# Consistent order estimation and the local geometry of mixtures

Elisabeth Gassiat, Ramon van Handel

# CONSISTENT ORDER ESTIMATION AND THE LOCAL GEOMETRY OF MIXTURES

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 $(\mathcal{M}_q)_{q\in\mathbb{N}}$, $\mathcal{M}_q \subset \mathcal{M}_{q+1}$. The smallest index $q^\star$ such that $\mathcal{M}_{q^\star}$ contains $f^\star$ is called the model order. We establish strong consistency of the penalized likelihood order estimator in a general setting with penalties of order $\eta(q)\log\log n$, where $\eta(q)$ is a dimensional quantity. Moreover, such penalties are shown to be minimal. In contrast to previous work, an a priori upper bound on the model order is not assumed.

The local dimension $\eta(q)$ of the model $\mathcal{M}_q$ is defined in terms of the bracketing entropy of a class of weighted densities, whose computation is a nonstandard problem which is of independent interest. We perform the requisite computations for the case of one-dimensional location mixtures, thus demonstrating the consistency of the penalized likelihood mixture order estimator. The proof requires a delicate analysis of the local geometry of the mixture family $\mathcal{M}_q$ in a neighborhood of $f^\star$, for $q > q^\star$. The extension to more general mixture models remains an open problem.

**1. Introduction.** Let $(X_k)_{k\in\mathbb{N}}$ be a sequence of random variables whose distribution $f^\star$ lies in one of a nested family of models $(\mathcal{M}_q)_{q\in\mathbb{N}}$, indexed (and ordered) by the integers. We define the model order as the smallest index $q^\star$ such that the true distribution of the model lies in the corresponding model class. Model order estimation from observed data is a statistical problem of significant practical interest. On the one hand, 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, in many cases the model order has a concrete interpretation in terms of the modelling of the underlying phenomenon (for example, the estimation of the number of distinct clusters in a data set, or the estimation of the number of regimes in an economic time series). For these reasons, order estimation problems appear in a wide variety of applications. Typical examples of order estimation problems include Markov order estimation, hidden Markov model order estimation, and mixture model order estimation. From both the theoretical and practical prespective, a challenging problem is to develop strongly consistent or-

*AMS 2000 subject classifications:* 62G20, 60F15, 60F10, 41A46, 41A25

*Keywords and phrases:* consistent order estimation, penalized likelihood, likelihood inequalities, uniform law of iterated logarithm, empirical process theory, bracketing entropy, location mixtures

der estimators which can be applied in a general setting and which do not suffer from restrictive assumptions which are rarely satisfied in applications, such as the availability of an a priori upper bound on the model order.

In this paper, we focus on independent and identically distributed sequences $(X_k)_{k\in\mathbb{N}}$ and on penalized likelihood order estimators of the form

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

where $\ell_n(f)$ is the log-likelihood of the sequence $(X_k)_{k\le n}$ with distribution $f$ and $\mathrm{pen}(n,q)$ is a given penalty function. In this setting, we aim to determine which penalties $\mathrm{pen}(n,q)$ give rise to strongly consistent order estimators (that is, such that $\hat{q}_n \to q^\star$ a.s. as $n\to\infty$). The investigation of strong consistency requires a detailed understanding of the fluctuations of the likelihood ratio statistic. The guiding motivation for this paper was to obtain an understanding of mixture order estimation problems, where the behavior of the likelihood ratio statistic is notoriously complicated due to a fundamental lack of identifiability. However, the main results of this paper establish consistency and inconsistency results for model order estimation problems in a very general setting, going far beyond the problem of mixture order estimation. In addition to these general results, we will obtain specific results for mixture models which require a rather delicate analysis of their geometric structure. The latter sheds light also on other statistical problems in mixture models (such as hypothesis testing) and is of independent interest.

1.1. *Previous work.* There are two main approaches towards studying strong consistency of the penalized likelihood model order estimator.

The first approach (which forms the foundation also for this paper) stems from the observation that the likelihood ratio statistic can be approximated by the square of an empirical process. In regular parametric models, this follows by a simple Taylor expansion argument, similar to the one used in the chi-square theory of likelihood ratio tests. The situation is more delicate in non-identifiable models, but such a correspondence was nonetheless obtained in a very general setting by one of us [10] (see also [17] for related results). Once this equivalence is established, the law of iterated logarithm implies directly that the likelihood ratio statistic has pathwise fluctuations of order $\log\log n$, thus giving rise to strongly consistent penalties of order $\mathrm{pen}(n,q) \sim \log\log n$. This approach has been employed in a variety of order estimation problems: for ARMA models in [14], for regular parametric models in [19], for Markov order estimation in [9], and for mixture order estimation in [15, 4]. However, the drawback of this approach is that the law of iterated logarithm only applies to the likelihood ratio statistic for a single model class, so that one has no control over the fluctuations of the likelihood ratio statistic uniformly

in the model order. For this reason, the results in the above references must assume that one has prior knowledge that the true model order is upper bounded by some known constant. As is pointed out in [7], this restriction is unsatisfactory as such an upper bound is rarely available in practice.

The second approach is entirely different in nature and is based on the approximation of penalized likelihood order estimators by minimum description length (MDL) order estimators, which can be studied using techniques from information theory (see [3], Chapter 15 for a primer). This approach was employed for the hidden Markov model order estimation problem in [9, 11, 5], and for Gaussian or Poisson mixture order estimation in [5]. In contrast to the first approach, the information theoretic approach does not require an a priori upper bound on the model order. On the other hand, the strongly consistent penalties obtained through this approach are typically of order $\mathrm{pen}(n,q) \sim \log n$ and grow rather rapidly in the model order $q$. Therefore, such penalties are substantially larger than those obtained through the first approach, and are therefore expected to be suboptimal in most cases (it should be noted that small penalties are highly desirable in practice, as they minimize the probability of underestimating the order). In addition, the computations involved in the information-theoretic approach are specific to particular families of densities (such as discrete distributions [9, 11] or Gaussian mixtures and Poisson mixtures [5]) and do not appear to admit a general consistency theorem that applies simultaneously to a large class of order estimation problems.

The inadequacies of these approaches was highlighted in the work of Csiszar and Shields [7, 6], who present a detailed study of the Markov order estimation problem. They establish consistency of the BIC Markov order estimator $\mathrm{pen}(n,q) = \frac{1}{2}\dim(q)\log n$, where $\dim(q)$ is the dimension of the parameter space of the model of order $q$, without a prior upper bound on the order. The analysis in these papers is very delicate, however, and relies heavily on the availability of an explicit expression for the maximum likelihood estimator for Markov chains. Such explicit expressions are rarely available in more general order estimation problems.

Recently, one of us has shown [24] that penalties of order $\log\log n$ already lead to strongly consistent estimators for Markov order estimation, even in the absence of a prior upper bound. This refinement of the results of [7, 6] requires an entirely different method of proof: the key idea is to use martingale concentration inequalities and techniques from empirical process theory to obtain a law of iterated logarithm for the likelihood ratio statistic which holds *uniformly* in the model order. In particular, this approach does not rely on an explicit expression for the maximum likelihood estimator, and is therefore much more generally applicable.

1.2. *Contributions of this paper.* The goal of this paper is to investigate which penalties give rise to strongly consistent model order estimators in the i.i.d. setting.

Our main contributions are threefold.

First, Theorem 2.4 establishes in a very general setting that the penalized likelihood model order estimator with penalty of order $\eta(q)\log\log n$ is strongly consistent in the absence of a prior upper bound. Here $\eta(q)$ is a dimensional quantity related to the bracketing entropy of a certain weighted class of densities $\mathcal{D}_q$ derived from the model class $\mathcal{M}_q$. The proof of this result is inspired by the method developed in [24] for Markov order estimation, though we follow a somewhat different approach here to obtain a much more widely applicable result.

Second, Theorem 2.10 shows that penalties of the form $C\,\eta(q)\log\log n$ give rise to inconsistent order estimators when the constant $C$ is chosen sufficiently small. This implies that penalties of order $\log\log n$ are in fact minimal when the aim is to achieve strong consistency. Thus our results essentially characterize those strongly consistent penalties which minimize the probability of underestimating the order. The main ingredient of the proof of is an exact characterization of the fluctuations of the generalized likelihood ratio test, which may be of independent interest.

Finally, we aim to apply our general results to the mixture order estimation problem. The key difficulty here is to compute the bracketing entropy of the weighted model classes $\mathcal{D}_q$. To our knowledge, this has hitherto remained an open problem: despite that one may find various claims [15, 4, 1] that the bracketing numbers are polynomial, no proof supports these claims. Moreover, as we aim to establish results that hold uniformly in the model order, it is of key importance that the constants that appear in estimated on the bracketing entropies are independent of the model order. The computation of entropies of weighted model classes appears to be a nonstandard problem in empirical process theory,[1] and the lack of identifiability in mixture models makes this a novel and rather delicate problem.

In Theorem 2.12, we establish explicit (polynomial) bounds on the bracketing entropies of the weighted classes $\mathcal{D}_q$ in the case of one-dimensional location mixtures. The requisite assumptions are mild and easily verified: we require only some smoothness assumpions and the existence of exponential moments. This result is of independent interest, and could have a variety of applications to other statistical problems involving mixtures. The proof of Theorem 2.12 requires a delicate analysis of the local geometry of mixture models in the neighborhood of the true distribution. We believe that essentially the same results should hold in much more

[1] The standard approach for dealing with weighted empirical processes is to employ a so-called peeling device (see [23]) to reduce the problem to the computation of entropies of unweighted model classes. Unfortunately, in parametric models, this gives rise to additional terms of order $\log n$ unless on can compute the *local* entropies of the model classes. In nonidentifiable models such as mixtures, whose geometry is notoriously delicate, such local entropy computations do not appear to be feasible. Therefore, the direct computation of the entropies of weighted model classes becomes essential in such models. Note that any additional $\log n$ factors would dominate entirely the correct $\log\log n$ growth rate of the likelihood, and would therefore lead to far from optimal results.

general mixture models, but we were not able to complete one key part of the proof. This remains a challenging open problem.

1.3. *Organization of this paper.* The remainder of this paper is organized as follows. Section 2 describes our main results: the strong consistency theorem (section 2.1), the inconsistency theorem (section 2.2), and the application to mixture order estimation (section 2.3). The proofs of the results in these sections are given in sections 3, 4, and 5, respectively. Finally, the Appendix recalls some inequalities for empirical processes that play a fundamental role in our proofs.

## 2. Main results.

2.1. *Consistent order estimation.* Let $(E, \mathcal{E}, \mu)$ be a measure space. For each $q \in \mathbb{N}$, let $\mathcal{M}_q$ be a given family of strictly positive probability densities with respect to $\mu$ (that is, we assume that $\int f d\mu = 1$ and that $f > 0$ $\mu$-a.e. for every $f \in \mathcal{M}_q$). Moreover, we assume that $(\mathcal{M}_q)_{q\in\mathbb{N}}$ is a nested family of models, that is, $\mathcal{M}_q \subset \mathcal{M}_{q+1}$ for all $q \in \mathbb{N}$. We also define $\mathcal{M} = \bigcup_{q\in\mathbb{N}} \mathcal{M}_q$.

Consider an i.i.d. sequence of $E$-valued random variables $(X_k)_{k\in\mathbb{N}}$ 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 $q^\star \in \mathbb{N}$ (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*. Neither $q^\star$ nor $f^\star$ are presumed to be known. Our aim is to estimate $q^\star$ from an observation sequence $(X_k)_{k\in\mathbb{N}}$. To this end, let

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

Evidently $\ell_n(f)$ is the log-likelihood of the i.i.d. sequence $(X_k)_{k\le n}$ under the measure where $X_k \sim f d\mu$. The *penalized likelihood order estimator* is defined by

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

where $\operatorname{pen}(n, q)$ is a penalty function. Our main 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 penalty. Let us emphasize that the maximum in the definition of $\hat{q}_n$ is taken over *all* model orders $q \in \mathbb{N}$, that is, we do not assume that an a priori upper bound on the order is available.

REMARK 2.1. To avoid measurability problems and other technical complications, we employ throughout this paper the simplifying convention that uncountable suprema (such as $\sup_{f\in\mathcal{M}_q} \ell_n(f)$) are interpreted as essential suprema with respect to the measure $\mathbf{P}^\star$. In applications the model classes $\mathcal{M}_q$ will typically be separable, so that the supremum and essential supremum coincide.

Let us begin by recalling the notion of bracketing.

DEFINITION 2.2. Given a class $\mathcal{Q}$ of measurable functions $g : E \to \mathbb{R}$, a finite collection of pairs of functions $\{g_i^L, g_i^U\}_{i=1,\ldots,N}$ is called a $(\mathcal{Q}, \delta)$-bracketing set if for every $g \in \mathcal{Q}$, there is a $j \in \{1, \ldots, N\}$ such that

$$g_j^L \le g \le g_j^U, \qquad \mathbf{E}^\star[(g_j^U(X_1) - g_j^L(X_1))^2] \le \delta^2.$$

We denote as $\mathcal{N}(\mathcal{Q}, \delta)$ the cardinality $N$ of the smallest $(\mathcal{Q}, \delta)$-bracketing set.

Our general consistency result is stated in terms of the bracketing numbers of a certain class of weighted densities $\mathcal{D}_q$ derived from $\mathcal{M}_q$, which we define presently. The significance of $\mathcal{D}_q$ follows immediately from the likelihood inequality obtained in Lemma 3.1 below, which plays a fundamental role in the proof.

DEFINITION 2.3. For any $q > q^\star$ and $f \in \mathcal{M}_q$ ($f \neq f^\star$), define

$$d_f = \frac{\sqrt{f/f^\star} - 1}{h(f, f^\star)}, \qquad h(f, f^\star)^2 = \int \bigl(\sqrt{f} - \sqrt{f^\star}\bigr)^2 d\mu$$

(that is, $h(f, f^\star)$ is the Hellinger distance between the densities $f$ and $f^\star$). Define $\mathcal{D}_q = \{d_f : f \in \mathcal{M}_q,\ f \neq f^\star\}$ for $q > q^\star$, and let $\mathcal{D} = \bigcup_{q > q^\star} \mathcal{D}_q$.

We can now formulate the main result of this section.

THEOREM 2.4. *Assume that the following hold.*

1. *There is an envelope function $D : E \to \mathbb{R}$ such that $|d| \le D$ for all $d \in \mathcal{D}$ and $R^2 = \mathbf{E}^\star(D(X_1)^2) < \infty$. Moreover, for every $q > q^\star$, we have*

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

2. *For every $q < q^\star$, the family*

$$\left\{ \log\left(\frac{f + f^\star}{2f^\star}\right) : f \in \mathcal{M}_q \right\}$$

*is a $\mathbf{P}^\star$-Glivenko-Cantelli class.*

*Define the penalty*

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

*where $\varpi$ is any function such that $\varpi(n) \to \infty$ as $n \to \infty$ (arbitrarily slowly), $n^{-1}\varpi(n) \log\log n \to 0$ as $n \to \infty$, and $\eta$ is any function such that $\eta(q) > \eta(q-1)$ for all $q > q^\star$ and such that for some constant $\beta > 0$, we have for every $q > q^\star$*

$$\sqrt{\beta\, q} \vee \int_0^R \sqrt{\log \mathcal{N}(\mathcal{D}_q, u)}\, du \leq \sqrt{\eta(q)} < \infty.$$

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

The proof of this theorem is given in section 3.

REMARK 2.5. The scaling function $\eta(q)$ is closely related to the dimension of the the model $\mathcal{M}_q$. Indeed, if $\mathcal{D}_q$ is a finite-dimensional family, one would typically expect that $\int_0^R \log^{1/2} \mathcal{N}(\mathcal{D}_q, u)\, du \;\propto\; (\dim \mathcal{D}_q)^{1/2}$. As the scaling factor $h(f, f^\star)^{-1}$ in the definition of $d_f$ becomes singular as $f \to f^\star$, one could think of $\eta(q)$ as the "local dimension" of $\mathcal{M}_q$ in a neighborhood of the true density $f^\star$.

REMARK 2.6. An alternative set of assumptions under which the conclusion of Theorem 2.4 holds can be obtained by adapting the method of proof used in [24] for the Markov order estimation problem. The key requirement here is that

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

for some constant $C$ and for any $\varepsilon, \delta > 0$ and $q \in \mathbb{N}$, where we have defined the Hellinger balls $\mathcal{M}_q(\varepsilon) = \{\sqrt{f} : f \in \mathcal{M}_q,\ h(f, f^\star) \leq \varepsilon\}$. In this case, a peeling device can be employed to avoid dealing with the weighted class $\mathcal{D}_q$. However, the proof relies crucially on the exact dependence of the local bracketing entropy $\log \mathcal{N}(\mathcal{M}_q(\varepsilon), \delta)$ on $\varepsilon/\delta$ given above. It is not sufficient to obtain a global entropy bound (that is, where the scaling in $\varepsilon$ is omitted), as additional logarithmic factors then appear in the proof which give rise to suboptimal penalties.

We must therefore choose between two alternatives: either establish (i) a *local* entropy bound directly on the model class $\mathcal{M}_q$, or (ii) a *global* entropy bound on the *weighted* model class $\mathcal{D}_q$. Alternative (i) implies essentially that the family $\mathcal{M}_q$, endowed with the Hellinger distance, has the same metric structure as a subset of $\mathbb{R}^{\eta(q)}$ endowed with the Euclidean metric. However, in the examples we have in mind, this is not typically the case. For example, finite mixture models (cf. section 2.3) possess a notoriously complicated geometry which is qualitatively different than that of Euclidean space, so that the local entropy approach is not well suited to such models. In this paper, we have therefore chosen to develop the more flexible alternative (ii). The interested reader may easily adapt the proof in [24] to obtain a version of Theorem 2.4 under assumptions corresponding to alternative (i).

Alternative definitions of the weighted model class $\mathcal{D}_q$ are possible, however, without changing the conclusion of Theorem 2.4. See Remark 3.2 below.

2.2. *Minimal penalties.* Theorem 2.4 shows that any penalty that increases faster than $\eta(q)\log\log n$ defines a strongly consistent order estimator. In this section, we will establish (under some mild additional assumptions) that this result is essentially optimal: we will show that the order estimator with penalty $\mathrm{pen}(n,q) = C\,\eta(q)\log\log n$ is inconsistent when the constant $C$ is chosen sufficiently small. Therefore Theorem 2.4 identifies in essence the minimal penalty that gives rise to strong consistency. The identification of the minimal penalty is relevant in practice, as such penalties minimize the probability of underestimating the order.

REMARK 2.7. Though the penalty $\mathrm{pen}(n,q) = C\,\eta(q)\log\log n$ gives an inconsistent order estimator for sufficiently small constants $C$, we have not yet discussed the consistency of this penalty for larger $C$. A careful reading of the proof of Theorem 2.4 shows that consistency is in fact achieved even for penalties of the form $\mathrm{pen}(n,q) = C\,\eta(q)\log\log n$, provided that $C \geq C_0$ for a certain threshold $C_0$ (see Remark 3.4). Unfortunately, the value of $C_0$ depends on $q^\star$, so that such penalties cannot be used for the purpose of order estimation in the absence of a prior upper bound on the order. It therefore appears that we can identify three regimes: that of inconsistent estimation ($\mathrm{pen}(n,q) = C\,\eta(q)\log\log n$ with $C$ small), consistent estimation with a prior upper bound on the order ($\mathrm{pen}(n,q) = C\,\eta(q)\log\log n$ with $C$ large), and consistent estimation without a prior upper bound on the order ($\mathrm{pen}(n,q) = \eta(q)\,\varpi(n)\log\log n$ with $\varpi(n)\to\infty$).

To state the main result of this section, we will need some additional notation.

DEFINITION 2.8. For $q \in \mathbb{N}$ and $\varepsilon > 0$, define the families

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

where the closure $\mathrm{cl}\,\mathcal{D}_q(\varepsilon)$ 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 $\mathcal{M}_q$. We will require some assumptions on the richness of neighborhoods of $\bar{\mathcal{D}}_q$.

DEFINITION 2.9. 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 2.10. *Assume 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.*

*Define the penalty*
$$\mathrm{pen}(n, q) = C\, \eta(q) \log\log n,$$
*where $\eta$ is any nonnegative function such that $\eta(q) > \eta(q^\star)$. 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 this theorem is given in section 4.

REMARK 2.11. The proof of Theorem 2.10 will show 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. Thus imposing a prior upper bound on the order does not alter the conclusion of Theorem 2.10.

2.3. *Application to mixtures.* The general mixture order estimation problem can be defined as follows. Let $\mathcal{P}$ be a given family of strictly positive probability densities with respect to $\mu$. For $q \in \mathbb{N}$, we define the model class
$$\mathcal{M}_q = \left\{ \sum_{i=1}^q \pi_i f_i : \pi_i \ge 0,\ \sum_{i=1}^q \pi_i = 1,\ f_i \in \mathcal{P} \right\}$$
to be the family of mixtures of $q$ elements of $\mathcal{P}$. In this setting, the model order is the smallest number of mixture components that is needed to describe the distribution of the data $(X_k)_{k \ge 0}$. We aim to apply Theorem 2.4 to obtain strongly consistent penalized likelihood mixture order estimators.

The key problem is evidently to bound the bracketing number $\mathcal{N}(\mathcal{D}_q, \delta)$. This appears to be a novel and nontrivial problem. As the normalizer $h(f, f^\star)^{-1}$ in the definition of the weighted class $\mathcal{D}_q$ becomes singular as $f \to f^\star$, it is the *local* geometry of the mixture family in a neighborhood of $f^\star$ that is of interest (in contrast to the unweighted entropy computations in [12, 13]). Unfortunately, the geometry of finite mixtures is notoriously complicated due to the lack of identifiability (see, for example, [8]), and we are not aware of any quantitative results in this direction. To complicate matters further, the application of Theorem 2.4 requires that our bounds on $\mathcal{N}(\mathcal{D}_q, \delta)$ hold uniformly in the model order $q$.

In the present paper, we provide a detailed analysis of the local geometry of one-dimensional location mixtures, which leads to the requisite bounds on $\mathcal{N}(\mathcal{D}_q, \delta)$ in this setting. Let $E = \mathbb{R}$ and let $\mu$ be the Lebesgue measure on $\mathbb{R}$. We fix a constant $T > 0$ and a strictly positive probability density $f_0$ with respect to $\mu$. We will consider mixtures of probability densities in the class

$$\mathcal{P} = \{f_\theta : \theta \in [-T, T]\}, \qquad f_\theta(x) = f_0(x - \theta) \ \forall\, x \in \mathbb{R}.$$

The mixtures in $\mathcal{M}_q$ are known as location mixtures, as each mixture component is obtained from the mother function $f_0$ by a shift of location.

To obtain our main result, we impose some regularity assumptions on $f_0$.

ASSUMPTION A. $f_0$ has three continuous derivatives, such that:

1. The functions $x \mapsto e^{tx} f_0(x)$, $x \mapsto e^{tx} f_0'(x)$, $x \mapsto e^{tx} f_0''(x)$ are in $L^1(d\mu)$ and $x \mapsto e^{tx} f_0(x)$, $x \mapsto e^{tx} f_0'(x)$ vanish at infinity for each $t \in \mathbb{R}$.
2. Define the functions $x \mapsto H_k(x) = \sup_{\theta \in [-T,T]} |\partial^k f_\theta(x)/\partial x^k| / f^\star$. Then $H_k \in L^4(f^\star d\mu)$ for $k = 0, 1, 2$ and $H_3 \in L^2(f^\star d\mu)$.

It is easily verified that Assumption A is satisfied for $f_0(x) = e^{-x^2/2\sigma^2}/\sqrt{2\pi\sigma^2}$ (and any $f^\star \in \mathcal{M}$), so that our results apply directly to Gaussian location mixtures.

We can now formulate the main result of this section.

THEOREM 2.12. *Suppose that Assumption A holds. Then there exist constants $C^\star$ and $\delta^\star$, which depend on $f^\star$ but not on $q$ or $\delta$, such that*

$$\mathcal{N}(\mathcal{D}_q, \delta) \le \left(\frac{C^\star}{\delta}\right)^{36q} \qquad \textit{for all } q > q^\star,\ \delta \le \delta^\star.$$

*Moreover, there is a function $D \in L^4(f^\star d\mu)$ such that $|d| \le D$ for all $d \in \mathcal{D}$.*

The proof of this result is given in section 5. Though the general approach of the proof should extend to a much larger class of mixture models, the details of the analysis of the local geometry of $\mathcal{M}$ rely on Laplace transform techniques which are specific to the location mixture model under consideration. The extension of our results to more general mixture models remains a challenging open problem.

REMARK 2.13. We have made no attempt to optimize the constants in Theorem 2.12. In particular, the factor 36 in the exponent can likely be improved.

Combining Theorems 2.4 and 2.12, we can now obtain the following result.

COROLLARY 2.14. *Suppose that Assumption A holds. Define the penalty*

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

*where $\omega$ is any function with $\omega(n) \to \infty$ and $n^{-1}\omega(n)\log\log n \to 0$ as $n \to \infty$. Then the penalized likelihood mixture order estimator is strongly consistent.*

On the other hand, Theorem 2.10 can be used to prove the following.

PROPOSITION 2.15. *Suppose that Assumption A holds. Define the penalty*

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

*If the constant $C > 0$ is chosen sufficiently small, then the penalized likelihood mixture order estimator is not strongly consistent.*

The proofs of Corollary 2.14 and Proposition 2.15 are given in section 5.

**3. Proof of Theorem 2.4.** Define the empirical process

$$\nu_n(g) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left\{g(X_i) - \mathbf{E}^\star(g(X_1))\right\}.$$

The proof of Theorem 2.4 is based on a simple likelihood ratio inequality, which relates the log-likelihood ratio $\ell_n(f) - \ell_n(f^\star)$ to the empirical process. Related inequalities appear in [10, 4], but the following form is perhaps the most natural.

LEMMA 3.1. *For any strictly positive probability density $f \neq f^\star$, we have*

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

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. □

REMARK 3.2. Along similar lines, one can prove the inequalities

$$\ell_n(f) - \ell_n(f^\star) \le \frac{1}{4}\left|\nu_n\left(\frac{\log(f/f^\star)}{\sqrt{D(f^\star||f)}}\right)\right|^2$$

(which improves on [4], Proposition A.1) and

$$\ell_n(f) - \ell_n(f^\star) \le \frac{1}{2}\left|\nu_n\left(\frac{\log(\{f+f^\star\}/2f^\star)}{\sqrt{D(f^\star||\{f+f^\star\}/2)}}\right)\right|^2,$$

where $D(f^\star||f) = \int \log(f^\star/f) f^\star d\mu$ is the relative entropy. By using these inequalities instead of Lemma 3.1, the proof below can be repeated to show that Theorem 2.4 still holds if we replace the definition of $d_f$ in Definition 2.3 by $d_f = \log(f/f^\star)/\sqrt{D(f^\star||f)}$ or by $d_f = \log(\{f+f^\star\}/2f^\star)/\sqrt{D(f^\star||\{f+f^\star\}/2)}$.

At the heart of the proof of Theorem 2.4 lies the following law of iterated logarithm, which holds uniformly in the model order $q > q^\star$.

THEOREM 3.3. *Assume that* $\sup_{d\in\mathcal{D}}|d| \le D$ *for a function* $D : E \to \mathbb{R}$ *with* $R^2 = \mathbf{E}^\star(D(X_1)^2) < \infty$, *and that* $\eta : \mathbb{N} \to \mathbb{R}_+$ *and* $\beta > 0$ *are defined such that*

$$\sqrt{\beta\, q} \vee \int_0^R \sqrt{\log \mathcal{N}(\mathcal{D}_q, u)}\, du \le \sqrt{\eta(q)} < \infty$$

*for every* $q > q^\star$. *Then*

$$\limsup_{n\to\infty} \frac{1}{\log\log n} \sup_{q>q^\star} \frac{1}{\eta(q)} \left\{ \sup_{f\in\mathcal{M}_q} \ell_n(f) - \sup_{f\in\mathcal{M}_{q^\star}} \ell_n(f) \right\} \le \tilde{C} \quad \mathbf{P}^\star\text{-a.s.}$$

*for a sufficiently large constant* $\tilde{C} > 0$ *(depending only on* $\beta$ *and* $R$*).*

PROOF. We proceed in several steps.

**Step 1** (*blocking and truncation*). By Lemma 3.1, we have

$$\sup_{f\in\mathcal{M}_q} \ell_n(f) - \sup_{f\in\mathcal{M}_{q^\star}} \ell_n(f) \le \sup_{f\in\mathcal{M}_q} \{\ell_n(f) - \ell_n(f^\star)\} \le \sup_{f\in\mathcal{M}_q} |\nu_n(d_f)|^2.$$

Therefore, we can estimate as follows:

$$\begin{aligned}
&\max_{n=2^N,\dots,2^{N+1}} \frac{1}{\log\log n} \sup_{q>q^\star} \frac{1}{\eta(q)} \left\{ \sup_{f\in\mathcal{M}_q} \ell_n(f) - \sup_{f\in\mathcal{M}_{q^\star}} \ell_n(f) \right\} \\
&\quad\le \left( \max_{n=2^N,\dots,2^{N+1}} \frac{1}{\sqrt{\log\log n}} \sup_{q>q^\star} \frac{1}{\sqrt{\eta(q)}} \sup_{f\in\mathcal{M}_q} |\nu_n(d_f)| \right)^2 \\
&\quad\le \left( \max_{n=2^N,\dots,2^{N+1}} \frac{1}{\sqrt{\log\log n}} \sup_{q>q^\star} \frac{1}{\sqrt{\eta(q)}} \sup_{f\in\mathcal{M}_q} |\nu_n(d_f \mathbf{1}_{D>a_N(q)})| \right. \\
&\qquad\left. + \max_{n=2^N,\dots,2^{N+1}} \frac{1}{\sqrt{\log\log n}} \sup_{q>q^\star} \frac{1}{\sqrt{\eta(q)}} \sup_{f\in\mathcal{M}_q} |\nu_n(d_f \mathbf{1}_{D\le a_N(q)})| \right)^2,
\end{aligned}$$

where we introduce

$$a_N(q) = \sqrt{\frac{2^N}{C_2 \eta(q) \log\log 2^N}},$$

and $C_2$ is a constant to be chosen later on.

**Step 2** (*the first term*). Note that for $n = 2^N, \dots, 2^{N+1}$

$$\begin{aligned}
&|\nu_n(d_f \mathbf{1}_{D>a_N(q)})| \\
&\quad= \frac{1}{\sqrt{n}} \left| \sum_{i=1}^n \left\{ d_f(X_i)\mathbf{1}_{D>a_N(q)}(X_i) - \mathbf{E}^\star(d_f(X_1)\mathbf{1}_{D>a_N(q)}(X_1)) \right\} \right| \\
&\quad\le \frac{1}{\sqrt{n}} \sum_{i=1}^n \left\{ D(X_i)\mathbf{1}_{D>a_N(q)}(X_i) + \mathbf{E}^\star(D(X_1)\mathbf{1}_{D>a_N(q)}(X_1)) \right\} \\
&\quad\le \frac{1}{a_N(q)\sqrt{2^N}} \sum_{i=1}^{2^{N+1}} \left\{ D(X_i)^2 + R^2 \right\}.
\end{aligned}$$

Therefore,

$$\begin{aligned}
&\limsup_{N\to\infty} \max_{n=2^N,\dots,2^{N+1}} \frac{1}{\sqrt{\log\log n}} \sup_{q>q^\star} \frac{1}{\sqrt{\eta(q)}} \sup_{f\in\mathcal{M}_q} |\nu_n(d_f \mathbf{1}_{D>a_N(q)})| \\
&\qquad\le \lim_{N\to\infty} \frac{2\sqrt{C_2}}{2^{N+1}} \sum_{i=1}^{2^{N+1}} \left\{ D(X_i)^2 + R^2 \right\} = 4R^2\sqrt{C_2} \quad \mathbf{P}^\star\text{-a.s.}
\end{aligned}$$

by the law of large numbers.

**Step 3** (*the second term*). Let $C_3$ be a constant to be chosen later on. Note that

$$\mathbf{P}^\star\left[\max_{n=2^N,\ldots,2^{N+1}}\frac{1}{\sqrt{\log\log n}}\sup_{q>q^\star}\frac{1}{\sqrt{\eta(q)}}\sup_{f\in\mathcal{M}_q}|\nu_n(d_f\mathbf{1}_{D\le a_N(q)})|>3C_3\right]\le$$
$$\sum_{q=q^\star+1}^{\infty}\mathbf{P}^\star\left[\max_{n=2^N,\ldots,2^{N+1}}\sup_{f\in\mathcal{M}_q}|S_n(d_f\mathbf{1}_{D\le a_N(q)})|>3C_3\sqrt{\eta(q)2^N\log\log 2^N}\right]$$

where $S_n(f)=\sqrt{n}\,\nu_n(f)$. By Proposition A.2, we have

$$\mathbf{P}^\star\left[\max_{n=2^N,\ldots,2^{N+1}}\frac{1}{\sqrt{\log\log n}}\sup_{q>q^\star}\frac{1}{\sqrt{\eta(q)}}\sup_{f\in\mathcal{M}_q}|\nu_n(d_f\mathbf{1}_{D\le a_N(q)})|>3C_3\right]\le$$
$$\sum_{q=q^\star+1}^{\infty}3\max_{n=2^N,\ldots,2^{N+1}}\mathbf{P}^\star\left[\sup_{d\in\mathcal{D}_q}|\nu_n(d\mathbf{1}_{D\le a_N(q)})|>C_3\sqrt{\tfrac{1}{2}\eta(q)\log\log 2^N}\right].$$

Now note that

$$\sup_{d\in\mathcal{D}_q}\|d\mathbf{1}_{D\le a_N(q)}\|_\infty\le a_N(q),\qquad\sup_{d\in\mathcal{D}_q}\mathbf{E}^\star[\{d(X_1)\mathbf{1}_{D\le a_N(q)}(X_1)\}^2]\le R^2,$$

and

$$C_3\sqrt{\tfrac{1}{2}\eta(q)\log\log 2^N}=\left(\frac{C_3}{R^2\sqrt{2C_2}}\right)\frac{R^2\sqrt{2^N}}{a_N(q)}\le\left(\frac{C_3}{R^2\sqrt{2C_2}}\right)\frac{R^2\sqrt{n}}{a_N(q)}$$

for all $n=2^N,\ldots,2^{N+1}$. Moreover, if $\{f_i^L,f_i^U\}$ is a $(\mathcal{D}_q,\delta)$-bracketing set, then $\{f_i^L\mathbf{1}_{D\le a_N(q)},f_i^U\mathbf{1}_{D\le a_N(q)}\}$ is a $(\mathcal{D}_q\mathbf{1}_{D\le a_N(q)},\delta)$-bracketing set. Therefore

$$\begin{aligned}&C\sqrt{C_1+1}\int_0^R\sqrt{\log\mathcal{N}(\mathcal{D}_q\mathbf{1}_{D\le a_N(q)},u)}\,du\\&\le C\sqrt{C_1+1}\int_0^R\sqrt{\log\mathcal{N}(\mathcal{D}_q,u)}\,du\\&\le C\sqrt{C_1+1}\sqrt{\eta(q)}\le C_3\sqrt{\tfrac{1}{2}\eta(q)\log\log 2^N}\end{aligned}$$

for all $q>q^\star$ simultaneously when $N$ is sufficiently large, regardless of the choice of $C_1$. Choosing $C_1=C_3/R^2\sqrt{2C_2}$, Proposition A.1 gives

$$\mathbf{P}^\star\left[\max_{n=2^N,\ldots,2^{N+1}}\frac{1}{\sqrt{\log\log n}}\sup_{q>q^\star}\frac{1}{\sqrt{\eta(q)}}\sup_{f\in\mathcal{M}_q}|\nu_n(d_f\mathbf{1}_{D\le a_N(q)})|>3C_3\right]$$
$$\le\sum_{q=q^\star+1}^{\infty}6\,\exp\left[-\frac{C_3^2\eta(q)\log\log 2^N}{2C^2(C_1+1)R^2}\right]$$

for $N$ sufficiently large. But we clearly have

$$\begin{aligned}\sum_{q=q^\star+1}^{\infty} 6\,\exp\left[-\frac{C_3^2\eta(q)\log\log 2^N}{2C^2(C_1+1)R^2}\right]&\\ \le \sum_{q=1}^{\infty} 6\left(e^{-C_3^2\log\log 2/2C^2(C_1+1)R^2}N^{-C_3^2/2C^2(C_1+1)R^2}\right)^{\eta(q)}&\\ \le 12\,e^{-\beta C_3^2\log\log 2/2C^2(C_1+1)R^2}N^{-\beta C_3^2/2C^2(C_1+1)R^2}&\end{aligned}$$

for $N$ sufficiently large. As $C_1 = C_3/R^2\sqrt{2C_2}$, we may choose $C_2 = C_3^2/2R^4$ and $C_3$ sufficiently large so that $\beta C_3^2/2C^2(C_1+1)R^2 = \beta C_3^2/4C^2R^2 > 1$. Then

$$\sum_{N=1}^{\infty}\mathbf{P}^\star\left[\max_{n=2^N,\ldots,2^{N+1}}\frac{1}{\sqrt{\log\log n}}\sup_{q>q^\star}\frac{1}{\sqrt{\eta(q)}}\sup_{f\in\mathcal{M}_q}|\nu_n(d_f\mathbf{1}_{D\le a_N(q)})| > 3C_3\right] < \infty,$$

so that in particular $\mathbf{P}^\star$-a.s.

$$\limsup_{N\to\infty}\max_{n=2^N,\ldots,2^{N+1}}\frac{1}{\sqrt{\log\log n}}\sup_{q>q^\star}\frac{1}{\sqrt{\eta(q)}}\sup_{f\in\mathcal{M}_q}|\nu_n(d_f\mathbf{1}_{D\le a_N(q)})| \le 3C_3$$

by the Borel-Cantelli lemma.

**Step 4** (*end of proof*). Putting it all together, we obtain

$$\begin{aligned}\limsup_{N\to\infty}\max_{n=2^N,\ldots,2^{N+1}}\frac{1}{\log\log n}\sup_{q>q^\star}\frac{1}{\eta(q)}\left\{\sup_{f\in\mathcal{M}_q}\ell_n(f) - \sup_{f\in\mathcal{M}_{q^\star}}\ell_n(f)\right\}&\\ \le C_3^2\left(3+2\sqrt{2}\right)^2 \quad \mathbf{P}^\star\text{-a.s.}&\end{aligned}$$

for $C_3 > 2CR/\sqrt{\beta}$. The proof is easily completed. □

We can now complete the proof of Theorem 2.4.

PROOF OF THEOREM 2.4. By Theorem 3.3 and easy manipulations, we have

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

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

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

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, note that as $\log x$ is a concave function, we have the basic inequality $\log x \le 2\log(\{x+1\}/2)$ for all $x > 0$. Therefore, we obtain $\mathbf{P}^\star$-a.s.

$$\begin{aligned}
\limsup_{n\to\infty} \frac{1}{n}\left\{\sup_{f\in\mathcal{M}_q} \ell_n(f) - \sup_{f\in\mathcal{M}_{q^\star}} \ell_n(f)\right\} &\le \limsup_{n\to\infty} \sup_{f\in\mathcal{M}_q} \frac{\ell_n(f) - \ell_n(f^\star)}{n} \\
&\le \lim_{n\to\infty} \sup_{f\in\mathcal{M}_q} \frac{1}{n}\sum_{i=1}^n 2\,\log\left(\frac{f(X_i)+f^\star(X_i)}{2f^\star(X_i)}\right) \\
&= 2 \sup_{f\in\mathcal{M}_q} \mathbf{E}^\star\left[\log\left(\frac{f(X_1)+f^\star(X_1)}{2f^\star(X_1)}\right)\right] \\
&= -2 \inf_{f\in\mathcal{M}_q} D\left(f^\star\,\middle\|\,\frac{f+f^\star}{2}\right)
\end{aligned}$$

for $q < q^\star$ using the Glivenko-Cantelli property, where $D(f^\star||f)$ denotes the relative entropy (see Remark 3.2). We now claim that

$$\min_{q<q^\star} \inf_{f\in\mathcal{M}_q} D\left(f^\star\,\middle\|\,\frac{f+f^\star}{2}\right) > 0.$$

Indeed, suppose this is not the case. Then for some $q < q^\star$, there is a sequence $(f_n)_{n\in\mathbb{N}} \subset \mathcal{M}_q$ such that $D(f^\star||\{f_n+f^\star\}/2) \to 0$ as $n\to\infty$. By Pinsker's inequality, this implies that $f_n \to f^\star$ in $L^1(d\mu)$, so that $f^\star \in \mathrm{cl}\,\mathcal{M}_q$ for some $q < q^\star$. But $\mathrm{cl}\,\mathcal{M}_q \subset \mathrm{cl}\,\mathcal{M}_{q^\star-1}$ and $f^\star \in \mathcal{M}_{q^\star}\setminus \mathrm{cl}\,\mathcal{M}_{q^\star-1}$ by assumption, giving a contradiction. Thus the claim is established. To complete the proof, it suffices to note that by assumption $\mathrm{pen}(n,q)/n \to 0$ as $n\to\infty$. Therefore $\mathbf{P}^\star$-a.s.

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

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

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

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. □

REMARK 3.4. If we were to choose $\varpi(n) = C_0$ to be constant, rather than $\varpi(n) \to \infty$ as required by Theorem 2.4 (see section 2.2), then we would obtain in the first equation display of the above proof the upper bound

$$\frac{\eta(q^\star+1)}{\eta(q^\star+1)-\eta(q^\star)}\frac{\tilde{C}}{C_0},$$

where $\tilde{C}$ depends on $\beta$ and $R$. To obtain a consistent order estimator, we must choose $C_0$ sufficiently large so that this constant is less than one. As typically $\sup_{q\geq 0}\{\eta(q+1)/(\eta(q+1)-\eta(q))\} = \infty$, however, we cannot control this constant without prior knowledge of $q^\star$. It is for this reason that we must require $\varpi(n) \to \infty$ to obtain a computable order estimator. Let us note that even if we were to have $\sup_{q\geq 0}\{\eta(q+1)/(\eta(q+1)-\eta(q))\} < \infty$ (that is, when $\eta(q)$ grows exponentially with $q$), it will still typically be the case that the bracketing numbers $\mathcal{N}(\mathcal{D}_q, \delta)$ depend on $f^\star$ (as the class $\mathcal{D}_q$ itself depends on $f^\star$), so that we cannot choose $C_0$ without prior knowledge of $f^\star$. The Markov order estimation problem treated in [24] is a special case where all these parameters can be chosen independent of $q^\star$ and $f^\star$, which accounts for the slightly smaller penalty used there.

**4. Proof of Theorem 2.10.** The main ingredient of the proof of Theorem 2.10 is a precise characterization of the fluctuations of the likelihood ratio test for two model classes $\mathcal{M}_p$ and $\mathcal{M}_q$, which may be of independent interest. The proof of Theorem 2.10 will follow easily from this result. In the sequel, we denote by $\langle f, g\rangle = \int fgf^\star d\mu$ be the Hilbert space inner product in $L^2(f^\star d\mu)$, and we denote by $\|g\|_2^2 = \langle g, g\rangle$ the corresponding Hilbert space norm.

THEOREM 4.1. *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* $L_0^2(f^\star d\mu) = \{g \in L^2(f^\star d\mu) : \|g\|_2 \le 1,\ \langle 1, g\rangle = 0\}$.

REMARK 4.2. 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 4.1 provides the exact characterization

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

However, only the first (lower bound) part of the theorem will be needed to prove Theorem 2.10. We provide the more precise version of the theorem here due to its independent interest: we are not aware of a similar characterization of the pathwise fluctuations of the likelihood ratio test in the literature.

The proof of Theorem 4.1 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 4.3. *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* $(I_n)_{n\ge 0}$ *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 [20], Theorem 4.2 or in [16], Theorem 9.

We will also need the following simple result on partial maxima.

LEMMA 4.4. *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 $n \to \infty$.*

PROOF. Fix $\alpha > 0$. As $\mathbf{E}[|X_1|^p] < \infty$ and $(X_i)_{i\geq 1}$ are i.i.d., we have

$$\sum_{n=1}^{\infty}\mathbf{P}[|X_n| > n^{1/p}\alpha] = \sum_{n=1}^{\infty}\mathbf{P}[|X_1|^p/\alpha^p > n] = \sum_{n=1}^{\infty}\int_{n-1}^{n}\mathbf{P}[|X_1|^p/\alpha^p > \lceil x\rceil]\,dx \leq \mathbf{E}[|X_1|^p]/\alpha^p < \infty.$$

By the Borel-Cantelli lemma, $|X_n| \leq n^{1/p}\alpha$ eventually a.s. Let $\tau < \infty$ a.s. be such that $|X_n| \leq n^{1/p}\alpha$ for all $n > \tau$. Then $\max_{i=1,\ldots,n}|X_i| \leq n^{1/p}\alpha \vee \max_{i=1,\ldots,\tau}|X_i|$ for all $n > \tau$. It follows directly that $\limsup_{n\to\infty} n^{-1/p}\max_{i=1,\ldots,n}|X_i| \leq \alpha$ a.s. But as $\alpha$ was arbitrary, this establishes the claim. □

We can now obtain the following asymptotic expansion of the log-likelihood, which provides an almost sure counterpart to the corresponding results in [10, 17].

PROPOSITION 4.5. *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*

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

*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 3.1,

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

Therefore, we can estimate

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

Now note that we can estimate

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

The first term on the right converges to zero $\mathbf{P}^\star$-a.s. as $n\to\infty$ by Theorem 4.3, while the second term is easily seen to equal $\sup_{d\in\mathcal{D}_q}\|d-\langle 1,d\rangle\|_2\le 1$. Therefore

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

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

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

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

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

But the reverse inequality clearly holds for all $n\ge 0$, 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 $2\log(1+x)=2x-x^2+x^2R(x)$, where $R(x)\to 0$ as $x\to 0$. Thus we can write, for any $f\in\mathcal{M}_q$,

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

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

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

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

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

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

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

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

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 4.4 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 4.6. *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*

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

PROOF. By Proposition 4.5, 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

$$\sup_{d\in\bar{\mathcal{D}}_q} (I_{\tau_n}(d))_+^2 - \sup_{f\in\mathcal{M}_q(4\sqrt{\log\log \tau_n/\tau_n})} (I_{\tau_n}(d_f))_+^2 \leq -\varepsilon \quad \text{for all } n \tag{4.1}$$

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

By Theorem 4.3 (which can be applied here as $\mathcal{N}(\mathcal{D}_q, \delta) = \mathcal{N}(\mathrm{cl}\,\mathcal{D}_q, \delta)$ for all $\delta > 0$), the process $(I_{\tau_n})_{n\geq 0}$ is $\mathbf{P}^\star$-a.s. relatively compact in $\ell_\infty(\mathrm{cl}\,\mathcal{D}_q)$ with

$$\inf_{g\in L_0^2(f^\star d\mu)} \sup_{d\in\mathrm{cl}\,\mathcal{D}_q} |I_{\tau_n}(d) - \langle d, g\rangle| \xrightarrow{n\to\infty} 0 \quad \mathbf{P}^\star\text{-a.s.} \tag{4.2}$$

Then there is a set of positive probability on which (4.1) and (4.2) 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

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

where we have used the elementary estimate $|a_+^2 - b_+^2| = |a_+ - b_+|(a_+ + b_+) \leq |a_+ - b_+|(|a_+ - b_+| + 2b_+) \leq |a - b|(|a - b| + 2|b|)$ for any $a, b \in \mathbb{R}$, and the fact that $\sup_{d \in \operatorname{cl} \mathcal{D}_q} |\langle d, g \rangle| \leq \sup_{d \in \operatorname{cl} \mathcal{D}_q} \|d\|_2 \|g\|_2 \leq 1$. Thus (4.1) gives

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

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

But as $d \mapsto \langle d, g \rangle$ is continuous in $L^2(f^\star d\mu)$ and $\operatorname{cl} \mathcal{D}_q(4\sqrt{\log\log \sigma_n/\sigma_n})$ 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

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

$$\sup_{d \in \bigcap_{n\geq 0} \operatorname{cl} \mathcal{D}_q(4\sqrt{\log\log \sigma_n/\sigma_n})} (\langle d, g \rangle)_+^2 = \sup_{d \in \bar{\mathcal{D}}_q} (\langle d, g \rangle)_+^2.$$

Thus we have a contradiction, completing the proof. □

We now obtain a converse to the previous result.

PROPOSITION 4.7. *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*

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

PROOF. Suppose that the result does not hold true. By Proposition 4.5, 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 4.6, 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

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

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

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

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 4.1.

PROOF OF THEOREM 4.1. We obtain separately the lower and upper bounds.
**Lower bound.** By Propositions 4.6 and 4.7, we have

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

Now fix any $g \in L_0^2(f^\star d\mu)$. By Theorem 4.3 (which can be applied here as $\mathcal{N}(\mathcal{D}_q,\delta) = \mathcal{N}(\operatorname{cl}\mathcal{D}_q,\delta) \geq \mathcal{N}(\bar{\mathcal{D}}_q,\delta)$ 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

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

$\mathbf{P}^\star$-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 4.6 and 4.7, we have

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

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

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

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

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

It follows as in the proof of Proposition 4.6 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 Theorem 2.10.

PROOF OF THEOREM 2.10. It suffices to prove that

$$\Gamma := \sup_{g\in L_0^2(f^\star d\mu)} \left\{ \sup_{d\in\bar{\mathcal{D}}_q^c} (\langle d,g\rangle)_+^2 - \sup_{d\in\bar{\mathcal{D}}_{q^\star}} (\langle d,g\rangle)_+^2 \right\} > 0. \tag{4.3}$$

Indeed, by Theorem 4.1, we have

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

Thus if (4.3) holds, then 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.

To prove (4.3), note that as $f/f^\star - 1 = (\sqrt{f/f^\star} - 1)(\sqrt{f/f^\star} + 1)$, we can estimate for any $f \in \mathcal{M}_q \backslash \{f^\star\}$ using Hölder's inequality

$$|\langle 1, d_f \rangle| = \left| \int d_f f^\star d\mu \right| \leq \int \left| d_f - \frac{f/f^\star - 1}{2h(f, f^\star)} \right| f^\star d\mu \leq \frac{h(f, f^\star)}{2}.$$

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

$$|\langle 1, d_0 \rangle| = \lim_{n \to \infty} \left| \int d_{f_n} f^\star d\mu \right| \leq \lim_{n \to \infty} \frac{h(f_n, f^\star)}{2} = 0.$$

Moreover, it is immediate that $\|d_0\|_2 \leq 1$. 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 (4.3) holds, and the proof is complete. □

## 5. Proof of Theorem 2.12.

5.1. *The local geometry of* $\mathcal{M}$. As $f^\star \in \mathcal{M}_{q^\star}$, we can clearly write

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

Without loss of generality, we will assume that

$$-T \leq \theta_1^\star < \theta_2^\star < \cdots < \theta_{q^\star}^\star \leq T.$$

In the following, let us fix some parameters $\bar{\theta}_1, \ldots, \bar{\theta}_{q^\star} \in [-T, T]$ such that

$$-T = \bar{\theta}_1 \leq \theta_1^\star < \bar{\theta}_2 < \theta_2^\star < \cdots < \bar{\theta}_{q^\star} < \theta_{q^\star}^\star \leq T.$$

The precise choice of $\bar{\theta}_1, \ldots, \bar{\theta}_{q^\star}$ only affects the constants in the proofs below, and is therefore irrelevant to our final result. We only presume that $\bar{\theta}_1, \ldots, \bar{\theta}_{q^\star}$ remain fixed throughout. Define the intervals $A_i = [\bar{\theta}_i, \bar{\theta}_{i+1}[$ for $i = 1, \ldots, q^\star - 1$ and $A_{q^\star} = [\bar{\theta}_{q^\star}, T]$. Then $A_1, \ldots, A_{q^\star}$ partition the parameter set $[-T, T]$ in such a way that each interval contains precisely one component of the mixture $f^\star$.

Let us define for each $k \geq 0$ and probability measure $\lambda$ on $[-T, T]$ the functions

$$D_k f_\theta(x) = \frac{\partial^k}{\partial \theta^k} f_\theta(x) = (-1)^k \frac{\partial^k}{\partial x^k} f_\theta(x), \qquad f_\lambda(x) = \int f_\theta(x)\, \lambda(d\theta).$$

Denote by $\mathfrak{P}(A)$ the space of probability measures supported on $A \subseteq [-T, T]$.

DEFINITION 5.1. Let us write

$$\mathfrak{D} = \{(\eta, \beta, \rho, \tau, \nu) : \eta, \beta \in \mathbb{R}^{q^\star},\ \rho, \tau \in \mathbb{R}_+^{q^\star},\ \nu \in \mathfrak{P}(A_1) \times \cdots \times \mathfrak{P}(A_{q^\star})\}.$$

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

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

and the nonnegative quantity

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

Denote by $\|\cdot\|_p$ the $L^p(f^\star d\mu)$-norm, that is, $\|f\|_p^p = \int f(x)^p f^\star(x) \mu(dx)$. We can now formulate the key result on the local geometry of the mixture class $\mathcal{M}$.

THEOREM 5.2. *Suppose that Assumption A holds. Then there exists a constant $c^\star > 0$ (depending on $f^\star$ and $\bar\theta_1, \ldots, \bar\theta_{q^\star}$ but not on $\eta, \beta, \rho, \tau, \nu$) such that*

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

This result has several important consequences. To provide some basic intuition, recall that the total variation distance between $f$ and $f^\star$ is defined as

$$\|f - f^\star\|_{\mathrm{TV}} = \int |f - f^\star|\, d\mu = \left\| \frac{f - f^\star}{f^\star} \right\|_1.$$

Let $f \in \mathcal{M}_q$, so that we may write $f = \sum_{i=1}^q \pi_i f_{\theta_i}$. Then Theorem 5.2 shows that

$$\|f - f^\star\|_{\mathrm{TV}} \geq c^\star \sum_{i=1}^{q^\star} \left\{ \left| \sum_{j:\theta_j \in A_i} \pi_j - \pi_i^\star \right| + \frac{1}{2} \sum_{j:\theta_j \in A_i} \pi_j (\theta_j - \theta_i^\star)^2 \right\}.$$

As $\|f - f^\star\|_{\mathrm{TV}} \leq 2h(f, f^\star)$ (see, for example, [21], chapter III, section 9) this provides control on the geometry of Hellinger neighborhoods of $f^\star$ when viewed as a subset of the mixture parameters $(\pi_1, \ldots, \pi_q, \theta_1, \ldots, \theta_q)$. Similarly, using Theorem 5.2 and a Taylor expansion of $f_\theta$, one can show that any limit point of $d_f$ as $f \to f^\star$ is in the closure of $\{\ell(\eta, \beta, \rho, \tau, \nu) : (\eta, \beta, \rho, \tau, \nu) \in \mathfrak{D}\}$. The proof of Theorem 2.12 is based on more precise variants of these ideas.

We now turn to the proof of Theorem 5.2. Define the bilateral Laplace transform

$$L[f](t) = \int e^{tx}\, f(x)\, dx$$

for all $t \in \mathbb{R}$. Note that under Assumption A, integration by parts gives

$$L[f'](t) = -t\, L[f](t), \qquad L[f''](t) = t^2\, L[f](t),$$

and we have

$$L[D_k f_\theta](t) = t^k\, e^{\theta t}\, L[f](t), \qquad k = 0, 1, 2.$$

The fundamental use of Laplace transform techniques in the proof of Theorem 5.2 is the main reason that our main result is restricted to location mixtures. We conjecture that the conclusion of Theorem 5.2 holds in a much more general setting, but a proof for general mixture models would likely require a different approach. Let us note that the use of Laplace transforms is somewhat reminiscent of the approach used in [22] to establish weak identifiability of finite mixtures.

PROOF OF THEOREM 5.2. Suppose that the conclusion of the theorem does not hold. Then there must exist a sequence of coefficients $(\eta^n, \beta^n, \rho^n, \delta^n, \nu^n) \in \mathfrak{D}$ such that $\|\ell(\eta^n, \beta^n, \rho^n, \delta^n, \nu^n)\|_1 / N(\eta^n, \beta^n, \rho^n, \delta^n, \nu^n)$ tends to 0.

Applying Taylor's theorem to $\theta \mapsto f_\theta$, we can write

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

where $\lambda_i^n$ is the probability measure on $A_i$ defined by

$$\int g(\theta)\, \lambda_i^n(d\theta) = \frac{\int g(\theta)\, (\theta - \theta_i^\star)^2\, \nu_i^n(d\theta)}{\int (\theta - \theta_i^\star)^2\, \nu_i^n(d\theta)}$$

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

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

$$c_i^n = \frac{\rho_i^n}{N(\eta^n, \beta^n, \rho^n, \delta^n, \nu^n)}, \qquad d_i^n = \frac{\frac{\tau_i^n}{2} \int (\theta - \theta_i^\star)^2\, \nu_i^n(d\theta)}{N(\eta^n, \beta^n, \rho^n, \delta^n, \nu^n)}.$$

Note that

$$\sum_{i=1}^{q^\star} \{|a_i^n| + |b_i^n| + |c_i^n| + |d_i^n|\} = 1$$

for all $n$. Moreover, the intervals $A_i$ have compact closure. We may therefore extract a subsequence such that the following hold:

1. There exist constants $a_i, b_i \in \mathbb{R}$ and $c_i, d_i \in \mathbb{R}_+$ (for $i = 1, \ldots, q^\star$) such that $\sum_{i=1}^{q^\star} \{|a_i| + |b_i| + |c_i| + |d_i|\} = 1$, and we have $a_i^n \to a_i$, $b_i^n \to b_i$, $c_i^n \to c_i$, and $d_i^n \to d_i$ as $n \to \infty$ for all $i = 1, \ldots, q^\star$.
2. There exist probability measures $\lambda_i$, $i = 1, \ldots, q^\star$, such that $\lambda_i$ is supported on $\mathrm{cl}\, A_i$ for each $i$, and $\lambda_i^n$ converges weakly to $\lambda_i$ for all $i = 1, \ldots, q^\star$.

Then $\ell(\eta^n, \beta^n, \rho^n, \delta^n, \nu^n)/N(\eta^n, \beta^n, \rho^n, \delta^n, \nu^n)$ converges pointwise along this subsequence to the function $h/f^\star$ defined by

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

But as $\|\ell(\eta^n, \beta^n, \rho^n, \delta^n, \nu^n)\|_1/N(\eta^n, \beta^n, \rho^n, \delta^n, \nu^n) \to 0$, 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 5.3. *The Laplace transform $L[h](t)$ exists for all $t \in \mathbb{R}$:*

$$\frac{L[h](t)}{L[f](t)} = \sum_{i=1}^{q^\star} \left\{ a_i\, e^{\theta_i^\star t} + b_i\, t\, e^{\theta_i^\star t} + c_i\, t^2\, e^{\theta_i^\star t} + d_i\, t^2\, e^{\theta_i^\star t} \int \phi((\theta - \theta_i^\star)t)\, \lambda_i(d\theta) \right\}.$$

*Here we defined the positive increasing convex function $\phi(u) = 2(e^u - u - 1)/u^2$.*

PROOF. The $a_i, b_i, c_i$ terms are easily computed using Assumption A and integration by parts. It remains to compute the Laplace transform of the function

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

We begin by noting that, using Assumption A,

$$\int\int\int_0^1 e^{tx}\,|D_2 f_{\theta_i^\star+u(\theta-\theta_i^\star)}|\,2(1-u)\,du\,\lambda_i(d\theta)\,dx = \int e^{tx}\,|f''(x)|\,dx \times \int\left\{\int_0^1 e^{t(\theta_i^\star+u(\theta-\theta_i^\star))}\,2(1-u)\,du\right\}\lambda_i(d\theta) < \infty$$

for every $t \in \mathbb{R}$. We may therefore apply Fubini's theorem, giving

$$L[\Xi_i](t) = e^{\theta_i^\star t}\,L[f''](t)\int\left\{\int_0^1 e^{tu(\theta-\theta_i^\star)}\,2(1-u)\,du\right\}\lambda_i(d\theta) = L[f](t)\,t^2\,e^{\theta_i^\star t}\int\phi((\theta-\theta_i^\star)t)\,\lambda_i(d\theta),$$

where we have computed the inner integral using integration by parts. □

By this lemma, and as $L[f](t) > 0$ and $L[h](t) = 0$ for all $t \in \mathbb{R}$, we must have

$$\Phi(t) = \sum_{i=1}^{q^\star}\left\{a_i\,e^{\theta_i^\star t} + b_i\,t\,e^{\theta_i^\star t} + c_i\,t^2\,e^{\theta_i^\star t} + d_i\,t^2\,e^{\theta_i^\star t}\,\Phi_i(t)\right\} = 0 \tag{5.1}$$

for all $t \in \mathbb{R}$, where we have defined

$$\Phi_i(t) = \int\phi((\theta-\theta_i^\star)t)\,\lambda_i(d\theta).$$

In the remainder of the proof, we argue that (5.1) can not hold, thus completing the proof by contradiction. We distinguish between three different cases.

**Case 1.** Suppose that $\lambda_{q^\star}([\theta_{q^\star}^\star, T]) > 0$. As $\phi$ is positive and increasing, we can estimate $\Phi_{q^\star}(t) \geq \lambda_{q^\star}([\theta_{q^\star}^\star, T]) > 0$ for all $t \geq 0$. Now divide (5.1) by $t^2\,e^{\theta_{q^\star}^\star t}\,\Phi_{q^\star}(t)$, and let $t \to +\infty$. As $\lambda_i$ is supported in $[-T, \bar{\theta}_{q^\star}]$ for every $i < q^\star$,

$$\frac{t^2\,e^{\theta_i^\star t}\,\Phi_i(t)}{e^{\theta_{q^\star}^\star t}} \leq t^2\,e^{-(\theta_{q^\star}^\star-\theta_i^\star)t}\,\phi((\bar{\theta}_{q^\star}-\theta_i^\star)t) \xrightarrow{t\to+\infty} 0 \tag{5.2}$$

for every $i < q^\star$. It follows easily that for some constant $K \geq 0$

$$0 = \limsup_{t\to+\infty}\frac{\Phi(t)}{t^2\,e^{\theta_{q^\star}^\star t}\,\Phi_{q^\star}(t)} = d_{q^\star} + K\,c_{q^\star}.$$

As $c_{q^\star}, d_{q^\star} \geq 0$, this clearly implies that $d_{q^\star} = 0$. Using (5.2), we now obtain $\lim_{t\to+\infty}\Phi(t)/t^2\,e^{\theta_{q^\star}^\star t} = c_{q^\star} = 0$, which implies $\lim_{t\to+\infty}\Phi(t)/t\,e^{\theta_{q^\star}^\star t} = b_{q^\star} = 0$, and consequently $\lim_{t\to+\infty}\Phi(t)/e^{\theta_{q^\star}^\star t} = a_{q^\star} = 0$.

**Case 2.** Suppose that $\lambda_{q^\star}([\theta^\star_{q^\star}, T]) = 0$ and $\lambda_{q^\star}(]\bar\theta_{q^\star}, \theta^\star_{q^\star}[) > 0$. It is easily established, using the dominated convergence theorem and equation (5.2), that $\lim_{t\to+\infty} \Phi(t)/t^2\, e^{\theta^\star_{q^\star} t} = c_{q^\star}$, so that $c_{q^\star} = 0$. Next, dividing (5.1) by $e^{\theta^\star_{q^\star} t}$ and taking two derivatives with respect to $t$, we obtain for all $t \geq 0$

$$0 = \frac{d^2}{dt^2}\left(\frac{\Phi(t)}{e^{\theta^\star_{q^\star} t}}\right) = d_{q^\star} \int e^{(\theta-\theta^\star_{q^\star})t}\,\lambda_{q^\star}(d\theta) + \sum_{i=1}^{q^\star-1} \frac{d^2}{dt^2}\Bigg\{ a_i\, e^{-(\theta^\star_{q^\star}-\theta^\star_i)t} + b_i\, t\, e^{-(\theta^\star_{q^\star}-\theta^\star_i)t} + c_i\, t^2\, e^{-(\theta^\star_{q^\star}-\theta^\star_i)t} + d_i\, t^2\, e^{-(\theta^\star_{q^\star}-\theta^\star_i)t}\,\Phi_i(t)\Bigg\},$$

where the derivative and integral may be exchanged by [25], Appendix A16. Now note that as $\lambda_{q^\star}(]\bar\theta_{q^\star}, \theta^\star_{q^\star}[) > 0$, there is an $\varepsilon > 0$ such that $\lambda_{q^\star}([\bar\theta_{q^\star} + \varepsilon, \theta^\star_{q^\star}[) > 0$. Therefore, as $\phi$ is positive and increasing, we can estimate for all $t \geq 0$

$$\int e^{(\theta-\theta^\star_{q^\star})t}\,\lambda_{q^\star}(d\theta) \geq e^{(\bar\theta_{q^\star}+\varepsilon-\theta^\star_{q^\star})t}\,\lambda_{q^\star}([\bar\theta_{q^\star}+\varepsilon, \theta^\star_{q^\star}[) > 0.$$

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

$$\begin{aligned}
&e^{-(\bar\theta_{q^\star}+\varepsilon-\theta^\star_{q^\star})t}\left|\frac{d^2}{dt^2} t^2\, e^{-(\theta^\star_{q^\star}-\theta^\star_i)t}\,\Phi_i(t)\right| \\
&\quad= e^{-(\bar\theta_{q^\star}+\varepsilon-\theta^\star_{q^\star})t} \times e^{-(\theta^\star_{q^\star}-\theta^\star_i)t} \times \Bigg|(\theta^\star_{q^\star}-\theta^\star_i)^2 \int t^2\phi((\theta-\theta^\star_i)t)\,\lambda_i(d\theta) \\
&\qquad\qquad - 2(\theta^\star_{q^\star}-\theta^\star_i)\int \frac{e^{(\theta-\theta^\star_i)t}-1}{\theta-\theta^\star_i}\,\lambda_i(d\theta) + \int e^{(\theta-\theta^\star_i)t}\,\lambda_i(d\theta)\Bigg| \\
&\quad\leq e^{-(\bar\theta_{q^\star}+\varepsilon-\theta^\star_i)t}\Bigg\{(\theta^\star_{q^\star}-\theta^\star_i)^2\, t^2\,\phi((\bar\theta_{i+1}-\theta^\star_i)t) \\
&\qquad\qquad + 2\,(\theta^\star_{q^\star}-\theta^\star_i)\,\frac{e^{(\bar\theta_{i+1}-\theta^\star_i)t}-1}{\bar\theta_{i+1}-\theta^\star_i} + e^{(\bar\theta_{i+1}-\theta^\star_i)t}\Bigg\},
\end{aligned}$$

which converges to zero as $t \to +\infty$ for every $i < q^\star$. It follows easily that

$$0 = \frac{1}{\int e^{(\theta-\theta^\star_{q^\star})t}\,\lambda_{q^\star}(d\theta)}\,\frac{d^2}{dt^2}\left(\frac{\Phi(t)}{e^{\theta^\star_{q^\star} t}}\right) \xrightarrow{t\to+\infty} d_{q^\star}.$$

Finally, using (5.2), we obtain $\lim_{t\to+\infty} \Phi(t)/t\, e^{\theta^\star_{q^\star} t} = b_{q^\star} = 0$, and this subsequently gives $\lim_{t\to+\infty} \Phi(t)/e^{\theta^\star_{q^\star} t} = a_{q^\star} = 0$.

**Case 3.** Suppose that $\lambda_{q^\star}(]\bar{\theta}_{q^\star}, T]) = 0$. Then $\Phi_{q^\star}(t) = \phi((\bar{\theta}_{q^\star} - \theta^\star_{q^\star})t)$ for all $t$. Using equation (5.2), we now obtain $\lim_{t\to+\infty} \Phi(t)/t^2\, e^{\theta^\star_{q^\star} t} = c_{q^\star} = 0$, then $\lim_{t\to+\infty} \Phi(t)/t\, e^{\theta^\star_{q^\star} t} = b_{q^\star} = 0$, and $\lim_{t\to+\infty} \Phi(t)/e^{\theta^\star_{q^\star} t} = a_{q^\star} = 0$. Finally,

$$\frac{t^2\, e^{\theta^\star_i t}\, \Phi_i(t)}{t^2\, e^{\theta^\star_{q^\star} t}\, \phi((\bar{\theta}_{q^\star} - \theta^\star_{q^\star})t)} \xrightarrow{t\to+\infty} 0$$

for all $i < q^\star$ by (5.2) and as $t^2\, \phi((\bar{\theta}_{q^\star} - \theta^\star_{q^\star})t) \to +\infty$ as $t \to +\infty$. Therefore

$$0 = \limsup_{t\to+\infty} \frac{\Phi(t)}{t^2\, e^{\theta^\star_{q^\star} t}\, \phi((\bar{\theta}_{q^\star} - \theta^\star_{q^\star})t)} = d_{q^\star}.$$

**End of proof.** We have now shown that $a_{q^\star}, b_{q^\star}, c_{q^\star}, d_{q^\star} = 0$, and we are left with $q^\star - 1$ terms in equation (5.1). But proceeding by induction, we find that $a_i, b_i, c_i, d_i = 0$ for all $i$. This is impossible, as $\sum_{i=1}^{q^\star}\{|a_i| + |b_i| + |c_i| + |d_i|\} = 1$ by construction. Thus we have a contradiction, and the proof is complete. □

5.2. *Proof of Theorem 2.12.* The proof of Theorem 2.12 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 5.4. *Define $S = (H_0 + H_1 + H_2)/c^\star$. Then $S \in L^4(f^\star d\mu)$, and*

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

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

$$\frac{f - f^\star}{f^\star} = \sum_{i=1}^{q^\star} \left\{ \left( \sum_{j:\theta_j \in A_i} \pi_j - \pi^\star_i \right) \frac{f_{\theta^\star_i}}{f^\star} + \sum_{j:\theta_j \in A_i} \pi_j \left( \frac{f_{\theta_j} - f_{\theta^\star_i}}{f^\star} \right) \right\}.$$

Taylor expansion gives $f_{\theta_j}(x) - f_{\theta^\star_i}(x) = D_1 f_{\theta^\star_i}(x)\,(\theta_j - \theta^\star_i) + \frac{1}{2} D_2 f_{\theta_{ij}(x)}(x)\,(\theta_j - \theta^\star_i)^2$ for some $\theta_{ij}(x) \in [-T, T]$. Using Assumption A, we find that

$$\left|\frac{f - f^\star}{f^\star}\right| \le \sum_{i=1}^{q^\star} \left\{ \left| \sum_{j:\theta_j \in A_i} \pi_j - \pi^\star_i \right| + \left| \sum_{j:\theta_j \in A_i} \pi_j (\theta_j - \theta^\star_i) \right| + \frac{1}{2} \sum_{j:\theta_j \in A_i} \pi_j (\theta_j - \theta^\star_i)^2 \right\} (H_0 + H_1 + H_2).$$

On the other hand, Theorem 5.2 gives

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

The proof follows directly. □

COROLLARY 5.5. *$|d| \le D$ 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 5.4. □

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

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

*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 5.5

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

Elementary arguments show that for all $x \geq -1$

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

Therefore, by Lemma 5.4,

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

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

Finally, we need one further approximation step.

LEMMA 5.7. *Let $q \in \mathbb{N}$ and $\alpha > 0$. Then for every $f \in \mathcal{M}_q$ such that $h(f, f^\star) \leq \alpha$, there exist $\eta, \beta, \rho \in \mathbb{R}^{q^\star}$, $\gamma \in \mathbb{R}^q$, and $\theta \in [-T, T]^q$ such that*

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

*and*

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

*where we have defined*

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

PROOF. As $f \in \mathcal{M}_q$, we can write $f = \sum_{j=1}^q \pi_j f_{\theta_j}$. Note that by Theorem 5.2

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

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

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

Taylor expansion gives $f_{\theta_j}(x)-f_{\theta_i^\star}(x)=D_1f_{\theta_i^\star}(x)\,(\theta_j-\theta_i^\star)+\frac{1}{2}D_2f_{\theta_i^\star}(x)\,(\theta_j-\theta_i^\star)^2+\frac{1}{6}D_3f_{\theta_{ij}(x)}(x)\,(\theta_j-\theta_i^\star)^3$ for some $\theta_{ij}(x)\in[-T,T]$. Therefore

$$\frac{f-f^\star}{f^\star}=L+\frac{1}{6}\sum_{i=1}^{q^\star}\sum_{j\in J:\theta_j\in A_i}\pi_j(\theta_j-\theta_i^\star)^3\,\frac{D_3f_{\theta_{ij}(x)}(x)}{f^\star},$$

where we have defined

$$L=\sum_{i=1}^{q^\star}\left\{\left(\sum_{j\in J:\theta_j\in A_i}\pi_j-\pi_i^\star\right)\frac{f_{\theta_i^\star}}{f^\star}+\sum_{j\in J:\theta_j\in A_i}\pi_j(\theta_j-\theta_i^\star)\,\frac{D_1f_{\theta_i^\star}}{f^\star}\right.$$
$$\left.+\frac{1}{2}\sum_{j\in J:\theta_j\in A_i}\pi_j(\theta_j-\theta_i^\star)^2\,\frac{D_2f_{\theta_i^\star}}{f^\star}\right\}+\sum_{j\notin J}\pi_j\,\frac{f_{\theta_j}}{f^\star}.$$

Now note that

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

where we have used Lemma 5.4. By Theorem 5.2, we obtain

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

Therefore, we can estimate

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

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

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

It remains to show that for our choice of $\ell = L/\|L\|_2$, the vectors $\eta, \beta, \rho, \gamma$ in the statement of the lemma satisfy the desired bounds. To this end, note that

$$\|L\|_2 \geq c^\star \sum_{i=1}^{q^\star} \left\{ \left| \sum_{j:\theta_j \in A_i} \pi_j - \pi_i^\star \right| + \left| \sum_{j:\theta_j \in A_i} \pi_j(\theta_j - \theta_i^\star) \right| + \frac{1}{2} \sum_{j:\theta_j \in A_i} \pi_j (\theta_j - \theta_i^\star)^2 \right\}$$

by Theorem 5.2, while we have

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

It follows immediately that $\sum_{i=1}^{q^\star} |\rho_i| \leq 1/c^\star$. 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

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

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

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

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

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

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

We can now complete the proof of Theorem 2.12.

PROOF OF THEOREM 2.12. Let $\alpha > 0$ be a constant to be chosen later on, and

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

Then clearly

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

We will estimate each term separately.

**Step 1** (*the first term*). Define the family of functions

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

where

$$\mathfrak{I}_{q,\alpha} = \Bigg\{ (\eta,\beta,\rho,\gamma,\theta) \in \mathbb{R}^{q^\star} \times \mathbb{R}^{q^\star} \times \mathbb{R}^{q^\star} \times \mathbb{R}^q \times [-T,T]^q : \sum_{i=1}^{q^\star} \{|\eta_i| + |\beta_i| + |\rho_i|\} \le \frac{3}{c^\star} + \frac{1+2T}{\sqrt{c^\star \alpha}},\ \sum_{j=1}^{q} |\gamma_j| \le \frac{1}{\sqrt{c^\star \alpha}} \Bigg\}.$$

From Lemmas 5.6 and 5.7, 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 $h(f,f^\star) \le \sqrt{2}$ for any $f$)

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

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

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

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

$$\mathcal{N}(\mathcal{D}_{q,\alpha}, \varepsilon + 2\alpha^{1/4}\|U\|_2) \le \mathcal{N}(\mathcal{L}_{q,\alpha}, \varepsilon) \quad \text{for } \varepsilon > 0.$$

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

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

$$|\ell - \ell'| \le (H_0 + H_1 + H_2) \sum_{i=1}^{q^\star} \{|\eta_i - \eta_i'| + |\beta_i - \beta_i'| + |\rho_i - \rho_i'|\} + H_0 \sum_{j=1}^{q} |\gamma_j - \gamma_j'| + \frac{H_1}{\sqrt{c^\star \alpha}} \max_{j=1,\dots,q} |\theta_j - \theta_j'|,$$

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

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

where $|||\cdot|||_{q,\alpha}$ is the Banach space norm on $\mathbb{R}^{3q^\star+2q}$ defined by

$$|||(\eta,\beta,\rho,\gamma,\theta)|||_{q,\alpha} = \sum_{i=1}^{q^\star}\{|\eta_i| + |\beta_i| + |\rho_i|\} + \sum_{j=1}^{q}|\gamma_j| + \frac{1}{\sqrt{c^\star\alpha}}\max_{j=1,\dots,q}|\theta_j|.$$

Note that if $|||(\eta,\beta,\rho,\gamma,\theta) - (\eta',\beta',\rho',\gamma',\theta')|||_{q,\alpha} \le \varepsilon'$, 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 $N_0(\mathfrak{I}_{q,\alpha}, |||\cdot|||_{q,\alpha}, \varepsilon')$ the cardinality of the smallest proper cover of $\mathfrak{I}_{q,\alpha}$ by $|||\cdot|||_{q,\alpha}$-balls or radius $\varepsilon'$ (the cover is called *proper* if each ball is centered at some point inside $\mathfrak{I}_{q,\alpha}$), then we have shown that

$$\mathcal{N}(\mathcal{L}_{q,\alpha}, \varepsilon) \le N_0(\mathfrak{I}_{q,\alpha}, |||\cdot|||_{q,\alpha}, \varepsilon/2\|V\|_2) \quad \text{for } \varepsilon > 0.$$

But note that $\mathfrak{I}_{q,\alpha}$ is included in the $|||\cdot|||_{q,\alpha}$-ball

$$\mathfrak{I}_{q,\alpha} \subset \mathfrak{B}_{q,\alpha} = \left\{ v \in \mathbb{R}^{3q^\star+2q} : |||v|||_{q,\alpha} \le \frac{3}{c^\star} + \frac{2+3T}{\sqrt{c^\star\alpha}} \right\}.$$

We now use the following standard facts:

1. $N_0(S,d,\varepsilon) \le N(S,d,\varepsilon/2)$ for any subset $S \subset B$ of a metric space $(B,d)$, where $N(S,d,\varepsilon)$ is the cardinality of the smallest (not necessarily proper) cover of $S$ consisting of $d$-balls of radius $\varepsilon$.
2. For any $n$-dimensional Banach space $(B, |||\cdot|||)$, the covering number of the $r$-ball $B(r) = \{x \in B : |||x||| \le r\}$ satisfies $N(B(r), |||\cdot|||, \varepsilon) \le (\frac{2r+\varepsilon}{\varepsilon})^n$.

Using these facts, we can estimate

$$\begin{aligned} N_0(\mathfrak{I}_{q,\alpha}, |||\cdot|||_{q,\alpha}, \varepsilon/2\|V\|_2) &\le N_0(\mathfrak{B}_{q,\alpha}, |||\cdot|||_{q,\alpha}, \varepsilon/2\|V\|_2) \\ &\le N(\mathfrak{B}_{q,\alpha}, |||\cdot|||_{q,\alpha}, \varepsilon/4\|V\|_2) \le \left(\frac{8\|V\|_2(\frac{3}{c^\star} + \frac{2+3T}{\sqrt{c^\star}}) + 1}{\varepsilon\sqrt{\alpha}}\right)^{3q^\star+2q} \end{aligned}$$

for $\varepsilon/4\|V\|_2 \le 1$, $\alpha \le 1$. Choosing $\varepsilon + 2\alpha^{1/4}\|U\|_2 = \delta$, we obtain

$$\mathcal{N}(\mathcal{D}_{q,\alpha}, \delta) \le \mathcal{N}(\mathcal{L}_{q,\alpha}, \delta - 2\alpha^{1/4}\|U\|_2) \le \left(\frac{8\|V\|_2(\frac{3}{c^\star} + \frac{2+3T}{\sqrt{c^\star}}) + 1}{(\delta - 2\alpha^{1/4}\|U\|_2)\sqrt{\alpha}}\right)^{3q^\star+2q}$$

for $\delta \le 4\|V\|_2$ and $\alpha \le (\delta/2\|U\|_2)^4 \wedge 1$.

**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

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

Defining

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

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 $|||\cdot|||_q$ is a Banach space norm on $\mathbb{R}^{2q}$). 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

$$\mathcal{N}(\mathcal{D}_q \backslash \mathcal{D}_{q,\alpha}, \delta) \le N_0(\Delta_q \times [-T,T]^q, |||\cdot|||_q, \alpha^4\delta^2/4\|W\|_2^2),$$

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

$$\mathcal{N}(\mathcal{D}_q \backslash \mathcal{D}_{q,\alpha}, \delta) \le \left(\frac{(24+16T)\|W\|_2^2}{\alpha^4\delta^2}\right)^{2q}$$

for $\delta \leq 8\|W\|_2$ and $\alpha \leq 1$.

**End of proof.** Choose $\alpha = (\delta/4\|U\|_2)^4$. Collecting the various estimates above, we find that we have for $\delta \leq \min(4\|U\|_2, 4\|V\|_2, 8\|W\|_2)$

$$\mathcal{N}(\mathcal{D}_q, \delta) \leq \left( \frac{16^2\|U\|_2^2\|V\|_2(\frac{3}{c^\star} + \frac{2+3T}{\sqrt{c^\star}}) + 32\|U\|_2^2}{\delta^3} \right)^{3q^\star+2q} + \left( \frac{4^{16}(24+16T)\|U\|_2^{16}\|W\|_2^2}{\delta^{18}} \right)^{2q}.$$

Using that $q > q^\star$, it now follows easily that there exist constants $C^\star$ and $\delta^\star$, depending only on $\|U\|_2$, $\|V\|_2$, $\|W\|_2$, $T$, and $c^\star$, such that

$$\mathcal{N}(\mathcal{D}_q, \delta) \leq \left( \frac{C^\star}{\delta} \right)^{36q} \quad \text{for all } \delta \leq \delta^\star.$$

This establishes the estimate given in the statement of the Theorem. To complete the proof, it remains to note that the existence of $D$ follows from Corollary 5.5. □

5.3. *Proof of Corollary 2.14.* The first condition of Theorem 2.4 follows directly from Theorem 2.12. To establish the second condition, note that for any $f, f' \in \mathcal{M}_q$, $f = \sum_{i=1}^q \pi_i f_{\theta_i}$, $f' = \sum_{i=1}^q \pi'_i f_{\theta'_i}$, we can estimate

$$\left| \log\left(\frac{f+f^\star}{2f^\star}\right) - \log\left(\frac{f'+f^\star}{2f^\star}\right) \right| \leq \left| \frac{f-f'}{f^\star} \right| \leq (H_0 + H_1) \left( \sum_{i=1}^q |\pi_i - \pi'_i| + \max_{i=1,\ldots,q} |\theta_i - \theta'_i| \right).$$

It follows as above that $\mathcal{N}(\{\log(\{f+f^\star\}/2f^\star) : f \in \mathcal{M}_q\}, \delta) < \infty$ for all $\delta > 0$, which implies that $\{\log(\{f+f^\star\}/2f^\star) : f \in \mathcal{M}_q\}$ is $\mathbf{P}^\star$-Glivenko-Cantelli.

Now define the functions $\eta(q)$ and $\varpi(n)$ as

$$\eta(q) = q \left[ 6 \int_0^{\|D\|_2} \sqrt{\log\left(\frac{C^\star}{\delta \wedge \delta^\star}\right)}\, du \right]^2,$$

$$\varpi(n) = \omega(n) \left[ 6 \int_0^{\|D\|_2} \sqrt{\log\left(\frac{C^\star}{\delta \wedge \delta^\star}\right)}\, du \right]^{-2}.$$

Then all the assumptions of Theorem 2.4 are satisfied, yielding the consistency of the penalized likelihood mixture order estimator. □

5.4. *Proof of Proposition 2.15.* We begin by characterizing $\bar{\mathcal{D}}_{q^\star}$.

LEMMA 5.8. *Suppose that Assumption A holds. Then we have*

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

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 5.2, 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

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

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 5.6 and 5.7, we can estimate

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

But using Theorem 5.2, we find that for $n$ sufficiently large

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

Thus we have

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

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 $\eta, \beta \in \mathbb{R}^{q^\star}$ 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 5.6, we obtain

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

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

REMARK 5.9. The proof of Lemma 5.8 in fact shows that $\bar{\mathcal{D}}_{q^\star} = \bar{\mathcal{D}}^c_{q^\star}$.

We can now complete the proof of Proposition 2.15.

PROOF OF PROPOSITION 2.15. We apply Theorem 2.10 with $q = q^\star + 1$. The requisite envelope and bracketing assumptions follow directly from Theorem 2.12. It therefore remains to show that $\bar{\mathcal{D}}^c_{q^\star+1} \backslash \bar{\mathcal{D}}_{q^\star}$ is nonempty.

Consider the function $f_t$ defined for $t > 0$ as follows:

$$f_t = \frac{\pi_1^\star}{2}\left(f_{\theta_1^\star + t} + f_{\theta_1^\star - t}\right) + \sum_{i=2}^{q^\star} \pi_i^\star f_{\theta_i^\star}.$$

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

$$\frac{f_t - f^\star}{t^2} = \frac{\pi_1^\star}{2} \frac{f_{\theta_1^\star + t} - 2\, f_{\theta_1^\star} + f_{\theta_1^\star - t}}{t^2} \xrightarrow{t\to 0} \frac{\pi_1^\star}{2} D_2 f_{\theta_1^\star}.$$

As in the proof of Lemma 5.8, we find that

$$\lim_{t\to 0} d_{f_t} = \lim_{t\to 0} \frac{(f_t - f^\star)/t^2 f^\star}{\|(f_t - f^\star)/t^2 f^\star\|_2} = \frac{D_2 f_{\theta_1^\star}}{\|D_2 f_{\theta_1^\star}\|_2} = d_0.$$

By construction, $d_0 \in \bar{\mathcal{D}}^c_{q^\star+1}$. But by Theorem 5.2, the functions $f_{\theta_i^\star}$, $D_1 f_{\theta_i^\star}$, and $D_2 f_{\theta_i^\star}$ ($i = 1, \ldots, q^\star$) are all linearly independent. Together with Lemma 5.8, 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. □

## APPENDIX A: EMPIRICAL PROCESS INEQUALITIES

**A.1. Two inequalities for the empirical process.** In the proof of Theorem 2.4 we will need two maximal inequalities for the empirical process. These inequalities follow rather easily from standard results in empirical process theory.

First, we need the following deviation inequality for the supremum of an empirical process, similar to [23], Theorem 5.11. A short proof is in section A.2.

PROPOSITION A.1. *Let $\mathcal{Q}$ be a family of measurable functions $f : E \to \mathbb{R}$. Assume that for some constants $R, K > 0$.*

$$\sup_{f\in\mathcal{Q}} \|f\|_\infty \le K, \qquad \sup_{f\in\mathcal{Q}} \mathbf{E}^\star[f(X_1)^2] \le R^2.$$

*Then we have*

$$\mathbf{P}^\star\left[\sup_{f\in\mathcal{Q}} |\nu_n(f)| \ge \alpha\right] \le 2\,\exp\left[-\frac{\alpha^2}{C^2(C_1+1)R^2}\right]$$

*for all $n \in \mathbb{N}$, $\alpha > 0$, and $C_1 > 0$ such that*

$$C\sqrt{C_1+1}\int_0^R \sqrt{\log \mathcal{N}(\mathcal{Q}, u)}\, du \le \alpha \le \frac{C_1 R^2\sqrt{n}}{K},$$

*where $C$ is a universal constant (the choice $C = 37.5$ works).*

We also need the following variant of Etemadi's inequality.

PROPOSITION A.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 $S_n(f) = \sqrt{n}\,\nu_n(f)$.*

The proof is given in section A.3 below.

**A.2. Proof of Proposition A.1.** The following Bernstein-type deviation inequality can be read off from [18], Corollary 6.9, together with the standard fact that the essential supremum of a family of random variables coincides with the essential supremum of a countable subfamily (cf. Remark 2.1).

THEOREM A.3. *Let $\mathcal{Q}$ be a family of measurable functions $f : E \to \mathbb{R}$. Assume that for some constants $R, K > 0$*

$$\sup_{f\in\mathcal{Q}} \|f\|_\infty \le K, \qquad \sup_{f\in\mathcal{Q}} \mathbf{E}^\star[f(X_1)^2] \le R^2.$$

*Then we have*

$$\mathbf{P}^\star\left[\sup_{f\in\mathcal{Q}} |\nu_n(f)| \ge E(n) + 7R\sqrt{2x} + \frac{4Kx}{3\sqrt{n}}\right] \le 2\,e^{-x}$$

*for every $n \in \mathbb{N}$ and $x > 0$, where we have defined*

$$E(n) = 27\int_0^R \sqrt{\log \mathcal{N}(\mathcal{Q}, u)}\,du + \frac{10K}{3\sqrt{n}}\,\log \mathcal{N}(\mathcal{Q}, R).$$

*for every $n \in \mathbb{N}$.*

The proof of Proposition A.1 reduces easily to this result.

PROOF OF PROPOSITION A.1. Let $\alpha = \sqrt{C^2(C_1+1)R^2x}$ and assume the given bounds on $\alpha$ hold. Then we can estimate

$$x = \frac{\alpha^2}{C^2(C_1+1)R^2} \le \frac{C_1R^2\sqrt{n}}{K} \times \frac{\alpha}{C^2(C_1+1)R^2} \le \frac{\alpha\sqrt{n}}{C^2K},$$

as well as

$$\alpha^2 \le \frac{C_1R^2\alpha\sqrt{n}}{K}.$$

On the other hand, as $\mathcal{N}(\mathcal{Q}, u)$ is nonincreasing in $u$, we have

$$CR\sqrt{C_1+1}\sqrt{\log \mathcal{N}(\mathcal{Q}, R)} \le C\sqrt{C_1+1}\int_0^R \sqrt{\log \mathcal{N}(\mathcal{Q}, u)}\,du \le \alpha.$$

We can therefore estimate

$$\begin{aligned} E(n) + 7R\sqrt{2x} + \frac{4Kx}{3\sqrt{n}} &\le \left\{\frac{27+7\sqrt{2}}{C\sqrt{C_1+1}} + \frac{10C_1}{3C^2(C_1+1)} + \frac{4}{3C^2}\right\}\alpha \\ &\le \left\{\frac{37}{C} + \frac{5}{C^2}\right\}\alpha \le \alpha \end{aligned}$$

provided that we choose $C$ such that $37/C + 5/C^2 \le 1$ (e.g., $C = 37.5$). The proof is completed by applying Theorem A.3. □

**A.3. Proof of Proposition A.2.** The proof of Proposition A.2 follows closely the proof of the classical Etemadi inequality, see [2], Appendix M19.

PROOF OF PROPOSITION A.2. Define the stopping time

$$\tau = \inf\left\{k \geq m : \sup_{f\in\mathcal{Q}} |S_k(f)| \geq 3\alpha\right\}.$$

Then we have

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

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

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

Therefore, we can estimate

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

As $\sup_{f\in\mathcal{Q}}|S_n(f)-S_k(f)|$ and $\{\tau=k\}$ are independent, we obtain

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

The proof is easily completed. □

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

[3] 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.

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

[5] 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.

[6] 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.

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

[8] DACUNHA-CASTELLE, D. AND GASSIAT, E. (1999). Testing the order of a model using locally conic parametrization: population mixtures and stationary ARMA processes. *Ann. Statist. 27*, 1178–1209.

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

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

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

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

[13] GHOSAL, S. AND VAN DER VAART, A. W. (2001). Entropies and rates of convergence for maximum likelihood and Bayes estimation for mixtures of normal densities. *Ann. Statist. 29*, 1233–1263.

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

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

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

[18] MASSART, P. (2007). *Concentration inequalities and model selection*. Lecture Notes in Mathematics, Vol. **1896**. Springer, Berlin. Lectures from the 33rd Summer School on Probability Theory held in Saint-Flour, July 6–23, 2003, With a foreword by Jean Picard.

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

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

[21] SHIRYAEV, A. N. (1996). *Probability*, Second ed. Graduate Texts in Mathematics, Vol. **95**. Springer-Verlag, New York.

[22] TEICHER, H. (1963). Identifiability of finite mixtures. *Ann. Math. Statist. 34*, 1265–1269.

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

[24] VAN HANDEL, R. (2009). On the minimal penalty for Markov order estimation. Preprint, available at http://arxiv.org/abs/0908.3666.

[25] 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

DEPARTMENT OF OPERATIONS RESEARCH
AND FINANCIAL ENGINEERING,
PRINCETON UNIVERSITY,
PRINCETON, NJ 08544, USA.
E-MAIL: rvan@princeton.edu