Graduate J. Math. **1** (2016), 79 – 91

Le Cam theory on the comparison of statistical models

Ester Mariucci

Abstract

We recall the main concepts of the Le Cam theory of statistical experiments, especially the notion of Le Cam distance and its properties. We also review classical tools for bounding such a distance before presenting some examples. A proof of the classical equivalence result between density estimation problems and Gaussian white noise models will be analyzed.

Key Words: Statistical experiments, Le Cam distance, deficiency, density estimation model.
MSC 2010. Primary 62B15; Secondary 62G20, 62G07.

1 Introduction

The theory of *Mathematical Statistics* is based on the notion of *statistical model*, also called *statistical experiment* or just *experiment*. A statistical model, as in its original formulation due to Blackwell (1951), is a triple

$$\mathscr{P} = (\Omega, \mathscr{T}, (P_\theta : \theta \in \Theta)),$$

where $(\Omega, \mathscr{T})$ is a sample space, Θ is a set called the *parameter space* and $(P_\theta : \theta \in \Theta)$ is a family of probability measures on $(\Omega, \mathscr{T})$. This definition is a mathematical abstraction intended to represent a concrete experiment; consider for example the following situation taken from the book of Le Cam and Yang (2000). A physicist decides to estimate the half life of Carbon 14, C^{14}. He supposes that the life of a C^{14} atom has an exponential distribution with parameter θ and, in order to develop his investigation, he takes a sample of n atoms of C^{14}. The physicist fixes in advance the duration of the experiment, say 2 hours, and then he counts the number of disintegrations. Formally, this leads to the definition of the statistical model $\mathscr{P}_1 = (\mathbb{N}, \mathcal{P}(\mathbb{N}), (P_\theta : \theta \in (0, \infty)))$ where P_θ represents the law of the random variable X counting the number of disintegrations observed in 2 hours. This is not the only way to proceed if we want to estimate the half life of Carbon 14. Indeed, the physicist could choose to consider the first random time Y after which a fixed number of disintegrations, say 10^6, have occurred. In this case he will represent the experiment via the statistical model $\mathscr{P}_2 = (\mathbb{R}_+, \mathscr{B}(\mathbb{R}_+), (Q_\theta : \theta \in (0, \infty)))$ where Q_θ is the law of the random variable Y. A natural question is then how much "statistical information" the considered experiments contain or, more precisely, when the experiment $\mathscr{P}_1$ will be more informative than $\mathscr{P}_2$ and conversely.

The quest for comparison of statistical experiments was initiated by the paper of Bohnenblust, Shapley and Sherman (1949) followed by the papers of Blackwell (1951, 1953) where the following definition was introduced: "$\mathscr{P}_1$ is more informative than $\mathscr{P}_2$" if for any bounded loss function L, $\|L\|_\infty \leq 1$, and any decision procedure ρ_2 in the experiment $\mathscr{P}_2$ there exists a decision procedure ρ_1 in the experiment $\mathscr{P}_1$ such that

$$R_\theta(\mathscr{P}_1, \rho_1, L) \leq R_\theta(\mathscr{P}_2, \rho_2, L), \quad \forall \theta \in \Theta.$$

Here we denote by $R_\theta(\mathscr{P}_1, \rho_1, L)$ and $R_\theta(\mathscr{P}_2, \rho_2, L)$ the *statistical risk* for the experiments $\mathscr{P}_1$ and $\mathscr{P}_2$, respectively.

However, this can lead to two models being non-comparable. This issue was solved by Le Cam who introduced the notion of deficiency $\delta(\mathscr{P}_1, \mathscr{P}_2)$. We will give a precise definition in the forthcoming sections. Here, we only remark two interesting properties:

- $\delta(\mathscr{P}_1, \mathscr{P}_2)$ is a well defined non-negative real number for every two given statistical models $\mathscr{P}_1$ and $\mathscr{P}_2$ sharing the same parameter space.
- For every loss function L with $0 \leq L \leq 1$ and every decision procedure ρ_2 available on Θ using $\mathscr{P}_2$, there exists a decision procedure ρ_1 in $\mathscr{P}_1$ such that for all $\theta \in \Theta$,

$$R_\theta(\mathscr{P}_1, \rho_1, L) \leq R_\theta(\mathscr{P}_2, \rho_2, L) + \delta(\mathscr{P}_1, \mathscr{P}_2).$$

This solves the issue mentioned above: It could be that both $\delta(\mathscr{P}_1, \mathscr{P}_2)$ and $\delta(\mathscr{P}_2, \mathscr{P}_1)$ are strictly positive, in which case they will not be comparable according to the first definition; nevertheless, we can still say "how much

information" we lose when passing from one model to the other one. Le Cam's theory has found applications in several problem in statistical decision theory and it has been developed, for example, for nonparametric regression, nonparametric density estimation problems, generalized linear models, diffusion models, Lévy models, spectral density estimation problem. Historically, the first results of asymptotic equivalence in a nonparametric context date from 1996 and are due to Brown and Low (1996) and Nussbaum (1996). The first two authors have shown the asymptotic equivalence of nonparametric regression and a Gaussian white noise model while the third one those of density estimation problems and Gaussian white noise models. Over the years many generalizations of these results have been proposed such as Brown et al. (2002), Carter (2006, 2007, 2009), Grama and Nussbaum (2002), Meister and Reiß (2013), Reiß (2008), Rohde (2004) and Schmidt-Hieber (2014) for nonparametric regression or Brown et al. (2004), Carter (2002), Jähnisch and Nussbaum (2003) and Mariucci (To appear) for nonparametric density estimation models. Another very active field of study is that of diffusion experiments. The first result of equivalence between diffusion models and Euler scheme was established in 1998, see Milstein and Nussbaum (1998). In later papers generalizations of this result have been considered (see Genon-Catalot and Laredo (2014) and Mariucci (2016b)) as well as different statistical problems always linked with diffusion processes (see, e.g., Dalalyan and Reiß (2006, 2007), Delattre and Hoffmann (2002) and Genon-Catalot, Laredo and Nussbaum (2002)). Among others we can also cite equivalence results for generalized linear models (see, e.g., Grama and Nussbaum (1998)), time series (see, e.g., Grama and Neumann (2006) and Milstein and Nussbaum (1998)), GARCH model (see, e.g., Buchmann and Müller (2012)), functional linear regression (see, e.g.,Meister (2011)), spectral density estimation (see, e.g. Golubev, Nussbaum and Zhou (2010)), volatility estimation (see, e.g. Reiß (2011)) and jump models (see, e.g., Mariucci (2015, 2016a)). Negative results are somewhat harder to come by; the most notable among them are Brown and Zhang (1998), Efromovich and Samarov (1996) and Wang (2002). Another new research direction that has been explored involves quantum statistical experiments (see, e.g., Buscemi (2012)).

The aim of this survey paper is to present some basic concepts of the Le Cam theory of asymptotic equivalences between statistical models. Our aim in this review is to give an accessible introduction to the subject. Therefore, we will not follow the most general approach to the theory, also because such an approach is already available in the literature, see e.g., Le Cam (1986), Le Cam and Yang (2000) and van der Vaart (2002). In order to achieve such a goal, the paper has been organized as follows. In Section 2 we recall the definition of the Le Cam distance and its statistical meaning. Particular attention has been payed to the interpretation of the Le Cam distance in terms of decision theory. In Section 3 we collect some classical tools to control the Le Cam distance before passing to some examples described in Section 4. Section 5 is devoted to show in details a proof of a classical result in Le Cam theory, namely the asymptotic equivalence between density estimation problems and Gaussian white noise models.

2 Deficiency and Le Cam distance

As we have already pointed out, a possible way to compare two given statistical models (having the same parameter space) could be to compare the corresponding risk functions or to ask "how much information" we lose when passing from one model to the other one, saying that there is no loss if we have at our disposal a mechanism able to convert the observations from the distribution $P_{1,\theta}$ to observations from $P_{2,\theta}$. If we adopt the latter point of view a natural formalization for such a mechanism is the notion of Markov kernel.

Definition 2.1. Let $(\mathscr{X}_i, \mathscr{T}_i)$, $i = 1, 2$, be two measurable spaces. A *Markov kernel* K with source $(\mathscr{X}_1, \mathscr{T}_1)$ and target $(\mathscr{X}_2, \mathscr{T}_2)$ is a map $K : \mathscr{X}_1 \times \mathscr{T}_2 \to [0, 1]$ with the following properties:

- The map $x \mapsto K(x, A)$ is $\mathscr{T}_1$-measurable for every $A \in \mathscr{T}_2$.
- The map $A \mapsto K(x, A)$ is a probability measure on $(\mathscr{X}_2, \mathscr{T}_2)$ for every $x \in \mathscr{X}_1$.

We will denote by $K : (\mathscr{X}_1, \mathscr{T}_1) \to (\mathscr{X}_2, \mathscr{T}_2)$ a Markov kernel with source $(\mathscr{X}_1, \mathscr{T}_1)$ and target $(\mathscr{X}_2, \mathscr{T}_2)$.

Starting from a Markov kernel $K : (\mathscr{X}_1, \mathscr{T}_1) \to (\mathscr{X}_2, \mathscr{T}_2)$ and a probability measure P_1 on $(\mathscr{X}_1, \mathscr{T}_1)$ one can construct a probability measure on $(\mathscr{X}_2, \mathscr{T}_2)$ in the following way:

$$KP_1(A) = \int K(x, A)P_1(dx), \quad \forall A \in \mathscr{T}_2.$$

Roughly speaking we can think that two models $\mathscr{P}_1$ and $\mathscr{P}_2$ contain "the same amount of information about θ" if there exist two Markov kernels, K_1 and K_2, not depending on θ, such that $K_1P_{1,\theta} = P_{2,\theta}$ and $K_2P_{2,\theta} = P_{1,\theta}$. This idea has been formalized in the sixties by Lucien Le Cam and led to the notion of the deficiency, hence to the introduction of a pseudo-metric on the class of all statistical experiments having the same parameter space.

The definition of the deficiency in its most general form involves the notion of "transition" which is a generalization of the concept of Markov kernel. In this paper, however, we prefer to keep things simpler and only focus on the case in which one has to deal with dominated statistical models having Polish sample spaces (see below for a definition). The advantage is that in this case the definition of deficiency simplifies and the abstract concept of transition coincides with that of Markov kernel (see Proposition 9.2 in Nussbaum (1996)).

Definition 2.2. A statistical model

$$\mathscr{P}_1 = (\mathscr{X}_1, \mathscr{T}_1, (P_{1,\theta} : \theta \in \Theta))$$

is called *Polish* if its sample space $(\mathscr{X}_1, \mathscr{T}_1)$ is a separable completely metrizable topological space. $\mathscr{P}_1$ is said to be *dominated* if there exists a σ-finite measure μ on $(\mathscr{X}_1, \mathscr{T}_1)$ such that, for all $\theta \in \Theta$, $P_{1,\theta}$ is absolutely continuous with respect to μ. The measure μ is called the *dominating measure*.

Example 2.3. Typical examples of Polish spaces in probability theory are the spaces $\mathbb{R}, \mathbb{R}^n, \mathbb{R}^\infty$, the space C_T of continuous functions on $[0, T]$ equipped with the supremum norm $d(x, y) = \sup_{0\leq t\leq T} |x_t - y_t|$, the space D of càdlàg functions equipped with the Skorokhod metric.

Definition 2.4. Let Q_1 and Q_2 be two probability measures defined on a measurable space Ω. The *total variation distance* between Q_1 and Q_2 is defined as the quantity:

$$\|Q_1 - Q_2\|_{TV} = \sup_{A\subseteq\Omega} |Q_1(A) - Q_2(A)| = \frac{1}{2} L_1(Q_1, Q_2),$$

where $L_1(Q_1, Q_2)$ denotes the L_1 norm between Q_1 and Q_2.

Definition 2.5. Let $\mathscr{P}_i = (\mathscr{X}_i, \mathscr{T}_i, (P_{i,\theta} : \theta \in \Theta))$, $i = 1, 2$, be two experiments. The *deficiency* $\delta(\mathscr{P}_1, \mathscr{P}_2)$ of $\mathscr{P}_1$ with respect to $\mathscr{P}_2$ is the number

$$\delta(\mathscr{P}_1, \mathscr{P}_2) = \inf_T \sup_{\theta\in\Theta} \|TP_{1,\theta} - P_{2,\theta}\|_{TV},$$

for an infimum taken over all Markov kernels $T : (\mathscr{X}_1, \mathscr{T}_1) \to (\mathscr{X}_2, \mathscr{T}_2)$ and $\|\cdot\|_{TV}$ denotes the total variation distance.

Definition 2.6. The *Le Cam distance* or Δ*-distance* between $\mathscr{P}_1$ and $\mathscr{P}_2$ is defined as

$$\Delta(\mathscr{P}_1, \mathscr{P}_2) = \max(\delta(\mathscr{P}_1, \mathscr{P}_2), \delta(\mathscr{P}_2, \mathscr{P}_1)).$$

The Le Cam distance is a pseudo-metric on the space of all statistical models: It satisfies the triangle inequality $\Delta(\mathscr{P}_1, \mathscr{P}_3) \leq \Delta(\mathscr{P}_1, \mathscr{P}_2) + \Delta(\mathscr{P}_2, \mathscr{P}_3)$ but the equality $\Delta(\mathscr{P}_1, \mathscr{P}_2) = 0$ does not imply that $\mathscr{P}_1$ and $\mathscr{P}_2$ actually coincide.

Concerning the glossary, when $\delta(\mathscr{P}_1, \mathscr{P}_2) = 0$ (i.e. if the experiment $\mathscr{P}_2$ can be reconstructed from the experiment $\mathscr{P}_1$ by a Markov kernel), we will say that $\mathscr{P}_2$ is *less informative* than $\mathscr{P}_1$, or that $\mathscr{P}_1$ is *better* than $\mathscr{P}_2$, or that $\mathscr{P}_1$ is *more informative* than $\mathscr{P}_2$. When $\Delta(\mathscr{P}_1, \mathscr{P}_2) = 0$ the models $\mathscr{P}_1$ and $\mathscr{P}_2$ are said to be *equivalent* and 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* when $\Delta(\mathscr{P}_{1,n}, \mathscr{P}_{2,n}) \to 0$ as $n \to \infty$.

A way to interpret the Le Cam distance between experiments is to see it as a numerical indicator of the cost needed to reconstruct one model from the other one and vice-versa, via Markov kernels. But, as we said in the introduction, a way to compare statistical models that seems just as natural is to compare the respective risk functions. Let us then highlight how the definition of the deficiency has a clear interpretation in terms of statistical decision theory. To that aim, we will start by recalling the standard framework:

- A *statistical model*, which is just an indexed set $\mathscr{P} = (\mathscr{X}, \mathscr{T}, (P_\theta : \theta \in \Theta))$ of probability measures all defined on the same measurable space $(\mathscr{X}, \mathscr{T})$, for some set $\mathscr{X}$ equipped with a σ-field $\mathscr{T}$. The elements of Θ are sometimes called the *states of Nature*.
- A space A of possible actions or decisions that the statistician can take after observing $x \in \mathscr{X}$. For example, in estimation problems we can take $A = \Theta$. To make sense of the notion of integral on A we need it to be equipped with a σ-field $\mathcal{A}$.
- A loss function $L : \Theta \times A \mapsto (-\infty, \infty]$, with the interpretation that action $z \in A$ incurs a loss $L(\theta, z)$ when θ is the true state of Nature.
- A *(randomized) decision rule* ρ in $\mathscr{P}$ is a Markov kernel $\rho : (\mathscr{X}, \mathscr{T}) \to (A, \mathscr{A})$.
- The *risk* is:

$$R_\theta(\mathscr{P}, \rho, L) = \int_{\mathscr{X}} \left(\int_A L(z, \theta)\rho(y, dz) \right) P_\theta(dy).$$

More precisely, the standard interpretation of risk is as follows. The statistician observes a value $x \in \mathscr{X}$ obtained from a probability measure P_θ. He does not know the value of θ and must take a decision $z \in A$. He does so by choosing a probability measure $\rho(x, \cdot)$ on A and picking a point in A at random according to $\rho(x, \cdot)$. If he has chosen z when the true distribution of x is P_θ, he suffers a loss $L(\theta, z)$. His average loss when x is observed is then $\int L(\theta, z)\rho(x, dz)$. His all over average loss when x is picked according to P_θ is the integral $\int \left(\int L(\theta, z)\rho(x, dz) \right) P_\theta(dx)$.

A very important result allowing to translate the notion of deficiency as described above in a decision theory language is the following:

Theorem 2.7 (See Le Cam (1964) or Theorem 2, page 20 in Le Cam (1986))**.** *Let $\varepsilon > 0$ be fixed. $\delta(\mathscr{P}_1, \mathscr{P}_2) < \varepsilon$ if and only if: $\forall$ decision rule ρ_2 on $\mathscr{P}_2$ and for all bounded loss function L, $\|L\|_\infty \leq 1$, there exists a decision rule ρ_1 on $\mathscr{P}_1$ such that*

$$R_\theta(\mathscr{P}_1, \rho_1, L) < R_\theta(\mathscr{P}_2, \rho_2, L) + \varepsilon, \quad \forall\theta \in \Theta.$$

In other words we have that $\delta(\mathscr{P}_1, \mathscr{P}_2)$ is equal to

$$\inf_{\rho_1} \sup_{\rho_2} \sup_{\theta} \sup_{L} |R(\mathscr{P}_1, \rho_1, L, \theta) - R(\mathscr{P}_2, \rho_2, L, \theta)|,$$

where the last supremum is taken on the set of all loss functions L s.t. $0 \leq L(\theta, z) \leq 1$, $\forall z \in A$, $\forall\theta \in \Theta$ and ρ_i belongs to the set of all randomised decision procedures in the experiment $\mathscr{P}_i$, $i = 1, 2$.

Remark 2.8. An important consequence of the previous theorem is that if two sequences of experiments $(\mathscr{P}_{1,n})_{n\in\mathbb{N}}$ and $(\mathscr{P}_{2,n})_{n\in\mathbb{N}}$ are asymptotically equivalent in Le Cam's sense then asymptotic properties of any inference problem are the same for these experiments. This means that when two sequences of statistical experiments are proven to be asymptotically equivalent it is enough to choose the simplest one, to study there the inference problems one is interested in and to transfer the knowledge about such inference problems to the more complicated sequence, via Markov kernels.

2.1 How to transfer decision rules via randomisations

Let $\mathscr{P}_{i,n} = (\mathscr{X}_{i,n}, \mathscr{T}_{i,n}, (P_{i,n,\theta} : \theta \in \Theta))$, $i = 1, 2$, be two sequences of statistical models sharing the same parameter space Θ and having Polish sample spaces $(\mathscr{X}_{i,n}, \mathscr{T}_{i,n})$. Suppose that there exist Markov kernels K_n such that $\|K_n P_{1,n,\theta} - P_{2,n,\theta}\|_{TV} \to 0$ uniformly on the parameter space. Then, given a decision rule (or an estimator) $\pi_{2,n}$ on $\mathscr{P}_{2,n}$ we can define a decision rule $\pi_{1,n}$ on $\mathscr{P}_{1,n}$ that, asymptotically, has the same statistical risk as $\pi_{2,n}$. To show that let us start by considering the easier case in which both K_n and $\pi_{2,n}$ are deterministic. More precisely, we suppose that K_n is of the form $K_n(A) = \mathbb{I}_A S_n(x)$ for all $A \in \mathscr{T}_{2,n}$ for some functions S_n.

Then, we have (suppressing the index n to shorten notations):

$$\begin{aligned}
&\left| \int_{\mathscr{X}_1} L(\theta, \pi_1(y)) P_{1,\theta}(dy) - \int_{\mathscr{X}_2} L(\theta, \pi_2(y)) P_{2,\theta}(dy) \right| \\
&\leq \left| \int_{\mathscr{X}_1} L(\theta, \pi_1(y)) P_{1,\theta}(dy) - \int_{\mathscr{X}_2} L(\theta, \pi_2(y)) K P_{1,\theta}(dy) \right| \\
&\quad + \left| \int_{\mathscr{X}_2} L(\theta, \pi_2(y)) \Big[K P_{1,\theta}(dy) - P_{2,\theta}(dy) \Big] \right| \\
&\leq \left| \int_{\mathscr{X}_1} L(\theta, \pi_1(y)) P_{1,\theta}(dy) - \int_{\mathscr{X}_1} L(\theta, \pi_2(S(y))) P_{1,\theta}(dy) \right| \\
&\quad + \|L\|_\infty \|K P_1 - P_2\|_{TV}.
\end{aligned}$$

In particular, assuming that the loss function L is bounded by 1 and defining

$$\pi_1(y) = \pi_2(S(y))$$

one finds that

$$\left| \int_{\mathscr{X}_1} L(\theta, \pi_1(y)) P_{1,\theta}(dy) - \int_{\mathscr{X}_2} L(\theta, \pi_2(y)) P_{2,\theta}(dy) \right| \leq \|K P_1 - P_2\|_{TV} \to 0,$$

that is, the decision rule $\pi_{1,n}(y) = \pi_{2,n}(S_n(y))$ has asymptotically the same risk as $\pi_{2,n}$. The same kind of computations work in the general case in which the K_n's are not deterministic and $(\pi_{2,n})$ is a sequence of decision rule having $(A_n, \mathscr{A}_n)$ as action's spaces. In this case one can show that the randomized sequence of decision rules

$$\pi_{1,n}(y, C) = \int_{\mathscr{X}_{2,n}} \pi_{2,n}(x, C) K(y, dx), \quad \forall y \in \mathscr{X}_{1,n},\ \forall C \in \mathscr{A}_n$$

has asymptotically the same risk as $\pi_{2,n}$.

Remark 2.9. Let P_i be a probability measure on $(E_i, \mathcal{E}_i)$ and K_i a Markov kernel on $(G_i, \mathcal{G}_i)$. One can then define a Markov kernel K on $(\prod_{i=1}^n E_i, \otimes_{i=1}^n \mathcal{G}_i)$ in the following way:

$$K(x_1, \ldots, x_n; A_1 \times \cdots \times A_n) = \prod_{i=1}^n K_i(x_i, A_i),$$

for all x_i in E_i and for all A_i in $\mathcal{G}_i$. Clearly $K \otimes_{i=1}^n P_i = \otimes_{i=1}^n K_i P_i$.

3 How to control the Le Cam distance

Even if the definition of deficiency has a perfectly reasonable statistical meaning, it is not easy to compute: Explicit computations have appeared but they are rare (see Hansen and Torgersen (1974) and Torgersen (1972, 1974) and Section 1.9 in Shiryaev and Spokoiny (2000)). More generally, one may hope to find more easily some upper bounds for the Δ-distance. We collect below some useful techniques for this purpose.

Property 3.1. *Let $\mathscr{P}_j = (\mathscr{X}, \mathscr{T}, (P_{j,\theta}; \theta \in \Theta))$, $j = 1, 2$, 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, $\Delta(\mathscr{P}_1, \mathscr{P}_2) \leq \Delta_0(\mathscr{P}_1, \mathscr{P}_2)$.*

In particular, Property 3.1 allows us to bound the Δ-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 (and classical) results.

Fact 3.2 (see Le Cam 1969, p. 35)**.** *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,

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

An important property is the following:

Property 3.3. *If μ and ν are product measures defined on the same measurable space, $\mu = \bigotimes_{j=1}^m \mu_j$ and $\nu = \bigotimes_{j=1}^m \nu_j$, then*

$$H^2(\mu,\nu) = 2\left[1 - \prod_{j=1}^m \left[1 - \frac{H^2(\mu_j,\nu_j)}{2}\right]\right].$$

Proof. See, e.g., Zolotarev (1983), p. 279. □

Thus one can express the distance between distributions of vectors with independent components in terms of the component-wise distances. A consequence of Property 3.3 is:

Property 3.4. *If μ and ν are product measures defined on the same measurable space, $\mu = \bigotimes_{j=1}^m \mu_j$ and $\nu = \bigotimes_{j=1}^m \nu_j$, then*

$$H^2(\mu,\nu) \leq \sum_{i=1}^m H^2(\mu_i,\nu_i).$$

Proof. See, e.g., Strasser (1985), Lemma 2.19. □

Property 3.5. *The Hellinger distance between two normal distributions $\mu \sim \mathcal{N}(m_1,\sigma_1^2)$ and $\nu \sim \mathcal{N}(m_2,\sigma_2^2)$ is:*

$$\begin{aligned} H^2(\mu,\nu) &= 2\left[1 - \left[\frac{2\sigma_1\sigma_2}{\sigma_1^2+\sigma_2^2}\right]^{1/2} \exp\left[-\frac{(m_1-m_2)^2}{4(\sigma_1^2+\sigma_2^2)}\right]\right] \\ &\leq 2\left|1 - \frac{\sigma_1^2}{\sigma_2^2}\right| + \frac{(m_1-m_2)^2}{2\sigma_2^2}. \end{aligned}$$

Proof. See, e.g., Mariucci (2015), Fact 1.5. □

3.1 The likelihood process

Another way to control the Le Cam distance lies in the deep relation linking the equivalence between experiments to the proximity of the distributions of the related likelihood ratios.

Let $\mathscr{P}_j = (\mathscr{X}_j, \mathscr{T}_j, (P_{j,\theta} : \theta \in \Theta))$ be a statistical model dominated by P_{j,θ_0}, $\theta_0 \in \Theta$, and let $\Lambda_j(\theta) = \frac{dP_{j,\theta}}{dP_{j,\theta_0}}$ be the density of $P_{j,\theta}$ with respect to P_{j,θ_0}. In particular, one can see $\Lambda_j(\theta)$ as a real random variable defined on the probability space $(\mathscr{X}_j, \mathscr{T}_j)$, i.e. one can see $(\Lambda_j(\theta))_{\theta\in\Theta}$ as a stochastic process. For that reason we introduce the notation $\Lambda_{\mathscr{P}_j} := (\Lambda_j(\theta), \theta \in \Theta)$ and we call $\Lambda_{\mathscr{P}_j}$ the *likelihood process*.

A key result of the theory of Le Cam is the following.

Property 3.6. *Let $\mathscr{P}_j = (\mathscr{X}_j, \mathscr{T}_j, (P_{j,\theta} : \theta \in \Theta))$, $j = 1, 2$, be two experiments. If the family $(P_{j,\theta} : \theta \in \Theta)$ is dominated by P_{j,θ_0}, then $\mathscr{P}_1$ and $\mathscr{P}_2$ are equivalent if and only if their likelihood processes under the dominating measures P_{1,θ_0} and P_{2,θ_0} coincide.*

Proof. see Strasser (1985), Corollary 25.9. □

Suppose now that there are two processes $(\Lambda_j^{n,*}(\theta))_{\theta\in\Theta}$, $j = 1, 2$ defined on a same probability space $(\mathscr{X}^*, \mathscr{T}^*, \Pi^*)$ and such that the law of $(\Lambda_j^n(\theta))_{\theta\in\Theta}$ under P_{j,θ_0} is equal to the law of $(\Lambda_j^{n,*}(\theta))_{\theta\in\Theta}$ under Π^*, $j = 1, 2$. Then, the following holds (see Le Cam and Yang (2000), Lemma 6).

Property 3.7. *If $\Lambda_{\mathscr{P}_1}$ and $\Lambda_{\mathscr{P}_2}$ are the likelihood processes associated with the experiments $\mathscr{P}_1$ and $\mathscr{P}_2$, then*

$$\Delta(\mathscr{P}_1^n, \mathscr{P}_2^n) \leq \sup_{\theta\in\Theta} \mathbb{E}_{\Pi^*}\left|\Lambda_1^{n,*}(\theta) - \Lambda_2^{n,*}(\theta)\right|.$$

3.2 Sufficiency and Le Cam distance

A very useful tool, when comparing statistical models having different sample spaces, is to look for a sufficient statistic. The introduction of the term *sufficient statistic* is usually attributed to R.A. Fisher who gave several definitions of the concept. We cite here the presentation of the subject from Le Cam (1964). Fisher's most relevant statement seems to be the requirement "...that the statistic chosen should summarize the whole of the relevant information supplied by the sample." Such a requirement may be made precise in various ways, the following three interpretations are the most common.

(i) *The classical, or operational definition of sufficiency*, claims that a statistic S is sufficient if, given the value of S, one can proceed to a post-experimental randomization reproducing variables which have the same distributions as the originally observable variables.

(ii) *The Bayesian interpretation.* A statistic S is sufficient if for every a priori distribution of the parameter the a posteriori distributions of the parameter given S is the same as if the entire result of the experiment was given.

(iii) *The decision theoretical concept.* A statistic S is sufficient if for every decision problem and every decision procedure made available by the experiment there is a decision procedure, depending on S only, which has the same performance characteristics.

The study of sufficiency in an abstract way can be found in Halmos and Savage (1949). The last section of such a work is named "The value of sufficient statistics in statistical methodology" and starts with the following observation:

> *We gather from conversations with some able and prominent mathematical statisticians that there is doubt and disagreement about just what a sufficient statistic is sufficient to do, and in particular about in what sense if any it contains "all the information in a sample".*

In Bahadur (1954) a continuation of the work of Halmos and Savage (1949) can be found. A particular effort was done to highlight the interest of using sufficient

statistics in statistical methodology. One of the main results in Bahadur (1954) is Theorem 7.1 establishing the equivalence of the decision theoretical concept of sufficiency and the operational concept in terms of conditional probabilities. We mention this fact because of its similarity with the result of Le Cam, here stated as Theorem 2.7, that is the core of the theory of comparison of statistical experiments.

Formally, let $\mathscr{P} = (\mathscr{X}, \mathscr{T}, (P_\theta : \theta \in \Theta))$ be a statistical model. A *statistic* is a measurable map from a measurable space $(\mathscr{X}, \mathscr{T})$ to another measurable one $(\mathscr{X}_2, \mathscr{T}_2)$. We denote by $S_\# P_\theta$ the image law of S under P_θ, i.e $S_\# P_\theta(B) = P_\theta(S^{-1}(B))$, for all $B \in \mathscr{T}_2$.

Definition 3.8. S is a *sufficient statistic* for $(P_\theta : \theta \in \Theta)$ if for any $A \in \mathscr{T}$ there exists a function ϕ_A, with $\phi_A \circ S$ $\mathscr{T}$-measurable, such that

$$P_\theta(A \cap S^{-1}(B)) = \int_B \phi_A(y) S_\# P_\theta(dy),$$

for all A in $\mathscr{T}$, B in $\mathscr{T}_2$ and θ in Θ. An arbitrary subalgebra $\mathscr{T}_0$ of $\mathscr{T}$ is said to be *sufficient* for $(P_\theta : \theta \in \Theta)$ if for all $A \in \mathscr{T}$ there exists a $\mathscr{T}_0$-measurable function ϕ_A such that

$$P_\theta(A \cap A_0) = \int_{A_0} \phi_A(x) P_\theta(dx), \quad \forall A_0 \in \mathscr{T}_0, \ \forall \theta \in \Theta.$$

The set $\{S^{-1}(B) : B \in \mathscr{T}_2\}$ is called the *subalgebra induced by the statistic* S.

Property 3.9 (See, e.g. Bahadur (1954))**.** *A statistic S is sufficient for $(P_\theta : \theta \in \Theta)$ if the subalgebra induced by S is sufficient for $(P_\theta : \theta \in \Theta)$.*

In accordance with the notation introduced in Section 2, we will state Theorem 7.1 in Bahadur (1954) as follows (recall that $(A, \mathscr{A})$ denotes the *action/decision space.*)

Theorem 3.10 (See Theorem 7.1, Bahadur (1954))**.** *If the subalgebra $\mathscr{T}_0$ of $\mathscr{T}$ is sufficient for $(P_\theta : \theta \in \Theta)$, then for every decision rule $\rho : (\mathscr{X}, \mathscr{T}) \mapsto (A, \mathscr{A})$ there exists a decision rule $\pi : (\mathscr{X}, \mathscr{T}_0) \mapsto (A, \mathscr{A})$ such that*

$$P_\theta \rho(C) = P_\theta \pi(C), \quad \forall C \in \mathscr{A}, \ \forall \theta \in \Theta.$$

Before focusing on the relation between the notion of sufficient statistic and the one of equivalence between statistical models, let us recall the Neyman-Fisher factorization theorem, a powerful tool for identifying sufficient statistics for a given dominated family of probabilities. Let $(P_\theta : \theta \in \Theta)$ be a family of probabilities on $(\Omega, \mathscr{T})$, absolutely continuous with respect to a σ-finite measure μ, and denote by $p_\theta := \frac{dP_\theta}{d\mu}$ the density.

Theorem 3.11. *A statistic $S : (\Omega, \mathscr{T}) \to (\mathscr{X}, \mathscr{B})$ is sufficient for $(P_\theta : \theta \in \Theta)$ if and only if there exists a $\mathscr{B}$-measurable function g_θ $\forall \theta \in \Theta$ and a $\mathscr{T}$-measurable function $h \neq 0$ such that*

$$p_\theta(x) = g_\theta(S(x)) h(x), \quad \mu\text{-a.s. } \forall x \in \Omega.$$

An important result linking the Le Cam distance with the existence of a sufficient statistic is the following:

Property 3.12. *Let $\mathscr{P}_i = (\mathscr{X}_i, \mathscr{T}_i, (P_{i,\theta} : \theta \in \Theta))$, $i = 1, 2$, be two statistical models. Let $S : \mathscr{X}_1 \to \mathscr{X}_2$ be a sufficient statistic such that the distribution of S under $P_{1,\theta}$ is equal to $P_{2,\theta}$. Then $\Delta(\mathscr{P}_1, \mathscr{P}_2) = 0$.*

Proof. In order to prove that $\delta(\mathscr{P}_1, \mathscr{P}_2) = 0$ it is enough to consider the Markov kernel $M : (\mathscr{X}_1, \mathscr{T}_1) \to (\mathscr{X}_2, \mathscr{T}_2)$ defined as $M(x, B) := \mathbb{I}_B(S(x))$ $\forall x \in \mathscr{X}_1$ and $\forall B \in \mathscr{T}_2$. Conversely, to show that $\delta(\mathscr{P}_2, \mathscr{P}_1) = 0$ one can consider the Markov kernel $K : (\mathscr{X}_2, \mathscr{T}_2) \to (\mathscr{X}_1, \mathscr{T}_1)$ defined as $K(y, A) = \mathbb{E}_{P_{2,\theta}}(\mathbb{I}_A | S = y)$, $\forall y \in \mathscr{X}_2$, $\forall A \in \mathscr{T}_1$. Since S is a sufficient statistic, the Markov kernel K does not depend on θ. Denoting by $S_\# P_{1,\theta}$ the distribution of S under $P_{1,\theta}$, one has:

$$\begin{aligned} K P_{2,\theta}(A) &= \int K(y, A) P_{2,\theta}(dy) \\ &= \int \mathbb{E}_{P_{2,\theta}}(\mathbb{I}_A | S = y) S_\# P_{1,\theta}(dy) = P_{1,\theta}(A). \end{aligned}$$

□

For asymptotic arguments, one also needs an appropriate version of the notion of sufficiency.

Definition 3.13. Let $\mathscr{P}_n = (\mathscr{X}_n, \mathscr{T}_n, (P_{n,\theta} : \theta \in \Theta))$ be a sequence of statistical models. The sequence of subalgebras $\tilde{\mathscr{T}}_n$ of $\mathscr{T}_n$ is *asymptotically sufficient* for $(P_{n,\theta} : \theta \in \Theta)$ if $\Delta(\mathscr{P}_n, \mathscr{P}_{n|_{\tilde{\mathscr{T}}_n}}) \to 0$, where $\mathscr{P}_{n|_{\tilde{\mathscr{T}}_n}}$ denotes the restriction of the experiment $\mathscr{P}_n$ to $\tilde{\mathscr{T}}_n$, i.e. $\mathscr{P}_{n|_{\tilde{\mathscr{T}}_n}} = (\mathscr{X}_n, \tilde{\mathscr{T}}_n, (\tilde{P}_{n,\theta} : \theta \in \Theta))$, where $\tilde{P}_{n,\theta}(A) = P_{n,\theta}(A)$, for all $A \in \tilde{\mathscr{T}}_n$.

This is a stronger notion than asymptotic equivalence; indeed, let $\mathscr{P}_{1,n}$ and $\mathscr{P}_{2,n}$ be two sequences of experiments having the same parameter space. Then, by the triangle inequality, it is clear that if there exist two sequences $S_{1,n}$ and $S_{2,n}$ of asymptotically sufficient statistics in $\mathscr{P}_{1,n}$ and $\mathscr{P}_{2,n}$ respectively, taking values in the same measurable space, and such that

$$\sup_{\theta \in \Theta} \| S_{1,n\#} P_{1,\theta} - S_{2,n\#} P_{2,\theta} \|_{TV} \to 0 \quad \text{as } n \to \infty,$$

then the sequences $\mathscr{P}_{1,n}$ and $\mathscr{P}_{2,n}$ are asymptotically equivalent. We also recall that an important generalization of the notion of sufficiency is the notion of insufficiency. The discussion of this concept is beyond the purposes of this paper, the reader is referred to Le Cam (1974) or Chapter 5 in Le Cam (1986) for an exhaustive treatment of the subject.

4 Examples

To better understand what is the typical form of an asymptotic equivalence result let us analyze some examples. As a toy example let us start by considering the following parametric case.

Example 4.1. Let $\mathscr{P}_{1,n}$ be the statistical model associated with the observation of a vector X of n independent Gaussian random variables $\mathscr{N}(\theta, 1)$. Here the inference concerns θ and the parameter space Θ will be an interval of $\mathbb{R}$. Formally

$$\mathscr{P}_{1,n} = (\mathbb{R}^n, \mathscr{B}(\mathbb{R}^n), (P_{1,\theta} : \theta \in \Theta)),$$

where P_θ is the law of X.

Then, let us denote by $\mathscr{P}_{2,n}$ the experiment associated with the observation of the empirical mean relative to the previous random variables, i.e.

$$\mathscr{P}_{2,n} = (\mathbb{R}, \mathscr{B}(\mathbb{R}), (P_{2,\theta} : \theta \in \Theta)),$$

where $P_{2,\theta}$ is the law of a Gaussian random variable of mean θ and variance $1/n$. By means of the Neyman-Fisher factorization theorem it is easy to see that the application $S : \mathbb{R}^n \to \mathbb{R}$, $S(x_1, \ldots, x_n) = \frac{\sum_{i=1}^n x_i}{n}$ is a sufficient statistic. An immediate application of Property 3.12 implies that $\Delta(\mathscr{P}_{1,n}, \mathscr{P}_{2,n}) = 0$ for all n.

Before passing to some examples in a nonparametric framework, let us recall a result due to Carter and concerning the asymptotic equivalence between a multinomial and a Gaussian multivariate experiment. The parameter space will be a subset of $\mathbb{R}^m$ and the reason for which we focus on such a result lies on its being a very useful tool in establishing global asymptotic equivalence results for density estimation problems.

Example 4.2. Let $X = (X_1, \ldots, X_m)$ be a random vector having a multinomial distribution of parameters n and $(p_1, \ldots, p_m)$ with $p_i \geq 0$ for $i = 1, \ldots, m$ and $\sum_{i=1}^m p_i = 1$.

Denote by $\mathscr{P}$ the statistical model associated with a multinomial distribution $P_\theta = \mathcal{M}(n; (\theta_1, \ldots, \theta_m))$ with parameters $\theta = (\theta_1, \ldots, \theta_m)$ that belong to $\Theta_R \subset \mathbb{R}^m$, a set consisting of all vectors of probabilities such that

$$\frac{\max_i \theta_i}{\min_i \theta_i} \leq R.$$

The main result in Carter (2002) is a bound of the Le Cam distance between statistical models associated with multinomial distributions and multivariate normal distributions with the same means and covariances as the multinomial ones. More precisely, let us denote by $\mathscr{Q}$ the statistical model associated with a family of multivariate normal distributions $Q_\theta = \mathscr{N}(\mu, \Sigma)$, $\theta \in \Theta_R$, where

$$\mu = (n\theta_1, \ldots, n\theta_m), \quad \Sigma = (\sigma_{i,j})_{i,j=1,\ldots,m}$$

with $\sigma_{i,j} = n\theta_i(1 - \theta_i)\delta_{i=j} - n\theta_i\theta_j\delta_{i\neq j}$.

Theorem 4.3 (see Carter (2002), p. 709)**.** *With the notations above,*

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

for a constant C_R that depends only on R.

Another interesting result contained in Carter (2002) is the approximation of $\mathscr{Q}$ by a Gaussian experiment with independent coordinates. Let us denote by $\tilde{\mathscr{Q}}$ the statistical model associated with m independent Gaussian random variables $\mathscr{N}(\sqrt{\theta_i}, 1/(4n))$, $i = 1, \ldots, m$.

Theorem 4.4 (see Carter (2002), p. 717–719)**.** *With the notations above,*

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

for a constant C_R that depends only on R.

Let us now consider some examples in a nonparametric framework. More precisely, we will recall the results of Brown and Low (1996) and Nussbaum (1996) that are the first asymptotic equivalence results for nonparametric experiments.

Example 4.5. In Brown and Low (1996), the authors consider the problem of estimating the function f from a continuously observed Gaussian process $y(t)$, $t \in [0, 1]$, which satisfies the SDE

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

where dW_t is a Gaussian white noise. They find that the statistical model associated with the continuous observation of (y_t) is asymptotically equivalent to the statistical model associated with its discrete counterpart, i.e. the nonparametric regression:

$$y_i = f(t_i) + \sigma(t_i)\xi_i, \quad i = 1, \ldots, n.$$

The time grid is uniform, $t_i = \frac{i-1}{n}$, and the ξ_i's are standard normal variables; they assume that f varies in a nonparametric subset $\mathscr{F}$ of $L_2[0, 1]$ defined by some smoothness conditions and n tends to infinity not too slowly. More precisely, the drift function $f(\cdot)$ is unknown and such that, for B a positive constant, one has:

$$\sup\left\{|f(t)| : t \in [0, 1], f \in \mathscr{F}\right\} = B < \infty.$$

Moreover, defining

$$\bar{f}_n(t) = \begin{cases} f\left(\frac{i}{n}\right) & \text{if} \quad \frac{i-1}{n} \leq t < \frac{i}{n}, \quad i = 1, \ldots, n; \\ f(1) & \text{if} \quad t = 1; \end{cases}$$

one asks:

$$\lim_{n\to\infty} \sup_{f\in\mathcal{F}} n \int_0^1 \frac{(f(t) - \bar{f}_n(t))^2}{\sigma^2(t)} dt = 0.$$

The diffusion coefficient $\sigma^2(\cdot) > 0$ is supposed to be a known absolutely continuous function on $[0, 1]$ such that

$$\left|\frac{d}{dt} \ln \sigma(t)\right| \leq C, \quad t \in [0, 1],$$

for some positive constant C.

Example 4.6. In Nussbaum (1996) the author 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 $f(\cdot)$ on $[0,1]$ with respect to the Lebesgue measure. The densities $f(\cdot)$ 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 Nussbaum (1996) 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.

In the last years, asymptotic equivalence results have also been established for discretely observed stochastic processes. As an example, let us present the result in Mariucci (2015), very close in spirit to the one of Brown and Low (1996).

Example 4.7. Let $\{X_t\}_{t\geq 0}$ be a sequence of time inhomogeneous jump-diffusion processes defined by

$$X_t = \eta + \int_0^t f(s)ds + \varepsilon_n \int_0^t \sigma(s)dW_s + \sum_{i=1}^{N_t} Y_i, \quad t \in [0,T],$$

where:

- η is some random initial condition;
- $W = \{W_t\}_{t\geq 0}$ is a standard Brownian motion;
- $N = \{N_t\}_{t\geq 0}$ is an inhomogeneous Poisson process with intensity function $\lambda(\cdot)$, independent of W;
- $(Y_i)_{i\geq 1}$ is a sequence of i.i.d. real random variables with distribution G, independent of W and N;
- $\sigma^2(\cdot)$ is supposed to be known. The horizon of observation T is finite and $\varepsilon_n \to 0$ as $n \to \infty$.
- $f(\cdot)$ belongs to some non-parametric class $\mathscr{F}$.
- $\lambda(\cdot)$ and G are also unknown and belong to nonparametric classes Λ and $\mathscr{G}$, respectively.

In Mariucci (2015), the problem of estimating f from high frequency observations of $\{X_t\}_{t\geq 0}$ is considered. More precisely, we suppose to observe $\{X_t\}_{t\geq 0}$ at discrete times $0 = t_0 < t_1 < \cdots < t_n = T$ such that $\Delta_n = \max_{1\leq i\leq n}\Big\{|t_i - t_{i-1}|\Big\} \downarrow 0$ as n goes to infinity. Let $\mathscr{P}_n$ be the statistical model associated with the continuous observation of $\{X_t\}_{t\in[0,T]}$ and $\mathscr{Q}_n$ the one associated with the observations $(X_{t_i})_{i=0}^n$. Finally, let $\mathscr{W}_n$ be the Gaussian white noise model associated with the continuous observation of the Gaussian process

$$dy_t = f(t)dt + \varepsilon_n \sigma(t)dW_t, \quad y_0 = \eta, \quad t \in [0,T].$$

Suppose that $\mathscr{F}$ is a subclass of α-Hölder, uniformly bounded functions on $\mathbb{R}$ and the nuisance parameters $\sigma(\cdot)$ and $\lambda(\cdot)$ satisfy the following conditions:

- There exist two constants m and M such that $0 < m \leq \sigma(\cdot) \leq M < \infty$ and $\sigma(\cdot)$ is derivable with derivative $\sigma'(\cdot)$ in $L_\infty(\mathbb{R})$.
- There exists a constant $L < \infty$ such that for all $\lambda \in \Lambda$, $\|\lambda\|_{L_2([0,T])} < L$.

Then, under the assumption that $\Delta_n^{2\alpha}\varepsilon_n^{-2} \to 0$ as $n \to \infty$, the three models $\mathscr{P}_n$, $\mathscr{Q}_n$ and $\mathscr{W}_n$ are asymptotically equivalent. A bound for $\Delta(\mathscr{Q}_n, \mathscr{W}_n)$ and $\Delta(\mathscr{P}_n, \mathscr{Q}_n)$ is given by

$$\Delta_n^{\beta/2} + T\Delta_n^{2\alpha}\varepsilon_n^{-2} + T\Delta_n,$$

where $\beta = 1$ if $\mathscr{G}$ is a subclass of discrete distributions with support on $\mathbb{Z}$ and $\beta = 1/2$ if $\mathscr{G}$ is a subclass of absolutely continuous distributions with respect to the Lebesgue measure on $\mathbb{R}$ with uniformly bounded densities on a fixed neighborhood of 0. In particular, this result tells us that the jumps of the process $\{X_t\}_{t\geq 0}$ can be ignored when the goal is the estimation of the drift function $f(\cdot)$. Moreover, the proof is constructive: an explicit Markov kernel is constructed to filter the jumps out.

5 Density estimation problems and Gaussian white noise models: A constructive proof

In this Section, following Carter (2002) (see p. 720-725), we will detail how one can prove, in a constructive way, the asymptotic equivalence between a density estimation problem and a Gaussian white noise model, as presented in Example 4.6. However, with respect to the work of Nussbaum (1996), we will ask some stronger hypotheses on the parameter space in order to simplify the proofs. More precisely, for fixed $\gamma \in (0,1]$ and K, ε, M strictly positive constants, we will consider a functional parameter space of the form

$$\mathscr{F}_{(\gamma,K,\varepsilon,M)} = \Big\{ f \in C^1(I) : \varepsilon \leq f(x) \leq M, \\ |f'(x) - f'(y)| \leq K|x-y|^\gamma, \ \forall x,y \in [0,1] \Big\}.$$

As in Example 4.6, $\mathscr{P}_{1,n}$ will be a density estimation problem:

$$(Y_i)_{1\leq i\leq n} \text{ i.i.d. r.v. with density } f \in \mathscr{F}_{(\gamma,K,\varepsilon,M)} \quad (5.1)$$

and $\mathscr{P}_{2,n}$ a Gaussian with noise model:

$$dy_t = \sqrt{f(t)}dt + \frac{1}{2\sqrt{n}}dW_t, \quad t \in [0,1], \quad f \in \mathscr{F}_{(\gamma,K,\varepsilon,M)}.$$

The idea of Carter was to use the bound on the distance between multinomial and Gaussian normal variables as presented in Example 4.2 to make assertions about density estimation experiments. The intuition is to see the multinomial experiment as the result of grouping independent observations from a continuous density into m subsets, say J_i, $i = 1, \dots, m$. Using the square root as a variance-stabilizing transformation, these multinomial variables can be asymptotically approximated by m normal variables with constant variances. These normal variables, in turn, are approximations to the increments of the process (y_t) over the sets J_i. In Subsection 5.1 we will analyze how to obtain a asymptotically equivalent multinomial experiment starting from $\mathscr{P}_{1,n}$. Assuming the results of Carter stated here as Theorems 4.3 and 4.4 we will then obtain a bound of the Δ-distance between such a multinomial experiment and one associated with independent Gaussian random variables. In Subsection 5.2 we will explain how to show the asymptotic equivalence between an adequate normal approximation with independent coordinates and $\mathscr{P}_{2,n}$.

5.1 Density estimation problems and multinomial experiments

Let us consider a partition of $[0,1]$ in m intervals $J_i = [(i-1)/m, i/m]$ and denote by $S : [0,1]^n \to \{1, \dots, n\}^m$ the application mapping the n-tuple $(x_1, \dots, x_n)$ to the m-tuple $(\#\{j : x_j \in J_1\}, \dots, \#\{j : x_j \in J_m\})$, where the writing $\#\{j : x_j \in J_i\}$ stands for the number of x_j belonging to the interval J_i. Let $P_f^{\otimes n}$ be the law of $(Y_1, \dots, Y_n)$ as in (5.1). The law of S under $P_f^{\otimes n}$ is a multinomial distribution $\mathscr{M}(n; \theta_1, \dots, \theta_m)$, $\theta_i = \int_{J_i} f(x)dx$, $i = 1, \dots, m$. In particular this means that an appropriate multinomial experiment is more informative than $\mathscr{P}_{1,n}$. More precisely, we have proven that the statistical model associated with the multinomial distribution $(\mathscr{M}(n; \theta_1, \dots, \theta_m) : f \in \mathscr{F})$, denoted by $\mathscr{P}_m$, is such that $\delta(\mathscr{P}_{1,n}, \mathscr{P}_m) = 0$.

Let us now investigate the quantity $\delta(\mathscr{P}_m, \mathscr{P}_{1,n})$. A trivial observation is that the total variation distance between the multinomial distribution $\mathscr{M}(n; \theta_1, \dots, \theta_m)$ and the law $P_f^{\otimes n}$ is always 1, hence, in order to prove that $\delta(\mathscr{P}_m, \mathscr{P}_{1,n}) \to 0$ we need to construct a non trivial Markov kernel. We will divide the proof in three main steps.

Step 1: We denote by x_i^* the midpoints of the intervals J_i, i.e. $x_i^* = \frac{2i-1}{2m}$, and we introduce a discrete random variable X^* concentrated at the points x_i^* with masses θ_i. Let us then denote by $\mathscr{P}^*$ the statistical model associated with the observation of n independent realizations of X^*. Then, by means of a "sufficient statistic" argument we can get $\Delta(\mathscr{P}_m, \mathscr{P}^*) = 0$. Indeed, consider the application

$$S : \{x_1^*, \dots, x_m^*\}^n \to \{1, \dots, n\}^m$$

mapping $(y_1, \dots, y_n)$ to

$$(\#\{j : y_j = x_1^*\}, \dots, \#\{j : y_j = x_m^*\})$$

and observe that the density h of the law of n independent realizations of X^* with respect to the counting measure is given by

$$\begin{aligned} h(y_1, \dots, y_n) &= \prod_{i=1}^n \mathbb{P}(X^* = y_i) \\ &= \theta_1^{\#\{j : y_j = x_1^*\}} \times \cdots \times \theta_m^{\#\{j : y_j = x_m^*\}}. \end{aligned}$$

By means of the Neyman-Fisher factorization theorem, we conclude that S is a sufficient statistic, thus $\Delta(\mathscr{P}_m, \mathscr{P}^*) = 0$.

Step 2: Starting from n realizations of X^* we want to obtain something close to n independent realizations of P_f, the law of Y_1 as in (5.1). To that aim we define an approximation of f as follows:

$$\hat{f}_m(x) = \sum_{j=1}^m V_j(x)\theta_j,$$

where V_j's are piecewise linear functions interpolating the values in the points x_j^* as in Figure 1.

In particular, $\hat{f}_m$ is a piecewise linear function that can be written as

$$\hat{f}_m(x) = \begin{cases} m\theta_1 \mathbb{I}_{[0, x_1^*]}(x), & \text{if } i = 1, \\ \Big(m - m^2|x - x_i^*|\Big)\mathbb{I}_{[x_i^*, x_{i+1}^*]}(x) & \text{if } 1 < i < m, \\ m\theta_m \mathbb{I}_{[x_m^*, 1]}(x), & \text{if } i = m. \end{cases}$$

We then consider the Markov kernel

$$M(k, A) = \sum_{j=1}^m \mathbb{I}_{\{x_j^*\}}(k) \int_A V_j(y)dy,$$

for all k in $\mathbb{N}$ and A in $\mathscr{B}([0,1])$. Denoting by P^* the law of the random variable X^*, we have

$$\begin{aligned} MP^*(A) &= \sum_{k \in \mathbb{N}} M(k, A)\mathbb{P}(X^* = k) = \sum_{i=1}^m \theta_i M(x_i^*, A) \\ &= \sum_{i=1}^m \theta_i \int_A V_i(y)dy = \int_A \hat{f}_m(y)dy. \end{aligned}$$

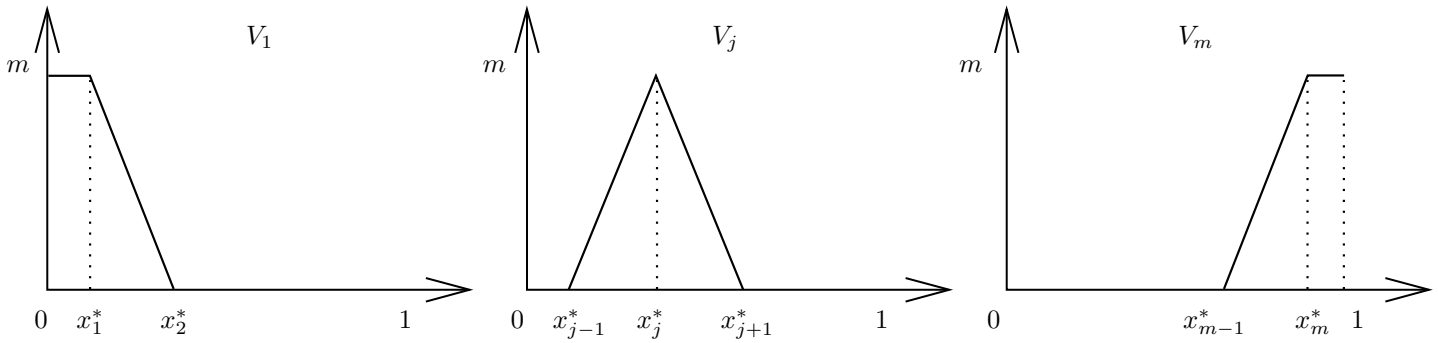
Let $\hat{\mathscr{P}}_m$ be the statistical model associated with the observation of n i.i.d. random variables $(\hat{Y}_i)_{1 \le i \le n}$ having $\hat{f}_m$ as a density with respect to the Lebesgue measure on $[0,1]$. Applying Remark 2.9 we get $\delta(\mathscr{P}^*, \hat{\mathscr{P}}_m) = 0$.

Step 3: We are only left to check that $\delta(\hat{\mathscr{P}}_m, \mathscr{P}_{1,n}) \to 0$. This is actually the case and we can show that

$$\Delta(\mathscr{P}_{1,n}, \hat{\mathscr{P}}_m) = O\Big(\sqrt{n}\Big(m^{-3/2} + m^{-1-\gamma}\Big)\Big).$$

Indeed, the total variation distance between the family of probabilities associated with the experiments $\mathscr{P}_{1,n}$

Fig. 1: The definition of the V_j functions.



and $\hat{\mathscr{P}}_m$ is bounded by $\sqrt{n}H(f,\hat{f}_m)$. Since $f(x) \geq \varepsilon$ for all $x \in [0,1]$ one can write:

$$H^2(f,\hat{f}_m) = \int_0^1 \left(\frac{f(x)-\hat{f}_m(x)}{\sqrt{f(x)}+\sqrt{\hat{f}_m(x)}}\right)^2 dx \leq \frac{1}{4\varepsilon}\int_0^1 \Big(f(x)-\hat{f}_m(x)\Big)^2 dx.$$

In order to control the L_2 distance between f and $\hat{f}_m$ we will split $\int_0^1 \Big(f(x)-\hat{f}_m(x)\Big)^2 dx$ as follows:

$$\int_0^1 \Big(f(x)-\hat{f}_m(x)\Big)^2 dx = \int_0^{1/2m} \Big(f(x)-m\theta_1\Big)^2 dx + \int_{1/2m}^{1-1/2m} \Big(f(x)-\hat{f}_m(x)\Big)^2 dx + \int_{1-1/2m}^1 \Big(f(x)-m\theta_m\Big)^2 dx.$$

An application of the mean theorem allows us to deduce $\int_{J_i}\Big(f(x)-m\theta_i\Big)^2 dx = O(m^{-3})$, $i=1,\ldots,m$. To control the term $\int_{1/2m}^{1-1/2m}\Big(f(x)-\hat{f}_m(x)\Big)^2 dx$, let us consider the Taylor expansion of f at points x_i^*, where x denotes a point in J_i , $i=2,\ldots,m-1$:

$$f(x) = f(x_i^*) + f'(x_i^*)(x-x_i^*) + R_i(x). \qquad (5.2)$$

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

$$|R_i(x)| = \Big|f(x)-f(x_i^*)-f'(x_i^*)(x-x_i^*)\Big| = \Big|f'(\xi_i)-f'(x_i^*)\Big||\xi_i-x_i^*| \leq Km^{-1-\gamma}, \quad (5.3)$$

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 $J_{i-1}\cup J_i$), indeed:

$$\begin{aligned}\hat{f}'_m(x_i^*) &= -m(\hat{f}_m(x_i^*)-\hat{f}_m(x_{i+1}^*)) \\ &= -m^2\left(\int_{\frac{i-1}{m}}^{\frac{i}{m}} f(s)ds - \int_{\frac{i}{m}}^{\frac{i+1}{m}} f(s)ds\right) \\ &= m^2\int_{\frac{i-1}{m}}^{\frac{i}{m}} \Big(f(s+1/m)-f(s)\Big)ds = m\int_{J_i} f'(\xi_s)ds\end{aligned}$$

for some $\xi_s \in [s, s+1/m]$. Applying the mean theorem to the function $g(s) = f'(\xi_s)$ we get that $\int_{J_i} f'(\xi_s)ds = \frac{1}{m}f'(t)$ for some $t \in J_i \cup J_{i+1}$. The fact that $\hat{f}'_m(x_i^*) = f'(t)$, 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:

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

Using (5.2) and the fact that $\int_{J_i}(x-x_i^*)dx = 0$, we get:

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

Collecting all the pieces together we find

$$\int_0^1 \Big(f(x)-\hat{f}_m(x)\Big)^2 dx = O\Big(m^{-3}+m^{-2\gamma-2}\Big),$$

hence we can conclude that

$$\Delta(\mathscr{P}_{1,n},\hat{\mathscr{P}}_m) = O\Big(\sqrt{n}(m^{-3/2}+m^{-1-\gamma})\Big).$$

5.2 Independent Gaussian random variables and Gaussian white noise experiments

In Subsection 5.1 we have seen how to reduce a density estimation problem to an adequate multinomial experiment. An application of the results of Carter (2002)

recalled in Example 4.2 allows us to obtain an asymptotic equivalence between the statistical model associated with the observation of n i.i.d. random variables of density $f : [0,1] \to \mathbb{R}$ with respect to the Lebesgue measure and an experiment in which one observes $m = m_n$ Gaussian and independent random variables $\mathscr{N}(\sqrt{\theta_i}, 1/4n)$, $i = 1, \dots, m$. Of course, such a Gaussian experiment is equivalent to $\mathscr{N}_m$, the statistical model associated with independent Gaussian random variables $\mathscr{N}\Big(\sqrt{\frac{\theta_i}{m}}, \frac{1}{4nm}\Big)$, $i = 1, \dots, m$. We claim that $\mathscr{N}_m$ is asymptotically equivalent to the white noise model $\mathscr{P}_{2,n}$ associated with the continuous observation of a trajectory of a Gaussian process $(y_t)_{t\in[0,1]}$ solution of the SDE:

$$dy_t = \sqrt{f(t)}dt + \frac{1}{2\sqrt{n}}dW_t, \quad t \in [0,1], \tag{5.4}$$

where $(W_t)_t$ is a standard Brownian motion. We will divide the proof in two steps. Denote by $\mathscr{N}_m^*$ the statistical model associated with the observation of $(y_t)_t$ over the intervals J_i, $i = 1, \dots, m$, i.e. $\mathscr{N}_m^*$ is the experiment associated with m independent Gaussian random variables $\mathscr{N}\Big(\int_{J_i}\sqrt{f(s)}ds, \frac{1}{4nm}\Big)$, $i = 1, \dots, m$. Firstly, we will show that $\mathscr{N}_m$ is asymptotically equivalent to $\mathscr{N}_m^*$, then that observing $(y_t)_t$ is asymptotically equivalent to observing its increments.

Step 1: By means of Property 3.5 we get

$$\begin{aligned}\Delta(\mathscr{N}_m, \mathscr{N}_m^*) &\le \sqrt{2mn}\sqrt{\sum_{i=1}^m \left(\int_{J_i}\sqrt{f(t)}dt - \sqrt{\frac{\theta_i}{m}}\right)^2} \\ &= \sqrt{2n}\sqrt{\sum_{i=1}^m \frac{1}{m}\left(m\int_{J_i}\Big(\sqrt{f(t)} - \sqrt{m\theta_i}\Big)dt\right)^2}\end{aligned}$$

Denote by $E_i = |m\int_{J_i}\Big(\sqrt{f(t)} - \sqrt{m\theta_i}\Big)dt|$. By the triangular inequality, we bound E_i by $F_i + G_i$ where:

$$F_i = \left|\sqrt{m\theta_i} - \sqrt{f(x_i^*)}\right| \text{ and } G_i = \left|\sqrt{f(x_i^*)} - m\int_{J_i}\sqrt{f(y)}dy\right|.$$

Using the same trick as in Step 3 of Subsection 5.1, we can bound:

$$\begin{aligned}F_i &= \frac{|m\theta_i - f(x_i^*)|}{\sqrt{m\theta_i} + \sqrt{f(x_i^*)}} \le \frac{|m\theta_i - f(x_i^*)|}{2\sqrt{\varepsilon}} \\ &= \frac{m}{2\sqrt{\varepsilon}}\left|\int_{J_i}\Big(f(s) - f(x_i^*)\Big)ds\right| \le \frac{1}{2\sqrt{\varepsilon}}\|R_i\|_\infty,\end{aligned}$$

where we have used the fact that $\int_{J_i}(x - x_i^*) = 0$ and R_i denotes the remainder in the Taylor expansion of f in x_i^*, as in (5.2). On the other hand,

$$\begin{aligned}G_i &= m\left|\int_{J_j}\Big(\sqrt{f(x_i^*)} - \sqrt{f(y)}\Big)dy\right| \\ &= m\left|\int_{J_i}\left(\frac{f'(x_i^*)}{2\sqrt{f(x_i^*)}}(x - x_i^*) + \tilde{R}_i(y)\right)dy\right| \le \|\tilde{R}_i\|_\infty,\end{aligned}$$

where $\tilde{R}_i$ is the remainder in the Taylor expansion of $\sqrt{f}$ in x_i^*. We observe that if f belongs to the functional class $\mathscr{F}_{(\gamma,K,\varepsilon,M)}$ then $\sqrt{f}$ is still bounded away from zero and from infinity with a Hölder continuous derivative, more precisely $\sqrt{f} \in \mathscr{F}_{(\gamma,K/\sqrt{\varepsilon},\sqrt{\varepsilon},\sqrt{M})}$. In particular, we deduce that $\|\tilde{R}_i\|_\infty$ has the same magnitude as $\sqrt{\frac{M}{\varepsilon}}\|R_i\|_\infty$. Thanks to (5.3) we know that $\|R_i\|_\infty \le Km^{-1-\gamma}$ for any $i = 2, \dots, m-1$ and $\|R_i\|_\infty = O(m^{-3/2})$ for $i \in \{1, m\}$. Hence the quantities F_i and G_i are of the same order and we find that

$$\Delta(\mathscr{N}_m, \mathscr{N}_m^*) = O\Big(\sqrt{n}\Big(m^{-1-\gamma} + m^{-\frac{3}{2}}\Big)\Big).$$

Step 2: Since $\mathscr{N}_m^*$ is the model associated with the observation of the increments $(\bar{Y}_i)_{1\le i\le n}$ of the process $(y_t)_t$ defined as in (5.4) it is clear that $\delta(\mathscr{P}_{2,n}, \mathscr{N}_m^*) = 0$. Let us now discuss how to bound $\delta(\mathscr{N}_m^*, \mathscr{P}_{2,n})$. We start by introducing a new stochastic process:

$$y_t^* = \sum_{i=1}^m \bar{Y}_i\int_0^t V_i(y)dy + \frac{1}{2\sqrt{nm}}\sum_{i=1}^m B_i(t), \quad t \in [0,1],$$

where the functions V_i are defined as in Figure 1 and $B_i(t)$ are independent centered Gaussian processes independent of (W_t) and with variances

$$\mathrm{Var}(B_i(t)) = \int_0^t V_i(y)dy - \left(\int_0^t V_i(y)dy\right)^2.$$

These processes can be constructed from a standard Brownian bridge $B(t)$, independent of (W_t), via

$$B_i(t) = B\left(\int_0^t V_i(y)dy\right).$$

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

$$\begin{aligned}\mathbb{E}[y_t^*] &= \sum_{i=1}^m \mathbb{E}[\bar{Y}_i]\int_0^t V_i(y)dy \\ &= \sum_{i=1}^m\left(\int_{J_i}\sqrt{f(y)}dy\right)\int_0^t V_i(y)dy, \\ \mathrm{Var}[y_t^*] &= \sum_{i=1}^m \mathrm{Var}[\bar{Y}_i]\left(\int_0^t V_i(y)dy\right)^2 + \frac{1}{4nm}\sum_{i=1}^m \mathrm{Var}(B_i(t)) \\ &= \frac{1}{4nm}\int_0^t\sum_{i=1}^m V_i(y)dy = \frac{t}{4n}.\end{aligned}$$

One can compute in the same way the covariance of (y_t^*) and deduce that

$$Y_t^* = \int_0^t \widehat{\sqrt{f}}_m(y)dy + \int_0^t \frac{1}{2\sqrt{n}} dW_s^*, \quad t \in [0,1],$$

where (W_t^*) is a standard Brownian motion and

$$\widehat{\sqrt{f}}_m(x) := \sum_{i=1}^m \left(\int_{J_i} \sqrt{f(y)}dy \right) V_i(x).$$

Applying Fact 3.5, we get that the total variation distance between the process $(y_t^*)_{t\in[0,1]}$ constructed from the random variables $\bar{Y}_i$, $i = 1, \ldots, m$ and the Gaussian process $(y_t)_{t\in[0,1]}$ is bounded by

$$\sqrt{4n \int_0^1 \left(\widehat{\sqrt{f}}_m(y) - \sqrt{f(y)} \right)^2 dy}.$$

Since $f \in \mathscr{F}_{(\gamma,K,\varepsilon,M)}$ implies $\sqrt{f} \in \mathscr{F}_{(\gamma,K\sqrt{M}/\sqrt{\varepsilon},\sqrt{\varepsilon},\sqrt{M})}$, the same kind of computations made in Step 3 of Subsection 5.1 allows us to conclude that

$$\begin{aligned} \Delta(\mathscr{N}_m^*, \mathscr{P}_{2,n}) &= \delta(\mathscr{N}_m^*, \mathscr{P}_{2,n}) \qquad (5.5) \\ &= O\Big(\sqrt{n}\Big(m^{-3/2} + m^{-1-\gamma}\Big)\Big). \end{aligned}$$

5.3 The choice of m

In Subsection 5.1 we have proven that the cost needed to pass from the model associated to the observation of n i.i.d. random variables with unknown density $f \in \mathscr{F}_{(\gamma,K,\varepsilon,M)}$ to an adequate multinomial approximation $\mathcal{M}(n; \theta_1, \ldots, \theta_m)$ is of the order of $\sqrt{n}\Big(m^{-3/2}+m^{-1-\gamma}\Big)$. Using Theorem 4.3 we can take a further step obtaining a Gaussian approximation (with independent coordinates) starting from the multinomial one. This comes to the price of $\frac{m \ln m}{\sqrt{n}}$. Finally, in Subsection 5.2 we have found that for appropriate choices of m there is an asymptotic equivalence between such a Gaussian approximation and the Gaussian with noise model $\mathscr{P}_{2,n}$. A bound for the rate of convergence of the Δ-distance up to constants is, again, given by $\sqrt{n}\Big(m^{-3/2}+m^{-1-\gamma}\Big)$. In particular we deduce that

$$\Delta(\mathscr{P}_{1,n}, \mathscr{P}_{2,n}) = \begin{cases} O\Big(n^{-\frac{\gamma}{2(\gamma+2)}} \log n\Big), & if 0 < \gamma \leq \frac{1}{2}, \\ 0(n^{-\frac{1}{10}} \log n) & if \frac{1}{2} < \gamma \leq 1. \end{cases}$$

after the choice $m = n^{1/(2+\gamma)}$.

Acknowledgments: The research leading to these results has received funding from the European Research Council under ERC Grant Agreement 320637.

References

Bahadur, R. R. (1954). 'Sufficiency and statistical decision functions'. In: *Ann. Math. Statistics* 25, pp. 423–462.

Blackwell, D. (1951). 'Comparison of experiments'. In: *Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability, 1950*. Berkeley and Los Angeles: University of California Press, pp. 93–102.

— (1953). 'Equivalent comparisons of experiments'. In: *Ann. Math. Statistics* 24, pp. 265–272.

Bohnenblust, A., L. Shapley and S. Sherman (1949). 'Reconnaissance in game theory'. In: *Rand Research Memorandum 1949/208*.

Brown, L. D. and M. G. Low (1996). 'Asymptotic equivalence of nonparametric regression and white noise'. In: *Ann. Statist.* 24.6, pp. 2384–2398.

Brown, L. D. and C.-H. Zhang (1998). 'Asymptotic nonequivalence of nonparametric experiments when the smoothness index is 1/2'. In: *Ann. Statist.* 26.1, pp. 279–287.

Brown, L. D., T. T. Cai, M. G. Low and C.-H. Zhang (2002). 'Asymptotic equivalence theory for nonparametric regression with random design'. In: *Ann. Statist.* 30.3. Dedicated to the memory of Lucien Le Cam, pp. 688–707.

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

Buchmann, B. and G. Müller (2012). 'Limit experiments of GARCH'. In: *Bernoulli* 18.1, pp. 64–99.

Buscemi, F. (2012). 'Comparison of quantum statistical models: equivalent conditions for sufficiency'. In: *Comm. Math. Phys.* 310.3, pp. 625–647.

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

— (2006). 'A continuous Gaussian approximation to a nonparametric regression in two dimensions'. In: *Bernoulli* 12.1, pp. 143–156.

— (2007). 'Asymptotic approximation of nonparametric regression experiments with unknown variances'. In: *Ann. Statist.* 35.4, pp. 1644–1673.

— (2009). 'Asymptotically sufficient statistics in nonparametric regression experiments with correlated noise'. In: *J. Probab. Stat.* Art. ID 275308, 19.

Dalalyan, A. and M. Reiß (2006). 'Asymptotic statistical equivalence for scalar ergodic diffusions'. In: *Probab. Theory Related Fields* 134.2, pp. 248–282.

— (2007). 'Asymptotic statistical equivalence for ergodic diffusions: the multidimensional case'. In: *Probab. Theory Related Fields* 137.1-2, pp. 25–47.

Delattre, S. and M. Hoffmann (2002). 'Asymptotic equivalence for a null recurrent diffusion'. In: *Bernoulli* 8.2, pp. 139–174.

Efromovich, S. and A. Samarov (1996). 'Asymptotic equivalence of nonparametric regression and white noise model has its limits'. In: *Statist. Probab. Lett.* 28.2, pp. 143–145.

Genon-Catalot, V. and C. Laredo (2014). 'Asymptotic equivalence of nonparametric diffusion and Euler scheme experiments'. In: *The Annals of Statistics* 42.3, pp. 1145–1165.

Genon-Catalot, V., C. Laredo and M. Nussbaum (2002). 'Asymptotic equivalence of estimating a Poisson intensity and a positive diffusion drift'. In: *Ann. Statist.* 30.3. Dedicated to the memory of Lucien Le Cam, pp. 731–753.

Golubev, G. K., M. Nussbaum and H. H. Zhou (2010). 'Asymptotic equivalence of spectral density estimation and Gaussian white noise'. In: *Ann. Statist.* 38.1, pp. 181–214.

Grama, I. and M. Nussbaum (1998). 'Asymptotic equivalence for nonparametric generalized linear models'. In: *Probab. Theory Related Fields* 111.2, pp. 167–214.

— (2002). 'Asymptotic equivalence for nonparametric regression'. In: *Math. Methods Statist.* 11.1, pp. 1–36.

Grama, I. G. and M. H. Neumann (2006). 'Asymptotic equivalence of nonparametric autoregression and nonparametric regression'. In: *Ann. Statist.* 34.4, pp. 1701–1732.

Halmos, P. R. and L. J. Savage (1949). 'Application of the Radon-Nikodym theorem to the theory of sufficient statistics'. In: *Ann. Math. Statistics* 20, pp. 225–241.

Hansen, O. H. and E. N. Torgersen (1974). 'Comparison of linear normal experiments'. In: *The Annals of Statistics*, pp. 367–373.

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

Le Cam, L. (1964). 'Sufficiency and approximate sufficiency'. In: *Ann. Math. Statist.* 35, pp. 1419–1455.

Le Cam, L. (1969). *Théorie asymptotique de la décision statistique*. Séminaire de Mathématiques Supérieures, No 33 (Été, 1968). Les Presses de l'Université de Montréal, Montreal, Que., p. 140.

— (1974). 'On the information contained in additional observations'. In: *Ann. Statist.* 2, pp. 630–649.

— (1986). *Asymptotic methods in statistical decision theory*. Springer Series in Statistics. New York: Springer-Verlag, pp. xxvi+742.

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

Mariucci, E. (2015). 'Asymptotic equivalence for inhomogeneous jump diffusion processes and white noise'. In: *ESAIM Probab. Stat.* 19, pp. 560–577.

— (2016a). 'Asymptotic equivalence for pure jump Lévy processes with unknown Lévy density and Gaussian white noise'. In: *Stochastic Process. Appl.* 126.2, pp. 503–541.

— (2016b). 'Asymptotic equivalence of discretely observed diffusion processes and their Euler scheme: small variance case'. In: *Stat. Inference Stoch. Process.* 19.1, pp. 71–91.

— (To appear). 'Asymptotic equivalence for density estimation and Gaussian white noise: an extension'. In: *Annales de l'ISUP*. ArXiv:1503.05019.

Meister, A. (2011). 'Asymptotic equivalence of functional linear regression and a white noise inverse problem'. In: *Ann. Statist.* 39.3, pp. 1471–1495.

Meister, A. and M. Reiß (2013). 'Asymptotic equivalence for nonparametric regression with non-regular errors'. In: *Probab. Theory Related Fields* 155.1-2, pp. 201–229.

Milstein, G. and M. Nussbaum (1998). 'Diffusion approximation for nonparametric autoregression'. In: *Probab. Theory Related Fields* 112.4, pp. 535–543.

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

Reiß, M. (2008). 'Asymptotic equivalence for nonparametric regression with multivariate and random design'. In: *Ann. Statist.* 36.4, pp. 1957–1982.

— (2011). 'Asymptotic equivalence for inference on the volatility from noisy observations'. In: *Ann. Statist.* 39.2, pp. 772–802.

Rohde, A. (2004). 'On the asymptotic equivalence and rate of convergence of nonparametric regression and Gaussian white noise'. In: *Statist. Decisions* 22.3, pp. 235–243.

Schmidt-Hieber, J. (2014). 'Asymptotic equivalence for regression under fractional noise'. In: *The Annals of Statistics* 42.6, pp. 2557–2585.

Shiryaev, A. N. and V. G. Spokoiny (2000). *Statistical experiments and decisions*. Vol. 8. Advanced Series on Statistical Science & Applied Probability. Asymptotic theory. River Edge, NJ: World Scientific Publishing Co. Inc., pp. xvi+281.

Strasser, H. (1985). *Mathematical theory of statistics*. Vol. 7. de Gruyter Studies in Mathematics. Statistical experiments and asymptotic decision theory. Berlin: Walter de Gruyter & Co., pp. xii+492.

Torgersen, E. N. (1972). 'Comparison of translation experiments'. In: *The Annals of Mathematical Statistics* 43.5, pp. 1383–1399.

— (1974). 'Comparison of experiments by factorization'. In: *Stat.Res.Report, Univ. of Oslo*.

van der Vaart, A. (2002). 'The statistical work of Lucien Le Cam'. In: *Ann. Statist.* 30.3. Dedicated to the memory of Lucien Le Cam, pp. 631–682.

Wang, Y. (2002). 'Asymptotic nonequivalence of Garch models and diffusions'. In: *Ann. Statist.* 30.3. Dedicated to the memory of Lucien Le Cam, pp. 754–783.

Zolotarev, V. M. (1983). 'Probability metrics'. In: *Teoriya Veroyatnostei i ee Primeneniya* 28.2, pp. 264–287.

Mariucci Ester
E-mail address: e.mariucci@math.leidenuniv.nl

University of Leiden, The Netherlands