f\in\mathcal{M}_q:\ell_n(f)-\ell_n(f^\star)\geq 0} \nu_n(d_f) \leq \sqrt{\frac{8\log\log n}{n}} \sup_{d\in\mathcal{D}_q} I_n(d).
\end{aligned}$$

Now note that we can estimate

$$\sup_{d\in\mathcal{D}_q} I_n(d) \leq \inf_{g\in L_0^2(f^\star d\mu)} \sup_{d\in\mathcal{D}_q} |I_n(d) - \langle d,g\rangle| + \sup_{d\in\mathcal{D}_q} \sup_{g\in L_0^2(f^\star d\mu)} \langle d,g\rangle.$$

The first term on the right converges to zero $\mathbf{P}^\star$-a.s. as $n\to\infty$ by Theorem 5.3, while the second term is easily seen to equal $\sup_{d\in\mathcal{D}_q}\|d-\langle 1,d\rangle\|_2 \leq 1$. Therefore

$$\sup_{f\in\mathcal{M}_q:\ell_n(f)-\ell_n(f^\star)\geq 0} h(f,f^\star) \leq (1+\varepsilon)\sqrt{\frac{8\log\log n}{n}}$$

eventually as $n\to\infty$ $\mathbf{P}^\star$-a.s. for any $\varepsilon>0$. In particular, we find that

$$\{f\in\mathcal{M}_q : \ell_n(f)-\ell_n(f^\star)\geq 0\} \subseteq \left\{ f\in\mathcal{M}_q : h(f,f^\star) \leq 4\sqrt{\log\log n/n} \right\}$$

eventually as $n\to\infty$ $\mathbf{P}^\star$-a.s. This implies that $\mathbf{P}^\star$-a.s. eventually as $n\to\infty$

$$\sup_{f\in\mathcal{M}_q} \ell_n(f) - \ell_n(f^\star) \leq \sup_{f\in\mathcal{M}_q:h(f,f^\star)\leq 4\sqrt{\log\log n/n}} \{\ell_n(f) - \ell_n(f^\star)\}.$$

But the reverse inequality clearly holds for all $n\geq 0$, so that in fact

$$\sup_{f\in\mathcal{M}_q} \ell_n(f) - \ell_n(f^\star) = \sup_{f\in\mathcal{M}_q(4\sqrt{\log\log n/n})} \{\ell_n(f) - \ell_n(f^\star)\}$$

eventually as $n \to \infty$ $\mathbf{P}^\star$-a.s.

Step 2 (*Taylor expansion*). Taylor expansion gives $2\log(1+x) = 2x - x^2 + x^2 R(x)$, where $R(x) \to 0$ as $x \to 0$. Thus we can write, for any $f \in \mathcal{M}_q$,

$$\begin{aligned}\ell_n(f) - \ell_n(f^\star) = \sum_{i=1}^n 2\,\log(1 + h(f,f^\star)\,d_f(X_i)) = \\ 2\,h(f,f^\star)\sum_{i=1}^n \left\{ d_f(X_i) + \frac{1}{2}\,h(f,f^\star)\right\} - h(f,f^\star)^2 \sum_{i=1}^n (d_f(X_i))^2 \\ - n\,h(f,f^\star)^2 + h(f,f^\star)^2 \sum_{i=1}^n (d_f(X_i))^2 R(h(f,f^\star)\,d_f(X_i)).\end{aligned}$$

Using that $\mathbf{E}^\star(d_f(X_1)) = -h(f,f^\star)/2$, we therefore have

$$\begin{aligned}\frac{1}{\log\log n}\,\{\ell_n(f) - \ell_n(f^\star)\} = \\ 2\,I_n(d_f)\,h(f,f^\star)\sqrt{\frac{2n}{\log\log n}} - h(f,f^\star)^2\,\frac{2n}{\log\log n} + R_{f,n}\,\frac{n\,h(f,f^\star)^2}{\log\log n}\end{aligned}$$

where we have defined

$$R_{f,n} = \frac{1}{n}\sum_{i=1}^n \{1 - (d_f(X_i))^2\} + \frac{1}{n}\sum_{i=1}^n (d_f(X_i))^2 R(h(f,f^\star)\,d_f(X_i)).$$

It follows easily that

$$\begin{aligned}\left| \sup_{f\in\mathcal{M}_q(4\sqrt{\log\log n/n})} \left\{ 2\,I_n(d_f)\,h(f,f^\star)\sqrt{\frac{2n}{\log\log n}} - h(f,f^\star)^2\,\frac{2n}{\log\log n}\right\}\right. \\ \left. - \frac{1}{\log\log n}\left\{ \sup_{f\in\mathcal{M}_q} \ell_n(f) - \ell_n(f^\star)\right\}\right| \\ \le \sup_{f\in\mathcal{M}_q(4\sqrt{\log\log n/n})} |R_{f,n}|\,\frac{n\,h(f,f^\star)^2}{\log\log n} \le 16 \sup_{f\in\mathcal{M}_q(4\sqrt{\log\log n/n})} |R_{f,n}|\end{aligned}$$

eventually as $n \to \infty$ $\mathbf{P}^\star$-a.s.

Step 3 (*end of proof*). We can easily estimate

$$\begin{aligned}\sup_{f\in\mathcal{M}_q(4\sqrt{\log\log n/n})} |R_{f,n}| \le \sup_{f\in\mathcal{M}_q} \left| \frac{1}{n}\sum_{i=1}^n \{(d_f(X_i))^2 - 1\}\right| \\ + \left(\sup_{|x|\le 4\sqrt{\log\log n/n}\,\max_{i=1,\ldots,n} D(X_i)} |R(x)|\right) \frac{1}{n}\sum_{i=1}^n (D(X_i))^2.\end{aligned}$$

As $\mathcal{N}(\mathcal{D}_q,\delta)<\infty$ for every $\delta>0$, the class $\{d^2 : d\in\mathcal{D}_q\}$ can be covered by a finite number of brackets with arbitrary small $L^1(f^\star d\mu)$-norm and is therefore $\mathbf{P}^\star$-Glivenko-Cantelli. Moreover, by construction $\mathbf{E}^\star[(d_f(X_i))^2]=1$ for all $f\in\mathcal{M}_q$. Therefore, the first term in this expression converges to zero as $n\to\infty$ $\mathbf{P}^\star$-a.s. On the other hand, by Lemma 5.4 and the fact that $D\in L^{2+\alpha}(f^\star d\mu)$, we have $\mathbf{P}^\star$-a.s.

$$\sqrt{\log\log n/n}\max_{i=1,\ldots,n}D(X_i)=\frac{\sqrt{\log\log n}}{n^{\alpha/2(2+\alpha)}}\,n^{-1/(2+\alpha)}\max_{i=1,\ldots,n}D(X_i)\xrightarrow{n\to\infty}0.$$

Therefore the second term converges to zero also, and the proof is complete. □

PROPOSITION 5.7. *Let $q\geq q^\star$. Assume that*

$$\int_0^1\sqrt{\log\mathcal{N}(\mathcal{D}_q,u)}\,du<\infty,$$

and that $|d|\leq D$ for all $d\in\mathcal{D}_q$ with $D\in L^{2+\alpha}(f^\star d\mu)$ for some $\alpha>0$. Then

$$\liminf_{n\to\infty}\left\{\sup_{d\in\bar{\mathcal{D}}_q}(I_n(d))_+^2-\frac{1}{\log\log n}\left\{\sup_{f\in\mathcal{M}_q}\ell_n(f)-\ell_n(f^\star)\right\}\right\}\geq 0\quad\mathbf{P}^\star\text{-a.s.}$$

PROOF. By Proposition 5.6, we have

$$\begin{aligned}
&\liminf_{n\to\infty}\left\{\sup_{d\in\bar{\mathcal{D}}_q}(I_n(d))_+^2-\frac{1}{\log\log n}\left\{\sup_{f\in\mathcal{M}_q}\ell_n(f)-\ell_n(f^\star)\right\}\right\}\\
&\quad\geq\liminf_{n\to\infty}\left\{\sup_{d\in\bar{\mathcal{D}}_q}(I_n(d))_+^2-\sup_{f\in\mathcal{M}_q(4\sqrt{\log\log n/n})}\sup_{p\geq 0}\left\{2\,I_n(d_f)\,p-p^2\right\}\right\}\\
&\quad=\liminf_{n\to\infty}\left\{\sup_{d\in\bar{\mathcal{D}}_q}(I_n(d))_+^2-\sup_{f\in\mathcal{M}_q(4\sqrt{\log\log n/n})}(I_n(d_f))_+^2\right\}.
\end{aligned}$$

Suppose that the right hand side is negative with positive probability. Then there is an $\varepsilon>0$ and a sequence $\tau_n\uparrow\infty$ of random times such that

$$\sup_{d\in\bar{\mathcal{D}}_q}(I_{\tau_n}(d))_+^2-\sup_{f\in\mathcal{M}_q(4\sqrt{\log\log\tau_n/\tau_n})}(I_{\tau_n}(d_f))_+^2\leq-\varepsilon\quad\text{for all } n\tag{5.1}$$

with positive probability. We will show that this entails a contradiction.

By Theorem 5.3 (which can be applied here as $\mathcal{N}(\mathcal{D}_q,\delta)=\mathcal{N}(\operatorname{cl}\mathcal{D}_q,\delta)$ for all $\delta>0$), the process $(I_{\tau_n})_{n\geq 0}$ is $\mathbf{P}^\star$-a.s. relatively compact in $\ell_\infty(\operatorname{cl}\mathcal{D}_q)$ with

$$\inf_{g\in L_0^2(f^\star d\mu)}\sup_{d\in\operatorname{cl}\mathcal{D}_q}|I_{\tau_n}(d)-\langle d,g\rangle|\xrightarrow{n\to\infty}0\quad\mathbf{P}^\star\text{-a.s.}\tag{5.2}$$

Then there is a set of positive probability on which (5.1) and (5.2) hold simultaneously. We now concentrate our attention on a single sample path in this set. For any such path, we can clearly find a further subsequence $\sigma_n \uparrow \infty$ such that $\sup_{d\in\mathrm{cl}\,\mathcal{D}_q} |I_{\sigma_n}(d) - \langle d, g\rangle| \to 0$ as $n \to \infty$ for some $g \in L_0^2(f^\star d\mu)$. Therefore

$$\sup_{d\in\mathrm{cl}\,\mathcal{D}_q} |(I_{\sigma_n}(d))_+^2 - (\langle d, g\rangle)_+^2| \le \sup_{d\in\mathrm{cl}\,\mathcal{D}_q} |I_{\sigma_n}(d) - \langle d, g\rangle|^2 + 2 \sup_{d\in\mathrm{cl}\,\mathcal{D}_q} |I_{\sigma_n}(d) - \langle d, g\rangle| \sup_{d\in\mathrm{cl}\,\mathcal{D}_q} |\langle d, g\rangle| \xrightarrow{n\to\infty} 0,$$

where we have used the elementary estimate $|a_+^2 - b_+^2| = |a_+ - b_+|(a_+ + b_+) \le |a_+ - b_+|(|a_+ - b_+| + 2b_+) \le |a - b|(|a - b| + 2|b|)$ for any $a, b \in \mathbb{R}$, and the fact that $\sup_{d\in\mathrm{cl}\,\mathcal{D}_q} |\langle d, g\rangle| \le \sup_{d\in\mathrm{cl}\,\mathcal{D}_q} \|d\|_2 \|g\|_2 \le 1$. Thus (5.1) gives

$$\liminf_{n\to\infty} \left\{ \sup_{d\in\bar{\mathcal{D}}_q} (\langle d, g\rangle)_+^2 - \sup_{f\in\mathcal{M}_q(4\sqrt{\log\log\sigma_n/\sigma_n})} (\langle d_f, g\rangle)_+^2 \right\} = \liminf_{n\to\infty} \left\{ \sup_{d\in\bar{\mathcal{D}}_q} (I_{\sigma_n}(d))_+^2 - \sup_{f\in\mathcal{M}_q(4\sqrt{\log\log\sigma_n/\sigma_n})} (I_{\sigma_n}(d_f))_+^2 \right\} \le -\varepsilon.$$

But as $d \mapsto \langle d, g\rangle$ is continuous in $L^2(f^\star d\mu)$ and $\mathrm{cl}\,\mathcal{D}_q(4\sqrt{\log\log\sigma_n/\sigma_n})$ is compact in $L^2(f^\star d\mu)$ (which follows from $\mathcal{N}(\mathcal{D}_q, \delta) < \infty$ for all $\delta > 0$), we have

$$\sup_{f\in\mathcal{M}_q(4\sqrt{\log\log\sigma_n/\sigma_n})} (\langle d_f, g\rangle)_+^2 = \sup_{d\in\mathrm{cl}\,\mathcal{D}_q(4\sqrt{\log\log\sigma_n/\sigma_n})} (\langle d, g\rangle)_+^2 \xrightarrow{n\to\infty} \sup_{d\in\bigcap_{n\ge0}\mathrm{cl}\,\mathcal{D}_q(4\sqrt{\log\log\sigma_n/\sigma_n})} (\langle d, g\rangle)_+^2 = \sup_{d\in\bar{\mathcal{D}}_q} (\langle d, g\rangle)_+^2.$$

Thus we have a contradiction, completing the proof. □

We now obtain a converse to the previous result.

PROPOSITION 5.8. *Let $q \ge q^\star$. Assume that*

$$\int_0^1 \sqrt{\log \mathcal{N}(\mathcal{D}_q, u)}\, du < \infty,$$

and that $|d| \le D$ for all $d \in \mathcal{D}_q$ with $D \in L^{2+\alpha}(f^\star d\mu)$ for some $\alpha > 0$. Then

$$\limsup_{n\to\infty} \left\{ \sup_{d\in\bar{\mathcal{D}}_q^c} (I_n(d))_+^2 - \frac{1}{\log\log n} \left\{ \sup_{f\in\mathcal{M}_q} \ell_n(f) - \ell_n(f^\star) \right\} \right\} \le 0 \quad \mathbf{P}^\star\text{-a.s.}$$

PROOF. Suppose that the result does not hold true. By Proposition 5.6, there is an $\varepsilon > 0$ and a sequence $\tau_n \uparrow \infty$ of random times such that

$$\sup_{d\in\bar{\mathcal{D}}_q^c} (I_{\tau_n}(d))_+^2 - \sup_{f\in\mathcal{M}_q(4\sqrt{\log\log\tau_n/\tau_n})} \left\{ - h(f,f^\star)^2 \, \frac{2\tau_n}{\log\log\tau_n} \right.$$
$$\left. + 2\, I_{\tau_n}(d_f)\, h(f,f^\star) \sqrt{\frac{2\tau_n}{\log\log\tau_n}} \right\} \geq \varepsilon \quad \text{for all } n$$

with positive probability. Proceeding as in the proof of Proposition 5.7, we can then show that there is a sequence of times $\sigma_n \uparrow \infty$ and some $g \in L_0^2(f^\star d\mu)$ such that

$$\limsup_{n\to\infty} \left\{ \sup_{d\in\bar{\mathcal{D}}_q^c} (\langle d,g\rangle)_+^2 - \sup_{f\in\mathcal{M}_q(4\sqrt{\log\log\sigma_n/\sigma_n})} \left\{ - h(f,f^\star)^2 \, \frac{2\sigma_n}{\log\log\sigma_n} \right.\right.$$
$$\left.\left. + 2\,\langle d_f, g\rangle\, h(f,f^\star) \sqrt{\frac{2\sigma_n}{\log\log\sigma_n}} \right\} \right\} \geq \varepsilon.$$

We will show that this entails a contradiction.

Let $d_0 \in \bar{\mathcal{D}}_q$ be a continuously accessible point. Then there exists an $\alpha_0 > 0$ (depending on d_0) and a path $(f_\alpha)_{\alpha\in]0,\alpha_0]}$ such that $h(f_\alpha, f^\star) = \alpha$ for all $\alpha \in]0,\alpha_0]$ and $d_{f_\alpha} \to d_0$ in $L^2(f^\star d\mu)$ as $\alpha \to 0$. Now choose the sequence

$$\alpha_n = \{(\langle d_0, g\rangle)_+ + \sigma_n^{-1}\} \sqrt{\frac{\log\log\sigma_n}{2\sigma_n}}.$$

As $(\langle d_0, g\rangle)_+ \leq \|d_0\|_2 \|g\|_2 \leq 1$, we clearly have

$$0 < \alpha_n < \alpha_0 \wedge 4\sqrt{\log\log\sigma_n/\sigma_n}$$

for all n sufficiently large. In particular $f_{\alpha_n} \in \mathcal{M}_q(4\sqrt{\log\log\sigma_n/\sigma_n})$, so that

$$\sup_{f\in\mathcal{M}_q(4\sqrt{\log\log\sigma_n/\sigma_n})} \left\{ 2\,\langle d_f, g\rangle\, h(f,f^\star) \sqrt{\frac{2\sigma_n}{\log\log\sigma_n}} - h(f,f^\star)^2 \, \frac{2\sigma_n}{\log\log\sigma_n} \right\}$$
$$\geq 2\,\langle d_{f_{\alpha_n}}, g\rangle\, \{(\langle d_0,g\rangle)_+ + \sigma_n^{-1}\} - \{(\langle d_0,g\rangle)_+ + \sigma_n^{-1}\}^2.$$

Therefore, we have

$$\limsup_{n\to\infty} \left\{ \sup_{d\in\bar{\mathcal{D}}_q^c} (\langle d,g\rangle)_+^2 - \sup_{f\in\mathcal{M}_q(4\sqrt{\log\log\sigma_n/\sigma_n})} \left\{ - h(f,f^\star)^2 \, \frac{2\sigma_n}{\log\log\sigma_n} \right.\right.$$
$$\left.\left. + 2\,\langle d_f, g\rangle\, h(f,f^\star) \sqrt{\frac{2\sigma_n}{\log\log\sigma_n}} \right\} \right\} \leq \sup_{d\in\bar{\mathcal{D}}_q^c} (\langle d,g\rangle)_+^2 - (\langle d_0,g\rangle)_+^2$$

for any continuously accessible element $d_0 \in \bar{\mathcal{D}}_q$. But clearly we can choose d_0 to make the right hand side of this expression arbitrarily small. Thus we have the desired contradiction, completing the proof. □

We can now complete the proof of Theorem 2.5.

PROOF OF THEOREM 2.5. We obtain separately the lower and upper bounds.

Lower bound. By Propositions 5.7 and 5.8, we have

$$\limsup_{n\to\infty} \frac{1}{\log\log n} \left\{ \sup_{f\in\mathcal{M}_q} \ell_n(f) - \sup_{f\in\mathcal{M}_p} \ell_n(f) \right\} \ge \limsup_{n\to\infty} \left\{ \sup_{d\in\bar{\mathcal{D}}_q^c} (I_n(d))_+^2 - \sup_{d\in\bar{\mathcal{D}}_p} (I_n(d))_+^2 \right\} \quad \mathbf{P}^\star\text{-a.s.}$$

Now fix any $g \in L_0^2(f^\star d\mu)$. By Theorem 5.3 (which applies here as $\mathcal{N}(\mathcal{D}_q,\delta) = \mathcal{N}(\operatorname{cl}\mathcal{D}_q,\delta) \ge \mathcal{N}(\bar{\mathcal{D}}_q,\delta)$ for all $\delta > 0$), there is a sequence $\tau_n \uparrow \infty$ of random times such that $I_{\tau_n} \to \langle\,\cdot\,, g\rangle$ in $\ell_\infty(\bar{\mathcal{D}}_q)$ $\mathbf{P}^\star$-a.s. Therefore

$$\sup_{d\in\bar{\mathcal{D}}_q^c} (I_{\tau_n}(d))_+^2 - \sup_{d\in\bar{\mathcal{D}}_p} (I_{\tau_n}(d))_+^2 \xrightarrow{n\to\infty} \sup_{d\in\bar{\mathcal{D}}_q^c} (\langle d,g\rangle)_+^2 - \sup_{d\in\bar{\mathcal{D}}_p} (\langle d,g\rangle)_+^2 \quad \mathbf{P}^\star\text{-a.s.},$$

so that certainly

$$\limsup_{n\to\infty} \frac{1}{\log\log n} \left\{ \sup_{f\in\mathcal{M}_q} \ell_n(f) - \sup_{f\in\mathcal{M}_p} \ell_n(f) \right\} \ge \sup_{d\in\bar{\mathcal{D}}_q^c} (\langle d,g\rangle)_+^2 - \sup_{d\in\bar{\mathcal{D}}_p} (\langle d,g\rangle)_+^2$$

$\mathbf{P}^\star$-a.s. But as this inequality holds for every $g \in L_0^2(f^\star d\mu)$, taking the supremum over g gives the requisite lower bound.

Upper bound. By Propositions 5.7 and 5.8, we have

$$\limsup_{n\to\infty} \frac{1}{\log\log n} \left\{ \sup_{f\in\mathcal{M}_q} \ell_n(f) - \sup_{f\in\mathcal{M}_p} \ell_n(f) \right\} \le \limsup_{n\to\infty} \left\{ \sup_{d\in\bar{\mathcal{D}}_q} (I_n(d))_+^2 - \sup_{d\in\bar{\mathcal{D}}_p^c} (I_n(d))_+^2 \right\} \quad \mathbf{P}^\star\text{-a.s.}$$

It is elementary that for any $d, d' \in \bar{\mathcal{D}}_q$ and $g \in L_0^2(f^\star d\mu)$

$$\begin{aligned}
&(I_n(d))_+^2 - (I_n(d'))_+^2 \\
&\quad \le |(I_n(d))_+^2 - (\langle d,g\rangle)_+^2| + |(I_n(d'))_+^2 - (\langle d',g\rangle)_+^2| + (\langle d,g\rangle)_+^2 - (\langle d',g\rangle)_+^2 \\
&\quad \le 2 \sup_{d\in\bar{\mathcal{D}}_q} |(I_n(d))_+^2 - (\langle d,g\rangle)_+^2| + (\langle d,g\rangle)_+^2 - (\langle d',g\rangle)_+^2.
\end{aligned}$$

Taking the supremum over $d \in \bar{\mathcal{D}}_q$ and the infimum over $d' \in \bar{\mathcal{D}}_p^c$, we find that

$$\begin{aligned}
&\sup_{d\in\bar{\mathcal{D}}_q} (I_n(d))_+^2 - \sup_{d\in\bar{\mathcal{D}}_p^c} (I_n(d))_+^2 \\
&\quad\le 2 \sup_{d\in\bar{\mathcal{D}}_q} |(I_n(d))_+^2 - (\langle d, g\rangle)_+^2| + \sup_{d\in\bar{\mathcal{D}}_q} (\langle d, g\rangle)_+^2 - \sup_{d\in\bar{\mathcal{D}}_p^c} (\langle d, g\rangle)_+^2 \\
&\quad\le 2 \sup_{d\in\bar{\mathcal{D}}_q} |(I_n(d))_+^2 - (\langle d, g\rangle)_+^2| \\
&\qquad\qquad + \sup_{g\in L_0^2(f^\star d\mu)} \left\{ \sup_{d\in\bar{\mathcal{D}}_q} (\langle d, g\rangle)_+^2 - \sup_{d\in\bar{\mathcal{D}}_p^c} (\langle d, g\rangle)_+^2 \right\}.
\end{aligned}$$

But as this holds for any $g \in L_0^2(f^\star d\mu)$, we finally obtain

$$\begin{aligned}
\sup_{d\in\bar{\mathcal{D}}_q} (I_n(d))_+^2 - \sup_{d\in\bar{\mathcal{D}}_p^c} (I_n(d))_+^2 \le 2 \inf_{g\in L_0^2(f^\star d\mu)} \sup_{d\in\bar{\mathcal{D}}_q} |(I_n(d))_+^2 - (\langle d, g\rangle)_+^2| \\
+ \sup_{g\in L_0^2(f^\star d\mu)} \left\{ \sup_{d\in\bar{\mathcal{D}}_q} (\langle d, g\rangle)_+^2 - \sup_{d\in\bar{\mathcal{D}}_p^c} (\langle d, g\rangle)_+^2 \right\}.
\end{aligned}$$

It follows as in the proof of Proposition 5.7 that the first term in this expression converges to zero $\mathbf{P}^\star$-a.s. The requisite upper bound follows immediately. □

Finally, we now complete the proof of Corollary 2.6

PROOF OF COROLLARY 2.6. It evidently suffices to prove that

$$\Gamma := \sup_{g\in L_0^2(f^\star d\mu)} \left\{ \sup_{d\in\bar{\mathcal{D}}_q^c} (\langle d, g\rangle)_+^2 - \sup_{d\in\bar{\mathcal{D}}_{q^\star}} (\langle d, g\rangle)_+^2 \right\} > 0. \tag{5.3}$$

To this end, note that by direct computation

$$\langle 1, d_f\rangle = \frac{\int \sqrt{f f^\star}\, d\mu - 1}{h(f, f^\star)} = -\frac{h(f, f^\star)}{2}.$$

Choose $(f_n)_{n\ge 0} \subset \mathcal{M}_q \backslash \{f^\star\}$ such that $h(f_n, f^\star) \to 0$ and $d_{f_n} \to d_0 \in \bar{\mathcal{D}}_q$, then

$$\langle 1, d_0\rangle = \lim_{n\to\infty} \langle 1, d_{f_n}\rangle = -\lim_{n\to\infty} \frac{h(f_n, f^\star)}{2} = 0.$$

Moreover, it is immediate that $\|d_0\|_2 \le 1$. We have therefore shown that $\bar{\mathcal{D}}_q \subset L_0^2(f^\star d\mu)$. Now choose $g \in \bar{\mathcal{D}}_q^c \backslash \bar{\mathcal{D}}_{q^\star}$. As $\bar{\mathcal{D}}_{q^\star}$ is closed, it follows directly that

$$\sup_{d\in\bar{\mathcal{D}}_q^c} (\langle d, g\rangle)_+^2 = 1, \qquad \sup_{d\in\bar{\mathcal{D}}_{q^\star}} (\langle d, g\rangle)_+^2 < 1.$$

Therefore (5.3) holds, and the proof is complete. □

5.3. *Proof of Theorem 3.1.*

PROOF OF THEOREM 3.1. The assumption implies that

$$\mathcal{N}(\mathcal{D},\varepsilon) \le \left(\frac{C_0}{\varepsilon \wedge \varepsilon_0}\right)^q \quad \text{for every } \varepsilon > 0.$$

If $\varepsilon < \|R\|_2/4$, then

$$\frac{\varepsilon}{\varepsilon \wedge \varepsilon_0} \le 1 \vee \frac{\|R\|_2}{4\varepsilon_0}.$$

Defining $C = C_0(1 \vee \|R\|_2/4\varepsilon_0)$, we find that

$$\mathcal{N}(\mathcal{D},\varepsilon) \le \left(\frac{C}{\varepsilon}\right)^q \quad \text{for every } \varepsilon < \|R\|_2/4.$$

The remainder of the proof is devoted to establishing that

$$\mathcal{N}(\mathcal{H}(\delta),\rho) \le \left(\frac{8C\delta}{\rho}\right)^{q+1}$$

for all $\delta,\rho > 0$ such that $\rho/\delta < 4 \wedge 2\|R\|_2$, which is the desired result.

Fix $\varepsilon,\delta > 0$ and let $N = \mathcal{N}(\mathcal{D},\varepsilon)$. Then there exist $l_1,u_1,\ldots,l_N,u_N$ such that $\|u_i - l_i\|_2 \le \varepsilon$ for all i and for every f, there is an i such that $l_i \le d_f \le u_i$. Choose f such that $r^{-n}\delta \le h(f,f^\star) \le r^{-n+1}\delta$ (with $r > 1$). Then there is an i so that

$$(r^{-n} l_i \wedge r^{-n+1} l_i)\,\delta + 1 \le \sqrt{f/f^\star} \le (r^{-n} u_i \vee r^{-n+1} u_i)\,\delta + 1.$$

Note that

$$\begin{aligned}
\|u_i\, r^{-n}\delta - l_i\, r^{-n}\delta\|_2 &\le r^{-n}\delta\varepsilon,\\
\|u_i\, r^{-n+1}\delta - l_i\, r^{-n+1}\delta\|_2 &\le r^{-n+1}\delta\varepsilon,\\
\|u_i\, r^{-n+1}\delta - l_i\, r^{-n}\delta\|_2 &\le (r-1)r^{-n}\delta + r^{-n+1}\delta\varepsilon,\\
\|u_i\, r^{-n}\delta - l_i\, r^{-n+1}\delta\|_2 &\le (r-1)r^{-n}\delta + r^{-n+1}\delta\varepsilon,
\end{aligned}$$

where the latter two estimates follow from $l_i \le d_f \le u_i$, $\|d_f\|_2 = 1$, and

$$\begin{aligned}
(u_i - l_i)\, r^{-n}\delta &\le u_i\, r^{-n+1}\delta - l_i\, r^{-n}\delta - d_f\,(r-1)r^{-n}\delta \le (u_i - l_i)\, r^{-n+1}\delta,\\
(u_i - l_i)\, r^{-n}\delta &\le u_i\, r^{-n}\delta - l_i\, r^{-n+1}\delta + d_f\,(r-1)r^{-n}\delta \le (u_i - l_i)\, r^{-n+1}\delta.
\end{aligned}$$

As $|a \vee b - c \wedge d| \le |a-c| + |a-d| + |b-c| + |b-d|$, we can estimate

$$\|(r^{-n} u_i \vee r^{-n+1} u_i)\,\delta - (r^{-n} l_i \wedge r^{-n+1} l_i)\,\delta\|_2 \le 2(r-1)r^{-n}\delta + 4r^{-n+1}\delta\varepsilon.$$

Therefore, we have shown that

$$\mathcal{N}(\{\sqrt{f/f^\star} : r^{-n}\delta \le h(f,f^\star) \le r^{-n+1}\delta\}, 2(r-1)r^{-n}\delta + 4r^{-n+1}\delta\varepsilon) \le \mathcal{N}(\mathcal{D},\varepsilon)$$

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

$$\mathcal{N}(\{\sqrt{f/f^\star} : r^{-n}\delta \le h(f,f^\star) \le r^{-n+1}\delta\}, \rho) \le \mathcal{N}(\mathcal{D}, \tfrac{1}{4}r^{n-1}\rho/\delta - \tfrac{1}{2}(1-1/r))$$

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

Note that, by finiteness of the bracketing entropies, we can choose an envelope function $R \ge \sup_f |d_f|$ such that $\|R\|_2 < \infty$. Then we evidently have

$$1 - r^{-n}\delta R \le \sqrt{f/f^\star} \le 1 + r^{-n}\delta R$$

whenever $h(f,f^\star) \le r^{-n}\delta$. Therefore

$$\mathcal{N}(\{\sqrt{f/f^\star} : h(f,f^\star) \le r^{-\lceil H \rceil}\delta\}, 2r^{-H}\delta\|R\|_2) = 1$$

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

$$\begin{aligned}
&\mathcal{N}(\{\sqrt{f/f^\star} : h(f,f^\star) \le \delta\}, 2r^{-H}\delta\|R\|_2) \\
&\quad \le 1 + \sum_{n=1}^{\lceil H \rceil} \mathcal{N}(\{\sqrt{f/f^\star} : r^{-n}\delta \le h(f,f^\star) \le r^{-n+1}\delta\}, 2r^{-H}\delta\|R\|_2) \\
&\quad \le 1 + \sum_{n=1}^{\lceil H \rceil} \mathcal{N}(\mathcal{D}, \{r^{n-H-1}\|R\|_2 - (1-1/r)\}/2)
\end{aligned}$$

whenever $\delta > 0$, $r > 1$, $H > 0$ such that $\|R\|_2 > (1-1/r)r^H$. In particular,

$$\mathcal{N}(\{\sqrt{f/f^\star} : h(f,f^\star) \le \delta\}, 2r^{-H}\delta\|R\|_2) \le 1 + \sum_{n=1}^{\lceil H \rceil} \mathcal{N}(\mathcal{D}, r^{n-H-1}\|R\|_2/4)$$

whenever $\delta > 0$, $r > 1$, $H > 0$ such that $\|R\|_2 \ge 2(1-1/r)r^H$, where we have used that the bracketing number is a nonincreasing function of the bracket size.

Now recall that

$$\mathcal{N}(\mathcal{D},\varepsilon) \le \left(\frac{C}{\varepsilon}\right)^q \quad \text{for every } 0 < \varepsilon < \|R\|_2/4,$$

where $q, C \ge 1$. Thus

$$\mathcal{N}(\{\sqrt{f/f^\star} : h(f,f^\star) \le \delta\}, 2r^{-H}\delta\|R\|_2) \le 1 + \sum_{n=1}^{\lceil H \rceil} r^{-(n-1)q} \left(\frac{8C}{2r^{-H}\|R\|_2}\right)^q$$

whenever $\delta > 0$, $r > 1$, $H > 0$ such that $\|R\|_2 \ge 2(1-1/r)r^H$. But

$$\sum_{n=1}^{\lceil H \rceil} r^{-(n-1)q} \le \frac{1}{1-1/r^q} \le \frac{1}{1-1/r} \le \frac{\|R\|_2}{2(1-1/r)r^H} \frac{4C}{2r^{-H}\|R\|_2}$$

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

$$\mathcal{N}(\{\sqrt{f/f^\star} : h(f,f^\star) \le \delta\}, 2r^{-H}\delta\|R\|_2) \le \frac{\|R\|_2}{2(1-1/r)r^H} \left(\frac{8C}{2r^{-H}\|R\|_2}\right)^{q+1}$$

whenever $\delta > 0$, $r > 1$, $H > 0$ such that $\|R\|_2 \ge 2(1-1/r)r^H$.

We now fix $\delta, \rho > 0$ such that $\rho/\delta < 4 \wedge 2\|R\|_2$, and choose

$$r = \frac{4}{4-\rho/\delta}, \qquad H = \frac{\log(2\|R\|_2\delta/\rho)}{\log r}.$$

Clearly $r > 1$ and $H > 0$. Moreover, note that our choice of r and H implies that $\|R\|_2 = 2(1-1/r)r^H$ and $\rho = 2r^{-H}\delta\|R\|_2$. We have therefore shown that

$$\mathcal{N}(\{\sqrt{f/f^\star} : h(f,f^\star) \le \delta\}, \rho) \le \left(\frac{8C\delta}{\rho}\right)^{q+1}$$

for all $\delta, \rho > 0$ such that $\rho/\delta < 4 \wedge 2\|R\|_2$. □

5.4. *Proof of Theorem 3.3.*

5.4.1. *The local geometry of mixtures.* Define the Euclidean balls $B(\theta,\varepsilon) = \{\theta' \in \mathbb{R}^d : \|\theta - \theta'\| < \varepsilon\}$, denote by $\langle u, v\rangle$ the inner product of two vectors $u, v \in \mathbb{R}^d$, and denote by $\langle A, u\rangle = \{\langle \theta, u\rangle : \theta \in A\} \subseteq \mathbb{R}$ the inner product of a set $A \subseteq \mathbb{R}^d$ with a vector $u \in \mathbb{R}^d$.

LEMMA 5.9. *It is possible to choose a bounded convex neighborhood A_i of $\theta_i^\star$ for every $i = 1, \ldots, q^\star$ such that, for some linearly independent family $u_1, \ldots, u_d \in \mathbb{R}^d$, the sets $\{\langle A_i, u_j\rangle : i = 1, \ldots, q^\star\}$ are disjoint for every $j = 1, \ldots, d$.*

PROOF. We first claim that one can choose linearly independent $u_1, \ldots, u_d$ such that $|\{\langle \theta_i^\star, u_j\rangle : i = 1, \ldots, q^\star\}| = q^\star$ for every $j = 1, \ldots, d$. Indeed, note that the set $\{u \in \mathbb{R}^d : |\{\langle \theta_i^\star, u\rangle : i = 1, \ldots, q^\star\}| < q^\star\}$ 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 $\mathrm{SO}(d)$, and let $u_i = Te_i$ 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 $A_i = B(\theta_i^\star, \varepsilon/4)$ with $\varepsilon = \min_k \min_{i \ne j} |\langle \theta_i^\star - \theta_j^\star, u_k\rangle|$. □

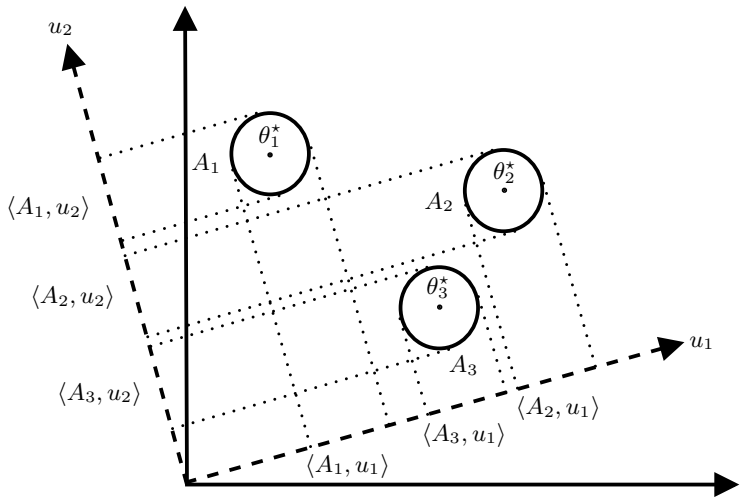

FIG 2. *Illustration of the construction of the sets A_i for a mixture with $d = 2$ and $q^\star = 3$. The sets A_i are chosen in such a way that their projections on some linearly independent vectors u_1, u_2 are disjoint. Note that the choice of u_1, u_2 is not arbitrary (e.g., consider the projections on the coordinate axes).*

We now fix once and for all a family of neighborhoods $A_1, \ldots, A_{q^\star}$ that satisfy the conditions of Lemma 5.9. The precise choice of these sets only affects the constants in the proofs below and is therefore irrelevant to our final result; we only presume that $A_1, \ldots, A_{q^\star}$ remain fixed throughout the proofs. Let us also define $A_0 = \mathbb{R}^d \backslash (A_1 \cup \cdots \cup A_{q^\star})$. Then $\{A_0, \ldots, A_{q^\star}\}$ partitions the parameter set $\mathbb{R}^d$ in such a way that each bounded element A_i, $i = 1, \ldots, q^\star$ contains precisely one component of the mixture $f^\star$, while the unbounded element A_0 contains no components of $f^\star$. This construction is illustrated in Figure 2.

Let us define for each finite measure λ on $\mathbb{R}^d$ the function

$$f_\lambda(x) = \int f_\theta(x)\, \lambda(d\theta).$$

We also define the derivatives $D_1 f_\theta(x) \in \mathbb{R}^d$ and $D_2 f_\theta(x) \in \mathbb{R}^{d\times d}$ as

$$[D_1 f_\theta(x)]_i = \frac{\partial}{\partial \theta^i} f_\theta(x), \qquad [D_2 f_\theta(x)]_{ij} = \frac{\partial^2}{\partial \theta^i \partial \theta^j} f_\theta(x).$$

Denote by $\mathfrak{P}(A)$ the space of probability measures supported on $A \subseteq \mathbb{R}^d$, and denote by M_+^d the family of all $d \times d$ positive semidefinite (symmetric) matrices.

DEFINITION 5.10. Let us write

$$\mathfrak{D} = \{(\eta,\beta,\rho,\tau,\nu) : \eta_1,\ldots,\eta_{q^\star} \in \mathbb{R},\ \beta_1,\ldots,\beta_{q^\star} \in \mathbb{R}^d,\ \rho_1,\ldots,\rho_{q^\star} \in M_+^d,\\ \tau_0,\ldots,\tau_{q^\star} \ge 0,\ \nu_0 \in \mathfrak{P}(A_0),\ldots,\nu_{q^\star} \in \mathfrak{P}(A_{q^\star})\}.$$

Then we define for each $(\eta,\beta,\rho,\tau,\nu) \in \mathfrak{D}$ the function

$$\ell(\eta,\beta,\rho,\tau,\nu) = \tau_0 \frac{f_{\nu_0}}{f^\star} + \sum_{i=1}^{q^\star} \left\{ \eta_i \frac{f_{\theta_i^\star}}{f^\star} + \beta_i^* \frac{D_1 f_{\theta_i^\star}}{f^\star} + \mathrm{Tr}\left[\rho_i \frac{D_2 f_{\theta_i^\star}}{f^\star}\right] + \tau_i \frac{f_{\nu_i}}{f^\star} \right\},$$

and the nonnegative quantity

$$N(\eta,\beta,\rho,\tau,\nu) = \tau_0 + \sum_{i=1}^{q^\star} |\eta_i + \tau_i| + \sum_{i=1}^{q^\star} \left\| \beta_i + \tau_i \int (\theta - \theta_i^\star)\, \nu_i(d\theta) \right\| + \\ \sum_{i=1}^{q^\star} \mathrm{Tr}[\rho_i] + \sum_{i=1}^{q^\star} \frac{\tau_i}{2} \int \|\theta - \theta_i^\star\|^2 \nu_i(d\theta).$$

We now formulate the key result on the local geometry of the mixture class $\mathcal{M}$.

THEOREM 5.11. *Suppose that*

1. $f_0 \in C^2$ *and* $f_0(x)$, $D_1 f_0(x)$ *vanish as* $\|x\| \to \infty$.
2. $\|[D_1 f_0]_i / f^\star\|_1 < \infty$ *and* $\|[D_2 f_0]_{ij} / f^\star\|_1 < \infty$ *for all* $i,j = 1,\ldots,d$.

Then there exists a constant $c^\star > 0$ *such that*

$$\|\ell(\eta,\beta,\rho,\tau,\nu)\|_1 \ge c^\star\, N(\eta,\beta,\rho,\tau,\nu) \quad \textit{for all } (\eta,\beta,\rho,\tau,\nu) \in \mathfrak{D}.$$

[*The constant* $c^\star$ *may depend on* $f^\star$ *and* $A_1,\ldots,A_{q^\star}$ *but not on* η,β,ρ,τ,ν.]

Before we turn to the proof, let us introduce a notion that is familiar in quantum physics. If (Ω,Σ) is a measurable space, call the map $\lambda : \Sigma \to \mathbb{R}^{d\times d}$ a *state*[2] 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. $\mathrm{Tr}[\lambda(\Omega)] = 1$.

[2]Our terminology is in analogy with the usual notion of a state on the C^*-algebra $\mathbb{C}^{d\times d} \otimes C_{\mathbb{C}}(\Omega)$, where Ω is a compact metric space and $C_{\mathbb{C}}(\Omega)$ is the algebra of complex-valued continuous functions on Ω. Such states are precisely 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 ξ_i such that $\langle \xi_i, \lambda(\Omega)\xi_i\rangle > 0$. 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 λ_n converges weakly to some state λ on B in the sense that $\int \mathrm{Tr}[M(\theta)\lambda_n(d\theta)] \to \int \mathrm{Tr}[M(\theta)\lambda(d\theta)]$ 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 $[\lambda_n]_{ij}$ 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 5.11. Suppose that the conclusion of the theorem does not hold. Then there must exist a sequence of coefficients $(\eta^n, \beta^n, \rho^n, \tau^n, \nu^n) \in \mathfrak{D}$ with

$$\frac{\|\ell(\eta^n, \beta^n, \rho^n, \tau^n, \nu^n)\|_1}{N(\eta^n, \beta^n, \rho^n, \tau^n, \nu^n)} \xrightarrow{n\to\infty} 0.$$

Let us fix such a sequence throughout the proof.

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

$$\begin{aligned}
&\eta_i^n \frac{f_{\theta_i^\star}}{f^\star} + \beta_i^{n*} \frac{D_1 f_{\theta_i^\star}}{f^\star} + \mathrm{Tr}\left[\rho_i^n \frac{D_2 f_{\theta_i^\star}}{f^\star}\right] + \tau_i^n \frac{f_{\nu_i^n}}{f^\star} \\
&= (\eta_i^n + \tau_i^n)\, \frac{f_{\theta_i^\star}}{f^\star} + \left(\beta_i^n + \tau_i^n \int (\theta - \theta_i^\star)\, \nu_i^n(d\theta)\right)^* \frac{D_1 f_{\theta_i^\star}}{f^\star} + \mathrm{Tr}\left[\rho_i^n \frac{D_2 f_{\theta_i^\star}}{f^\star}\right] \\
&\quad + \frac{\tau_i^n}{2} \int \|\theta - \theta_i^\star\|^2\, \nu_i^n(d\theta) \int \mathrm{Tr}\left[\left\{\int_0^1 \frac{D_2 f_{\theta_i^\star + u(\theta-\theta_i^\star)}}{f^\star}\, 2(1-u)\, du\right\} \lambda_i^n(d\theta)\right]
\end{aligned}$$

where λ_i^n is the state on A_i defined by

$$\int \mathrm{Tr}[M(\theta)\, \lambda_i^n(d\theta)] = \frac{\int \mathrm{Tr}[M(\theta)\, (\theta - \theta_i^\star)(\theta - \theta_i^\star)^*]\, \nu_i^n(d\theta)}{\int \|\theta - \theta_i^\star\|^2\, \nu_i^n(d\theta)}$$

(it is clearly no loss of generality to assume that ν_i^n has no mass at $\theta_i^\star$ for any i, n, so that everything is well defined). We now define the coefficients

$$a_i^n = \frac{\eta_i^n + \tau_i^n}{N(\eta^n, \beta^n, \rho^n, \tau^n, \nu^n)}, \qquad b_i^n = \frac{\beta_i^n + \tau_i^n \int (\theta - \theta_i^\star)\, \nu_i^n(d\theta)}{N(\eta^n, \beta^n, \rho^n, \tau^n, \nu^n)},$$
$$c_i^n = \frac{\rho_i^n}{N(\eta^n, \beta^n, \rho^n, \tau^n, \nu^n)}, \qquad d_i^n = \frac{\frac{\tau_i^n}{2} \int \|\theta - \theta_i^\star\|^2\, \nu_i^n(d\theta)}{N(\eta^n, \beta^n, \rho^n, \tau^n, \nu^n)}$$

for $i = 1, \ldots, q^\star$, and

$$a_0^n = \frac{\tau_0^n}{N(\eta^n, \beta^n, \rho^n, \tau^n, \nu^n)}.$$

Note that

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

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

1. There exist $a_i \in \mathbb{R}$, $b_i \in \mathbb{R}^d$, $c_i \in M_+^d$, and $a_0, d_i \geq 0$ (for $i = 1, \ldots, q^\star$) with $|a_0| + \sum_{i=1}^{q^\star} \{|a_i| + \|b_i\| + \mathrm{Tr}[c_i] + |d_i|\} = 1$, such that $a_0^n \to a_0$ and $a_i^n \to a_i$, $b_i^n \to b_i$, $c_i^n \to c_i$, $d_i^n \to d_i$ as $n \to \infty$ for all $i = 1, \ldots, q^\star$.
2. There exists a sub-probability measure ν_0 supported on A_0, such that ν_0^n converges vaguely to ν_0 as $n \to \infty$.
3. There exist states λ_i supported on $\mathrm{cl}\, A_i$ for $i = 1, \ldots, q^\star$, such that λ_i^n converges weakly to λ_i as $n \to \infty$ for every $i = 1, \ldots, q^\star$.

It follows that the functions $\ell(\eta^n, \beta^n, \rho^n, \tau^n, \nu^n)/N(\eta^n, \beta^n, \rho^n, \tau^n, \nu^n)$ converge pointwise along this subsequence to the function $h/f^\star$ defined by

$$h = a_0\, f_{\nu_0} + \sum_{i=1}^{q^\star} \Bigg\{ a_i\, f_{\theta_i^\star} + b_i^*\, D_1 f_{\theta_i^\star} + \mathrm{Tr}[c_i\, D_2 f_{\theta_i^\star}] + d_i \int \mathrm{Tr}\left[\left\{ \int_0^1 D_2 f_{\theta_i^\star + u(\theta - \theta_i^\star)}\, 2(1-u)\, du \right\} \lambda_i(d\theta) \right] \Bigg\}.$$

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

To proceed, we need the following lemma.

LEMMA 5.12. *The Fourier transform $F[h](s) := \int e^{\mathrm{i}\langle x, s\rangle} h(x) dx$ is given by*

$$F[h](s) = F[f_0](s) \Bigg[a_0 \int e^{\mathrm{i}\langle \theta, s\rangle}\, \nu_0(d\theta) + \sum_{i=1}^{q^\star} \Bigg\{ a_i\, e^{\mathrm{i}\langle \theta_i^\star, s\rangle} + \mathrm{i}\langle b_i, s\rangle\, e^{\mathrm{i}\langle \theta_i^\star, s\rangle} - \langle s, c_i s\rangle\, e^{\mathrm{i}\langle \theta_i^\star, s\rangle} - d_i\, e^{\mathrm{i}\langle \theta_i^\star, s\rangle} \int \phi(\mathrm{i}\langle \theta - \theta_i^\star, s\rangle)\, \langle s, \lambda_i(d\theta) s\rangle \Bigg\} \Bigg]$$

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

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

We begin by noting that

$$\int\int\int_0^1 |[D_2 f_{\theta_i^\star + u(\theta-\theta_i^\star)}(x)]_{jk}|\, 2(1-u)\, du\, dx\, |[\lambda_i]_{kj}|(d\theta) = \\ \|[\lambda_i]_{kj}\|_{\mathrm{TV}} \int |[D_2 f_0(x)]_{jk}|\, dx < \infty.$$

We may therefore apply Fubini's theorem, giving

$$F[[\Xi_i]_{jk}](s) = -F[f_0](s)\, s_j s_k\, e^{\mathrm{i}\langle\theta_i^\star, s\rangle} \int \Big\{ \int_0^1 e^{\mathrm{i}u\langle\theta-\theta_i^\star, s\rangle} 2(1-u) du \Big\} [\lambda_i(d\theta)]_{kj} \\ = -F[f_0](s)\, s_j s_k\, e^{\mathrm{i}\langle\theta_i^\star, s\rangle} \int \phi(\mathrm{i}\langle\theta-\theta_i^\star, s\rangle)\, [\lambda_i(d\theta)]_{kj},$$

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 5.9. As $F[h](s) = 0$ for all $s \in \mathbb{R}^d$, we obtain

$$\Phi^\ell(\mathrm{i}t) := a_0\, \Phi_0^\ell(\mathrm{i}t) + \sum_{i=1}^{q^\star} e^{\mathrm{i}t\langle\theta_i^\star, u_\ell\rangle} \big\{ a_i + \mathrm{i}t\langle b_i, u_\ell\rangle - t^2\langle u_\ell, c_i u_\ell\rangle - d_i\, t^2\, \Phi_i^\ell(\mathrm{i}t) \big\} = 0$$

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

$$\Phi_i^\ell(\mathrm{i}t) = \int \phi(\mathrm{i}t\langle\theta-\theta_i^\star, u_\ell\rangle)\, \langle u_\ell, \lambda_i(d\theta) u_\ell\rangle$$

for $i = 1, \ldots, q^\star$, and

$$\Phi_0^\ell(\mathrm{i}t) = \int e^{\mathrm{i}t\langle\theta, u_\ell\rangle}\, \nu_0(d\theta).$$

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 5.12 and the fact that $F[f_0](s)$ is nonvanishing in a sufficiently small neighborhood of the origin.

As all λ_i have compact support, it is easily seen that for every $i = 1, \ldots, q^\star$, the function $\Phi_i^\ell(z)$ is defined for all $z \in \mathbb{C}$ by a convergent power series. The function $\Psi^\ell(\mathrm{i}t) := \Phi^\ell(\mathrm{i}t) - a_0\, \Phi_0^\ell(\mathrm{i}t)$ is therefore an entire function with $|\Psi^\ell(z)| \le k_1 e^{k_2|z|}$ for some $k_1, k_2 > 0$ and all $z \in \mathbb{C}$. But as $\Phi^\ell(\mathrm{i}t) = 0$ for $t \in [-\iota, \iota]$, it follows from [20], Theorem 7.2.2 that $a_0\, \Phi_0^\ell(\mathrm{i}t)$ is the Fourier transform of a finite measure with compact support. Thus we may assume without loss of generality that the law of $\langle\theta, u_\ell\rangle$ under the sub-probability ν_0 is compactly supported for every $\ell = 1, \ldots, d$, so by linear independence ν_0 must be compactly supported. Therefore, the function

$\Phi^\ell(z)$ is defined for all $z \in \mathbb{C}$ by a convergent power series. But as $\Phi^\ell(z)$ vanishes for $z \in \mathrm{i}[-\iota,\iota]$, we must have $\Phi^\ell(z) = 0$ for all $z \in \mathbb{C}$, and in particular

(5.4)
$$\Phi^\ell(t) = a_0\,\Phi_0^\ell(t) + \sum_{i=1}^{q^\star} e^{t\langle\theta_i^\star,u_\ell\rangle}\big\{a_i + t\langle b_i,u_\ell\rangle + t^2\langle u_\ell,c_iu_\ell\rangle + d_i\,t^2\,\Phi_i^\ell(t)\big\} = 0$$

for all $t \in \mathbb{R}$ and $\ell = 1,\ldots,d$. In the remainder of the proof, we argue that (5.4) can not hold, thus completing the proof by contradiction.

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

PROPOSITION 5.13. *Fix $\ell \in \{1,\ldots,d\}$, and define*

$$\tilde\Phi_0^\ell(t) := a_0\,\Phi_0^\ell(t) + \sum_{i=1}^{q^\star} a_i\,e^{t\langle\theta_i^\star,u_\ell\rangle}.$$

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

$$\Phi^{\ell,j}(t) := \tilde\Phi_0^\ell(t) + \sum_{i=1}^{j} e^{t\langle\theta^\star_{(\ell i)},u_\ell\rangle}\big\{t\langle b_{(\ell i)},u_\ell\rangle + t^2\langle u_\ell,c_{(\ell i)}u_\ell\rangle + d_{(\ell i)}\,t^2\,\Phi^\ell_{(\ell i)}(t)\big\}.$$

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

PROOF. Let us write for simplicity $\theta_i^\ell = \langle\theta_i^\star,u_\ell\rangle$, and denote by λ_i^ℓ and ν_0^ℓ the finite measures on $\mathbb{R}$ defined such that $\int f(x)\lambda_i^\ell(dx) = \int f(\langle\theta,u_\ell\rangle)\langle u_\ell,\lambda_i(d\theta)u_\ell\rangle$ 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 $\nu_0^\ell(\{\theta_i^\ell\}) = 0$ for all $i = 1,\ldots,q^\star$. This entails no loss of generality: the former can always be attained by relabeling of the points $\theta_i^\star$, while $\tilde\Phi_0^\ell$ is unchanged if we replace ν_0^ℓ and a_i by $\nu_0^\ell(\,\cdot\,\cap\mathbb{R}\backslash\{\theta_1^\ell,\ldots,\theta_{q^\star}^\ell\})$ and $a_i + a_0\,\nu_0^\ell(\{\theta_i^\ell\})$, respectively. Note that

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

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:

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

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

$$\liminf_{t\to\infty} \frac{|\tilde{\Phi}_0^\ell(t)|}{e^{t\theta_{j+1}^\ell}} > 0,$$

where we have used that ν_0^ℓ has no mass at $\{\theta_1^\ell, \dots, \theta_{q^\star}^\ell\}$. On the other hand, as ϕ is positive and increasing and as λ_i is supported on $\operatorname{cl} A_i$, we can estimate

$$0 \le \frac{t^2\, e^{t\theta_i^\ell}\, \Phi_i^\ell(t)}{e^{t\theta_{j+1}^\ell}} \le t^2\, e^{-t(\theta_{j+1}^\ell - \theta_i^\ell)}\, \phi(t\{\theta_j^{\ell+} - \theta_i^\ell\})\, \lambda_i^\ell(\mathbb{R}) \xrightarrow{t\to\infty} 0$$

for $i = 1, \dots, j$. But then we must have

$$0 = \liminf_{t\to\infty} \frac{|\Phi^{\ell,j}(t)|}{e^{t\theta_{j+1}^\ell}} > 0,$$

which yields the desired contradiction.

Step 2. We claim that the following hold:

$$d_j \lambda_j^\ell([\theta_j^\ell, \infty[) = 0, \quad \langle u_\ell, c_j u_\ell\rangle = 0, \quad \text{and} \quad a_0\, \nu_0^\ell([\theta_j^\ell, \infty[) = 0.$$

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

$$\begin{aligned} a_0\, \Phi_0^\ell(t) + e^{t\theta_j^\ell}\big\{t^2\langle u_\ell, c_j u_\ell\rangle + d_j\, t^2\, \Phi_j^\ell(t)\big\} \ge \\ t^2\, e^{t\theta_j^\ell} \left\{ \frac{\varepsilon^2}{4}\, a_0\, \nu_0^\ell([\theta_j^\ell, \infty[) + \langle u_\ell, c_j u_\ell\rangle + d_j\, \lambda_j^\ell([\theta_j^\ell, \infty[) \right\} > 0 \end{aligned}$$

for all $t \ge 0$. On the other hand, it is easily seen that

$$\frac{1}{t^2\, e^{t\theta_j^\ell}} \left[\sum_{i=1}^{j} e^{t\theta_i^\ell}\big\{a_i + t\langle b_i, u_\ell\rangle\big\} + \sum_{i=1}^{j-1} e^{t\theta_i^\ell}\big\{t^2\langle u_\ell, c_i u_\ell\rangle + d_i\, t^2\, \Phi_i^\ell(t)\big\} \right] \xrightarrow{t\to\infty} 0.$$

But this would imply that

$$0 = \lim_{t\to\infty} \frac{\Phi^{\ell,j}(t)}{a_0\, \Phi_0^\ell(t) + e^{t\theta_j^\ell}\{t^2\langle u_\ell, c_j u_\ell\rangle + d_j\, t^2\, \Phi_j^\ell(t)\}} = 1,$$

which yields the desired contradiction.

Step 3. We claim that the following hold:

$$d_j\, \lambda_j^\ell([\theta_j^{\ell-}, \theta_j^\ell[) = 0 \quad \text{and} \quad a_0\, \nu_0^\ell([\theta_j^{\ell-}, \theta_j^\ell[) = 0.$$

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

$$0 = \frac{d^2}{dt^2}\left(\frac{\Phi^{\ell,j}(t)}{e^{t\theta_j^\ell}}\right) = d_j \int e^{t(\theta-\theta_j^\ell)}\,\lambda_j^\ell(d\theta) + a_0 \int e^{t(\theta-\theta_j^\ell)}\,(\theta-\theta_j^\ell)^2\,\nu_0^\ell(d\theta) + \sum_{i=1}^{j-1} \frac{d^2}{dt^2}\, e^{-t(\theta_j^\ell-\theta_i^\ell)}\big\{a_i + t\langle b_i, u_\ell\rangle + t^2\langle u_\ell, c_i u_\ell\rangle + d_i\, t^2\,\Phi_i^\ell(t)\big\},$$

where the derivative and integral may be exchanged by [28], Appendix A16. We now note that as $a_0, d_j \geq 0$, we can estimate for $t \geq 0$

$$d_j \int e^{t(\theta-\theta_j^\ell)}\,\lambda_j^\ell(d\theta) + a_0 \int e^{t(\theta-\theta_j^\ell)}\,(\theta-\theta_j^\ell)^2\,\nu_0^\ell(d\theta) \geq e^{t(\theta_j^{\ell-}-\theta_j^\ell)}\left\{d_j\,\lambda_j^\ell([\theta_j^{\ell-},\theta_j^\ell[) + a_0\int_{[\theta_j^{\ell-},\theta_j^\ell[}(\theta-\theta_j^\ell)^2\,\nu_0^\ell(d\theta)\right\} > 0.$$

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

$$\begin{aligned}
&e^{-t(\theta_j^{\ell-}-\theta_j^\ell)}\left|\frac{d^2}{dt^2}\,e^{-t(\theta_j^\ell-\theta_i^\ell)}\,t^2\,\Phi_i^\ell(t)\right| \\
&\quad= e^{-t(\theta_j^{\ell-}-\theta_j^\ell)} \times e^{-t(\theta_j^\ell-\theta_i^\ell)} \times \Bigg|(\theta_j^\ell-\theta_i^\ell)^2\int t^2\phi(t\{\theta-\theta_i^\ell\})\,\lambda_i^\ell(d\theta) \\
&\qquad\quad - 2(\theta_j^\ell-\theta_i^\ell)\int\frac{e^{t(\theta-\theta_i^\ell)}-1}{\theta-\theta_i^\ell}\,\lambda_i^\ell(d\theta) + \int e^{t(\theta-\theta_i^\ell)}\,\lambda_i^\ell(d\theta)\Bigg| \\
&\quad\leq e^{-t(\theta_j^{\ell-}-\theta_i^\ell)}\Bigg\{(\theta_j^\ell-\theta_i^\ell)^2\,t^2\,\phi(t\{\theta_i^{\ell+}-\theta_i^\ell\}) \\
&\qquad\quad + 2\,(\theta_j^\ell-\theta_i^\ell)\,\frac{e^{t(\theta_i^{\ell+}-\theta_i^\ell)}-1}{\theta_i^{\ell+}-\theta_i^\ell} + e^{t(\theta_i^{\ell+}-\theta_i^\ell)}\Bigg\}\,\lambda_i^\ell(\mathbb{R}),
\end{aligned}$$

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

$$0 = \lim_{t\to\infty}\frac{\frac{d^2}{dt^2}\left(\Phi^{\ell,j}(t)/e^{t\theta_j^\ell}\right)}{d_j\int e^{t(\theta-\theta_j^\ell)}\,\lambda_j^\ell(d\theta) + a_0\int e^{t(\theta-\theta_j^\ell)}\,(\theta-\theta_j^\ell)^2\,\nu_0^\ell(d\theta)} = 1,$$

which yields the desired contradiction.

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

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

It is therefore easily seen that

$$0 = \lim_{t\to\infty} \frac{\Phi^{\ell,j}(t)}{t\, e^{t\theta_j^\ell}} = \langle b_j, u_\ell \rangle.$$

Thus the proof is complete. □

We can now perform the induction by starting from (5.4) and applying Proposition 5.13 repeatedly. This yields $d_j\langle u_\ell, \lambda_j(\mathbb{R}^d)u_\ell\rangle = \langle u_\ell, c_j u_\ell\rangle = \langle b_j, u_\ell\rangle = 0$ for all $j = 1, \ldots, q^\star$ and $\ell = 1, \ldots, d$. As $u_1, \ldots, u_d$ are linearly independent and $c_j \in M_+^d$, this implies that $b_j = 0$, $c_j = 0$ and $d_j = 0$ for all $j = 1, \ldots, q^\star$, so that

$$a_0 \int e^{\mathrm{i}\langle\theta,s\rangle}\, \nu_0(d\theta) + \sum_{i=1}^{q^\star} a_i\, e^{\mathrm{i}\langle\theta_i^\star,s\rangle} = 0$$

for all $s \in \mathbb{R}^d$ (this follows as above by Lemma 5.12, $h \equiv 0$, $F[f_0](s) \neq 0$ for s in a neighborhood of the origin, and using analyticity). But by the uniqueness of Fourier transforms, this implies that the signed measure $a_0\,\nu_0 + \sum_{i=1}^{q^\star} a_i\,\delta_{\{\theta_i^\star\}}$ has no mass. As ν_0 is supported on A_0, this implies that $a_j = 0$ for all $j = 1, \ldots, q^\star$. We have therefore shown that $a_i, b_i, c_i, d_i = 0$ for all $i = 1, \ldots, q^\star$. But recall that $|a_0| + \sum_{i=1}^{q^\star}\{|a_i| + \|b_i\| + \mathrm{Tr}[c_i] + |d_i|\} = 1$, so that evidently $a_0 = 1$.

To complete the proof, it remains to note that

$$\int \frac{\ell(\eta^n, \beta^n, \rho^n, \tau^n, \nu^n)}{N(\eta^n, \beta^n, \rho^n, \tau^n, \nu^n)}\, f^\star d\mu = \sum_{i=0}^{q^\star} a_i^n \xrightarrow{n\to\infty} 1.$$

But this is impossible, as

$$\left\| \frac{\ell(\eta^n, \beta^n, \rho^n, \tau^n, \nu^n)}{N(\eta^n, \beta^n, \rho^n, \tau^n, \nu^n)} \right\|_1 \xrightarrow{n\to\infty} 0$$

by construction. Thus we have the desired contradiction. □

5.4.2. *Proof of Theorem 3.3.* The proof of Theorem 3.3 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 5.14. *Define* $S = (H_0 + H_1 + H_2)\, d/c^\star$. *Then* $S \in L^4(f^\star d\mu)$, *and*

$$\frac{|f/f^\star - 1|}{\|f/f^\star - 1\|_1} \leq S \quad \textit{for all } f \in \mathcal{M}.$$

PROOF. That $S \in L^4(f^\star d\mu)$ 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

$$\frac{f-f^\star}{f^\star} = \sum_{j:\theta_j \in A_0} \pi_j \frac{f_{\theta_j}}{f^\star} + \sum_{i=1}^{q^\star} \left\{ \left(\sum_{j:\theta_j \in A_i} \pi_j - \pi_i^\star \right) \frac{f_{\theta_i^\star}}{f^\star} + \sum_{j:\theta_j \in A_i} \pi_j \, \frac{f_{\theta_j} - f_{\theta_i^\star}}{f^\star} \right\}.$$

Taylor expansion gives

$$f_{\theta_j}(x) - f_{\theta_i^\star}(x) = (\theta_j - \theta_i^\star)^* D_1 f_{\theta_i^\star}(x) + \frac{1}{2} \int_0^1 (\theta_j - \theta_i^\star)^* D_2 f_{\theta_i^\star + u(\theta_j - \theta_i^\star)}(x) \, (\theta_j - \theta_i^\star) \, 2(1-u) \, du.$$

Using Assumption A, we find that

$$\left| \frac{f-f^\star}{f^\star} \right| \le \left[\sum_{j:\theta_j \in A_0} \pi_j + \sum_{i=1}^{q^\star} \left\{ \left| \sum_{j:\theta_j \in A_i} \pi_j - \pi_i^\star \right| + \left\| \sum_{j:\theta_j \in A_i} \pi_j (\theta_j - \theta_i^\star) \right\| + \frac{1}{2} \sum_{j:\theta_j \in A_i} \pi_j \|\theta_j - \theta_i^\star\|^2 \right\} \right] (H_0 + H_1 + H_2) \, d.$$

On the other hand, Theorem 5.11 gives

$$\left\| \frac{f-f^\star}{f^\star} \right\|_1 \ge c^\star \left[\sum_{j:\theta_j \in A_0} \pi_j + \sum_{i=1}^{q^\star} \left\{ \left| \sum_{j:\theta_j \in A_i} \pi_j - \pi_i^\star \right| + \left\| \sum_{j:\theta_j \in A_i} \pi_j (\theta_j - \theta_i^\star) \right\| + \frac{1}{2} \sum_{j:\theta_j \in A_i} \pi_j \|\theta_j - \theta_i^\star\|^2 \right\} \right].$$

The proof follows directly. □

COROLLARY 5.15. *$|d| \le D$ for all $d \in \mathcal{D}$, where $D = 2S \in L^4(f^\star d\mu)$.*

PROOF. Using $\|f - f^\star\|_{\mathrm{TV}} \le 2h(f, f^\star)$ and $|\sqrt{x} - 1| \le |x - 1|$, we find

$$|d_f| = \frac{|\sqrt{f/f^\star} - 1|}{h(f, f^\star)} \le \frac{|f/f^\star - 1|}{\frac{1}{2}\|f/f^\star - 1\|_1} \le 2S,$$

where we have used Lemma 5.14. □

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

LEMMA 5.16. *For any $f \in \mathcal{M}$, we have*

$$\left| \frac{\sqrt{f/f^\star} - 1}{h(f, f^\star)} - \frac{f/f^\star - 1}{\sqrt{\chi^2(f||f^\star)}} \right| \leq \{4\|S\|_4^2 S + 2S^2\}\, h(f, f^\star),$$

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

PROOF. Let us define the function R as

$$\sqrt{\frac{f}{f^\star}} - 1 = \frac{1}{2}\left\{ \frac{f - f^\star}{f^\star} + R \right\}.$$

Then we have

$$\frac{\sqrt{f/f^\star} - 1}{h(f, f^\star)} - \frac{f/f^\star - 1}{\sqrt{\chi^2(f||f^\star)}} = \frac{f/f^\star - 1 + R}{\|f/f^\star - 1 + R\|_2} - \frac{f/f^\star - 1}{\|f/f^\star - 1\|_2} = \frac{(f/f^\star - 1 + R)\{\|f/f^\star - 1\|_2 - \|f/f^\star - 1 + R\|_2\} + R\|f/f^\star - 1 + R\|_2}{\|f/f^\star - 1 + R\|_2 \, \|f/f^\star - 1\|_2},$$

so that by the reverse triangle inequality and Corollary 5.15

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

Now note that for all $x \geq -1$

$$-\frac{x^2}{2} \leq -\frac{(\sqrt{1+x} - 1)^2}{2} = \sqrt{1+x} - 1 - \frac{x}{2} \leq 0.$$

Therefore, by Lemma 5.14,

$$|R| \leq \left(\frac{f - f^\star}{f^\star} \right)^2 \leq S^2 \left\| \frac{f - f^\star}{f^\star} \right\|_1^2 \leq S^2 \left\| \frac{f - f^\star}{f^\star} \right\|_1 \left\| \frac{f - f^\star}{f^\star} \right\|_2.$$

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

Finally, we need one further approximation step.

LEMMA 5.17. *Let $q \in \mathbb{N}$ and $\alpha > 0$. Then for every $f \in \mathcal{M}_q$ such that $h(f, f^\star) \leq \alpha$, it is possible to choose coefficients $\eta_i \in \mathbb{R}$, $\beta_i \in \mathbb{R}^d$, $\rho_i \in M_+^d$ for*

$i = 1, \ldots, q^\star$, and $\gamma_i \geq 0$, $\theta_i \in \Theta$ for $i = 1, \ldots, q$, such that $\sum_{i=1}^{q^\star} \mathrm{rank}[\rho_i] \leq q \wedge dq^\star$,

$$\sum_{i=1}^{q^\star} |\eta_i| \leq \frac{1}{c^\star} + \frac{1}{\sqrt{c^\star \alpha}}, \qquad \sum_{i=1}^{q^\star} \|\beta_i\| \leq \frac{1}{c^\star} + \frac{2T}{\sqrt{c^\star \alpha}},$$
$$\sum_{i=1}^{q^\star} \mathrm{Tr}[\rho_i] \leq \frac{1}{c^\star}, \qquad \sum_{j=1}^{q} |\gamma_j| \leq \frac{1}{\sqrt{c^\star \alpha} \wedge c^\star},$$

and

$$\left| \frac{f/f^\star - 1}{\sqrt{\chi^2(f\|f^\star)}} - \ell \right| \leq \frac{d^{3/2}\sqrt{2}}{3(c^\star)^{5/4}} \left\{ \|H_3\|_2 \, S + H_3 \right\} \alpha^{1/4},$$

where we have defined

$$\ell = \sum_{i=1}^{q^\star} \left\{ \eta_i \frac{f_{\theta_i^\star}}{f^\star} + \beta_i^* \frac{D_1 f_{\theta_i^\star}}{f^\star} + \mathrm{Tr}\left[\rho_i \frac{D_2 f_{\theta_i^\star}}{f^\star} \right] \right\} + \sum_{j=1}^{q} \gamma_j \frac{f_{\theta_j}}{f^\star}.$$

PROOF. As $f \in \mathcal{M}_q$, we can write $f = \sum_{j=1}^q \pi_j f_{\theta_j}$. Note that by Theorem 5.11

$$h(f, f^\star) \geq \frac{c^\star}{4} \sum_{i=1}^{q^\star} \sum_{j : \theta_j \in A_i} \pi_j \|\theta_j - \theta_i^\star\|^2.$$

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

$$J = \bigcup_{i=1,\ldots,q^\star} \left\{ j : \theta_j \in A_i, \ \|\theta_j - \theta_i^\star\|^2 \leq 2\sqrt{\alpha/c^\star} \right\}.$$

Taylor expansion gives

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

where $|R_{ji}| \leq \frac{1}{6} d^{3/2} \|\theta_j - \theta_i^\star\|^3 H_3$. We can therefore write

$$\frac{f - f^\star}{f^\star} = L + \sum_{i=1}^{q^\star} \sum_{j \in J : \theta_j \in A_i} \pi_j R_{ji},$$

where we have defined

$$L = \sum_{i=1}^{q^\star} \Bigg\{ \Bigg(\sum_{j \in J : \theta_j \in A_i} \pi_j - \pi_i^\star \Bigg) \frac{f_{\theta_i^\star}}{f^\star} + \sum_{j \in J : \theta_j \in A_i} \pi_j (\theta_j - \theta_i^\star)^* \frac{D_1 f_{\theta_i^\star}}{f^\star} + \frac{1}{2} \sum_{j \in J : \theta_j \in A_i} \pi_j (\theta_j - \theta_i^\star)^* \frac{D_2 f_{\theta_i^\star}}{f^\star} (\theta_j - \theta_i^\star) \Bigg\} + \sum_{j \notin J} \pi_j \frac{f_{\theta_j}}{f^\star}.$$

Now note that

$$\begin{aligned} \left| \frac{f/f^\star - 1}{\sqrt{\chi^2(f||f^\star)}} - \frac{L}{\|L\|_2} \right| &\leq \frac{|f/f^\star - 1|}{\|f/f^\star - 1\|_2} \frac{\|f/f^\star - 1 - L\|_2}{\|L\|_2} + \frac{|f/f^\star - 1 - L|}{\|L\|_2} \\ &\leq \frac{\|f/f^\star - 1 - L\|_2 \, S + |f/f^\star - 1 - L|}{\|L\|_2}, \end{aligned}$$

where we have used Lemma 5.14. By Theorem 5.11, we obtain

$$\|L\|_2 \geq \|L\|_1 \geq \frac{c^\star}{2} \sum_{i=1}^{q^\star} \sum_{j \in J : \theta_j \in A_i} \pi_j \|\theta_j - \theta_i^\star\|^2.$$

Therefore, we can estimate

$$\frac{|f/f^\star - 1 - L|}{\|L\|_2} \leq \frac{d^{3/2} H_3}{3c^\star} \frac{\sum_{i=1}^{q^\star} \sum_{j \in J : \theta_j \in A_i} \pi_j \|\theta_j - \theta_i^\star\|^3}{\sum_{i=1}^{q^\star} \sum_{j \in J : \theta_j \in A_i} \pi_j \|\theta_j - \theta_i^\star\|^2} \leq \left(\frac{4\alpha}{c^\star} \right)^{1/4} \frac{d^{3/2} H_3}{3c^\star}$$

where we have used the definition of J. Setting $\ell = L/\|L\|_2$, we obtain

$$\left| \frac{f/f^\star - 1}{\sqrt{\chi^2(f||f^\star)}} - \ell \right| \leq \frac{d^{3/2}\sqrt{2}}{3(c^\star)^{5/4}} \left\{ \|H_3\|_2 \, S + H_3 \right\} \alpha^{1/4}.$$

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

$$\eta_i = \frac{1}{\|L\|_2} \Bigg(\sum_{j \in J : \theta_j \in A_i} \pi_j - \pi_i^\star \Bigg), \qquad \beta_i = \frac{1}{\|L\|_2} \sum_{j \in J : \theta_j \in A_i} \pi_j (\theta_j - \theta_i^\star),$$

$$\rho_i = \frac{1}{2\|L\|_2} \sum_{j \in J : \theta_j \in A_i} \pi_j (\theta_j - \theta_i^\star)(\theta_j - \theta_i^\star)^*, \qquad \gamma_j = \frac{\pi_j \mathbf{1}_{j \notin J}}{\|L\|_2}.$$

Clearly $\mathrm{rank}[\rho_i] \le \#\{j : \theta_j \in A_i\} \wedge d$, so $\sum_{i=1}^{q^\star} \mathrm{rank}[\rho_i] \le q \wedge dq^\star$. Moreover,

$$\|L\|_2 \ge c^\star \Bigg[\sum_{j:\theta_j \in A_0} \pi_j + \sum_{i=1}^{q^\star} \Bigg\{ \Bigg| \sum_{j:\theta_j \in A_i} \pi_j - \pi_i^\star \Bigg| + \Bigg\| \sum_{j:\theta_j \in A_i} \pi_j (\theta_j - \theta_i^\star) \Bigg\| + \frac{1}{2} \sum_{j:\theta_j \in A_i} \pi_j \|\theta_j - \theta_i^\star\|^2 \Bigg\} \Bigg]$$

by Theorem 5.11. It follows that $\sum_{i=1}^{q^\star} \mathrm{Tr}[\rho_i] \le 1/c^\star$. Now note that for $j \notin J$ such that $\theta_j \in A_i$, we have $\|\theta_j - \theta_i^\star\|^2 > 2\sqrt{\alpha/c^\star}$ by construction. Therefore

$$\|L\|_2 \ge c^\star \Bigg[\sum_{j \notin J : \theta_j \in A_0} \pi_j + \frac{1}{2} \sum_{i=1}^{q^\star} \sum_{j \notin J : \theta_j \in A_i} \pi_j \|\theta_j - \theta_i^\star\|^2 \Bigg] \ge (\sqrt{c^\star \alpha} \wedge c^\star) \sum_{j \notin J} \pi_j.$$

It follows that $\sum_{j=1}^{q} |\gamma_j| \le 1/(\sqrt{c^\star \alpha} \wedge c^\star)$. Next, we note that

$$\sum_{i=1}^{q^\star} \Bigg| \sum_{j \in J : \theta_j \in A_i} \pi_j - \pi_i^\star \Bigg| \le \sum_{i=1}^{q^\star} \Bigg| \sum_{j : \theta_j \in A_i} \pi_j - \pi_i^\star \Bigg| + \sum_{j \notin J : \theta_j \notin A_0} \pi_j.$$

Therefore $\sum_{i=1}^{q^\star} |\eta_i| \le 1/c^\star + 1/\sqrt{c^\star \alpha}$. Finally, note that

$$\sum_{i=1}^{q^\star} \Bigg\| \sum_{j \in J : \theta_j \in A_i} \pi_j (\theta_j - \theta_i^\star) \Bigg\| \le \sum_{i=1}^{q^\star} \Bigg\| \sum_{j : \theta_j \in A_i} \pi_j (\theta_j - \theta_i^\star) \Bigg\| + 2T \sum_{j \notin J : \theta_j \notin A_0} \pi_j.$$

Therefore $\sum_{i=1}^{q^\star} \|\beta_i\| \le 1/c^\star + 2T/\sqrt{c^\star \alpha}$. The proof is complete. □

We can now complete the proof of Theorem 3.3.

PROOF OF THEOREM 3.3. Let $\alpha > 0$ be a constant to be chosen later on, and

$$\mathcal{D}_{q,\alpha} = \{d_f : f \in \mathcal{M}_q,\ f \ne f^\star,\ h(f, f^\star) \le \alpha\}.$$

Then clearly

$$\mathcal{N}(\mathcal{D}_q, \delta) \le \mathcal{N}(\mathcal{D}_{q,\alpha}, \delta) + \mathcal{N}(\mathcal{D}_q \backslash \mathcal{D}_{q,\alpha}, \delta).$$

We will estimate each term separately.

Step 1 (*the first term*). Define

$$\mathbb{M}_q = \{(m_1, \ldots, m_{q^\star}) \in \mathbb{Z}_+^{q^\star} : m_1 + \cdots + m_{q^\star} = q \wedge dq^\star\}.$$

For every $m \in \mathbb{M}_q$, we define the family of functions

$$\mathcal{L}_{q,m,\alpha} = \left\{ \sum_{i=1}^{q^\star} \left\{ \eta_i \frac{f_{\theta_i^\star}}{f^\star} + \beta_i^* \frac{D_1 f_{\theta_i^\star}}{f^\star} + \sum_{j=1}^{m_i} \rho_{ij}^* \frac{D_2 f_{\theta_i^\star}}{f^\star} \rho_{ij} \right\} + \sum_{j=1}^{q} \gamma_j \frac{f_{\theta_j}}{f^\star} : \right.$$
$$\left. (\eta, \beta, \rho, \gamma, \theta) \in \mathfrak{I}_{q,m,\alpha} \right\},$$

where

$$\mathfrak{I}_{q,m,\alpha} = \left\{ (\eta, \beta, \rho, \gamma, \theta) \in \mathbb{R}^{q^\star} \times (\mathbb{R}^d)^{q^\star} \times (\mathbb{R}^d)^{m_1} \times \cdots \times (\mathbb{R}^d)^{m_{q^\star}} \times \mathbb{R}^q \times \Theta^q : \right.$$
$$\sum_{i=1}^{q^\star} |\eta_i| \leq \frac{1}{c^\star} + \frac{1}{\sqrt{c^\star \alpha}}, \qquad \sum_{i=1}^{q^\star} \|\beta_i\| \leq \frac{1}{c^\star} + \frac{2T}{\sqrt{c^\star \alpha}},$$
$$\left. \sum_{i=1}^{q^\star} \sum_{j=1}^{m_i} \|\rho_{ij}\|^2 \leq \frac{1}{c^\star}, \qquad \sum_{j=1}^{q} |\gamma_j| \leq \frac{1}{\sqrt{c^\star \alpha} \wedge c^\star} \right\}.$$

Define the family of functions

$$\mathcal{L}_{q,\alpha} = \bigcup_{m \in \mathbb{M}_q} \mathcal{L}_{q,m,\alpha}$$

From Lemmas 5.16 and 5.17, 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 $h(f, f^\star) \leq \sqrt{2}$ for any f)

$$|d - \ell| \leq \{4\|S\|_4^2 S + 2S^2\} (\alpha \wedge \sqrt{2}) + \frac{d^{3/2}\sqrt{2}}{3(c^\star)^{5/4}} \{\|H_3\|_2 S + H_3\} \alpha^{1/4}.$$

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

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

where $U \in L^2(f^\star d\mu)$ 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

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

Of course, we will ultimately choose ε, α 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 ℓ is defined by the parameters $(\eta,\beta,\rho,\gamma,\theta)\in\mathfrak{I}_{q,m,\alpha}$ and ℓ' is defined by the parameters $(\eta',\beta',\rho',\gamma',\theta')\in\mathfrak{I}_{q,m,\alpha}$. Note that

$$\sum_{i=1}^{q^\star}\sum_{j=1}^{m_i}\left|\rho_{ij}^*\frac{D_2f_{\theta_i^\star}}{f^\star}\rho_{ij}-(\rho'_{ij})^*\frac{D_2f_{\theta_i^\star}}{f^\star}\rho'_{ij}\right|\le\frac{2d}{\sqrt{c^\star}}\,H_2\sum_{i=1}^{q^\star}\sum_{j=1}^{m_i}\|\rho_{ij}-\rho'_{ij}\|.$$

We can therefore estimate

$$|\ell-\ell'|\le H_0\sum_{i=1}^{q^\star}|\eta_i-\eta'_i|+H_1\sqrt{d}\sum_{i=1}^{q^\star}\|\beta_i-\beta'_i\|+H_0\sum_{j=1}^{q}|\gamma_j-\gamma'_j|+$$
$$\frac{\sqrt{d}}{\sqrt{c^\star\alpha}\wedge c^\star}\,H_1\max_{j=1,\ldots,q}\|\theta_j-\theta'_j\|+\frac{2d\sqrt{dq^\star}}{\sqrt{c^\star}}\,H_2\left[\sum_{i=1}^{q^\star}\sum_{j=1}^{m_i}\|\rho_{ij}-\rho'_{ij}\|^2\right]^{1/2}.$$

where we have used that $|f_\theta-f_{\theta'}|/f^\star\le\|\theta-\theta'\|\,H_1\sqrt{d}$ by Taylor expansion. Therefore, writing $V=(H_0+H_1+H_2)\,d\sqrt{dq^\star}$, we have

$$|\ell-\ell'|\le V\,|\!|\!|(\eta,\beta,\rho,\gamma,\theta)-(\eta',\beta',\rho',\gamma',\theta')|\!|\!|_{q,m,\alpha},$$

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

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

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

$$\mathcal{N}(\mathcal{L}_{q,m,\alpha},\varepsilon)\le N(\mathfrak{I}_{q,m,\alpha},|\!|\!|\cdot|\!|\!|_{q,m,\alpha},\varepsilon/2\|V\|_2)\quad\text{for }\varepsilon>0.$$

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

$$\mathcal{N}(\mathcal{L}_{q,m,\alpha},\varepsilon)\le\left(\frac{4\|V\|_2(6+3T)/(\sqrt{c^\star\alpha}\wedge c^\star)+\varepsilon}{\varepsilon}\right)^{(1+d)q^\star+d(q\wedge dq^\star)+(1+d)q}.$$

In particular, if $\varepsilon \le 1$ and $\alpha \le c^\star$, then

$$\mathcal{N}(\mathcal{L}_{q,m,\alpha}, \varepsilon) \le \left(\frac{(24+12T)\|V\|_2/\sqrt{c^\star} + \sqrt{c^\star}}{\varepsilon\sqrt{\alpha}} \right)^{3(d+1)q}.$$

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

$$\#\mathbb{M}_q = \binom{q^\star + q \wedge dq^\star - 1}{q \wedge dq^\star} \le e^{q \wedge dq^\star} \left(\frac{q^\star + q \wedge dq^\star - 1}{q \wedge dq^\star} \right)^{q \wedge dq^\star} \le 2^{3q},$$

where we have used that $q \ge q^\star$. We therefore obtain

$$\begin{aligned} \mathcal{N}(\mathcal{D}_{q,\alpha}, \delta) &\le \sum_{m \in \mathbb{M}_q} \mathcal{N}(\mathcal{L}_{q,m,\alpha}, \delta - 2\alpha^{1/4}\|U\|_2) \\ &\le \left(\frac{24(2+T)\|V\|_2/\sqrt{c^\star} + \sqrt{c^\star}}{(\delta - 2\alpha^{1/4}\|U\|_2)\sqrt{\alpha}} \right)^{3(d+1)q} \end{aligned}$$

whenever $\delta \le 1$ and $\alpha \le (\delta/2\|U\|_2)^4 \wedge c^\star$.

Step 2 (*the second term*). For $f, f' \in \mathcal{M}_q$ with $h(f, f^\star) > \alpha$ and $h(f', f^\star) > \alpha$,

$$\begin{aligned} |d_f - d_{f'}| &= \frac{|(\sqrt{f/f^\star} - 1)\|\sqrt{f'/f^\star} - 1\|_2 - (\sqrt{f'/f^\star} - 1)\|\sqrt{f/f^\star} - 1\|_2|}{h(f, f^\star)\, h(f', f^\star)} \\ &\le \frac{\|\sqrt{f'/f^\star} - \sqrt{f/f^\star}\|_2 |\sqrt{f/f^\star} - 1| + \sqrt{2}\,|\sqrt{f/f^\star} - \sqrt{f'/f^\star}|}{\alpha^2}, \end{aligned}$$

where we have used that $h(f, f^\star) \le \sqrt{2}$ for any f. Now note that

$$\left|\sqrt{a} - \sqrt{b}\right|^2 \le \left|\sqrt{a} - \sqrt{b}\right| \left(\sqrt{a} + \sqrt{b}\right) = |a - b|$$

for any $a, b \ge 0$. We can therefore estimate

$$|d_f - d_{f'}| \le \frac{\|(f - f')/f^\star\|_1^{1/2}(\sqrt{H_0} + 1) + \sqrt{2}\,|(f - f')/f^\star|^{1/2}}{\alpha^2},$$

where we have used that $|\sqrt{f/f^\star} - 1| \le \sqrt{H_0} + 1$ for any $f \in \mathcal{M}$. Now note that if we write $f = \sum_{i=1}^q \pi_i f_{\theta_i}$ and $f' = \sum_{i=1}^q \pi_i' f_{\theta_i'}$, then we can estimate

$$\left| \frac{f - f'}{f^\star} \right| \le H_0 \sum_{i=1}^q |\pi_i - \pi_i'| + H_1\sqrt{d} \max_{i=1,\dots,q} \|\theta_i - \theta_i'\|.$$

Defining

$$W = (\sqrt{H_0} + 1)\|H_0 + H_1\sqrt{d}\|_1^{1/2} + \sqrt{2}\,(H_0 + H_1\sqrt{d})^{1/2},$$

we obtain

$$|d_f - d'_f| \le \frac{W}{\alpha^2}\,|||(\pi,\theta)-(\pi',\theta')|||_q^{1/2}, \qquad |||(\pi,\theta)|||_q = \sum_{i=1}^q |\pi_i| + \max_{i=1,\ldots,q}\|\theta_i\|$$

(clearly $|||\cdot|||_q$ defines a norm on $\mathbb{R}^{(d+1)q}$). Now note that if $|||(\pi,\theta)-(\pi',\theta')|||_q \le \varepsilon$, then we obtain a bracket $d'_f - \varepsilon^{1/2}W/\alpha^2 \le d_f \le d'_f + \varepsilon^{1/2}W/\alpha^2$ of size $\|(d'_f + \varepsilon^{1/2}W/\alpha^2) - (d'_f - \varepsilon^{1/2}W/\alpha^2)\|_2 = 2\varepsilon^{1/2}\|W\|_2/\alpha^2$. Therefore

$$\mathcal{N}(\mathcal{D}_q\backslash\mathcal{D}_{q,\alpha},\delta) \le N(\Delta_q\times\Theta^q, |||\cdot|||_q, \alpha^4\delta^2/4\|W\|_2^2),$$

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

$$\mathcal{N}(\mathcal{D}_q\backslash\mathcal{D}_{q,\alpha},\delta) \le \left(\frac{8(1+T)\|W\|_2^2 + (c^\star)^4}{\alpha^4\delta^2}\right)^{(d+1)q}$$

for $\delta\le 1$ and $\alpha\le c^\star$.

End of proof. Choose $\alpha = (\delta/4\|U\|_2)^4$. Collecting the various estimates above, we find that for $\delta \le 1\wedge 4(c^\star)^{1/4}$ (as $\|U\|_2\ge\|S\|_1\ge 1$ by Lemma 5.14)

$$\begin{aligned}\mathcal{N}(\mathcal{D}_q,\delta) &\le \left(\frac{768(2+T)\|U\|_2^2\|V\|_2/\sqrt{c^\star} + 32\|U\|_2^2\sqrt{c^\star}}{\delta^3}\right)^{3(d+1)q} \\ &\qquad + \left(\frac{4^{18}(1+T)\|U\|_2^{16}\|W\|_2^2 + 4^{16}\|U\|_2^{16}(c^\star)^4}{\delta^{18}}\right)^{(d+1)q} \\ &\le \left(\frac{c_0^\star\,(T\vee 1)^{1/6}\,(\|U\|_2\vee\|V\|_2\vee\|W\|_2)}{\delta}\right)^{18(d+1)q}\end{aligned}$$

where $c_0^\star = 12(c^\star)^{-1/12} + 2(c^\star)^{1/12} + 4(c^\star)^{4/18} + 8$. It follows that

$$\mathcal{N}(\mathcal{D}_q,\delta) \le \left(\frac{C^\star(T\vee 1)^{1/6}(\|H_0\|_4^4\vee\|H_1\|_4^4\vee\|H_2\|_4^4\vee\|H_3\|_2^2)}{\delta}\right)^{18(d+1)q}$$

for all $\delta\le\delta^\star$, where $C^\star$ and $\delta^\star$ are constants that depend only on $c^\star$, d, and $q^\star$. This establishes the estimate given in the statement of the Theorem. The proof of the second half of the Theorem follows from Corollary 5.15 and $\|H_0\|_4\ge 1$. □

5.5. *Proof of Theorem 4.1.* The proof of Theorem 4.1 is based on Theorem 2.3 and the following result.

PROPOSITION 5.18. *Let $\mathcal{M}^n$ for $n \geq 1$ be a family of strictly positive probability densities with respect to a reference measure μ such that $\mathcal{M}^n \subseteq \mathcal{M}^{n+1}$ for all n. Define $\mathcal{M} = \bigcup_n \mathcal{M}^n$, and let $f^\star$ be another probability density with respect to μ such that $f^\star \notin \operatorname{cl}\mathcal{M}$, where $\operatorname{cl}\mathcal{M}$ denotes the $L^1(d\mu)$-closure of $\mathcal{M}$. Let $\mathcal{H}^n = \{\sqrt{f/f^\star} : f \in \mathcal{M}^n\}$, and suppose there exist $K(n) \geq 1$ and $p \geq 1$ so that*

$$\mathcal{N}(\mathcal{H}^n, \delta) \leq \left(\frac{K(n)}{\delta}\right)^p$$

for all $\delta \leq 1$ and $n \geq 1$, where $\mathcal{N}(\mathcal{H}^n, \delta)$ is the minimal number of brackets of $L^2(f^\star d\mu)$-width δ needed to cover $\mathcal{H}^n$. Let $(X_i)_{i\in\mathbb{N}}$ be i.i.d. with distribution $f^\star d\mu$. If in addition $\log K(n) = o(n)$, then we have

$$\limsup_{n\to\infty} \sup_{f\in\mathcal{M}^n} \frac{1}{n}\sum_{j=1}^n \log\left(\frac{f(X_j)}{f^\star(X_j)}\right) < 0 \quad \textit{a.s.}$$

PROOF. As in the proof of Theorem 5.1, we have

$$\frac{1}{n}\sum_{j=1}^n \log\left(\frac{f(X_j)}{f^\star(X_j)}\right) \leq 4n^{-1/2}\nu_n(\log(\{\bar{f}/f^\star\}^{1/2})) - 2D(f^\star||\bar{f}).$$

The following claim will be proved below:

$$\lim_{n\to\infty} \sup_{f\in\mathcal{M}^n} n^{-1/2}\nu_n(\log(\{\bar{f}/f^\star\}^{1/2})) = 0 \quad \text{a.s.}$$

Using the claim, the proof is easily completed: indeed, we then have

$$\limsup_{n\to\infty} \sup_{f\in\mathcal{M}^n} \frac{1}{n}\sum_{j=1}^n \log\left(\frac{f(X_j)}{f^\star(X_j)}\right) \leq -2 \inf_{f\in\mathcal{M}} D(f^\star||\bar{f}) < 0 \quad \text{a.s.},$$

where the last inequality follows from Pinsker's inequality and $f^\star \notin \operatorname{cl}\mathcal{M}$.

It therefore remains to prove the claim. To this end we apply [25], Theorem 5.11 as in the proof of [25], Theorem 7.4 (cf. Theorem 5.1 above), which yields

$$\mathbf{P}\left[\sup_{f\in\mathcal{M}^n} |n^{-1/2}\nu_n(\log(\{\bar{f}/f^\star\}^{1/2}))| \geq \alpha\right] \leq C\, e^{-n\alpha^2/C}$$

for every $\alpha > 0$ such that $C\sqrt{p}\,(1 + \sqrt{\log K(n)}) \leq \alpha\sqrt{n} \leq 32\sqrt{n}$ and $n \geq 1$, where C is a universal constant. As $\log K(n) = o(n)$, we have

$$\sum_{n\geq 1} \mathbf{P}\left[\sup_{f\in\mathcal{M}^n} |n^{-1/2}\nu_n(\log(\{\bar{f}/f^\star\}^{1/2}))| \geq \alpha\right] < \infty$$

for all $0 < \alpha \leq 32$, so the claim follows from the Borel-Cantelli lemma. □

We can now complete the proof of Theorem 4.1.

PROOF OF THEOREM 4.1. By Theorem 2.3 and easy manipulations, $\mathbf{P}^\star$-a.s.

$$\begin{aligned}
&\limsup_{n\to\infty} \sup_{q>q^\star} \frac{1}{\mathrm{pen}(n,q)-\mathrm{pen}(n,q^\star)} \left\{ \sup_{f\in\mathcal{M}_q^n} \ell_n(f) - \sup_{f\in\mathcal{M}_{q^\star}^n} \ell_n(f) \right\} \\
&\le \limsup_{n\to\infty} \sup_{q>q^\star} \frac{\eta(q)\{\log K(2n)\vee\log\log n\}}{\mathrm{pen}(n,q)-\mathrm{pen}(n,q^\star)} \times \\
&\quad \limsup_{n\to\infty} \frac{1}{\log K(2n)\vee\log\log n} \sup_{q>q^\star} \frac{1}{\eta(q)} \left\{ \sup_{f\in\mathcal{M}_q^n} \ell_n(f) - \sup_{f\in\mathcal{M}_{q^\star}^n} \ell_n(f) \right\} = 0.
\end{aligned}$$

Therefore, $\mathbf{P}^\star$-a.s. eventually as $n\to\infty$

$$\sup_{f\in\mathcal{M}_q^n} \ell_n(f) - \mathrm{pen}(n,q) < \sup_{f\in\mathcal{M}_{q^\star}^n} \ell_n(f) - \mathrm{pen}(n,q^\star)$$

for all $q > q^\star$. It follows that $\limsup_{n\to\infty} \hat{q}_n \le q^\star$ $\mathbf{P}^\star$-a.s., that is, the penalized likelihood order estimator does not asymptotically overestimate the order.

On the other hand, we note that for every $q < q^\star$

$$\limsup_{n\to\infty} \frac{1}{n} \left\{ \sup_{f\in\mathcal{M}_q^n} \ell_n(f) - \sup_{f\in\mathcal{M}_{q^\star}^n} \ell_n(f) \right\} \le \limsup_{n\to\infty} \sup_{f\in\mathcal{M}_q^n} \frac{1}{n} \sum_{j=1}^n \log\left(\frac{f(X_j)}{f^\star(X_j)}\right)$$

which is strictly negative $\mathbf{P}^\star$-a.s. by Proposition 5.18, where we have used that $\log K(n) = o(n)$ and that $\mathcal{N}(\mathcal{H}_q^n(2),\delta) \le \mathcal{N}(\mathcal{H}_{q^\star}^n(2),\delta) \le (2K(n)/\delta)^{\eta(q^\star)}$ for all $\delta \le 2$ and n sufficiently large. As $\mathrm{pen}(n,q)/n \to 0$ as $n\to\infty$ for $q < q^\star$

$$\limsup_{n\to\infty} \max_{q<q^\star} \frac{1}{n} \left\{ \sup_{f\in\mathcal{M}_q^n} \ell_n(f) - \mathrm{pen}(n,q) - \sup_{f\in\mathcal{M}_{q^\star}^n} \ell_n(f) + \mathrm{pen}(n,q^\star) \right\} < 0$$

$\mathbf{P}^\star$-a.s. In particular, we find that $\mathbf{P}^\star$-a.s. eventually as $n\to\infty$

$$\sup_{f\in\mathcal{M}_q^n} \ell_n(f) - \mathrm{pen}(n,q) < \sup_{f\in\mathcal{M}_{q^\star}^n} \ell_n(f) - \mathrm{pen}(n,q^\star)$$

for all $q < q^\star$. It follows that $\liminf_{n\to\infty} \hat{q}_n \ge q^\star$ $\mathbf{P}^\star$-a.s., that is, the penalized likelihood order estimator does not asymptotically underestimate the order. □

Finally, let us prove Corollary 4.3.

PROOF OF COROLLARY 4.3. It is shown in the proof of Corollary 2.6 that

$$\Gamma := \sup_{g \in L_0^2(f^\star d\mu)} \left\{ \sup_{d \in \bar{\mathcal{D}}_q^c} (\langle d, g \rangle)_+^2 - \sup_{d \in \bar{\mathcal{D}}_{q^\star}} (\langle d, g \rangle)_+^2 \right\} > 0.$$

By Theorem 2.5, we have

$$\limsup_{n \to \infty} \frac{1}{\mathrm{pen}(n,q) - \mathrm{pen}(n,q^\star)} \left\{ \sup_{f \in \mathcal{M}_q} \ell_n(f) - \sup_{f \in \mathcal{M}_{q^\star}} \ell_n(f) \right\} \geq$$
$$\frac{1}{C\{\eta(q) - \eta(q^\star)\}} \sup_{g \in L_0^2(f^\star d\mu)} \left\{ \sup_{d \in \bar{\mathcal{D}}_q^c} (\langle d, g \rangle)_+^2 - \sup_{d \in \bar{\mathcal{D}}_{q^\star}} (\langle d, g \rangle)_+^2 \right\} \quad \mathbf{P}^\star\text{-a.s.}$$

Therefore, choosing $C < \Gamma / \{\eta(q) - \eta(q^\star)\}$, we find that

$$\sup_{f \in \mathcal{M}_q} \ell_n(f) - \mathrm{pen}(n,q) > \sup_{f \in \mathcal{M}_{q^\star}} \ell_n(f) - \mathrm{pen}(n,q^\star)$$

infinitely often $\mathbf{P}^\star$-a.s., so that $\hat{q}_n \neq q^\star$ infinitely often $\mathbf{P}^\star$-a.s. □

5.6. *Proof of Proposition 4.4.* The proofs of the consistency results in Propositions 4.4 and 4.5 follow almost immediately from Theorem 4.1, Corollary 3.4, and Example 3.5. The main difficulty is to establish the condition $\bar{\mathcal{D}}_q^c \backslash \bar{\mathcal{D}}_{q^\star} \neq \varnothing$ of Corollary 4.3, which is needed to prove the inconsistency part of Proposition 4.4. To this end, we will need the following lemma characterizing $\bar{\mathcal{D}}_{q^\star}$ (here we adopt the same notations as in section 3.2).

LEMMA 5.19. *Suppose that Assumption A holds. Then we have*

$$\bar{\mathcal{D}}_{q^\star} = \left\{ \frac{L}{\|L\|_2} : L = \sum_{i=1}^{q^\star} \left\{ \eta_i \frac{f_{\theta_i^\star}}{f^\star} + \beta_i^* \frac{D_1 f_{\theta_i^\star}}{f^\star} \right\},\ \eta_i \in \mathbb{R},\ \beta_i \in \mathbb{R}^d,\ \sum_{i=1}^{q^\star} \eta_i = 0 \right\}.$$

PROOF. Let $(f_n)_{n \geq 1} \subset \mathcal{M}_{q^\star}$ be such that $h(f_n, f^\star) \to 0$ and $d_{f_n} \to d_0 \in \bar{\mathcal{D}}_{q^\star}$. By Theorem 5.11, we may assume without loss of generality that $f_n = \sum_{i=1}^{q^\star} \pi_i^n f_{\theta_i^n}$ with $\theta_i^n \to \theta_i^\star$ and $\pi_i^n \to \pi_i^\star$ for every $i = 1, \ldots, q^\star$. Taylor expansion gives

$$\frac{f_n - f^\star}{f^\star} = L_n + R_n, \qquad |R_n| \leq \frac{d}{2} H_2 \sum_{i=1}^{q^\star} \pi_i^n \|\theta_i^n - \theta_i^\star\|^2,$$

where

$$L_n = \sum_{i=1}^{q^\star} \left\{ (\pi_i^n - \pi_i^\star) \frac{f_{\theta_i^\star}}{f^\star} + \pi_i^n (\theta_i^n - \theta_i^\star)^* \frac{D_1 f_{\theta_i^\star}}{f^\star} \right\}.$$

Proceeding as in Lemmas 5.16 and 5.17, we can estimate

$$\left\| d_{f_n} - \frac{L_n}{\|L_n\|_2} \right\|_2 \le 2\|S\|_4^2\{2\|S\|_2+1\}\, h(f_n, f^\star) + \{\|S\|_2+1\}\, \frac{\|R_n\|_2}{\|L_n\|_2}.$$

But using Theorem 5.11, we find that for n sufficiently large

$$\|L_n\|_2 \ge \|L_n\|_1 \ge c^\star \sum_{i=1}^{q^\star} \pi_i^n \|\theta_i^n - \theta_i^\star\|.$$

Thus we have

$$\frac{\|R_n\|_2}{\|L_n\|_2} \le \frac{d\|H_2\|_2}{2c^\star} \frac{\sum_{i=1}^{q^\star} \pi_i^n \|\theta_i^n - \theta_i^\star\|^2}{\sum_{i=1}^{q^\star} \pi_i^n \|\theta_i^n - \theta_i^\star\|} \le \frac{d\|H_2\|_2}{2c^\star} \max_{i=1,\ldots,q^\star} \|\theta_i^n - \theta_i^\star\| \xrightarrow{n\to\infty} 0.$$

We have therefore shown that $L_n/\|L_n\|_2 \to d_0$ in $L^2(f^\star d\mu)$. Now define

$$\eta_i^n = \frac{\pi_i^n - \pi_i^\star}{Z_n}, \quad \beta_i^n = \frac{\pi_i^n(\theta_i^n - \theta_i^\star)}{Z_n}, \qquad Z_n = \sum_{i=1}^{q^\star} \{|\pi_i^n - \pi_i^\star| + \|\pi_i^n(\theta_i^n - \theta_i^\star)\|\}.$$

As $\sum_{i=1}^{q^\star}\{|\eta_i^n| + \|\beta_i^n\|\} = 1$ for all n, we may extract a subsequence such that $\eta_i^n \to \eta_i$, $\beta_i^n \to \beta_i$, and $\sum_{i=1}^{q^\star}\{|\eta_i| + \|\beta_i\|\} = 1$. We obtain immediately

$$d_0 = \frac{L}{\|L\|_2}, \qquad L = \sum_{i=1}^{q^\star} \left\{ \eta_i \frac{f_{\theta_i^\star}}{f^\star} + \beta_i^* \frac{D_1 f_{\theta_i^\star}}{f^\star} \right\}.$$

Clearly $\sum_{i=1}^{q^\star} \eta_i = 0$. Thus we have shown that any $d_0 \in \bar{\mathcal{D}}_{q^\star}$ has the desired form.

It remains to show that any function of the desired form is in fact an element of $\bar{\mathcal{D}}_{q^\star}$. To this end, fix $\eta_i \in \mathbb{R}$, $\beta_i \in \mathbb{R}^d$ with $\sum_{i=1}^{q^\star} \eta_i = 0$, and define f_t for $t > 0$ as

$$f_t = \sum_{i=1}^{q^\star} (\pi_i^\star + t\eta_i)\, f_{\theta_i^\star + \beta_i t/\pi_i^\star}.$$

Clearly $f_t \in \mathcal{M}_{q^\star}$ for all t sufficiently small, and $f_t \to f^\star$ as $t \to 0$. But

$$\frac{f_t - f^\star}{t} = \sum_{i=1}^{q^\star} \pi_i^\star \frac{f_{\theta_i^\star + \beta_i t/\pi_i^\star} - f_{\theta_i^\star}}{t} + \sum_{i=1}^{q^\star} \eta_i\, f_{\theta_i^\star + \beta_i t/\pi_i^\star}.$$

Therefore clearly

$$\frac{1}{t}\frac{f_t - f^\star}{f^\star} \xrightarrow{t\to 0} \sum_{i=1}^{q^\star} \left\{ \eta_i \frac{f_{\theta_i^\star}}{f^\star} + \beta_i^* \frac{D_1 f_{\theta_i^\star}}{f^\star} \right\} = L.$$

Using Lemma 5.16, we obtain

$$\lim_{t\to 0} d_{f_t} = \lim_{t\to 0} \frac{(f_t - f^\star)/tf^\star}{\|(f_t - f^\star)/tf^\star\|_2} = \frac{L}{\|L\|_2}.$$

Thus any function of the desired form is in $\bar{\mathcal{D}}_{q^\star}$, and the proof is complete. □

REMARK 5.20. The proof of Lemma 5.19 in fact shows that $\bar{\mathcal{D}}_{q^\star} = \bar{\mathcal{D}}^c_{q^\star}$.

We can now complete the proof of Propositions 4.4 and 4.5.

PROOF OF PROPOSITION 4.4. We begin by proving consistency of the penalty $\mathrm{pen}(n,q) = q\,\omega(n)$. Note that by Corollary 3.4, the assumption of Corollary 4.2 holds with $\eta(q) = 18(d+1)q + 1 \le 19(d+1)q$. Thus consistency of $\mathrm{pen}(n,q) = q\,\omega(n)$ follows from Corollary 4.2 using $\varpi(n) = \omega(n)/19(d+1)$.

To prove inconsistency of the penalty $\mathrm{pen}(n,q) = C\,q\,\log\log n$ with $C > 0$ sufficiently small, it suffices to show that $\bar{\mathcal{D}}^c_{q^\star+1} \backslash \bar{\mathcal{D}}_{q^\star}$ is nonempty. Indeed, if this is the case then we can apply Corollary 4.3 with $q = q^\star + 1$, where the requisite entropy assumption follows immediately from Theorem 4.1.

Fix $v \in \mathbb{R}^d$, and consider the function f_t defined for $t > 0$ as follows:

$$f_t = \frac{\pi^\star_1}{2}\,(f_{\theta^\star_1 + vt} + f_{\theta^\star_1 - vt}) + \sum_{i=2}^{q^\star} \pi^\star_i f_{\theta^\star_i}.$$

Clearly $f_t \in \mathcal{M}_{q^\star+1}$ for all t sufficiently small, $f_t \to f^\star$ as $t \to 0$, and

$$\frac{f_t - f^\star}{t^2} = \frac{\pi^\star_1}{2}\,\frac{f_{\theta^\star_1+vt} - 2\,f_{\theta^\star_1} + f_{\theta^\star_1 - vt}}{t^2} \xrightarrow{t\to 0} \frac{\pi^\star_1}{2}\,v^* D_2 f_{\theta^\star_1} v.$$

As in the proof of Lemma 5.19, we find that

$$\lim_{t\to 0} d_{f_t} = \lim_{t\to 0} \frac{(f_t - f^\star)/t^2 f^\star}{\|(f_t - f^\star)/t^2 f^\star\|_2} = \frac{v^* D_2 f_{\theta^\star_1} v}{\|v^* D_2 f_{\theta^\star_1} v\|_2} = d_0.$$

By construction, $d_0 \in \bar{\mathcal{D}}^c_{q^\star+1}$. But by Theorem 5.11, the functions $f_{\theta^\star_i}$, $D_1 f_{\theta^\star_i}$, and $v^* D_2 f_{\theta^\star_i} v$ ($i = 1, \ldots, q^\star$) are all linearly independent. Together with Lemma 5.19, this shows that $d_0 \notin \bar{\mathcal{D}}_{q^\star}$. Thus $d_0 \in \bar{\mathcal{D}}^c_{q^\star+1} \backslash \bar{\mathcal{D}}_{q^\star}$, and the proof is complete. □

PROOF OF PROPOSITION 4.5. By Example 3.5, the assumption of Theorem 4.1 holds with $\eta(q) = 18(d+1)q + 1$ and $\log K(n) = \log C^\star_1 + C^\star_2 T(n)^2$. The desired consistency results now follow immediately from Theorem 4.1. □

Acknowledgments. The authors would like to thank Jean Bretagnolle for providing an enlightening counterexample which guided some of our proofs. We also thank Michel Ledoux for suggesting some helpful references.

REFERENCES

[1] AZAIS, J.-M., GASSIAT, E., AND MERCADIER, C. (2009). The likelihood ratio test for general mixture models with possibly structural parameter. *ESAIM P and S 3*, 301–327.

[2] BARRON, A., RISSANEN, J., AND YU, B. (1998). The minimum description length principle in coding and modeling. *IEEE Trans. Inform. Theory 44*, 2743–2760.

[3] BICKEL, P. AND CHERNOFF, H. (1993). Asymptotic distribution of the likelihood ratio statistic in a prototypical non regular problem. In *Statistics and Probability: a Raghu Raj Bahadur festschrift*. Wiley Eastern Ltd., New Delhi, 83–96.

[4] BILLINGSLEY, P. (1999). *Convergence of probability measures*, Second ed. John Wiley & Sons Inc., New York.

[5] CAPPÉ, O., MOULINES, E., AND RYDÉN, T. (2005). *Inference in hidden Markov models*. Springer Series in Statistics. Springer, New York. With Randal Douc's contributions to Chapter 9 and Christian P. Robert's to Chapters 6, 7 and 13, With Chapter 14 by Gersende Fort, Philippe Soulier and Moulines, and Chapter 15 by Stéphane Boucheron and Elisabeth Gassiat.

[6] CHAMBAZ, A. (2006). Testing the order of a model. *Ann. Statist. 34*, 1166–1203.

[7] CHAMBAZ, A., GARIVIER, A., AND GASSIAT, E. (2009). A MDL approach to HMM with Poisson and Gaussian emissions. Application to order indentification. *Journal of Stat. Planning and Inf. 139*, 962–977.

[8] CSISZAR, I. (2002). Large-scale typicality of Markov sample paths and consistency of MDL order estimators. *IEEE Trans. Info. Theory 48*, 1616–1628. Special issue on Shannon theory: perspective, trends, and applications.

[9] CSISZAR, I. AND SHIELDS, P. C. (2000). The consistency of BIC Markov order estimator. *Annals of Stat. 28*, 1601–1619.

[10] FINESSO, L. (1990). Consistent estimation of the order for Markov and hidden Markov chains. Ph.D. Thesis, Univ. of Maryland.

[11] GASSIAT, E. (2002). Likelihood ratio inequalities with applications to various mixtures. *Ann. Inst. H. Poincaré Probab. Statist. 38*, 897–906.

[12] GASSIAT, E. AND BOUCHERON, S. (2003). Optimal error exponents in hidden Markov model order estimation. *IEEE Trans. Info. Theory 48*, 964–980.

[13] GENOVESE, C. R. AND WASSERMAN, L. (2000). Rates of convergence for the Gaussian mixture sieve. *Ann. Statist. 28*, 1105–1127.

[14] HANNAN, E. J. AND QUINN, B. G. (1979). The determination of the order of an autoregression. *J. Roy. Statist. Soc. Ser. B 41*, 190–195.

[15] HARTIGAN, J. A. (1985). A failure of likelihood asymptotics for normal mixtures. In *Proceedings of the Berkeley conference in honor of Jerzy Neyman and Jack Kiefer, Vol. II (Berkeley, Calif., 1983)*. Wadsworth Statist./Probab. Ser. Wadsworth, Belmont, CA, 807–810. MR822066 (87i:62046)

[16] KERIBIN, C. (2000). Consistent estimation of the order of mixture models. *Sankhya Ser. A 62*, 49–66.

[17] LEDOUX, M. AND TALAGRAND, M. (1989). Comparison theorems, random geometry and some limit theorems for empirical processes. *Ann. Probab. 17*, 596–631.

[18] LIU, X. AND SHAO, Y. (2003). Asymptotics for likelihood ratio tests under loss of identifiability. *Ann. Statist. 31*, 807–832.

[19] LIU, X. AND SHAO, Y. (2004). Asymptotics for the likelihood ratio test in a two-component normal mixture model. *J. Statist. Plann. Inference* **123**, 1, 61–81. MR2058122 (2005c:62036)

[20] LUKACS, E. (1970). *Characteristic functions*, second ed. Griffin, London.

[21] MAUGIS, C. AND MICHEL, B. (2011). A non asymptotic penalized criterion for gaussian mixture model selection. *ESAIM: Probability and Statistics*. to appear.

[22] NISHII, R. (1988). Maximum likelihood principle and model selection when the true model is unspecified. *J. Multivariate Anal. 27*, 392–403.

[23] OSSIANDER, M. (1987). A central limit theorem under metric entropy with L_2 bracketing. *Ann. Probab. 15*, 897–919.

[24] RISSANEN, J. (1986). Stochastic complexity and modeling. *Ann. Statist. 14*, 1080–1100.

[25] VAN DE GEER, S. A. (2000). *Applications of empirical process theory*. Cambridge Series in Statistical and Probabilistic Mathematics, Vol. **6**. Cambridge University Press, Cambridge.

[26] VAN DER VAART, A. W. (1998). *Asymptotic statistics*. Cambridge Series in Statistical and Probabilistic Mathematics, Vol. **3**. Cambridge University Press, Cambridge.

[27] VAN HANDEL, R. (2011). On the minimal penalty for Markov order estimation. *Probab. Th. Rel. Fields*. to appear.

[28] WILLIAMS, D. (1991). *Probability with martingales*. Cambridge Mathematical Textbooks. Cambridge University Press, Cambridge.

LABORATOIRE DE MATHÉMATIQUES,
UNIVERSITÉ PARIS-SUD,
BÂTIMENT 425,
91405 ORSAY CEDEX, FRANCE.
E-MAIL: elisabeth.gassiat@math.u-psud.fr

SHERRERD HALL, ROOM 227,
PRINCETON UNIVERSITY,
PRINCETON, NJ 08544, USA.
E-MAIL: rvan@princeton.edu