

ESTIMATION IN DISCRETE PARAMETER MODELS

CHRISTINE CHOIRAT AND RAFFAELLO SERI

CONTENTS

1. Introduction	3
2. Examples of Discrete and Mixed Parameter Models	4
3. m -estimators in Discrete Parameter Models	7
3.1. Consistency of m -estimators	7
3.2. Asymptotic Distribution of the m -estimator	8
3.3. The MLE and Bayes Estimators in Discrete Parameter Models	15
4. Optimality and Efficiency	18
4.1. Risk Functions	18
4.2. Information Inequalities	20
4.3. Optimality and Efficiency	23
5. Some Alternative Estimators	27
5.1. MLE and Other Estimators under Continuity	28
5.2. Discretized Continuous Estimators	28
5.3. Estimators Obtained by Linear Convex Combinations	29
6. Proofs	30
References	34

Abstract: In some estimation problems, especially in applications dealing with information theory, signal processing and biology, theory provides us with additional information allowing us to restrict the parameter space to a finite number of points. In this case, we speak of discrete parameter models.

Even though the problem is quite old and has interesting connections with testing and model selection, asymptotic theory for these models has hardly ever been studied. Therefore, we discuss consistency, asymptotic distribution theory, information inequalities and their relations with efficiency and superefficiency for a general class of m -estimators. Then, we extend some of the previous results to other classes of estimators.

Keywords: Discrete Parameter Space, Detection, Large Deviations, Information Inequalities, Efficiency, Superefficiency.

JEL Classification: C13, C12, C44.

AMS 2000 Subject Classification: Primary: 62F12; Secondary: 62F03, 93E10, 90C10, 62F30.

Acknowledgments: The authors would like to thank Lucien Birgé, Mehmet Caner, Jean-Pierre Florens, Christian Gouriéroux, Christian Hess, Marc Hoffmann, Pierre Jacob, Søren Johansen, Rasul A. Khan, Oliver B. Linton, Christian P. Robert, Keunkwan Ryu, Igor Vajda and the participants to seminars at Université

Paris 9 Dauphine, CREST and Institut Henri Poincaré to ESEM 2001 in Lausanne, to XXXIVèmes Journées de Statistique 2002 in Bruxelles, to BS/IMSC 2004 in Barcelona, and to ESEM 2004 in Madrid. All the remaining errors are our responsibility.

1. INTRODUCTION

It is well known that the rate of convergence of estimators depends critically on the dimension of the parameter space: in regular parametric problems, when the parameter space is a compact subset of an Euclidean space and the objective function satisfies some smoothness requirements, the mean squared error of estimators converges at a n -rate leading to the so-called \sqrt{n} -consistency; in nonparametric problems, the parameter space is infinite dimensional and the convergence rate is usually slower than \sqrt{n} (see [50] for a general discussion).

This does not exhaust the range of results presented in the literature: in very special cases, n -consistency (also called superconsistency in Econometrics) can arise like, for instance, in the estimation of the support of a density (see e.g. [35]) or in the estimation of an autoregressive coefficient under the unit root hypothesis. The case of $\sqrt[3]{n}$ -consistency can also be encountered, as in the semiparametric maximum score estimator of Manski ([87, 88]) and the other cases treated in [73].

Sometimes, especially in applications dealing with signal processing and biology, theory provides us with some additional information allowing us to restrict the parameter space to a finite number of points: in these cases, we speak of *discrete parameter models*. In this class of models, previous works have shown that the rate of convergence of m -estimators is often exponential ([52, 117, 119, 120]). There are also other cases in which the parameter space is given by the Cartesian product of a subset of a finite dimensional Euclidean space and a finite set: then, we speak of *mixed parameter models*. This case has also been dealt with in the literature (see [108, 109]), the result being that part of the parameters converges at a \sqrt{n} -rate, while the remainder converges at an exponential rate.

Statistical inference when the parameter space is reduced to a lattice was first considered by Hammersley ([52]) in a seminal paper: he derived a lower bound for the variance of a consistent estimator analogous to the Cram er-Rao and Chapman-Robbins lower bounds (see Section 4.2). However, since the author was motivated by the measurement of the mean weight of insulin, he focused mainly on the case of a Gaussian distribution with known variance and unknown integer mean (see [52], p. 192); this case was further developed by Khan ([69, 70, 71, 72]). The Poisson case also met some attention in the literature and was dealt with in [52] (p. 199), [90, 113].

Starting with Hammersley's paper, the properties of specific estimators were investigated and topics such as asymptotic distributional results and information inequalities were left aside. General treatments of admissibility and related topics are in [105, 43, 59, 91] (see also the book [14]). Special cases have been dealt with in [67] (p. 424, in the special case of a translation integral parameter and of integral data under the quadratic loss), [52, 69, 44, 70, 71, 72] (for the case of the Gaussian distribution) and [18] (for the case of the discrete uniform distribution). Other papers dealing with optimality in discrete parameter spaces are [115, 116, 118, 121, 42]. Optimality of estimation under a discrete parameter space was also considered by Vajda ([117, 119, 120]) in a nonorthodox setting inspired by R enyi's theory of random search. Bayesian encompassing has been studied in [36]. Moreover, in the estimation of complex statistical models (see [48, 28], Ch. 4) and in the calculation of efficiency rates (see [5, 78, 23]), approximating a general parameter space by a sequence of finite sets has proved to be a valuable tool.

A few papers showed the practical importance of discrete parameter models in Signal Processing, Automatic Control and Information Theory and derived some bounds on the performance of the estimators (see [76, 77, 84, 56, 57, 55, 9, 8, 10, 11]). More recently, the topic has received new interest in the Information Theory literature (see [101, 66], and the review paper [58]), and in Stochastic Integer Programming (see [38, 74, 123]).

However, no general formula for the convergence rate has ever been obtained, no optimality proof under generic conditions has been provided and no general discussion of efficiency and superefficiency in discrete parameter models has appeared in the literature. In the present paper, we provide a full answer to these problems in the case of discrete parameter models for samples of independent and identically distributed random variables: the related case of mixed parameter models and the extension to dependent and not identically distributed random variables will be considered in forthcoming works. Therefore, after introducing some examples of discrete parameter models in Section 2, in Section 3 we investigate the properties of a class of m -estimators. In particular, in Section 3.1, we derive some conditions for strong consistency; then, in Section 3.2, we calculate an approximate asymptotic distribution for the estimator and we establish its convergence rate. These results are specialized to the case of the maximum likelihood estimator (MLE) and extended to Bayes estimators in Section 3.3. In Section 4, we derive upper bounds for the convergence rate in the standard and in the minimax contexts, and we discuss the relations between information inequalities, efficiency and superefficiency. In particular, we prove that, under the zero-one loss function, no estimator is efficient in the class of consistent estimators for any value of $\theta_0 \in \Theta$ (θ_0 being here the true value of the parameter) and no estimator attains the information inequality we derive. But the MLE still has some appealing properties since it is minimax efficient and attains the minimax information inequality bound. In Section 5, we extend some of the previous results to other classes of estimators.

2. EXAMPLES OF DISCRETE AND MIXED PARAMETER MODELS

The following examples are intended to show the relevance of discrete and mixed parameter spaces in Applied and Theoretical Statistics. In particular, they show that the proposed statistical analysis of discrete parameter models solves some long-standing problems in Statistics, Optimization, Information Theory and Signal Processing. Remark that Examples 6, 7 and 8 are instances of mixed parameter models and their study will be deferred to a forthcoming companion paper.

We recall that a *statistical model* is a collection of probability measures $\mathcal{P} = \{\mathbb{P}_\theta, \theta \in \Theta\}$ where Θ is the *parameter space*. Θ is a subset of an Euclidean or of a more abstract space.

Example 1. [Tumour transplantability] We consider tumour transplantability in mice. For a certain type of mating, the probability of a tumour “taking” when transplanted from the grandparents to the offspring is equal to $\left(\frac{3}{4}\right)^\theta$ where θ is an integer equal to the number of genes determining transplantability. For another type of mating, the probability is $\left(\frac{1}{2}\right)^\theta$. We aim at estimating θ knowing that n_0 transplants take out of n . The likelihood is given by:

$$\ell_n(\theta) = \binom{n}{n_0} \cdot k^{\theta n_0} \cdot (1 - k^\theta)^{n - n_0}, \quad \theta \in \mathbb{N}, k \in \left\{\frac{1}{2}, \frac{3}{4}\right\}.$$

In this case the parameter space is discrete and the maximum likelihood estimator can be shown to be:

$$\hat{\theta}^n = \text{ni} \left[\frac{\ln \left(\frac{n_0}{n} \right)}{\ln k} \right]$$

where $\text{ni}[x]$ is the integer nearest to x (see [52], p. 236).

Example 2. [Exponential family restricted to a lattice] Consider a random variable X distributed according to an exponential family where the natural parameter θ is restricted to a lattice $\{\theta_0 + \varepsilon \cdot N, N \in \mathbb{N}^k\}$, for fixed θ_0 and ε (see [83], p. 759). The likelihood for a sample of n independent and identically distributed realizations is:

$$\ell_n(\theta) = \exp \left\{ \sum_{i=1}^n T(x_i) \cdot \theta - n \cdot \psi(\theta) \right\} \cdot \prod_{i=1}^n h(x_i).$$

The case of a Gaussian distribution has been considered in [52] (p. 192) and [69, 71], the Poisson case in [52] (p. 199), [90, 113]. In particular, [52] uses the Gaussian model to estimate the molecular weight of insulin, assumed to be an integer (however, see the remarks of Tweedie in the discussion of the same paper).

Example 3. [Stochastic discrete optimization] We consider the optimization problem of the form $\min_{x \in S} g(x)$, where $g(x) = \mathbb{E}G(x, W)$ is an integral functional, \mathbb{E} is the mean under probability \mathbb{P} , $G(x, w)$ is a real valued function of two variables x and w , W is a random variable having probability distribution \mathbb{P} and S is a finite set. We approximate this problem through the sample average function $\hat{g}_n(x) \triangleq \frac{1}{n} \sum_{i=1}^n G(x, W_i)$ and the associated problem $\min_{x \in S} \hat{g}_n(x)$. See [74] for some theoretical results and a discussion of the stochastic knapsack problem and [123] for an up-to-date bibliography.

Example 4. [Approximate inference] In many applied cases, the requirement that the true model generating the data corresponds to a point belonging to the parameter space appears to be too strong and unlikely. Moreover, the objective is often to recover a model reproducing some stylized facts from the original data. In these cases, approximation of a continuous parameter space with a finite number of points allows for obtaining such a model under weaker assumptions. This situation arises, for example, in Signal Processing and Automatic Control applications ([56, 57, 55, 9, 10, 11]) and is reminiscent of some related statistical techniques, such as the *discretization* device of Le Cam ([78], Chapitre III), or the *sieve estimation* of Grenander ([48]; see also [40], Remark 5).

Example 5. [M -ary hypotheses testing and related fields] In Information Theory, discrete parameter models are quite common, and their estimation is a generalization of binary hypothesis testing that goes under the names of *M -ary hypotheses* (or *multihypothesis*) *testing*, *classification* or *detection* (see the examples in [92]). Consider a received waveform $r(t)$ described by the equation $r(t) = m(t) + \sigma n(t)$ for $t \geq 0$, where $m(t)$ is a deterministic signal, $n(t)$ is an additive Gaussian white noise and σ is the noise intensity. The set of possible signals is restricted to a finite number of alternatives, say $\{m_0(t), \dots, m_J(t)\}$: the chosen signal is usually the one that maximizes the loglikelihood of the sample, or an alternative criterion function. For example, if the loglikelihood of the process based on the observation

window $[0, T]$ is used, we have:

$$\hat{m}_j(\cdot) = \arg \max_{j=0, \dots, J} \frac{1}{\sigma^2} \left[\int_0^T m_j(t) r(t) dt - \frac{1}{2} \int_0^T m_j^2(t) dt \right].$$

Much more complex cases can be dealt with; see [58] for an introduction.

Example 6. [Ryu's plant management model] In [108], a plant management model is developed aiming at modelling nuclear power plant data. A plant is subject to an exogenous process of failures and each failure has a repair cost: the administrator can choose to shut down the operation of the plant when the failure rate appears too high, but he incurs in a cost of preventive maintenance. The problem is to choose the policy optimally in order to minimize the sum of preventive maintenance costs and repair costs. It turns out that, under a particular information structure, the optimal strategy is to stop the plant when a certain number of failures θ from the previous preventive shut down is observed: clearly, θ is an integer parameter to be estimated from the data, together with a set of other continuous parameters.

Example 7. [Global positioning system] The GPS (Global Positioning System) is a navigation system that enables positioning anywhere on the earth at any time. It calculates the position of the user through triangulation using electromagnetic signals: as a result, the range measurements \mathbf{y} can be approximated by a linear function of the position of the user α and the ambiguous cycles of the carrier signals β . Since β is an integer multiple of the wavelengths, the parameter vector $(\alpha^T, \beta^T)^T$ is composed of a continuous and of a discrete part. A complete treatment of this problem is in [53, 54]. We remark that similar applications of mixed parameter models can be found in radar imaging, MRI and communications.

Example 8. [Model selection] In this case, several models indexed by $i = 1, \dots, I$ are considered; each model belongs to the parametric family $\mathcal{P}_i = \{\mathbb{P}_{\beta^i}, \beta^i \in B\}$. Here, we have supposed that the parameter space B is the same for any alternative model: however this hypothesis can easily be removed. We can therefore embed the estimation problem in the *nesting model*¹

$$\mathcal{P} = \{\mathbb{P}_{\theta^*}, \theta^* = (\beta, \theta) \in B \times \{1, \dots, I\} \text{ and } \beta = \beta^i \text{ if } \theta = i\},$$

with likelihood function:

$$\ell_n(\theta^*) = \prod_{i=1}^I [\ell_n(\beta^i)]^{1_{\{\theta=i\}}}.$$

Example 9. [Number of transmission sources] Suppose that θ sources transmit signals independently in time according to a Poisson process with parameter μ . In the interval $(0, t]$, $K(t)$ sources are observed; each of these sources transmits $M_i(t)$ signals in $(0, t]$. The problem is to estimate the unknown parameter θ : the parameter μ can be assumed to be known or unknown (see [114, 124, 49]). The same kind of situation arises in the fishing problem ([98]), the proofreading problem, the debugging software problem, etc.

¹[46] (Vol. 2, Sect. 22.2.7) deal with model selection through nesting models in a slightly different framework not based on discrete parameters.

Example 10. [Capture-recapture models] Consider a population of N animals and suppose to catch, mark and release a sample of n_1 animals. Then a second sample of size n_2 is taken from the population and we observe m_2 marked animals in it. It turns out that m_2 is the realization of a random variable M_2 that has an hypergeometric distribution. Therefore, the likelihood of the unknown parameter $N \in \mathbb{N}$ is:

$$\ell(N) = \frac{\binom{n_1}{m_2} \binom{N-n_1}{n_2-m_2}}{\binom{N}{n_2}}.$$

Clearly, more complex models can be considered (see [83]).

As already mentioned, Examples 6, 7 and 8 will not be treated in this paper. Also the models of Examples 9 and 10 cannot be considered along with the other ones. Indeed, Example 9 is often formulated as an optimal stopping problem. The estimator $\hat{\theta}(t)$ (depending on time t) and the observation length T are chosen in order to minimize the cost function $\mathbb{E} \left(\hat{\theta}(T) - \theta \right)^2 + cT$, given by the sum of a cost term depending on the estimation error and of a penalty term positively depending on the observation time T . Example 10 differs from the previous ones in another more peculiar sense. The asymptotics $\frac{n}{N} \rightarrow 1$ (the so-called *saturation sampling*) is clearly consistent, since in this case all the population is sampled; however, such a sampling scheme is overwhelmingly expensive. The real interest is for estimators that do well when n is small with respect to N . Therefore, it is customary to consider $N \rightarrow \infty$ and $n \rightarrow \infty$ (with a different rate such that $\frac{n}{N} \rightarrow 0$). For any well-behaved estimator, as long as $N \rightarrow \infty$, we expect $\hat{N} \rightarrow \infty$; therefore, consistency is often replaced by the α -consistency property ($\frac{\hat{N}-N}{N^\alpha} \rightarrow 0$, for $\alpha > 0$). Under additional assumptions, it is possible to show that the estimator \hat{N} is α -consistent and asymptotically normal. This shows well that the asymptotic behavior is nonstandard, as pointed out in [83] among others.

3. m -ESTIMATORS IN DISCRETE PARAMETER MODELS

In this Section, we consider an estimator obtained maximizing an objective function of the form:²

$$Q_n(\theta) = \frac{1}{n} \sum_{i=1}^n \ln q(y_i; \theta);$$

in what follows, we allow for misspecification.

3.1. Consistency of m -estimators . In the case of a discrete parameter space, uniform convergence reduces to pointwise convergence. Therefore, m -estimators are strongly consistent under less stringent conditions than in the standard case; in particular, no condition is needed on the continuity or differentiability of the objective function. The following hypothesis is used in order to prove consistency in the case of independent and identically distributed replications.

Obj The data $(Y_i)_{i=1}^n$ are realizations of independent and identically distributed $(\mathfrak{Y}, \mathcal{Y})$ -valued random variables having probability measure \mathbb{P}_0 .

²The expression m -estimator stands for *maximum likelihood type estimator*, in the spirit of Huber ([60]), and not for *maximum (or extremum) estimator*.

The estimator $\hat{\theta}^n$ is obtained maximizing over the set $\Theta = \{\theta_0, \theta_1, \dots, \theta_J\}$, of finite cardinality, the objective function:

$$Q_n(\theta) \triangleq \frac{1}{n} \sum_{i=1}^n \ln q(y_i; \theta).$$

The function q is \mathcal{Y} -measurable for each $\theta \in \Theta$ and satisfies the following L^1 -domination condition:

$$\mathbb{E}_0 |\ln q(Y; \theta)| < +\infty, \quad \forall \theta \in \Theta,$$

where \mathbb{E}_0 denotes the expectation taken under the true probability measure \mathbb{P}_0 .

Moreover, θ_0 is the point of Θ maximizing $\mathbb{E}_0 \ln q(Y; \theta)$ and θ_0 is globally identified (see [106, 125]).

Remark 1. (i) The assumption of a finite parameter space seems restrictive with respect to the more general assumption of Θ being countable (see e.g. [52]). However, **Obj** is compatible with the convex hull of Θ being compact, as in standard asymptotic theory. Indeed, the cases analysed in Hammersley have convex likelihood functions and this is a well known substitute for compactness of Θ (see [4], p. 1116, and [93], p. 2133; see [27], for consistency with neither convexity nor compactness). Moreover, the restriction to finite parameter spaces seems to be necessary to derive the asymptotic distribution of m -estimators.

(ii) The assumption of independence and identical distribution can clearly be relaxed in specific cases. In order to prove consistency, it can be substituted with an hypothesis of ergodic stationarity (or even “asymptotic mean stationarity”, see [47]): then, the SLLN of Proposition 1 could be substituted by the Shannon-McMillan-Breiman Theorem of [13, 2].

(iii) The relative position of the points of Θ is unimportant and the choice of θ_0 as the maximizer is arbitrary and is made only for practical purposes. Remark that θ_0 has no link with \mathbb{P}_0 apart from being the pseudo-true value of $\ln q$ with respect to \mathbb{P}_0 on the parameter space Θ .

Proposition 1. *Under Assumption **Obj**, the m -estimator $\hat{\theta}^n$ is a \mathbb{P}_0 -strongly consistent estimator of θ_0 and is $\mathcal{Y}^{\otimes n}$ -measurable.*

Remark 2. A similar result of consistency for discrete parameter spaces has already been provided by [112] (p. 446), by [20, 21] (pp. 325-333), by [13] (p. 1294) as an application of the Shannon-McMillan-Breiman Theorem of information theory, by [126] (Sect. 2.1) as a preliminary result of his work on partial likelihood, and by [89] (p. 96, Section 7.1.6).

3.2. Asymptotic Distribution of the m -estimator. For a discrete parameter space, the finite sample distribution of the m -estimator $\hat{\theta}^n$ is a multinomial distribution and it tends to a Dirac mass concentrated at θ_0 . Since the determination of this asymptotic distribution is an interesting and open problem, we derive in this Section upper and lower bounds and asymptotic estimates for probabilities of the form $\mathbb{P}_0(\hat{\theta}^n = \theta_i)$.

To simplify the following discussion, we introduce the processes:

$$(1) \quad \begin{cases} Q_n(\theta_j) \triangleq \frac{1}{n} \cdot \sum_{i=1}^n \ln q(y_i; \theta_j), \\ \mathbf{X}_k^{(i)} \triangleq [\ln q(Y_k; \theta_i) - \ln q(Y_k; \theta_j)]_{j=0, \dots, J, j \neq i}, \quad i = 1, \dots, J \\ \mathbf{X}_k \triangleq \mathbf{X}_k^{(0)} = [\ln q(Y_k; \theta_0) - \ln q(Y_k; \theta_j)]_{j=1, \dots, J}. \end{cases}$$

The probability of the estimator $\hat{\theta}^n$ taking on the value θ_i can be written as:

$$(2) \quad \mathbb{P}_0(\hat{\theta}^n = \theta_i) = \mathbb{P}_0(Q_n(\theta_i) > Q_n(\theta_j), \forall j \neq i) = \mathbb{P}_0\left(\sum_{k=1}^n \mathbf{X}_k^{(i)} \in \text{int } \mathbb{R}_+^J\right).$$

The Inversion Formula for the characteristic function ([86], Theorem 1.6.2, p. 23) can be used in order to have:

$$\begin{aligned} \mathbb{P}_0(\hat{\theta}^n = \theta_i) &= \mathbb{P}_0(Q_n(\theta_i) > Q_n(\theta_j), \forall j \neq i) = \mathbb{P}_0\left(\frac{1}{\sqrt{n}} \sum_{k=1}^n \mathbf{X}_k^{(i)} \in \text{int } \mathbb{R}_+^J\right) \\ &= \lim_{T \rightarrow \infty} \frac{1}{(2\pi)^J} \cdot \int_T^T \cdots \int_T^T \left[\prod_{j=1}^J \frac{1}{\iota \cdot \lambda_j} \right] \cdot \left[\varphi_{\mathbf{X}^{(i)}}\left(\frac{\lambda}{\sqrt{n}}\right) \right]^n d\lambda, \end{aligned}$$

where ι is the imaginary unit. However this method does not seem to be general enough to give widely applicable formulas.

On the other hand, since this probability can be written in the form:

$$(3) \quad \mathbb{P}_0(\hat{\theta}^n = \theta_i) = \mathbb{P}_0\left(\frac{1}{\sqrt{n}} \sum_{k=1}^n [X_{k,j}^{(i)} - \mathbb{E}_0 X_{k,j}^{(i)}] > -\sqrt{n} \mathbb{E}_0 X_{k,j}^{(i)}, \forall j \neq i\right),$$

where $X_{k,j}^{(i)}$ is the j -th element of $\mathbf{X}_k^{(i)}$, it would be tempting to use an Edgeworth expansion for a sum of independent random vectors, as given in [15] (Corollary 20.4, p. 215; see also [51], Lemma 5.4, p. 263). Though, an asymptotic Edgeworth expansion of the probability $\mathbb{P}_0(\sqrt{n}\bar{\mathbf{X}} > \mathbf{x})$ fails to converge or to adequately describe the probability tail for values of \mathbf{x} of size $n^{1/6}$ or smaller (see [51], p. 325 and [100], p. 194) and this rules out this approach. However, if the objective is to derive the exact distribution of $\hat{\theta}^n$, then an Edgeworth expansion can be used as an infinite series but the conditions for the application of such a result are indeed very restrictive. In the case $J = 1$, under Assumption Obj, an infinite Edgeworth expansion holds for $\mathbb{P}_0(\hat{\theta}^n = \theta_1)$ under conditions that can be found in [30] (p. 223) or [51] (p. 45) and that are too strict for many applications. Furthermore, moderate deviations do not lend themselves to approximate probabilities of the type of equation (3) since they only work if \mathbf{x} is of size $o(n^{1/2})$ (see e.g. [107], pp. 299-300).

The only approaches that have been successful in our experience are large deviations (in logarithmic and exact form) and saddlepoint approximations. Therefore, in Section 3.2.2, we derive some results on the asymptotic behavior of $\mathbb{P}_{\theta^*}(\hat{\theta}^n = \theta_i)$ using Large Deviations Principles (LDP). Remark that we could have defined the probability in (2) as $\mathbb{P}_{\theta^*}(Q_n(\theta_i) \geq Q_n(\theta_j), \forall j \neq i)$ or through any other combination of equality and inequality signs: this introduces some arbitrariness in the calculation of the asymptotic distribution of $\hat{\theta}^n$. A condition to be stated in the following (Cra- (i)) allows to prove a result ensuring that the difference is asymptotically negligible (see [51], Theorem 2.3, p. 57), but the result is not strong

enough to guarantee the same asymptotic distribution for alternative choices of the inequality signs: the point is made clearer in Section 3.2.2. However, we will give some alternative conditions (see Proposition 2) under which this difference is asymptotically irrelevant.

Section 3.2.1 introduces some definitions and hypotheses and discusses a preliminary result used in the large deviations estimate.

In Section 3.2.3, we refine the previous expressions for the convergence rate of $\hat{\theta}^n$ using the theory of exact asymptotics for large deviations for a random variable \mathbf{Y}_k as developed by [7, 94, 95, 96, 62]. The aim is to find asymptotic results, as $n \rightarrow \infty$, of the form:

$$(4) \quad \mathbb{P}_0 \left(\sum_{i=1}^n \mathbf{Y}_k \in n \cdot \Gamma \right) = n^\gamma \cdot e^{-n \cdot \inf_{\mathbf{y} \in \Gamma} I(\mathbf{y})} \cdot (d_0 + o(1)),$$

where $I(\cdot)$ is the rate function for which the LDP holds, and we have $\mathbf{Y}_k = \mathbf{X}_k^{(i)}$ and $\Gamma = \text{int } \mathbb{R}_+^J$. We only have to determine if a result such as (4) holds, and in this case, what the values of γ and d_0 are. At last, Section 3.2.5 derives saddlepoint approximations for probabilities of the form (2).

3.2.1. Definitions, Hypotheses and Preliminary Results. As concerns the asymptotic distribution of the m -estimator $\hat{\theta}^n$, we shall need some concepts and functions derived from large deviations theory (see [32]); we recall that the processes $Q_n(\theta_j)$, \mathbf{X}_k and $\mathbf{X}_k^{(i)}$ have been introduced in (1). Then we define the moment generating functions

$$M^{(i)}(\lambda) \triangleq \mathbb{E}_0 \left[e^{\sum_{j=0, \dots, J, j \neq i} \lambda_j \cdot [\ln q(Y; \theta_i) - \ln q(Y; \theta_j)]} \right] = \mathbb{E}_0 \left[e^{\lambda^\top \mathbf{X}^{(i)}} \right],$$

the logarithmic moment generating functions

$$\Lambda^{(i)}(\lambda) \triangleq \ln M^{(i)}(\lambda) = \ln \mathbb{E}_0 \left[e^{\sum_{j=0, \dots, J, j \neq i} \lambda_j \cdot [\ln q(Y; \theta_i) - \ln q(Y; \theta_j)]} \right] = \ln \mathbb{E}_0 \left[e^{\lambda^\top \mathbf{X}^{(i)}} \right],$$

and the Cram er $\frac{1}{2}$ transforms

$$\Lambda^{(i),*}(\mathbf{y}) \triangleq \sup_{\lambda \in \mathbb{R}^J} \left[\langle \mathbf{y}, \lambda \rangle - \Lambda^{(i)}(\lambda) \right],$$

where $\langle \cdot, \cdot \rangle$ is the scalar product. Remark that in what follows, $M(\lambda)$, $\Lambda(\lambda)$ and $\Lambda^*(\mathbf{y})$ are respectively shortcuts for $M^{(0)}(\lambda)$, $\Lambda^{(0)}(\lambda)$ and $\Lambda^{(0),*}(\mathbf{y})$. Moreover, for a function $f : E \rightarrow \overline{\mathbb{R}}$, we will need the definition of the *effective domain* of f , $\mathcal{D}_f \triangleq \{x \in E : f(x) < \infty\}$.

The following hypothesis will be needed to obtain the asymptotic distribution of $\hat{\theta}^n$.

η -Int: There exists a $\delta > 0$, such that, for any $\eta \in (-\delta, \delta)$ we have:

$$\mathbb{E}_0 \left[\frac{q(Y; \theta_j)}{q(Y; \theta_k)} \right]^\eta < +\infty, \quad \forall j, k = 0, \dots, J.$$

In the following (Lemma 1), we will show that the latter hypothesis is equivalent to the Cram er $\frac{1}{2}$ condition $\mathbf{0} \in \text{int}(\mathcal{D}_{\Lambda^{(i)}})$, for any $i = 0, \dots, J$.

Remark 3. In what follows, this Assumption could be substituted by a condition as in [97] (Assumptions H1 and H2).

Steep–(i): $\Lambda^{(i)}(\lambda)$ is *steep*, that is, $\lim_{n \rightarrow \infty} \left\| \frac{\partial \Lambda^{(i)}(\mathbf{x})}{\partial \mathbf{x}} \right\| = \infty$ whenever $\{\mathbf{x}_n\}_n$ is a sequence in $\text{int}(\mathcal{D}_{\Lambda^{(i)}})$ converging to a boundary point of $\text{int} \mathcal{D}_{\Lambda^{(i)}}$.

Remark 4. (i) We recall that f is *essentially smooth* if $\text{int}(\mathcal{D}_f)$ is nonempty, f is differentiable in $\text{int}(\mathcal{D}_f)$ and if $\lim_{n \rightarrow \infty} \left\| \frac{\partial f(\mathbf{x})}{\partial \mathbf{x}} \right\| = \infty$ whenever $\{\mathbf{x}_n\}$ is a sequence in $\text{int}(\mathcal{D}_f)$ converging to a boundary point of $\text{int}(\mathcal{D}_f)$ (see [95], p. 903, [32], p. 44). Under Assumptions **Obj**, η –**Int** and **Steep–(i)**, $\Lambda^{(i)}(\cdot)$ is essentially smooth (see e.g. [32], p. 44).

(ii) A sufficient condition for **Steep–(i)** and essential smoothness is openness of $\mathcal{D}_{\Lambda^{(i)}}$ (see [95], p. 905, and [62], pp. 505–506).

Meet–(i): $\text{int}(\mathbb{R}_+^J \cap \mathcal{S}^{(i)}) \neq \emptyset$, where $\mathcal{S}^{(i)}$ is the closure of the convex hull of the support of the law of $\mathbf{X}^{(i)}$.

L^r –**Sum–(i):** The variable $[n \cdot (Q_n(\theta_i) - Q_n(\theta_j))]_{j=0, \dots, J, j \neq i}$ has a density in L^r with $1 < r \leq +\infty$.

Remark 5. The assumption L^r –**Sum–(i)** can be substituted by a general nonlattice condition such as the forthcoming **Cra–(i)**, but the substitution would give rise to an oscillating error term (that is an $O(1)$ –term such that $\limsup O(1) \neq \liminf O(1)$ as $n \rightarrow \infty$; see [62], p. 521).

Cra–(i): (*Cramér’s $\frac{1}{2}$ condition*) $\limsup_{\|\lambda\| \rightarrow \infty} |M^{(j)}(\iota\lambda)| < 1, \forall j = 0, \dots, J$, where ι is the imaginary unit and $M^{(j)}(\iota\lambda)$ is a formal expression for the characteristic function.

We will need the following Lemma showing the equivalence between Assumption η –**Int** and the so-called *Cramér’s $\frac{1}{2}$ condition* $\mathbf{0} \in \text{int}(\mathcal{D}_{\Lambda^{(i)}})$, for any $i = 0, \dots, J$ (beware of confusion with Assumption **Cra–(i)**, also called in the literature *Cramér’s $\frac{1}{2}$ condition*).

Lemma 1. *Under Assumption **Obj**, the following conditions are equivalent:*

- (i):** Assumption η –**Int** holds;
- (ii):** $\mathbf{0} \in \text{int}(\mathcal{D}_{\Lambda^{(i)}})$, for any $i = 0, \dots, J$.

As concerns the saddlepoint approximation of Section 3.2.5, we need the following hypothesis.

s–TiltCF–(i):

$$\left| \mathbb{E}_0 \left[\prod_{j=0, \dots, J, j \neq i} \left(\frac{q(Y; \theta_i)}{q(Y; \theta_j)} \right)^{u_j + \iota t_j} \right] \right| < (1 - \delta) \cdot \left| \mathbb{E}_0 \left[\prod_{j=0, \dots, J, j \neq i} \left(\frac{q(Y; \theta_i)}{q(Y; \theta_j)} \right)^{u_j} \right] \right| < \infty$$

for $\mathbf{u} \in \text{int}(\mathcal{D}_{\Lambda^{(i)}})$, $\delta > 0$ and $c < |\mathbf{t}| < C \cdot n^{\frac{s-3}{2}}$ (ι denotes the imaginary unit). Moreover, the distribution of $\mathbf{X}^{(i)}$ is not concentrated on an affine subspace of \mathbb{R}^J .

3.2.2. Large Deviations Asymptotics. In this Section we consider large deviation asymptotics. We remark that, in what follows, $\text{int}(\mathbb{R}_+^J)^c$ stands for $\text{int} \left\{ \left[(\mathbb{R}_+)^J \right]^c \right\}$.

Proposition 2. (i) *For $i = 1, \dots, J$, under Assumption **Obj**, the following result hold:*

$$\mathbb{P}_0 \left(\hat{\theta}^n = \theta_i \right) \geq \exp \left\{ -n \cdot \inf_{\mathbf{y} \in \text{int}(\mathbb{R}_+^J)} \Lambda^{(i),*}(\mathbf{y}) + o_{\text{inf}}(n) \right\}$$

where $o_{\inf}(n)$ is a function such that $\liminf_{n \rightarrow \infty} \frac{o_{\inf}(n)}{n} = 0$.

(ii) Under Assumptions **Obj** and η -**Int**:

$$\mathbb{P}_0 \left(\hat{\theta}^n = \theta_i \right) \leq \exp \left\{ -n \cdot \inf_{\mathbf{y} \in \mathbb{R}_+^J} \Lambda^{(i),*}(\mathbf{y}) - o_{\sup}(n) \right\},$$

where $o_{\sup}(n)$ is a function such that $\limsup_{n \rightarrow \infty} \frac{o_{\sup}(n)}{n} = 0$.

(iii) Under Assumptions **Obj**, η -**Int**, **Meet**-(i) and **Steep**-(i):

$$\begin{aligned} \mathbb{P}_0 \left(\hat{\theta}^n = \theta_i \right) &= \exp \left\{ -(n + o(n)) \cdot \inf_{\mathbf{y} \in \text{int}(\mathbb{R}_+^J)} \Lambda^{(i),*}(\mathbf{y}) \right\} \\ &= \exp \left\{ -(n + o(n)) \cdot \inf_{\mathbf{y} \in \mathbb{R}_+^J} \Lambda^{(i),*}(\mathbf{y}) \right\}. \end{aligned}$$

Proposition 3. Under Assumption **Obj**, the following inequality holds:

$$\mathbb{P}_0 \left(\hat{\theta}^n \neq \theta_0 \right) \geq H \cdot \exp \left\{ -n \cdot \inf_{\mathbf{y} \in \text{int}(\mathbb{R}_+^J)^c} \Lambda^*(\mathbf{y}) + o_{\inf}(n) \right\},$$

where H is the finite cardinality of the set $\arg \inf_{\mathbf{y} \in \text{int}(\mathbb{R}_+^J)^c} \Lambda^*(\mathbf{y})$ and $o_{\inf}(n)$ is a function such that $\liminf_{n \rightarrow \infty} \frac{o_{\inf}(n)}{n} = 0$.

Under Assumptions **Obj** and η -**Int**:

$$\mathbb{P}_0 \left(\hat{\theta}^n \neq \theta_0 \right) \leq H \cdot \exp \left\{ -n \cdot \inf_{\mathbf{y} \in \mathbb{R}_+^J} \Lambda^{(i),*}(\mathbf{y}) + o_{\sup}(n) \right\},$$

where $o_{\sup}(n)$ is a function such that $\limsup_{n \rightarrow \infty} \frac{o_{\sup}(n)}{n} = 0$.

Remark 6. The Proposition allows us to obtain an upper bound on the bias of the m -estimator. Indeed,

$$\mathbb{E}_0 \hat{\theta}^n = \sum_{j=0}^J \theta_j \cdot \mathbb{P}_0 \left(\hat{\theta}^n = \theta_j \right) \leq \theta_0 + \sup_{j \neq 0} |\theta_j - \theta_0| \cdot \mathbb{P}_0 \left(\hat{\theta}^n \neq \theta_0 \right),$$

and:

$$\text{Bias} \left(\hat{\theta}^n \right) \leq \sup_{j \neq 0} |\theta_j - \theta_0| \cdot \mathbb{P}_0 \left(\hat{\theta}^n \neq \theta_0 \right).$$

3.2.3. Exact Asymptotics of Large Deviations. The exact asymptotic behavior of the probability $\mathbb{P}_0 \left(\hat{\theta}^n = \theta_i \right)$ can be derived, under some additional conditions, from a better study of the neighborhood of the contact point between the set $(\mathbb{R}_+)^J$ and the level sets of the Cramér transform $\Lambda^{(i),*}(\cdot)$. Therefore, in the following, we will consider the probabilities of equation (2), that is $\mathbb{P}_0 \left(\hat{\theta}^n = \theta_i \right) = \mathbb{P}_0 \left(\sum_{k=1}^n \mathbf{X}_k^{(i)} \in \mathbb{R}_+^J \right)$, for $i = 1, \dots, J$. The probability $\mathbb{P}_0 \left(\hat{\theta}^n = \theta_0 \right)$ will be obtained by subtraction.

Suppose that Γ is a $\Lambda^{(i),*}$ -continuity set, that is the large deviation estimate has the form:

$$\lim_{n \rightarrow \infty} \frac{1}{n} \ln \mathbb{P}_0 \left(\sum_{k=1}^n \mathbf{X}_k^{(i)} \in \Gamma \right) = - \inf_{\mathbf{y} \in \Gamma} \Lambda^{(i),*}(\mathbf{y});$$

see Proposition 2 (**iii**) for a condition ensuring that this holds true for $\Gamma = \mathbb{R}_+^J$. A point $\mathbf{y}^{(i)}$ is a *dominating point* (see [96], p. 79) of Γ if:

- $\mathbf{y}^{(i)} \in \partial\Gamma$ and $\inf_{\mathbf{y} \in \Gamma} [\Lambda^{*,(i)}(\mathbf{y})] = \Lambda^{*,(i)}(\mathbf{y}^{(i)})$;
- the equation $\frac{\partial \Lambda^{(i)}(\lambda)}{\partial \lambda} = \mathbf{y}^{(i)}$ has a solution $\lambda(\mathbf{y}^{(i)}) \triangleq \lambda^{(i)} \in \mathcal{D}_{\Lambda^{(i)}}$;
- $\Gamma \subset \Pi^+(\lambda^{(i)}, \mathbf{y}^{(i)})$, where $\Pi^+(\alpha, \mathbf{x}) \triangleq \{\mathbf{z} \in \mathbb{R}^J \mid \mathbf{z}^\top \alpha \geq \mathbf{x}^\top \alpha\}$ is an half-space.

Then, we have the following result.

Proposition 4. *Under Assumptions **Obj**, η -**Int**, **Meet**-(i) and **Steep**-(i), for sufficiently large n , the following result holds:*

$$c_1 \frac{e^{-n \cdot \lambda^{(i), \top} \mathbf{y}^{(i)}} \cdot [M^{(i)}(\lambda^{(i)})]^n}{n^{J/2}} \leq \mathbb{P}_0(\hat{\theta}^n = \theta_i) \leq c_2 \frac{e^{-n \cdot \lambda^{(i), \top} \mathbf{y}^{(i)}} \cdot [M^{(i)}(\lambda^{(i)})]^n}{n^{1/2}},$$

for $i = 1, \dots, J$ and for some $0 < c_1 \leq c_2 < +\infty$.

Now, we try to specialize these estimates to obtain the precise convergence rates for probabilities of the form $\mathbb{P}_0(\hat{\theta}^n = \theta_i)$. To do so, we introduce the following definitions.

We define the *level surface* $S_a^{(i)} \triangleq \{\mathbf{y} \in \mathbb{R}^J \mid \Lambda^{(i),*}(\mathbf{y}) = a\}$ and the *level set* $V_a^{(i)} \triangleq \{\mathbf{y} \in \mathbb{R}^J \mid \Lambda^{(i),*}(\mathbf{y}) \geq a\}$ and we set $\Lambda^{(i),*} \triangleq \Lambda^{(i),*}(\mathbf{y}^{(i)})$; we take a new coordinate system centered at $\mathbf{y}^{(i)}$, we make the change of variable $\mathbf{z} = \mathbf{R}^{(i)} \cdot \mathbf{y}$, where $\mathbf{R}^{(i)}$ is the rotation that takes $\lambda^{(i)}$ into $\mathbf{R}^{(i)} \cdot \lambda^{(i)} \triangleq (0, \dots, 0, \|\lambda^{(i)}\|)$. Therefore, the new direction $z_J^{(i)}$ points in the direction normal to $S_{\Lambda^{(i),*}}^{(i)}$. The directions $(z_1^{(i)}, \dots, z_J^{(i)})$ form an orthonormal basis and the set Γ becomes $\{\mathbf{R}^{(i)} \cdot \mathbf{y} \mid \mathbf{y} \in \Gamma - \mathbf{y}^{(i)}\}$. In the following we will use the new basis: in order to avoid complications, no new notation will be adopted for the translated and rotated Γ and $\Lambda^{(i),*}$, but they will be expressed in terms of the vector $\mathbf{y}_{(J)} \triangleq (y_1, \dots, y_{J-1})$ and $\mathbf{y} = (\mathbf{y}_{(J)}, y_J)$.

Then, in this new system, near $\mathbf{y}^{(i)} = \mathbf{0}$, $y_J = g(\mathbf{y}_{(J)})$ describes $\partial\Gamma$ and $y_J = f(\mathbf{y}_{(J)})$ describes $S_{\Lambda^{(i),*}}^{(i)}$.

Proposition 5. *Under Assumptions **Obj**, η -**Int**, **Meet**-(i) and **Steep**-(i), we have:*

(**I**): $\mathbf{y}^{(i)} = \mathbf{0}$ and $\lambda^{(i)} \in \text{int}(\mathbb{R}_+^J)$: under Assumption L^r -**Sum**-(i), there exists a constant d_0 such that:

$$\mathbb{P}_0(\hat{\theta}^n = \theta_i) = \frac{e^{-n \cdot [\lambda^{(i), \top} \mathbf{y}^{(i)} - \Lambda^{(i)}(\lambda^{(i)})]}}{n^{J/2}} \cdot (d_0 + O(n^{-1})),$$

where:

$$d_0 \leq \frac{(\det \mathbf{H}_{\Lambda^{(i)}}(\lambda^{(i)}))^{-1/2}}{2^{J/2} \pi^{1/2} \|\lambda^{(i)}\|^J} \left(\frac{\|\lambda^{(i)}\|^2 - \max_j (\lambda_j^{(i)})^2}{\max_j (\lambda_j^{(i)})^2} \right)^{\frac{J-1}{2}} \frac{\Gamma(J)}{\Gamma(\frac{J-1}{2})},$$

$$d_0 \geq \frac{(\det \mathbf{H}_{\Lambda^{(i)}}(\lambda^{(i)}))^{-1/2}}{2^{J/2} \pi^{1/2} \|\lambda^{(i)}\|^J} \left(\frac{\min_j (\lambda_j^{(i)})^2}{\|\lambda^{(i)}\|^2 - \min_j (\lambda_j^{(i)})^2} \right)^{\frac{J-1}{2}} \frac{\Gamma(J)}{\Gamma(\frac{J-1}{2})}.$$

(II): $\mathbf{y}^{(i)}$ belongs to the interior of a face of \mathbb{R}_+^J : then, under Assumption **Cra**–(i), we have:

$$\mathbb{P}_0 \left(\hat{\theta}^n = \theta_i \right) = \frac{e^{-n \cdot [\lambda^{(i), \top} \mathbf{y}^{(i)} - \Lambda^{(i)}(\lambda^{(i)})]}}{n^{1/2}} \cdot (d_0 + o(n^{-\delta}))$$

for a certain $\delta > 0$. If, moreover, Assumption L^r –**Sum**–(i) holds and the Hessian $\mathbf{H}_f(\mathbf{0})$ is negative definite:

$$d_0 = \frac{(\det \mathbf{H}_{\Lambda^{(i)}}^{-1}(\lambda^{(i)}))^{-1/2} \{ \det[\|\lambda^{(i)}\| \cdot (\mathbf{H}_g(\mathbf{0}) - \mathbf{H}_f(\mathbf{0}))] \}^{-1/2}}{(2\pi)^{1/2} \|\lambda^{(i)}\|}$$

and $o(n^{-\delta})$ can be substituted with $O(n^{-1})$.

Remark 7. These results can be useful to derive an asymptotic expression for $\mathbb{P}_0(\hat{\theta}^n \neq \theta_0)$. Indeed, following the proof of Proposition 3, this probability can be written as:

$$\mathbb{P}_0 \left(\sum_{k=1}^n \mathbf{X}_k \in \text{int}(\mathbb{R}_+^J)^c \right) = (1 + o(1)) \cdot \sum_{h=1}^H \mathbb{P}_0 \left(\sum_{k=1}^n \mathbf{X}_k \in \text{int} \Gamma_h \right).$$

The points of $\arg \inf_{\mathbf{y} \in \text{int}(\mathbb{R}_+^J)^c} \Lambda^*(\mathbf{y})$ belong to $(J-1)$ -dimensional faces of $(\mathbb{R}_+^J)^c$ and therefore, Proposition 5 can be applied. The obtained formulas are, however, overwhelmingly complicated but in the case $H = 1$.

3.2.4. *Asymptotics for $J = 1$.* When $J = 1$, the previous results can be obtained under the following hypothesis.

Min: When $J = 1$, there is a positive value $\mu \in \text{int}(\mathcal{D}_{\Lambda^{(1)}})$ such that:

$$\left. \frac{\partial \Lambda^{(1)}(\lambda)}{\partial \lambda} \right|_{\lambda=\mu} = 0.$$

Moreover, the law of $\ln \frac{q(Y; \theta_1)}{q(Y; \theta_0)}$ is nonlattice (see [32], p. 110).

The following Proposition allows us to obtain a precise estimate of the convergence rate.

Proposition 6. *Under Assumptions **Obj**, η –**Int**, **Meet**–(i), **Steep**–(i) and **Min**, with $\Theta = \{\theta_0, \theta_1\}$ and $J = 1$, we have:*

$$\begin{aligned} \mathbb{P}_0 \left(\hat{\theta}^n = \theta_1 \right) &= \mathbb{P}_0 \left(\hat{\theta}^n \neq \theta_0 \right) = \frac{e^{n \cdot \Lambda^{(1)}(\mu)}}{\mu \cdot \sqrt{\Lambda^{(1),''}(\mu)} 2\pi n} \cdot (1 + o(1)) \\ &= \frac{e^{-n \cdot \Lambda^{(1),*}(0)}}{(\Lambda^{(1),*})'(0)} \cdot \sqrt{\frac{(\Lambda^{(1),*})''(0)}{2\pi n}} \cdot (1 + o(1)). \end{aligned}$$

Remark 8. A refinement of the previous asymptotic rates can be obtained using results in [16, 7].

3.2.5. *Saddlepoint Approximation.* In this Section we consider a different kind of approximation of the probabilities $\mathbb{P}_0(\hat{\theta}^n = \theta_i)$.

Theorem 1. Under Assumptions **Obj** and s -**TiltCF**-(i), for $i \neq 0$, it is possible to choose \mathbf{u} such that, for every $\mathbf{v} \in \left[\left(\text{int } \mathbb{R}_+^J \right) \ominus \frac{\partial \Lambda^{(i)}(\mathbf{u})}{\partial \mathbf{u}} \right]$, $\mathbf{u}^\top \mathbf{v} \geq 0$ and:

$$\begin{aligned} \mathbb{P}_0 \left(\hat{\theta}^n = \theta_i \right) &= \exp \left(n \left[\Lambda^{(i)}(\mathbf{u}) - \mathbf{u} \cdot \frac{\partial \Lambda^{(i)}(\mathbf{u})}{\partial \mathbf{u}} \right] \right) \\ &\cdot \left[e_{s-3} \left(\mathbf{u}, \text{int } \mathbb{R}_+^J \ominus \mathbb{E}_0 \mathbf{X}^{(i)} \right) + \delta \left(\mathbf{u}, \text{int } \mathbb{R}_+^J \ominus \mathbb{E}_0 \mathbf{X}^{(i)} \right) \right] \end{aligned}$$

where:

$$\begin{aligned} e_{s-3} \left(\mathbf{u}, \text{int } \mathbb{R}_+^J \ominus \mathbb{E}_0 \mathbf{X}^{(i)} \right) &= \int_{\text{int } \mathbb{R}_+^J \ominus \frac{\partial \Lambda^{(i)}(\mathbf{u})}{\partial \mathbf{u}}} \frac{\exp \left(-n \mathbf{u} \cdot \mathbf{y} - n \|\mathbf{y}^*\|^2 / 2 \right)}{(2\pi/n)^{J/2} \Delta^{1/2}} \\ &\cdot \left[1 + \sum_{i=1}^{s-3} n^{-i/2} Q_{i\mathbf{u}}(\sqrt{n}\mathbf{y}^*) \right] d\mathbf{y} \\ Q_{\ell\mathbf{u}}(\mathbf{x}) &= \sum_{m=1}^{\ell} \frac{1}{m!} \sum^* \sum^{**} \left(\frac{\kappa_{\nu_1 n} \cdots \kappa_{\nu_m n}}{\nu_1! \cdots \nu_m!} \right) \cdot H_{I_1}(x_1) \cdots H_{I_d}(x_d) \\ \left| \delta \left(\mathbf{u}, \text{int } \mathbb{R}_+^J \ominus \mathbb{E}_0 \mathbf{X}^{(i)} \right) \right| &\leq C \cdot n^{-\frac{s-2}{2}} \end{aligned}$$

and $\mathbf{V} = \frac{\partial^2 \Lambda^{(i)}(\mathbf{u})}{\partial \mathbf{u}^2}$, $\mathbf{y}^* = \mathbf{V}^{-1/2} \mathbf{y}$, $\|\mathbf{y}^*\|^2 = \mathbf{y}^* \cdot \mathbf{y}^* = \mathbf{y}^\top \mathbf{V}^{-1} \mathbf{y}$, $\Delta = |\mathbf{V}|$, H_m is the usual Hermite-Chebyshev polynomial of degree m , \sum^* denotes the sum over all m -tuples of positive integers (j_1, \dots, j_m) satisfying $j_1 + \dots + j_m = \ell$, \sum^{**} denotes the sum over all m -tuples (ν_1, \dots, ν_m) with $\nu_i = (\nu_{1i}, \dots, \nu_{di})$, satisfying $(\nu_{1i} + \dots + \nu_{di} = j_i + 2, i = 1, \dots, m)$, and $I_h = \nu_{h1} + \dots + \nu_{hm}$, $h = 1, \dots, d$. Remark that $Q_{\ell\mathbf{u}}$ depends on \mathbf{u} through the cumulants calculated at \mathbf{u} .

Remark 9. (i) The main question that this Theorem leaves open is the choice of the point \mathbf{u} . Usually this point is chosen as a solution $\hat{\mathbf{u}}$ of $\mathbf{m}(\hat{\mathbf{u}}) = \hat{\mathbf{x}}$: this corresponds to a saddlepoint in $\kappa(\mathbf{u})$. [31] (Section 6) and [85] (p. 480) give some conditions for $J = 1$, [64] (p. 23) and [12] (p. 153) give conditions for general J . [65] suggests that the most common solution is to choose $\hat{\mathbf{x}}$ and $\hat{\mathbf{u}}$ ($\hat{\mathbf{x}}$ belonging to the boundary of $[\text{int } \mathbb{R}_+^J \ominus \mathbb{E}_0 \mathbf{X}^{(i)}]$ and $\hat{\mathbf{u}}$ solving $\mathbf{m}(\hat{\mathbf{u}}) = \hat{\mathbf{x}}$), such that for every $\mathbf{v} \in \left[\text{int } \mathbb{R}_+^J \ominus \frac{\partial \Lambda^{(i)}(\mathbf{u})}{\partial \mathbf{u}} \right]$, $\hat{\mathbf{u}}^\top \mathbf{v} \geq 0$. This is the same as a dominating point in [94, 95, 96]: therefore, η -**Int**, **Meet**-(i) and **Steep**-(i), for sufficiently large n , imply the existence of this point for any i .

(ii) To obtain a more manageable formula for e_{s-3} , it can be useful to remark that it is a Laplace type integral (see [17], p. 321), i.e. an integral of the form $\int \exp \{ \lambda \phi(\mathbf{x}) \} g(\mathbf{x}) d\mathbf{x}$ with real λ tending to infinity, and can therefore be approximated as in [17].

3.3. The MLE and Bayes Estimators in Discrete Parameter Models . In this Section, we show how the previous results can be applied to the MLE and Bayes estimators under the zero-one loss function. The MLE is defined by:

$$\hat{\theta}^n \triangleq \arg \max_{\theta \in \Theta} \prod_{i=1}^n f_{Y_i}(y_i; \theta_k) = \arg \max_{\theta \in \Theta} \left[\frac{1}{n} \sum_{i=1}^n \ln f_{Y_i}(y_i; \theta) \right].$$

This corresponds to the *minimum-error-probability estimate* of [101] and to the *Bayesian estimator* of [119, 120]. On the other hand, using the prior densities

given by $\pi(\theta)$ for $\theta \in \Theta$, the posterior densities of the Bayesian estimator are given by:

$$\mathbb{P}\{\theta_k | \mathbf{Y}\} = \frac{\prod_{i=1}^n f_{Y_i}(y_i; \theta_k) \pi(\theta_k)}{\sum_{j=0}^J \prod_{i=1}^n f_{Y_i}(y_i; \theta_j) \pi(\theta_j)}.$$

The Bayes estimator relative to zero-one loss $\check{\theta}^n$ (see Section 4.1 for a definition) is the mode of the posterior distribution and is given by:

$$\begin{aligned} \check{\theta}^n &\triangleq \arg \max_{\theta \in \Theta} \ln \mathbb{P}\{\theta | \mathbf{Y}\} \\ (5) \quad &= \arg \max_{\theta \in \Theta} \left[\frac{1}{n} \sum_{i=1}^n \ln f_{Y_i}(y_i; \theta) + \frac{\ln \pi(\theta)}{n} \right]. \end{aligned}$$

Remark that the MLE coincides with the Bayes estimator corresponding to the uniform distribution $\pi(\theta) = (J+1)^{-1}$ for any $\theta \in \Theta$.

The hypothesis **Obj** can be substituted by the following ones.

Hom: The parametric statistical model \mathcal{P} is formed by a set of probability measures on a measurable space (Ω, \mathcal{A}) indexed by a parameter θ ranging over a parameter space $\Theta = \{\theta_0, \theta_1, \dots, \theta_J\}$, of finite cardinality. Let $(\mathfrak{Y}, \mathcal{Y})$ a measurable space and μ a positive σ -finite measure defined on $(\mathfrak{Y}, \mathcal{Y})$ such that, for every $\theta \in \Theta$, \mathbb{P}_θ is equivalent to μ ; the densities $f_Y(Y; \theta)$ are \mathcal{Y} -measurable for each $\theta \in \Theta$.

The data $(Y_i)_{i=1}^n$ are independent and identically distributed realizations of the probability measure \mathbb{P}_0 .

L^1 -**Dom:** The log density satisfies the following L^1 -domination condition:

$$\mathbb{E}_0 |\ln f_Y(Y; \theta_i)| < +\infty, \quad \forall \theta_i \in \Theta,$$

where \mathbb{E}_0 denotes the expectation taken under the true probability measure \mathbb{P}_0 .

Id: θ_0 is the point of Θ maximizing $\mathbb{E}_0 \ln f_Y(Y; \theta)$ and $\theta_0 \in \Theta$ is globally identified.

Remark 10. The previous two assumptions are standard requirements entailing that the likelihood function is asymptotically maximized at θ_0 only.

In order to obtain the consistency of Bayes estimators, we need the following hypothesis on the behavior of the prior distribution.

Bay: The prior distribution verifies $\pi(\theta) > 0$ for any $\theta \in \Theta$.

Proposition 1 holds for the MLE under Assumptions **Hom**, L^1 -**Dom** and **Id**, while for Bayes estimators **Bay** is required too. Remark that, under correct specification (that is when the true parameter value belongs to Θ), a standard Wald's argument (see e.g. Lemma 2.2 in [93], p. 2124) shows that $\mathbb{E}_{\theta_0} \ln f_Y(Y; \theta)$ is maximized for $\theta = \theta_0$.

As concerns the asymptotic distribution of the MLE, we have to consider the case in which $q(y; \theta)$ is given by $f_Y(y; \theta)$, $Q_n(\theta)$ by the loglikelihood function $L_n(\theta)$ and \mathbf{X}_k and $\mathbf{X}_k^{(i)}$ by the loglikelihood processes:

$$\begin{cases} L_n(\theta_j) \triangleq \frac{1}{n} \cdot \sum_{i=1}^n \ln f_{Y_i}(y_i; \theta_j), \\ \mathbf{X}_k^{(i)} \triangleq [\ln f_{Y_k}(Y_k; \theta_i) - \ln f_{Y_k}(Y_k; \theta_j)]_{j=0, \dots, J, j \neq i}, \\ \mathbf{X}_k \triangleq [\ln f_{Y_k}(Y_k; \theta_0) - \ln f_{Y_k}(Y_k; \theta_j)]_{j=1, \dots, J}. \end{cases}$$

Also $M(\lambda)$ and $M^{(i)}(\lambda)$ are consequently defined. Propositions 2 and 3 hold when Assumption **Obj** is substituted by Assumptions **Hom**, L^1 -**Dom** and **Id**.

When the model is correctly specified, it is interesting to stress an interpretation of the moment generating function in discrete parameter models. We remark that the moment generating functions can be written as follows:

$$\begin{aligned} M^{(i)}(\lambda) &\triangleq \mathbb{E}_{\theta_0} \left[e^{\sum_{j=0, \dots, J, j \neq i} \lambda_j \cdot [\ln f_Y(Y; \theta_i) - \ln f_Y(Y; \theta_j)]} \right] \\ (6) \quad &= \int f_Y(y; \theta_i)^{\sum_{j=0, \dots, J, j \neq i} \lambda_j} \cdot \prod_{j=1, \dots, J, j \neq i} f_Y(y; \theta_j)^{-\lambda_j} \cdot f_Y(y; \theta_0)^{1-\lambda_0} \mu(dy). \end{aligned}$$

Therefore, in this case, the moment generating function $M^{(i)}(\lambda)$ reduces to the so-called Hellinger transform $H_\gamma(\theta_0, \dots, \theta_J)$ (see [81], p. 27) for a certain linear transformation of λ in γ :

$$H_\gamma(\theta_0, \dots, \theta_J) \triangleq \int \prod_{j=0}^J [\mathbb{P}_{\theta_j}(dy)]^{\gamma_j} = \int \left[\prod_{j=0}^J f_Y(y; \theta_j)^{\gamma_j} \right] \mu(dy), \quad \sum_{j=0}^J \gamma_j = 1.$$

Moreover, due to its convexity, $H_\gamma(\theta_0, \dots, \theta_J)$ is surely finite for γ belonging to the closed simplex in \mathbb{R}^{J+1} .

Other quantities strictly related are the *Hellinger integral of order* $u \in (0, 1)$ (see e.g. [63], p. 192, [110], p. 363):

$$H_u(\mathbb{P}_{\theta_0}, \mathbb{P}_{\theta_1}) \triangleq \mathbb{E}_{\theta_0} \left(\frac{f_Y(Y; \theta_1)}{f_Y(Y; \theta_0)} \right)^u = \int f_Y(y; \theta_1)^u \cdot f_Y(y; \theta_0)^{1-u} \mu(dy),$$

the *Hellinger arc* (see [41]), that is the set of measures defined by:

$$\left\{ \frac{\int f_Y(y; \theta_1)^u \cdot f_Y(y; \theta_0)^{1-u} \mu(dy)}{\int f_Y(y; \theta_1)^u \cdot f_Y(y; \theta_0)^{1-u} \mu(dy)}, u \in [0, 1] \right\},$$

the *error function of order* $u \in (0, 1)$ ([1], p. 208):

$$uE(\mathbb{P}_{\theta_0} | \mathbb{P}_{\theta_1}) \triangleq \frac{1}{u-1} \cdot \ln \left(\int f_Y(y; \theta_1)^u \cdot f_Y(y; \theta_0)^{1-u} \mu(dy) \right),$$

the γ -*deviation of order* $\gamma \in [-1, 2]$ ([3]) on the space of finite positive measures:

$$D_\gamma(\mu, \nu) \triangleq \int \frac{\gamma d\mu + (1-\gamma) d\nu - (d\mu)^\gamma (d\nu)^{(1-\gamma)}}{\gamma(1-\gamma)}$$

and the *Chernoff index* ([25], pp. 500, 507, [26], p. 17, [29], p. 314):

$$D(Y) \triangleq -\ln \left[\inf_{1 > u > 0} \int f_Y(y; \theta_1)^u \cdot f_Y(y; \theta_0)^{1-u} \mu(dy) \right].$$

Proposition 4 holds if Assumption **Obj** is substituted by Assumptions **Hom**, L^1 -**Dom** and **Id**, and if η -**Int** and **Steep**-(*i*) hold true. However, hypothesis **Meet**-(*i*) is unnecessary: indeed, the fact that $\text{int}(\mathbb{R}_+^J \cap \mathcal{S}^{(i)}) \neq \emptyset$ can be proved showing that $\mathbf{0} \in \text{int}(\mathcal{S}^{(i)})$. This is equivalent to the existence, for $j = 1, \dots, J, j \neq i$, of two sets A_j^* and A_j^{**} of positive μ -measure and included in the support of Y , such that for $y_j^* \in A_j^*$ and $y_j^{**} \in A_j^{**}$:

$$f_Y(y_j^*; \theta_i) > f_Y(y_j^*; \theta_j), \quad f_Y(y_j^{**}; \theta_i) < f_Y(y_j^{**}; \theta_j).$$

Now, this follows easily remarking that these densities have to integrate to 1, are as different according to Assumption **Id** and have the same support according to Assumption **Hom**.

Proposition 5 holds if Assumption **Obj** is replaced by Assumptions **Hom**, L^1 -**Dom** and **Id**.

In order to derive the asymptotic distribution of Bayes estimators, we consider equation (5) and we let $\ln \pi^{(i)} \triangleq \left[\ln \frac{\pi(\theta_i)}{\pi(\theta_j)} \right]_{j=0, \dots, J, j \neq i}$. Then, we can write:

$$\begin{aligned} \mathbb{P}_0(\check{\theta}^n = \theta_i) &= \mathbb{P}_0 \left(\sum_{k=1}^n \mathbf{X}_k^{(i)} + \ln \pi^{(i)} \in \text{int } \mathbb{R}_+^J \right) \\ &= \mathbb{P}_0 \left(\sum_{k=1}^n \mathbf{X}_k^{(i)} \in \prod_{j=0, \dots, J, j \neq i} \left(\ln \frac{\pi(\theta_i)}{\pi(\theta_j)}, +\infty \right) \right), \end{aligned}$$

and we can use the previous large deviations or saddlepoint formulas, simply changing the set over which the inf is taken. However, care is needed since both formulas hold under the hypothesis:

$$\mathbb{E}_0 \mathbf{X}_k^{(i)} + \frac{1}{n} \cdot \ln \pi^{(i)} \in \text{int } (\mathbb{R}_+^J)^c.$$

In the case $J = 1$, the similarity of these formulas with the corresponding ones for a Neyman-Pearson test is striking; this revives the interpretation of a Neyman-Pearson test as a Bayesian estimation problem. Therefore, our analysis can be seen as a (minor) extension of the theory of hypothesis testing to a larger number of alternatives.

4. OPTIMALITY AND EFFICIENCY

In this Section, we are interested in the problem of efficiency, with special reference to maximum likelihood and Bayes estimators.

In the Statistics literature, efficiency (or superefficiency) can be defined comparing the behavior of the estimator with respect to a lower bound or, alternatively, to a class of estimators. In the regular case, the two concepts almost coincide (despite superefficiency). However, in the present case, the two concepts diverge dramatically and we need more care in the derivation of the information inequalities and in the statement of the efficiency properties.

In what follows, we will suppose that the true parameter value belongs to Θ : this will be reflected in the probabilities that will be written as $\mathbb{P}_0 = \mathbb{P}_{\theta_0}$. Indeed, efficiency statement for misspecified models are quite difficult to interpret.

Since the variance (or the MSE) as a measure of efficiency seems to be well suited only in cases in which the limiting distribution is Gaussian, in Section 4.1, we present some alternative risk functions; then we derive in Section 4.2 some lower bounds for these risk functions and we prove in Section 4.3 some optimality and efficiency results for the Bayes and ML estimators.

4.1. Risk Functions. An interesting problem concerns the choice of a measure of efficiency for the MLE in discrete parameter models: in his seminal paper, [52] derives a generalization of Cram er-Rao inequality for the variance that is also valid when the parameter space is countable. The same inequality has been derived, in

slightly more generality, in [24, 19]. Therefore, we will consider the following cost and risk functions:

$$\begin{aligned}\mathcal{C}_1(\tilde{\theta}^n, \theta_0) &= (\tilde{\theta}^n - \theta_0)^2, \\ \mathcal{R}_1(\tilde{\theta}^n, \theta_0) &\triangleq \mathbb{E}_{\theta_0} \mathcal{C}_1(\tilde{\theta}^n, \theta_0) = \text{MSE}(\tilde{\theta}^n).\end{aligned}$$

However, this choice is well-suited only in cases in which the variance or the MSE are good measures of risk. This is indeed the case if the limiting distribution of the normalized estimator is normal. Following the discussion by Lindley in [52], we consider also a different cost function $\mathcal{C}_2(\theta, \theta_0)$:

$$\mathcal{C}_2(\tilde{\theta}^n, \theta_0) = 1_{\{\tilde{\theta}^n \neq \theta_0\}};$$

the risk function is therefore given by the probability of missclassification:

$$\mathcal{R}_2(\tilde{\theta}^n, \theta_0) \triangleq \mathbb{E}_{\theta_0} \mathcal{C}_2(\tilde{\theta}^n, \theta_0) = \mathbb{P}_{\theta_0}(\tilde{\theta}^n \neq \theta_0).$$

Moreover, [69] shows that the estimator of the integral mean of a Gaussian sample with known variance is minimax efficient and admissible for the zero-one loss even in finite samples, while it is not with respect to the quadratic loss in finite samples.

The previous cost function has the drawback of weighting in the same way points of the parameter space that lies at different distances with respect to the true value θ_0 . In many cases, a more general loss function can be considered, as suggested in [46] (Vol. 1, p. 51) for multiple tests:

$$\mathcal{C}_3(\tilde{\theta}^n, \theta_0) = \begin{cases} 0 & \text{if } \tilde{\theta}^n = \theta_0 \\ a_j(\theta_0) & \text{if } \tilde{\theta}^n = \theta_j \end{cases}$$

where $a_j(\theta_0) > 0$ for $j = 1, \dots, J$. The risk function is therefore given by the weighted probability of missclassification:

$$\begin{aligned}\mathcal{R}_3(\tilde{\theta}^n, \theta_0) &\triangleq \mathbb{E}_{\theta_0} \mathcal{C}_3(\tilde{\theta}^n, \theta_0) = \mathbb{E}_{\theta_0} \left[\sum_{j=1}^J a_j(\theta_0) \cdot 1_{\{\tilde{\theta}^n = \theta_j\}} \right] \\ &= \sum_{j=1}^J a_j(\theta_0) \cdot \mathbb{P}_{\theta_0} \{ \tilde{\theta}^n = \theta_j \}.\end{aligned}$$

The $a_j(\theta_0)$'s can be tuned in order to give more or less weight to different points of the parameter space.

At last, we define the *Bayes risk* (under the zero-one loss function) associated with a prior distribution π on the parameter space Θ . In particular, we consider the Bayes risk under the risk function $\mathcal{R}_2(\tilde{\theta}^n, \theta_0)$ as:

$$r_2(\tilde{\theta}^n, \pi) \triangleq \sum_{j=0}^J \pi(\theta_j) \cdot \mathcal{R}_2(\tilde{\theta}^n, \theta_j) = \sum_{j=0}^J \pi(\theta_j) \cdot \mathbb{P}_{\theta_j}(\tilde{\theta}^n \neq \theta_j).$$

If $\pi(\theta_j) = (J+1)^{-1}$ we define $\mathbb{P}_e \triangleq r_2(\tilde{\theta}^n, \pi)$ as the *average probability of error*. Remark that this is indeed the measure of error used by [119, 120].

4.2. Information Inequalities. This Section contains lower bounds for the previously introduced risk functions and in particular for the risk function $\mathcal{R}_2(\tilde{\theta}^n, \theta_0)$ corresponding to the zero-one loss. In the specific case of discrete parameters, these generalize and unify the lower bounds proposed in [52, 24, 68, 50]. Lower bounds for more general cost functions can be obtained using, for example, Markov inequality. Indeed, if w_n is a strictly positive Borel function increasing on \mathbb{R}_+^* , then:

$$\begin{aligned}\mathbb{P}_{\theta_0}\left(\|\tilde{\theta}^n - \theta_0\| \geq k_n\right) &\leq \frac{\mathbb{E}_{\theta_0}\left[w_n\left(\|\tilde{\theta}^n - \theta_0\|\right)\right]}{w_n(k_n)}, \\ \mathbb{P}_{\theta_0}\left(\|\tilde{\theta}^n - \theta_0\| \geq k_n\right) &\leq \frac{\mathbb{E}_{\theta_0}\left[w_n\left(\frac{\|\tilde{\theta}^n - \theta_0\|}{k_n}\right)\right]}{w_n(1)}.\end{aligned}$$

In the following, first of all, a lower bound is proved and then, we obtain a minimax version of the same result. We will sometimes refer to the former as *Chapman-Robbins lower bound* (and to the related efficiency concept as *Chapman-Robbins efficiency*) since it recalls the lower bound proposed by these two authors in their 1951 paper. Then, from these results, we derive lower bounds for the MSE, for the weighted probability of missclassification and for the Bayes risk.

4.2.1. A Lower Bound for the Probability of Missclassification. The Proposition of this Section is intended to play the role of Cramér- $\frac{1}{2}$ -Rao and Chapman-Robbins lower bounds for the variance. It corresponds essentially to Stein's Lemma in hypothesis testing; the reduction of an estimation problem to a test between two hypotheses is a standard technique in the derivation of efficiency lower bounds and is attributed to Farrell ([33]; see also [79], for a related technique). Here, Stein's Lemma is applied as in Theorem 9.2 in [61] (p. 96), taking into account the fact that the parameter space is made up of more than two points. Moreover, a version of the same bound for estimators respecting condition (8) is provided; this corresponds to similar results proposed in [34, 22, 39].

Proposition 7. *Under Assumptions **Hom** and **Id**, for a strongly consistent estimator $\tilde{\theta}^n$:*

$$\begin{aligned}\lim_{n \rightarrow \infty} \frac{1}{n} \ln \mathcal{R}_2(\tilde{\theta}^n, \theta_0) &= \lim_{n \rightarrow \infty} \frac{1}{n} \ln \mathbb{P}_{\theta_0}(\tilde{\theta}^n \neq \theta_0) \\ (7) \qquad \qquad \qquad &\geq \sup_{\theta_1 \in \Theta \setminus \{\theta_0\}} \mathbb{E}_{\theta_1} \ln \left(\frac{f_Y(Y; \theta_0)}{f_Y(Y; \theta_1)} \right)\end{aligned}$$

On the other hand, if

$$(8) \qquad \qquad \qquad \limsup_{n \rightarrow \infty} \mathbb{P}_{\theta_j} \left\{ \tilde{\theta}^n \neq \theta_j \right\} < 1,$$

then:

$$\liminf_{n \rightarrow \infty} \frac{1}{n} \ln \mathcal{R}_2(\tilde{\theta}^n, \theta_0) \geq \sup_{\theta_1 \in \Theta \setminus \{\theta_0\}} \mathbb{E}_{\theta_1} \ln \left(\frac{f_Y(Y; \theta_0)}{f_Y(Y; \theta_1)} \right).$$

Remark 11. (i) Remark that this inequality only holds for estimators that are consistent or respect condition (8), while the one of Proposition 8 holds for any estimator.

(ii) The *inaccuracy rate* considered in [68] can be defined as (see also [6, 111, 37, 102] for related concepts):

$$e(\varepsilon, \theta_0, \tilde{\theta}^n) \triangleq -\liminf_{n \rightarrow \infty} \frac{1}{n} \cdot \ln \mathbb{P}_{\theta_0} \left(\|\tilde{\theta}^n - \theta_0\| > \varepsilon \right);$$

for any ε small enough ($\varepsilon < \min_{\theta_1 \in \Theta \setminus \{\theta_0\}} \|\theta_1 - \theta_0\|$), we have:

$$e(\varepsilon, \theta_0, \tilde{\theta}^n) = -\liminf_{n \rightarrow \infty} \frac{1}{n} \cdot \ln \mathbb{P}_{\theta_0} \left(\tilde{\theta}^n \neq \theta_0 \right) = -\liminf_{n \rightarrow \infty} \frac{1}{n} \cdot \ln \mathcal{R}_2 \left(\tilde{\theta}^n, \theta_0 \right),$$

and this leads to the following upper bound:

$$e(\varepsilon, \theta_0, \tilde{\theta}^n) \leq \inf_{\theta_1 \in \Theta \setminus \{\theta_0\}} \mathbb{E}_{\theta_1} \ln \left(\frac{f_Y(Y; \theta_1)}{f_Y(Y; \theta_0)} \right).$$

Since the bound of Proposition 1.1 of [68] coincides with the one implied by equation (7) and the attainment of this limit, that is Chapman-Robbins optimality, coincides with IR-optimality (see [68], p. 649), this enables us to use the results in [68] in order to prove that the MLE in discrete parameter models does not attain the Chapman-Robbins lower bound under the risk function \mathcal{R}_2 (Proposition 11).

4.2.2. *A Minimax Lower Bound for the Probability of Missclassification.* The following result is a minimax lower bound on the probability of missclassification. It is based on Neyman-Pearson Lemma and Chernoff's Bound.

Proposition 8. *Under Assumptions Hom and Id, for any estimator $\tilde{\theta}^n$, we have:*

$$\begin{aligned} & \liminf_{n \rightarrow \infty} \frac{1}{n} \ln \sup_{\theta_0 \in \Theta} \mathcal{R}_2 \left(\tilde{\theta}^n, \theta_0 \right) \\ &= \liminf_{n \rightarrow \infty} \frac{1}{n} \ln \sup_{\theta_0 \in \Theta} \mathbb{P}_{\theta_0} \left(\tilde{\theta}^n \neq \theta_0 \right) \geq \sup_{\theta_1 \in \Theta \setminus \{\theta_0\}} \sup_{\theta_0 \in \Theta} \Lambda^*(0) \\ (9) \quad &= \sup_{\theta_1 \in \Theta \setminus \{\theta_0\}} \sup_{\theta_0 \in \Theta} \ln \left[\inf_{1 > u > 0} \int f_Y(y; \theta_1)^u \cdot f_Y(y; \theta_0)^{1-u} \mu(dy) \right]. \end{aligned}$$

Remark 12. (i) Following [50] (Theorem 3.1, p. 50) and replacing in its proof Cauchy-Schwarz inequality with Hölder inequality, we could obtain the information inequality:

$$\begin{aligned} & \sup_{\theta_0 \in \Theta} \mathcal{R}_2 \left(\tilde{\theta}^n, \theta_0 \right) = \sup_{\theta_0 \in \Theta} \mathbb{P}_{\theta_0} \left(\tilde{\theta}^n \neq \theta_0 \right) \\ & \geq \sup_{\theta_1 \in \Theta \setminus \{\theta_0\}} \sup_{\theta_0 \in \Theta} \sup_{u > 1} \frac{1}{2^u \cdot \left[\mathbb{E}_{\theta_0} \left(\frac{f_Y(Y; \theta_1)}{f_Y(Y; \theta_0)} \right)^u \right]^{n \cdot (u-1)}}. \end{aligned}$$

It is not difficult to see that the lower bound of Proposition 8 is sharper than this one.

(ii) The previous Proposition allows us to obtain an expression for the *minimax Bahadur risk* (also called (*minimax*) *rate of inaccuracy*, see [5, 75]) analogous to Chernoff's Bound, thus providing a minimax version of the result in Remark 11 (ii).

(iii) Other methods to derive similar minimax inequalities are Fano's Inequality and Assouad's Lemma (see [80], Lemmas 8 and 9): however, in the present case they do not allow us to obtain tight bounds (for Fano's Inequality see [99], Section 5.1, pp. 132-133, and compare with the bound derived in our Proposition). This is apparently due to the fact that usual applications of these methods rely on the

approximation of the parameter space with a finite set of points Θ whose cardinality increases with n . Clearly, this cannot be done in the present case.

4.2.3. *Lower Bounds for the MSE.* The probability of the event $\left\{\left\|\hat{\theta}^n - \theta_0\right\| \geq \varepsilon\right\}$ for $\varepsilon > 0$ has often an exponential limit behavior (see [5], Lemma 5.2, p. 245, [37], p. 762), in the sense that there exists $\rho > 0$ such that:

$$\mathbb{P}_{\theta_0}\left(\left\|\hat{\theta}^n - \theta_0\right\| \geq \varepsilon\right) \leq \rho^n$$

for all sufficiently large n ; however, in the standard setting this result does not contrast with \sqrt{n} -consistency. Here we show that in discrete parameter models the asymptotic exponential behavior of the missclassification probability extends to the MSE too.

Indeed, the results of the previous Sections can easily be converted in corresponding results for the risk function $\mathcal{R}_1(\tilde{\theta}^n, \theta_0)$. The MSE of a generic estimator $\tilde{\theta}^n$ can be shown to be:

$$\begin{aligned} \text{MSE}\left(\tilde{\theta}^n\right) &= \mathbb{E}_{\theta_0}\left(\tilde{\theta}^n - \theta_0\right)^2 = \sum_{j=0}^J\left(\theta_j - \theta_0\right)^2 \cdot \mathbb{P}_{\theta_0}\left(\tilde{\theta}^n = \theta_j\right) \\ &\leq \sup_{\theta_1 \in \Theta \setminus \{\theta_0\}}\left|\theta_1 - \theta_0\right|^2 \cdot \mathbb{P}_{\theta_0}\left(\tilde{\theta}^n \neq \theta_0\right), \\ \text{MSE}\left(\tilde{\theta}^n\right) &\geq \inf_{\theta_1 \in \Theta \setminus \{\theta_0\}}\left|\theta_1 - \theta_0\right|^2 \cdot \mathbb{P}_{\theta_0}\left(\tilde{\theta}^n \neq \theta_0\right). \end{aligned}$$

This shows that the MSE of a generic estimator is given by:

$$(10) \quad \mathbb{E}_{\theta_0}\left(\tilde{\theta}^n - \theta_0\right)^2 = K_n \cdot \mathbb{P}_{\theta_0}\left(\tilde{\theta}^n \neq \theta_0\right),$$

where K_n is a function of n belonging to the interval

$$\left[\inf_{\theta_1 \in \Theta \setminus \{\theta_0\}}\left|\theta_1 - \theta_0\right|^2, \sup_{\theta_1 \in \Theta \setminus \{\theta_0\}}\left|\theta_1 - \theta_0\right|^2\right].$$

This means that

$$\begin{aligned} \lim_{n \rightarrow \infty} \frac{1}{n} \ln \text{MSE}\left(\tilde{\theta}^n\right) &= \lim_{n \rightarrow \infty} \frac{1}{n} \ln \mathbb{P}_{\theta_0}\left(\tilde{\theta}^n \neq \theta_0\right), \\ \liminf_{n \rightarrow \infty} \frac{1}{n} \ln \sup_{\theta_0 \in \Theta} \text{MSE}\left(\tilde{\theta}^n\right) &= \liminf_{n \rightarrow \infty} \frac{1}{n} \ln \sup_{\theta_0 \in \Theta} \mathbb{P}_{\theta_0}\left(\tilde{\theta}^n \neq \theta_0\right) \end{aligned}$$

and the lower bounds of Proposition 7 and 8 hold also in this case (see [50], p. 51 for a minimax bound for $\text{MSE}\left(\tilde{\theta}^n\right)$ that, according to our Remark 12, is not tight).

4.2.4. *Lower Bounds for the Weighted Probability of Missclassification.* It is simple to give a lower bound to the weighted probability of missclassification $\mathcal{R}_3\left(\hat{\theta}^n, \theta_0\right) \triangleq$

$\sum_{j=1}^J a_j(\theta_0) \cdot \mathbb{P}_{\theta_0} \left\{ \tilde{\theta}^n = \theta_j \right\}$, using the inequalities:

$$\begin{aligned} \max_{\theta_j \in \Theta} \ln a_j + \max_{\theta_j \in \Theta \setminus \{\theta_0\}} \ln \mathbb{P}_{\theta_0} \left\{ \tilde{\theta}^n = \theta_j \right\} &\geq \ln \mathcal{R}_3 \left(\tilde{\theta}^n, \theta_0 \right) \\ &\geq \min_{\theta_j \in \Theta} \ln a_j + \max_{\theta_j \in \Theta \setminus \{\theta_0\}} \ln \mathbb{P}_{\theta_0} \left\{ \tilde{\theta}^n = \theta_j \right\}, \\ \ln J + \max_{\theta_j \in \Theta \setminus \{\theta_0\}} \ln \mathbb{P}_{\theta_0} \left\{ \tilde{\theta}^n = \theta_j \right\} &\geq \ln \mathcal{R}_2 \left(\tilde{\theta}^n, \theta_0 \right) \\ &\geq \max_{\theta_j \in \Theta \setminus \{\theta_0\}} \ln \mathbb{P}_{\theta_0} \left\{ \tilde{\theta}^n = \theta_j \right\}. \end{aligned}$$

Indeed, these imply:

$$\lim_{n \rightarrow \infty} \frac{1}{n} \ln \mathcal{R}_3 \left(\tilde{\theta}^n, \theta_0 \right) = \lim_{n \rightarrow \infty} \frac{1}{n} \max_{\theta_j \in \Theta \setminus \{\theta_0\}} \ln \mathbb{P}_{\theta_0} \left\{ \tilde{\theta}^n = \theta_j \right\} = \lim_{n \rightarrow \infty} \frac{1}{n} \ln \mathcal{R}_2 \left(\tilde{\theta}^n, \theta_0 \right).$$

The same equality holds also for the \liminf .

4.2.5. Lower Bounds for the Bayes Risk and for \mathbb{P}_e . We consider the Bayes risk under the risk function $\mathcal{R}_2 \left(\tilde{\theta}^n, \theta_0 \right)$ and the prior π as:

$$r_2 \left(\tilde{\theta}^n, \pi \right) \triangleq \sum_{j=0}^J \pi(\theta_j) \cdot \mathcal{R}_2 \left(\tilde{\theta}^n, \theta_j \right) = \sum_{j=0}^J \pi(\theta_j) \cdot \mathbb{P}_{\theta_j} \left(\tilde{\theta}^n \neq \theta_j \right).$$

Under Assumption **Bay**, the lower bound in logarithmic asymptotics can be obtained from the following inequalities:

$$\begin{aligned} \ln \max_{\theta_j \in \Theta} \mathcal{R}_2 \left(\tilde{\theta}^n, \theta_j \right) &\geq \ln r_2 \left(\tilde{\theta}^n, \pi \right), \\ \ln r_2 \left(\tilde{\theta}^n, \pi \right) &\geq \ln \min_{\theta_j \in \Theta} \pi(\theta_j) + \ln \max_{\theta_j \in \Theta} \mathcal{R}_2 \left(\tilde{\theta}^n, \theta_j \right), \end{aligned}$$

leading to

$$(11) \quad \lim_{n \rightarrow \infty} \frac{1}{n} \ln r_2 \left(\tilde{\theta}^n, \pi \right) = \lim_{n \rightarrow \infty} \frac{1}{n} \ln \max_{\theta_0 \in \Theta} \mathcal{R}_2 \left(\tilde{\theta}^n, \theta_0 \right).$$

Then, Proposition 8 holds also for the Bayes risk: clearly this bound is independent of the prior distribution π (provided it is strictly positive, i.e. **Bay** holds) and also holds for the probability of error \mathbb{P}_e . Remark that this inequality can be seen as an asymptotic version of van Trees inequality (see e.g. [45]) for a different risk function.

4.3. Optimality and Efficiency. In this Section, we establish some optimality results for the MLE in discrete parameter models.

The situation is much more intricate than in regular statistical models under the quadratic loss function, in which efficiency coincides with the attainment of the Cramér-Rao lower bound (despite superefficiency). Therefore, we propose the following definition. We denote by $\mathcal{R} = \mathcal{R} \left(\tilde{\theta}^n, \theta_0 \right)$ the risk function of the estimator $\tilde{\theta}^n$ evaluated at θ_0 , and by $\tilde{\Theta}$ a class of estimators.

Definition 1. The estimator $\tilde{\theta}^n$ is *efficient with respect to (wrt) $\tilde{\Theta}$ and wrt \mathcal{R} at θ_0* if:

$$(12) \quad \mathcal{R} \left(\tilde{\theta}^n, \theta_0 \right) \leq \mathcal{R} \left(\tilde{\theta}^n, \theta_0 \right), \quad \forall \tilde{\theta}^n \in \tilde{\Theta}.$$

The estimator $\bar{\theta}^n$ is *minimax efficient wrt $\tilde{\Theta}$ and wrt \mathcal{R}* if:

$$(13) \quad \sup_{\theta_0 \in \Theta} \mathcal{R}(\bar{\theta}^n, \theta_0) \leq \sup_{\theta_0 \in \Theta} \mathcal{R}(\tilde{\theta}^n, \theta_0), \quad \forall \tilde{\theta}^n \in \tilde{\Theta}.$$

The estimator $\bar{\theta}^n$ is *superefficient wrt $\tilde{\Theta}$ and wrt \mathcal{R}* if for every $\tilde{\theta}^n \in \tilde{\Theta}$:

$$\mathcal{R}(\bar{\theta}^n, \theta_0) \leq \mathcal{R}(\tilde{\theta}^n, \theta_0),$$

for every $\theta_0 \in \Theta$ and there exists at least a value $\theta_0^* \in \Theta$ such that the inequality is replaced by a strict inequality for $\theta_0 = \theta_0^*$.

The estimator $\bar{\theta}^n$ is *asymptotically CR-efficient wrt \mathcal{R} at θ_0* if it attains the Chapman-Robbins lower bound of Proposition 7 at θ_0 (say $\text{CR-}\mathcal{R}(\theta_0)$) in the asymptotic form:

$$\liminf_{n \rightarrow \infty} \frac{1}{n} \ln \mathcal{R}(\bar{\theta}^n, \theta_0) = \ln \text{CR-}\mathcal{R}(\theta_0).$$

The estimator $\bar{\theta}^n$ is *asymptotically minimax CR-efficient wrt \mathcal{R}* if it attains the minimax Chapman-Robbins lower bound of Proposition 8 (say $\text{CR-}\mathcal{R}_{\max}$) in the asymptotic form:

$$\liminf_{n \rightarrow \infty} \frac{1}{n} \ln \sup_{\theta_0 \in \Theta} \mathcal{R}(\bar{\theta}^n, \theta_0) = \ln \text{CR-}\mathcal{R}_{\max}.$$

The estimator $\bar{\theta}^n$ is *asymptotically CR-superefficient wrt \mathcal{R}* if:

$$\liminf_{n \rightarrow \infty} \frac{1}{n} \ln \mathcal{R}(\bar{\theta}^n, \theta_0) \leq \ln \text{CR-}\mathcal{R}(\theta_0),$$

for every $\theta_0 \in \Theta$ and there exists at least a value $\theta_0^* \in \Theta$ such that the inequality is replaced by a strict inequality for $\theta_0 = \theta_0^*$.

Remark 13. (i) Efficiency, superefficiency and minimax efficiency wrt $\tilde{\Theta}$ and wrt \mathcal{R} can be defined also asymptotically. As an example, the estimator $\bar{\theta}^n$ is *asymptotically efficient wrt $\tilde{\Theta}$ and wrt \mathcal{R} at θ_0* if:

$$\liminf_{n \rightarrow \infty} \frac{1}{n} \ln \mathcal{R}(\bar{\theta}^n, \theta_0) \leq \liminf_{n \rightarrow \infty} \frac{1}{n} \ln \mathcal{R}(\tilde{\theta}^n, \theta_0), \quad \forall \tilde{\theta}^n \in \tilde{\Theta}.$$

However, since we will not need these concepts, we will limit ourselves to the previous ones.

(ii) As in Remark 11 **(ii)**, it is easy to see that IR-optimality and CR-efficiency wrt \mathcal{R}_2 coincide.

The efficiency landscape offered by discrete parameter models will be illustrated by Example 11. This shows that, even in the simplest case, i.e. the estimation of the integer mean of a Gaussian random variable with known variance, the MLE does not attain the lower bound on the missclassification probability but it attains the minimax lower bound. Moreover, simple estimators are built that outperform the MLE for certain values of the true parameter value θ_0 .

Example 11. Let us consider the estimation of the mean of a Gaussian distribution whose variance σ^2 is known: we suppose that the true mean is α , while the parameter space is $\{-\alpha, \alpha\}$, where α is known.

The maximum likelihood estimator $\hat{\theta}^n$ takes the value $-\alpha$ if the sample mean takes

on its value on $(-\infty, 0)$ and α if it falls in $[0, +\infty)$ (the position of 0 is a convention). Therefore:

$$\begin{aligned}\mathbb{P}_{\theta_0}(\hat{\theta}^n \neq \theta_0) &= \mathbb{P}_{\theta_0}(\hat{\theta}^n = -\alpha) = \int_{-\infty}^0 \frac{e^{-\frac{(\bar{y}-\mu)^2}{2\sigma^2/n}}}{\sqrt{2\pi\sigma^2/n}} d\bar{y} = \int_{-\infty}^{-\frac{\sqrt{n}\mu}{\sigma}} \frac{e^{-\frac{t^2}{2}}}{\sqrt{2\pi}} dt \\ &= \Phi\left(-\frac{\sqrt{n}\mu}{\sigma}\right) = \frac{e^{-\frac{n\alpha^2}{2\sigma^2}} \sigma}{\sqrt{2\pi n} \alpha} \cdot \left(1 + O\left(\frac{1}{n}\right)\right).\end{aligned}$$

Proposition 6 allows also for recovering the right convergence rate. Indeed, we have:

$$\begin{aligned}\Lambda^{(1)}(\lambda) &= -\lambda(1-\lambda) \cdot \frac{2\alpha^2}{\sigma^2}, & \eta &= \frac{1}{2}, \\ (\Lambda^{(1)})'(0) &= -\frac{2\alpha^2}{\sigma^2}, & (\Lambda^{(1)})''(0) &= \frac{2\alpha^2}{\sigma^2};\end{aligned}$$

therefore, we get:

$$\mathbb{P}_{\theta_0}(\hat{\theta}^n \neq \alpha) = \mathbb{P}_{\theta_0}(\hat{\theta}^n = -\alpha) = \frac{e^{-\frac{n\alpha^2}{2\sigma^2}} \sigma}{\sqrt{2\pi n} \alpha} \cdot (1 + o(1)).$$

On the other hand, the lower bound of Proposition 7 yields:

$$\liminf_{n \rightarrow \infty} \frac{1}{n} \ln \mathbb{P}_{\theta_0}(\hat{\theta}^n \neq \theta_0) \geq -\frac{2n \cdot \alpha^2}{\sigma^2},$$

and the lower bound of Proposition 8 yields:

$$\liminf_{n \rightarrow \infty} \frac{1}{n} \sup_{\theta_0 \in \{-\alpha, \alpha\}} \ln \mathbb{P}_{\theta_0}(\hat{\theta}^n \neq \theta_0) \geq -\frac{n \cdot \alpha^2}{2\sigma^2}.$$

Therefore, the MLE attains the minimax lower bound but not the classical one.

In the following, we will show that estimators can be pointwise more efficient than the MLE; consider the estimator defined by:

$$\tilde{\theta}^n = \begin{cases} \theta_0 & \text{if } L_n(\theta_0) \geq L_n(\theta_1) + k \cdot n, \\ \theta_1 & \text{else.} \end{cases}$$

Then, the behavior of the estimator is characterized by the probabilities:

$$\begin{aligned}\mathbb{P}_{\theta_0}(\tilde{\theta}^n = \theta_0) &= \Phi\left(\frac{k \cdot n \cdot \sigma^2 + 2\alpha^2 \cdot n}{2\alpha\sigma\sqrt{n}}\right), \\ \mathbb{P}_{\theta_1}(\tilde{\theta}^n = \theta_0) &= \Phi\left(\frac{k \cdot n \cdot \sigma^2 - 2\alpha^2 \cdot n}{2\alpha\sigma\sqrt{n}}\right).\end{aligned}$$

We have (weak) consistency if:

$$(14) \quad 2\left(\frac{\alpha}{\sigma}\right)^2 > k > -2\left(\frac{\alpha}{\sigma}\right)^2.$$

The risk $\mathcal{R}_2(\tilde{\theta}^n, \theta_0)$ under θ_0 is then:

$$\begin{aligned}\mathbb{P}_{\theta_0}(\tilde{\theta}^n \neq \theta_0) &= \Phi\left[-\frac{k \cdot \sigma^2 + 2\alpha^2}{2\alpha\sigma} \cdot \sqrt{n}\right] \\ &= \frac{e^{-\frac{(k \cdot \sigma^2 + 2\alpha^2)^2 \cdot n}{8\alpha^2\sigma^2}}}{\sqrt{2\pi n}} \cdot \frac{2\alpha\sigma}{k\sigma^2 + 2\alpha^2} \cdot \left(1 + O\left(\frac{1}{n}\right)\right);\end{aligned}$$

since this can simply be made smaller than the probability of error of the MLE,³ this implies that the MLE is not pointwise efficient.

Now, we show that this estimator cannot converge faster than the Chapman-Robbins lower bound without losing its consistency. Indeed, $\mathbb{P}_{\theta_0}(\tilde{\theta}^n \neq \theta_0)$ is smaller than the Chapman-Robbins lower bound if:

$$k^2 + 4k \left(\frac{\alpha}{\sigma}\right)^2 - 12 \left(\frac{\alpha}{\sigma}\right)^4 \geq 0,$$

and this is never true under (14). If this estimator is pointwise more efficient than the MLE under θ_0 , then under θ_1 , this estimator has risk $\mathcal{R}_2(\tilde{\theta}^n, \theta_1)$ given by:

$$\mathbb{P}_{\theta_1}(\tilde{\theta}^n \neq \theta_1) = \Phi \left[\frac{k \cdot \sigma^2 - 2\alpha^2}{2\alpha\sigma} \cdot \sqrt{n} \right] = \frac{e^{-\frac{(k \cdot \sigma^2 - 2\alpha^2)^2 \cdot n}{8\alpha^2 \sigma^2}}}{\sqrt{2\pi n}} \cdot \frac{2\alpha\sigma}{k\sigma^2 + 2\alpha^2} \cdot \left(1 + O\left(\frac{1}{n}\right)\right),$$

and this is greater than for the MLE. This shows that a faster convergence rate can be obtained in some points, the price to pay being a worse convergence rate elsewhere in Θ .

4.3.1. *Optimality wrt classes of estimators.* In the following Section, we show some optimality properties of Bayes and ML estimators. We start with an important and well-known fact.

Proposition 9. *Under Hom, L^1 -Dom, Id and Bay, the Bayes risk $r_2(\tilde{\theta}^n, \pi)$ (under the zero-one loss function) associated with a prior distribution π is strictly minimized by the posterior mode corresponding to the prior π , for any finite n .*

The following Proposition shows that the MLE is admissible and minimax under the zero-one loss and minimizes the average probability of error. Moreover, it shows that no estimator can have a convergence rate faster than the one prescribed by the Chapman-Robbins inequality and that estimators that are more efficient than the MLE at a certain point $\theta_0 \in \Theta$ are less efficient in at least another point $\theta_1 \in \Theta$. Roughly stated, for discrete parameter models, the Chapman-Robbins inequality acts as a pointwise (in θ_0) limit for the performance of estimators but it is not uniform over the parameter space (that is it does not correspond with its minimax version). Estimators can be more efficient than minimax efficient ones only on portions of the parameter space, but are strictly less efficient elsewhere.

Proposition 10. *Under Assumptions Hom, L^1 -Dom and Id, the MLE is admissible and minimax efficient wrt the class of all estimators and wrt \mathcal{R}_2 and minimizes the average probability of error \mathbb{P}_e .*

Remark 14. If we change the loss function, the result does not necessarily transpose. As an example, using the asymptotic equalities of Section 4.2.3 for \mathcal{R}_1 and \mathcal{R}_3 , it is simple to show that with respect to the risk functions $\mathcal{R}_1(\tilde{\theta}^n, \theta_0)$ and $\mathcal{R}_3(\tilde{\theta}^n, \theta_0)$, the estimator is only asymptotically minimax efficient (see [69], for a similar result in a related setup).

³It is enough to take k such that:

$$n \geq \frac{\ln\left(\frac{2\alpha^2}{k\sigma^2 + 2\alpha^2}\right)}{\frac{k^2\sigma^2 + 4k\alpha^2}{8\alpha^2}} + O\left(\frac{1}{n}\right).$$

4.3.2. *Optimality wrt the Information Inequalities.* In this Subsection, we will show that the MLE does not attain the Chapman-Robbins lower bound in the form of Proposition 7 but that it attains the minimax form of Proposition 8 and that efficiency and minimax efficiency are generally incompatible.

Therefore, the situation described in Example 11 is general, for it is possible to show that the MLE is generally inefficient with respect to the lower bounds exposed in Proposition 7: since we have shown in Remark 11 (ii) that IR-optimality and CR-efficiency are equivalent in this setting, we use the method developed in [68] to assess IR-optimality.

Proposition 11. *Under Assumptions **Hom**, L^1 -**Dom** and **Id**:*

- (i) *the MLE is not asymptotically CR-efficient wrt \mathcal{R}_2 at θ_0 ;*
- (ii) *the MLE is asymptotically minimax CR-efficient wrt \mathcal{R}_2 .*
- (iii) *an estimator that is asymptotically CR-efficient wrt \mathcal{R}_2 at θ_0 is not asymptotically minimax CR-efficient wrt \mathcal{R}_2 .*

Remark 15. (i) Indeed, to obtain part (i) of the Proposition, the hypothesis of homogeneity of the probability measures is not even necessary, but it is used in order to derive part (ii): it can be removed along the lines of [68].

(ii) Using the asymptotic equalities of Section 4.2.3 for \mathcal{R}_1 and \mathcal{R}_3 , it is trivial to extend the results of this Proposition to these risk functions.

4.3.3. *The Evil of Superefficiency.* Ever since it was discovered by Hodges, the problem of superefficiency has been dealt with extensively in regular statistical problems (see e.g., [82, 122]). However, these proofs do not transpose to discrete parameter estimation problems, since they are mostly based on the equivalence of prior probability measures with the Lebesgue measure and on properties of Bayes estimators that do not hold in this case. Moreover, the discussion of the previous Sections has shown that, in discrete parameter problems, CR-efficiency and efficiency with respect to a class of estimators do not coincide. The following Proposition yields a solution to the superefficiency problem.

Proposition 12. *Under Assumptions **Hom**, L^1 -**Dom** and **Id**:*

- (i) *no estimator $\tilde{\theta}^n$ is asymptotically CR-superefficient wrt \mathcal{R}_2 at $\theta_0 \in \Theta$;*
- (ii) *no estimator $\tilde{\theta}^n$ is superefficient wrt the MLE and \mathcal{R}_2 .*

Remark 16. The results of the previous Proposition can be extended to the risk functions \mathcal{R}_1 and \mathcal{R}_3 , but in this case they hold only asymptotically.

5. SOME ALTERNATIVE ESTIMATORS

In this Section, we propose a more informal and less general presentation of some alternative estimators that may, in specific situations, have some advantages over the previously described m -estimator based on a finite parameter space. Section 5.1 recalls the asymptotics of standard \sqrt{n} -consistent and asymptotically normal (\sqrt{n} -CAN) estimation theory. Section 5.2 proposes a class of simple estimators obtained looking for the point in Θ that is nearest to a \sqrt{n} -CAN estimator: this estimator has clear computational advantages, but requires smoothness assumptions on the objective function.

At last, Section 5.3 proposes an estimator obtained as a convex linear combination of an estimator under discreteness with an exponential convergence rate and an estimator under continuity with an asymptotic normal distribution: modulating

a parameter, we obtain an estimator with almost the same convergence rate of the estimator under discreteness and an asymptotically normally distributed estimator.

5.1. MLE and Other Estimators under Continuity. The idea is to consider an extended statistical model $\mathcal{P} = \{\mathbb{P}_\theta, \theta \in \Theta'\}$ where Θ' is a set with nonempty interior containing the finite set Θ : clearly, this method works only when the density function $f = (d\mathbb{P}/d\mu) : \Theta \rightarrow \mathbb{R}$ can be extended to a function defined on Θ' . The conditions under which estimators obtained with a continuous parameter space are \sqrt{n} -CAN are well-known: we remark that they requires much stronger continuity and differentiability Assumptions than the ones stated here. We address the interested reader to [93] or [46] (Vol. 2, Ch. 24).

5.2. Discretized Continuous Estimators. In this Section, we consider a large family of estimators obtained improving \sqrt{n} -CAN estimators: the implementability of such a method depends crucially on the remarks of the previous Section. In what follows, we suppose that the set $\Theta = \{\theta_0, \dots, \theta_J\}$ is a subset of the real line: this can be partly generalized, but the topic will not be treated in this paper.

Let $\check{\theta}^n$ be a \sqrt{n} -CAN estimator of the form

$$\sqrt{n}(\check{\theta}^n - \theta_0) \xrightarrow{\mathcal{D}} \mathcal{N}(0, I_0);$$

clearly any maximum likelihood, method of moments or generalized method of moments estimator can serve the purpose. We define a discretized estimator as:

$$\bar{\theta}^n \triangleq \arg \min_{\theta_i \in \Theta} \|\check{\theta}^n - \theta_i\|^2;$$

this is a *classical minimum distance (CMD) estimator* (see [93], p. 2116), in the sense that it minimizes the distance between the reduced form parameter estimator, say $\check{\theta}^n$, and the structural parameter $\theta_i \in \Theta$. Therefore:

$$\begin{aligned} \mathbb{P}_{\theta_0}(\bar{\theta}^n = \theta_i) &\triangleq \mathbb{P}_{\theta_0} \left(\|\check{\theta}^n - \theta_i\|^2 < \|\check{\theta}^n - \theta_j\|^2, \forall \theta_j \in \Theta \setminus \{\theta_i\} \right) \\ &= \mathbb{P}_{\theta_0} \left(\frac{(\theta_i + \max_{\theta_j < \theta_i} \theta_j)}{2} < \check{\theta}^n < \frac{(\theta_i + \min_{\theta_j > \theta_i} \theta_j)}{2}, \forall \theta_j \in \Theta \setminus \{\theta_i\} \right) \\ &= \left[\Phi \left(\sqrt{n/I_0} \left(\frac{\theta_i + \min_{\theta_j > \theta_i} \theta_j - 2\theta_0}{2} \right) \right) - \Phi \left(\sqrt{n/I_0} \left(\frac{\theta_i + \max_{\theta_j < \theta_i} \theta_j - 2\theta_0}{2} \right) \right) \right] \cdot (1 + o(1)), \end{aligned}$$

where in the case of the continuous MLE, I_0 is the Fisher information, otherwise we expect it to have an Eicker-White sandwich form (see [125], p. 5). Using the asymptotic approximation $\Phi(-t) \sim (1 + O(t^{-2})) \cdot \phi(t)/t$, the probability of missclassification can be obtained as:

$$\begin{aligned} \mathbb{P}_{\theta_0}(\bar{\theta}^n \neq \theta_0) &= 1 - \mathbb{P}_{\theta_0}(\bar{\theta}^n = \theta_0) \\ &= \left[\Phi \left(\sqrt{n/I_0} \left(\frac{\theta_0 - \min_{\theta_j > \theta_0} \theta_j}{2} \right) \right) + \Phi \left(\sqrt{n/I_0} \left(\frac{\max_{\theta_j < \theta_0} \theta_j - \theta_0}{2} \right) \right) \right] \cdot (1 + o(1)) \\ &= \sqrt{\frac{2I_0}{\pi n}} \cdot \left[\frac{e^{-\frac{n}{8I_0}(\theta_0 - \min_{\theta_j > \theta_0} \theta_j)^2}}{(\min_{\theta_j > \theta_0} \theta_j - \theta_0)} + \frac{e^{-\frac{n}{8I_0}(\max_{\theta_j < \theta_0} \theta_j - \theta_0)^2}}{(\theta_0 - \max_{\theta_j < \theta_0} \theta_j)} \right] \cdot (1 + o(1)). \end{aligned}$$

This estimator can be generalized to the case of a distance $d(\cdot, \cdot)$ (instead of the norm $\|\cdot\|$), such that, for any $\theta_j \in \Theta \setminus \{\theta_0\}$ and for any θ^* in an open neighborhood

of θ_0 , the following Taylor limited development holds:

$$d(\theta^*, \theta_j) = d(\theta_0, \theta_j) + \frac{\partial d(\theta_0, \theta_j)}{\partial \theta_0} \cdot (\theta^* - \theta_0) + o(\|\theta^* - \theta_0\|).$$

In what follows, we write $d_j \triangleq d(\theta_0, \theta_j)$ and $d'_j \triangleq \partial d(\theta_0, \theta_j) / \partial \theta_0$ for short. In this case:

$$\bar{\theta}^n \triangleq \arg \min_{\theta_i \in \Theta} d(\check{\theta}^n, \theta_i);$$

therefore:

$$\begin{aligned} \mathbb{P}_{\theta_0}(\bar{\theta}^n = \theta_i) &\triangleq \mathbb{P}_{\theta_0} \left(d(\check{\theta}^n, \theta_i) < d(\check{\theta}^n, \theta_j), \forall \theta_j \in \Theta \setminus \{\theta_i\} \right) \\ &= \mathbb{P}_{\theta_0} \left(d_i + d'_i \cdot (\check{\theta}^n - \theta_0) < d_j + d'_j \cdot (\check{\theta}^n - \theta_0) + o_{\mathbb{P}}(n^{-1/2}), \forall \theta_j \in \Theta \setminus \{\theta_i\} \right) \\ &= \mathbb{P}_{\theta_0} \left((d'_i - d'_j) \cdot (\check{\theta}^n - \theta_0) < (d_j - d_i) + o_{\mathbb{P}}(n^{-1/2}), \forall \theta_j \in \Theta \setminus \{\theta_i\} \right) \\ &= \mathbb{P}_{\theta_0} \left(\max_{\{j|d'_i < d'_j\}} \frac{d_j - d_i}{d'_i - d'_j} < \check{\theta}^n - \theta_0 + o_{\mathbb{P}}(n^{-1/2}) < \min_{\{j|d'_i > d'_j\}} \frac{d_j - d_i}{d'_i - d'_j} \right) \\ &= \left[\Phi \left(\sqrt{n/I_0} \min_{\{j|d'_i > d'_j\}} \frac{d_j - d_i}{d'_i - d'_j} \right) - \Phi \left(\sqrt{n/I_0} \max_{\{j|d'_i < d'_j\}} \frac{d_j - d_i}{d'_i - d'_j} \right) \right] \cdot (1 + o(1)). \end{aligned}$$

The probability of missclassification is:

$$\begin{aligned} \mathbb{P}_{\theta_0}(\bar{\theta}^n \neq \theta_0) &= 1 - \mathbb{P}_{\theta_0}(\bar{\theta}^n = \theta_0) \\ &= \left[\Phi \left(\sqrt{n/I_0} \max_{\{j|d'_0 > d'_j\}} \frac{-d_j}{d'_0 - d'_j} \right) + \Phi \left(\sqrt{n/I_0} \max_{\{j|d'_0 < d'_j\}} \frac{d_j}{d'_0 - d'_j} \right) \right] \cdot (1 + o(1)) \\ &= \sqrt{\frac{I_0}{2\pi n}} \cdot \left[\frac{e^{-\frac{n}{2I_0} \min_{\{j|d'_0 > d'_j\}} \left(\frac{d_j}{d'_0 - d'_j} \right)^2}}{\min_{\{j|d'_0 > d'_j\}} \frac{d_j}{d'_0 - d'_j}} - \frac{e^{-\frac{n}{2I_0} \min_{\{j|d'_0 < d'_j\}} \left(\frac{d_j}{d'_j - d'_0} \right)^2}}{\min_{\{j|d'_0 < d'_j\}} \frac{d_j}{d'_j - d'_0}} \right] \cdot (1 + o(1)). \end{aligned}$$

5.3. Estimators Obtained by Linear Convex Combinations. In this Section, we propose an alternative class of estimators of θ_0 ; they have the characteristic of being asymptotically normal, without losing the exponential convergence rate of m -estimator.

We consider a new estimator formed by a linear convex combination of an m -estimator of θ under discreteness of Θ , say $\hat{\theta}^{n,D}$, and of the continuous one, say $\hat{\theta}^{n,C}$:

$$\hat{\theta}^n = (1 - \lambda) \cdot \hat{\theta}^{n,D} + \lambda \cdot \hat{\theta}^{n,C}.$$

First of all, we consider the behavior of the MSE of this new estimator:

$$\begin{aligned} \mathbb{E} \left[\hat{\theta}^n - \theta_0 \right]^2 &= (1 - \lambda)^2 \cdot \mathbb{E} \left(\hat{\theta}^{n,D} - \theta_0 \right)^2 + \lambda^2 \cdot \mathbb{E} \left(\hat{\theta}^{n,C} - \theta_0 \right)^2 \\ &\quad + 2(1 - \lambda)\lambda \cdot \mathbb{E} \left[\left(\hat{\theta}^{n,D} - \theta_0 \right) \left(\hat{\theta}^{n,C} - \hat{\theta}^{n,D} \right) \right] \\ &\leq \left\{ \sqrt{\mathbb{E} \left(\hat{\theta}^{n,D} - \theta_0 \right)^2} + \lambda \left[\sqrt{\mathbb{E} \left(\hat{\theta}^{n,C} - \theta_0 \right)^2} - \sqrt{\mathbb{E} \left(\hat{\theta}^{n,D} - \theta_0 \right)^2} \right] \right\}^2; \end{aligned}$$

from the previous results (see also [109]), it is clear that there exist $C_1 > 0$, $\rho \in]0, 1[$ and $R > 0$ such that:

$$\begin{aligned}\mathbb{E} \left[\hat{\theta}^{n,D} - \theta_0 \right]^2 &= C_1^2 \cdot \rho^{2n} \cdot (1 + o(1)), \\ \mathbb{E} \left[\hat{\theta}^{n,C} - \theta_0 \right]^2 &= \frac{C_2^2}{n} \cdot (1 + o(1));\end{aligned}$$

therefore:

$$\mathbb{E} \left[\hat{\theta}^n - \theta_0 \right]^2 \leq \left\{ C_1 \cdot \rho^n + \lambda \left[\frac{C_2}{\sqrt{n}} - C_1 \cdot \rho^n \right] \right\}^2 \cdot (1 + o(1)) = \frac{\lambda^2 \cdot C_2^2}{n} \cdot (1 + o(1)).$$

If we choose a value of λ depending on n , say λ_n , we can increase the speed of convergence of $\mathbb{E} \left[\hat{\theta}^n - \theta_0 \right]^2$ to 0 as close as possible to ρ^n even if $\hat{\theta}^n$ keeps a Gaussian asymptotic distribution. If we take $\lambda_n \triangleq \frac{\sqrt{n}}{o(\rho^{-n})}$, we get:

$$\begin{aligned}\mathbb{E} \left[\hat{\theta}^n - \theta_0 \right]^2 &\leq \left\{ C_1 \cdot \rho^n + \lambda \left[\frac{C_2}{\sqrt{n}} - C_1 \cdot \rho^n \right] \right\}^2 \cdot (1 + o(1)) \\ &\sim \left\{ \frac{1}{o(\rho^{-n})} - \frac{\sqrt{n}}{o(\rho^{-2n})} \right\}^2 \sim \frac{1}{o(\rho^{-2n})}, \\ \frac{\sqrt{n}}{\lambda_n} \left(\hat{\theta}^n - \theta_0 \right) &= \frac{\sqrt{n}}{\lambda_n} \left[(1 - \lambda_n) \cdot \left(\hat{\theta}^{n,D} - \theta_0 \right) + \lambda_n \cdot \left(\hat{\theta}^{n,C} - \theta_0 \right) \right] \\ &\sim \sqrt{n} \cdot \left(\hat{\theta}^{n,C} - \theta_0 \right) + o_{\mathbb{P}}(1).\end{aligned}$$

$\hat{\theta}^n$ is clearly consistent and the normalized estimator $\frac{\sqrt{n}}{\lambda_n} \left(\hat{\theta}^n - \theta_0 \right)$ is asymptotically distributed as a normal random variable.

Therefore, a good candidate for λ_n is given by a function like $\lambda_n = \frac{\sqrt{n}}{f(n)}$ where $f(n) = o(\rho^{-n})$, $(f(n))^{-1} = o\left(\frac{1}{\sqrt{n}}\right)$ and f is a decreasing function of n : for instance, we could take $f(n) = n^K$ for $K > 1/2$ or $f(n) = \kappa^{-n}$ for $\kappa > \rho$. Remark that the latter form requires the knowledge of ρ while the former does not. As an example, taking $\lambda_n = C_3 \cdot \sqrt{n} \cdot (\rho + \varepsilon)^n$ for any $\varepsilon > 0$ such that $(\rho + \varepsilon) \in]0, 1[$, we get:

$$\begin{aligned}\mathbb{E} \left[\hat{\theta}^n - \theta_0 \right]^2 &\leq (C_2 \cdot C_3)^2 \cdot (\rho + \varepsilon)^{2n} \cdot (1 + o(1)), \\ (\rho + \varepsilon)^{-n} \left(\hat{\theta}^n - \theta_0 \right) &= C_3 \cdot \sqrt{n} \cdot \left(\hat{\theta}^{n,C} - \theta_0 \right) + O_{\mathbb{P}} \left[\left(\frac{\rho}{\rho + \varepsilon} \right)^n \right].\end{aligned}$$

6. PROOFS

Proof of Proposition 1. Under **Obj**, Kolmogorov's SLLN implies that \mathbb{P}_{θ^*} – as:

$$\frac{1}{n} \sum_{i=1}^n \ln q(Y_i; \theta_j) \rightarrow \mathbb{E}_{\theta^*} \ln q(Y; \theta_j),$$

and for \mathbb{P}_{θ^*} – as any sequence of realizations, $\hat{\theta}^n$ converges to θ_0 . Measurability follows from the fact that the following set belongs to $\mathcal{Y}^{\otimes n}$:

$$\left\{ \omega \in \Omega \left| \sup_{\theta \in \Theta} \frac{1}{n} \sum_{i=1}^n \ln q(y_i; \theta) \leq t \right. \right\} = \bigcap_{\theta_j \in \Theta} \left\{ \omega \in \Omega \left| \frac{1}{n} \sum_{i=1}^n \ln q(y_i; \theta_j) \leq t \right. \right\}.$$

Proof of Lemma 1. Clearly **(ii)** implies η -**Int** for a certain $\eta > 0$. On the other hand, suppose that η -**Int** holds; then, applying recursively Hölder inequality:

$$\Lambda^{(i)}(\lambda) \triangleq \ln \mathbb{E}_{\theta^*} \left[\prod_{j=0, \dots, J, j \neq i} \left(\frac{q(Y; \theta_i)}{q(Y; \theta_j)} \right)^{\lambda_j} \right] \leq \sum_{j=0, \dots, J, j \neq i} \frac{1}{J} \cdot \ln \mathbb{E}_{\theta^*} \left[\left(\frac{q(Y; \theta_i)}{q(Y; \theta_j)} \right)^{J \cdot \lambda_j} \right]$$

and choosing the λ_j 's adequately, we get **(ii)**.

Proof of Proposition 2. The first two results are straightforward applications of Cramér's Theorem in \mathbb{R}^d (see e.g. [32], Corollary 6.1.6, p. 253). Indeed, it is known that the lower bound holds without any supplementary assumption, while the upper bound requires a Cramér's $\frac{1}{2}$ condition $\mathbf{0} \in \text{int}(\mathcal{D}_{\Lambda^{(i)}})$: indeed, from Lemma 1, this is equivalent to Assumption η -**Int**. Then, a full LDP holds:

$$\begin{aligned} \liminf_{n \rightarrow \infty} \frac{1}{n} \ln \mathbb{P}_{\theta^*}(\hat{\theta}^n = \theta_i) &\geq - \inf_{\mathbf{y} \in \text{int} \mathbb{R}_+^J} \sup_{\lambda \in \mathbb{R}^J} \left\{ \langle \mathbf{y}, \lambda \rangle - \Lambda^{(i)}(\lambda) \right\}, \\ \limsup_{n \rightarrow \infty} \frac{1}{n} \ln \mathbb{P}_{\theta^*}(\hat{\theta}^n = \theta_i) &\leq - \inf_{\mathbf{y} \in \mathbb{R}_+^J} \sup_{\lambda \in \mathbb{R}^J} \left\{ \langle \mathbf{y}, \lambda \rangle - \Lambda^{(i)}(\lambda) \right\}. \end{aligned}$$

In order to prove the final result, we have to show that \mathbb{R}_+^J is a $\Lambda^{(i),*}$ -continuity set, that is $\inf_{\mathbf{y} \in \text{int} \mathbb{R}_+^J} \Lambda^{(i),*}(\mathbf{y}) = \inf_{\mathbf{y} \in \mathbb{R}_+^J} \Lambda^{(i),*}(\mathbf{y})$. It is enough to apply part (ii) in Lemma on page 903 of [95].

Proof of Proposition 3. First of all, we remark that $\mathbb{P}_{\theta^*}(\hat{\theta}^n \neq \theta_0) = \mathbb{P}_{\theta^*}(\sum_{k=1}^n \mathbf{X}_k \in \text{int}(\mathbb{R}_+^J)^c)$. Therefore, we can apply large deviations principles, with the candidate rate function $\Lambda^*(\mathbf{y})$; this is a strictly convex function on $\text{int} \mathcal{D}_{\Lambda^*}$ globally minimized at

$$\mathbf{y}' = [\mathbb{E}_{\theta^*}(\ln f_Y(Y; \theta_0) - \ln f_Y(Y; \theta_j))]_{j=1, \dots, J}.$$

By Assumption **Obj**, \mathbf{y}' is finite and belongs to $\text{int} \mathbb{R}_+^J$. From the strict convexity of the level sets of $\Lambda^*(\mathbf{y})$, the set $\arg \inf_{\mathbf{y} \in \text{int}(\mathbb{R}_+^J)^c} \Lambda^*(\mathbf{y})$ has at most finite cardinality H . Moreover, since large deviations theory allows us to ignore the part of $\text{int}(\mathbb{R}_+^J)^c$ where $\Lambda^*(\mathbf{y}) \geq \varepsilon + \inf_{\mathbf{y} \in \text{int}(\mathbb{R}_+^J)^c} \Lambda^*(\mathbf{y})$, we can replace $(\mathbb{R}_+^J)^c$ with a collection of H disjoint sets, say Γ_h , $h = 1, \dots, H$, each of them containing in its interior one and only one of the points of $\arg \inf_{\mathbf{y} \in \text{int}(\mathbb{R}_+^J)^c} \Lambda^*(\mathbf{y})$ (see [62], p. 508):

$$\begin{aligned} (15) \quad \mathbb{P}_{\theta^*} \left(\sum_{k=1}^n \mathbf{X}_k \in \text{int}(\mathbb{R}_+^J)^c \right) &= (1 + o(1)) \cdot \mathbb{P}_{\theta^*} \left(\sum_{k=1}^n \mathbf{X}_k \in \text{int} \bigcup_{h=1}^H \Gamma_h \right) \\ &= (1 + o(1)) \cdot \sum_{h=1}^H \mathbb{P}_{\theta^*} \left(\sum_{k=1}^n \mathbf{X}_k \in \text{int} \Gamma_h \right). \end{aligned}$$

As before, the bounds derive from Cramér's $\frac{1}{2}$'s Theorem in \mathbb{R}^d . Remarking that the contribution of any Γ_h is the same and recalling (15), we get the results.

Proof of Proposition 4. The hypotheses of the Theorem on page 904 of [95] are easily verified. This shows that a unique dominating point $\mathbf{y}^{(i)}$ exists and implies, through Proposition on page 161 of [94] (according to the "Remarks on the hypotheses" in [95], p. 905, the "lattice" conditions are not necessary), that the stated bracketing of $\mathbb{P}_{\theta^*}(\hat{\theta}^n = \theta_i)$ holds.

Proof of Proposition 5. The results used in the following require Assumptions **A1-4** from [62]; in particular, his Assumptions **A1** coincides with L^r -**Sum**-(i).

A3 derives from Proposition 4 under Assumptions **Obj**, η -**Int**, **Steep**–(i) and **Meet**–(i). Remark that fulfillment of **A3** entails that also **A2** is satisfied (see [62], p. 508). When needed, Assumption **A4**, corresponding to a property of $\mathbf{H}_g(\mathbf{0})$ and $\mathbf{H}_f(\mathbf{0})$, is stated explicitly in the Proposition.

(I) $\mathbf{y}^{(i)} = \mathbf{0}$ and $\lambda^{(i)} \in \text{int } \mathbb{R}_+^J$: in particular, in the notation of Theorem 1.5 of [62], Γ is contained in a cone with $\beta = 1$ and

$$C^* = \sqrt{\max_j (\lambda_j^{(i)})^2 / \left[\|\lambda^{(i)}\|^2 - \max_j (\lambda_j^{(i)})^2 \right]}$$

and contains the cone with $\beta = 1$ and

$$C^{**} = \sqrt{\left[\|\lambda^{(i)}\|^2 - \min_j (\lambda_j^{(i)})^2 \right] / \min_j (\lambda_j^{(i)})^2}$$

(remark that $C^{**} > C^*$). This yields the results.

(II) $\mathbf{y}^{(i)}$ belongs to the interior of a face of \mathbb{R}_+^J : in this case, the formula derives from [94] (p. 165, CASE (i)). If, moreover, the Hessian $\mathbf{H}_f(\mathbf{0})$ is negative definite (that is \mathbb{R}_+^J is not flush with the level set of $\Lambda^{(i),*}$) we use Theorem 1.4 of [62].

Proof of Proposition 6. Under Assumptions **Obj**, η -**Int**, **Meet**–(i) and **Steep**–(i), according to Proposition 2 (iii) we have $\mathbb{P}_0 \{Q_n(\theta_1) \geq Q_n(\theta_0)\} = \mathbb{P}_0 \{Q_n(\theta_1) > Q_n(\theta_0)\} \cdot (1 + o(1))$ and we can study the behavior of:

$$\begin{aligned} \mathbb{P}_0(\hat{\theta}^n \neq \theta_0) &= \mathbb{P}_0(\hat{\theta}^n = \theta_1) = \mathbb{P}_0 \{Q_n(\theta_1) \geq Q_n(\theta_0)\} \\ &= \mathbb{P}_0 \{Q_n(\theta_1) - Q_n(\theta_0) \in [0, +\infty)\}. \end{aligned}$$

Min implies that the conditions of Theorem 3.7.4 in [32] (p. 110) are verified, in particular the existence of a positive $\mu \in \text{int } (\mathcal{D}_{\Lambda^{(1)}})$ solution to the equation $0 = (\Lambda^{(1)})'(\mu)$. From Lemma 2.2.5 (c) in [32], this implies $\Lambda^{(1)}(\mu) = -\Lambda^{(1),*}(0)$, and the result follows.

Proof of Theorem 1. We remark that the function $\kappa(\cdot)$ in [65] (p. 1117) is given by:

$$\begin{aligned} \kappa(\mathbf{u}) &= \ln \mathbb{E}_0 \exp \left[\mathbf{u} \cdot \left(\mathbf{X}^{(i)} - \mathbb{E}_0 \mathbf{X}^{(i)} \right) \right] \\ &= \ln \mathbb{E}_0 \exp \left[\mathbf{u} \cdot \mathbf{X}^{(i)} \right] - \mathbf{u} \cdot \mathbb{E}_0 \mathbf{X}^{(i)} \\ &= \Lambda^{(i)}(\mathbf{u}) - \mathbf{u} \cdot \mathbb{E}_0 \mathbf{X}^{(i)}. \end{aligned}$$

Therefore, we write the mean $\mathbf{m}(\mathbf{u})$ and covariance matrix $\mathbf{V}(\mathbf{u})$ as:

$$\begin{aligned} \mathbf{m}(\mathbf{u}) &= \kappa'(\mathbf{u}) = \frac{\partial \kappa(\mathbf{u})}{\partial \mathbf{u}} = \frac{\partial \Lambda^{(i)}(\mathbf{u})}{\partial \mathbf{u}} - \mathbb{E}_0 \mathbf{X}^{(i)}, \\ \mathbf{V}(\mathbf{u}) &= \kappa''(\mathbf{u}) = \frac{\partial^2 \kappa(\mathbf{u})}{\partial \mathbf{u}^2} = \frac{\partial^2 \Lambda^{(i)}(\mathbf{u})}{\partial \mathbf{u}^2}. \end{aligned}$$

From (2), we have:

$$\begin{aligned} \mathbb{P}_0(\hat{\theta}^n = \theta_i) &= \mathbb{P}_0 \left(\sum_{k=1}^n \mathbf{X}_k^{(i)} \in \text{int } (\mathbb{R}_+^J) \right) \\ &= \mathbb{P}_0 \left\{ \frac{1}{n} \cdot \sum_{k=1}^n \left(\mathbf{X}_k^{(i)} - \mathbb{E}_0 \mathbf{X}^{(i)} \right) \in \text{int } (\mathbb{R}_+^J) \oplus \mathbb{E}_0 \mathbf{X}^{(i)} \right\}. \end{aligned}$$

Hypothesis s –**TiltCF**–(i) implies (S.1)–(S.4) of [65] (see e.g. [104], p. 735). Since $\mathbb{E}_0 \mathbf{X}^{(i)}$ is strictly negative by **Obj**, $\text{int } \mathbb{R}_+^J \ominus \mathbb{E}_0 \mathbf{X}^{(i)}$ does not contain $\mathbf{0}$ and, according to Theorem 1 in [65] (p. 1118), the result of the Theorem follows.

Proof of Proposition 7. First of all, we prove (7). We suppose that

$$\int \ln \frac{f_Y(y; \theta_1)}{f_Y(y; \theta_0)} f_Y(y; \theta_1) \mu(dy) < \infty,$$

otherwise the inequality is trivial. Then, for any $\theta_1 \in \Theta \setminus \{\theta_0\}$, we apply Lemma 3.4.7 in [32] (p. 94) with $\alpha_n = \mathbb{P}_{\theta_1} \{\hat{\theta}^n \neq \theta_1\}$ and $\beta_n = \mathbb{P}_{\theta_0} \{\hat{\theta}^n \neq \theta_0\}$; since $\hat{\theta}^n$ is strongly consistent, α_n is ultimately less than any $\varepsilon > 0$ and the bound holds.

The second part can be proved as follows. Define the sets:

$$\begin{aligned} A_n(j) &= \{\omega : \tilde{\theta}^n = \theta_j\} \\ B_n(j) &= \left\{ \omega : \frac{1}{n} \ln \left(\frac{f_Y(Y; \theta_j)}{f_Y(Y; \theta_0)} \right) \leq \mathbb{E}_{\theta_j} \ln \left(\frac{f_Y(Y; \theta_j)}{f_Y(Y; \theta_0)} \right) + \varepsilon \right\} \end{aligned}$$

Therefore, we have:

$$\begin{aligned} \mathbb{P}_{\theta_0} \{\tilde{\theta}^n \neq \theta_0\} &= \mathbb{E}_{\theta_0} \mathbf{1}\{\tilde{\theta}^n \neq \theta_0\} = \mathbb{E}_{\theta_j} \frac{f_Y(Y; \theta_0)}{f_Y(Y; \theta_j)} \mathbf{1}\{\tilde{\theta}^n \neq \theta_0\} \\ &\geq \mathbb{E}_{\theta_j} \frac{f_Y(Y; \theta_0)}{f_Y(Y; \theta_j)} \mathbf{1}\{A_n(j)\} \\ &\geq \mathbb{E}_{\theta_j} \mathbf{1}\{A_n(j)\} \mathbf{1}\{B_n(j)\} \cdot \exp \left\{ -n \cdot \left[\mathbb{E}_{\theta_j} \ln \left(\frac{f_Y(Y; \theta_j)}{f_Y(Y; \theta_0)} \right) + \varepsilon \right] \right\} \\ &\geq \left[1 - \mathbb{P}_{\theta_j} \{A_n^c(j)\} - \mathbb{P}_{\theta_j} \{B_n^c(j)\} \right] \cdot \exp \left\{ -n \cdot \left[\mathbb{E}_{\theta_j} \ln \left(\frac{f_Y(Y; \theta_j)}{f_Y(Y; \theta_0)} \right) + \varepsilon \right] \right\} \\ &\geq \left[1 - \mathbb{P}_{\theta_j} \{\tilde{\theta}^n \neq \theta_j\} - \mathbb{P}_{\theta_j} \{B_n^c(j)\} \right] \cdot \exp \left\{ -n \cdot \left[\mathbb{E}_{\theta_j} \ln \left(\frac{f_Y(Y; \theta_j)}{f_Y(Y; \theta_0)} \right) + \varepsilon \right] \right\}. \end{aligned}$$

This implies:

$$\begin{aligned} \liminf_{n \rightarrow \infty} \frac{1}{n} \ln \mathbb{P}_{\theta_0} \{\tilde{\theta}^n \neq \theta_0\} &\geq -\mathbb{E}_{\theta_j} \ln \left(\frac{f_Y(Y; \theta_j)}{f_Y(Y; \theta_0)} \right) - \varepsilon \\ &\quad + \liminf_{n \rightarrow \infty} \frac{1}{n} \ln \left[1 - \mathbb{P}_{\theta_j} \{\tilde{\theta}^n \neq \theta_j\} - \mathbb{P}_{\theta_j} \{B_n^c(j)\} \right]. \end{aligned}$$

Now, since $\lim_{n \rightarrow \infty} \mathbb{P}_{\theta_j} \{B_n^c(j)\} = 0$ and $\limsup_{n \rightarrow \infty} \mathbb{P}_{\theta_j} \{\tilde{\theta}^n \neq \theta_j\} < 1$, the third term in the right hand side goes to zero; since ε is arbitrary, the result follows.

Proof of Proposition 8. From the Neyman-Pearson Lemma, we have:

$$\begin{aligned} \sup_{\theta_0 \in \Theta} \mathbb{P}_{\theta_0} (\tilde{\theta}^n \neq \theta_0) &\geq \max \left\{ \mathbb{P}_{\theta_0} (\tilde{\theta}^n \neq \theta_0), \mathbb{P}_{\theta_1} (\tilde{\theta}^n \neq \theta_1) \right\} \\ &\geq \frac{1}{2} \cdot \left\{ \mathbb{P}_{\theta_0} (\tilde{\theta}^n \neq \theta_0) + \mathbb{P}_{\theta_1} (\tilde{\theta}^n \neq \theta_1) \right\} \\ &\geq \frac{1}{2} \cdot \left\{ \mathbb{P}_{\theta_0} \left(\frac{L_n(\theta_0)}{L_n(\theta_1)} < 1 \right) + \mathbb{P}_{\theta_1} \left(\frac{L_n(\theta_0)}{L_n(\theta_1)} \geq 1 \right) \right\}, \end{aligned}$$

for an arbitrary couple of different alternatives θ_0 and θ_1 in Θ . Then we can use Chernoff's Bound ([32], p. 93): the second expression derives from the equality $\Lambda^*(0) = -\inf_{\lambda \in \mathbb{R}} \Lambda(\lambda)$.

Proof of Proposition 10. In order to prove that the MLE is admissible and minimax we use the Bayesian method. Using the prior densities given by $\pi(\theta_k) = (J+1)^{-1}$, the Bayes estimator relative to zero-one loss $\check{\theta}^n$ coincides with the MLE $\hat{\theta}^n$. Therefore, respectively from Lemma 2.10 and Proposition 6.3 in [103], $\hat{\theta}^n$ is minimax and admissible (see also Property 2.2 in [46], Vol. 1, p. 60). The fact that the MLE minimizes the average probability of error derives from Proposition 9.

Proof of Proposition 11. (i) In order to prove the first statement, we apply Lemma 2.4 in [68] (p. 653). Clearly \mathcal{P} is closed in total variation, since it is finite, and is not exponentially convex; indeed, under hypothesis **Hom**, there exist $\theta_1, \theta_2 \in \Theta$ and $\alpha \in [0, 1]$, such that the probability measure $\mathbb{P}_{\theta(\alpha)}$ defined as

$$\mathbb{P}_{\theta(\alpha)}(dx) = \frac{(f_{\theta_1}(x))^\alpha \cdot (f_{\theta_2}(x))^{1-\alpha}}{\int (f_{\theta_1}(x))^\alpha \cdot (f_{\theta_2}(x))^{1-\alpha} \cdot \mu(dx)} \mu(dx)$$

does not belong to \mathcal{P} . Therefore, from Lemma 2.4 (iii) in [68], there exists $\theta'_1, \theta'_2 \in \Theta$ such that equation (2.12) in [68] holds and, as a consequence of Lemma 2.4 (i) in [68], the MLE fails to be an inaccuracy rate optimal estimator at least at one of the points θ'_1, θ'_2 . This means that, say for θ'_1 :

$$\liminf_{n \rightarrow \infty} \frac{1}{n} \ln \mathbb{P}_{\theta'_1} \left\{ \left| \hat{\theta}^n - \theta'_1 \right| > \varepsilon \right\} > \sup_{\theta \in \Theta, |\theta - \theta'_1| > \varepsilon} \mathbb{E}_\theta \ln \left(\frac{f_Y(Y; \theta'_1)}{f_Y(Y; \theta)} \right),$$

and this implies that the Chapman-Robbins bound is not attained at θ'_1 .

(ii) The second statement follows easily from the results of [66] (Theorem 2) on $\lim_{n \rightarrow \infty} \frac{1}{n} \ln r_2(\hat{\theta}^n, \pi)$, using equation (11). Indeed, the MLE attains the lower bound (9) and is therefore asymptotically minimax efficient.

(iii) If the estimator is asymptotically CR-efficient wrt \mathcal{R}_2 at θ_0 , this means that at θ_0 it is more efficient than the MLE and therefore it has to be less efficient elsewhere (since from Proposition 10 the MLE minimizes the probability of error). Therefore, it cannot be minimax CR-efficient.

Proof of Proposition 12. (i) It is enough to follow the proof of Proposition 7 and to reason by contradiction.

(ii) This is simply another way of stating Proposition 10.

REFERENCES

- [1] J. Aczél and Z. Daróczy. *On measures of information and their characterizations*. Academic Press [Harcourt Brace Jovanovich Publishers], New York, 1975.
- [2] P.H. Algoet and T.M. Cover. A sandwich proof of the Shannon-McMillan-Breiman theorem. *Ann. Probab.*, 16(2):899–909, 1988.
- [3] S.-I. Amari. *Differential-geometrical methods in statistics*, volume 28 of *Lecture Notes in Statistics*. Springer-Verlag, New York, 1985.
- [4] P.K. Andersen and R.D. Gill. Cox's regression model for counting processes: a large sample study. *Ann. Statist.*, 10(4):1100–1120, 1982.
- [5] R.R. Bahadur. On the asymptotic efficiency of tests and estimates. *Sankhyā*, 22:229–252, 1960.
- [6] R.R. Bahadur. Rates of convergence of estimates and test statistics. *Ann. Math. Statist.*, 38:303–324, 1967.
- [7] R.R. Bahadur and R. Ranga Rao. On deviations of the sample mean. *Ann. Math. Statist.*, 31:1015–1027, 1960.
- [8] Y. Baram. A sufficient condition for consistent discrimination between stationary gaussian models. *IEEE Trans. Automatic Control*, 23(5):958–960, 1978.

- [9] Y. Baram and N.R. Sandell, Jr. An information theoretic approach to dynamical systems modeling and identification. In *Proceedings of the 1977 IEEE Conference on Decision and Control (New Orleans, La., 1977)*, Vol. 1, pages 1113–1118. Inst. Electrical Electron. Engrs., New York, 1977.
- [10] Y. Baram and N.R. Sandell, Jr. Consistent estimation on finite parameter sets with application to linear systems identification. *IEEE Trans. Automat. Control*, 23(3):451–454, 1978.
- [11] Y. Baram and N.R. Sandell, Jr. An information theoretic approach to dynamical systems modeling and identification. *IEEE Trans. Automatic Control*, AC-23(1):61–66, 1978.
- [12] O. Barndorff-Nielsen. *Information and exponential families in statistical theory*. John Wiley & Sons Ltd., Chichester, 1978.
- [13] A.R. Barron. The strong ergodic theorem for densities: generalized Shannon-McMillan-Breiman theorem. *Ann. Probab.*, 13(4):1292–1303, 1985.
- [14] J.O. Berger. *Statistical decision theory and Bayesian analysis*. Springer Series in Statistics. Springer-Verlag, New York, 1993.
- [15] R.N. Bhattacharya and R. Ranga Rao. *Normal approximation and asymptotic expansions*. John Wiley & Sons, New York-London-Sydney, 1976.
- [16] D. Blackwell and J.L. Hodges, Jr. The probability in the extreme tail of a convolution. *Ann. Math. Statist.*, 30:1113–1120, 1959.
- [17] N. Bleistein and R.A. Handelsman. *Asymptotic expansions of integrals*. Dover Publications Inc., New York, 1986.
- [18] C.R. Blyth. Necessary and sufficient conditions for inequalities of Cramér-Rao type. *Ann. Statist.*, 2:464–473, 1974.
- [19] C.R. Blyth and D.M. Roberts. On inequalities of Cramér-Rao type and admissibility proofs. In *Proceedings of the Sixth Berkeley Symposium on Mathematical Statistics and Probability (Univ. California, Berkeley, Calif., 1970/1971)*, Vol. I: *Theory of statistics*, pages 17–30, Berkeley, Calif., 1972. Univ. California Press.
- [20] P.E. Caines. A note on the consistency of maximum likelihood estimates for finite families of stochastic processes. *Ann. Statist.*, 3:539–546, 1975.
- [21] P.E. Caines. *Linear stochastic systems*. Wiley Series in Probability and Mathematical Statistics: Probability and Mathematical Statistics. John Wiley & Sons Inc., New York, 1988.
- [22] A. Chambaz. Estimating and testing the order of a model. Technical report, Université Paris-Sud, 2002.
- [23] G. Chamberlain. Econometric applications of maxmin expected utility. *Journal of Applied Econometrics*, 15(6):625–644, 2000.
- [24] D.G. Chapman and H. Robbins. Minimum variance estimation without regularity assumptions. *Ann. Math. Statistics*, 22:581–586, 1951.
- [25] H. Chernoff. A measure of asymptotic efficiency for tests of a hypothesis based on the sum of observations. *Ann. Math. Statistics*, 23:493–507, 1952.
- [26] H. Chernoff. Large-sample theory: parametric case. *Ann. Math. Statist.*, 27:1–22, 1956.
- [27] C. Choirat, C. Hess, and R. Seri. A functional version of the Birkhoff ergodic theorem for a normal integrand: a variational approach. *Ann. Probab.*, 31(1):63–92, 2003.
- [28] E. Clément. *Modélisation statistique en finance et estimation de processus de diffusion*. PhD thesis, Université Paris 9 Dauphine, 1995.
- [29] T.M. Cover and J.A. Thomas. *Elements of information theory*. Wiley Series in Telecommunications. John Wiley & Sons Inc., New York, 1991.
- [30] H. Cramér. *Mathematical Methods of Statistics*. Princeton Mathematical Series, vol. 9. Princeton University Press, Princeton, N. J., 1946.
- [31] H.E. Daniels. Saddlepoint approximations in statistics. *Ann. Math. Statist.*, 25:631–650, 1954.
- [32] A. Dembo and O. Zeitouni. *Large deviations techniques and applications*, volume 38 of *Applications of Mathematics (New York)*. Springer-Verlag, New York, 1998.
- [33] R.H. Farrell. On the best obtainable asymptotic rates of convergence in estimation of a density function at a point. *Ann. Math. Statist.*, 43:170–180, 1972.
- [34] L. Finesso, C.-C. Liu, and P. Narayan. The optimal error exponent for Markov order estimation. *IEEE Trans. Inform. Theory*, 42(5):1488–1497, 1996.
- [35] C. Flinn and J.J. Heckman. New methods for analyzing structural models of labor force dynamics. *J. Econometrics*, 18(1):115–168, 1982.

- [36] J.-P. Florens and J.-F. Richard. Encompassing in finite parametric spaces. Working Paper Series 89-03, Institute of Statistics and Decision Sciences, Duke University, 1989.
- [37] J.C. Fu. Large sample point estimation: a large deviation theory approach. *Ann. Statist.*, 10(3):762–771, 1982.
- [38] A. Futschik and G. Pflug. Confidence sets for discrete stochastic optimization. *Ann. Oper. Res.*, 56:95–108, 1995.
- [39] E. Gassiat and S. Boucheron. Optimal error exponents in hidden Markov models order estimation. *IEEE Trans. Inform. Theory*, 49(4):964–980, 2003.
- [40] S. Geman and C.-R. Hwang. Nonparametric maximum likelihood estimation by the method of sieves. *Ann. Statist.*, 10(2):401–414, 1982.
- [41] V. Genon-Catalot and D. Picard. *Éléments de statistique asymptotique*, volume 11 of *Mathématiques & Applications (Berlin) [Mathematics & Applications]*. Springer-Verlag, Paris, 1993.
- [42] A.M. Geršanov. Optimal estimation of a discrete parameter. *Teor. Veroyatnost. i Primenen.*, 24(1):220–224, 1979.
- [43] A.M. Geršanov and S.K. Šamroni. Randomized estimation in problems with a discrete parameter space. *Teor. Veroyatnost. i Primenen.*, 21(1):195–200, 1976.
- [44] M. Ghosh and G. Meeden. Admissibility of the mle of the normal integer mean. *Sankhyā Ser. B*, 40(1-2):1–10, 1978.
- [45] R.D. Gill and B.Y. Levit. Applications of the Van Trees inequality: a Bayesian Cramér-Rao bound. *Bernoulli*, 1(1-2):59–79, 1995.
- [46] C. Gouriéroux and A. Monfort. *Statistics and Econometric Models*. Cambridge University Press, 1995.
- [47] R.M. Gray and J.C. Kieffer. Asymptotically mean stationary measures. *Ann. Probab.*, 8(5):962–973, 1980.
- [48] U. Grenander. *Abstract inference*. John Wiley & Sons Inc., New York, 1981.
- [49] P. Hall. On Starr and Vardi's estimates of the number of transmission sources. *J. Appl. Probab.*, 19(1):52–63, 1982.
- [50] P. Hall. On convergence rates in nonparametric problems. *International Statistical Review*, 57(1):45–58, 1989.
- [51] P. Hall. *The bootstrap and Edgeworth expansion*. Springer Series in Statistics. Springer-Verlag, New York, 1992.
- [52] J.M. Hammersley. On estimating restricted parameters. *J. Roy. Statist. Soc. Ser. B.*, 12:192–229; discussion, 230–240, 1950.
- [53] A. Hassibi and S. Boyd. Integer parameter estimation in linear models with application to GPS. In *Proceedings of IEEE Conference on Decision and Control*, volume 3, pages 3245–51, Kobe, Japan, December 1996.
- [54] A. Hassibi and S. Boyd. Integer parameter estimation in linear models with applications to GPS. *IEEE Trans. Signal Process.*, 46(11):2938–2952, 1998.
- [55] R.M. Hawkes and J.B. Moore. Performance bounds for adaptive estimation. *Proc. IEEE*, 64(8):1143–1150, 1976.
- [56] R.M. Hawkes and J.B. Moore. Performance of Bayesian parameter estimators for linear signal models. *IEEE Trans. Automatic Control*, AC-21(4):523–527, 1976.
- [57] R.M. Hawkes and J.B. Moore. An upper bound on the mean-square error for Bayesian parameter estimators. *IEEE Trans. Information Theory*, IT-22(5):610–615, 1976.
- [58] A.E. Hero. *Digital Signal Processing Handbook*, chapter Signal detection and classification. Boca Raton: CRC Press LLC, 1999.
- [59] F.C. Hsuan. A stepwise Bayesian procedure. *Ann. Statist.*, 7(4):860–868, 1979.
- [60] P.J. Huber. The 1972 Wald lecture. Robust statistics: A review. *Ann. Math. Statist.*, 43:1041–1067, 1972.
- [61] I.A. Ibragimov and R.Z. Has'minskii. *Statistical estimation*, volume 16 of *Applications of Mathematics*. Springer-Verlag, New York, 1981.
- [62] M. Iltis. Sharp asymptotics of large deviations in \mathbf{R}^d . *J. Theoret. Probab.*, 8(3):501–522, 1995.
- [63] J. Jacod and A.N. Shiryaev. *Limit theorems for stochastic processes*, volume 288 of *Grundlehren der Mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences]*. Springer-Verlag, Berlin, 1987.

- [64] J.L. Jensen. *Saddlepoint approximations*, volume 16 of *Oxford Statistical Science Series*. The Clarendon Press Oxford University Press, New York, 1995.
- [65] B.-Y. Jing and J. Robinson. Saddlepoint approximations for marginal and conditional probabilities of transformed variables. *Ann. Statist.*, 22(3):1115–1132, 1994.
- [66] F. Kanaya and T.S. Han. The asymptotics of posterior entropy and error probability for Bayesian estimation. *IEEE Trans. Inform. Theory*, 41(6, part 2):1988–1992, 1995.
- [67] S. Karlin. Admissibility for estimation with quadratic loss. *Ann. Math. Statist.*, 29:406–436, 1958.
- [68] A.D.M. Kester and W.C.M. Kallenberg. Large deviations of estimators. *Ann. Statist.*, 14(2):648–664, 1986.
- [69] R.A. Khan. On some properties of Hammersley’s estimator of an integer mean. *Ann. Statist.*, 1:756–762, 1973.
- [70] R.A. Khan. A note on the admissibility of Hammersley’s estimator of an integer mean. *Canad. J. Statist.*, 6(1):113–119, 1978.
- [71] R.A. Khan. A note on Hammersley’s estimator of an integer mean. *J. Statist. Plann. Inference*, 88(1):37–45, 2000.
- [72] R.A. Khan. A note on Hammersley’s inequality for estimating the normal integer mean. *Int. J. Math. Math. Sci.*, (34):2147–2156, 2003.
- [73] J. Kim and D. Pollard. Cube root asymptotics. *Ann. Statist.*, 18(1):191–219, 1990.
- [74] A.J. Kleywegt, A. Shapiro, and T. Homem-de Mello. The sample average approximation method for stochastic discrete optimization. *SIAM J. Optim.*, 12(2):479–502 (electronic), 2001/02.
- [75] A.P. Korostelev and S.L. Leonov. Minimax efficiency in the sense of Bahadur for small confidence levels. *Problemy Peredachi Informatsii*, 32(4):3–15, 1996.
- [76] D.G. Lainiotis. A class of upper bounds on probability of error for multi-hypothesis pattern recognition. *IEEE Trans. Information Theory*, IT-15:730–731, 1969.
- [77] D.G. Lainiotis. On a general relationship between estimation, detection, and the Bhattacharyya coefficient. *IEEE Trans. Information Theory*, IT-15:504–505, 1969.
- [78] L.M. Le Cam. *Théorie asymptotique de la décision statistique*. Séminaire de Mathématiques Supérieures, No. 33 (Été, 1968). Les Presses de l’Université de Montréal, Montreal, Que., 1969.
- [79] L.M. Le Cam. Convergence of estimates under dimensionality restrictions. *Ann. Statist.*, 1:38–53, 1973.
- [80] L.M. Le Cam. *Asymptotic methods in statistical decision theory*. Springer Series in Statistics. Springer-Verlag, New York, 1986.
- [81] L.M. Le Cam and G.L. Yang. *Asymptotics in statistics*. Springer Series in Statistics. Springer-Verlag, New York, 1990.
- [82] L.M. LeCam. On some asymptotic properties of maximum likelihood estimates and related Bayes’ estimates. *Univ. California Publ. Statist.*, 1:277–329, 1953.
- [83] B.G. Lindsay and K. Roeder. A unified treatment of integer parameter models. *J. Amer. Statist. Assoc.*, 82(399):758–764, 1987.
- [84] L.A. Liporace. Variance of Bayes estimates. *IEEE Trans. Information Theory*, IT-17:665–669, 1971.
- [85] R. Lugannani and S. Rice. Saddle point approximation for the distribution of the sum of independent random variables. *Adv. in Appl. Probab.*, 12(2):475–490, 1980.
- [86] E. Lukacs and R.G. Laha. *Applications of characteristic functions*. Griffin’s Statistical Monographs & Courses, No. 14. Hafner Publishing Co., New York, 1964.
- [87] C.F. Manski. Maximum score estimation of the stochastic utility model of choice. *J. Econometrics*, 3(3):205–228, 1975.
- [88] C.F. Manski. Semiparametric analysis of discrete response. Asymptotic properties of the maximum score estimator. *J. Econometrics*, 27(3):313–333, 1985.
- [89] C.F. Manski. *Analog estimation methods in econometrics*. Monographs on Statistics and Applied Probability. Chapman and Hall, New York, 1988.
- [90] G.P. McCabe, Jr. Sequential estimation of a Poisson integer mean. *Ann. Math. Statist.*, 43:803–813, 1972.
- [91] G. Meeden and M. Ghosh. Admissibility in finite problems. *Ann. Statist.*, 9(4):846–852, 1981.

- [92] M. Nafe and A. Tewfik. Reduced complexity m-ary hypotheses testing in wireless communications. In *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing, Seattle, Washington, 1998, Vol. 6*, pages 3209–3212. Inst. Electrical Electron. Engrs., New York, 1998.
- [93] W.K. Newey and D. McFadden. Large sample estimation and hypothesis testing. In *Handbook of econometrics, Vol. IV*, volume 2 of *Handbooks in Econom.*, pages 2111–2245. North-Holland, Amsterdam, 1994.
- [94] P.E. Ney. Dominating points and the asymptotics of large deviations for random walk on \mathbf{R}^d . *Ann. Probab.*, 11(1):158–167, 1983.
- [95] P.E. Ney. Convexity and large deviations. *Ann. Probab.*, 12(3):903–906, 1984.
- [96] P.E. Ney. Notes on dominating points and large deviations. *Resenhas*, 4(1):79–91, 1999.
- [97] P.E. Ney and S.M. Robinson. Polyhedral approximation of convex sets with an application to large deviation probability theory. *J. Convex Anal.*, 2(1-2):229–240, 1995.
- [98] J. Neyman. On the problem of estimating the number of schools of fish. *Univ. California Publ. Statist.*, 1:21–36, 1949.
- [99] M. Nikouline and V. Solev. *L'héritage de Kolmogorov en mathématiques*, chapter Problème de l'estimation et ε -entropie de Kolmogorov. Echelles. Belin, 2004.
- [100] P.C.B. Phillips. Finite sample theory and the distributions of alternative estimators of the marginal propensity to consume. *Rev. Econom. Stud.*, 47(1):183–224, 1980.
- [101] H.V. Poor and S. Verdú. A lower bound on the probability of error in multihypothesis testing. *IEEE Trans. Inform. Theory*, 41(6, part 2):1992–1994, 1995.
- [102] A. Puhalskii and V. Spokoiny. On large-deviation efficiency in statistical inference. *Bernoulli*, 4(2):203–272, 1998.
- [103] C.P. Robert. *The Bayesian choice*. Springer Texts in Statistics. Springer-Verlag, New York, 1994.
- [104] J. Robinson, T. Höglund, L. Holst, and M.P. Quine. On approximating probabilities for small and large deviations in \mathbf{R}^d . *Ann. Probab.*, 18(2):727–753, 1990.
- [105] D.S. Robson. Admissible and minimax integer-valued estimators of an integer-valued parameter. *Ann. Math. Statist.*, 29:801–812, 1958.
- [106] T.J. Rothenberg. Identification in parametric models. *Econometrica*, 39:577–591, 1971.
- [107] L.V. Rozovsky. Asymptotic expansions for probabilities of large deviations. *Probab. Theory Relat. Fields*, 73(2):299–318, 1986.
- [108] K. Ryu. Structural duration analysis of management data. *J. Econometrics*, 57(1-3):91–115, 1993.
- [109] K. Ryu. Econometric analysis of mixed parameter models. *Journal of Economic Theory and Econometrics*, 5, 1999.
- [110] A.N. Shiryaev. *Probability*, volume 95 of *Graduate Texts in Mathematics*. Springer-Verlag, New York, 1996.
- [111] G.L. Sievers. Estimates of location: a large deviation comparison. *Ann. Statist.*, 6(3):610–618, 1978.
- [112] S.D. Silvey. A note on maximum-likelihood in the case of dependent random variables. *J. Roy. statist. Soc. Ser. B*, 23:444–452, 1961.
- [113] A.E. Stark. Some estimators of the integer-valued parameter of a Poisson variate. *J. Amer. Statist. Assoc.*, 70(351, part 1):685–689, 1975.
- [114] N. Starr. Optimal and adaptive stopping based on capture times. *J. Appl. Probability*, 11:294–301, 1974.
- [115] I. Vajda. On the statistical decision problems with discrete parameter space. *Kybernetika (Prague)*, 3:110–126, 1967.
- [116] I. Vajda. On the statistical decision problems with finite parameter space. *Kybernetika (Prague)*, 3:451–466, 1967.
- [117] I. Vajda. Rate of convergence of the information in a sample concerning a parameter. *Czechoslovak Math. J.*, 17 (92):225–231, 1967.
- [118] I. Vajda. On the convergence of information contained in a sequence of observations. In *Proc. Colloquium on Information Theory (Debrecen, 1967), Vol. II*, pages 489–501. János Bolyai Math. Soc., Budapest, 1968.
- [119] I. Vajda. A discrete theory of search. I. *Apl. Mat.*, 16:241–255, 1971.
- [120] I. Vajda. A discrete theory of search. II. *Apl. Mat.*, 16:319–335, 1971.

- [121] I. Vajda. On the convergence of Bayes empirical decision functions. In *Proceedings of the Prague Symposium on Asymptotic Statistics (Charles Univ., Prague, 1973)*, Vol. II, pages 413–425, Prague, 1974. Charles Univ.
- [122] A.W. van der Vaart. Superefficiency. In *Festschrift for Lucien Le Cam*, pages 397–410. Springer, New York, 1997.
- [123] M.H. van der Vlerk. Stochastic programming bibliography. World Wide Web, <http://mally.eco.rug.nl/spbib.html>, 1996-2003.
- [124] Y. Vardi. On a stopping time of Starr and its use in estimating the number of transmission sources. *J. Appl. Probab.*, 17(1):235–242, 1980.
- [125] H. White. Maximum likelihood estimation of misspecified models. *Econometrica*, 50(1):1–25, 1982.
- [126] W.H. Wong. Theory of partial likelihood. *Ann. Statist.*, 14(1):88–123, 1986.

DIPARTIMENTO DI ECONOMIA, UNIVERSITÀ DEGLI STUDI DELL'INSUBRIA, VIA RAVASI 2, 21100 VARESE, ITALY

E-mail address: `cchoirat@eco.uninsubria.it`

DIPARTIMENTO DI ECONOMIA, UNIVERSITÀ DEGLI STUDI DELL'INSUBRIA, VIA RAVASI 2, 21100 VARESE, ITALY

E-mail address: `rseri@eco.uninsubria.it`