Strong identifiability and optimal minimax rates for finite mixture estimation

Philippe Heinrich, Jonas Kahn

Submitted to the Annals of Statistics

STRONG IDENTIFIABILITY AND OPTIMAL MINIMAX RATES FOR FINITE MIXTURE ESTIMATION

By Philippe Heinrich and Jonas Kahn

Université Lille 1
Laboratoire Paul Painlevé Bât. M2
Cité Scientifique
59655 Villeneuve d'Ascq, FRANCE

Abstract We study the rates of estimation of finite mixing distributions, that is, the parameters of the mixture. We prove that under some regularity and strong identifiability conditions, around a given mixing distribution with m_0 components, the optimal local minimax rate of estimation of a mixing distribution with m components is $n^{-1/(4(m-m_0)+2)}$. This corrects a previous paper by Chen (1995) in The Annals of Statistics.

By contrast, it turns out that there are estimators with a (non-uniform) pointwise rate of estimation of $n^{-1/2}$ for all mixing distributions with a finite number of components.

1. Introduction. Finite mixture models have been applied since Pearson (1894) in various fields including astronomy, biology, genetics, economy, social sciences and engineering (McLachlan and Peel, 2000).

Finite mixtures and their estimation naturally arise mostly in three cases. One is model-based clustering. Here the aim is to divide the data into k clusters and assign (new) data to a cluster. A possible approach is to consider that data point from each cluster is generated according to a probability distribution, so that the whole data is generated by mixture with k components (McLachlan and Peel, 2000; Teh, 2010).

The second, more traditional case, is the statistical description of possibly heterogeneous data where the underlying mixing distribution has no particular meaning. In that case, mixtures are a tool to describe efficiently the "true" probability distribution and control the convergence rate of mixture estimators to it (van de Geer, 1996; Ghosal and van der Vaart, 2001; Genovese and Wasserman, 2000).

MSC 2010 subject classifications: Primary 62G05; secondary 62G20.

Keywords and phrases: Local asymptotic normality, convergence of experiments, maximum likelihood estimate, Wasserstein metric, mixing distribution, mixture model, rate of convergence, strong identifiability, pointwise rate, superefficiency.

In the third case, the goal is the mixing distribution itself: its support points and proportions are the parameters we want to estimate. They typically correspond to the phenomenon that is studied. This is the case we are interested in, on the basis of observations drawn from the mixture.

Some works try to bridge the gap between the estimation of the mixture and the one of the mixing distribution, usually at least through estimation of the number of components – the order – in the finite mixture. In particular, Rousseau and Mengersen (2011) have proved that their Bayesian estimator of an overfitted mixture tends to empty the extra components, and Gassiat and van Handel (2013) have given the minimal penalty on the maximum-likelihood estimator of the order that yields strong consistency.

One could expect that a good estimator for the mixture would be a good estimator for the mixing model. However, this is not so clear. The situation is reminiscent of the difference between estimation and identification in model selection, where Yang (2005) has proved that no procedure can be optimal for both. Moreover, rates of convergence can be very different, as illustrated in an infinite-dimensional case by Bontemps and Gadat (2014).

Optimal rates are a key information in estimating the mixture parameters. These were unknown (see e.g. Titterington, Smith and Makov, 1985) till the work of Chen (1995), who established a $n^{-1/4}$ local minimax rate, under reasonable identifiability conditions, for one-dimensional-parameter mixtures. This result is somewhat surprising since the rate does not depend on the number of components. It turns out to be erroneous, because of its Lemma 2.

Our article aims at giving correct statements and proofs and its consequences.

The main part consists in finding the correct exponent in the local minimax rate; we do so in Theorem 3.2 and Theorem 3.3. The rate gets worse with more components, which is consistent with the behaviour when there are infinitely many components, such as deconvolution: Fan (1991) had proved that the L^2-convergence rate was polylogarithmic in general, and Caillerie et al. (2013) and Dedecker and Michel (2013) have generalized this kind of rates to the more relevant (for us) Wasserstein metrics. The most original technical tool we shall use is a coarse-graining tree on the parameter indices.

In addition, the optimal local minimax rate and the optimal pointwise rate of estimation everywhere are not the same. This discrepancy is unusual in statistics, and probably the reason why the $n^{-1/4}$ rate went unchallenged for twenty years. Specifically, if instead of comparing all pairs of mixtures in a ball, we allow only one mixture in it, we get (21) which corrects Lemma 2

of Chen. As a consequence, Theorem 2 of Chen is valid by dropping uniformity: for any fixed mixing distribution say G, the estimator considered there will converge at rate $n^{-1/4}$, but with a multiplicative constant that depends on G. It then becomes a statement on the optimal pointwise rate of estimation everywhere, and can even be strengthened to $n^{-1/2}$ as we show in Theorem 4.1.

The paper by Chen (1995) has been widely cited and used. Apart from applied papers citing it that may have relied on the theoretical guarantees (see e.g. Kuhn et al., 2014; Liu and Hancock, 2014), there are essentially two ways it could play a role. Firstly, when it is used as part of a proof, secondly when it is used as a benchmark.

The first case covers papers that generalize Chen's result in other settings, and re-use its theorems and proofs. For example, Ishwaran, James and Sun (2001) propose a Bayesian estimator that achieves the $n^{-1/4}$ frequentist rate, and use Chen (1995, Lemma 2) in their analysis. More recently, Nguyen (2013) generalizes those results to mixtures with an abstract parameter space and indefinite number of components. But Nguyen (2013, Theorem 1) generalizes Chen (1995, Lemma 2) while transposing the proof with the mistake. The main results of both these articles hold however: they do not need the full strength of Chen (1995, Lemma 2), but merely the weaker version (21). Ho and Nguyen (2016) prove such a sufficient version.

These two papers also use Chen's (1995) article as a benchmark. However, the optimal pointwise rate everywhere would probably be a better reference point in their case, as in many others. In particular, it seems likely that a Bayesian estimator could converge pointwise at speed $n^{-1/2}$ everywhere. We have not checked whether the proof by Ishwaran, James and Sun (2001) can be improved, or if another prior is necessary.

This use as a benchmark is very usual, as expected for this kind of optimality result (see e.g. Zhu and Zhang, 2006, 2004). Let us point in particular to a result by Martin (2012). He achieves almost $n^{-1/2}$ rate for the predictive recursion algorithm, and tries to explain the discrepancy with Chen (1995) by the fact that the parameters are constrained to live in a finite space for his algorithm. In fact, since his rate is pointwise, it fits with the continuous case.

Since early versions of this article have been made available, there have been interesting new developments: Ho et al. (2016); Ho and Nguyen (2015) have made explicit the system of equations underlying the minimax argument, for any finite number of parameters, and solved important special cases. The strong identifiability conditions we shall use in this article ensure

that the system of equations is generic.

In Section 2, we give the notations and define and discuss the regularity assumptions we use. In Section 3, we state and discuss the main theorem, giving the optimal local minimax rate. In Section 4, pointwise rate everywhere is investigated. We try to give some intuition in both of these sections. In Section 5, we also dwell on the interpretation and practical consequences of having different rates, and conclude with open questions. In Section 6, we give and explain the meaning of the key intermediate results and prove the main theorems from here. In Section 7, we prove those key intermediate results. In particular, we introduce the most original tool of our proofs: the coarse-graining tree that allows to patch the mistake in the article by Chen (1995).

Some auxiliary and technical results are detailed in appendices grouped in a supplemental part (Heinrich and Kahn, 2015).

2. Notations and regularity conditions.

2.1. *Basic notations.* Throughout the paper, the parameter set Θ, of diameter $\operatorname{Diam}\Theta$, is always assumed to be a compact subset of $\mathbb{R}$ with nonempty interior. Let $\mathcal{G}_m$ (resp. $\mathcal{G}_{\leqslant m}$) be the set of (resp. at most) m-mixing or m-point support distributions G on Θ. We set also $\mathcal{G}_{<\infty} = \cup_{\ell \geqslant 1} \mathcal{G}_\ell$.

As usual, we compare two mixing distributions G and G' using transportation distances, or L^q-Wasserstein metrics, with $q \geq 1$. They completely bypass identifiability issues that would arise with the square error on parameters. The definition is:

$$W_q(G, G') = \inf_{\Pi} \left[\int_{\Theta^2} |\theta - \theta'|^q \mathrm{d}\Pi(\theta, \theta') \right]^{1/q} \tag{1}$$

where the infimum is taken over probability measures Π on $\Theta \times \Theta$ with marginals G and G'. By Jensen's inequality, $W_q \geq W_{q'}$ if $q > q'$ and moreover, $W_q^q \leq W_{q'}^{q'} (\operatorname{Diam}\Theta)^{q-q'}$. We will usually work with the strongest available Wasserstein metric for our results. Endowed with the metric W_q, the space $\mathcal{G}_{\leqslant m}$ is compact.

In the special case of W_1, we will also use its dual representation, where $|f|_{\mathrm{Lip}}$ stands for the Lipschitz seminorm of f (e.g. Dudley, 2002, section 11.8):

$$W_1(G, G') = \sup_{|f|_{\mathrm{Lip}} \leqslant 1} \int_\Theta f(\theta) \mathrm{d}(G - G')(\theta). \tag{2}$$

Given $G = \sum_{j=1}^m \pi_j \delta_{\theta_j} \in \mathcal{G}_m$ and a family $\{f(x,\theta)\}_{\theta\in\Theta}$ of probability densities on $\mathbb{R}$ w.r.t. some σ-finite measure λ, we can define a *finite mixture with m components*:

$$f(x,G) = \int_\Theta f(x,\theta)\mathrm{d}G(\theta) = \sum_{j=1}^m \pi_j f(x,\theta_j). \tag{3}$$

To compare mixture distribution functions $F(x,G)$ and $F(x,G')$, we will use the Kolmogorov metric $\|F(\cdot,G) - F(\cdot,G')\|_\infty$. Here of course we have by definition $F(x,\theta) = \int_{-\infty}^x f(y,\theta)\mathrm{d}\lambda(y)$, which extends to $F(x,G)$ by linearity provided that $f(y,\delta_\theta) = f(y,\theta)$ where δ_θ denotes the Dirac measure at θ.

2.2. *Regularity conditions.*

(p,α)-smoothness. Hereafter, $f^{(p)}(x,\theta)$ or $f^{(p)}(\cdot,\theta)$ denote the p-th derivative of f always taken w.r.t. the variable θ.

DEFINITION 2.1. *The family $\{f(\cdot,\theta), \theta\in\Theta\}$ w.r.t. some σ-finite measure λ is* (p,α)-smooth *if*

$$E_{p,\alpha}\left(\theta,\theta',\theta''\right) = \int_{\mathbb{R}} \left|\frac{f^{(p)}(x,\theta')}{f(x,\theta'')}\right|^\alpha f(x,\theta)\,\mathrm{d}\lambda(x) \tag{4}$$

is a well-defined $[0,\infty]$-valued continuous function on Θ^3, and if there exists $\varepsilon > 0$ such that

$$|\theta' - \theta''| < \varepsilon \implies \forall\theta\in\Theta, \quad E_{p,\alpha}(\theta,\theta',\theta'') < \infty. \tag{5}$$

These smoothness conditions are easy to check in practice, and general enough. For example, all exponential families satisfy them, as shown in the supplemental part (Heinrich and Kahn, 2015, E.2). They will be useful for proving local asymptotic normality (Le Cam, 1986) of relevant families.

k-strong identifiability ($k\in\mathbb{N}$). Chen (1995) introduced a notion of strong identifiability. We will need a slightly more general version.

DEFINITION 2.2. *The family $\{F(\cdot,\theta), \theta\in\Theta\}$ of distribution functions is* k-strongly identifiable *if for any finite set of say m distinct points $\theta_j\in\Theta$,*

$$\left\|\sum_{p=0}^k\sum_{j=1}^m a_{p,j}F^{(p)}(\cdot,\theta_j)\right\|_\infty = 0 \implies \|a\| = \max_{p,j}|a_{p,j}| = 0,$$

where $\|\cdot\|_\infty$ denotes the supremum norm with respect to the variable x.

Chen's strong identifiability corresponds to 2-strong identifiability. Let us show why this notion is useful. Consider $G_n = \frac{1}{2}(\delta_{n^{-1}} + \delta_{-n^{-1}})$ in $\mathcal{G}_2$. We see that $F(\cdot, G_n) = F(\cdot, 0) + n^{-2}F^{(2)}(\cdot, 0)/2 + o(n^{-2})$, provided we can expand around $\theta = 0$. Then 2-strong identifiability ensures that $\|F(\cdot, 0) - F(\cdot, G_n)\|_\infty$ is of order n^{-2}, as shown in Proposition 2.3 below, whereas simple (1-strong) identifiability would say nothing. We will need k-strong identifiability with a higher k if more terms cancel.

PROPOSITION 2.3. *Let $\{F(\cdot, \theta), \theta \in \Theta\}$ be k-strongly identifiable family of distribution functions with $F^{(k)}(x, \theta)$ continuous in θ. Set for $\varepsilon > 0$*

$$\Theta_\varepsilon^m = \left\{ (\theta_j)_{1 \leqslant j \leqslant m} \subset \Theta : \min_{j \neq j'} |\theta_j - \theta_{j'}| \geqslant \varepsilon \right\}.$$

Then for all $a = (a_{p,j})_{\substack{0 \leqslant p \leqslant k \\ 1 \leqslant j \leqslant m}}$,

$$\inf_{\Theta_\varepsilon^m} \left\| \sum_{p=0}^{k} \sum_{j=1}^{m} a_{p,j} F^{(p)}(\cdot, \theta_j) \right\|_\infty \underset{\varepsilon,k,m}{\succcurlyeq} \|a\|,$$

where $\underset{\varepsilon,k,m}{\succcurlyeq}$ means "more than", up to some constant $C(\varepsilon, k, m) > 0$.

PROOF. The function $(a, (\theta_j)_{1 \leqslant j \leqslant m}) \mapsto \left\| \sum_{p=0}^{k} \sum_{j=1}^{m} a_{p,j} F^{(p)}(\cdot, \theta_j) \right\|_\infty$ is lower semi-continuous on the compact set $\{a : \|a\| = 1\} \times \Theta_\varepsilon^m$, so that it admits a minimum. By k-strong identifiability, it is nonzero. □

We expect the strong identifiability to be rather generic, and hence the statements of this paper often meaningful. In particular, Chen (1995, Theorem 3) has proved that location and scale families with smooth densities are 2-strongly identifiable. The theorem and the proof straightforwardly generalize to our setting. We merely state the result.

THEOREM 2.4. *Let $k \geqslant 1$. Let f be a probability density w.r.t. the Lebesgue measure on $\mathbb{R}$. Assume that f is $k-1$ times differentiable with*

$$\lim_{x \to \pm\infty} f^{(p)}(x) = 0 \text{ for } p \in [\![0, k-1]\!].$$

Consider $f(x, \theta) = f(x - \theta)$, with $\theta \in \Theta \subset \mathbb{R}$. Then the corresponding distributions family $\{F(\cdot, \theta), \theta \in \Theta\}$ is k-strongly identifiable. If $\Theta \subset (0, \infty)$, the result stays true with $f(x, \theta) = \frac{1}{\theta} f\left(\frac{x}{\theta}\right)$.

For more general conditions see the article by Holzmann, Munk and Stratmann (2004), that also generalize well to k-strong identifiability.

3. Assumptions and main results on local asymptotic minimax rate. The statistical estimation will always be done in the model $\mathcal{G}_{\leqslant \mathbf{m}}$ and, for local statements, around a fixed mixture $G_0 \in \mathcal{G}_{m_0}$ with $m_0 \leqslant \mathbf{m}$. Set once and for all

$$d_0 = \mathbf{m} - m_0. \tag{6}$$

Lower bounds on local asymptotic minimax rates.

ASSUMPTION $\mathrm{A}(k, \theta_0)$. *For all $(p, \alpha) \in [\![1, 2k+2]\!] \times [\![1, 4]\!]$, the family of densities $\{f(\cdot, \theta)\}_{\theta \in \Theta}$ is (p, α)-smooth and satisfies, for some point θ_0 in the interior of Θ,*

$$\int |f^{(2k+1)}(\cdot, \theta_0)| \mathrm{d}\lambda > 0.$$

Typically, k will be d_0 and θ_0 a support point of G_0. These conditions allow to prove local asymptotic normality (Le Cam, 1986) for relevant families. This will give some insight on the reason why the lower bound on the rate holds, and on how the mixtures behave when we change the parameters in the least sensitive direction. The condition on the support point guarantees identifiability locally for the families, and we need more derivatives than usual, since there will be cancellations in the first terms.

REMARK 3.1. *When comparing sequences we will write $a_n \preccurlyeq b_n$ or $a_n = O(b_n)$ for $a_n \leqslant C b_n$ where $C > 0$ does not depend on n. We will furthermore use $a_n \asymp b_n$ for $b_n \preccurlyeq a_n \preccurlyeq b_n$. If needed, the dependence of C on other parameters, say u, θ will be stressed by subscripts: $a_n \underset{u,\theta}{\preccurlyeq} b_n$ or $a_n \underset{u,\theta}{\asymp} b_n$.*

In what follows, $\mathbb{E}_G$ will denote the expectation w.r.t. the measure $\mathrm{d}\mathbb{P}_G = f(\cdot, G)\mathrm{d}\lambda$. And all the mixing distribution estimators denoted by $\widehat{G}_n$ below will be based on i.i.d. n-samples.

THEOREM 3.2. *Recall (6) and set $\varepsilon_n = n^{-1/(4d_0+2)+\kappa}$ for some $\kappa > 0$. Let θ_0 be a support point of G_0. Under Assumption $\mathrm{A}(d_0, \theta_0)$, for any sequence of estimators $\widehat{G}_n$, we have*

$$\sup_{\substack{G \in \mathcal{G}_{\mathbf{m}} \\ W_1(G, G_0) < \varepsilon_n}} \mathbb{E}_G \left[W_1(G, \widehat{G}_n) \right] \succcurlyeq n^{-1/(4d_0+2)}.$$

Let us give some intuition. The data we have access to is the empirical distribution F_n where n is the sample size, and which gets closer to the true

mixture $F(\cdot, G)$ at rate $n^{-1/2}$. Hence two mixing distributions G and G' can be told apart only if $\|F(\cdot, G) - F(\cdot, G')\|_\infty$ is at least of order $n^{-1/2}$.

As an example, let $G_0 = \delta_0$ and consider two-component mixing distributions around, say $G_n = \frac{1}{2}\left(\delta_{-2n^{-1/6}} + \delta_{2n^{-1/6}}\right)$ and $G'_n = \frac{4}{5}\delta_{-n^{-1/6}} + \frac{1}{5}\delta_{4n^{-1/6}}$. Both have 0 as first moment, and $4n^{-1/3}$ as second moment but the third moment is zero for G_n and $12n^{-1/2}$ for G'_n. A Taylor expansion in $\theta = 0$ up to the third order gives then $F(\cdot, G_n) = o(n^{-1/6})$ and $F(\cdot, G'_n) = o(n^{-1/6})$. So that no test can reliably tell G_n from G'_n with an n-sample. On the other hand, we clearly have $W_1(G_n, G'_n) = n^{-1/6}$ for all n. So that the minimax rate for two-mixing distributions cannot be better than $n^{-1/6}$.

This moment matching argument can be made rigorous and precise with two tools. One is Lindsay's Hankel trick (Lindsay, 1989, Theorem 2A), also used by Dacunha-Castelle and Gassiat (1997) to estimate the order of a mixture. The other is local asymptotic normality property (LAN) developed by Le Cam (1986). Section 6 uses them to build a LAN family with scale factor $n^{1/(4d_0+2)}$ which gives Theorem 3.2 via Theorem 6.1.

Upper bounds on local asymptotic minimax rates.

ASSUMPTION B(k). *The family of densities $\{f(\cdot, \theta)\}_{\theta\in\Theta}$ satisfies*

- *For all x, $F(x, \theta) = \int_{-\infty}^{x} f(\cdot, \theta)\mathrm{d}\lambda$ is k-differentiable w.r.t. θ,*
- *$\{F(\cdot, \theta), \theta \in \Theta\}$ is k-strongly identifiable,*
- *There is a uniform continuity modulus $\omega(\cdot)$ such that*

$$\sup_x \left|F^{(k)}(x, \theta) - F^{(k)}(x, \theta')\right| \leqslant \omega(\theta - \theta') \quad \textit{with} \quad \lim_{h\to 0} \omega(h) = 0.$$

The latter condition holds if $\sup_{x,\theta} |F^{(k+1)}(x, \theta)|$ exists and is finite. These differentiability conditions should be compared with the usual parametric case, where differentiability in quadratic mean, or twice differentiability in θ for a less technical condition, is enough to get $n^{-1/2}$ local minimax rate. We will need B($2\mathbf{m}$) to prove a global minimax rate of $n^{-1/(4\mathbf{m}-2)}$ (see (9) in Theorem 3.3), and B(1) for a pointwise rate of $n^{-1/2}$ everywhere (Theorem 4.1).

THEOREM 3.3. *Let $\widehat{G}_n(\mathbf{m})$ be "the"* minimum distance estimator*, that is any mixing distribution in $\mathcal{G}_{\leqslant\mathbf{m}}$ such that*

$$\|F(\cdot, \widehat{G}_n(\mathbf{m})) - F_n\|_\infty = \inf_{G\in\mathcal{G}_{\leqslant\mathbf{m}}} \|F(\cdot, G) - F_n\|_\infty. \tag{7}$$

Under Assumption $\mathrm{B}(2\mathbf{m})$*, there is* $\varepsilon > 0$ *such that, with* $q = 2d_0 + 1$*:*

$$\sup_{\substack{G \in \mathcal{G}_{\leqslant \mathbf{m}} \\ W_q(G,G_0)<\varepsilon}} \mathbb{E}_G \left[W_q(\widehat{G}_n(\mathbf{m}), G) \right] \preccurlyeq n^{-1/(2q)}, \tag{8}$$

and more globally, with $r = 2\mathbf{m} - 1$*:*

$$\sup_{G \in \mathcal{G}_{\leqslant \mathbf{m}}} \mathbb{E}_G \left[W_r(\widehat{G}_n(\mathbf{m}), G) \right] \preccurlyeq n^{-1/(2r)}. \tag{9}$$

REMARK 3.4. *Since* $G \mapsto \|F(\cdot, G) - F_n\|_\infty$ *is lower semi-continuous on the compact metric space* $(\mathcal{G}_{\leqslant \mathbf{m}}, W_q)$*, the infimum in* (7) *is attained . The minimum distance estimator is discussed by Deely and Kruse (1968) and Chen.*

Theorem 3.3 is proved it by establishing a uniform control of the ratio $\|F(\cdot, G) - F(\cdot, G')\|_\infty / W_q(G, G')^q$ in Theorem 6.3. To do so, we consider sequences of couples (G_n, G'_n) minimizing the relevant ratios, and expand $F(\cdot, G_n) - F(\cdot, G'_n)$ as a weighted sum on the relevant derivatives $F^{(p)}(\cdot, \theta_{j,n})$. A difficulty arises since distinct support points $\theta_{j,n}$ may converge to the same θ_j, leading to cancellations in the sums. Forgetting this case was the mistake in the proof of Chen (1995, Lemma 2). We overcome the issue in Section 7: we build clusters of support points whose pairwise distances decrease at a given rate and structured as nodes of a coarse-graining tree. We may then use Taylor expansions on each node and its descendants (Lemma 7.4).

REMARKS 3.5. *It is worth noticing the following from Theorems 3.2-3.3:*

- *They together imply that the optimal local asymptotic minimax rate is* $n^{-1/(4d_0+2)}$ *for estimating a mixture with at most* $\mathbf{m}$ *components around a mixture with* m_0 *components, for any transportation distance* W_p *with* $p \in [\![1, 2d_0 + 1]\!]$*.*
- *The rate is driven by* d_0*, that is, it gets harder to estimate the parameters of a mixture when it is close to a mixture with less components.*
- *The worst case is when* $m_0 = 1$*, yielding a global minimax rate of estimation* $n^{-1/(4\mathbf{m}-2)}$*. The rate gets worse when more components are allowed. So that the nonparametric rates for estimating mixtures with an infinite number of components like in deconvolution appear natural.*
- *On the other hand, when the number of components is known, that is* $\mathbf{m} = m_0$*, we have the usual local minimax rate* $n^{-1/2}$*.*
- *The global minimax rate on the mixtures with exactly* $\mathbf{m}$ *components stays at* $n^{-1/(4\mathbf{m}-2)}$*, because* $\mathcal{G}_{\mathbf{m}}$ *is not compact, and Theorem 3.2 still apply in the vicinity of* m_0*-component mixtures.*

4. On pointwise rate and superefficiency. The slow rate $n^{-1/(4\mathbf{m}-2)}$ in (9) might be a little surprising when for example some Bayesian estimators have $n^{-1/4}$ rate of convergence (Ishwaran, James and Sun, 2001). However this convergence rate is not the local minimax rate, but is closer to a pointwise rate of convergence, that is the speed at which an estimator converges to a fixed G when n increases. The difference with local minimax may be viewed as the loss of uniformity in G. We study here the optimal pointwise rates everywhere.

One motivation for local minimax results was to make clear how the Hodges' estimator (van der Vaart, 1998, ch.8) and other superefficient estimators could cohabit with Cramér-Rao bound, and how much they could improve on it.

Specifically, a superefficient estimator can have a better pointwise convergence rate than any regular estimator, but not a better local minimax convergence rate (Hájek, 1972). Moreover, it turns out that they can only have a better pointwise rate on a Lebesgue-null set (van der Vaart, 1998, ch.8).

Now, the set of parameters (weights and support points) defining $\mathcal{G}_{<\mathbf{m}}$ is a Lebesgue-null w.r.t. the one defining $\mathcal{G}_{\leqslant\mathbf{m}}$. Hence, we might expect that, by biasing the estimators toward the low numbers of components, we might attain better pointwise rates on $\mathcal{G}_{<\mathbf{m}}$, up to $n^{-1/2}$, which is the value when the number of components is known. By letting $\mathbf{m}$ go to infinity, we would have this pointwise rate for all finite mixing distributions. It turns out this is indeed the case.

THEOREM 4.1. *Consider for each $m \geqslant 1$ the minimum distance estimator $\widehat{G}_n(m)$ in $\mathcal{G}_{\leqslant m}$ as defined in Theorem 3.3, with $\widehat{G}_n(\infty)$ arbitrary. Fix $\kappa \in (0, 1/2)$ and set*

$$\widehat{m}_n = \min\left\{m \geqslant 1 : \|F(\cdot, \widehat{G}_n(m)) - F_n\|_\infty \leqslant n^{-1/2+\kappa}\right\}. \tag{10}$$

Under Assumption B(1)*, for any finite mixing distribution $G \in \mathcal{G}_{<\infty}$,*

$$\mathbb{E}_G\left[W_1(\widehat{G}_n(\widehat{m}_n), G)\right] \underset{G,\kappa}{\preccurlyeq} n^{-1/2}.$$

REMARKS 4.2.
- *Since the typical distance between empirical and theoretical distribution functions is $n^{-1/2}$, this $\widehat{m}_n$ in* (10) *is the lowest number of components that is not clearly insufficient.*
- *The rate $n^{-1/2}$ cannot be improved since it is the rate if the number of components is known beforehand.*

- *This is slightly stronger than just checking that we find the right number of components and then applying Theorem 3.3, because we need much less regularity. Only Assumption* B(1) *is required, instead of* B($2m$). *That is, we do not need more smoothness when the number of components increases. Under the hood we rely on the bound* (20) *instead of Theorem 6.3.*
- *The estimation of the number of components $\widehat{m}_n$ and the estimation of $\widehat{G}$ within $\mathcal{G}_{\widehat{m}_n}$ are not associated. For example, we may estimate $\widehat{m}_n$ with Equation* (10)*, and then use the maximum likelihood estimator $\widehat{G}$ on $\mathcal{G}_{\widehat{m}_n}$. Conversely, we may estimate the number of components using Gassiat and van Handel's (2013) penalized maximum likelihood estimator.*

5. Practical consequences and perspectives. Disagreement between local minimax rate and pointwise rate everywhere might be rare enough that it is worth recalling first what it means.

The asymptotic rate of convergence to a given G will be the pointwise rate $C_G n^{-1/2}$ where C_G is some positive constant. However, the estimator will enter this asymptotic regime only after a long time. More precisely, it enters this regime when G is not any more in any of the balls used in the local minimax bound. Alternatively, we may view this situation as the constant C_G exploding when G is close to specific G_0.

In our case, imagine we have a mixing distribution with three components, with all support points within distance $\delta > 0$ of some θ_0. Then about $\delta^{-(4(3-1)+2)} = \delta^{-10}$ observations are necessary to get an estimator with an error of δ. In particular, if G and G' are two such three-component mixing distributions, chosen to have the same first four moments, and $\tilde{G}$ and $\tilde{G}'$ are the same mixing distributions, rescaled to be ten times closer, we will need 10^{10} as many data points to tell them apart as for G and G'.

As a consequence, if the components of the mixing distribution to be estimated are not far apart one from the other, it is quite often impossible to get enough data points to get an appropriate estimate. An experimentalist with any leeway in what he measures (use of different markers, say) might then wish to ensure that the peaks are far apart, even at the cost of many data points.

We end the section by some thoughts on possible further work. This article contains the proof that the optimal local minimax rate of estimation around a mixing distribution with m_0 components among mixing distributions with $\mathbf{m}$ components is $n^{-1/(4(\mathbf{m}-m_0)+2)}$, when the parameter space Θ is a compact subset of $\mathbb{R}$.

We think that extension to a multivariate Θ should be workable, much like Nguyen (2013) did for the former erroneous result. On the other hand, non-compactness of Θ would probably bring about technical difficulties, and cases where the result would not hold. Stronger forms of identifiability would probably be required in general, to avoid problems with limits. Moreover, for many natural higher-dimensional families, strong identifiability does not hold, so that the results would be different.

Finally, another line of inquiry are the results that might be expected in a Bayesian framework. The most natural equivalent to the convergence rate of the *a posteriori* distribution to the real parameter is the pointwise rate of convergence. Hence the question: can we build Bayesian estimators where the *a posteriori* distributions converge at rate $n^{-1/2}$ everywhere? Of course, the convergence would not be uniform.

6. Key tools and proofs.

6.1. *Local asymptotic normality for Theorem 3.2.* All the densities considered in the sequel are w.r.t. some given dominating σ-finite measure on $\mathbb{R}$. We call *experiment* a family $\mathcal{E}$ of densities.

THEOREM 6.1. *Let $G_0 \in \mathcal{G}_{m_0}$ with a support point θ_0 in the interior of Θ. There is a family $\{G_n(u)\}_{n\geqslant 0, u\in\mathbb{R}}$ in $\mathcal{G}_{\mathbf{m}}$ with the following properties:*

a. For all distinct u, u' in $\mathbb{R}$, we have together

$$W_1(G_n(u), G_n(u')) \underset{u,u'}{\gtrsim} n^{-1/(4d_0+2)} \underset{u}{\gtrsim} W_1(G_n(u), G_0);$$

b. Assume $\mathrm{A}(d_0, \theta_0)$ for the family $\{f(\cdot,\theta)\}_{\theta\in\Theta}$ and set the product density $f_{n,u} = \otimes_{i=1}^n f(\cdot, G_n(u))$. There is an increasing real sequence $U_n \to \infty$ such that the sequence of experiments $\mathcal{E}_n = (f_{n,u})_{u\in[-U_n,U_n]}$ is locally asymptotically normal *(LAN) : there are random variables Z_n, asymptotically $\mathcal{N}(0,1)$, and numbers $\Gamma_n > 0$ such that for all $u \in \mathbb{R}$,*

$$\mathrm{Log}\left(\frac{f_{n,u}(X)}{f_{n,0}(X)}\right) - uZ_n\sqrt{\Gamma_n} + \frac{u^2}{2}\Gamma_n \xrightarrow[n\to\infty]{P} 0 \tag{11}$$

where X is a n-sample of density $f_{n,0}$.
In addition, we have $\liminf_n \Gamma_n > 0$ and $\limsup_n \Gamma_n < \infty$.

REMARK 6.2. *We want only an example of this slow convergence, and it should be somewhat typical. That is why we have chosen the regularity*

conditions to make the proof easy, while still being easy to check, in particular for exponential families.

In particular, in Assumption $\mathrm{A}(d_0, \theta_0)$*, it could probably be possible to lower* α *in* (p, α)*-smoothness to* $2 + \varepsilon$ *and still get the uniform bound we use in the law of large numbers below. Similarly, less differentiability might be necessary if we tried to imitate differentiability in quadratic mean.*

Conversely, under possibly more stringent regularity conditions, Γ_n *is expected to converge to* $\mathbb{E}_{G_0}\left|\frac{f^{(2d_0+1)}(\cdot,\theta_0)}{f(\cdot,G_0)}\right|^2$ *up to a multiplicative constant.*

Proof of Theorem 6.1. Write the mixing distribution G_0 as

$$G_0 = \sum_{j=1}^{m_0-1} \pi_j \delta_{\theta_j} + \pi_0 \delta_{\theta_0} \tag{12}$$

with θ_0 in the interior of Θ. Let $u \in \mathbb{R}$ and replace the Dirac measure δ_{θ_0} in (12) with a mixing distribution $H_n(u)$:

$$G_n(u) = \sum_{j=1}^{m_0-1} \pi_j \delta_{\theta_j} + \pi_0 H_n(u). \tag{13}$$

We want to choose $H_n(u)$ close to δ_{θ_0}. To this end, set $\mu_0 = 1$ and $\mu_{2d-1} = u$ with $d = d_0 + 1$,. Choose in addition numbers $\mu_1, \ldots, \mu_{2d-2}$ such that the $k \times k$-Hankel matrices $(M_k)_{i,j} = \mu_{i+j-2}$ satisfy $\det M_k > 0$ for $k \in [\![1, d-1]\!]$. Then, by Lindsay's Theorem 2A (1989), there is a unique mixing distribution $H(u) = \sum_{j=m_0}^{\mathbf{m}} \pi_j(u) \delta_{h_j(u)}$ with exactly d support-points $h_j(u)$ and first moments μ_k up to order $2d-1$ satisfying

$$\sum_{j=m_0}^{\mathbf{m}} \pi_j(u) h_j(u)^k = \mu_k, \quad k \in [\![0, 2d-1]\!]. \tag{14}$$

Define then $H_n(u)$ by shifting and rescaling the support points of $H(u)$:

$$H_n(u) = \sum_{j=m_0}^{\mathbf{m}} \pi_j(u) \delta_{\theta_0 + \varepsilon_n h_j(u)} \quad \text{with} \quad \varepsilon_n = n^{-1/(4d-2)}.$$

Now, using the dual representation (1) of W_1, we see that

$$W_1(G_n(u), G_0) = \pi_0 W_1(H_n(u), \delta_{\theta_0}) = \pi_0 \varepsilon_n W_1(H(u), \delta_0)$$

and likewise, $W_1(G_n(u), G_n(u'))$ equals $\pi_0 \varepsilon_n W_1(H(u), H(u'))$ so that Theorem 6.1.a follows.

To guarantee that the points $\theta_0 + \varepsilon_n h_j(u)$ involved in $G_n(u)$ stay inside Θ uniformly in u, let us show that the functions $h_j(\cdot)$ are continuous. Consider the map

$$\varphi(\pi_1, \dots, \pi_d, h_1, \dots, h_d) = \left(\sum_1^d \pi_j, \sum_1^d \pi_j h_j, \sum_1^d \pi_j h_j^2, \dots, \sum_1^d \pi_j h_j^{2d-1} \right)$$

on the set $\{(\pi_1, \dots, \pi_d, h_1, \dots, h_d) : \pi_1 > 0, \dots, \pi_d > 0, h_1 < \dots < h_d\}$. The uniqueness in Theorem 2A by Lindsay (1989) implies that φ is injective. Moreover, its Jacobian is non-zero, as it can be seen by recurrence on d:

$$J(\varphi) = (-1)^{\frac{(d-1)d}{2}} \pi_1 \cdots \pi_d \prod_{1 \leqslant j < k \leqslant d} (h_j - h_k)^4.$$

Thus the inverse of φ is locally continuous, so that, in particular, the $h_j(u)$ are all continuous.

Set now

$$\mathbf{h}(U) = \max_{j \leqslant d} \max_{|u| \leqslant U} |h_j(u)| \tag{15}$$

which is finite for any $U > 0$ and choose a positive sequence (U_n) such that

$$U_n \to \infty \quad \text{and} \quad \varepsilon_n \mathbf{h}(U_n) \to 0.$$

We can now prove local asymptotic normality (11). Let $X = (X_{1,n}, \dots, X_{n,n})$ be an i.i.d. sample with density $f_{n,0}$. Since we proceed along the lines of Chen (1995), the proof is only sketched here. Write the log-likelihood ratio as

$$\operatorname{Log} \left(\frac{f_{n,u}(X)}{f_{n,0}(X)} \right) = \sum_{i=1}^n \operatorname{Log} \left(1 + Y_{i,n}(u)\right)$$

with

$$Y_{i,n}(u) = \frac{f\left(X_{i,n}, G_n(u)\right) - f\left(X_{i,n}, G_n(0)\right)}{f(X_{i,n}, G_n(0))}.$$

The main steps are as follows, see Heinrich and Kahn (2015, Sections A.1, A.2 and A.3) for the details:

Step 1. Use linearity of $G \mapsto f(\cdot, G)$ and Taylor expansions up to the order $2d-1$ with remainder on $Y_{i,n}(u)$ at θ_0 to show that the r.v.'s

$$Z_{i,n} = \pi_0 \frac{f^{(2d-1)}(X_{i,n}, \theta_0)}{f(X_{i,n}, G_n(0))}$$

are centered under $f_{n,0}$.

Step 2. Define $\Gamma_n = \mathbb{E}_{G_n(0)}|Z_{1,n}|^2$ and $Z_n = n^{-1/2}\Gamma_n^{-1/2}\sum_{i=1}^n Z_{i,n}$ and prove that Z_n is asymptotically $\mathcal{N}(0,1)$ via Lyapunov Theorem for triangular arrays.

Step 3. Show the following convergences for all u:

$$A_n(u) := \sum_{i=1}^n Y_{i,n}(u) - uZ_n\sqrt{\Gamma_n} \quad \xrightarrow{L^2} \quad 0,$$

$$B_n(u) := \sum_{i=1}^n Y_{i,n}(u)^2 - u^2\Gamma_n \quad \xrightarrow{L^1} \quad 0,$$

$$C_n(u) := \sum_{i=1}^n |Y_{i,n}(u)|^3 \quad \xrightarrow{L^1} \quad 0.$$

Then derive the LAN property from the equality

$$\mathrm{Log}\left(\frac{f_{n,u}(X)}{f_{n,0}(X)}\right) - uZ_n\sqrt{\Gamma_n} + \frac{u^2}{2}\Gamma_n = A_n(u) + \frac{1}{2}B_n(u) + O_P(C_n(u)).$$

□

6.2. *Proof ofTheorem 3.2.* Let us show how Theorem 6.1 entails Theorem 3.2 using just two points and contiguity (Le Cam, 1960). Consider any sequence of estimators $\widehat{G}_n$, and $G_n(u)$ for $u = 0, 1$ as defined in Theorem 6.1. It's enough to show that for large n,

$$\sup_{G\in\{G_n(0),G_n(1)\}} \mathbb{E}_G[W_1(G,\widehat{G}_n)] \succcurlyeq n^{-1/(4d_0+2)}. \tag{16}$$

Recall that we set here $\varepsilon_n = n^{-1/(4d_0+2)+\kappa}$ and note that $G_n(0)$ and $G_n(1)$ are in the ball $\{G : W_1(G, G_0) < \varepsilon_n\}$ for large n, by Theorem 6.1.a.

Consider the probability measures $P_{n,u}$ with the densities $f_{n,u}$ of Theorem 6.1.b for $u = 0, 1$ and set $g_n = \dfrac{f_{n,1}(X)}{f_{n,0}(X)}\exp(-Z_n\sqrt{\Gamma_n} + \Gamma_n/2)$. Then,

$$P_{n,1}(A) = \mathrm{e}^{-\frac{\Gamma_n}{2}}\int_A \mathrm{e}^{Z_n\sqrt{\Gamma_n}} g_n \,\mathrm{d}P_{n,0} \geqslant \frac{\mathrm{e}^{-\frac{\Gamma_n}{2}}}{2}\int_{A\cap\{Z_n>0\}\cap\{g_n>1/2\}} \mathrm{d}P_{n,0}.$$

We have $g_n \xrightarrow{P_{n,0}} 1$ by (11) so that $P_{n,0}(g_n \leqslant 1/2) \leqslant 1/16$ for large n. Since Z_n is asymptotically $\mathcal{N}(0,1)$, we also have $P_{n,0}(Z_n \leqslant 0) \leqslant 1/2 + 1/16$. Thus $P_{n,0}(\{Z_n > 0\} \cap \{g_n > 1/2\})$ is at least $3/8$ for n large enough and

$$P_{n,0}(A) \geqslant \frac{3}{4} \implies P_{n,1}(A) \succcurlyeq \mathrm{e}^{-\Gamma_n/2}. \tag{17}$$

Now, choose $A = \{W_1(G_n(1), \widehat{G}_n) \geqslant an^{-1/(4d_0+2)}\}$ where $a > 0$ is such that $W_1(G_n(1), G_n(0)) \geqslant 2an^{-1/(4d_0+2)}$, by Theorem 6.1.a. By the triangle's inequality, the complement A^c is included in $\{W_1(G_n(0), \widehat{G}_n) \geqslant an^{-1/(4d_0+2)}\}$. Now, either we have $P_{n,0}(A^c) \geqslant \frac{1}{4}$ and, for $G = G_n(0)$, we get

$$\mathbb{E}_G[W_1(G, \widehat{G}_n)\mathbf{1}_{A^c}] \geqslant \frac{a}{4}n^{-1/(4d_0+2)},$$

or we have $P_{n,0}(A) \geqslant \frac{3}{4}$ and by using (17), for $G = G_n(1)$, we get

$$\mathbb{E}_G[W_1(G, \widehat{G}_n)\mathbf{1}_A] \succcurlyeq \mathrm{e}^{-\Gamma_n/2}an^{-1/(4d_0+2)},$$

so that (16) is proved since $\limsup_n \Gamma_n < \infty$.

6.3. *Comparison between distances for Theorem 3.3 and Theorem 4.1.* The key technical tool is

THEOREM 6.3. • *Let $G_0 \in \mathcal{G}_{m_0}$. Under Assumption* B($2\mathbf{m}$)*, there are $\varepsilon > 0$ and $\delta > 0$ such that, with $q = 2d_0 + 1$:*

$$\inf_{\substack{G \neq G' \in \mathcal{G}_{\leqslant \mathbf{m}} \\ W_q(G,G_0) < \varepsilon \\ W_q(G',G_0) < \varepsilon}} \frac{\|F(\cdot, G) - F(\cdot, G')\|_\infty}{W_q(G, G')^q} > \delta, \tag{18}$$

and more globally, with $r = 2\mathbf{m} - 1$:

$$\inf_{G \neq G' \in \mathcal{G}_{\leqslant \mathbf{m}}} \frac{\|F(\cdot, G) - F(\cdot, G')\|_\infty}{W_r(G, G')^r} > \delta. \tag{19}$$

- *Let $G_0 \in \mathcal{G}_{m_0}$. Under Assumption* B(1)*, there are $\varepsilon > 0$ and $\delta > 0$ such that*

$$\inf_{\substack{G \neq G' \in \mathcal{G}_{\leqslant m_0} \\ W_1(G,G_0) < \varepsilon \\ W_1(G',G_0) < \varepsilon}} \frac{\|F(\cdot, G) - F(\cdot, G')\|_\infty}{W_1(G, G')} > \delta. \tag{20}$$

- *Let now $G_0 \in \mathcal{G}_{\leqslant m_0}$. Under Assumption* B(2)*, there are $\varepsilon > 0$ and $\delta > 0$ such that*

$$\inf_{\substack{G \in \mathcal{G}_{\leqslant m_0} \\ W_1(G,G_0) < \varepsilon}} \frac{\|F(\cdot, G) - F(\cdot, G_0)\|_\infty}{W_1(G, G_0)^2} > \delta. \tag{21}$$

The proof of Theorem 6.3 is postponed to Section 7 where the novel ingredient, a coarse-graining tree, is constructed to prove (18) and (19), the most difficult points. These one entail Theorem 3.3. The two related bounds (20) and (21) hold under weaker differentiability assumptions, but are less general. Bound (20) covers the case where the number of components in the mixture is known, and is used for the proof of Theorem 4.1. Bound (21) is the valid weaker version of Lemma 2 by Chen (1995), which is sufficient for the use other authors have made of it. Here, we only compare mixtures in a ball with the mixture at the center of the ball.

For the proofs of Theorem 3.3 and Theorem 4.1, we need in addition:

LEMMA 6.4. *Let $q, d \geqslant 1$ and $G \in \mathcal{G}_{\leqslant m}$. Assume that the minimum distance estimators $\widehat{G}_n := \widehat{G}_n(m)$ defined in Theorem 3.3 satisfy for some constant $C > 0$ and on some event A,*

$$W_q(\widehat{G}_n, G)^d \leqslant C\|F(\cdot, \widehat{G}_n) - F(\cdot, G)\|_\infty.$$

Then

$$\mathbb{E}_G\left[W_q(\widehat{G}_n, G)\right] \leqslant \left(2\pi C^2\right)^{1/2d} n^{-1/2d} + \mathrm{Diam}(\Theta)\mathbb{P}_G(A^c).$$

Moreover, $\mathbb{P}_G(A^c)$ is at most $2\mathrm{e}^{-2nz^2}$ if A is either $\{\|F_n - F(\cdot, G)\|_\infty \leqslant z\}$ or $\{\|F(\cdot, \widehat{G}_n) - F(\cdot, G)\|_\infty \leqslant 2z\}$.

SKETCH OF PROOF. Bound $W_q(\widehat{G}_n, G)$ by $\mathrm{Diam}(\Theta)$ on A^c, use the definition (7) and the triangle's inequality to bound $\|F(\cdot, \widehat{G}_n) - F(\cdot, G)\|_\infty$ by $2\|F_n - F(\cdot, G)\|_\infty$, then use Jensen's inequality on A and bound $\mathbb{E}_{G_1}\|F_n - F(\cdot, G)\|_\infty$ (and $\mathbb{P}_G(A^c)$) by applying DKW's inequality (Massart, 1990). □

6.4. *Proof of Theorem 3.3.* Let $\varepsilon, \delta > 0$ such (18) holds. Set

$$z = \frac{1}{2} \inf_{\substack{G, G' \in \mathcal{G}_{\leqslant \mathbf{m}} \\ W_q(G, G_0) \leqslant \varepsilon/2 \\ W_q(G', G_0) \geqslant \varepsilon}} \|F(\cdot, G) - F(\cdot, G')\|_\infty.$$

The infimum is taken over a compact set and is thus attained. We have $z > 0$ by identifiability (coming from Assumption B($2\mathbf{m}$)).

Consider $A = \{\|F(\cdot, G) - F(\cdot, \widehat{G}_n(\mathbf{m}))\|_\infty \leqslant z\}$. If G is in $\mathcal{G}_{\leqslant \mathbf{m}}$ with $W_q(G, G_0) \leqslant \varepsilon/2$ then $\widehat{G}_n(\mathbf{m})$ must satisfy $W_q(\widehat{G}_n(\mathbf{m}), G_0) < \varepsilon$ on the event A so that by (18),

$$\text{On } A, \quad W_q(\widehat{G}_n(\mathbf{m}), G)^q < \frac{1}{\delta}\|F(\cdot, \widehat{G}_n(\mathbf{m})) - F(\cdot, G)\|_\infty.$$

Applying Lemma 6.4 with $d = q = 2d_0 + 1$ and $C = 1/\delta$ yields

$$\mathbb{E}_G[W_q(\widehat{G}_n(\mathbf{m}), G)] \leqslant \left(\frac{2\pi}{\delta^2}\right)^{1/2q} n^{-1/2q} + 2\,\mathrm{Diam}(\Theta)\mathrm{e}^{-nz^2/2}$$

so that bound (8) is proved. Bound (9) is obtained likewise from (19).

6.5. *Proof of Theorem 4.1.* Consider a mixing distribution $G_0 \in \mathcal{G}_{m_0}$. Under Assumption B(1), let $\varepsilon, \delta > 0$ such (20) holds. Fix $\kappa \in (0, \frac{1}{2})$ and set

$$z = n^{-1/2+\kappa} \wedge \frac{1}{4} \inf_{\substack{G \in \mathcal{G}_{\leqslant m_0} \\ W_1(G, G_0) \geqslant \varepsilon}} \|F(\cdot, G) - F(\cdot, G_0)\|_\infty .$$

By compactness and identifiability, the infimum in z is attained and positive. On the event $A = \{\|F(\cdot, G_0) - F_n\|_\infty \leqslant z\}$, the minimum distance estimator $\widehat{G}_n(m_0)$ in $\mathcal{G}_{\leqslant m_0}$, defined in Theorem 3.3, satisfies

$$\|F(\cdot, \widehat{G}_n(m_0)) - F_n\|_\infty \leqslant \|F(\cdot, G_0) - F_n\|_\infty \leqslant z \leqslant n^{-1/2+\kappa}.$$

so that $\hat{m}_n$ is at most m_0 by (10); thus we have $\widehat{G}_n(\hat{m}_n) \in \mathcal{G}_{\leqslant m_0}$. Next, $\widehat{G}_n(\hat{m}_n)$ must satisfy $W_1(\widehat{G}_n(\hat{m}_n), G_0) < \varepsilon$ on A since

$$\|F(\cdot, \widehat{G}_n(\hat{m}_n)) - F(\cdot, G_0)\|_\infty \leqslant 2\|F(\cdot, G_0) - F_n\|_\infty \leqslant 2z,$$

by the triangle's inequality. Applying then (20) on A, we get

$$W_1(\widehat{G}_n(\hat{m}_n), G_0) < \frac{1}{\delta} \|F(\cdot, \widehat{G}_n(\hat{m}_n)) - F(\cdot, G_0)\|_\infty.$$

Now, apply Lemma 6.4 with $q = d = 1$, $C = 1/\delta$ and A as above, so that

$$\mathbb{E}_{G_0}[W_1(\widehat{G}_n(\hat{m}_n), G_0)] \leqslant \sqrt{\frac{2\pi}{\delta}} n^{-1/2} + 2\,\mathrm{Diam}(\Theta) \exp(-2n^{2\kappa}).$$

7. The coarse-graining tree and the proof of Theorem 6.3.

7.1. *Proof of* (18)*: the coarse graining tree.* Let $G_0 \in \mathcal{G}_{m_0}$. We have to show that, under Assumption B($2\mathbf{m}$), there is $\varepsilon > 0$ such that

$$L := \inf_{\substack{G \neq G' \in \mathcal{G}_{\leqslant \mathbf{m}} \\ W_q(G, G_0) < \varepsilon \\ W_q(G', G_0) < \varepsilon}} \frac{\|F(\cdot, G) - F(\cdot, G')\|_\infty}{W_q(G, G')^q} > 0 \quad \text{with } q = 2d_0 + 1. \tag{22}$$

Assume on the contrary that $L = 0$ and choose mixing distributions G_n and G'_n in $\mathcal{G}_{\leqslant \mathbf{m}}$ with $W_q(G_n, G_0) \vee W_q(G'_n, G_0) < 1/n$ such that for each $n \geqslant 1$, the ratios $\|F(\cdot, G_n) - F(\cdot, G'_n)\|_\infty / W_q(G_n, G'_n)^q$ are less than $1/n$. We shall prove, up to selecting subsequences, the following contradiction:

$$\left\|F(\cdot, G_n) - F(\cdot, G'_n)\right\|_\infty \succcurlyeq W_q(G_n, G'_n)^q. \tag{23}$$

Some notations. We may and do assume that there are integers m, m' at most $\mathbf{m}$ such that $(G_n) \subset \mathcal{G}_m$ and $(G'_n) \subset \mathcal{G}_{m'}$. We can then write $G_n = \sum_{j=1}^{m} \pi_{j,n}\delta_{\theta_{j,n}}$ and $G'_n = \sum_{j=1}^{m'} \pi'_{j,n}\delta_{\theta'_{j,n}}$ and set

$$(\varpi_{j,n}, \vartheta_{j,n}) = \begin{cases} (\pi_{j,n}, \theta_{j,n}) & \text{if } j \leqslant m \\ (-\pi'_{j-m,n}, \theta'_{j-m,n}) & \text{if } j > m \end{cases},$$

so that the signed measure $G_n - G'_n = \displaystyle\sum_{j=1}^{m+m'} \varpi_{j,n}\delta_{\vartheta_{j,n}}$ has total mass zero.

The discrepancy orders of the $\vartheta_{j,n}$'s. We shall first classify the differences between the $\vartheta_{j,n}$'s in an intrinsic way:

LEMMA 7.1. *For a suitable subsequence of $G_n - G'_n$, there is a finite number S of "scaling" sequences*

$$0 \equiv \varepsilon_0(n) < \varepsilon_1(n) < \cdots < \varepsilon_S(n) \equiv 1 \text{ with } \varepsilon_s(n) = o\big(\varepsilon_{s+1}(n)\big),$$

such that for all $j, j' \in [\![1, m+m']\!]$ there is a unique $s(j, j') \in [\![0, S]\!]$ satisfying

$$\big|\vartheta_{j,n} - \vartheta_{j',n}\big| \asymp \varepsilon_{s(j,j')}(n).$$

The proof is given in Heinrich and Kahn (2015, Appendix C). It follows from the definition of $s(j, j')$ that $s(j, j') \leqslant \max(s(j, j"), s(j', j"))$ and thus $s(\cdot, \cdot)$ defines an ultrametric on $[\![1, m + m']\!]$. The ultrametric makes any two balls either included one into the other, or disjoint, and allows us to build a coarse-graining tree :

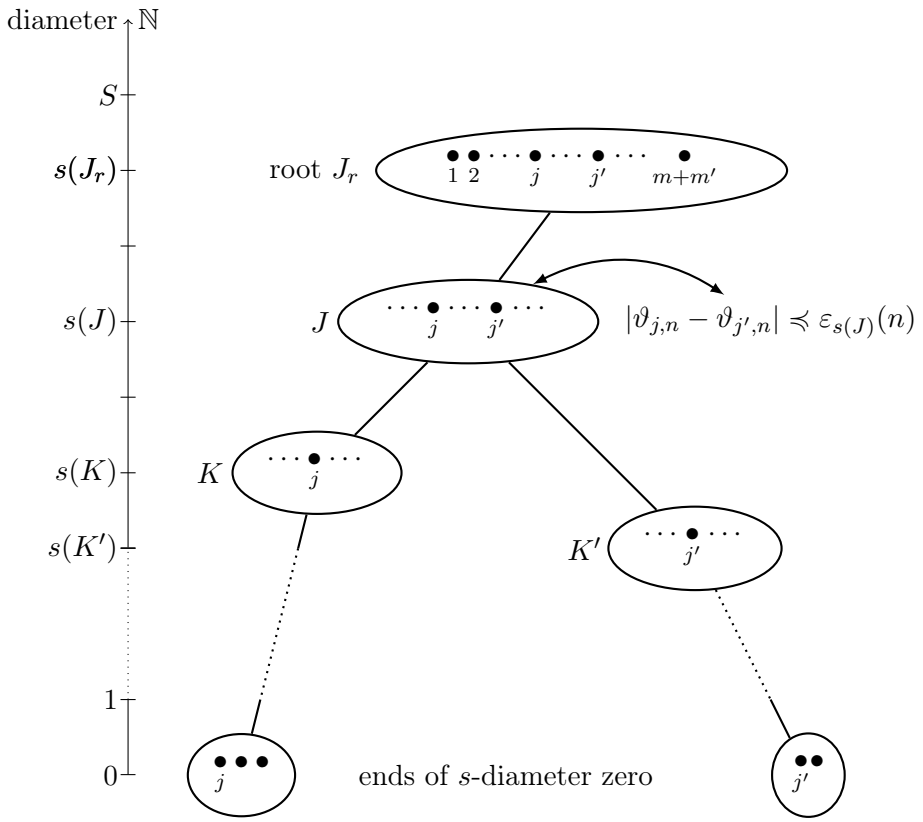
DEFINITION 7.2. *The* coarse-graining tree $\mathcal{T}$ *is the collection of distinct balls $J = \{s(\cdot, j) \leqslant s\}$, called* nodes*, when j ranges over $[\![1, m + m']\!]$ and s over $[\![0, S]\!]$. Moreover:*

- *The* root *of $\mathcal{T}$ is $J_r = [\![1, m + m']\!]$,*
- *The* parent *$J^\uparrow$ of a node J is defined by*

$$(J \subset I \subsetneq J^\uparrow, I \in \mathcal{T}) \implies I = J,$$

- *The* set of children *of a node J is* $\mathrm{Child}(J) = \{I \in \mathcal{T} : I^\uparrow = J\}$,
- *The* set of descendants *of a node J is* $\mathrm{Desc}(J) = \{I \in \mathcal{T} : I^\uparrow \subset J\}$,
- *The* diameter *of a node J is $s(J) = \max_{j,j' \in J} s(j, j')$.*

Let us show how the tree $\mathcal{T}$ looks like with a partial representation :



Note that the ends are not necessarily singletons since the ultrametric $s(\cdot,\cdot)$ does not separate points. Note also that j and j' are in different children K and K' so that $|\vartheta_{j,n} - \vartheta_{j',n}|$ is actually exactly of order $\varepsilon_{s(J)}(n)$.

The Wassertsein distances $W_q(G_n, G'_n)$ through the coarse graining tree $\mathcal{T}$. In what follows n is skipped in the ϑ_j's, ϖ_j's and ε_s's. Set for short

$$\varpi_J = \sum_{j\in J} \varpi_j \quad \text{and} \quad \varepsilon_J = \varepsilon_{s(J)}. \tag{24}$$

LEMMA 7.3. *For any $q \geqslant 1$, we have*

$$W_q\left(G_n, G'_n\right)^q \asymp \max_{J\in \mathrm{Desc}(J_r)} |\varpi_J|\, \varepsilon_{J^\uparrow}^q.$$

PROOF. Consider any coupling Π between G_n and G'_n and set

$$\begin{aligned} \Pi(J, J') &= \Pi\left(\{\vartheta_j\}_{j\in J\cap[\![1,m]\!]} \times \{\vartheta_j\}_{j\in J'\cap[\![m+1,m+m']\!]}\right), \\ w_q(J, J') &= \sum_{(j,j')\in J\times J'} \Pi(\{j\},\{j'\})|\vartheta_j - \vartheta_{j'}|^q. \end{aligned} \tag{25}$$

Set also $\pi_J = \sum_{j\in J\cap[\![1,m]\!]} \varpi_j$ and $\pi'_J = -\sum_{j\in J\cap[\![m+1,m+m']\!]} \varpi_j$. These define the marginal distributions of Π and we have $\Pi(J,J) \leqslant \pi_J \wedge \pi'_J$. Note also that $|\varpi_J| = \pi_J \vee \pi'_J - \pi_J \wedge \pi'_J$. With $J^c = J_r \setminus J$, this gives

$$\Pi(J,J^c) \vee \Pi(J^c,J) \geqslant |\varpi_J|.$$

Notice moreover that if $(j,j') \in J \times J^c$, then $|\vartheta_j - \vartheta_{j'}| \succcurlyeq \varepsilon_{J\uparrow}$. Hence the lower bound of Lemma 7.3 follows from

$$w_q(J_r,J_r) \geqslant w_q(J,J^c) + w_q(J^c,J) \succcurlyeq |\varpi_J|\, \varepsilon^q_{J\uparrow}.$$

Conversely, for the upper bound, we show recursively that for all node J,

$$w_q(J,J) \preccurlyeq \max_{K\in \mathrm{Desc}(J)} |\varpi_K| \varepsilon^q_{K\uparrow}. \tag{26}$$

This is obviously true if J is an end node, with the value of zero. Assume that (26) holds for children K of a given node J. We may develop J on its children:

$$w_q(J,J) = \sum_{K\in \mathrm{Child}(J)} \Big[w_q(K,K) + \sum_{\substack{K'\in \mathrm{Child}(J)\\ K'\neq K}} w_q(K,K') \Big].$$

Furthermore, we get $w_q(K,K') \preccurlyeq \Pi(K,K')\, \varepsilon^q_J$ from (25) and

$$\Pi(K,K') \leqslant \Pi(K,K^c) \leqslant \pi_K - \Pi(K,K),$$

and if the coupling Π is chosen (see Heinrich and Kahn (2015, Lemma B.2) for a construction) such that $\Pi(K,K) = \pi_K \wedge \pi'_K$ for all node K, then it follows that

$$\Pi(K,K') \leqslant |\varpi_K|,$$

and thus

$$w_q(J,J) \preccurlyeq \sum_{K\in \mathrm{Child}(J)} \left[w_q(K,K) + |\varpi_K| \varepsilon^q_J \right].$$

The recurrence hypothesis on children K yields then (26). □

Expanding $F(x,G_n) - F(x,G'_n)$ through the coarse graining tree. The dependence on n is skipped in the following notations. Consider the additive set-function $J \mapsto F(x,J) = \sum_{j\in J} \varpi_j F(x,\vartheta_j)$ and note that $F(x,J_r)$ is equal to $F(x,G_n) - F(x,G'_n)$.

LEMMA 7.4. *Choose ϑ_J in $\{\vartheta_j : j \in J\}$ for each node J of $\mathcal{T}$.*
There are a vector $a_J = (a_J(p))_{0\leqslant p\leqslant 2\mathbf{m}}$ and a remainder $R(x,J)$ such that

$$F(x,J) = \sum_{p=0}^{2\mathbf{m}} a_J(p)\varepsilon_J^p F^{(p)}(x,\vartheta_J) + R(x,J), \tag{27}$$

where:

(a) $a_J(0) = \varpi_J$ and $\|a_J\| \preccurlyeq 1$,
(b) There is an integer $p_J < |J|$ such that $\|a_J\| \asymp |a_J(p_J)|$,
(c) The norm $\|a_J\|$ is bounded from below by a quantity linked to W_q:

$$\|a_J\| \succcurlyeq \max_{K\in \mathrm{Desc}(J)} \left[|\varpi_K| \left(\frac{\varepsilon_{K^\uparrow}}{\varepsilon_J} \right)^{|J|-1} \right],$$

(d) $R(x,J) = o\left(\|a_J\|\,\varepsilon_J^{2m}\right)$ uniformly in x.

As a remark, the lower bound on $F(x,J_r)$ will stem from points (c) and (d). Points (a) and (b) are mainly there for transmitting recurrence hypotheses. They control the size of $F(x,J)$, together with point (c). The behaviour of $F(x,J)$ only depends on the first $|J|$ terms in the sum. However, the sum goes to $2\mathbf{m}$ so that it is useful when $J = J_r$.

PROOF. The proof uses Taylor expansions at θ_K for a given generation of children K together with separation and order properties of the coarse-graining tree $\mathcal{T}$. Recall notation (24).

If K is an end of the tree $\mathcal{T}$, then all the θ_j for $j \in K$ are equal, and $F(x,K) = \varpi_K F(x,\theta_K)$. Choose $a_K(p) = \varpi_K \mathbf{1}_{\{p=0\}}$ and $R(x,K) = 0$ so that the equality (27) holds for the end node K with all the desired estimates (a), (b), (c) and (d).

Assume now that J has children K, each of them satisfying (27) with all the estimates (a), (b), (c) and (d):

$$F(x,K) = \sum_{\ell=0}^{2\mathbf{m}} a_K(\ell)\varepsilon_K^\ell F^{(\ell)}(x,\vartheta_K) + R(x,K), \tag{28}$$

We want to transmit (28) and the estimates to the parent J. Suppose without loss of generality that $\vartheta_J \leqslant \vartheta_K$ and apply Taylor's formula with remainder

to $F^{(\ell)}(x,\vartheta_K)$ at ϑ_J for all $\ell \in [\![0, 2\mathbf{m}]\!]$:

$$F^{(\ell)}(x,\vartheta_K) - \sum_{p=\ell}^{2\mathbf{m}-1} \frac{(\vartheta_K-\vartheta_J)^{p-\ell}}{(p-\ell)!} F^{(p)}(x,\vartheta_J) = \int_{\vartheta_J}^{\vartheta_K} \frac{(\vartheta_K-\theta)^{2\mathbf{m}-1-\ell}}{(2\mathbf{m}-1-\ell)!} F^{(2m)}(x,\theta)\mathrm{d}\theta.$$

Substract the term $\frac{(\vartheta_K-\vartheta_J)^{2\mathbf{m}-\ell}}{(2\mathbf{m}-\ell)!} F^{(2\mathbf{m})}(x,\vartheta_J)$ from either side so that

$$\begin{aligned} F^{(\ell)}(x,\vartheta_K) - \sum_{p=\ell}^{2\mathbf{m}} \frac{(\vartheta_K-\vartheta_J)^{p-\ell}}{(p-\ell)!} F^{(p)}(x,\vartheta_J) &= \int_{\vartheta_J}^{\vartheta_K} \frac{(\vartheta_K-\theta)^{2\mathbf{m}-1-\ell}}{(2\mathbf{m}-1-\ell)!} \left[F^{(2\mathbf{m})}(x,\theta) - F^{(2\mathbf{m})}(x,\vartheta_J)\right] \mathrm{d}\theta \\ &= (\vartheta_K-\vartheta_J)^{2\mathbf{m}-\ell} O\left(\sup_{\theta\in[\vartheta_J,\vartheta_K]} |F^{(2\mathbf{m})}(x,\theta) - F^{(2\mathbf{m})}(x,\vartheta_J)| \right). \end{aligned}$$

The modulus of continuity of $F^{(2\mathbf{m})}(x,\cdot)$ from $\mathrm{B}(2\mathbf{m})$ then yields

$$F^{(\ell)}(x,\vartheta_K) = \sum_{p=\ell}^{2\mathbf{m}} \frac{(\vartheta_K-\vartheta_J)^{p-\ell}}{(p-\ell)!} F^{(p)}(x,\vartheta_J) + o\left((\vartheta_K-\vartheta_J)^{2\mathbf{m}-\ell}\right).$$

The normalised discrepancies $\phi_K := (\vartheta_K-\vartheta_J)/\varepsilon_J$ for $K \in \mathrm{Child}(J)$ are by definition at most of order 1 so that

$$F^{(\ell)}(x,\vartheta_K) = \sum_{p=\ell}^{2\mathbf{m}} \varepsilon_J^{p-\ell} \frac{\phi_K^{p-\ell}}{(p-\ell)!} F^{(p)}(x,\vartheta_J) + \varepsilon_J^{2\mathbf{m}-\ell} o(1).$$

Subtitute in (28) and change the order of summation:

$$F(x,K) = \sum_{p=0}^{2\mathbf{m}} \left[\sum_{\ell=0}^{p} a_K(\ell) \left(\frac{\varepsilon_K}{\varepsilon_J}\right)^{\ell} \frac{\phi_K^{p-\ell}}{(p-\ell)!} \right] \varepsilon_J^p F^{(p)}(x,\vartheta_J) + R(x,K) + \varepsilon_J^{2\mathbf{m}} \max_{0\leqslant\ell\leqslant 2\mathbf{m}} \left| a_K(\ell) \left(\frac{\varepsilon_K}{\varepsilon_J}\right)^{\ell} \right| o(1).$$

Add up over the children K of J to obtain (27) for J, that is

$$F(x,J) = \sum_{p=0}^{2m} a_J(p) \varepsilon_J^p F^{(p)}(x,\vartheta_J) + R(x,J),$$

with $a_J(p) = \sum_{K\in\mathrm{Child}(J)} \sum_{\ell=0}^{p} a_K(\ell) \left(\frac{\varepsilon_K}{\varepsilon_J}\right)^{\ell} \frac{\phi_K^{p-\ell}}{(p-\ell)!}$ and

$$R(x,J) = \sum_{K\in\mathrm{Child}(J)} \left[R(x,K) + \varepsilon_J^{2\mathbf{m}} \max_{0\leqslant\ell\leqslant 2\mathbf{m}} \left| a_K(\ell) \left(\frac{\varepsilon_K}{\varepsilon_J}\right)^{\ell} \right| o(1) \right]. \tag{29}$$

We have now to prove the estimates (a), (b), (c) and (d) for the defined coefficients $a_J(p)$ and remainder $R(x,J)$. Keep in mind that these estimates are assumed to be true for the children K and set for short

$$M_{p,K} := \max_{0\leqslant\ell\leqslant p} \left| a_K(\ell) \left(\frac{\varepsilon_K}{\varepsilon_J}\right)^{\ell} \right|.$$

Proof of (a) *for* J. It is immediate from the definition of $a_J(p)$ that

$$a_J(0) = \sum_{K\in\mathrm{Child}(J)} a_K(0) = \sum_{K\in\mathrm{Child}(J)} \varpi_K = \varpi_J,$$

and, using (a) for K together with $\varepsilon_K \leqslant \varepsilon_J$, we get

$$|a_J(p)| \preccurlyeq \max_{K\in\mathrm{Child}(J)} M_{p,K} \preccurlyeq 1. \tag{30}$$

Proof of (b) *for* J. It's enough to establish

$$\max_{|J|\leqslant p\leqslant 2\mathbf{m}} |a_J(p)| \preccurlyeq \max_{K\in\mathrm{Child}(J)} M_{|K|-1,K} \asymp \max_{0\leqslant p<|J|} |a_J(p)|. \tag{31}$$

To prove the l.h.s. of (31), note from (30) that $|a_J(p)| \preccurlyeq \max_{K\in\mathrm{Child}(J)} M_{p,K}$. Moreover, for all $p \geqslant |K|$,

$$M_{p,K} \leqslant M_{|K|-1,K} + \max_{|K|\leqslant\ell\leqslant p} \left| a_K(\ell) \left(\frac{\varepsilon_K}{\varepsilon_J}\right)^{\ell} \right| \leqslant M_{|K|-1,K} + \|a_K\| \left(\frac{\varepsilon_K}{\varepsilon_J}\right)^{|K|}$$

and we have $\|a_K\| \asymp \max_{0\leqslant\ell<|K|} |a_K(\ell)|$ by (b) so that, even for $p < |K|$,

$$M_{p,K} \preccurlyeq \left(1 + \frac{\varepsilon_K}{\varepsilon_J}\right) M_{|K|-1,K} \preccurlyeq M_{|K|-1,K}. \tag{32}$$

Taking the supremum over $K \in \mathrm{Child}(J)$ and over p give the l.h.s. of (31).

To prove the r.h.s. of (31), write $a_J(p) = a_J^{(1)}(p) + a_J^{(2)}(p)$ with

$$\begin{aligned} a_J^{(1)}(p) &= \sum_{K\in\mathrm{Child}(J)} \sum_{\ell=0}^{|K|-1} a_K(\ell) \left(\frac{\varepsilon_K}{\varepsilon_J}\right)^{\ell} \frac{\phi_K^{p-\ell}}{(p-\ell)!} \mathbf{1}_{p\geqslant\ell}, \\ a_J^{(2)}(p) &= \sum_{K\in\mathrm{Child}(J)} \sum_{\ell=|K|}^{p} a_K(\ell) \left(\frac{\varepsilon_K}{\varepsilon_J}\right)^{\ell} \frac{\phi_K^{p-\ell}}{(p-\ell)!} \mathbf{1}_{p\geqslant\ell}. \end{aligned}$$

Note that $\{\phi_K\}_{K\in\mathrm{Child}(J)}$ is ε-separated since $|\phi_K-\phi_{K'}| = |\vartheta_K-\vartheta_{K'}|/\varepsilon_J \asymp 1$ for $K \neq K'$. Set $\lambda_{K,\ell} = a_K(\ell)(\varepsilon_K/\varepsilon_J)^\ell$ and apply Corollary D.2 of Heinrich and Kahn (2015) to $A(\{\phi_K\}_{K\in\mathrm{Child}(J)})$ and $\Lambda = (\lambda_{K,\ell})_{K\in\mathrm{Child}(J),\,0\leqslant\ell<|K|}$:

$$\max_{0\leqslant p<|J|} |a_J^{(1)}(p)| \asymp \max_{\substack{K\in\mathrm{Child}(J)\\ 0\leqslant\ell<|K|}} \left| a_K(\ell)\left(\frac{\varepsilon_K}{\varepsilon_J}\right)^\ell \right| = \max_{K\in\mathrm{Child}(J)} M_{|K|-1,K}, \tag{33}$$

which is the r.h.s. of (31) with $a_J^{(1)}(\cdot)$ instead of $a_J(\cdot)$.

We show now that the $|a_J^{(2)}(p)|$ are in fact negligible so that the r.h.s. of (31) will follows. Indeed, easy bounds on $a_J^{(2)}(p)$ yield

$$\max_{0\leqslant p<|J|} |a_J^{(2)}(p)| \preccurlyeq \max_{K\in\mathrm{Child}(J)} \left[\|a_K\| \left(\frac{\varepsilon_K}{\varepsilon_J}\right)^{|K|} \right],$$

whereas, as a by-product of (33), using $\|a_K\| \asymp \max_{0\leqslant\ell<|K|} |a_K(\ell)|$,

$$\max_{0\leqslant p<|J|} |a_J^{(1)}(p)| \succcurlyeq \max_{K\in\mathrm{Child}(J)} \left[\|a_K\| \left(\frac{\varepsilon_K}{\varepsilon_J}\right)^{|K|-1} \right].$$

Proof of (c) *for* J. From the r.h.s. of (31), and (a) for K, we deduce

$$\|a_J\| \succcurlyeq \max_{K\in\mathrm{Child}(J)} \left[\|a_K\| \left(\frac{\varepsilon_K}{\varepsilon_J}\right)^{|K|-1} \right] \vee \max_{K\in\mathrm{Child}(J)} |\varpi_K|. \tag{34}$$

Here we used $M_{|K|-1,K} \geqslant M_{0,K} = |\varpi_K|$. Combining (34) with (c) for children K gives

$$\|a_J\| \succcurlyeq \max_{K\in\mathrm{Child}(J)} \max_{F\in\mathrm{Desc}(K)} \left[|\pi_F| \left(\frac{\varepsilon_{F^\uparrow}}{\varepsilon_J}\right)^{|K|-1} \right] \vee \max_{K\in\mathrm{Child}(J)} |\varpi_K|.$$

Now, bound the exponent $|K|$ by $|J|$ to derive (c) for J.

Proof of (d) *for* J. Split (29) as $R(x,J) = R^{(1)}(x,J) + R^{(2)}(x,J)$ with

$$\begin{aligned} R^{(1)}(x,J) &= \sum_{K\in\mathrm{Child}(J)} R(x,K), \\ R^{(2)}(x,J) &= \varepsilon_J^{2\mathbf{m}} \sum_{K\in\mathrm{Child}(J)} M_{2\mathbf{m},K}\, o(1). \end{aligned}$$

Note that (31) and (32) give $\max_{K\in\mathrm{Child}(J)} M_{2\mathbf{m},K} \asymp \|a_J\|$; moreover,

$$\left\|R^{(1)}(\cdot,J)\right\|_\infty \preccurlyeq \max_{K\in\mathrm{Child}(J)} \left[o\left(\|a_K\|\,\varepsilon_K^{2\mathbf{m}}\right)\right]$$

by assumption (d) for K, so that by triangle inequality,

$$\|R(\cdot,J)\|_\infty \preccurlyeq \varepsilon_J^{2\mathbf{m}} \left\{ \max_{K\in\mathrm{Child}(J)} \left[o\left(\|a_K\| \left(\frac{\varepsilon_K}{\varepsilon_J}\right)^{2\mathbf{m}} \right)\right] + \|a_J\|\, o(1) \right\}.$$

By (34), $\|a_J\|$ dominates $\|a_K\|\,(\varepsilon_K/\varepsilon_J)^{2\mathbf{m}}$ and thus (d) follows for J. □

Concluding the proof of (23). We shall show that

$$\left\|F(\cdot,G_n) - F(\cdot,G_n')\right\|_\infty \succcurlyeq \max_{J\in\mathrm{Desc}(J_r)} |\varpi_J| \varepsilon_{J^\uparrow}^{2d_0+1}. \tag{35}$$

Recall that $F(x,G_n) - F(x,G_n') = F(x,J_r)$ and distinguish two cases:

Case $\varepsilon_{J_r} \to 0$**.** All the $\vartheta_{j,n}$'s converge to a single support point of G_0 so that $m_0 = 1$. Apply directly Lemma 7.4 to the root node $J := J_r$:

$$F(x,J) = \sum_{p=0}^{2\mathbf{m}} a_J(p)\varepsilon_J^p F^{(p)}(x,\vartheta_J) + R(x,J),$$

so that by the triangle's inequality, Proposition 2.3 and (d) for J,

$$\|F(\cdot,J)\|_\infty \succcurlyeq \max_{0\leqslant p\leqslant 2\mathbf{m}} \left|a_J(p)\varepsilon_J^p\right| - o\left(\|a_J\|\,\varepsilon_J^{2\mathbf{m}}\right).$$

By (b), the optimal p is at most $|J|-1$, and since $|J|\leqslant 2\mathbf{m}$, we get

$$\|F(\cdot,J)\|_\infty \succcurlyeq \|a_J\|\,\varepsilon_J^{|J|-1}.$$

Now, the estimate (c) for J yields further

$$\|F(\cdot,J)\|_\infty \succcurlyeq \max_{K\in\mathrm{Desc}(J)} |\varpi_K| \varepsilon_{K^\uparrow}^{2\mathbf{m}-1}.$$

But this estimate is nothing else than (35) since $m_0 = \mathbf{m} - d_0$ is one.

Case $\varepsilon_{J_r} \equiv 1$**.** This case means either there are more than one support point in the limit G_0 ($m_0 > 1$) or there is only one support point for G_0 but with possible sequences $\theta_{j,n}$ converging to other points (vanishing weights $\varpi_{j,n}$ may exist).

Here, all the $\varepsilon_{J_r}^p$'s are of the same order (actually identical), so the scheme used for the case when $\varepsilon_{J_r} \to 0$ does not work. It works however for the children J of J_r:

$$\begin{aligned} F(x, J_r) &= \sum_{J \in \mathrm{Child}(J_r)} F(x, J) \\ &= \sum_{J \in \mathrm{Child}(J_r)} \sum_{p=0}^{2\mathbf{m}} a_J(p) \varepsilon_J^p F^{(p)}(x, \vartheta_J) + \sum_{J \in \mathrm{Child}(J_r)} R(x, J), \end{aligned}$$

so that by the triangle's inequality, Proposition 2.3 and (d) for J,

$$\|F(\cdot, J_r)\|_\infty \succcurlyeq \max_{J \in \mathrm{Child}(J_r)} \max_{0 \leqslant p \leqslant 2\mathbf{m}} \left| a_J(p) \varepsilon_J^p \right| - \max_{J \in \mathrm{Child}(J_r)} o\left(\|a_J\| \, \varepsilon_J^{2\mathbf{m}} \right).$$

The optimal p is at most $|J| - 1$ by (b), so that $\|F(\cdot, J_r)\|_\infty$ dominates $\max_{J \in \mathrm{Child}(J_r)} \|a_J\| \, \varepsilon_J^{|J|-1}$, whereas for $p = 0$ we get that $\|F(\cdot, J_r)\|_\infty$ dominates $\max_{J \in \mathrm{Child}(J_r)} |\varpi_J|$, by (a). Together with $\varepsilon_{J\uparrow} \equiv 1$ and (c) for J, we obtain

$$\|F(\cdot, J_r)\|_\infty \succcurlyeq \max_{J \in \mathrm{Child}(J_r)} \max_{K \in \mathrm{Desc}(J) \cup \{J\}} |\varpi_K| \varepsilon_{K\uparrow}^{|J|-1}$$

which is nothing else than

$$\left\| F(\cdot, G_n) - F(\cdot, G_n') \right\|_\infty \succcurlyeq \max_{J \in \mathrm{Desc}(J_r)} |\varpi_J| \varepsilon_{J\uparrow}^{|J|-1}. \tag{36}$$

Now, note that a descendant J of J_r of maximal cardinality must be a child of J_r, call it $J_\star$. Since G_n and G_n' converge to $G_0 \in \mathcal{G}_{m_0}$, the root J_r has at least m_0 children, each of them containing at least two points. Thus, we have

$$|J_r| \geqslant |J_\star| + 2(m_0 - 1).$$

Since $|J_r|$ is at most $2\mathbf{m}$, we deduce further that $|J_\star| \leqslant 2\mathbf{m} - 2m_0 + 2$. From (36), we finally arrive at (35), exacly as in the case $\varepsilon_{J_r} \to 0$.

Now, recall that $q = 2\mathbf{m} - 2m_0 + 1$. Lemma 7.3 together with (35) ensure that, whatever the case, $\varepsilon_{J_r} \to 0$ or $\varepsilon_{J_r} \equiv 1$,

$$\left\| F(\cdot, G_n) - F(\cdot, G_n') \right\|_\infty \succcurlyeq W_q(G_n, G_n')^q$$

which is the stated contradiction (23).

7.2. *From local to global: how* (18) *implies* (19). We have to show that, under Assumption B($2\mathbf{m}$), for $r = 2\mathbf{m} - 1$,

$$L := \inf_{G \neq G' \in \mathcal{G}_{\leqslant \mathbf{m}}} \frac{\|F(\cdot, G) - F(\cdot, G')\|_\infty}{W_r(G, G')^r} > 0.$$

From the definition of L, we can select mixing distributions G_n and G'_n in $\mathcal{G}_{\leqslant \mathbf{m}}$ such that $\|F(\cdot, G_n) - F(\cdot, G'_n)\|_\infty / W_r(G_n, G'_n)^r$ converges to L. Since the set $\mathcal{G}_{\leqslant \mathbf{m}} \times \mathcal{G}_{\leqslant \mathbf{m}}$ is compact, we can assume that (G_n, G'_n) converges to some limit (G_∞, G'_∞). Set $w = W_r(G_\infty, G'_\infty)$.

Case $w > 0$**.** This case does not depend on (18). By identifiability, there is $x_0 \in \mathbb{R}$ such that $\Delta_0 := |F(x_0, G_\infty) - F(x_0, G'_\infty)| > 0$. Then, for all n,

$$\frac{\|F(\cdot, G_n) - F(\cdot, G'_n)\|_\infty}{W_r(G_n, G'_n)^r} \geqslant \frac{|F(x_0, G_n) - F(x_0, G'_n)|}{W_r(G_n, G'_n)^r}. \tag{37}$$

The numerator of the r.h.s. of (37) goes to Δ_0 by the triangle's inequality and since the function $\theta \mapsto F(x_0, \theta)$ is Lipschitz w.r.t. the metric W_1 and thus also w.r.t. W_r. As a consequence, we get $L \geqslant \Delta_0 / w^r > 0$.

Case $w = 0$**.** Consider (18) with $G_0 = G_\infty$. For n larger than some n_0, all $W_q(G_n, G_0)$ and $W_q(G'_n, G_0)$ are less than ε so that by (18),

$$\inf_{n \geqslant n_0} \frac{\|F(\cdot, G_n) - F(\cdot, G'_n)\|_\infty}{W_q(G_n, G'_n)^q} > \delta.$$

Since we have $W_q(\cdot, \cdot)^q \mathrm{Diam}(\Theta)^{r-q} \geqslant W_r(\cdot, \cdot)^r$ for $r \geqslant q$, we get

$$\inf_{n \geqslant n_0} \frac{\|F(\cdot, G_n) - F(\cdot, G'_n)\|_\infty}{W_r(G_n, G'_n)^r} > \frac{\delta}{\mathrm{Diam}(\Theta)^{r-q}}$$

which gives $L \geqslant \delta / \mathrm{Diam}(\Theta)^{r-q}$ in the limit and (19) in that case.

7.3. *Completing the proof of Theorem 6.3: the easy cases* (20) *and* (21). For the proof of (20), we can simply make use of Theorem 3.1 of Ho and Nguyen (2015). Alternatively, a detailed proof with our notations is available in the supplemental part (Heinrich and Kahn, 2015, B.2).

For the proof of (21), we can follow the proof of Chen (1995, Lemma 2) which holds here, because the γ_j defined in his paper are all non-negative, and at least one is nonzero.

Acknowledgements. We thank Sébastien Gadat for numerous suggestions that have greatly improved the presentation of the paper, and Élisabeth Gassiat for helpful discussions.

This work was supported in part by the Labex CEMPI (ANR-11-LABX-0007-01).

SUPPLEMENTARY MATERIAL

Auxiliary results and technical details
(doi: 10.1214/00-AOASXXXXSUPP; .pdf). This supplemental part gathers some proof details on some assertions given in the paper.

References.

Bontemps, D. and Gadat, S. (2014). Bayesian methods for the shape invariant model. *Electron. J. Stat.* **8** 1522–1568.

Caillerie, C., Chazal, F., Dedecker, J. and Michel, B. (2013). Deconvolution for the Wasserstein metric and geometric inference. In *Geometric Science of Information* 561–568. Springer.

Chen, J. (1995). Optimal Rate of Convergence for Finite Mixture Models. *The Annals of Statistics* **23** 221-233.

Dacunha-Castelle, D. and Gassiat, E. (1997). The estimation of the order of a mixture model. *Bernoulli* 279–299.

Dedecker, J. and Michel, B. (2013). Minimax rates of convergence for Wasserstein deconvolution with supersmooth errors in any dimension. *Journal of Multivariate Analysis* **122** 278–291.

Deely, J. J. and Kruse, R. L. (1968). Construction of Sequences Estimating the Mixing Distribution. *The Annals of Mathematical Statistics* **39** 286–288.

Dudley, R. M. (2002). *Real analysis and probability. Cambridge Studies in Advanced Mathematics* **74**. Cambridge University Press, Cambridge Revised reprint of the 1989 original.

Fan, J. (1991). On the optimal rates of convergence for nonparametric deconvolution problems. *The Annals of Statistics* 1257–1272.

Gassiat, E. and van Handel, R. (2013). Consistent order estimation and minimal penalties. *IEEE Trans. Inform. Theory* **59** 1115–1128.

Genovese, C. R. and Wasserman, L. (2000). Rates of convergence for the Gaussian mixture sieve. *Ann. Statist.* **28** 1105–1127.

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.

Hájek, J. (1972). Local asymptotic minimax and admissibility in estimation. In *Proceedings of the Sixth Berkeley Symposium on Mathematical Statistics and Probability (Univ. California, Berkeley, Calif., 1970/1971), Vol. I: Theory of statistics* 175–194. Univ. California Press, Berkeley, Calif.

Heinrich, P. and Kahn, J. (2015). Supplement to 'Minimax rates for finite mixture estimation': Auxiliary results and technical details.

Ho, N. and Nguyen, X. (2015). Identifiability and optimal rates of convergence for parameters of multiple types in finite mixtures. *arXiv preprint arXiv:1501.02497*.

Ho, N. and Nguyen, X. (2016). On strong identifiability and convergence rates of parameter estimation in finite mixtures. *Electron. J. Stat.* **10** 271–307. MR3466183

Ho, N., Nguyen, X. et al. (2016). Convergence rates of parameter estimation for some weakly identifiable finite mixtures. *The Annals of Statistics* **44** 2726–2755.

Holzmann, H., Munk, A. and Stratmann, B. (2004). Identifiability of finite mixtures-with applications to circular distributions. *Sankhyā: The Indian Journal of Statistics* 440–449.

Ishwaran, H., James, L. F. and Sun, J. (2001). Bayesian model selection in finite mixtures by marginal density decompositions. *J. Amer. Statist. Assoc.* **96** 1316–1332.

Kuhn, M. A., Feigelson, E. D., Getman, K. V., Baddeley, A. J., Broos, P. S., Sills, A., Bate, M. R., Povich, M. S., Luhman, K. L., Busk, H. A. et al. (2014). The Spatial Structure of Young Stellar Clusters. I. Subclusters. *The Astrophysical Journal* **787** 107.

Le Cam, L. (1960). Locally asymptotically normal families of distributions. *Univ. California Publ. Statist.* **3** 37–98.

Le Cam, L. (1986). *Asymptotic Methods in Statistical Decision Theory. Springer Series in Statistics*. Springer New York.

Lindsay, B. G. (1989). Moment matrices: applications in mixtures. *The Annals of Statistics* **17** 722–740.

Liu, M. and Hancock, G. R. (2014). Unrestricted Mixture Models for Class Identification in Growth Mixture Modeling. *Educational and Psychological Measurement* **74** 557–584.

Martin, R. (2012). Convergence rate for predictive recursion estimation of finite mixtures. *Statistics & Probability Letters* **82** 378–384.

Massart, P. (1990). The tight constant in the Dvoretzky-Kiefer-Wolfowitz inequality. *Annals of probability* **18** 1269–1283.

McLachlan, G. and Peel, D. (2000). *Finite mixture models. Wiley Series in Probability and Statistics: Applied Probability and Statistics*. Wiley-Interscience, New York.

Nguyen, X. (2013). Convergence of latent mixing measures in finite and infinite mixture models. *Ann. Statist.* **41** 370–400.

Pearson, K. (1894). Contributions to the theory of mathematical evolution. *Philosophical Transactions of the Royal Society of London A* **185** 71–110.

Rousseau, J. and Mengersen, K. (2011). Asymptotic behaviour of the posterior distribution in overfitted mixture models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **73** 689–710.

Teh, Y. W. (2010). Dirichlet Process. In *Encyclopedia of Machine Learning* (C. Sammut and G. Webb, eds.) 280-287. Springer US.

Titterington, D. M., Smith, A. F. M. and Makov, U. E. (1985). *Statistical analysis of finite mixture distributions. Wiley Series in Probability and Mathematical Statistics: Applied Probability and Statistics*. John Wiley & Sons, Ltd., Chichester.

van de Geer, S. (1996). Rates of convergence for the maximum likelihood estimator in mixture models. *J. Nonparametr. Statist.* **6** 293–310.

van der Vaart, A. W. (1998). *Asymptotic statistics. Cambridge Series in Statistical and Probabilistic Mathematics* **3**. Cambridge University Press, Cambridge.

Yang, Y. (2005). Can the strengths of AIC and BIC be shared? A conflict between model indentification and regression estimation. *Biometrika* **92** 937–950.

Zhu, H.-T. and Zhang, H. (2004). Hypothesis testing in mixture regression models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **66** 3–16.

Zhu, H. and Zhang, H. (2006). Asymptotics for estimation and testing procedures under loss of identifiability. *Journal of Multivariate Analysis* **97** 19–45.

E-mail: philippe.heinrich@math.univ-lille1.fr E-mail: jonas.kahn@math.univ-lille1.fr