Asymptotic equivalence for density estimation and Guassian white noise: an extension

Ester Mariucci

Pub. Inst. Stat. Univ. Paris
60, fasc.1-2, 2016, 23-34

ASYMPTOTIC EQUIVALENCE FOR DENSITY ESTIMATION AND GAUSSIAN WHITE NOISE: AN EXTENSION

ESTER MARIUCCI

ABSTRACT. The aim of this paper is to present an extension of the well-known asymptotic equivalence between density estimation experiments and a Gaussian white noise model. Our extension consists in enlarging the nonparametric class of the admissible densities. More precisely, we propose a way to allow densities defined on any subinterval of $\mathbb{R}$, and also some discontinuous or unbounded densities are considered (so long as the discontinuity and unboundedness patterns are somehow known a priori). The concept of equivalence that we shall adopt is in the sense of the Le Cam distance between statistical models. The results are constructive: all the asymptotic equivalences are established by constructing explicit Markov kernels.

1. INTRODUCTION

When looking for asymptotic results for some statistical model it is often useful to profit from a global asymptotic equivalence, in the Le Cam sense, in order to be allowed to work in a simpler but equivalent model. Indeed, proving an asymptotic equivalence result means that one can transfer asymptotic risk bounds for any inference problem from one model to the other, at least for bounded loss functions. Roughly speaking, saying that two models, $\mathscr{P}_1$ and $\mathscr{P}_2$, are equivalent means that they contain the same amount of information about the parameter that we are interested in. For the basic concepts and a detailed description of the notion of asymptotic equivalence, we refer to [6, 7]. A short review of this topic will be given in the Appendix.

In recent years, numerous papers have been published on the subject of nonparametric asymptotic equivalence. For a non exhaustive list of the main ones among them, see, for example, the introduction in [8]. In this paper, we will focus on nonparametric density estimation experiments.

The seminal paper in this subject is due to Nussbaum [9]. There, the asymptotic equivalence between an experiment given by n observations of a density f on $[0,1]$ and a Gaussian white noise model:

$$dy_t = \sqrt{f(t)}dt + \frac{1}{2\sqrt{n}}dW_t, \quad t \in [0,1],$$

was established. Over the years several generalizations of this result have been proposed such as [1, 5, 2]. In [1], the authors obtained the global asymptotic equivalence between a Poisson process with variable intensity and a Gaussian white noise experiment with drift problem. Via Poissonization, this result was also extended to density estimation models. In [5] the authors proved the global asymptotic equivalence between a nonparametric model associated with the observation of independent but not identically distributed random variables on the unit interval and a bivariate Gaussian white noise model. More closely related to our work is the result of Carter in [2]. In that paper, he proposed a new approach to establish the same normal approximations to density estimations experiments as in [9]. While the

Date: October 6, 2015.

result in [9] is obtained by means of Poissonization, in [2] the key step is to connect the density estimation problem to a multinomial experiment and to simplify the latter with a multivariate normal experiment.

The purpose of the present work is to generalize [9] and [2]. More precisely, the density estimation experiments that we consider consist of n independent observations $(Y_i)_{i=1}^n$ defined on a interval $I \subseteq \mathbb{R}$ from some unknown distribution P_f^g having density (with respect to the Lebesgue measure on I) $\frac{dP_f^g}{dx}(x) = f(x)g(x)$. In particular, we do not require $I \subseteq \mathbb{R}$ to be bounded as is generally done in the existing literature. The function g is supposed to be known whereas f is unknown and belongs to a certain nonparametric functional class $\mathscr{F}$. Formally, the statistical model we consider is

$$\mathscr{P}_n^g = \left(\mathbb{R}^n, \mathscr{B}(\mathbb{R}^n), \{\otimes_{i=1}^n P_f^g : f \in \mathscr{F}\}\right). \tag{1}$$

The exact assumptions on f and g will be specified in Section 2. Here, let us only stress the fact that f has to be bounded away from zero and infinity and sufficiently regular, whereas g can be both unbounded and discontinuous. The advantage with respect to the earlier works is that this framework allows us to treat densities of the form $h = fg$ not necessarily bounded nor smooth. See Section 3.1 for a discussion about the hypotheses.

Finally, let us introduce the Gaussian white noise model. For that, let us denote by $(C, \mathscr{C})$ the space of continuous mappings from I into $\mathbb{R}$ endowed with its standard filtration and by $\mathbb{W}_f^g$ the law induced on $(C, \mathscr{C})$ by a stochastic process satisfying:

$$dY_t = \sqrt{f(t)g(t)}dt + \frac{dW_t}{2\sqrt{n}}, \quad t \in I, \tag{2}$$

where $(W_t)_{t\in\mathbb{R}}$ is a Brownian motion on $\mathbb{R}$ conditional on $W_0 = 0$. Then we set

$$\mathscr{W}_n^g = \left(C, \mathscr{C}, \{\mathbb{W}_f^g : f \in \mathscr{F}\}\right). \tag{3}$$

Let Δ be the Le Cam pseudo-distance between statistical models having the same parameter space. For the convenience of the reader a formal definition is given in Section 4.2. Our main result is then as follows (see Theorem 3.1 for the precise statement):

Main result 1.1. *Let I be a possibly infinite subinterval of $\mathbb{R}$ and let $\mathscr{F}$ consist of functions bounded away from 0 and ∞, satisfying the regularity assumptions stated in Section 2. Then, we have*

$$\lim_{n\to\infty} \Delta(\mathscr{P}_n^g, \mathscr{W}_n^g) = 0. \tag{4}$$

In some special cases an explicit upper bound for the rate of convergence in (4) is available; see, e.g. Corollary 3.2. The structure of the proof follows Carter's in [2], but we detach from it on several aspects. The basic idea is to use his multinomial-multivariate normal approximation, but some technical points have to be taken into account. One of these is that I may be infinite, so that, in particular, the subintervals J_i in which it is partitioned cannot be of equal length. We choose intervals J_i of varying length, according to the quantiles of ν_0, the measure having density g with respect to Lebesgue. This kind of partitions was already considered in [8].

The paper is organized as follows. Section 2 fixes the assumptions on the parameter space $\mathscr{F}$. Section 3 contains the statement of the main results and a discussion while Section 4 is devoted to the proofs. The paper includes an Appendix recalling the definition and some useful properties of the Le Cam distance.

Fix a finite measure ν_0 concentrated on a possibly infinite interval $I \subset \mathbb{R}$, admitting a density $g > 0$ with respect to Lebesgue. The class of functions $\mathscr{F}$ will be considered as a class of probability densities with respect to ν_0, i.e. $\int_I f(x)g(x)dx = 1$. For each $f \in \mathscr{F}$, let ν (resp. $\hat{\nu}_m$) be the measure having f (resp. $\hat{f}_m$) as a density with respect to ν_0 where, for every $f \in \mathscr{F}$, $\hat{f}_m(x)$ is defined as follows. Given a positive integer m, let $J_1 = I \cap (-\infty, v_1]$, $J_j := (v_{j-1}, v_j]$ for $j = 2, \ldots, m-1$ and $J_m = I \cap (v_m, \infty)$ where the v_j's are the quantiles for ν_0, i.e.

$$\mu_n := \nu_0(J_j) = \frac{\nu_0(I)}{m}, \quad \forall j = 1, \ldots, m. \tag{5}$$

Define $x_j^* := \frac{\int_{J_j} x\nu_0(dx)}{\mu_n}$, $j = 1, \ldots, m$ and introduce a sequence of continuous functions $0 \le V_j \le \frac{1}{\mu_n}$, $j = 1, \ldots, m$, defined in the following way.

- V_1 is supported on $I \cap (-\infty, x_2^*]$ and:

$$V_1|_{I\cap(-\infty,x_1^*]} \equiv \frac{1}{\mu_n}; \quad \int_{x_1^*}^{x_2^*} V_1(x)\nu_0(dx) = \frac{\nu_0((x_1^*, v_1])}{\mu_n}; \quad V_1(x_2^*) = 0.$$

- For $j = 2, \ldots, m-1$, V_j is supported in $[x_{j-1}^*, x_{j+1}^*]$ and:

$$V_j|_{[x_{j-1}^*,x_j^*]} \equiv 1 - V_{j-1}|_{[x_{j-1}^*,x_{j+1}^*]}; \quad \int_{x_j^*}^{x_{j+1}^*} V_j(x)\nu_0(dx) = \frac{\nu_0((x_j^*, v_j])}{\mu_n}; \quad V_j(x_{j+1}^*) = 0.$$

- For $j = m$, V_m is supported on $I \setminus (-\infty, x_{m-1}^*)$ and:

$$V_m|_{[x_{m-1}^*,x_m^*]} \equiv 1 - V_{m-1}|_{[x_{m-1}^*,x_m^*]} \quad \text{and} \quad V_m|_{I\cap(-\infty,x_m^*)} \equiv \frac{1}{\mu_n}.$$

(It is immediate to check that such a choice is always possible). Observe that, by construction,

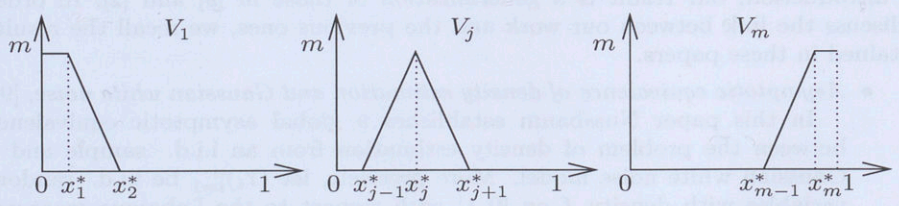
$$\sum_{j=1}^{m} V_j(x)\mu_n = 1, \quad \forall x \in I \quad \text{and} \quad \int_I V_j(y)\nu_0(dy) = 1.$$

Define

$$\hat{f}_m(x) = \sum_{j=1}^{m} V_j(x) \int_{J_j} f(y)\nu_0(dy).$$

The same construction of the V_j's already appeared in our previous work [8]. Their definition is modelled on the following example:

Example 2.1. Let ν_0 be the Lebesgue measure on $[0,1]$. Then $v_j = \frac{j-1}{m}$ and $x_j^* = \frac{2j-1}{2m}$, $j = 1, \ldots, m$. The standard choice for V_j (based on the construction by [2]) is given by the piecewise linear functions interpolating the values in the points x_j^* specified above:



We now explain the assumptions we will need to make on the parameter f. We require that:

(H1) There exist constants $\kappa, M > 0$ such that $\kappa \leq f(y) \leq M$, for all $y \in I$ and $f \in \mathscr{F}$.

The m introduced above will be considered as a function of n, $m = m_n$. We can thus consider $\widehat{\sqrt{f}}_m$, the approximation of $\sqrt{f}$ constructed as $\hat{f}_m$ above and introduce the quantities:

$$H_m^2(f) := \int_I \left(\sqrt{f(x)} - \sqrt{\hat{f}_m(x)}\right)^2 \nu_0(dx),$$

$$A_m^2(f) := \int_I \left(\widehat{\sqrt{f}}_m(y) - \sqrt{f(y)}\right)^2 \nu_0(dy),$$

$$B_m^2(f) := \sum_{j=1}^{m} \left(\int_{J_j} \frac{\sqrt{f(y)}}{\sqrt{\nu_0(J_j)}} \nu_0(dy) - \sqrt{\nu(J_j)}\right)^2.$$

We will assume the existence of a sequence of discretizations $m = m_n$ and functions V_j, $j = 1, \ldots, m$, such that:

(C1) $\lim_{n\to\infty} n \sup_{f\in\mathscr{F}} \left(H_m^2(f) + A_m^2(f) + B_m^2(f)\right) = 0.$

3. Main results and discussion

Using the notation introduced in Section 2, we now state our main result in terms of the models $\mathscr{P}_n^g$ and $\mathscr{W}_n^g$ defined in (1) and (3), respectively.

Theorem 3.1. *Let ν_0 be a finite measure concentrated on an (possibly infinite) interval $I \subset \mathbb{R}$ having density $g > 0$ with respect to Lebesgue. Suppose that there exist a sequence $m = m_n$ and functions V_j, $j = 1, \ldots, m$, such that every $f \in \mathscr{F}$ satisfies conditions (H1) and (C1). Then, for n big enough we have:*

$$\Delta(\mathscr{P}_n^g, \mathscr{W}_n^g) = O\left(\sqrt{n} \sup_{f\in\mathscr{F}} \left(A_m(f) + B_m(f) + H_m(f)\right) + \frac{m \ln m}{\sqrt{n}}\right).$$

Corollary 3.2. *Let I be a compact subset of $\mathbb{R}$ and let ν_0 be the Lebesgue measure on I. For fixed $\gamma \in (0, 1]$ and K, κ, M strictly positive constants, consider the functional class*

$$\mathscr{F}_{(\gamma,K,\kappa,M)} = \left\{f \in C^1(I) : \kappa \leq f(x) \leq M,\ |f'(x) - f'(y)| \leq K|x-y|^\gamma,\ \forall x, y \in I\right\}.$$

Suppose $\mathscr{F} \subset \mathscr{F}_{(\gamma,K,\kappa,M)}$. Then

$$\Delta(\mathscr{P}_n^g, \mathscr{W}_n^g) = O\left(n^{-\frac{\gamma}{\gamma+2}} \log n\right).$$

3.1. **Existing literature and discussion.** As it has already been highlighted in the introduction, our result is a generalization of those in [9] and [2]. In order to discuss the link between our work and the previous ones, we recall the results contained in these papers.

- *Asymptotic equivalence of density estimation and Gaussian white noise*, [9].
  In this paper Nussbaum establishes a global asymptotic equivalence between the problem of density estimation from an i.i.d. sample and a Gaussian white noise model. More precisely, let $(Y_i)_{i=1}^n$ be i.i.d. random variables with density f on $[0, 1]$ with respect to the Lebesgue measure. The densities f are the unknown parameters and they are supposed to belong to a certain nonparametric class $\mathscr{F}$ subject to a Hölder restriction: $|f(x) - f(y)| \leq C|x - y|^\alpha$ with $\alpha > \frac{1}{2}$ and a positivity restriction: $f(x) \geq \varepsilon > 0$. Let us denote by $\mathscr{P}_{1,n}$ the statistical model associated with

the observation of the Y_i's. Furthermore, let $\mathscr{P}_{2,n}$ be the experiment in which one observes a stochastic process $(Y_t)_{t\in[0,1]}$ such that

$$dY_t = \sqrt{f(t)}dt + \frac{1}{2\sqrt{n}}dW_t, \quad t \in [0,1],$$

where $(W_t)_{t\in[0,1]}$ is a standard Brownian motion. Then the main result in [9] is that $\Delta(\mathscr{P}_{1,n}, \mathscr{P}_{2,n}) \to 0$ as $n \to \infty$.

This is done by first showing that the result holds for certain subsets $\mathscr{F}_n(f_0)$ of the class $\mathscr{F}$ described above. Then it is shown that one can estimate the f_0 rapidly enough to fit the various pieces together. Without entering into any detail, let us just mention that the key steps are a Poissonization technique and the use of a functional KMT inequality.

- *Deficiency distance between multinomial and multivariate normal experiments,* [2].

  In this paper Carter establishes a global asymptotic equivalence between a density estimation model and a Gaussian white noise model by bounding the Le Cam distance between multinomial and multivariate normal random variables. More precisely, let us denote by $\mathcal{M}(n,\theta)$ the multinomial distribution, where $\theta := (\theta_1, \ldots, \theta_m)$. Denote the covariance matrix nV_θ: Its (i,j)th element equals to $n\theta_i(1-\theta_i)\delta_{i,j} - n\theta_i\theta_j$.

  The main result is an upper bound for the Le Cam distance $\Delta(\mathcal{M},\mathcal{N})$ between the models $\mathcal{M} := \{\mathcal{M}(n,\theta) : \theta \in \Theta\}$ and $\mathcal{N} := \{\mathcal{N}(n\theta, nV_\theta) : \theta \in \Theta\}$, under some regularity assumptions on Θ. In particular, Carter proves that

$$\Delta(\mathcal{M},\mathcal{N}) \le C'_\Theta \frac{m \ln m}{\sqrt{n}} \quad \text{provided} \quad \sup_{\theta\in\Theta} \frac{\max_i \theta_i}{\min_i \theta_i} \le C_\Theta < \infty,$$

  for a constant C'_Θ that depends only on C_Θ. From this inequality Carter can recover mostly the same results as [9] under stronger regularity assumptions on $\mathscr{F}$: $\mathscr{F}$ is a class of smooth, differentiable densities f on the interval $[0,1]$ such that there exist strictly positive constants ε, M, γ such that $\varepsilon \le f \le M$ and

$$|f'(x) - f'(y)| \le M|x-y|^\gamma, \quad \text{for all } x, y \in [0,1].$$

  Let us briefly explain how one can use a bound on the distance between multinomial and multivariate normal variables to make assertions about density estimation experiments. The idea is to see the multinomial experiment as the result of grouping independent observations from a continuous density into subsets. Using the square root as a variance-stabilizing transformation, these multinomial variables can be asymptotically approximated by normal variables with constant variances. These normal variables, in turn, are approximations to the increments of the Brownian motion processes over the sets in the partition.

Our work can be seen as a generalization of these two works: to see that it is enough to take $g(x) = \mathbb{I}_{[0,1]}(x)$ as in Corollary 3.2. However, it differs from Nussbaum and Carter's results in several aspects. First of all, we do not need to ask the random variables to be defined on $[0,1]$, allowing the observations to be defined on a possibly infinite interval I of $\mathbb{R}$. Secondly, in our setting the positivity restriction on the densities can be removed. Indeed, as an example, we can consider the class of densities of the form

$$h(x) = \frac{f(x)}{x^2}\mathbb{I}_{[A,\infty)}(x), \quad \forall A > 0$$

where f belongs to the functional class $\mathscr{F}_{(\gamma,K,\kappa,M)}$ as defined in Corollary 3.2. A proof that this kind of densities satisfies Hypotheses (H1) and (C1) can be found in [8], Section 5.2.

More generally, density functions h that can be written in form of a product are commonly used in statistics. One could cite as a simple case the problem of a parametric estimation for a Weibull density, see, e.g. [4, 3]. Generally speaking, the present work can be useful whenever the random variables Y_i's do not admit a smooth density h with respect to Lebesgue, but nevertheless one has some information on the discontinuity structure, namely one knows g in the decomposition $h(x) = f(x)g(x)$.

4. Proofs

4.1. **Proof of Theorem 3.1.** We will proceed in four steps.

Step 1. By means of Facts 4.4 and 4.5, we get

$$\left\| \bigotimes_{i=1}^{n} P_f^g - \bigotimes_{i=1}^{n} P_{\hat{f}_m}^g \right\|_{TV} \leq H\left(\bigotimes_{i=1}^{n} P_f^g, \bigotimes_{i=1}^{n} P_{\hat{f}_m}^g \right) \leq \sqrt{nH^2(P_f^g, P_{\hat{f}_m}^g)}.$$

Hence, denoting by $\hat{\mathscr{P}}_n^g$ the statistical model associated with the family of probabilities $\{ \otimes_{i=1}^n P_{\hat{f}_m}^g : f \in \mathscr{F} \}$:

$$\Delta(\mathscr{P}_n^g, \hat{\mathscr{P}}_n^g) \leq \sup_{f \in \mathscr{F}} \sqrt{n \int_I \left(\sqrt{f(x)} - \sqrt{\hat{f}_m(x)} \right)^2 g(x) dx}. \tag{6}$$

Step 2. Following the same approach as in [2], we introduce an auxiliary multinomial experiment to get closer to a normal one representing the increments of $(Y_t)_{t \in I}$ defined as in (2). The multinomial experiment is linked with the density estimation model in the following way: Let $(Y_i)_{i=1}^n$ be a sequence of i.i.d. random variables with density fg with respect to Lebesgue and define the multinomial experiment by grouping their observations into subsets. More precisely, let us introduce the random variables:

$$Z_i = \sum_{j=1}^{n} \mathbb{I}_{J_i}(Y_j), \ i = 1, \ldots, m.$$

Observe that the law of the vector $(Z_1, \ldots, Z_m)$ is multinomial $\mathcal{M}(n; \gamma_1, \ldots, \gamma_m)$ where

$$\gamma_i = \int_{J_i} f(x)g(x)dx, \quad i = 1, \ldots, m.$$

Let us denote by $\mathcal{M}_m$ the statistical model associated with the observation of $(Z_1, \ldots, Z_m)$. Clearly $\delta(\mathscr{P}_n^g, \mathcal{M}_m) = 0$. Indeed, $\mathcal{M}_m$ is the image experiment by the random variable $S : I^n \to \{1, \ldots, n\}^m$ defined as

$$S(x_1, \ldots, x_n) = \Big(\#\{j : x_j \in J_1\}; \ldots; \#\{j : x_j \in J_m\} \Big),$$

where $\#A$ denotes the cardinal of the set A. To conclude the second step we now prove that the multinomial experiment is as informative as $\hat{\mathscr{P}}_n^g$.

Lemma 4.1.

$$\delta(\mathcal{M}_m, \hat{\mathscr{P}}_n^g) = 0.$$

Proof. We need to produce an explicit Markov kernel that allows to approximate the density $\hat{f}_m g$ given an observation from the multinomial model which is given by

$$K\big((k_1, \ldots, k_m), A\big) = \int_A \mathbb{E}\Big[V_{X_{(k_1, \ldots, k_m)}}(x) \Big] \nu_0(dx), \quad \forall (k_1, \ldots, k_m) \in \mathbb{N}, \ \sum_i k_i = n, \ A \subset \mathbb{R}$$

where $X_{(k_1,\dots,k_m)} \in \{1,\dots,m\}$ is a randomly chosen integer obtained assigning to j the weight $\frac{k_j}{n}$. □

Step 3. Let us denote by $\mathcal{N}_m$ the statistical model associated with the observation of m independent Gaussian variables $\mathcal{N}(\sqrt{n\gamma_i}, \frac{1}{4})$, $i = 1,\dots,m$. Since $\frac{\max\gamma_i}{\min\gamma_i} \le \frac{M}{\kappa}$, one can apply Theorem 4.9 obtaining

$$\Delta(\mathcal{M}_m, \mathcal{N}_m) = O\Big(\frac{m \ln m}{\sqrt{n}}\Big).$$

Here the O depends only on M and κ.

Step 4. Finally, we conclude the proof of Theorem 3.1, by showing that

$$\Delta(\mathcal{N}_m, \mathcal{W}_n^g) \le 2\sqrt{n} \sup_{f\in\mathcal{F}} \big(A_m(f) + B_m(f)\big). \tag{7}$$

As a preliminary remark note that $\mathcal{W}_n^g$ is equivalent to the model that observes a trajectory from:

$$d\bar{y}_t = \sqrt{f(t)}g(t)dt + \frac{\sqrt{g(t)}}{2\sqrt{n}}dW_t, \quad t \in I.$$

In order to prove (7) we proceed in the following way: First of all, we prove that $\mathcal{N}_m$ is equivalent to the model that observes the increments on the intervals J_i of $(\bar{y}_t)_{t\in I}$. Secondly, we show that the increments of $(\bar{y}_t)_{t\in I}$ are more informative than another Gaussian process, say $(Y_t^*)_{t\in I}$, that turns out to be very close to $(\bar{y}_t)_{t\in I}$ in the total variation distance. We then conclude the asymptotic equivalence between $\mathcal{N}_m$ and $\mathcal{W}_n^g$ observing that the increments of $(\bar{y}_t)_{t\in I}$ are obviously less informative than $\mathcal{W}_n^g$.

Let us denote by $\bar{Y}_j$ the increments of the process $(\bar{y}_t)$ over the intervals J_j, $j = 1,\dots,m$, i.e.

$$\bar{Y}_j := \bar{y}_{v_j} - \bar{y}_{v_{j-1}} \sim \mathcal{N}\Big(\int_{J_j} \sqrt{f(y)}\nu_0(dy), \frac{\nu_0(J_j)}{4n}\Big)$$

and denote by $\bar{\mathcal{N}}_m$ the statistical model associated with the distributions of these increments. As announced we start by bounding the Le Cam distance between $\mathcal{N}_m$ and $\bar{\mathcal{N}}_m$ showing that

$$\Delta(\mathcal{N}_m, \bar{\mathcal{N}}_m) \le 2\sqrt{n} \sup_{f\in\mathcal{F}} B_m(f), \quad \text{for all m.} \tag{8}$$

In this regard, remark that the experiment $\bar{\mathcal{N}}_m$ is equivalent to another experiment, say $\mathcal{N}_m^{\#}$, that observes m independent Gaussian random variables of means equal to $\frac{2\sqrt{n}}{\sqrt{\nu_0(J_j)}}\int_{J_j}\sqrt{f(y)}\nu_0(dy)$, $j = 1,\dots,m$ and variances identically 1. Hence, using also Property 4.3, Facts 4.4–4.6 we get:

$$\Delta(\mathcal{N}_m, \bar{\mathcal{N}}_m) \le \Delta(\mathcal{N}_m, \mathcal{N}_m^{\#}) \le \sup_{f\in\mathcal{F}} \sqrt{\sum_{j=1}^m \Big(\frac{2\sqrt{n}}{\sqrt{\nu_0(J_j)}}\int_{J_j}\sqrt{f(y)}\nu_0(dy) - 2\sqrt{n\nu(J_j)}\Big)^2}.$$

Using similar ideas as in Section 8.2 of [2] and Lemma 3.2 of [8], we introduce a new stochastic process constructed from the random variables $\bar{Y}_j$'s. To that end define

$$Y_t^* = \sum_{j=1}^m \bar{Y}_j \int_{I\cap(-\infty,t]} V_j(y)\nu_0(dy) + \frac{1}{2\sqrt{n}}\sum_{j=1}^m \sqrt{\nu_0(J_j)}B_j(t), \quad t \in I, \tag{9}$$

where the $(B_j(t))_t$ are independent centered Gaussian processes conditional on $B_j(0) = 0$ with variances

$$\mathrm{Var}(B_j(t)) = \int_{I\cap(-\infty,t]} V_j(y)\nu_0(dy) - \Big(\int_{I\cap(-\infty,t]} V_j(y)\nu_0(dy)\Big)^2.$$

By construction, (Y_t^*) is a Gaussian process with mean and variance given by, respectively:

$$\mathbb{E}[Y_t^*] = \sum_{j=1}^m \mathbb{E}[\bar{Y}_j] \int_{I\cap(-\infty,t]} V_j(y)\nu_0(dy) = \sum_{j=1}^m \Big(\int_{J_j} \sqrt{f(y)}\nu_0(dy)\Big) \int_{I\cap(-\infty,t]} V_j(y)\nu_0(dy)$$

$$\begin{aligned}\mathrm{Var}[Y_t^*] &= \sum_{j=1}^m \mathrm{Var}[\bar{Y}_j]\Big(\int_{I\cap(-\infty,t]} V_j(y)\nu_0(dy)\Big)^2 + \frac{1}{4n}\sum_{j=1}^m \nu_0(J_j)\mathrm{Var}(B_j(t)) \\ &= \frac{1}{4n}\int_{I\cap(-\infty,t]} \sum_{j=1}^m \nu_0(J_j)V_j(y)\nu_0(dy) = \frac{1}{4n}\int_{I\cap(-\infty,t]} 1\nu_0(dy) = \frac{\nu_0(I\cap(-\infty,t])}{4n}.\end{aligned}$$

Therefore,

$$Y_t^* = \int_{I\cap(-\infty,t]} \widehat{\sqrt{f}}_m(y)\nu_0(dy) + \int_{I\cap(-\infty,t]} \frac{\sqrt{g(t)}}{2\sqrt{n}} W_t, \quad t \in I,$$

where

$$\widehat{\sqrt{f}}_m(x) := \sum_{j=1}^m \Big(\int_{J_j} \sqrt{f(y)}\nu_0(dy)\Big) V_j(x).$$

Applying Fact 4.7, we get that the total variation distance between the process $(Y_t^*)_{t\in I}$ constructed from the random variables $\bar{Y}_j$, $j = 1, \ldots, m$ and the Gaussian process $(\bar{y}_t)_{t\in I}$ is bounded by

$$\sqrt{4n\int_I \big(\widehat{\sqrt{f}}_m(y) - \sqrt{f(y)}\big)^2\nu_0(dy)},$$

as wanted.

4.2. **Proof of Corollary 3.2.** We start by proving a Lemma needed for the proof of Corollary 3.2. Since we are supposing that $g(x) = \mathbb{I}_I(x)$, we may take for the V_j the standard choice of triangular-trapezoidal functions (see Example 2.1 for a picture). Furthermore, $\mu_n = \nu_0(J_j) = \frac{1}{m}|I|$. For easiness of notations, in the proof we will also assume $I = [0, 1]$.

Lemma 4.2. *If $f \in \mathscr{F}_{(\gamma,K,\kappa,M)}$ and ν_0 is the Lebesgue measure on $[0, 1]$, then*

$$\|f - \hat{f}_m\|^2_{L_2(\nu_0)} \le O\Big(m^{-3} + m^{-2-2\gamma}\Big),$$

with the O depending on K, M and κ.

Proof. Let us consider the Taylor expansion of f at points x_i^*, where x denotes a point in J_i , $i = 1, \ldots, m$:

$$f(x) = f(x_i^*) + f'(x_i^*)(x - x_i^*) + R(x). \tag{10}$$

The smoothness condition on f allows us to bound the error R as follows:

$$\begin{aligned}|R(x)| &= \big|f(x) - f(x_i^*) - f'(x_i^*)(x - x_i^*)\big| \\ &= \big|f'(\xi_i) - f'(x_i^*)\big||\xi_i - x_i^*| \le Km^{-1-\gamma},\end{aligned}$$

where ξ_i is a certain point in J_i.

By the linear character of $\hat{f}_m$, we can write:

$$\hat{f}_m(x) = \hat{f}_m(x_i^*) + \hat{f}_m'(x_i^*)(x - x_i^*)$$

where $\hat{f}'_m$ denotes the left or right derivative of $\hat{f}_m$ in x_i^* depending whether $x < x_i^*$ or $x > x_i^*$. Let us observe that $\hat{f}'_m(x_i^*) = f'(\chi_i)$ for some $\chi_i \in J_i \cup J_{i+1}$ (here, we are considering right derivatives; for left ones, this would be $J_{i-1} \cup J_i$). To see that, take $x \in J_i \cap [x_i^*, x_{i+1}^*]$ and introduce the function $h(x) := f(x) - l(x)$ where

$$l(x) = \frac{x - x_i^*}{x_{i+1}^* - x_i^*}\big(\hat{f}_m(x_{i+1}^*) - \hat{f}_m(x_i^*)\big) + \hat{f}_m(x_i^*) = (x - x_i^*)\hat{f}'_m(x_i^*) + \hat{f}_m(x_i^*).$$

Then, using the fact that $\int_{J_i}(x - x_i^*)\nu_0(dx) = 0$ together with $\int_{J_{i+1}}(x - x_i^*)\nu_0(dx) = (x_{j+1}^* - x_j^*)\mu_n$, we get

$$\int_{J_i} h(x)\nu_0(dx) = 0 = \int_{J_{i+1}} h(x)\nu_0(dx).$$

In particular, by means of the mean theorem, one can conclude that there exist two points $p_i \in J_i$ and $p_{i+1} \in J_{i+1}$ such that

$$h(p_i) = \frac{\int_{J_i} h(x)\nu_0(dx)}{\nu_0(J_i)} = \frac{\int_{J_{i+1}} h(x)\nu_0(dx)}{\nu_0(J_{i+1})} = h(p_{i+1}).$$

As a consequence, we can deduce that there exists $\chi_i \in [p_i, p_{i+1}] \subseteq J_i \cup J_{i+1}$ such that $h'(\chi_i) = 0$, hence $f'(\chi_i) = l'(\chi_i) = \hat{f}'_m(x_i^*)$.

The fact that $\hat{f}'_m(x_i^*) = f'(t)$ for some $t \in J_i \cup J_{i+1}$, allows us to exploit the Hölder condition. Indeed, if $x \in J_i$, $i = 1, \ldots, m$, then there exists $t \in J_i \cup J_{i+1}$ such that:

$$\begin{aligned}|f(x) - \hat{f}_m(x)| &\le |f(x_i^*) - \hat{f}_m(x_i^*)| + |f'(x_i^*) - f'(t)||t - x_i^*| + |R(x)| \\ &\le |f(x_i^*) - \hat{f}_m(x_i^*)| + K|t - x_i^*|^{\gamma+1} + Km^{-1-\gamma} \le |f(x_i^*) - \hat{f}_m(x_i^*)| + 3Km^{-1-\gamma}.\end{aligned}$$

Using (10) and the fact that $\int_{J_i}(x - x_i^*)\nu_0(dx) = 0$, one gets:

$$|f(x_i^*) - \hat{f}_m(x_i^*)| = \frac{1}{\nu_0(J_i)}\left|\int_{J_i}(f(x_i^*) - f(x))\nu_0(dx)\right| \le Km^{-1-\gamma}.$$

Moreover, observe that, for all $x \in J_i$, $i = 1, \ldots, m$, $\left|f(x) - \frac{\nu(J_i)}{\nu_0(J_i)}\right|$, is bounded by $3Km^{-1-\gamma} + m^{-1}M$, indeed:

$$\begin{aligned}\left|f(x) - \frac{\nu(J_i)}{\nu_0(J_i)}\right| &= |f(x) - \hat{f}_m(x_i^*)| \le |f(x) - \hat{f}_m(x)| + |\hat{f}_m(x) - \hat{f}_m(x_i^*)| \\ &\le 3Km^{-1-\gamma} + |\hat{f}'_m(x_i^*)(x - x_i^*)| \le 3Km^{-1-\gamma} + Mm^{-1}.\end{aligned}$$

Collecting all the pieces together we find

$$\int_I \big(f(x) - \hat{f}_m(x)\big)^2\nu_0(dx) \le 2m^{-1}n\Big(3Km^{-1-\gamma} + Mm^{-1}\Big)^2 + 18K^2m^{-2-2\gamma}.$$

□

Proof of Corollary 3.2. By means of the fact that $f(x) \ge \kappa$ for all $x \in I$ one can write:

$$\begin{aligned}\int_I \Big(\sqrt{f(x)} - \sqrt{\hat{f}_m(x)}\Big)^2 dx &= \int_I \Big(\frac{f(x) - \hat{f}_m(x)}{\sqrt{f(x)} + \sqrt{\hat{f}_m(x)}}\Big)^2 dx \\ &\le \frac{1}{4\kappa}\int_I \big(f(x) - \hat{f}_m(x)\big)^2 dx.\end{aligned}$$

A straightforward application of Lemma 4.2 gives

$$H_m^2(f) = O\Big(m^{-3} + m^{-2-2\gamma}\Big).$$

The same bound holds for $A_m^2(f)$ since if $f \in \mathscr{F}_{(\gamma,K,\kappa,M)}$ then $\sqrt{f} \in \mathscr{F}_{(\gamma,\frac{K}{\sqrt{\kappa}},\sqrt{\kappa},\sqrt{M})}$. Moreover, one can see that B_m converges with the same rate as A_m. This may be done by explicit computations, see [8], Lemma 3.10 for more details. □

Background

Le Cam theory of statistical experiments. A *statistical model* or *experiment* is a triplet $\mathscr{P}_j = (\mathscr{X}_j, \mathscr{A}_j, \{P_{j,\theta}; \theta \in \Theta\})$ where $\{P_{j,\theta}; \theta \in \Theta\}$ is a family of probability distributions all defined on the same σ-field $\mathscr{A}_j$ over the *sample space* $\mathscr{X}_j$ and Θ is the *parameter space*. The *deficiency* $\delta(\mathscr{P}_1, \mathscr{P}_2)$ of $\mathscr{P}_1$ with respect to $\mathscr{P}_2$ quantifies "how much information we lose" by using $\mathscr{P}_1$ instead of $\mathscr{P}_2$ and it is defined as $\delta(\mathscr{P}_1, \mathscr{P}_2) = \inf_K \sup_{\theta\in\Theta} ||KP_{1,\theta} - P_{2,\theta}||_{TV}$, where TV stands for "total variation" and the infimum is taken over all "transitions" K (see [6], page 18). The general definition of transition is quite involved but, for our purposes, it is enough to know that (possibly randomized) Markov kernels are special cases of transitions. By $KP_{1,\theta}$ we mean the image measure of $P_{1,\theta}$ via the Markov kernel K, that is

$$KP_{1,\theta}(A) = \int_{\mathscr{X}_1} K(x, A) P_{1,\theta}(dx), \quad \forall A \in \mathscr{A}_2.$$

The experiment $K\mathscr{P}_1 = (\mathscr{X}_2, \mathscr{A}_2, \{KP_{1,\theta}; \theta \in \Theta\})$ is called a *randomization* of $\mathscr{P}_1$ by the Markov kernel K. When the kernel K is deterministic, that is $K(x, A) = \mathbb{I}_{S(x)\in A}$ for some random variable $S : (\mathscr{X}_1, \mathscr{A}_1) \to (\mathscr{X}_2, \mathscr{A}_2)$, the experiment $K\mathscr{P}_1$ is called the *image experiment by the random variable* S. The Le Cam distance is defined as the symetrization of δ and it defines a pseudometric. When $\Delta(\mathscr{P}_1, \mathscr{P}_2) = 0$ the two statistical models are said to be *equivalent*. Two sequences of statistical models $(\mathscr{P}_1^n)_{n\in\mathbb{N}}$ and $(\mathscr{P}_2^n)_{n\in\mathbb{N}}$ are called *asymptotically equivalent* if $\Delta(\mathscr{P}_1^n, \mathscr{P}_2^n)$ tends to zero as n goes to infinity. A very interesting feature of the Δ-distance is that it can be also translated in terms of statistical decision theory. Let $\mathscr{D}$ be any (measurable) decision space and let $L : \Theta \times \mathscr{D} \mapsto [0, \infty)$ denote a loss function. Let $\|L\| = \sup_{(\theta,z)\in\Theta\times\mathscr{D}} L(\theta, z)$. Let π_i denote a (randomized) decision procedure in the i-th experiment. Denote by $R_i(\pi_i, L, \theta)$ the risk from using procedure π_i when L is the loss function and θ is the true value of the parameter. Then, an equivalent definition of the deficiency is:

$$\delta(\mathscr{P}_1, \mathscr{P}_2) = \inf_{\pi_1} \sup_{\pi_2} \sup_{\theta\in\Theta} \sup_{L:\|L\|=1} \big|R_1(\pi_1, L, \theta) - R_2(\pi_2, L, \theta)\big|.$$

Thus $\Delta(\mathscr{P}_1, \mathscr{P}_2) < \varepsilon$ means that for every procedure π_i in problem i there is a procedure π_j in problem j, $\{i, j\} = \{1, 2\}$, with risks differing by at most ε, uniformly over all bounded L and $\theta \in \Theta$. In particular, when minimax rates of convergence in a nonparametric estimation problem are obtained in one experiment, the same rates automatically hold in any asymptotically equivalent experiment. There is more: When explicit transformations from one experiment to another are obtained, statistical procedures can be carried over from one experiment to the other one.

There are various techniques to bound the Le Cam distance. We report below only the properties that are useful for our purposes. For the proofs see, e.g., [6, 10].

Property 4.3. *Let $\mathscr{P}_j = (\mathscr{X}, \mathscr{A}, \{P_{j,\theta}; \theta \in \Theta\})$, $j = 1, 2$, be two statistical models having the same sample space and define $\Delta_0(\mathscr{P}_1, \mathscr{P}_2) := \sup_{\theta\in\Theta} \|P_{1,\theta} - P_{2,\theta}\|_{TV}$. Then, $\Delta(\mathscr{P}_1, \mathscr{P}_2) \le \Delta_0(\mathscr{P}_1, \mathscr{P}_2)$.*

In particular, Property 4.3 allows us to bound the Le Cam distance between statistical models sharing the same sample space by means of classical bounds for the total variation distance. To that aim, we collect below some useful results.

Fact 4.4. *Let P_1 and P_2 be two probability measures on $\mathscr{X}$, dominated by a common measure ξ, with densities $g_i = \frac{dP_i}{d\xi}$, $i = 1, 2$. Define*

$$L_1(P_1, P_2) = \int_{\mathscr{X}} |g_1(x) - g_2(x)| \xi(dx),$$

$$H(P_1, P_2) = \left(\int_{\mathscr{X}} \left(\sqrt{g_1(x)} - \sqrt{g_2(x)} \right)^2 \xi(dx) \right)^{1/2}.$$

Then,

$$\frac{H^2(P_1, P_2)}{2} \leq \|P_1 - P_2\|_{TV} = \frac{1}{2} L_1(P_1, P_2) \leq H(P_1, P_2).$$

Fact 4.5. *Let P and Q be two product measures defined on the same sample space: $P = \otimes_{i=1}^n P_i$, $Q = \otimes_{i=1}^n Q_i$. Then*

$$H^2(P, Q) \leq \sum_{i=1}^n H^2(P_i, Q_i).$$

Fact 4.6. *Let $Q_1 \sim \mathcal{N}(\mu_1, \sigma_1^2)$ and $Q_2 \sim \mathcal{N}(\mu_2, \sigma_2^2)$. Then*

$$\|Q_1 - Q_2\|_{TV} \leq \sqrt{2\left(1 - \frac{\sigma_1^2}{\sigma_2^2}\right)^2 + \frac{(\mu_1 - \mu_2)^2}{2\sigma_2^2}}.$$

Fact 4.7. *For $i = 1, 2$, let Q_i, $i = 1, 2$, be the law on $(C, \mathscr{C})$ of two Gaussian processes of the form*

$$X_t^i = \int_0^t h_i(s) ds + \int_0^t \sigma(s) dW_s, \ t \in I$$

where $h_i \in L_2(\mathbb{R})$ and $\sigma \in \mathbb{R}_{>0}$. Then:

$$L_1(Q_1, Q_2) \leq \sqrt{\int_I \frac{\left(h_1(s) - h_2(s)\right)^2}{\sigma^2(s)} ds}.$$

Property 4.8. *Let $\mathscr{P}_i = (\mathscr{X}_i, \mathscr{A}_i, \{P_{i,\theta}, \theta \in \Theta\})$, $i = 1, 2$, be two statistical models. Let $S : \mathscr{X}_1 \to \mathscr{X}_2$ be a sufficient statistics such that the distribution of S under $P_{1,\theta}$ is equal to $P_{2,\theta}$. Then $\Delta(\mathscr{P}_1, \mathscr{P}_2) = 0$.*

Finally, we recall the following result that allows us to bound the Le Cam distance between multinomial and Gaussian variables. According with the notation used throughout the paper, $\mathscr{M}(n, \theta)$ stands for a multinomial distribution of parameters (n, θ).

Theorem 4.9. *(See [2], Theorem 1 and Sections 7.1, 7.2) Let $\mathscr{P} = \{P_\theta : \theta \in \Theta_R\}$, where $P_\theta = \mathscr{M}(n, \theta)$ and $\Theta_R \subset \mathbb{R}^m$ consists of all vectors of probabilities such that*

$$\frac{\max \theta_i}{\min \theta_i} \leq R.$$

Let $\mathscr{Q} = \{Q_\theta : \theta \in \Theta_R\}$ where Q_θ is the multivariate normal distribution with vector mean $(\sqrt{n\theta_1}, \ldots, \sqrt{n\theta_m})$ and diagonal covariance matrix $\frac{1}{4} I_m$. Then

$$\Delta(\mathscr{P}, \mathscr{Q}) \leq C_R \frac{m \ln m}{\sqrt{n}}$$

for a constant C_R that depends only on R.

Acknowledgements. I would like to thank my Ph.D supervisor, Sana Louhichi, for several fruitful discussions. I am also very grateful to Markus Reiss for some very insightful exchanges from which the main idea behind this paper emerged.

References

[1] L. D. Brown, A. V. Carter, M. G. Low, and C.-H. Zhang. "Equivalence theory for density estimation, Poisson processes and Gaussian white noise with drift". In: *Ann. Statist.* 32.5 (2004), pp. 2074–2097.

[2] A. V. Carter. "Deficiency distance between multinomial and multivariate normal experiments". In: *Ann. Statist.* 30.3 (2002). Dedicated to the memory of Lucien Le Cam, pp. 708–730.

[3] J. Diebolt, L. Gardes, S. Girard, and A. Guillou. "Bias-reduced estimators of the Weibull tail-coefficient". In: *TEST* 17.2 (2008), pp. 311–331.

[4] L. Gardes and S. Girard. "Comparison of Weibull tail-coefficient estimators". In: *REVSTAT* 4.2 (2006), pp. 163–188.

[5] M. Jähnisch and M. Nussbaum. "Asymptotic equivalence for a model of independent non identically distributed observations". In: *Statist. Decisions* 21.3 (2003), pp. 197–218.

[6] L. Le Cam. *Asymptotic methods in statistical decision theory.* Springer Series in Statistics. Springer-Verlag, New York, 1986, pp. xxvi+742.

[7] L. Le Cam and G. L. Yang. *Asymptotics in statistics.* Second. Springer Series in Statistics. Some basic concepts. Springer-Verlag, New York, 2000, pp. xiv+285.

[8] E. Mariucci. " Asymptotic equivalence for pure jump Lévy processes with unknown Lévy density and Gaussian white noise". In: *Stochastic Process. Appl.* (To appear).

[9] M. Nussbaum. "Asymptotic equivalence of density estimation and Gaussian white noise". In: *Ann. Statist.* 24.6 (1996), pp. 2399–2430.

10] H. Strasser. *Mathematical theory of statistics.* Vol. 7. de Gruyter Studies in Mathematics. Berlin: Walter de Gruyter & Co., 1985, pp. xii+492.

Laboratoire LJK, Université Joseph Fourier UMR 5224, 51, Rue des Mathématiques, Campus le Saint Martin d'Hères, BP 53 38041 Grenoble Cedex 09

E-mail address: Ester.Mariucci@imag.fr