The Annals of Statistics
2001, Vol. 29, No. 1, 94–123

CONVEX MODELS, MLE AND MISSPECIFICATION[1]

BY VALENTIN PATILEA

Université d'Orléans

We analyze the asymptotic behavior of maximum likelihood estimators (MLE) in convex dominated models when the true distribution generating the independent data does not necessarily belong to the model. Inspired by the Hellinger distance and its properties, we introduce a family of divergences (contrast functions) which allow a unified treatment of well- and misspecified convex models. Convergence and rates of convergence of the MLE with respect to our divergences are obtained from inequalities satisfied by these divergences and results from empirical process theory (uniform laws of large numbers and maximal inequalities). As a particular case we recover existing results for Hellinger convergence of MLE in well-specified convex models. Four examples are considered: mixtures of discrete distributions, monotone densities, decreasing failure rate distributions and a finite-dimensional parametric model.

1. Introduction. Consider an i.i.d. sample $X_1, X_2, \dots,$ defined on a probability space $(\Omega, \mathscr{F}, \mathbf{P})$ and distributed according to Q, a probability measure on the sample space $(\mathscr{X}, A)$. It is assumed that Q has a density q with respect to some σ-finite measure μ. In order to estimate q we consider $\mathscr{P}$, a statistical model on $(\mathscr{X}, A)$ dominated by μ. In this paper we are interested in the case where $\mathscr{P}$ is a convex model, that is, a convex set of densities. Typically, a convex set of densities $\mathscr{P}$ is either written in a parametric form $\mathscr{P} = \{p_\theta, \theta \in \Theta\}$, with Θ a convex set of some linear space and $\theta \to p_\theta$ linear, or defined through restrictions on monotonicity, regularity, symmetry, *etc.* The most important class of convex sets of densities with linear parameterization is represented by mixture models. While any convex set of densities defined through restrictions can also be written as a mixture model, such parameterization may have less statistical meaning than the original conditions specifying the convex set of densities. For notational simplicity, even if $\mathscr{P}$ is given parametrically, we will most of the time omit the parameter in the following.

The true density q is not necessarily assumed to belong to the statistical model. Moreover, even if the model depends on the sample size, it may not approach q asymptotically. In other words we allow model $\mathscr{P}$ to be *misspecified* [see Dahlhaus and Wefelmeyer (1996) for a recent reference on misspecification]. The model $\mathscr{P}$ is supposed to be compatible with the observed data in the sense that it contains probability measures dominating Q, that is, there exists $p \in \mathscr{P}$ such that $p \cdot \mu$ dominates Q (where $p \cdot \mu$ denotes the probability measure having the density p with respect to μ).

Received December 1997; revised October 2000.

[1] Supported by PARC Grant 93/98-164 from the Belgian government.

AMS 2000 *subject classifications*. Primary 62A10; secondary 62G20, 62F12.

Key words and phrases. Convex models, maximum likelihood, misspecification, empirical process.

The estimation method we consider is maximum likelihood. We therefore assume that there exists $p_0 \in \mathscr{P}$, which we call the *(pseudo-)true density*, such that

$$\int \log \frac{p_0}{p}\, dQ \geq 0 \qquad \forall p \in \mathscr{P}; \tag{1.1}$$

the above expectation could be $+\infty$. Herein we use the following rule: we write "pseudo" in brackets when we consider both cases, well-specified and misspecified models; when we do not use brackets we refer only to misspecified models. Clearly, if $\mathscr{P}$ is well-specified, that is $q \in \mathscr{P}$, then q satisfies (1.1). When the model is misspecified, necessarily $P_0 = p_0 \cdot \mu$ dominates Q. If

$$\sup_{p\in\mathscr{P}} \int \log p\, dQ \quad \text{is finite} \tag{1.2}$$

and attained, then the (pseudo-)true density is defined as a density attaining the supremum. However, in general we do not impose (1.2). If the model is misspecified but there exists $p_0 \in \mathscr{P}$ such that

$$\int \log \frac{q}{p_0}\, dQ = \inf_{p\in\mathscr{P}} \int \log \frac{q}{p}\, dQ < \infty,$$

then p_0 satisfies (1.1). In this case p_0 can be interpreted as the element of $\mathscr{P}$ which is the closest, in the sense of Kullback-Leibler divergence, to the true density q. If $\mathscr{P}$ is given in a parametric form, then $\theta_0 \in \Theta$ with the property $p_{\theta_0} = p_0$ will be called a *(pseudo-)true parameter*. Pfanzagl (1990) gives some mild sufficient conditions for the existence of a pseudo-true parameter.

If there exists $p_0 \in \mathscr{P}$ satisfying (1.1), the fact that $\mathscr{P}$ is a convex model ensures the uniqueness of p_0. More precisely, if p_0 and p_0' satisfy (1.1), then $p_0 = p_0'$ Q-a.s. If the model is given in a parametric form, in general, we do not necessarily have the uniqueness of the (pseudo-)true parameter.

We assume that $\mathscr{P}$ is such that, almost surely, there exists a maximum likelihood estimator (MLE), that is, there exists $\hat{p} \in \mathscr{P}$ satisfying

$$\int \log \hat{p}\, dQ_n = \sup_{p\in\mathscr{P}} \int \log p\, dQ_n < \infty, \tag{1.3}$$

where Q_n denotes the empirical distribution built from the first n observations. More generally, we could consider η_n-MLE defined such that $\int \log \hat{p}\, dQ_n \geq \sup_{p\in\mathscr{P}} \int \log p\, dQ_n - \eta_n$, where $\eta_n \to 0$ as $n \to \infty$. All the results of this paper can be adapted to η_n-MLE. For the sake of simplicity we assume (1.3). If the model is given in a parametric form, the parameter value $\hat{\theta} \in \Theta$ such that $p_{\hat{\theta}} = \hat{p}$ will be also called a MLE.

In this paper we analyze the convergence of a MLE $\hat{p}$ (or $p_{\hat{\theta}}$) towards the (pseudo-)true density. MLE under misspecification has been considered, amongst others, by Huber (1967) and Pfanzagl (1969, 1990). However, the approach we consider herein is inspired by the Hellinger convergence results for MLE in well-specified models [see van de Geer (1993, 2000)]. Recall that the Hellinger distance between two densities p_1 and p_2 is defined as

$h(p_1, p_2) = [(1/2)\int(\sqrt{p_1} - \sqrt{p_2})^2 d\mu]^{1/2}$ and if $P_2 = p_2 \cdot \mu$ dominates $p_1 \cdot \mu$, then

$$h(p_1, p_2) = \left[\frac{1}{2}\int\left(\sqrt{\frac{p_1}{p_2}} - 1\right)^2 dP_2\right]^{1/2}. \tag{1.4}$$

Typically, if the true density belongs to the model, Hellinger convergence results can be obtained using two types of ingredients: (i) "basic inequalities" involving, on one side, the squared Hellinger distance between the estimator and the true density and, on the other side, an empirical process; and (ii) the behavior of the increments of the empirical process derived from entropy calculations. The Hellinger metric appears to be a convenient tool for investigating the properties of the ML estimators in infinite-dimensional models containing the true density. Nevertheless, it cannot be used in the same manner under misspecification since, in general, the usual basic inequalities do not hold. A quick way to understand the difficulty is to remark that when the model is "wrong", the true density no longer appears in the Hellinger distance between the MLE and its limit, the pseudo-true density.

Inspired by (1.4), we propose the divergence

$$h_0(p, p_0) = \left[\frac{1}{2}\int\left(\sqrt{\frac{p}{p_0}} - 1\right)^2 dQ\right]^{1/2},$$

where p_0 is the pseudo-true density and Q is the true distribution ($p_0 \neq q = dQ/d\mu$), as a natural substitute of the Hellinger distance between p and p_0 in the case of misspecification; see also Patilea (1997). Note that $h_0(p, p_0)$ coincides with the Hellinger distance $h(p, p_0)$ when the model contains the true density (i.e., $p_0 = q$) and $p_0 \cdot \mu$ dominates $p \cdot \mu$. We show in this paper that the divergence h_0 allows a unified study of the asymptotics of the MLE for well- and misspecified convex models. While preparing a revision for this paper, the Editor drew our attention to the fact that van de Geer (2000) also studies misspecified models. In this independent work we complete and extend her Lemma 10.14. We derive the properties of h_0 by considering it to be an element of the family of divergences h_α, $\alpha \in [0, 1)$ defined as follows: if p_1 and p_2 are nonnegative measurable functions, then

$$h_\alpha^2(p_1, p_2) = \frac{1}{2}\int_{\alpha p_1 + (1-\alpha)p_2 > 0}\left(\sqrt{\frac{p_1}{\alpha p_1 + (1-\alpha)p_2}} - 1\right)^2 dQ. \tag{1.5}$$

The framework we consider below is slightly more general than the usual framework of convex models: $\mathscr{P}$ is called a convex model if $\mathscr{P}$ is a set of nonnegative measurable functions (not necessarily densities), either specified as a convex set of functions, or given in a parametric form $\mathscr{P} = \{p_\theta, \theta \in \Theta\}$ with $\theta \to p_\theta$ concave. Clearly, if for all $\theta \in \Theta$, p_θ integrates to one, then $\theta \to p_\theta$ is necessarily linear. By abuse, we call the elements of $\mathscr{P}$ densities and p_0, $\hat{p}$ (pseudo-)true density and MLE, respectively. Again it seems that the convexity assumption plays a crucial role in the asymptotic study of MLE or, more generally, of M-estimators [see, e.g., Koltchinskii (1997) for a list of references on asymptotic results in convex models]. In particular, the convexity of a model in a parametric form allows little use to be made of the parameterization and thus avoids possible difficulties associated with it such as, for instance, the identification problem.

The paper is organized as follows. In Section 2 we study the properties of the family of divergences h_α, $\alpha \in [0, 1)$. In Section 3, using "basic inequalities" satisfies h_α and a uniform law of large numbers obtained under entropy conditions, we extend some results of van de Geer (1993) [see also van de Geer (2000), Section 4.1] and we prove h_α-convergence for ML estimators. When the model has a parametric form we show that, under some mild conditions, h_α-convergence is equivalent to convergence in the parameter space. Rates of convergence in probability for MLE are obtained in Section 4 as counterparts of the results of van de Geer [(2000), Chapter 7] using the behavior of an empirical process indexed by a class of transformations of the elements in $\mathscr{P}$. Moreover, under some mild conditions, we show that when $h_\alpha(p, p_0)$ is close to zero, $\int_{p>0} \log(p_0/p)\, dQ$ has almost the order of $h_\alpha^2(p, p_0)$. This is an extension of a result of Wong and Shen (1995) proving that the square of the Hellinger distance almost dominates the Kullback-Leibler contrast. In Section 5 we present four examples of convex models: mixtures of discrete distributions, monotone densities, decreasing failure rate distributions and a finite-dimensional model. For each model we give some properties of the pseudo-true density and we deduce the rates of convergence for MLE.

2. A divergence for misspecified convex models. As announced in the introduction, the properties of h_0 are derived from those of the family of divergences h_α, $\alpha \in [0, 1)$ defined in (1.5). Hereafter, we use the notation

$$p_\alpha = \alpha p + (1 - \alpha)\, p_0.$$

The rationale for introducing the convex combination p_α in the definition of $h_\alpha(p, p_0)$ is to avoid the difficulty produced by small values of p_0. Finally, we show that $h_0(p, p_0)$ is not much affected by the possible unboundedness of p/p_0. Nevertheless, studying h_α, $\alpha \in (0, 1)$ seems useful for better understanding the properties of h_0 and the behavior of the MLE under misspecification. Most of the properties of h_α are based on the following lemma.

LEMMA 2.1. *For any* $\alpha \in [0, 1)$ *and* $p \in \mathscr{P}$,

$$\int \frac{p}{p_\alpha}\, dQ \le 1 \le \int_{p>0} \frac{p_\alpha}{p}\, dQ \tag{2.1}$$

($\int_{p>0}(p_\alpha/p)\, dQ$ *could be* $+\infty$*). The second inequality is strict if* $p \neq p_0$ *on* $\{q > 0\}$.

PROOF. We obtain the first inequality as consequence of equation (2.2) in Pfanzagl (1990). Fix $\alpha \in [0, 1)$ and define $m(x, p) = p(x)/p_\alpha(x)$, $x \in \mathscr{X}$, $p \in \mathscr{P}$. It is easy to see that $p \to m(x, p)$ is concave for any $x \in \mathscr{X}$, provided that $\mathscr{P}$ is a convex model. Moreover, $(x, p) \to m(x, p)$ satisfies conditions (i) and (iii) of the lemma in Pfanzagl [(1990, page 1872] with (z, α) replaced by (x, p). Finally, note that equation (2.1) in Pfanzagl (1990) considered with $L(y) = \ln y$ is a direct consequence of the definition of p_0 [see (1.1) above] and the concavity of the logarithm function. For the second inequality in (2.1) above use Jensen's inequality. □

The previous lemma ensures that, for any $\alpha \in [0, 1)$, $h_\alpha(p, p_0)$, $p \in \mathscr{P}$ is uniformly bounded. We continue with some properties of h_α.

LEMMA 2.2. *For any $\alpha \in [0, 1)$ and $p \in \mathscr{P}$ we have*

$$0 \le h_\alpha^2(p, p_0) \le \int \left(1 - \sqrt{\frac{p}{p_\alpha}}\right) dQ. \tag{2.2}$$

Consequently, if $\hat{p}$ is a MLE, then

$$0 \le h_\alpha^2(\hat{p}, p_0) \le \int \left(\sqrt{\frac{\hat{p}}{\hat{p}_\alpha}} - 1\right) d(Q_n - Q), \tag{2.3}$$

where $\hat{p}_\alpha = \alpha \hat{p} + (1 - \alpha) p_0$.

PROOF. From Lemma 2.1 we have, for all $p \in \mathscr{P}$,

$$0 \le h_\alpha^2(p, p_0) = \frac{1}{2} \int \left(\frac{p}{p_\alpha} - 1\right) dQ - \int \left(\sqrt{\frac{p}{p_\alpha}} - 1\right) dQ \le \int \left(1 - \sqrt{\frac{p}{p_\alpha}}\right) dQ.$$

Moreover, from the definition of $\hat{p}$ and the convexity of the model we have that, for any β

$$0 \le \int \log \left(\frac{\hat{p}}{\hat{p}_\alpha}\right)^\beta dQ_n \le \int \left[\left(\frac{\hat{p}}{\hat{p}_\alpha}\right)^\beta - 1\right] dQ_n. \tag{2.4}$$

Take $\beta = 1/2$ and deduce (2.3). □

The "basic inequality" (2.3) provides an upper bound for the contrast, measured by h_α, between the MLE $\hat{p}$ and the (pseudo-)true density p_0. As a consequence, we can use the behavior of the empirical process indexed by the family $\{\sqrt{p/p_\alpha}\, \mathbb{1}_{\{q>0\}},\ p \in \mathscr{P}\}$ and deduce convergence and rates of convergence for $h_\alpha(\hat{p}, p_0)$ ($\mathbb{1}_A$ denotes the indicator function of the set A). The next lemma provides another interesting basic inequality for studying the asymptotics of MLE in possibly "wrong" models. Inequality (2.6) with $\alpha = 0$ is also proved by van de Geer [(2000), Lemma 10.14] in the case in which q/p_0 is bounded by a constant.

LEMMA 2.3. *Let $\alpha \in [0, 1)$. For any $p \in \mathscr{P}$,*

$$h_\alpha^2(p, p_0) \le \int \left(1 - \frac{p}{p_{\frac{1+\alpha}{2}}}\right) dQ = -\frac{1}{2} \int \frac{p - p_\alpha}{p_{\frac{1+\alpha}{2}}} dQ. \tag{2.5}$$

Consequently, if $\hat{p}$ *is a MLE, then*

$$h_\alpha^2(\hat{p}, p_0) \le \int \left(\frac{\hat{p}}{\hat{p}_{\frac{1+\alpha}{2}}} - 1 \right) d(Q_n - Q). \tag{2.6}$$

PROOF. For the first inequality write

$$-\frac{1}{2}\int \frac{p - p_\alpha}{p_{\frac{1+\alpha}{2}}}\, dQ = \frac{1}{2}\int \left(1 - \frac{p}{p_\alpha}\right) dQ + \frac{1}{2}\int \frac{(\sqrt{p} - \sqrt{p_\alpha})^2}{p_\alpha}\, \frac{(\sqrt{p} + \sqrt{p_\alpha})^2}{2\, p_{\frac{1+\alpha}{2}}}\, dQ.$$

As the first term on the right-hand side is nonnegative (see Lemma 2.1), it remains to prove that the last term dominates $h_\alpha^2(p, p_0)$. For this it suffices to remark that $\sqrt{p} + \sqrt{p_\alpha} \ge \sqrt{2\, p_{\frac{1+\alpha}{2}}}$. Finally, combine (2.5) and (2.4) in order to obtain (2.6). □

Note that the families $\{(p/p_{\frac{1+\alpha}{2}})\mathbb{1}_{\{q>0\}},\ p \in \mathscr{P}\}$, $\alpha \in [0, 1)$ are uniformly bounded by 2. Let us also remark that in the case $p_0 = q$ and $\alpha = 0$, inequality (2.6) is slightly weaker than the one proved in Lemma 4.5 of van de Geer (2000) where $h_0(\hat{p}, p_0)$ is replaced by the Hellinger distance $h(\hat{p}, p_0)$. This is because in well-specified models $h_0(\hat{p}, p_0) \le h(\hat{p}, p_0)$ and the equality holds if and only if $p_0 \cdot \mu$ dominates $\hat{p} \cdot \mu$. Nevertheless, this condition is satisfied in the common examples of convex models. Next, we show that, modulo a scaling factor, $h_\alpha^2(p, p_0)$ is dominated by the Kullback-Leibler contrast between p and p_0.

LEMMA 2.4. *Let* $\alpha \in [0, 1)$. *If* $p \in \mathscr{P}$, *then*

$$h_\alpha^2(p, p_0) \le \frac{1}{2}\int \log \frac{p_\alpha}{p} dQ \le \frac{1}{2}\int \log \frac{p_0}{p} dQ \le \frac{1}{2(1-\alpha)} \int \log \frac{p_\alpha}{p} dQ. \tag{2.7}$$

PROOF. From $\log x \ge (x-1)/x$, $\forall x > 0$, deduce that

$$\frac{1}{2}\int \log \frac{p_\alpha}{p} dQ \ge \int \frac{\sqrt{p_\alpha} - \sqrt{p}}{\sqrt{p_\alpha}} dQ = \int \left(1 - \sqrt{\frac{p}{p_\alpha}}\right) dQ.$$

The first inequality is then a consequence of (2.2). The other two inequalities are obvious. □

Note that setting $\alpha = 0$ in (2.7) yields

$$\frac{1}{2}\int \left(\sqrt{\frac{p}{p_0}} - 1\right)^2 dQ \le \frac{1}{2}\int \log \frac{p_0}{p}\, dQ,$$

which is an extension to our framework of the well-known inequality between the Hellinger distance and the Kullback-Leibler divergence. More precisely, we can write

$$\int \log \sqrt{\frac{p_0}{p}}\, dQ \ge \int \frac{\sqrt{p_0/p} - 1}{\sqrt{p_0/p}}\, dQ = \int \left(1 - \sqrt{\frac{p}{p_0}}\right) dQ, \tag{2.8}$$

and the last quantity is exactly $h^2(p, p_0)$, that is, the square of the Hellinger distance between p and p_0, provided that $p_0 = dQ/d\mu$ and $\mathscr{P}$ is a set, not necessarily convex, of densities ($\int p d\mu = 1$, for any $p \in \mathscr{P}$). When the model is "wrong" ($p_0 \neq dQ/d\mu$), the last quantity in (2.8) is not necessarily equal to $h_0^2(p, p_0)$. Nevertheless, the convexity of the model allows us to write

$$\int \left(1 - \sqrt{\frac{p}{p_0}}\right) dQ \geq h_0^2(p, p_0) = \frac{1}{2} \int \left(\sqrt{\frac{p}{p_0}} - 1\right)^2 dQ,$$

regardless of whether $p_0 = dQ/d\mu$. Moreover, for this last inequality the functions $p \in \mathscr{P}$ need not integrate to one.

In general, h_α, $\alpha \in [0, 1)$ is not a distance (e.g., h_α is not necessarily symmetric). However, h_α has other interesting properties which we present below. In particular, modulo a multiplicative factor, h_α, $\alpha \in (0, 1)$ is symmetric and satisfies the triangle inequality. It seems that h_0 does not necessarily share these properties.

LEMMA 2.5. *Let p_i, $i = 1, 2, 3$, be nonnegative measurable functions defined on $(\mathscr{X}, A)$ and $\alpha \in (0, 1)$.*

(a) *Let $\alpha' \in [0, 1)$ with $\alpha \geq \alpha'$. If $\alpha' = 0$ assume also that $p_2 \cdot \mu$ dominates Q. Then,*

$$h_\alpha(p_1, p_2) \leq h_{\alpha'}(p_1, p_2). \tag{2.9}$$

(b) *There exist C_1 and C_2, constants depending on α, such that*

$$h_\alpha(p_1, p_2) \leq C_1 h_\alpha(p_2, p_1)$$

and

$$h_\alpha(p_1, p_3) \leq C_2\,(h_\alpha(p_1, p_2) + h_\alpha(p_2, p_3)). \tag{2.10}$$

PROOF. See the Appendix.

Observe that $h_\alpha(p, p_0)$ can be included in a larger family of divergences suggested by Professor Rolin (private communication):

$$d_{s,\alpha}(p, p_0) = \frac{1}{4} \int \phi_s\left(\frac{p}{p_\alpha}\right) dQ, \qquad p \in \mathscr{P},$$

where, for $s > 1$, $\phi_s(z) = s(s-1)^{-1}[z - 1 - s(z^{1/s} - 1)]$, $z > 0$. We have $h_\alpha^2 = d_{2,\alpha}$. For instance, it is easy to extend inequality (2.3) to

$$0 \leq d_{s,\alpha}(\hat{p}, p_0) \leq \frac{s^2}{4(s-1)} \int \left(\left(\frac{\hat{p}}{\hat{p}_\alpha}\right)^{1/s} - 1\right) d(Q_n - Q).$$

However, in this paper we will stick to h_α.

Assume now that the model $\mathscr{P}$ is written in a parametric form. If $\hat{\theta}$ denotes a MLE, we look for conditions under which the h_α-convergence of $p_{\hat{\theta}}$ towards p_{θ_0} implies the convergence of $\hat{\theta}$ to θ_0, the (pseudo-)true parameter, in a topology of the parameter space Θ. For simplicity, we assume that θ_0 is identifiable, that is $p_\theta = p_{\theta_0}\ Q - a.s.$ implies $\theta = \theta_0$. We present some possible extensions after the next lemma. The following result is a version of Lemma 5.2 of van de Geer (1993) and shows that h_α, $\alpha \in [0, 1)$ is an appropriate tool for proving convergence in the parameter space when the parameterization satisfies some mild conditions. Note that the proof of the lemma below does not require convexity of the statistical model.

LEMMA 2.6. *Let $\mathscr{P} = \{p_\theta,\ \theta \in \Theta\}$ be a model* (*not necessarily convex*) *and* $\alpha \in [0, 1)$. *Assume that there exists a unique* (*pseudo-*)*true density p_{θ_0} and that θ_0 is identifiable. Moreover, $\Theta \subset \Theta^*$ with (Θ^*, τ) a first countable compact space. Suppose that, for any $x \in \mathscr{X}$, the map $\theta \to p_\theta(x) \geq 0$ is defined on Θ^* and is measurable. Moreover, assume that, for any $\theta \in \Theta^*$, the map $\theta' \to p_{\theta'}(x)$ is continuous at θ, for Q-almost all $x \in \mathscr{X}$* (*the exceptional set may depend on* θ). *Then, $h_\alpha(p_{\theta_n}, p_{\theta_0}) \to 0$ for some sequence $\{\theta_n\} \subset \Theta$ iff $\theta_n \to \theta_0$.*

PROOF. Fix $\theta \in \Theta^*$ with $\theta \neq \theta_0$. Let $V_m(\theta)$, $m \geq 1$, be a sequence of decreasing open neighborhoods of $\theta \in \Theta^*$, not containing θ_0. For any $m \geq 1$ define

$$v_{\theta,m} = ess \inf_{\theta' \in V_m(\theta)} \left| \sqrt{\frac{p_{\theta'}}{p_{\theta',\alpha}}} - 1 \right| \geq 0, \tag{2.11}$$

where $p_{\theta',\alpha} = \alpha p_{\theta'} + (1-\alpha)p_0$ and "$ess \inf$" stands for the essential infimum. Clearly, $v_{\theta,m}$, $m \geq 1$, is an increasing sequence of measurable functions, dominated, Q-almost everywhere, by $|\sqrt{p_\theta / p_{\theta,\alpha}} - 1|$. Moreover, for any m, the essential infimum $v_{\theta,m}$ is larger than the pointwise infimum

$$\inf_{\theta' \in V_m(\theta)} \left| \sqrt{\frac{p_{\theta'}(x)}{p_{\theta',\alpha}(x)}} - 1 \right|$$

which, for any $x \in \mathscr{X}$ such that the function $\theta' \to p_{\theta'}(x)$ is continuous at θ, grows to $|\sqrt{p_\theta(x)/p_{\theta,\alpha}(x)} - 1|$. From this and a monotone convergence argument we obtain

$$\lim_{m\to\infty} \int v_{\theta,m}^2 dQ = h_\alpha^2(p_\theta, p_{\theta_0}) > 0$$

[for $\theta \in \Theta^* \setminus \Theta$, $h_\alpha(p_\theta, p_{\theta_0})$ has the same definition as above].

Now fix some U an open neighborhood of θ_0 and note that U^c (its complement) is a compact set. For any $\theta \in U^c$ take V_θ an open neighborhood of θ such that

$$\int v_\theta^2 \, dQ > 0,$$

where v_θ is defined as in (2.11). We choose a finite subcover $V_{\theta_1}, \dots, V_{\theta_s}$ of U^c and we have

$$\inf_{\theta \in U^c} h_\alpha^2(p_\theta, p_{\theta_0}) \geq \min_{1 \leq i \leq s} \int v_{\theta_i}^2 dQ > 0.$$

Since U was arbitrary we obtain that $h_\alpha(p_{\theta_n}, p_{\theta_0}) \to 0$ implies $\theta_n \to \theta_0$. On the other hand, use (2.1) and deduce that $\int \sqrt{p_\theta / p_{\theta,\alpha}}\, dQ \leq 1$, $\theta \in \Theta$. Let $\theta_n \to \theta_0$. The continuity of the maps $\theta \to p_\theta(x)$ and Fatou's lemma then imply that $\int \sqrt{p_{\theta_n} / p_{\theta_n,\alpha}}\, dQ \to 1$. Finally, inequality (2.2) yields $h_\alpha(p_{\theta_n}, p_{\theta_0}) \to 0$. □

The previous result can be extended to the case where the (pseudo-)true parameter is not identifiable. Define $\Theta_0 = \{\theta \in \Theta,\ p_\theta = p_{\theta_0}\ Q\text{-a.s.}\}$. The continuity of p_θ with respect to the parameter implies that Θ_0 is a closed subset of Θ^*. We obtain that $h_\alpha(p_{\theta_n}, p_{\theta_0}) \to 0$ ensures $\theta_n \to \Theta_0$, that is, for any U open set containing Θ_0, $\theta_n \in U$ for n sufficiently large.

3. Convergence of MLE. The next step is to find conditions under which $h_\alpha(\hat{p}, p_0)$ [or $h_\alpha(p_{\hat{\theta}}, p_{\theta_0})$] decreases to zero. Such convergence can be obtained from the inequalities deduced above, between $h_\alpha(\hat{p}, p_0)$ and certain empirical processes, and strong uniform laws of large numbers (SULLN). Here, and in the following, we assume that there are no measurability problems with the quantities we manipulate. For instance we may suppose that the classes of functions we consider below are permissible in the sense of Pollard (1984). For a more general approach we could follow van der Vaart and Wellner (1996). Let us briefly recall some basic facts.

If $\mathscr{G}$ is a family of measurable real-valued functions defined on $(\mathscr{X}, A)$ and P is a probability on this space, $H_1(\delta, \mathscr{G}, P)$ denotes the δ-entropy of $\mathscr{G}$ with respect to the $L_1(P)$-norm, that is, $H_1(\delta, \mathscr{G}, P) = \log N_1(\delta, \mathscr{G}, P)$ with

$$N_1(\delta, \mathscr{G}, P) = \min\left\{J;\ \exists\, g_1, \dots, g_J,\ \text{ such that}\ \forall g \in \mathscr{G},\ \exists\, g_j \text{ with} \int |g_j - g|\, dP \leq \delta\right\}.$$

Recall also that the envelope of $\mathscr{G}$ is defined as $G = \sup_{g \in \mathscr{G}} |g|$. The following result can be found in van de Geer (2000), Section 3.6 [see also Pollard (1984), Section II.5].

THEOREM 3.1 (SULLN). *Let $X_1, X_2, \dots$ i.i.d. with distribution Q on $(\mathscr{X}, A)$ and let Q_n denote the empirical distribution based on the first n observations. Assume the envelope condition $\int G\, dQ < \infty$ and suppose that $n^{-1} H_1(\delta, \mathscr{G}, Q_n) \to_{\mathbf{P}} 0$, for all $\delta > 0$. Then*

$$\sup_{g \in \mathscr{G}} \left| \int g\, d(Q_n - Q) \right| \to 0 \quad \textit{almost surely.}$$

For $\alpha \in [0, 1)$ and $p \in \mathscr{P}$, define $g_p = (p / p_{\frac{1+\alpha}{2}}) \mathbb{1}_{\{q>0\}}$ and $g'_p = \sqrt{p / p_\alpha}\, \mathbb{1}_{\{q>0\}}$. Consider the families $\mathscr{G}_\alpha = \{g_p,\ p \in \mathscr{P}\}$ and $\mathscr{G}'_\alpha = \{g'_p,\ p \in \mathscr{P}\}$ and let G_α and G'_α denote the corresponding envelopes.

PROPOSITION 3.2. (a) *If $\alpha \in [0, 1)$ and, for any $\delta > 0$*

$$n^{-1} H_1(\delta, \mathscr{G}_\alpha, Q_n) \to_{\mathbf{P}} 0, \tag{3.1}$$

then $h_\alpha(\hat{p}, p_0) \to 0$, almost surely.

(b) *If* $\alpha \in (0, 1)$ *and, for any* $\delta > 0$

$$n^{-1} H_1(\delta, \mathscr{G}'_\alpha, Q_n) \to_{\mathbf{P}} 0, \tag{3.2}$$

then $h_\alpha(\hat{p}, p_0) \to 0$, *almost surely. If* (3.2) *is satisfied for* $\alpha = 0$ *and* G'_0 *is integrable, then* $h_0(\hat{p}, p_0) \to 0$, *almost surely.*

Proof. (a) Clearly G_α is bounded, thus integrable. This, together with (3.1), ensures the SULLN for the family $\mathscr{G}$. Finally, use Lemma 2.3.
(b) Use Lemma 2.2 and the SULLN for the family $\mathscr{G}'_\alpha$. □

Many statistical models are written in a parametric form $\mathscr{P} = \{p_\theta,\ \theta \in \Theta\}$ with Θ a topological space that is a subset of a first countable compact space. Moreover, for any $\theta \in \Theta$, the map $\theta' \to p_{\theta'}(x)$ is continuous at θ, for μ-almost all $x \in \mathscr{X}$; the exceptional set may depend on θ [see, e.g., Wang (1985) and Pfanzagl (1988)]. It can be proved that in such cases conditions (3.1) or (3.2) are automatically fulfilled, and this without assuming a convex model. The following result can be found in van de Geer (1993), Lemma 5.1.

Lemma 3.3. *Let* $\Theta \subset \Theta^\star$ *where* $(\Theta^\star, \tau)$ *is a first countable compact space. Consider* $\{p_\theta,\ \theta \in \Theta^\star\}$ *a family of functions as in Lemma* 2.6. *Let* $g(\cdot; x) : [0, \infty) \to \mathbb{R}$, $x \in \mathscr{X}$, *be a family of continuous transformations. Denote* $g_\theta(x) = g(p_\theta(x); x)$ *and* $\tilde{\mathscr{G}} = \{g_\theta,\ \theta \in \Theta^*\}$ *and assume that* $\tilde{\mathscr{G}}$ *is a family of measurable, uniformly bounded real-valued functions defined on* $(\mathscr{X}, A)$. *Then* $n^{-1} H_1(\delta, \tilde{\mathscr{G}}, Q_n) \to_{\mathbf{P}} 0$.

The previous lemma and Proposition 3.2 show that establishing h_α-convergence of MLE in convex models given in a parametric form with the densities continuous in the parameter is an easy matter, regardless of whether the model is well-specified or not. Let us note this in the following corollary.

Corollary 3.4. *Assume that* $\Theta \subset \Theta^\star$ *with* $(\Theta^\star, \tau)$ *a first countable compact space and that* $\mathscr{P} = \{p_\theta,\ \theta \in \Theta\}$ *is a convex model. Moreover,* $\theta \to p_\theta$, $\theta \in \Theta^\star$, *are defined and satisfy the conditions of Lemma* 3.3. *Then, for any* $\alpha \in [0, 1)$, $h_\alpha(p_{\hat{\theta}}, p_0) \to 0$, *almost surely. If furthermore* θ_0 *is identifiable then* $\hat{\theta} \to \theta_0$, *almost surely.*

Proof. Define $g(y; x) = 2y[(1+\alpha)y + (1-\alpha)p_{\theta_0}(x)]^{-1}\, \mathbb{1}_{\{q>0\}}(x)$ and obtain the result from Proposition 3.2 for $\mathscr{G} = \mathscr{G}_\alpha$, $\alpha \in [0, 1)$, Lemma 3.3 and Lemma 2.6. □

Extensions to the case where θ_0 is not identifiable can be also considered (see the comments after Lemma 2.6). Note that the definition of the MLE $\hat{\theta}$ does not change, that is, it is still defined as the maximizer of the likelihood over Θ and not over $\Theta^\star$.

4. h_α-rates of convergence. Inequalities (2.3) and (2.6) can also be used to obtain the rates of convergence in probability for $h_\alpha(\hat{p}, p_0)$ and other interesting quantities in possibly misspecified convex models. For this purpose we use the approach presented in Chapter 7 of van de Geer (2000); see also van der Vaart and Wellner [(1996), Section 3.2]. The idea is to analyze the behavior of the empirical processes indexed by the families $\mathscr{G}'_\alpha = \{\sqrt{p/p_\alpha}\, \mathbb{1}_{\{q>0\}},\ p \in \mathscr{P}\}$ and $\mathscr{G}_\alpha = \{(p/p_{\frac{1+\alpha}{2}})\, \mathbb{1}_{\{q>0\}},\ p \in \mathscr{P}\}$, respectively, in $L_2(Q)$ neighborhoods of the function $\mathbb{1}_{\{q>0\}}$. We deduce rates of convergence from the entropy with bracketing of the families $\mathscr{G}'_\alpha$ and $\mathscr{G}_\alpha$.

For some (large) universal constant c_1 and $\mathscr{G}$ a family of real-valued functions, define

$$J_B(\delta, \mathscr{G}, Q) = \int_{\frac{\delta^2}{c_1}}^{\delta} H_B^{1/2}(u, \mathscr{G}, Q)\, du \vee \delta, \qquad \delta > 0,$$

with $H_B(\delta, \mathscr{G}, Q)$ the δ-entropy with bracketing with respect to the $L_2(Q)$-norm, that is $H_B(\delta, \mathscr{G}, Q) = \log N_B(\delta, \mathscr{G}, Q)$ where

$$N_B(\delta, \mathscr{G}, Q) = \min\left\{ J;\ \exists \{(g_j^L, g_j^U)\}_{j=1}^J \text{ such that } \forall g \in \mathscr{G},\ \exists\, g_j^L \le g \le g_j^U \right.$$
$$\left. \text{with } \int (g_j^L - g_j^U)^2 dQ \le \delta^2 \right\}.$$

Results on the empirical process [see, e.g., van de Geer (2000), Chapter 5] prove that the behavior of the increments at $g_0 \in \mathscr{G}$ of the empirical process indexed by $\mathscr{G}$ depends on J_B.

Note that small values of p_0 may cause difficulties in computing entropies of $\mathscr{G}_\alpha$ and $\mathscr{G}'_\alpha$. A possible remedy is to truncate the elements of the families $\mathscr{G}_\alpha$ and $\mathscr{G}'_\alpha$ [see also van de Geer (2000), Section 7.3]. For $\sigma \ge 0$, define

$$\mathscr{G}_\alpha^\sigma = \left\{ (p/p_{\frac{1+\alpha}{2}})\, \mathbb{1}_{\{q>0\}}\, \mathbb{1}_{\{p_0>\sigma\}},\ p \in \mathscr{P} \right\} \tag{4.1}$$

and

$$\mathscr{G}_\alpha^{'\sigma} = \{\sqrt{p/p_\alpha}\, \mathbb{1}_{\{q>0\}}\, \mathbb{1}_{\{p_0>\sigma\}},\ p \in \mathscr{P}\}.$$

After a little algebra, it can be proved that, for any $\alpha, \alpha' \in [0, 1)$, there exists $C = C(\alpha, \alpha') > 0$ such that

$$H_B(C\delta, \mathscr{G}_\alpha^\sigma, Q) \le H_B(\delta, \mathscr{G}_{\alpha'}^\sigma, Q), \qquad \delta > 0, \sigma \ge 0. \tag{4.2}$$

Moreover, for any $\alpha \in (0, 1)$

$$H_B(C'\delta, \mathscr{G}_0^\sigma, Q) \le H_B(\delta, \mathscr{G}_\alpha^{'\sigma}, Q), \qquad \delta > 0, \sigma \ge 0, \tag{4.3}$$

with $C' = C'(\alpha) > 0$ [see Patilea (1997) for the details]. The inequalities (4.2) and (4.3) indicate that for proving a general result on h_α-rates, it suffices to use the family $\mathscr{G}_0^\sigma$ and the inequality (2.6). The families $\mathscr{G}_\alpha^{'\sigma}$, $\alpha \in (0, 1)$ may serve for obtaining entropy bounds. Note that the entropy calculations in the case of misspecification are usually not more complicated than in the case of well-specified models [see, e.g., van der Vaart and Wellner (1996) for some general results on entropy bounds].

Before applying empirical process results for $\mathscr{G}_0^{\sigma}$, we should note that, for any $p \in \mathscr{P}$ ($\mathscr{P}$ not necessarily convex)

$$\left\| p/p_{1/2} - 1 \right\|_{L_2(Q)} \leq 2h_0(p, p_0). \tag{4.4}$$

Indeed, $p_{1/2} = (1/2)\, p + (1/2)\, p_0$ and thus

$$\left\| p/p_{1/2} - 1 \right\|^2_{L_2(Q)} = \frac{1}{4} \int \frac{\left(\sqrt{p} - \sqrt{p_0}\right)^2}{p_0} \, \frac{p_0}{p_{1/2}} \left(\frac{\sqrt{p} + \sqrt{p_0}}{\sqrt{p_{1/2}}} \right)^2 dQ.$$

Since $p_0/p_{1/2} \leq 2$ and $(\sqrt{p} + \sqrt{p_0})/\sqrt{p_{1/2}} \leq 2$, we obtain (4.4).

PROPOSITION 4.1. *Consider $\sigma(\delta) \geq 0$, $\delta > 0$ and a family $\mathscr{G} = \mathscr{G}_0^{\sigma(\delta)}$ or $\mathscr{G} = \mathscr{G}'^{\,\sigma(\delta)}_{\alpha'}$, $\alpha' \in (0, 1)$. Assume that Φ is a function of $\delta > 0$ satisfying*:

(i) $\Phi(\delta) \geq J_B(\delta, \mathscr{G}, Q)$;

(ii) $\Phi(\delta)/\delta^2$ *is non-increasing.*

Consider $\{\delta_n\}$ a sequence of positive real numbers such that

$$\delta_n^2 \geq Q\left(\{p_0 \leq \sigma(\delta_n)\}\right) \tag{4.5}$$

and

$$\sqrt{n}\, \delta_n^2 \geq \Phi(\delta_n).$$

Then, for any $\alpha \in [0, 1)$, $h_\alpha^2(\hat{p}, p_0) = O_{\mathbf{P}}(\delta_n^2)$.

PROOF. See the Appendix.

The proof of the proposition above relies on the uniform inequality proved in Theorem 5.11 of van de Geer (2000). Note that when $\sigma(\cdot) \equiv 0$ we can also obtain exponential bounds for $\mathbf{P}(h_\alpha(\hat{p}, p_0) > \delta_n)$. It seems difficult to express the condition (4.5) only in terms of q and Q because, in general, very little can be said about the sets $\{p_0 \leq \sigma\}$, with σ small, or about the behavior of q/p_0 on such sets. In the example presented in subsection 5.1, the properties of the model indicate a set $K_n(\sigma)$ such that $\{p_0 \leq \sigma\} \subset K_n(\sigma)$. Therefore, we will replace (4.5) by the (stronger) condition $\delta_n^2 \geq Q\left(K_n(\sigma(\delta_n))\right)$.

As a consequence of the h_α-rates of convergence we can obtain the rates of convergence for several quantities representing measures of the performances of the MLE. These quantities can be also used for further investigation of the properties of the MLE [see Corollary 4.4; see also (5.6) and the subsequent comments].

COROLLARY 4.2. *If the conditions of Proposition* 4.1 *hold, then, for any $\alpha \in [0, 1)$,*

$$\int \left(\hat{p}/\hat{p}_{\frac{1+\alpha}{2}} - 1 \right) dQ_n, \quad \int \left(\hat{p}/\hat{p}_{\frac{1+\alpha}{2}} - 1 \right) dQ \quad \text{and} \quad \int \log \hat{p}/\hat{p}_{\frac{1+\alpha}{2}} \, dQ_n$$

are of order $O_{\mathbf{P}}(\delta_n^2)$. Moreover, if $\alpha \in (0, 1)$, then $\int (\hat{p}/\hat{p}_\alpha - 1)^2\, dQ$ is also of order $O_{\mathbf{P}}(\delta_n^2)$.

PROOF. For a shorter presentation we only consider the case $\sigma(\delta) \equiv 0$. Recall that $g_p = (p/p_{\frac{1+\alpha}{2}})\mathbb{1}_{\{q>0\}}$. From the definition of $\hat{p}$ and Lemma 2.1 we have that

$$0 \leq \int (g_{\hat{p}} - 1) dQ_n \leq \int (g_{\hat{p}} - 1) d(Q_n - Q).$$

As a consequence, $\forall \delta \geq \delta_n$,

$$\begin{aligned}
&\mathbf{P}\left(\int (g_{\hat{p}} - 1) dQ_n \geq \delta^2\right)\\
&\quad \leq \mathbf{P}\left(\int (g_{\hat{p}} - 1) d(Q_n - Q) \geq \delta^2\right)\\
&\quad = \mathbf{P}\left(\left\{\int (g_{\hat{p}} - 1) d(Q_n - Q) \geq \delta^2\right\} \cap \{h_\alpha(\hat{p}, p_0) > 2\delta\}\right)\\
&\qquad + \mathbf{P}\left(\left\{\int (g_{\hat{p}} - 1) d(Q_n - Q) \geq \delta^2\right\} \cap \{h_\alpha(\hat{p}, p_0) \leq 2\delta\}\right).
\end{aligned}$$

The first term on the right-hand side of the last inequality can be bounded as in Proposition 4.1. The second term can be bounded by

$$\mathbf{P}\left(\sup_{h_\alpha(p, p_0) \leq 2\delta} \int (g_p - 1) d(Q_n - Q) \geq \delta^2\right),$$

for which we can also apply the arguments of Proposition 4.1. Thus, we obtain the order of the first quantity. For the second it suffices to remark that $0 \leq \int (1 - g_{\hat{p}}) dQ \leq \int (g_{\hat{p}} - 1) d(Q_n - Q)$ and to use again the order of $\int (g_{\hat{p}} - 1) d(Q_n - Q)$. The order of the third quantity is obtained using the inequality $\log x \leq x - 1$, $x > 0$, and the definition of $\hat{p}$. The fourth quantity is dominated by $2(1/\sqrt{\alpha} + 1)^2 h_\alpha^2(\hat{p}, p_0)$. □

Apparently, we cannot use the previous propositions in order to obtain the rate of convergence of the log-likelihood ratio $\int_{\hat{p}>0} \log(p_0/\hat{p})\, dQ$. However, we can extend an inequality between the Hellinger distance and the Kullback-Leibler divergence [see Wong and Shen (1995)] to h_α. Using this inequality, which is of interest in its own right, we obtain the above log-likelihood convergence rate under mild conditions. Recall that Wong and Shen [(1995), page 357] prove that, under some integrability conditions [see condition (4.6)],

$$\int p_1 \log(p_1/p_2) d\mu = O(\varepsilon^2 \log(1/\varepsilon)),$$

where $\varepsilon^2 = \int (p_1^{1/2} - p_2^{1/2})^2 d\mu$.

PROPOSITION 4.3. *Consider $\alpha \in [0, 1)$, $\mathscr{P}$ a set of nonnegative measurable functions and Q a finite measure defined on $(\mathscr{X}, A)$. Let $\{p_n\}$, $\{p'_n\} \subset \mathscr{P}$ and assume that there exists $\gamma > 0$ such that, for any $n \geq 1$*

$$a_n = \int_{p_n > 0} \left(\frac{p_{n,\alpha}}{p_n}\right)^\gamma dQ < \infty, \tag{4.6}$$

where $p_{n,\alpha} = \alpha p_n + (1-\alpha)p'_n$. *Define*

$$h_n = \int_{p_{n,\alpha}>0} \left(\sqrt{\frac{p_n}{p_{n,\alpha}}} - 1\right)^2 dQ,$$

and assume $h_n > 0$, $n \geq 1$, *and* $h_n \to 0$. *Moreover, suppose that*

$$c_n = \frac{\log(a_n \vee 1)}{\log(1/h_n)}, \qquad n \geq 1, \tag{4.7}$$

is a bounded sequence of nonnegative real numbers. If $C = \limsup_{n\to\infty} c_n$, *then*

$$0 \leq \limsup_{n\to\infty} \frac{1}{h_n^2 \log(1/h_n)} \int_{p_n>0} \left(\log \frac{p_{n,\alpha}}{p_n} + \frac{p_n}{p_{n,\alpha}} - 1\right) dQ \leq \frac{2(2+C)}{\gamma}.$$

PROOF. Consider some $\varepsilon \in (0,1)$, to be specified below. Since the function

$$f(t) = \frac{\log(1/t) + t - 1}{(\sqrt{t}-1)^2}, \qquad t > 0,$$

is decreasing we deduce that, for any $p, p' \in \mathscr{P}$,

$$0 \leq \log \frac{p_\alpha}{p} + \frac{p}{p_\alpha} - 1 \leq \frac{\log(1/\varepsilon) + \varepsilon - 1}{(\sqrt{\varepsilon}-1)^2} \left(\sqrt{\frac{p}{p_\alpha}} - 1\right)^2,$$

provided that $p/p_\alpha \geq \varepsilon$; herein $p_\alpha = \alpha p + (1-\alpha)p'$. We may write

$$\int_{p>0} \left(\log \frac{p_\alpha}{p} + \frac{p}{p_\alpha} - 1\right) dQ = \int_{p/p_\alpha \geq \varepsilon} \left(\log \frac{p_\alpha}{p} + \frac{p}{p_\alpha} - 1\right) dQ$$
$$+ \int_{0<p/p_\alpha<\varepsilon} \left(\log \frac{p_\alpha}{p} + \frac{p}{p_\alpha} - 1\right) dQ \stackrel{\text{def}}{=} I_1 + I_2.$$

Moreover, from the monotonicity of the function f defined above we have

$$I_1 \leq \frac{\log(1/\varepsilon) + \varepsilon - 1}{(\sqrt{\varepsilon}-1)^2} \int_{p/p_\alpha \geq \varepsilon} \left(\sqrt{\frac{p}{p_\alpha}} - 1\right)^2 dQ \leq \frac{\log(1/\varepsilon) + \varepsilon}{(\sqrt{\varepsilon}-1)^2} h^2,$$

where $h^2 = \int_{p_\alpha>0} (\sqrt{p/p_\alpha} - 1)^2 dQ$. Use the fact that $\log x / x^\gamma$ is decreasing for $x \geq e^{1/\gamma}$ and deduce

$$\begin{aligned} I_2 &= \int_{0<p/p_\alpha<\varepsilon} \log \frac{p_\alpha}{p}\, dQ + \int_{0<p/p_\alpha<\varepsilon} \left(\frac{p}{p_\alpha} - 1\right) dQ \\ &\leq \frac{\log(1/\varepsilon)}{(1/\varepsilon)^\gamma} \int_{0<p/p_\alpha<\varepsilon} \left(\frac{p_\alpha}{p}\right)^\gamma dQ + (\varepsilon - 1) Q(\{0 < p/p_\alpha < \varepsilon\}) \\ &\leq (a \vee 1) \frac{\log(1/\varepsilon)}{(1/\varepsilon)^\gamma}, \qquad \varepsilon \leq e^{-1/\gamma}, \end{aligned}$$

where $a = \int_{p>0} (p_\alpha/p)^\gamma\, dQ < \infty$. Consequently,

$$\int_{p>0} \left(\log \frac{p_\alpha}{p} + \frac{p}{p_\alpha} - 1\right) dQ \leq \frac{\log(1/\varepsilon) + \varepsilon}{(\sqrt{\varepsilon}-1)^2} h^2 + (a \vee 1)\, \varepsilon^\gamma \log(1/\varepsilon).$$

Take $\varepsilon \leq e^{-1/\gamma}$ such that $\varepsilon^{\gamma}(a \vee 1) = h^2$ [thus, necessarily, $h^2 \leq (a \vee 1)/e$]. We obtain

$$\begin{aligned}
&\int_{p>0} \left(\log \frac{p_\alpha}{p} + \frac{p}{p_\alpha} - 1 \right) dQ \\
&\quad \leq h^2 \left[\frac{\log(1/\varepsilon) + \varepsilon}{(\sqrt{\varepsilon} - 1)^2} + \log(1/\varepsilon) \right] \\
&\quad = \frac{h^2}{\gamma} \log \frac{1}{h} \left[\frac{2 + \frac{\log(a \vee 1)}{\log(1/h)} + \frac{\gamma \left[h^2/(a \vee 1) \right]^{1/\gamma}}{\log(1/h)}}{\left([h^2/(a \vee 1)]^{1/2\gamma} - 1 \right)^2} + 2 + \frac{\log(a \vee 1)}{\log(1/h)} \right].
\end{aligned}$$

Finally, replace p, p',h, a by p_n, p'_n,h_n, a_n, respectively, and let $n \to \infty$. □

Note that there exists C_1, $C_2 > 0$ depending only on α, γ such that $C_1 b_n \leq a_n \leq C_2 b_n$, with a_n defined in (4.6) and $b_n = \int_{p_n>0} (p'_n/p_n)^{\gamma} \, dQ$. Moreover, if $\{a_n\}$ is bounded by $\left\{ C h_n^{-\beta} \right\}$, for some $C, \beta > 0$, then $\{c_n\}$ defined in (4.7) is bounded. Hereafter, i.o. means infinitely often.

COROLLARY 4.4. *Consider a convex model $\mathscr{P}$. Let $p_0 \in \mathscr{P}$ be the* (*pseudo-*) *true density and $\{\widehat{p}\}$ a sequence of MLE. Suppose that the conditions of Proposition* 4.1 *hold and let $\{\delta_n\}$ be a corresponding rate for $h_\alpha(\hat{p}, p_0)$. Moreover, assume that*:

(i) *there exists $\gamma > 0$ and $\{a_n\}$ such that*

$$\limsup_{n \to \infty} \frac{\log a_n}{\log \delta_n^{-1}} < \infty$$

and

$$\mathbf{P} \left(\int_{\widehat{p}>0} \left(\frac{p_0}{\widehat{p}} \right)^{\gamma} dQ > a_n, \ i.o. \right) = 0; \tag{4.8}$$

(ii) $Q(\{\widehat{p} = 0\}) = O_{\mathbf{P}}(\delta_n^2 \log \delta_n^{-1})$.

Then,

$$\int_{\widehat{p}>0} \log \frac{p_0}{\widehat{p}} \, dQ = O_{\mathbf{P}}(\delta_n^2 \log \delta_n^{-1}).$$

PROOF. From Proposition 4.1 we have $h_0^2(\hat{p}, p_0) = O_{\mathbf{P}}(\delta_n^2)$. Moreover, from Corollary 4.2 we have $\int \left(\widehat{p}/\widehat{p}_{1/2} - 1 \right) dQ = O_{\mathbf{P}}(\delta_n^2)$. We deduce

$$\int_{\widehat{p}>0} \left(\frac{\widehat{p}}{\widehat{p}_{1/2}} - 1 \right) dQ = O_{\mathbf{P}}(\delta_n^2 \log \delta_n^{-1}).$$

Since $h^2_{1/2} \le h^2_0$ (see Lemma 2.5), apply Proposition 4.3 for $\alpha = 1/2$, $p'_n = p_0$, $p_n = \widehat{p}$ and deduce $\int_{\widehat{p}>0} \log\big(\widehat{p}_{1/2}/\widehat{p}\big)\, dQ = O_{\mathbf{P}}(\delta_n^2 \log \delta_n^{-1})$. We can write

$$\begin{aligned}\int_{\widehat{p}>0} \log \frac{\widehat{p}_{1/2}}{\widehat{p}}\, dQ &= \int_{\widehat{p}>0} \log \frac{\widehat{p}_{1/2}}{p_0}\, dQ + \int_{\widehat{p}>0} \log \frac{p_0}{\widehat{p}}\, dQ \\ &= \int \log \frac{\widehat{p}_{1/2}}{p_0}\, dQ - \log \frac{1}{2}\, Q(\{\widehat{p} = 0\}) + \int_{\widehat{p}>0} \log \frac{p_0}{\widehat{p}}\, dQ \\ &\le -\log \frac{1}{2}\, Q(\{\widehat{p} = 0\}) + \int_{\widehat{p}>0} \log \frac{p_0}{\widehat{p}}\, dQ.\end{aligned}$$

On the other hand, use the concavity of the logarithm function and deduce

$$\frac{1}{2} \int_{\widehat{p}>0} \log \frac{p_0}{\widehat{p}}\, dQ \le \int_{\widehat{p}>0} \log \frac{\widehat{p}_{1/2}}{\widehat{p}}\, dQ.$$

The last two inequalities yield the order of $\int_{\widehat{p}>0} \log(p_0/\widehat{p})\, dQ$. □

Conditions (i) and (ii) of the previous corollary depend on the properties of the model and the true distribution Q. Note that condition (ii) does not represent a serious constraint. In any case, the MLE should be positive at all observation points and this will usually imply $Q(\{\widehat{p} = 0\}) = O_{\mathbf{P}}(n^{-1})$ (see Section 5.2 for an example). Moreover, $Q(\{\widehat{p} = 0\}) = 0$ provided that the densities of the model have the same support. A possible strategy for verifying condition (i) is the following. Let $\varphi \ge 1$ be a function defined on the sample space and assume that there exists $\gamma > 0$ such that $\int \varphi^\gamma dQ < \infty$. Consider

$$\mathscr{P}' = \{p \in \mathscr{P},\ (p_0/p)\, \mathbb{1}_{\{p>0\}} \le \varphi\}$$

and deduce that $\mathbf{P}(\hat{p} \notin \mathscr{P}',\ i.o.) = 0$ as a consequence of the almost sure convergence of $h_\alpha(\hat{p}, p_0)$ (see Section 5.1 for an example). In some cases it may be necessary to make φ and $\mathscr{P}'$ to depend on n. More precisely, one will have to look for a suitable sequence $\{a_n\}$ and $\gamma > 0$ and bound $\mathbf{P}(\int_{\widehat{p}>0} [p_0/\widehat{p}]^\gamma\, dQ > a_n)$ using the properties of p_0, $\widehat{p}$ and conditions on Q. The Borel-Cantelli lemma will then, with appropriates bounds, imply that $\mathbf{P}(\int_{\widehat{p}>0} [p_0/\widehat{p}]^\gamma\, dQ > a_n,\ i.o.) = 0$ (see Section 5.2 for an example).

5. Examples.

We consider four examples of convex models: mixtures of discrete distributions, monotone densities, decreasing failure rate distributions and a finite-dimensional parametric model. Using the general facts developed above we (re)obtain asymptotic properties of MLE when the model is possibly misspecified. See also Pfanzagl (1990) for a different approach.

EXAMPLE 1 (Mixtures of discrete distributions). In this example the sample space is the set of nonnegative integers and the dominating measure is the counting measure. Consider the class of mixtures of power series distributions (PSD hereafter); see Noack (1950). Recall that a family of PSD is defined as follows: let $a(\cdot) : [0, R] \to (0, \infty]$, $a(\eta) = \sum_{x\ge 0} a_x \eta^x$ with $a_x \ge 0$, $x = 0, 1, \ldots$

and R the radius of convergence of the power series. Without loss of generality we may assume $a_0 = 1$. A power series distribution P_η is defined by the probabilities $p_\eta(x) = P_\eta(\{x\}) = a_x \eta^x / a(\eta)$, $x \geq 0$. For the sake of simplicity, we consider hereafter that the function $a(\cdot)$ is such that there exists $b_l \geq 0$, $l \geq 1$ with $b_1 > 0$ and

$$\log a(\eta) = \sum_{l \geq 1} b_l \eta^l. \tag{5.1}$$

In particular, this implies $a_x > 0$ for all $x \geq 0$. The class of PSD satisfying (5.1) contains distributions such as Poisson, negative binomial and Hermite. The mixture models considered in this subsection are built by mixing PSD defined by the same function $a(\cdot)$. Thus, $\mathscr{P} = \{p_\theta,\ \theta \text{ probability on } H\}$ with

$$p_\theta(\cdot) = \int_H p_\eta(\cdot) d\theta(\eta)$$

and H a closed interval included in the set of points η for which the power series converges. By a slight abuse of notation, we identify η with the unit point mass at η. Lüxmann-Ellinghaus (1987) showed that (5.1) ensures the identification of the mixture model, that is $p_\theta(x) = p_{\theta'}(x)$, $\forall x \geq 0$, implies that the mixing distributions θ and θ' coincide. The properties of the (nonparametric) MLE in such models have been studied by Simar (1976), Lindsay (1983), Pfanzagl (1988) and Milhaud and Mounime (1996), among others. In the case of misspecification, one can use the empirical distribution and thus avoid the bias due to considering a wrong model. Even if the statistician is able to detect misspecification, he may still consider his PSD mixture model, say for interpretation purposes and/or for estimating the tail probabilities where no observation occurred.

The first issue to be analyzed in our framework is the existence of the pseudo-true mixing distribution θ_0. Note that $\sup_\theta \int \log p_\theta \, dQ$ finite (the supremum is taken over all probabilities on H) guarantees the existence of a pseudo-true mixing distribution [see the proposition of Pfanzagl (1990)]. Below we provide some mild conditions ensuring a finite supremum.

LEMMA 5.1. (a) *Consider a family of PSD satisfying* (5.1) *and a distribution Q on the nonnegative integers. If $\int x \log(x+1)\, dQ(x) < \infty$, then* $\sup_\theta \int \log p_\theta \, dQ$ *is finite*.

(b) *In the case of a mixture of Poisson distributions*, $\sup_\theta \int \log p_\theta \, dQ$ *is finite iff* $\int \log(x+1)\, dQ(x) < \infty$.

PROOF. See the Appendix.

Consider a misspecified mixture model, in particular a PSD mixture model, and a true distribution of the observations such that there exists a unique pseudo-true density. Let θ_0 be a pseudo-true mixing distribution, that is, $\int \log(p_{\theta_0}/p_\theta)\, dQ \geq 0$, $\forall \theta$ mixing distribution. We can deduce from Lemma

2.1 that

$$\int \frac{p_\eta}{p_{\theta_0}}\, dQ \le 1 \quad \forall \eta \in H. \tag{5.2}$$

Let S be the set of $\eta \in H$ satisfying (5.2) with equality. Integrating with respect to θ_0 and using the Fubini theorem we obtain $\theta_0(S) = 1$. In the case of misspecified PSD mixture models we obtain that θ_0 is discrete and its support has no limit point in $[0, R)$. Indeed, for any $\eta \in S$ we have

$$\sum_{x\ge 0} \frac{q(x)}{\int u^x/a(u)d\theta_0(u)} \eta^x = \sum_{x \ge 0} a_x \eta^x.$$

If S has a limit point in $[0, R)$, then identify the coefficients of the two series above and deduce that Q is necessarily a mixture of PSD. In particular, in the case of a misspecified PSD mixture model defined for a finite interval H the pseudo-true mixing distribution θ_0 is necessarily finitely supported.

It can be proved [see Lindsay (1983)] that in a PSD mixture model as considered herein there exists a unique MLE. Let it be denoted by $\hat{\theta}$.

COROLLARY 5.2. *Let Q be the probability generating the independent observations. Consider a PSD mixture model $\mathscr{P}$ and assume that there exists θ_0 unique (pseudo-)true mixing probability. Then, almost surely, the MLE $\hat{\theta}$ weakly converges towards θ_0.*

PROOF. First remark that if $(\theta_n)_{n\ge 1}$ is a sequence of probabilities on H and $\theta_n \to \theta_0$ weakly, then $p_{\theta_n}(x) \to p_{\theta_0}(x)$, $x = 0, 1, \dots$. Then apply Corollary 3.4. □

Now we look for the rates of convergence using Proposition 4.1. We confine our attention to the case $H = [0, M]$, $0 < M < 1 \le R$. Assume that there exists a unique (pseudo-)true mixing distribution θ_0. Note that, for any θ probability on $[0, M]$,

$$p_\theta(x) = a_x \int_{[0,M]} \eta^x a(\eta)^{-1}\, d\theta(\eta) \le \frac{1}{a_0} a_x M^x, \qquad x = 0, 1, \dots.$$

Moreover, using the Cauchy-Hadamard formula $1/R = \limsup_{x\to\infty} |a_x|^{1/x}$, we deduce that there exists $M \le M' < 1$ and $B > 0$, such that, for any θ, $p_\theta(x) \le BM'^x$, $\forall x \ge 0$. Consider a family $\mathscr{G}_0^{\sigma_n}$ as in (4.1) defined for a sequence $\sigma_n \to 0$. Note that $\{p_{\theta_0}(x) \ge \sigma_n\} \subset \{BM'^x \ge \sigma_n\} \subset \{x \le a_n\}$ where $a_n = a' \log(1/\sigma_n)$ and a' is some positive constant independent of the sequence $\{\sigma_n\}$ (provided that the sequence is bounded, say, by 1/2), but dependent on M'. Consequently, we have

$$\mathscr{G}_0^{\sigma_n} \subset \left\{ \frac{2p_\theta}{p_\theta + p_{\theta_0}} \mathbb{1}_{\{q>0\}} \mathbb{1}_{[0,a_n]},\ \theta \,\text{probability on } [0, M] \right\}.$$

We deduce that $H_B(\delta, \mathscr{G}_0^{\sigma_n}, Q)$, $\delta > 0$, can be bounded by the logarithm of the number of all functions defined from a set of $a' \log(1/\sigma_n)$ elements to a set of $(1 + 1/\delta)$ elements. Thus, $H_B(\delta, \mathscr{G}_0^{\sigma_n}, Q) \le A' \log(1/\sigma_n) \log(1/\delta)$, for some $A' > 0$, independent of σ_n and δ. Finally, we have

$$J_B(\delta, \mathscr{G}_0^{\sigma_n}, Q) \le A\, \delta \sqrt{\log(1/\sigma_n)} \sqrt{\log(1/\delta)}\,,$$

for some $A > 0$.

On the other hand, we bound $Q\left(\{p_{\theta_0} \le \sigma_n\}\right)$ by looking for a suitable set K_n such that $\{p_{\theta_0} \le \sigma_n\} \subset K_n$. Fix $0 < s \le M$ such that $\theta_0([s, M]) > 0$. Then, $\forall\, x \ge 0$, $p_{\theta_0}(x) \ge c' a_x s^x$, for some $c' > 0$. We note from the proof of Lemma 5.1 that if $a(\cdot)$ satisfies (5.1), then there exists $v > 0$ such that $a_x \ge (1/x!)^v$. In this case $p_{\theta_0}(x) \ge c' s^x/(x!)^v$. Stirling's formula indicates $K_n = \{x,\ \ C' x \log x \ge \log(1/\sigma_n)\}$, for some constant $C' > 0$.

Finally, we choose σ_n and δ_n ($\sigma_n = \sigma(\delta_n)$) such that

$$\sqrt{n}\, \delta_n \ge A \sqrt{\log(1/\sigma_n)} \sqrt{\log(1/\delta_n)} \tag{5.3}$$

and

$$\delta_n^2 \ge \sum_{x \ge Z_n} q(x) \ge Q\left(\{p_{\theta_0} \le \sigma_n\}\right), \tag{5.4}$$

where $Z_n = \inf K_n$ and A is some positive constant. The choice of the truncating sequence $\{\sigma_n\}$ has to be made by a trade-off between (5.3) and (5.4). A small value of σ_n allows faster rates in (5.4) but imposes slower rates in (5.3), while a larger value of σ_n produces the opposite effect. To simplify the problem, assume that the true distribution has a finite exponential moment, that is there exists $t_0 > 0$ such that $\int \exp(t_0 x)\, dQ(x) < \infty$. Then, (5.4) is satisfied if

$$\delta_n^2 \ge \exp(-t_0 Z_n). \tag{5.5}$$

Observe that, for any $\varepsilon > 0$, $Z_n = t_0^{-1} \log n - (2 + \varepsilon) t_0^{-1} \log\log n$, $\log 1/\sigma_n = C' Z_n \log Z_n$ and $\delta_n^2 = n^{-1} \log^{2+\varepsilon} n$ satisfy (5.3) and (5.5), up to some multiplicative constants. This allows us to state the following result.

COROLLARY 5.3. *Consider a class of PSD satisfying* (5.1) *and* $H = [0, M]$, $M < 1 \le R$. *Let* $\mathscr{P}$ *denote the corresponding PSD mixture model and* Q *the true distribution of the i.i.d. sample. Suppose that there exists a* (*pseudo-*)*true mixing distribution* θ_0. *Moreover, assume that there exists* $t_0 > 0$ *such that* $\int \exp(t_0 x)\, dQ(x) < \infty$. *Then, for any* $\alpha \in [0, 1)$ *and* $\varepsilon > 0$,

$$h_\alpha^2(p_{\hat\theta}, p_{\theta_0}) = O_{\mathbf{P}}\left(n^{-1} \log^{2+\varepsilon} n\right).$$

Now, we can apply Corollary 4.2 in order to obtain the rates of different quantities of interest. For instance,

$$\int \log \frac{2 p_{\hat\theta}}{p_{\hat\theta} + p_{\theta_0}}\, dQ_n \quad \text{and} \quad \int \left(\frac{p_{\hat\theta} - p_{\theta_0}}{p_{\hat\theta} + p_{\theta_0}}\right)^2 dQ \tag{5.6}$$

are of order $n^{-1}\log^{2+\varepsilon} n$, $\forall\varepsilon > 0$. The first quantity is used by van de Geer (2000), section 11.2, to prove asymptotic normality in general mixture models. The second can be used to recover the orders of the chi-square type norms considered by Lambert and Tierney (1984) (see their Proposition 3.1) and Milhaud and Mounime (1996). In both papers such rates represent the cornerstone for proving the asymptotic normality of $\sqrt{n}(p_{\hat{\theta}}(x) - p_{\theta_0}(x))$, $x = 0, 1, \dots$ in the case of mixtures models as considered herein. Future work will try to extend these results on the asymptotic normality to the case of misspecification.

On the other hand, from Corollary 4.4 we obtain

$$\int \log\left(\frac{p_{\theta_0}}{p_{\hat{\theta}}}\right) dQ = O_{\mathbf{P}}\left(n^{-1}\log^{3+\varepsilon} n\right)$$

$(Q(\{p_{\hat{\theta}} > 0\}) = 1$ in this example). Indeed, in order to verify (4.8), let $0 < s \le M$ such that $\theta_0([s, M]) \ge \eta > 0$. Take $\mathscr{P}'$ the set of PSD mixtures with $\theta([s, M]) \ge \eta/2$. Observe that for any element of $\mathscr{P}'$ we can write $p_\theta(x) \ge a_x s^x \eta/2a(M)$ and $p_\theta(x) \le a_x M^x/a_0$, $x \ge 0$. Hence, if the true distribution has a finite exponential moment, then there exists $\gamma > 0$ such that $\sup_{p_\theta \in \mathscr{P}'} \int_{p_\theta > 0} (p_{\theta_0}/p_\theta)^\gamma \, dQ < \infty$. If $\theta_0(\{s\}) = 0$, then Corollary 5.2 tells that $\mathbf{P}(p_{\hat{\theta}} \notin \mathscr{P}', \, i.o.) = 0$ and this implies (4.8).

Condition (5.1), which ensures the identifiability of mixtures of PSD, facilitates the proof of the existence of a pseudo-true density and the entropy calculations. However, since h_α is a divergence between densities, convergence results should also be obtainable when the (pseudo-)true mixing distribution θ_0 is not identifiable.

EXAMPLE 2. (Monotone densities). Let $\mathscr{P}$ be the set of all decreasing densities defined on $[0, \infty)$ (hereafter the dominating measure is the Lebesgue measure). The MLE in such a model, usually called the Grenander's estimator, was first described by Grenander (1956) and afterwards analyzed, amongst others, by Wang (1985) and Birgé (1989). It is defined as the slope of the least concave majorant of the empirical distribution function (df hereafter). There are (at least) two ways of parameterizing a model of decreasing densities. The first is as mixtures of uniforms on $[0, \eta]$, $\eta > 0$, with parameter θ the mixing distribution on $(0, \infty)$. Note that the mixing distribution θ is identifiable in this case. Another possible parameterization is $\theta = p$, $\Theta = \mathscr{P}$ and Θ is considered with the topology given by the distance considered by Wang (1985):

$$\begin{aligned} d(p_1, p_2) = \inf_{h \ge 0}\big\{ & p_1(x+h) - h \le p_2(x), \ \forall x \ge 0, \\ & \text{and } p_2(x) \le p_1(x-h) + h, \ \forall x \ge h \big\}. \end{aligned} \tag{5.7}$$

This metric could be extended to the set of decreasing subdensities on $[0, \infty)$; let us denote this set by $\overline{\mathscr{P}}$. A decreasing subdensity on $[0, \infty)$ is a decreasing nonnegative measurable function with integral at most one and it corresponds to a mixture of uniforms $[0, \eta]$ with mixing subdistribution (subprobability) concentrated on $(0, \infty)$. Wang's Lemma 4.1 states that $(\overline{\mathscr{P}}, d)$ is a compact metric space and that for any p and $\{p_n\}$ in $\overline{\mathscr{P}}$, $d(p_n, p) \to 0$ iff $p_n(x) \to p(x)$

at all nonzero continuity points of p. In particular, if p_n and p are densities such that $d(p_n, p) \to 0$, then $p_n \to p$ in $L_1(\mu)$ (by Scheffé's theorem). Below we show that convergence with respect to the d metric is equivalent to h_α-convergence.

LEMMA 5.4. *If* $\{p_{\theta_n}\}$*and* p_θ *are decreasing* (*sub*)*densities, and* $\{\theta_n\}$ *and* θ *the associated mixing* (*sub*)*distributions, then*

$$d(p_{\theta_n}, p_\theta) \to 0 \ \Leftrightarrow \theta_n \to \theta \text{ (vaguely) } \textit{weakly}.$$

If p_{θ_0} *is the* (*pseudo-*)*true decreasing density*, *then*, *for any* $\alpha \in [0, 1)$

$$d(p_{\theta_n}, p_{\theta_0}) \to 0 \Leftrightarrow h_\alpha(p_{\theta_n}, p_{\theta_0}) \to 0.$$

PROOF. We prove the first equivalence for subdensities. If θ_n converges to θ in the vague topology then, for any $x > 0$ such that $\theta(\{x\}) = 0$,

$$p_{\theta_n}(x) = \int_{\{x\le\eta\}} \eta^{-1}\theta_n(d\eta) \to \int_{\{x\le\eta\}} \eta^{-1}\theta(d\eta) = p_\theta(x)$$

[see, e.g., Billingsley (1968)]. For this implication we remark that $x > 0$ is a continuity point for p_θ iff $\theta(\{x\}) = 0$. For the converse implication, provided that θ_n does not converge vaguely to θ, Helly's selection theorem implies the existence of subsequence $\{\theta_{n_k}\}$ vaguely convergent to some subdistribution $\theta_1 \ne \theta$, θ_1. From this and the previous implication we obtain a contradiction. The first equivalence in the case of densities can be obtained in a similar way. The second equivalence is a consequence of Lemma 2.6. □

We have the following extension of Marshall's lemma [see Marshall (1970)].

LEMMA 5.5. *Consider* $X_1, X_2, \dots, X_n$ *an independent sample from a distribution* Q *on* $[0, \infty)$ *with df* $Q(\cdot)$. *Let* $\hat{p}^G$ *be the Grenander estimator and* $\hat{P}^G(\cdot)$ *the corresponding df. For any* n,

$$\sup_{x\ge 0} |\hat{P}^G(x) - P_0(x)| \le \sup_{x\ge 0} |Q_n(x) - Q(x)|, \tag{5.8}$$

where $Q_n(\cdot)$ *is the empirical df and* $P_0(\cdot)$ *denotes the smallest concave majorant of* $Q(\cdot)$. *Moreover, almost surely,* $\hat{p}^G$ *converges pointwise, except possibly at countable many points of* $[0, \infty)$, *and in* $L_1(\mu)$ *to* p_0, *the density corresponding to* $P_0(\cdot)$.

PROOF. Fix n and define $a = \sup_{x\ge 0} |Q_n(x) - Q(x)|$. Then $P_0(\cdot) + a$ is a concave function dominating the empirical df. From the definition of $\hat{P}^G(\cdot)$ we deduce $P_0(\cdot) + a \ge \hat{P}^G(\cdot)$. On the other hand, $P_0(\cdot) - a$ is the concave envelope of $Q(\cdot) - a$ and from $Q(\cdot) - a \le Q_n(\cdot)$ we deduce that $P_0(\cdot) - a \le \hat{P}^G(\cdot)$. From the Glivenko-Cantelli theorem and (5.8) we obtain that, almost surely, $\hat{P}^G(\cdot)$ converges uniformly to the smallest concave majorant of the true df. If we apply, for example, exercise C.9, page 20 of Roberts and Varberg (1973), we obtain that $\hat{p}^G$ converges to p_0, except possibly at countable many points of $[0, \infty)$. Scheffé's theorem gives the $L_1(\mu)$ convergence of $\hat{p}^G$, almost surely. □

The previous lemma indicates that the pseudo-true density in a misspecified decreasing densities model is the slope of the least concave majorant of the true df. This pseudo-true density always exists, even if (1.2) is not satisfied. The full justification of the fact that p_0 of Lemma 5.5 is indeed the pseudo-true density is given in Patilea (1997). A quick argument can be provided if we assume that $\int_{[0,1]} \log x \, dQ(x) > -\infty$. By simple geometry we have that $xp(x) \leq 1$, $\forall x \geq 0$, if p is a decreasing density on $[0, \infty)$ and from this we may deduce that $\sup_{p\in\mathscr{P}} \int \log p \, dQ$ is finite. Now we may apply the proposition of Pfanzagl (1990).

The facts of section 3 can be used in order to prove the h_α-convergence of the Grenander estimator. More interesting are the h_α-rates which we present below.

COROLLARY 5.6. *Let $\mathscr{P}$ be the family of decreasing densities on* $[0, \infty)$ *and $\hat{p}^G$ Grenander's estimator. Assume that there exists $\varepsilon \in (0, 1)$ such that*

$$\int_{\{p_0>1\}} p_0^{\varepsilon} dQ < \infty \tag{5.9}$$

and

$$\int_{\{p_0\leq 1\}} p_0^{-\varepsilon} dQ < \infty, \tag{5.10}$$

with p_0 the slope of the least concave majorant of the true df $Q(\cdot)$. Then, for any $\alpha \in [0, 1)$, $h_\alpha^2(\hat{p}^G, p_0) = O_{\mathbf{P}}(n^{-2/3})$.

PROOF. Following van de Geer [(2000), Example 7.4.2], consider the family of decreasing functions $\tilde{\mathscr{G}} = \{[2pp_0/(p + p_0)]^{1/2} \ \mathbb{1}_{\{q>0\}}, \, p \in \mathscr{P}\}$. Entropy bounds for $\tilde{\mathscr{G}}$ can be deduced from entropy bounds for the families $\tilde{\mathscr{G}}' = \{\tilde{g}\mathbb{1}_{\{p_0>1\}}, \ \tilde{g} \in \tilde{\mathscr{G}}\}$ and $\tilde{\mathscr{G}}'' = \{\tilde{g} \ \mathbb{1}_{\{p_0\leq 1\}}, \ \tilde{g} \in \tilde{\mathscr{G}}\}$. For $\tilde{\mathscr{G}}'$ we apply Lemma 7.11 of van de Geer for a family defined on $\{p_0 > 1\} \subset [0, 1]$, $F = (2p_0)^{1/2}$, $d\mu = p_0^{-1} dQ$ and using (5.9). We deduce that, for some $A_1 > 0$, $H_B(\delta, \tilde{\mathscr{G}}', \mu) = H_B(\delta, \ p_0^{-1/2}\tilde{\mathscr{G}}', Q) \leq A_1\delta^{-1}$, $\forall\delta > 0$ (we denote by $p_0^{-1/2}\tilde{\mathscr{G}}'$ the set $\{p_0^{-1/2}\tilde{g}, \tilde{g} \in \tilde{\mathscr{G}}'\}$). For the family $\tilde{\mathscr{G}}''$, we apply van de Geer's Lemma 7.10 with the same F and $d\mu$. Using (5.10) we deduce that, for some $A_2 > 0$, $H_B(\delta, \tilde{\mathscr{G}}'', \mu) = H_B(\delta, \, p_0^{-1/2}\tilde{\mathscr{G}}'', Q) \leq A_2\delta^{-1}$, $\forall\delta > 0$. Consequently, $H_B(\delta, \tilde{\mathscr{G}}, \mu) = H_B(\delta, p_0^{-1/2}\tilde{\mathscr{G}}, Q) \leq A\delta^{-1}$, with $A > 0$. Finally, apply Proposition 4.1 for $\mathscr{G} = p_0^{-1/2}\tilde{\mathscr{G}}$. □

It would be interesting to state the conditions (5.9) and (5.10) only in terms of the true distribution. Even if this seems to be difficult in a general framework, we can do it in some important cases. If q is unbounded but decreasing on some interval $I = (0, \delta)$, then $P_0(\cdot)$ and $Q(\cdot)$ coincide on any $(0, \delta') \subset I$

provided that the graph of $Q(\cdot)$ lies below the right-tangent to $Q(\cdot)$ at the point $(\delta', Q(\delta'))$. Therefore, in this case the condition (5.9) can be replaced by

$$\int_{I\cap\{q>1\}} q^{\varepsilon} dQ < \infty,$$

for some $\varepsilon \in (0, 1)$. If $q \le C_1$ on some $I = (0, \delta)$, with C_1 a positive constant, then $p_0 \le C_1 \vee (1/\delta)$ on I and thus (5.9) is satisfied. Finally, from $p_0(x)x \le 1$, $x > 0$ we deduce that (5.9) is satisfied if

$$\int_{\{x<1\}} x^{-\varepsilon} dQ < \infty.$$

Similar arguments can be used for replacing (5.10). If q is decreasing to zero on some $J = (\beta, b)$, where $b = \sup\{q > 0\}$, then there exists a neighborhood of b on which $p_0 = q$. This is because, when approaching b, $P_0(\cdot)$ and $Q(\cdot)$ will coincide as soon as the graph of $Q(\cdot)$ lies below the left-tangents to $Q(\cdot)$. Therefore, in this case (5.10) can be replaced by

$$\int_{J\cap\{q\le1\}} q^{-\varepsilon} dQ < \infty,$$

for some $\varepsilon \in (0, 1)$. If $q \ge C_2 > 0$ on some (finite) interval (β, b), with C_2 a constant, then $p_0 \ge C_2 \wedge [1 - Q(\beta)]\, b^{-1}$ and thus (5.10) is satisfied. Finally, (5.10) always implies

$$\int_{\{x\ge1\}} x^{\varepsilon} dQ < \infty. \tag{5.11}$$

As a consequence of Corollary 5.6, we can also deduce the rates of the quantities listed in Corollary 4.2. In order to apply Corollary 4.4, note that $\hat{p}^G(x) > 0$ for any $0 \le x \le X_{(n)}$, where $X_{(n)}$ denotes the largest observation. Therefore, we have $Q\left(\{\hat{p}^G = 0\}\right) = 1 - Q\left(X_{(n)}\right) = O_{\mathbf{P}}(n^{-1})$. This order is an easy consequence of the fact that $Q^n\left(X_{(n)}\right)$ is uniformly distributed on $[0, 1]$. Now, it remains to verify (4.8). Fix $y > 0$ a continuity point of p_0 such that $p_0(y) > 0$ and let $\mathscr{P}' = \{p,\ p(y) \ge p_0(y)/2\}$. Note that (5.9) ensures $\sup_{p\in\mathscr{P}'} \int_{[0,y]} (p_0/p)^{\varepsilon}\, dQ < \infty$ and that the h_α-convergence of $\hat{p}^G$ implies $\mathbf{P}(\hat{p}^G \notin \mathscr{P}',\ i.o.) = 0$. On the other hand, it is easy to see that $\hat{p}^G \ge 1/nX_{(n)}$. Therefore, for any $\gamma > 0$

$$\int_{\hat{p}^G>0} \left(\frac{p_0}{\hat{p}^G}\right)^{\gamma} \mathbb{1}_{[y,\infty)}\, dQ \le C\left(nX_{(n)}\right)^{\gamma},$$

for some $C > 0$. From (5.11) we deduce that the function $f(x) = x^{\varepsilon}(1 - Q(x))$, $x \ge 1$ is bounded and thus, for any $\beta > 0$, there exists $C', n_0 > 0$ such that

$$Q^n\left(n^{\beta}\right) \ge \exp\left(-C'n^{1-\varepsilon\beta}\right) \ge 1 - C'n^{1-\varepsilon\beta}, \qquad n \ge n_0.$$

Take $a_n = n$, $n \ge n_0$ and deduce

$$\mathbf{P}\left(\int_{\hat{p}^G>0} \left(\frac{p_0}{\hat{p}}\right)^{\gamma} \mathbb{1}_{[y,\infty)} dQ > a_n\right) \le \mathbf{P}(X_{(n)} > C''n^{1/\gamma-1}) = 1 - Q^n(C''n^{1/\gamma-1}),$$

for some $C'' > 0$ independent of n. If $\gamma < \varepsilon/(2+\varepsilon)$, then the Borel-Cantelli lemma yields

$$\mathbf{P}\left(\int_{\hat{p}^G>0}\left(\frac{p_0}{\hat{p}}\right)^{\gamma}\mathbb{1}_{[y,\infty)}\,dQ > a_n, \text{ i.o.}\right) = 0.$$

Thus, we can state the following result.

COROLLARY 5.7. *If the conditions of Corollary* 5.6 *are met*, *then*

$$\int_{\hat{p}^G>0}\log\frac{p_0}{\hat{p}^G}\,dQ = O_{\mathbf{P}}(n^{-2/3}\log n).$$

EXAMPLE 3. [Decreasing failure rate (DFR) distributions]. This kind of model has been examined, for example, by Marshall and Proschan (1965) and, more recently, Wang (1985). An absolutely continuous distribution on $[0,\infty)$ with density p and df $P(\cdot)$ is called DFR if the *hazard (or failure) rate* $\lambda(x) = p(x)/(1-P(x))$ is decreasing on $[0,\infty)$. DFR distributions arise, for example, as mixtures of exponentials. Let $\mathscr{P}_2$ be the set of the densities of DFR distributions. Note that $\mathscr{P}_2$ is convex [see Barlow et al. (1963)]. Since a DFR distribution has a decreasing density we consider on $\mathscr{P}_2$ the topology induced by the distance written in (5.7). The ML estimator for $\mathscr{P}_2$ was derived in Marshall and Proschan (1965) and it is determined only for x not exceeding the largest observation. Beyond this value MLE can be extended in any manner that preserves the DFR property. We denote such a MLE by $\hat{p}$ and its corresponding hazard rate by $\hat{\lambda}$.

As before, let $Q(\cdot)$ and $Q_n(\cdot)$ denote the true df and the empirical df, respectively. When $dQ/d\mu \notin \mathscr{P}_2$, we can deduce some information regarding the pseudo-true density p_0 from the form of the MLE derived by Marshall and Prochan (1965). Fix $X_{(0)} = 0$ and let $(X_{(1)},\dots,X_{(n)})$ denote the order statistics of the observed sample. A MLE $\hat{p}$ corresponds to a left-continuous hazard rate, constant between observations. More precisely, $\hat{\lambda}(x) = \hat{\lambda}(X_{(i)})$, $X_{(i-1)} < x \le X_{(i)}$, $i = 1,\dots,n-1$, where

$$\hat{\lambda}(X_{(i)}) = \max_{t\ge i}\min_{s\le i-1}\frac{t-s}{\sum_{j=s}^{t-1}(n-j)(X_{(j+1)}-X_{(j)})}. \tag{5.12}$$

Note that the MLE for $\mathscr{P}_2$ is such that $1/\hat{\lambda}(x)$ is the slope (from the left) of the largest convex minorant of $H_{Q_n}^{-1}(t)$ evaluated at the point $t = Q_n(x)$, where, for $P(\cdot)$ a df with support included in $[0,\infty)$ and $t \in [0,1]$,

$$H_P^{-1}(t) = \int_0^{P^{-1}(t)}(1-P(x))\,dx \tag{5.13}$$

$[P^{-1}(t) = \inf\{x,\ P(x) \ge t\}]$. The properties of the function $H_P^{-1}(\cdot)$ are described in section 5.3 of Barlow *et al.* (1972). The form of the MLE identifies p_0 the pseudo-true density as that corresponding to $\lambda_0(\cdot)$, where $1/\lambda_0(\cdot)$ is the slope of $H_{P_0}^{-1}(\cdot)$ which denotes the largest convex minorant of $H_Q^{-1}(\cdot)$. The arguments for proving that p_0 is the pseudo-true density are exactly the same as in the case of decreasing densities. Note that the pseudo-true density for $\mathscr{P}_2$ is not necessarily the same as the pseudo-true density in the misspecified decreasing densities model.

COROLLARY 5.8. *Let $1/\lambda_0(\cdot)$ be the slope of the largest convex minorant of $H_Q^{-1}(\cdot)$ defined as in* (5.13) *and p_0 the corresponding density. Then, almost surely, $h_\alpha(\hat{p}, p_0) \to 0$, $\alpha \in [0, 1)$. Moreover, $\hat{p}(x) \to p_0(x)$ and $\hat{\lambda}(x) \to \lambda_0(x)$ for all nonzero continuity points of $p_0(\cdot)$, and $\sup_{x\geq 0} |\hat{P}(x) - P_0(x)| \to 0$ where $\hat{P}(\cdot)$ is the MLE df and $P_0(\cdot)$ is the df corresponding to p_0. If conditions* (5.9) *and* (5.10) *hold, then $h_\alpha^2(\hat{p}, p_0) = O_{\mathbf{P}}(n^{-2/3})$, $\alpha \in [0, 1)$.*

PROOF. The h_α-convergence is a consequence of Corollary 3.4. The properties of the metric d [see (5.7)] yield the pointwise convergence of $\hat{p}$. This further implies the uniform convergence of the df $\hat{P}$ and the pointwise convergence of $\hat{\lambda}$. The h_α-rates are obtained as in the case of monotone densities. □

In this model the log-likelihood ratio $\int_{\hat{p}>0} \log(p_0/\hat{p})\, dQ$ depends on the version of the MLE $\hat{p}$. In order to recover the same order as in Corollary 5.7 for any $\hat{p}$, some additional conditions on Q seem necessary. We shall not further investigate this issue herein.

EXAMPLE 4. (A finite-dimensional convex model). The model below has been analyzed by Rolin (1992). Let $\mathscr{P} = \{p_\theta,\ \theta \in \Theta = [-1, 1]\}$ with $p_\theta(x) = (1 + \theta x)/2$, $x \in [-1, 1]$, and observe that $\sup_{\theta\in\Theta} \int \log p_\theta\, dQ$ is finite. By a concavity argument we obtain the existence of θ_0, the pseudo-true value of the parameter. For any $\theta \in \Theta$ consider the score function $s(\theta) = \int x/(1 + \theta x)\, dQ(x) \in [-\infty, \infty]$ and note that if $s(-1) > 0 > s(1)$, then $\theta_0 \in (-1, 1)$. Moreover, $\int x\, dQ(x)$ and θ_0 have the same sign. In particular, if $\int x\, dQ(x) = 0$, then $\theta_0 = 0$, that is the pseudo-true density is uniform on $[-1, 1]$. On the other hand, if $s(-1) > s(1) \geq 0$, then $\theta_0 = 1$. Finally, $\theta_0 = -1$ if $0 \geq s(-1) > s(1)$.

The MLE $\hat{\theta}$ exists, is unique and can be characterized via the empirical counterpart of the score function $s(\cdot)$ defined above (see Rolin (1992)). The almost sure convergence of $\hat{\theta}$ to θ_0 can be obtained as a consequence of Corollary 3.4.

If $\theta_0 \in (-1, 1)$ we may deduce $H_B(u, \mathscr{G}(\delta), Q) \leq C \log(\delta/u)$, $0 < u \leq \delta$, where $\mathscr{G}(\delta) = \{(2p_\theta/(p_\theta + p_{\theta_0}))\, \mathbb{1}_{\{q>0\}},\ \theta \in [\theta_0 - \delta, \theta_0 + \delta]\}$. Note that $p_\theta \in \mathscr{G}(\delta)$ is equivalent to $h_0(p_\theta, p_{\theta_0}) \leq C\delta$, with $C = C(\theta_0)$ some positive constant. Applying Proposition 4.1 for $\mathscr{G} = G(\delta)$ we deduce $h_0^2(p_{\hat{\theta}}, p_{\theta_0}) = O_{\mathbf{P}}(n^{-1})$. Moreover, the quantities in Corollary 4.2 are also of order $O_{\mathbf{P}}(n^{-1})$.

A more interesting situation occurs when $\theta_0 = 1$ or -1. Hereafter we consider $\theta_0 = 1$ (the case $\theta_0 = -1$ can be treated in a similar way) and we restrain the parameter space to $\Theta(\delta) = [1-\delta, 1]$, $0 < \delta < 1$. In order to simplify entropy computations let us assume from now on that

$$\sup_{\theta_1, \theta_2 \in \Theta(\delta)} \int \frac{1 + x}{(2 + (1 + \theta_1)x)(2 + (1 + \theta_2)x)}\, dQ(x) < \infty. \tag{5.14}$$

Under this assumption we may deduce, up to a constant, the same entropy bounds, and consequently the same rates, as in the case where θ_0 was in the interior of Θ. Observe that (5.14) is satisfied, for instance, when the model is well-specified (i.e., $dQ/d\mu = (1 + x)/2$). This case, an example of a well-specified model with the parameter on the boundary of the parameter space, was considered by Rolin (1992). In the case he considered, Rolin deduced that for any $\beta_n \to \infty$ and $y > 0$, $(\beta_n/\log\beta_n)s(1 - y/\beta_n) \to y/2$. Moreover, he showed that $\sqrt{n\log n}\,(1 - \hat\theta)$ converges in distribution to some nonnegative random variable. From this we may deduce

$$\int \left(\frac{2p_{\hat\theta}}{p_{\hat\theta} + p_{\theta_0}} - 1\right) dQ = \frac{\hat\theta - 1}{2}\, s\left(\frac{\hat\theta + 1}{2}\right) = O_{\mathbf{P}}\left(\frac{1}{n}\right),$$

and this order agrees with that obtained from Corollary 4.2. Rolin also proved that $(1/\log\beta_n)I(1 - y/\beta_n) \to y/2$ where

$$I(\theta) = \int \left(\frac{x}{1 + \theta x}\right)^2 dQ(x) = -s'(\theta).$$

From this we may write

$$\int \left(\frac{2p_{\hat\theta}}{p_{\hat\theta} + p_{\theta_0}} - 1\right)^2 dQ = \frac{(\hat\theta - 1)^2}{4}\, I\left(\frac{\hat\theta + 1}{2}\right) = O_{\mathbf{P}}\left(\frac{1}{n}\right)$$

which again coincides with the order given by Corollary 4.2.

The results obtained by Rolin (1992) are, of course, more precise for the case he considered. In particular he obtained the asymptotic distribution of the MLE in the non-standard case where the true parameter lies on the boundary of Θ. However, we are able to deduce all the orders above even under misspecification. The only simplifying assumption we impose is (5.14).

APPENDIX

PROOF OF LEMMA 2.5. (a) Denote $p_{1,\alpha} = \alpha p_1 + (1 - \alpha)p_2$. If $\alpha > \alpha'$, then $p_{1,\alpha}$ lies between $p_{1,\alpha'}$ and p_1. Indeed, $p_{1,\alpha} = \beta p_1 + (1-\beta)p_{1,\alpha'}$ with $\beta = (\alpha - \alpha')/(1-\alpha') \in (0, 1)$. Observe that $h_\alpha(p_1, p_2) = h_\beta(p_1, p_{1,\alpha'})$ and $h_{\alpha'}(p_1, p_2) = h_0(p_1, p_{1,\alpha'})$. Thus we have to prove that

$$h_\beta(p_1, p_{1,\alpha'}) \le h_0(p_1, p_{1,\alpha'}), \tag{A.1}$$

for $\beta \in (0, 1)$. Note that $h_0(p_1, p_{1,\alpha'})$ is finite for all $\alpha' \in (0, 1)$ since $p_1/p_{1,\alpha'} \le 1/\alpha'$ [if $\alpha' = 0$ and $h_0(p_1, p_2) = \infty$ there is nothing to be proved]. Moreover

$$h_\beta^2(p_1, p_{1,\alpha'}) = \frac{1}{2}\int_{p_1 \ne p_{1,\alpha'}} \frac{(\sqrt{p_1} - \sqrt{p_{1,\alpha'}})^2}{p_{1,\alpha'}} \left(\frac{\sqrt{p_1} - \sqrt{p_{\beta,\alpha'}}}{\sqrt{p_1} - \sqrt{p_{1,\alpha'}}}\right)^2 \frac{p_{1,\alpha'}}{p_{\beta,\alpha'}}\, dQ,$$

where $p_{\beta,\alpha'} = \beta p_1 + (1 - \beta)p_{1,\alpha'}$. Since $\{p_1 \ne p_{1,\alpha'}\} \cap \{q > 0\} \subset \{p_{1,\alpha'} > 0\} \cap \{q > 0\}$, $p_{1,\alpha'}/p_{\beta,\alpha'} \le 1/(1 - \beta)$ and

$$\frac{\sqrt{p_1} - \sqrt{p_{\beta,\alpha'}}}{\sqrt{p_1} - \sqrt{p_{1,\alpha'}}}\mathbb{1}_{\{p_1 \ne p_{1,\alpha'}\}} = (1 - \beta)\frac{\sqrt{p_1} + \sqrt{p_{1,\alpha'}}}{\sqrt{p_1} + \sqrt{p_{\beta,\alpha'}}}\mathbb{1}_{\{p_1 \ne p_{1,\alpha'}\}} \le \sqrt{1 - \beta}\,\mathbb{1}_{\{p_1 \ne p_{1,\alpha'}\}},$$

we obtain (A.1).

(b) Herein $\|\cdot\|$ stands for $\|\cdot\|_{L_2(Q)}$. For the sake of simplicity, assume that $p_{1,} p_2, p_3 > 0$ on $\{q > 0\}$. Otherwise, the equations below should be completed with the corresponding indicator functions. First, note that

$$\begin{aligned}\sqrt{2}\, h_\alpha(p_1, p_2) &= \left\| \sqrt{\frac{p_1}{\alpha p_1 + (1-\alpha)p_2}} - 1 \right\| \\ &= \frac{1-\alpha}{\alpha} \left\| \frac{\alpha(p_1 - p_2)}{\alpha p_1 + (1-\alpha)p_2} \, \frac{1}{1 + \sqrt{\frac{p_1}{\alpha p_1 + (1-\alpha)p_2}}} \right\| \\ &= \frac{1-\alpha}{\alpha} \left\| \left(1 - \sqrt{\frac{p_2}{(1-\alpha)p_2 + \alpha p_1}}\right) \frac{1 + \sqrt{\frac{p_2}{(1-\alpha)p_2 + \alpha p_1}}}{1 + \sqrt{\frac{p_1}{\alpha p_1 + (1-\alpha)p_2}}} \right\| \\ &\leq \frac{1-\alpha}{\alpha} \, \frac{1 + \sqrt{1-\alpha}}{\sqrt{1-\alpha}} \sqrt{2}\, h_{1-\alpha}(p_2, p_1).\end{aligned}$$

For the first inequality to be proved consider two cases. If $\alpha \in [1/2, 1)$, from a) and the inequality above we have

$$h_\alpha(p_1, p_2) \leq h_{1-\alpha}(p_1, p_2) \leq \frac{\sqrt{\alpha}}{1-\alpha}(1 + \sqrt{\alpha}) h_\alpha(p_2, p_1).$$

On the other hand, if $\alpha \in (0, 1/2)$ we may write

$$h_\alpha(p_1, p_2) \leq \frac{\sqrt{1-\alpha}}{\alpha}(1+\sqrt{1-\alpha}) h_{1-\alpha}(p_2, p_1) \leq \frac{\sqrt{1-\alpha}}{\alpha}(1+\sqrt{1-\alpha}) h_\alpha(p_2, p_1).$$

For the second inequality to be proved we remark that, by the usual triangle inequality

$$h_\alpha(p_3, p_1) \leq \frac{1}{\sqrt{2}} \left\| \sqrt{\frac{p_1}{\alpha p_1 + (1-\alpha)p_3}} - \sqrt{\frac{p_2}{\alpha p_2 + (1-\alpha)p_3}} \right\| + h_\alpha(p_2, p_3).$$

But the first term on the right hand side of the previous inequality equals

$$\begin{aligned}\frac{1}{\sqrt{2}} \Bigg\| \frac{(1-\alpha)(p_1 - p_2)}{\alpha p_1 + (1-\alpha)p_2} &\cdot \frac{p_3(\alpha p_1 + (1-\alpha)p_2)}{\sqrt{(\alpha p_1 + (1-\alpha)p_3)(\alpha p_2 + (1-\alpha)p_3)}} \\ &\times \frac{1}{(\sqrt{p_1(\alpha p_2 + (1-\alpha)p_3)} + \sqrt{p_2(\alpha p_1 + (1-\alpha)p_3)}\,)} \Bigg\|.\end{aligned}$$

The second ratio in the previous norm can be bounded by $1/(1-\alpha)\sqrt{\alpha}$. Moreover,

$$\left\| \frac{p_1}{\alpha p_1 + (1-\alpha)p_2} - 1 \right\| \leq \sqrt{2}\left(1 + \frac{1}{\sqrt{\alpha}}\right) h_\alpha(p_1, p_2).$$

Thus, we obtain $h_\alpha(p_1, p_3) \leq \big(1 + \sqrt{\alpha}\big) \left[\alpha(1-\alpha)\right]^{-1} \left(h_\alpha(p_1, p_2) + h_\alpha(p_2, p_3)\right)$. □

PROOF OF PROPOSITION 4.1. It suffices to prove the result for $\alpha = 0$ (see Lemma 2.5). We follow the lines of the proofs of Theorems 7.6 and 7.7 of van de Geer (2000). For any $n \geq 1$, let $A_n = \{p_0 \leq \sigma(\delta_n)\}$. Recall that $g_p = \left(p/p_{1/2}\right) \mathbb{1}_{\{q>0\}}$ and thus $\left|g_p - 1\right| \leq 1,\ p \in \mathscr{P}$. Using (4.5) we deduce that, for any $c > 0$

$$\mathbf{P}\left(h_0(\hat{p}, p_0) > (c+2)\delta_n\right) \leq \mathbf{P}(B_1) + \mathbf{P}(B_2) \leq \frac{1}{nc^2\delta_n^2} + \mathbf{P}(B_2),$$

where $B_1 = \{\int_{A_n}(g_{\hat{p}} - 1)d(Q_n - Q) > (c+2)\delta_n^2\}$ and B_2 denotes the set $B_1^c \cap \{h_0(\hat{p}, p_0) > (c+2)\delta_n\}$. Since the inequality

$$\tfrac{1}{2}\, h_0^2(\hat{p}, p_0) \leq \int_{A_n^c} (g_{\hat{p}} - 1)\, d\,(Q_n - Q)$$

holds on B_2 [see inequality (2.6)], we obtain

$$\mathbf{P}(B_2) \leq \mathbf{P}\left(\sup_{h_0(p,p_0)>(c+2)\delta_n} \int_{A_n^c} (g_p - 1)\, d\,(Q_n - Q) - \frac{1}{2} h_0^2(p, p_0) \geq 0\right) \overset{\text{def}}{=} \mathbf{P}(B_3).$$

Now, it remains to bound $\mathbf{P}(B_3)$. Define $S = \min\{s \in \mathbb{N},\ 2^{s+1}(c+2)\delta_n \geq 1\}$. Since $h_0(p, p_0) \leq 1,\ p \in \mathscr{P}$ we have

$$\begin{aligned}&\mathbf{P}(B_3)\\ &\quad \leq \sum_{s=0}^{S} \mathbf{P}\left(\sup_{h_0(p,p_0)\leq 2^{s+1}(c+2)\delta_n} \sqrt{n}\int_{A_n^c} (g_p - 1)\, d(Q_n - Q) \geq \sqrt{n}2^{2s-1}(c+2)^2\delta_n^2\right).\end{aligned}$$

Let $\rho_K^2(g) = 2K^2 \int [\exp(|g|/K) - 1 - |g|/K]\, dQ$, with $K > 0$ and g a real-valued function. From $e^x - 1 - x \leq x^2$, $x \in [0, 1]$ and (4.4), we deduce $\rho_4^2(g_p - 1) \leq 8h_0^2(p, p_0),\ p \in \mathscr{P}$. Next, proceed as in Theorem 7.6 of van de Geer (2000). That is, bound each term of the previous sum using a uniform inequality based on generalized entropy with bracketing which is defined in the same way as the entropy with bracketing but with $\rho_K(\cdot)$ replacing the $L_2(Q)$-norm. □

PROOF OF LEMMA 5.1. (a) Fix $x \geq 1$ and θ a probability on H which is not the Dirac mass concentrated at the origin. Consider an interval $[s, t] \subset H\backslash\{0\}$. We obtain

$$\begin{aligned}\log\ p_\theta(x) &\geq \log \int_{[s,t]} p_\eta(x)\, d\theta(\eta)\\ &= \log \theta([s, t]) + \log\left[\theta([s, t])^{-1} \int_{[s,t]} p_\eta(x)\, d\theta(\eta)\right]\\ &\geq \log \theta([s, t]) + \theta([s, t])^{-1} \int_{[s,t]} \log p_\eta(x)\, d\theta(\eta).\end{aligned}$$

where the last inequality is due to Jensen's inequality. A PSD defined by a function $a(\cdot)$ as in (5.1) satisfies $a_x \geq b_1^x/x!,\ x \geq 1$. Indeed, if $b(\eta)$ denotes the logarithm of $a(\eta)$, then $a^{(x)}(0) = a(0)\{[b'(0)]^x + c_x\}$, with $c_x \geq 0$, ($a^{(x)}$ denotes the xth derivative of the function a). Therefore,

$$\log\ p_\theta(x) \geq \log \theta([s,t]) + x\ \log b_1 - \log(x!) + x\ \log s - b(t)$$

and thus, if the right-hand side is integrable with respect to Q, then $\int \log p_\theta\, dQ > -\infty$. Due to Stirling's formula, $\forall x \geq 2$, there exists $\lambda_x \in (0,1)$ such that

$$\text{(A.2)} \qquad \log(x-1)! = (x - 1/2)\log x - x + (1/2)\log 2\pi + \lambda_x/(12x).$$

Thus $\int \log p_\theta\, dQ > -\infty$ if $\int x\ \log(x+1)\, dQ(x) < \infty$. Since for PSD mixtures we have $\int \log p_\theta\, dQ < 0$, we deduce that $\sup_\theta \int \log p_\theta\, dQ$ is finite.

(b) Assume that $\sup_\theta \int \log p_\theta\, dQ$ is finite and fix θ a probability on H with $\int \log p_\theta\, dQ > -\infty$. Since, $\forall\ x \geq 1$ and $\eta \in H$, $\eta^x \exp(-\eta) \leq x^x \exp(-x)$, use (A.2) and deduce that there exists a constant C such that, for any $x \geq 2$

$$\log\ p_\theta(x) \leq x\log x - x - \log(x\,!) + \log\ \theta[(0,\infty)] \leq -(1/2)\ \log(x+1) + C.$$

From this we have

$$-\infty < \int_{\{x\geq 2\}} \log p_\theta(x)\, dQ(x) \leq -\frac{1}{2}\int_{\{x\geq 2\}} \log(x+1)\ dQ(x) + C.$$

Thus, necessarily $\int \log p_\theta\, dQ > -\infty$ implies $\int \log(\,x+1)\, dQ(x) < \infty$. For the converse implication define, for example, $\theta(\{x\}) = 6/[\pi^2(x+1)^2]$, $x \geq 0$. Use again (A.2) and deduce that there exists a constant C' such that, for any $x \geq 1$

$$\begin{aligned}\log\ p_\theta(x) &> \log(6/\pi^2) - 2\ \log(x+1) + x\ \log\ x - x - \log(x!)\\ &> -\ (5/2)\ \log(x+1) + C'.\end{aligned}$$

Now, it is clear that $\sup_\theta \int \log p_\theta\, dQ$ is finite provided that $\int \log(x+1)\, dQ(x)$ is finite. □

Acknowledgments. This paper represents a part of my Ph.D. thesis at the Institute of Statistics, Louvain-la-Neuve. The work presented herein has been inspired and motivated by some lectures held by Professor van de Geer for Ph.D. students, in Utrecht. I thank Professor Rolin for his helpful advice. The suggestions of two referees greatly improved a former version of this paper.

REFERENCES

[1] BARLOW, R. E., BARTHOLOMEW, D. J., BREMNER, J. M. and BRUNK, H. D. (1972). *Statistical Inference under Order Restrictions*. Wiley, New York.

[2] BARLOW, R. E, MARSHALL, A. and PROCHAN, F. (1963). Properties of probability distributions with monotone hazard rates. *Ann. Math. Statist.* **34** 375–389.

[3] BILLINGSLEY, P. (1968). *Convergence of Probability Measures*. Wiley, New York.

[4] BIRGÉ, L. (1989). The Grenander estimator: a nonasymptotic approach. *Ann. Statist.* **17** 1532–1549.

[5] BIRGÉ, L. and MASSART, P. (1993). Rates of convergence for minimum contrast estimators. *Probab. Theory Related Fields* **97** 113–150.

[6] DAHLHAUS, R. and WEFELMEYER, W. (1996). Asymptotically optimal estimation in misspecified time series models. *Ann. Statist.* **24** 952–974.

[7] GRENANDER, U. (1956). On the theory of mortality measurement, part II. *Scand. Actuar. J.* **39** 125–153.

[8] HUBER, P. (1967). The behavior of maximum likelihood estimates under nonstandard conditions. *Proc. Fifth Berkeley Symp. Math. Statist. Probab.* **1** 221–233. Univ. California Press, Berkeley.

[9] LAMBERT, D. and TIERNEY, L. (1984). Asymptotic properties of the maximum likelihood estimates in mixed Poisson model. *Ann. Statist.* **12** 1388–1399.

[10] LINDSAY B. (1983). The geometry of mixture likelihood: a general theory. *Ann. Statist.* **11** 86–94.

[11] KOLTCHINSKII, V. I. (1997). *M*-estimation, convexity and quantiles. *Ann. Statist.* **25** 435–477.

[12] LÜXMANN-ELLINGHAUS, U. (1987). On the identifiability of mixture of infinitely divisible power series distributions. *Statist. Probab. Lett.* **5** 375–379.

[13] MARSHALL, A. W. (1970). Discussion of Barlow and van Zwet's paper "Asymptotic properties of isotonic estimators for the generalized failure rate function: strong consistency." In *Nonparametric Techniques in Statistical Inference* (M. L. Puri, ed.) 174–176. Cambridge Univ. Press.

[14] MARSHALL, A. W. and PROSCHAN, F. (1965). Maximum likelihood estimation for distribution with monotone failure rate. *Ann. Math. Statist.* **36** 69–77.

[15] NOACK, A. (1950). A class of random variables with discrete distribution *Ann. Math. Statist.* **21** 127–132.

[16] MILHAUD, X. and MOUNIME, S. (1996). A modified maximum likelihood estimator for infinite mixtures. Preprint, Univ. P. Sabatier, Toulouse.

[17] PATILEA, V. (1997). Convex models, NPLME and misspecification, Ph.D dissertation, part I. Institute of Statistics, Univ. catholique de Louvain.

[18] PFANZAGL J. (1969). On the measurability and consistency of minimum contrast estimates. *Metrika* 249–272.

[19] PFANZAGL J. (1988). Consistency of maximum likelihood estimators for certain nonparametric families, in particular: Mixtures. *J. Statist. Plann. Inference* **19** 137–158.

[20] PFANZAGL, J. (1990). Large deviations probabilities for certain nonparametric maximum likelihood estimators. *Ann. Statist.* **18** 1868–1877.

[21] POLLARD, D. (1984). *Convergence of Stochastic Processes*. Springer, New York.

[22] ROLIN, J.-M. (1992). Personal notes.

[23] ROBERTS, W. and VARBERG, D. (1973). *Convex Functions*. Academic Press, New York.

[24] SIMAR, L. (1976). Maximum likelihood estimation of a compound Poisson process. *Ann. Statist.* **4** 1200–1209.

[25] VAN DE GEER, S. (1993). Hellinger consistency of certain nonparametric maximum likelihood estimators. *Ann. Statist.* **21** 14–44.

[26] VAN DE GEER, S. (2000). *Empirical Processes in M-estimation*. Cambridge Univ. Press.

[27] VAN DER VAART, A. and WELLNER, J. (1996). *Weak Convergence and Empirical Process.* Springer, New York.

[28] WANG, J.-L. (1985). Strong consistency of approximate maximum likelihood estimators with applications to nonparametrics. *Ann. Statist.* **13** 932–946.

[29] WONG, W. H. and SHEN, X. (1995). Probabilities inequalities for likelihood ratios and convergence rates of sieve MLE's. *Ann. Statist.* **23** 339–362.

LABORATOIRE D'ECONOMIE D'ORLÉANS
UNIVERSITÉ D'ORLÉANS
ORLÉANS
FRANCE
E-MAIL: Valentin.Patilea@univ-orleans.fr