INSTITUT NATIONAL DE LA STATISTIQUE ET DES ETUDES ECONOMIQUES
Série des Documents de Travail du CREST
(Centre de Recherche en Economie et Statistique)

n° 2001-38

Estimation in Discrete Parameter Models

C. CHOIRAT[1]
R. SERI[2]

Les documents de travail ne reflètent pas la position de l'INSEE et n'engagent que leurs auteurs.

Working papers do not reflect the position of INSEE but only the views of the authors.

1
2

Estimation in Discrete Parameter Models

Christine Choirat* Raffaello Seri†

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.

Résumé : Dans certains problèmes d'estimation, et plus particulièrement dans ceux liés à la théorie de l'information, la théorie du signal et la biologie, la nature du problème nous apporte des informations supplémentaires qui permettent de restreindre l'espace des paramètres à un nombre fini de points. Dans un tel cas, on parle de modèles à paramètre discret.

Bien que ce problème soit assez ancien et bien qu'il ait des connexions intéressantes avec la théorie des tests et la sélection de modèle, la théorie asymptotique n'a presque jamais été étudiée. C'est pourquoi nous traitons la convergence, la distribution asymptotique, les inégalités d'information et leurs relations avec l'efficacité et la super-efficacité pour une classe générale de $m-$estimateurs. Puis, nous étendons certains des résultats précédents à d'autres classes d'estimateurs.

Mots Clés : Espace de paramètres discret, Détection, Grandes déviations, Inégalités d'information, Efficacité, Super-efficacité.

Classification JEL : C13, C12, C44.

Acknowledgments: The authors would like to thank Lucien Birgé, Mehmet Caner, Jean-Pierre Florens, Christian Gouriéroux, Christian Hess, Oliver B. Linton, Christian P. Robert, Keunkwan Ryu and the participants to seminars at Université Paris 9 Dauphine and CREST, and to ESEM 2001 in Lausanne. All the remaining errors are our responsibility.

*Centre de Recherche Viabilité, Jeux, Contrôle, Université Paris Dauphine, 75775 PARIS CEDEX 16, FRANCE, email: `choirat@viab.ufrmd.dauphine.fr`

†CREST-LFA, Timbre J320, 15 bd Gabriel Péri, 92245 MALAKOFF CEDEX, FRANCE, email: `seri@ensae.fr`

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 Hall, 1989 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 as, for instance, in the estimation of the support of a density (see Flinn and Heckman, 1982) or in the estimation of an autoregressive coefficient under the unit root hypothesis. $\sqrt[3]{n}-$consistency can also be encountered, as in the semiparametric maximum score estimator of Manski (1975, 1985) and the other cases treated in Kim and Pollard (1990).

Sometimes, especially in applications dealing with signal processing and biology, theory provides us with some additional information allowing 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 some estimators is exponential (Hammersley, 1950, Chapman and Robbins, 1951, Vajda, 1967, 1971a, b). 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 Ryu, 1993 and 1999), the result being that part of the parameters converges with 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 (1950) in a seminal paper: Hammersley derived a lower bound for the variance of a consistent estimator analogous to the Cramér-Rao and Chapman-Robbins lower bounds (see Section 2.4.2). 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 Hammersley, 1950, p. 192); this case was further developed by Khan (1973, 1998). The Poisson case also met some attention in the literature and was dealt with by Hammersley (1950, p. 199), McCabe (1972) and Stark (1975).

Starting with Hammersley's paper, the properties of specific estimators were investigated and topics such as asymptotic distributional results and information inequalities were neglected. Admissibility and minimax efficiency were dealt with by Robson (1958), Karlin (1958, p. 424, in the special case of a translation integral parameter and of integral data under the quadratic loss), Khan (1973, 1998, for the case of the Gaussian distribution), Ghosh and Meeden (1978, for the case of the Gaussian distribution) and Meeden and Ghosh (1981). Optimality of estimation under a discrete parameter space and integer-valued data

was also considered by Vajda (1967, 1971a, b) in a nonorthodox setting inspired by Rényi's theory of random search. 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 Lainiotis, 1969a, b, Liporace, 1971, Hawkes and Moore, 1976a, b, c, Baram and Sandell, 1978a, b). Moreover, in the estimation of complex statistical models (see Grenander, 1981, Clement, 1995, Ch. 4) and in the calculation of efficiency rates (see Bahadur, 1960, Le Cam, 1969, Chamberlain, 2000), approximating a general parameter space by a sequence of finite sets has proved to be a valuable tool.

More recently, the topic has received new interest in the Information Theory literature (see Poor and Verdù, 1995, Kanaya and Han, 1995, and the review by Hero, 1997), and in Stochastic Integer Programming (see Futschik and Pflug, 1995, Kleywegt et al., 1999, and van der Vlerk, 2001).

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 restrict ourselves to the case of discrete parameter models for samples of independent and identically distributed random variables: the related case of mixed parameter models will be considered in a forthcoming work, while the extension to dependent and not identically distributed random variables is left to the reader. Therefore, in Section 2, we investigate the properties of a class of $m-$estimators. In particular, in Section 2.1, we derive some conditions for strong consistency; then, in Section 2.2, 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 2.3. In Section 2.4, 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 $\theta_0 \in \Theta$ and that 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 3, we extend some of the previous results to other classes of estimators.

1.1 Some 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 1.6, 1.7 and 1.8 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 1.1 (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:*

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

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

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

where n.i. $[x]$ *is the integer nearest to* x *(see Hammersley, 1950, p. 236).*

Example 1.2 (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* $\left\{\theta_0 + \varepsilon \cdot N, N \in \mathbb{N}^k\right\}$, *for fixed* θ_0 *and* ε *(see Lindsay and Roeder, 1987, 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 by Hammersley (1950, p. 192) and Khan (1973, 1998), the Poisson case by Hammersley (1950, p. 199), McCabe (1972) and Stark (1975). In particular, Hammersley (1950) 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 1.3 (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 Kleywegt et al. (1999) for some theoretical results and a discussion of the stochastic knapsack problem and van der Vlerk (2001) for an up-to-date bibliography.*

Example 1.4 (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 (Hawkes and Moore, 1976a, b, c, Baram and Sandell, 1978a, b) and is reminiscent of some related statistical techniques (e.g., the* discretization *device of Le Cam, 1969, Chapitre III, or the* sieve estimation *of Grenander, 1981).*

Example 1.5 ($M-$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 Nafie and Tewfik, 1998). 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 $\{m_0(t), \ldots, m_J(t)\}$: 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:*

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

Much more complex cases can be dealt with; see Hero (1997) for an introduction.

Example 1.6 (Ryu's plant management model). *In Ryu (1993), 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 1.7 (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 $\left(\alpha^{\mathsf{T}}, \beta^{\mathsf{T}}\right)^{\mathsf{T}}$ is composed of a continuous and of a discrete part. A complete treatment of this problem is in Hassibi and Boyd (1996, 1998). We remark that similar applications of mixed parameter models can be found in radar imaging, MRI and communications.

Example 1.8 (Model selection). *In this case, several models indexed by $i = 1, ..., I$ are considered; each model belongs to the parametric family $\mathcal{P}_i = \left\{\mathbb{P}_{\beta^i},\ \beta^i \in B\right\}$. Here, we have supposed that the parameter space B is the same for any alternative model: however this hypothesis can be easily removed. We can therefore embed the estimation problem in the* nesting model[1]

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

with likelihood function:

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

Example 1.9 (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 Starr, 1974, Vardi, 1980 and Hall, 1982). The same kind of situation arises in the fishing problem (Neyman, 1949), the proofreading problem, the debugging software problem, etc.*

Example 1.10 (Capture-recapture models). *Consider a population of N animals that has a known percentage $p\left(= \frac{X}{N}\right)$ of marked animals: n animals are sampled without replacement and the observation $Y = y$ indicates the number of marked individuals in the sample; the unknown parameter is $N \in \mathbb{N}$. The likelihood is given by the hypergeometric density:*

$$f_Y(y; N) = \frac{\binom{X}{y}\binom{N-X}{n-y}}{\binom{N}{n}}.$$

Clearly, more complex models can be considered (see Lindsay and Roeder, 1987).

As already stated, Examples 1.6, 1.7 and 1.8 will not be treated in this paper. Also the models of Examples 1.9 and 1.10 cannot be considered along with the other ones. Indeed, Example 1.9 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 $\mathbb{E}\left(\hat{\theta}(T) - \theta\right)^2 + cT$,

[1] Gouriéroux and Monfort (1995, Vol. 2, Sect. 22.2.7) deal with model selection through nesting models in a slightly different framework not based on discrete parameters.

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 1.10 differs from the previous ones in another more peculiar sense. The asymptotics $\frac{n}{N} \to 1$ (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 $N \to \infty$ and $n \to \infty$ (with a different rate such that $\frac{n}{N} \to 0$). For any well-behaved estimator, as long as $N \to \infty$, we expect $\hat{N} \to \infty$; therefore, consistency is often replaced by the $\alpha-$*consistency* property ($\frac{\hat{N}-N}{N^{\alpha}} \to 0$, for $\alpha > 0$). Under additional assumptions, it is possible to show that the estimator $\hat{N}$ is $\alpha-$consistent and asymptotically normal. This shows well that the asymptotic behavior is nonstandard, as pointed out, among others, by Lindsay and Roeder (1987).

2 $m-$estimators in Discrete Parameter Models

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

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

in what follows, we allow for misspecification.

2.1 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 nevertheless needed 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 $\mathbb{P}_{\theta^*}$ where θ^* is the true parameter value.
The estimator $\hat{\theta}^n$ is obtained maximizing over the set $\Theta = \{\theta_0, \theta_1, ..., \theta_J\}$, of finite cardinality, the objective function:

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

[2] The expression $m-$estimator stands for *maximum likelihood type estimator*, in the spirit of Huber (1973), and not for *maximum* (or *extremum*) *estimator* (often written as $M-$*estimator*).

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

$$\mathbb{E}_{\theta^*} \left| \ln q\left(Y;\theta_j\right) \right| < +\infty, \quad \forall \theta_j \in \Theta,$$

where $\mathbb{E}_{\theta^*}$ denotes the expectation taken under the true probability measure $\mathbb{P}_{\theta^*}$.
Moreover, θ_0 is the point of Θ maximizing $\mathbb{E}_{\theta^*} \ln q\left(Y;\theta\right)$ and θ_0 is globally identified (see Rothenberg, 1971, White, 1982).

Remark 2.1. ***(i)*** *The assumption of a finite parameter space seems restrictive with respect to the more general assumption of Θ being countable (see Hammersley, 1950). However, Obj is compatible with the convex hull of Θ being compact, as in standard asymptotic theory. Indeed, the previously quoted cases have the characteristic of having convex likelihood functions and this is a well known substitute for compactness of Θ (see Andersen and Gill, 1982, p. 1116 and Newey and McFadden, 1994, p. 2133). Moreover, this restriction is 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 stationary ergodicity: then, the SLLN of Proposition 2.1 should be substituted with the Shannon-McMillan-Breiman Theorem of Barron (1985) or the extended Birkhoff Ergodic Theorem of Choirat, Hess and Seri (2001). As regards the asymptotic distribution, some results by Iltis (2000) allow for the extension of our results to realizations of a Markov process.*
(iii) *The relative position of the points of Θ is clearly uninteresting and the choice of θ_0 as the maximizer is clearly irrelevant and is made only for practical purposes. Remark that θ^* does not necessarily belong to the parameter space Θ.*

Proposition 2.1. *Under Assumption Obj, the $m-$estimator $\hat{\theta}^n$ is a $\mathbb{P}_{\theta^*}-$strongly consistent estimator of θ_0 and is $\mathcal{Y}^{\otimes n}-$measurable.*

Proof. Under Obj, Kolmogorov's SLLN implies that $\mathbb{P}_{\theta^*} -$ as:

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

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

$$\left\{ \left(y_1,...,y_n\right) \in \mathfrak{Y}^n \left| \sup_{\theta\in\Theta} \frac{1}{n}\sum_{i=1}^{n} \ln q\left(y_i;\theta\right) \leq t \right. \right\}$$
$$= \bigcap_{\theta_j\in\Theta} \left\{ \left(y_1,...,y_n\right) \in \mathfrak{Y}^n \left| \frac{1}{n}\sum_{i=1}^{n} \ln q\left(y_i;\theta_j\right) \leq t \right. \right\}.$$

□

Remark 2.2. *A similar proof of consistency has already been provided by Caines (1975), by Barron (1985, p. 1294), as an application of the Shannon-McMillan-Breiman Theorem of information theory, and by Wong (1986, Sect. 2.1), as a preliminary result of his work on partial likelihood.*

2.2 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 the true value of the parameter. Since the determination of this asymptotic distribution is an interesting problem, we derive in this Section upper and lower bounds and asymptotic estimates for probabilities of the form $\mathbb{P}_{\theta^*}\left(\hat{\theta}^n = \theta_i\right)$.

To simplify the following discussion, we introduce the processes:

$$\begin{cases} Q_n(\theta_j) \triangleq \frac{1}{n} \cdot \sum_{i=1}^n \ln q(y_i;\theta_j), \\ \mathbf{X}_k^{(i)} \triangleq [\ln q(Y_k;\theta_i) - \ln q(Y_k;\theta_j)]_{j=0,\ldots,J, j\neq i}, \\ \mathbf{X}_k \triangleq \mathbf{X}_k^{(0)} = [\ln q(Y_k;\theta_0) - \ln q(Y_k;\theta_j)]_{j=1,\ldots,J}. \end{cases} \tag{2.1}$$

In some cases, the Inversion Formula for the Characteristic Function (Lukacs and Laha, 1964, Theorem 1.6.2, p. 23) can be used in order to have:

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

However this method is not general enough to give widely applicable formulas.

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

$$\mathbb{P}_{\theta^*}\left(\hat{\theta}^n = \theta_i\right) = \mathbb{P}_{\theta^*}\left(\frac{1}{\sqrt{n}}\sum_{k=1}^n \left[X_{k,j}^{(i)} - \mathbb{E}_{\theta^*} X_{k,j}^{(i)}\right] > -\sqrt{n}\mathbb{E}_{\theta^*} X_{k,j}^{(i)}, \forall j \neq i\right), \tag{2.2}$$

it would be tempting to use an Edgeworth expansion for a sum of independent random vectors, as given by Bhattacharya and Rao (1976, Corollary 20.4, p. 215; see also Hall, 1992, Lemma 5.4, p. 263). Though, an asymptotic Edgeworth expansion of the probability $\mathbb{P}_{\theta^*}\left(\sqrt{n}\bar{\mathbf{X}} > \mathbf{x}\right)$ fails to converge or to adequately describe the probability tail for values of $\mathbf{x}$ of size $n^{1/6}$ or smaller (see Hall, 1992, p. 325 and Phillips, 1980, 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 $\mathbb{P}_{\theta^*}\left(\hat{\theta}^n = \theta_1\right)$ under conditions that can be found in Cramér (1946, p. 223) or Hall (1992, 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 (2.2) since they only work if $\mathbf{x}$ is of size $o\left(n^{1/2}\right)$ (see e.g. Rozovsky, 1986, pp. 299-300).

Therefore, in Section 2.2.3, we derive some results on the asymptotic behavior of $\mathbb{P}_{\theta^*}\left(\hat{\theta}^n = \theta_i\right)$ using large deviations principles (LDP). Indeed, such a probability can be written as:

$$\mathbb{P}_{\theta^*}\left(\hat{\theta}^n = \theta_i\right) = \mathbb{P}_{\theta^*}\left(Q_n\left(\theta_i\right) > Q_n\left(\theta_j\right), \forall j \neq i\right) = \mathbb{P}_{\theta^*}\left(\sum_{k=1}^{n} \mathbf{X}_k^{(i)} \in \operatorname{int} \mathbb{R}_+^J\right), \tag{2.3}$$

and is therefore amenable to large deviations approximation. Remark that we could have defined the same probability as $\mathbb{P}_{\theta^*}\left(Q_n\left(\theta_i\right) \geq Q_n\left(\theta_j\right), \forall j \neq i\right)$ 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$-(i)$) ensures that the difference is asymptotically negligible (see Hall, 1992, Theorem 2.3, p. 57), but the result is not strong enough to guarantee that the estimators have the same asymptotic distribution. However, we will give some alternative conditions (see Proposition 2.4) under which this difference is asymptotically irrelevant.

Section 2.2.1 introduces some definitions and hypotheses and discusses a preliminary result used in the large deviations estimate.

In Section 2.2.4, 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 Bahadur and Rao (1960), Ney (1983, 1984) and Iltis (1995). The aim is to find asymptotic results, as $n \to \infty$, of the form:

$$\mathbb{P}_{\theta^*}\left(\sum_{i=1}^{n} \mathbf{Y}_k \in n \cdot \Gamma\right) = n^{\gamma} \cdot e^{-n \cdot \inf_{\mathbf{y} \in \Gamma} I(\mathbf{y})} \cdot \left(d_0 + o\left(1\right)\right), \tag{2.4}$$

where $I\left(\cdot\right)$ is the rate function for which the LDP holds, and we have $\mathbf{Y}_k = \mathbf{X}_k^{(i)}$ and $\Gamma = \operatorname{int} \mathbb{R}_+^J$. We only have to determine if a result such as (2.4) holds, and in this case, what the values of γ and d_0 are.

Section 2.2.2 shows how these results can be obtained in greater generality when the parameter space is of the form $\Theta = \{\theta_0, \theta_1\}$, that is when $J = 1$.

2.2.1 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 Dembo and Zeitouni, 1998); we recall that the processes $Q_n\left(\theta_j\right)$, $\mathbf{X}_k$ and $\mathbf{X}_k^{(i)}$ have been introduced in equation (2.1). Then we define the moment generating functions

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

the logarithmic moment generating functions

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

and the Cramér transforms

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

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 $f : E \to \mathbb{R}_+$, we will need the definition of the *effective domain* of f, $\mathcal{D}_f \triangleq \{x \in E : f(x) < \infty\}$.

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

Min When $J = 1$, there is a positive value $\mu \in \operatorname{int} \mathcal{D}_{\Lambda^{(1)}}$ 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 Dembo and Zeitouni, 1998, p. 110).

$\eta-$**Int** There exists a $\delta > 0$, such that, for any $\eta \in]-\delta, \delta[$ we have:

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

In the following (Lemma 2.2), we will show that the latter hypothesis is equivalent to the Cramér condition $\mathbf{0} \in \operatorname{int} \mathcal{D}_{\Lambda^{(i)}}$, for any $i = 0, ..., J$.

Remark 2.3. *In what follows, this Assumption could be substituted by a condition as in Ney and Robinson (1995, Assumptions H1 and H2).*

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

Remark 2.4. ***(i)*** *Under Assumptions* **Obj** *and* $\eta-$**Int**, *since* $\Lambda^{(i)}$ *is convex,* $\operatorname{int} \mathcal{D}_{\Lambda^{(i)}}$ *is nonempty and* $\Lambda^{(i)}$ *is differentiable throughout* $\operatorname{int} \mathcal{D}_{\Lambda^{(i)}}$, $\Lambda^{(i)}(\cdot)$ *is also* essentially smooth.
(ii) *The assumption* **Steep**$-(i)$ *can be substituted, less generally, by the openness of* $\mathcal{D}_{\Lambda^{(i)}}$ *(see Ney, 1984, p. 905, and Iltis, 1995, pp. 505-506).*

Meet$-(i)$ $\operatorname{int}\left(\mathbb{R}_+^J \cap \mathcal{S}^{(i)}\right) \neq \varnothing$, where $\mathcal{S}^{(i)}$ is the closure of the convex hull of the support of the law of $\mathbf{X}^{(i)}$.

Com$-(i)$ $\Lambda^{(i)}(\lambda)$ converges in a neighborhood of each $\lambda \in K_i$, with K_i compact.

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

Remark 2.5. *The assumption $L^r-\mathsf{Sum}-(i)$ can be substituted by a general nonlattice condition such as the forthcoming $\mathsf{Cra}-(i)$, 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 $n \to \infty$; see Iltis, 1995, p. 521).*

$\mathsf{Cra}-(i)$ *(Cramér condition)* $\limsup_{\|\lambda\|\to\infty} \left|M^{(j)}(i\lambda)\right| < 1, \forall j = 0, ..., J$, where $M^{(j)}(i\lambda)$ is a formal expression for the characteristic function.

We will need the following Lemma showing the equivalence between Assumption $\eta-\mathsf{Int}$ and the so-called *Cramér condition* $\mathbf{0} \in \operatorname{int} \mathcal{D}_{\Lambda^{(i)}}$, for any $i = 0, ..., J$ (beware of confusion with Assumption $\mathsf{Cra}-(i)$, also called Cramér condition).

Lemma 2.2. *Under Assumption Obj, the following conditions are equivalent:*

(i) *Assumption $\eta-\mathsf{Int}$ holds;*

(ii) $\mathbf{0} \in \operatorname{int} \mathcal{D}_{\Lambda^{(i)}}$, *for any $i = 0, ..., J$.*

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

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

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

2.2.2 Distribution of $\hat{\theta}^n$ for $J = 1$

We first treat the case when $J = 1$, that is when $\Theta = \{\theta_0, \theta_1\}$, since it is simpler and allows for a more complete solution. In the case $J = 1$, under Assumption Obj, a strong form of Cramér's Theorem in $\mathbb{R}$ (see e.g. Dembo and Zeitouni, 1998, Corollary 2.2.19, p. 34) holds:

$$\mathbb{P}_{\theta^*}\left(\hat{\theta}^n = \theta_1\right) \leq \mathbb{P}_{\theta^*}\left(Q_n(\theta_1) \geq Q_n(\theta_0)\right) = \exp\left\{-(n + o(n)) \cdot \inf_{y\geq 0} \Lambda^{(1),*}(y)\right\}.$$

From the convexity property of $\Lambda^{(1),*}$ and the fact that the minimum is reached at $\mathbb{E}_{\theta^*}\left[\ln \frac{q(Y_k;\theta_1)}{q(Y_k;\theta_0)}\right] < 0$, we have $\inf_{y\geq 0} \Lambda^{(1),*}(y) = \Lambda^{(1),*}(0)$.

If $\Lambda^{(1),*}(0) = +\infty$, then $\mathbb{P}_{\theta^*}\left(\hat{\theta}^n = \theta_1\right) = \exp[o(n)]$. Otherwise, if $\eta-\mathsf{Int}$ is verified, $[0, +\infty)$ is a $\Lambda^{(1),*}-$continuity set and we get:

$$\mathbb{P}_{\theta^*}\left(\hat{\theta}^n = \theta_1\right) = \exp\left\{-(n + o(n)) \cdot \Lambda^{(1),*}(0)\right\}. \tag{2.5}$$

Moreover, Chebychev inequality (see equation (2.2.11) in Dembo and Zeitouni, 1998) implies:

$$\mathbb{P}_{\theta^*}\left(\hat{\theta}^n = \theta_1\right) \leq \inf_{\lambda\geq 0} \left[M^{(1)}(\lambda)\right]^n.$$

The following Proposition allows us to obtain a much more precise estimate of the convergence rate based on exact asymptotics for large deviation probabilities.

Proposition 2.3. *Under Assumptions Obj and Min, with* $\Theta = \{\theta_0, \theta_1\}$ *and* $J = 1$*, we have*

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

Proof. We verify the conditions of Theorem 3.7.4 in Dembo and Zeitouni (1998, p. 110). We want to study the behavior of

$$\mathbb{P}_{\theta^*}\left(\hat{\theta}^n \neq \theta_0\right) = \mathbb{P}_{\theta^*}\left(\hat{\theta}^n = \theta_1\right) = \mathbb{P}_{\theta^*}\left\{Q_n(\theta_1) \geq Q_n(\theta_0)\right\}$$
$$= \mathbb{P}_{\theta^*}\left\{Q_n(\theta_1) - Q_n(\theta_0) \in [0,+\infty)\right\},$$

and we start from the large deviations estimate (2.5). Then we look for the positive $\mu \in \operatorname{int}\mathcal{D}_{\Lambda^{(1)}}$ solution to the equation $0 = \left(\Lambda^{(1)}\right)'(\mu)$. From Lemma 2.2.5 (c) in Dembo and Zeitouni (1998), this implies $\Lambda^{(1)}(\mu) = -\Lambda^{(1),*}(0)$, and the result follows. □

Remark 2.6. *A refinement of the previous asymptotic rates can be obtained using results in Blackwell and Hodges (1959) and Bahadur and Rao (1960).*

2.2.3 Large Deviations Asymptotics in $\mathbb{R}^J$

In this Section we will extend the previous results for $J > 1$. We remark that, in what follows, $\operatorname{int}\left(\mathbb{R}_+^J\right)^c$ stands for $\operatorname{int}\left\{\left[(\mathbb{R}_+)^J\right]^c\right\}$.

Proposition 2.4. *For* $i = 1, \ldots, J$*, the following results hold. Under Assumption Obj:*

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

where $o_{\inf}(n)$ *is a function such that* $\liminf_{n\to\infty}\frac{o_{\inf}(n)}{n} = 0$.
Under Assumptions Obj and $\eta-$*Int:*

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

where $o_{\sup}(n)$ *is a function such that* $\limsup_{n\to\infty} \frac{o_{\sup}(n)}{n} = 0$.
Under Assumptions Obj, $\eta-$Int *and* Meet$-$ (i)*:*

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

Proof. The first two results are straightforward applications of Cramér's Theorem in $\mathbb{R}^d$ (see e.g. Dembo and Zeitouni, 1998, Corollary 6.1.6, p. 253). Indeed, it is known that the lower bound holds without any supplementary assumption:

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

The upper bound requires a Cramér condition $\mathbf{0} \in \operatorname{int}\mathcal{D}_{\Lambda^{(i)}}$: indeed, from Lemma 2.2, this is implied by Assumption $\eta-$Int. Then, a full LDP holds:

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

In order to prove the final result, we have to show that $\mathbb{R}_+^J$ is a $\Lambda^{(i),*}-$*continuity set*, that is $\inf_{\mathbf{y}\in \operatorname{int}\mathbb{R}_+^J} \Lambda^{(i),*}(\mathbf{y}) = \inf_{\mathbf{y}\in \mathbb{R}_+^J} \Lambda^{(i),*}(\mathbf{y})$. It is enough to apply part (ii) in Lemma on page 903 of Ney (1984). Remark however that the hypothesis of essential smoothness of f is unnecessary and that the proof proceeds along the same lines if we do not use the strict convexity of f^* but only its convexity. □

Proposition 2.5. *Under Assumption* Obj*, the following inequality holds:*

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

where H *is the finite cardinality of the set* $\arg\inf_{\mathbf{y}\in \operatorname{int}\left(\mathbb{R}_+^J\right)^c} \Lambda^*(\mathbf{y})$ *and* $o_{\inf}(n)$ *is a function such that* $\liminf_{n\to\infty} \frac{o_{\inf}(n)}{n} = 0$.
Under Assumptions Obj *and* $\eta-$Int*:*

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

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

Proof. First of all, we remark that $\mathbb{P}_{\theta^*}\left(\hat{\theta}^n \neq \theta_0\right) = \mathbb{P}_{\theta^*}\left(\sum_{k=1}^n \mathbf{X}_k \in \text{int}\left(\mathbb{R}_+^J\right)^c\right)$. 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

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

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 $\arg\inf_{\mathbf{y}\in\text{int}\left(\mathbb{R}_+^J\right)^c}\Lambda^*(\mathbf{y})$ has at most finite cardinality H. Moreover, as large deviations theory allows us to ignore the part of $\text{int}\left(\mathbb{R}_+^J\right)^c$ where $\Lambda^*(\mathbf{y}) \geq \varepsilon + \inf_{\mathbf{y}\in\text{int}\left(\mathbb{R}_+^J\right)^c}\Lambda^*(\mathbf{y})$, we can replace $\left(\mathbb{R}_+^J\right)^c$ with a collection of H disjoint sets, say Γ_h, $h = 1, \ldots, H$, each one containing in its interior one and only one of the points of $\arg\inf_{\mathbf{y}\in\text{int}\left(\mathbb{R}_+^J\right)^c}\Lambda^*(\mathbf{y})$ (see Iltis, 1995, p. 508):

$$\begin{aligned}\mathbb{P}_{\theta^*}\left(\sum_{k=1}^n \mathbf{X}_k \in \text{int}\left(\mathbb{R}_+^J\right)^c\right) &= (1+o(1))\cdot\mathbb{P}_{\theta^*}\left(\sum_{k=1}^n \mathbf{X}_k \in \text{int}\bigcup_{h=1}^H \Gamma_h\right) \qquad (2.6)\\ &= (1+o(1))\cdot\sum_{h=1}^H \mathbb{P}_{\theta^*}\left(\sum_{k=1}^n \mathbf{X}_k \in \text{int}\,\Gamma_h\right).\end{aligned}$$

As before, the bounds derive from Cramér's Theorem in $\mathbb{R}^d$. Remarking that the contribution of any Γ_h is the same and recalling (2.6), we get the results. □

Remark 2.7. *The Proposition allows us to obtain an upper bound on the bias of the MLE. Indeed,*

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

and:

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

2.2.4 Exact Asymptotics in $\mathbb{R}^J$

In higher dimensions, the exact asymptotic behavior of the probability $\mathbb{P}_{\theta^*}\left(\hat{\theta}^n = \theta_i\right)$ 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 (2.3) $\mathbb{P}_{\theta^*}\left(\hat{\theta}^n = \theta_i\right) = \mathbb{P}_{\theta^*}\left(\sum_{k=1}^n \mathbf{X}_k^{(i)} \in \mathbb{R}_+^J\right)$, for $i = 1, \ldots, J$. The probability $\mathbb{P}_{\theta^*}\left(\hat{\theta}^n = \theta_0\right)$ will be obtained by subtraction.

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

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

A point $\mathbf{y}^{(i)}$ is a *dominating point* (see Ney, 1984, p. 904) of Γ if:

- $\mathbf{y}^{(i)} \in \partial\Gamma \cap \operatorname{int}\mathcal{S}^{(i)}$ where $\mathcal{S}^{(i)}$ is the closure of the convex hull of the support of the law of $\mathbf{X}^{(i)}$;
- the equation $\operatorname{grad}\Lambda^{(i)}(\lambda) = \mathbf{y}^{(i)}$ has a unique solution $\lambda\left(\mathbf{y}^{(i)}\right) \triangleq \lambda^{(i)} \in \mathcal{D}_{\Lambda^{(i)}}$;
- $\Gamma \subset \Pi^{+}\left(\lambda^{(i)}, \mathbf{y}^{(i)}\right)$, where $\Pi^{+}(\alpha, \mathbf{x}) \triangleq \left\{\mathbf{z} \in \mathbb{R}^{J} \left| \mathbf{z}^{\mathsf{T}}\alpha \geq \mathbf{x}^{\mathsf{T}}\alpha \right.\right\}$ is an half-space.

The following Proposition holds.

Proposition 2.6. *Under Assumptions Obj, $\eta-$Int, Meet$-(i)$ and Steep$-(i)$, for sufficiently large n, the following result holds:*

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

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

Proof. The hypotheses of the Theorem on page 904 of Ney (1984) are easily verified. This shows that a unique dominating point $\mathbf{y}^{(i)}$ exists and implies, through Proposition on page 161 of Ney (1983) (see Ney, 1984, p. 905 for some remarks on the hypotheses), that the stated bracketing of $\mathbb{P}_{\theta^*}\left(\hat{\theta}^{n} = \theta_i\right)$ holds. □

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

We define the *level surface* $S_a^{(i)} \triangleq \left\{\mathbf{y} \in \mathbb{R}^{J} \left| \Lambda^{(i),*}(\mathbf{y}) = a \right.\right\}$ and the *level set* $V_a^{(i)} \triangleq \left\{\mathbf{y} \in \mathbb{R}^{J} \left| \Lambda^{(i),*}(\mathbf{y}) \geq a \right.\right\}$ and we set $\Lambda^{(i),*} \triangleq \Lambda^{(i),*}\left(\mathbf{y}^{(i)}\right)$; 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 $\mathbf{R}^{(i)} \cdot \lambda^{(i)} \triangleq \left(0, ..., 0, \left\|\lambda^{(i)}\right\|\right)$. Therefore, the new direction $z_J^{(i)}$ points in the direction normal to $S_{\Lambda^{(i),*}}^{(i)}$. The directions $\left(z_1^{(i)}, ..., z_J^{(i)}\right)$ form an orthonormal basis and the set Γ becomes $\left\{\mathbf{R}^{(i)} \cdot \mathbf{y} \left| \mathbf{y} \in \Gamma - \mathbf{y}^{(i)} \right.\right\}$. 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 $\mathbf{y}_{(J)} \triangleq (y_1, ..., y_{J-1})$ and $\mathbf{y} = \left(\mathbf{y}_{(J)}, y_J\right)$.

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

Proposition 2.7. *Under Assumptions Obj, $\eta-$Int, Meet$-(i)$ and Steep$-(i)$, we have:*

(I) $\mathbf{y}^{(i)} = \mathbf{0}$*:*

(i) $\lambda^{(i)} \in \operatorname{int} \mathbb{R}_+^J$*: under Assumptions* L^r-*Sum*$-(i)$ *and* *Com*$-(i)$*, we have:*

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

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

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

(ii) $\lambda^{(i)} \in \partial\mathbb{R}_+^J$*: under Assumption* *Cra*$-(i)$*, we have:*

$$0 \leq \mathbb{P}_{\theta^*}\left(\hat{\theta}^n = \theta_i\right) \leq \frac{e^{-n\cdot\left[\lambda^{(i),\top}\mathbf{y}^{(i)} - \Lambda^{(i)}\left(\lambda^{(i)}\right)\right]}}{n^{1/2}} \cdot \left(d_0 + o\left(n^{-\delta}\right)\right)$$

for a certain $\delta > 0$*.*

(iii) $\lambda^{(i)} \in \partial\mathbb{R}_+^J$ *and the Hessian* $\mathbf{H}_f(\mathbf{0})$ *is negative definite: under Assumptions* L^r-*Sum*$-(i)$ *and* *Com*$-(i)$*, we have:*

$$0 \leq \mathbb{P}_{\theta^*}\left(\hat{\theta}^n = \theta_i\right) \leq \frac{e^{-n\cdot\left[\lambda^{(i),\top}\mathbf{y}^{(i)} - \Lambda^{(i)}\left(\lambda^{(i)}\right)\right]}}{n^{1/2}} \cdot \left(d_0 + O\left(n^{-1}\right)\right),$$

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

(II) $\mathbf{y}^{(i)}$ *belongs to the interior of a face of* $\mathbb{R}_+^J$*:*

(i) *under Assumption* *Cra*$-(i)$*, we have:*

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

for a certain $\delta > 0$*.*

(ii) *the Hessian* $\mathbf{H}_f(\mathbf{0})$ *is negative definite: under Assumptions* L^r-*Sum*$-(i)$ *and* *Com*$-(i)$*, we have:*

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

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

Proof. Most of the results used in the following require Assumptions **A1-4** from Iltis (1995); in particular, his Assumptions **A1** and **A2** coincide respectively with $L^r-\mathsf{Sum}-(i)$ and $\mathsf{Com}-(i)$. **A3** derives from Proposition 2.6 under Assumptions Obj, $\eta-\mathsf{Int}$, $\mathsf{Steep}-(i)$ and $\mathsf{Meet}-(i)$. 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-i) $\mathbf{y}^{(i)} = \mathbf{0}$ and $\lambda^{(i)} \in \operatorname{int} \mathbb{R}_+^J$: in particular, in the notation of Theorem 1.5 of Iltis (1995), Γ is contained in a cone with $\beta = 1$ and

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

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

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

(remark that $C^{**} > C^*$). This yields the results.
(I-ii) $\mathbf{y}^{(i)} = \mathbf{0}$ and $\lambda^{(i)} \in \partial\mathbb{R}_+^J$: in this case, the formula derives from Ney (1983, p. 165, Case (i)).
(I-iii) $\mathbf{y}^{(i)} = \mathbf{0}$, $\lambda^{(i)} \in \partial\mathbb{R}_+^J$ and the Hessian $\mathbf{H}_f(\mathbf{0})$ is negative definite (that is the level set of $\Lambda^{(i),*}$ is not flush with $\mathbb{R}_+^J$): this is the case of Theorem 1.4 of Iltis (1995);
(II-i) $\mathbf{y}^{(i)}$ belongs to the interior of a face of $\mathbb{R}_+^J$: in this case, the formula derives from Ney (1983, p. 165, Case (i)).
(II-ii) $\mathbf{y}^{(i)}$ belongs to the interior of a face of $\mathbb{R}_+^J$ and the Hessian $\mathbf{H}_f(\mathbf{0})$ is negative definite (that is $\mathbb{R}_+^J$ is not flush with the level set of $\Lambda^{(i),*}$): this is the case of Theorem 1.4 of Iltis (1995). □

Remark 2.8. *These results can be useful to derive an asymptotic expression for* $\mathbb{P}_{\theta^*}\left(\hat{\theta}^n \neq \theta_0\right)$. *Indeed, following the proof of Proposition 2.5, this probability can be written as:*

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

Under Assumptions Obj *and* $\eta-\mathsf{Int}$, *a full* LDP *holds and we can apply exact asymptotics for* LDP. *Moreover, the points of* $\arg\inf_{\mathbf{y} \in \operatorname{int}\left(\mathbb{R}_+^J\right)^c} \Lambda^*(\mathbf{y})$ *belong to* $(J-1)-$*dimensional faces of* $\left(\mathbb{R}_+^J\right)^c$ *and therefore, Proposition 2.7-**(II-ii)** can be applied. The obtained formulas are, however, overwhelmingly complicated but in the case* $H = 1$.

The next Proposition yields a portmanteau result on the asymptotic behavior: this is a generalization of Theorems 1.4 and 1.5 of Iltis (1995).

Proposition 2.8. *Under Assumptions Obj, $\eta-$Int, Meet$-$ (i) and Steep$-$ (i), we have:*

$$\mathbb{P}_{\theta^*}\left(\hat{\theta}^n = \theta_i\right) = n^{(J-2)/2} \cdot e^{-n\cdot\Lambda^{(i),*}} \cdot \frac{\phi_{\mathbf{0},\mathbf{H}_{\Lambda^{(i)}}\left(\lambda^{(i)}\right)}(\mathbf{0})}{\left\|\lambda^{(i)}\right\|} \cdot (1+o(1))$$
$$\cdot \int_{\mathbb{R}^{J-1}} e^{n\cdot\left\|\lambda^{(i)}\right\|\cdot\left[f\left(\mathbf{y}_{(J)}\right)-g\left(\mathbf{y}_{(J)}\right)\right]} \mu\left(\mathbb{R}\times \mathrm{d}\mathbf{y}_{(J)}\right).$$

Proof. Let $\bar{s}$ be an integer to be chosen later and define Γ_n as the truncation of Γ given by:

$$\Gamma_n \triangleq \left\{\mathbf{y}\,\middle|\,\mathbf{y}_{(J)} \in \Lambda, y_J \geq g\left(\mathbf{y}_{(J)}\right)\right\}$$

and

$$\Lambda \triangleq \left\{\mathbf{y}_{(J)}\,\middle|\,\left.\frac{\partial\Lambda^{(i),*}}{\partial y_J}\right|_{y_J=f\left(\mathbf{y}_{(J)}\right)} \cdot \left[g\left(\mathbf{y}_{(J)}\right)-f\left(\mathbf{y}_{(J)}\right)\right] \leq \frac{\bar{s}\ln n}{n}\right\}.$$

We have:

$$\mathbb{P}_{\theta^*}\left(\sum_{i=k}^{n}\mathbf{X}_k^{(i)} \in n\Gamma\right) = \mathbb{P}_{\theta^*}\left(\sum_{i=k}^{n}\mathbf{X}_k^{(i)} \in n\Gamma_n\right) + \mathbb{P}_{\theta^*}\left(\sum_{i=k}^{n}\mathbf{X}_k^{(i)} \in n\Gamma\backslash\Gamma_n\right).$$

Through Taylor's theorem, we can write:

$$\begin{aligned}
&\Lambda^{(i),*}\left(\mathbf{y}_{(J)}, g\left(\mathbf{y}_{(J)}\right)\right)\\
&= \Lambda^{(i),*}\left(\mathbf{y}_{(J)}, f\left(\mathbf{y}_{(J)}\right)\right) + \left.\frac{\partial\Lambda^{(i),*}}{\partial y_J}\right|_{y_J=f\left(\mathbf{y}_{(J)}\right)} \cdot \left[g\left(\mathbf{y}_{(J)}\right)-f\left(\mathbf{y}_{(J)}\right)\right]\\
&+ O\left(\left.\frac{\partial^2\Lambda^{(i),*}}{\left(\partial y_J\right)^2}\right|_{y_J=f\left(\mathbf{y}_{(J)}\right)} \cdot \left[g\left(\mathbf{y}_{(J)}\right)-f\left(\mathbf{y}_{(J)}\right)\right]^2\right),
\end{aligned}$$

where $g(\mathbf{0}) = f(\mathbf{0}) = 0$. For $\mathbf{y} \in \Gamma\backslash\Gamma_n$, we have:

$$\begin{aligned}
\Lambda^{(i),*}(\mathbf{y}) &\geq \Lambda^{(i),*}\left(\mathbf{y}_{(J)}, g\left(\mathbf{y}_{(J)}\right)\right)\\
&= \Lambda^{(i),*}\left(\mathbf{y}_{(J)}, f\left(\mathbf{y}_{(J)}\right)\right) + \left.\frac{\partial\Lambda^{(i),*}}{\partial y_J}\right|_{y_J=f\left(\mathbf{y}_{(J)}\right)} \cdot \left[g\left(\mathbf{y}_{(J)}\right)-f\left(\mathbf{y}_{(J)}\right)\right]\\
&+ O\left(\left.\frac{\partial^2\Lambda^{(i),*}}{\left(\partial y_J\right)^2}\right|_{y_J=f\left(\mathbf{y}_{(J)}\right)} \cdot \left[g\left(\mathbf{y}_{(J)}\right)-f\left(\mathbf{y}_{(J)}\right)\right]^2\right)\\
&\geq \Lambda^{(i),*} + \frac{(\bar{s}-1)\cdot\ln n}{n}
\end{aligned}$$

for sufficiently large n. This implies that $\Gamma\backslash\Gamma_n$ is contained in

$$V_{\Lambda^{(i),*}+\frac{(\bar{s}-1)\cdot\ln n}{n}} \triangleq \left\{\mathbf{y} \in \mathbb{R}^J\,\middle|\,\Lambda^{(i),*}(\mathbf{y}) \geq \Lambda^{(i),*} + \frac{(\bar{s}-1)\cdot\ln n}{n}\right\}.$$

By Theorem 1.3 in Iltis (1995), choosing $\bar{s} \geq 2 + \frac{(J-2)}{2}$, we have:

$$\mathbb{P}_{\theta^*}\left(\sum_{i=k}^{n} \mathbf{X}_k^{(i)} \in n\Gamma \backslash \Gamma_n\right) \leq \mathbb{P}_{\theta^*}\left(\sum_{i=k}^{n} \mathbf{X}_k^{(i)} \in nV_{\Lambda^{(i),*}+\frac{(\bar{s}-1)\cdot \ln n}{n}}\right)$$
$$= e^{-n\cdot\Lambda^{(i),*}} n^{-1/2} O\left(n^{-1}\right).$$

Now we apply Theorem 1.2 in Iltis (1995) to Γ_n:

$$\mathbb{P}_{\theta^*}\left(\sum_{i=k}^{n} \mathbf{X}_k^{(i)} \in n\Gamma_n\right) \tag{2.7}$$
$$= \int_{\Gamma_n} n^{J/2} e^{-n\cdot\Lambda^{(i),*}(\mathbf{v}(\mathbf{y}))} e^{-n\cdot\lambda(\mathbf{v}(\mathbf{y}))^{\mathsf{T}}(\mathbf{y}-\mathbf{v}(\mathbf{y}))}$$
$$\times \left[\phi_{\mathbf{0},\mathbf{H}_{\Lambda^{(i)}}[\lambda(\mathbf{v}(\mathbf{y}))]}\left(\sqrt{n}\left(\mathbf{y}-\mathbf{v}\left(\mathbf{y}\right)\right)\right) + O\left(n^{-1/2}\right)\right] \mu\left(\mathrm{d}\mathbf{y}\right),$$

where $\mathbf{v}\left(\mathbf{y}\right)$ is a $\mathcal{C}^1-$function of $\mathbf{y}$ in Γ such that for some $s \geq 0$, we have:

$$\Lambda^{(i),*}\left(\mathbf{v}\left(\mathbf{y}\right)\right) - \Lambda^{(i),*}\left(\mathbf{y}^{(i)}\right) + \lambda\left(\mathbf{v}\left(\mathbf{y}\right)\right)^{\mathsf{T}}\left(\mathbf{y}-\mathbf{v}\left(\mathbf{y}\right)\right) \geq -\frac{s\cdot\ln n}{n}, \tag{2.8}$$
$$e^{-n\cdot\Lambda^{(i),*}(\mathbf{v}(\mathbf{y}))} \cdot e^{-n\cdot\lambda(\mathbf{v}(\mathbf{y}))^{\mathsf{T}}(\mathbf{y}-\mathbf{v}(\mathbf{y}))} \leq n^s \cdot e^{-n\cdot\Lambda^{(i),*}\left(\mathbf{y}^{(i)}\right)}$$

and $\phi_{\mathbf{0},\mathbf{H}_{\Lambda^{(i)}}[\lambda(\mathbf{v}(\mathbf{y}))]}\left(\sqrt{n}\left(\mathbf{y}-\mathbf{v}\left(\mathbf{y}\right)\right)\right)$ is the Gaussian probability density function defined by:

$$\phi_{\mathbf{0},\mathbf{V}}\left(\mathbf{x}\right) = (2\pi)^{-J/2}\left(\det \mathbf{V}\right)^{-1/2} e^{-(1/2)\mathbf{x}^{\mathsf{T}}\mathbf{V}^{-1}\mathbf{x}}$$

evaluated at $\mathbf{V} = \mathbf{H}_{\Lambda^{(i)}}\left[\lambda\left(\mathbf{v}\left(\mathbf{y}\right)\right)\right]$ and $\mathbf{x} = \sqrt{n}\left(\mathbf{y}-\mathbf{v}\left(\mathbf{y}\right)\right)$. Moreover, it can be shown that $\mathbf{V}^{-1} = \mathbf{H}_{\Lambda^{*,(i)}}\left[\mathbf{v}\left(\mathbf{y}\right)\right]$.
Here we take:

$$\mathbf{y} \triangleq \left(\mathbf{y}_{(J)}, y_J\right),$$
$$\mathbf{v}\left(\mathbf{y}\right) \triangleq \left(\mathbf{y}_{(J)}, f\left(\mathbf{y}_{(J)}\right)\right),$$

and these vectors satisfy equation (2.8).
Now, (2.7) can be rewritten as:

$$\mathbb{P}_{\theta^*}\left(\sum_{i=1}^{n} \mathbf{X}_k^{(i)} \in n\Gamma_n\right) = n^{J/2}\cdot e^{-n\cdot\Lambda^{(i),*}} \cdot \phi_{\mathbf{0},\mathbf{H}_{\Lambda^{(i)}}[\lambda(\mathbf{0})]}\left(\mathbf{0}\right)\cdot\left(1+o\left(1\right)\right)$$
$$\cdot \int_{\Lambda}\int_{g\left(\mathbf{y}_{(J)}\right)}^{+\infty} e^{-n\cdot\lambda_J\left(\mathbf{y}_{(J)},f\left(\mathbf{y}_{(J)}\right)\right)\cdot\left(y_J-f\left(\mathbf{y}_{(J)}\right)\right)} \mu\left(\mathrm{d}y_J \times \mathrm{d}\mathbf{y}_{(J)}\right);$$

performing the y_J-integration:

$$\mathbb{P}_{\theta^*}\left(\sum_{i=1}^{n} \mathbf{X}_k^{(i)} \in n\Gamma_n\right) = n^{(J-2)/2}\cdot e^{-n\cdot\Lambda^{(i),*}} \cdot \phi_{\mathbf{0},\mathbf{H}_{\Lambda^{(i)}}[\lambda(\mathbf{0})]}\left(\mathbf{0}\right)\cdot\left(1+o\left(1\right)\right)$$
$$\cdot \int_{\Lambda} \frac{e^{n\cdot\lambda_J\left(\mathbf{y}_{(J)},f\left(\mathbf{y}_{(J)}\right)\right)\cdot\left[f\left(\mathbf{y}_{(J)}\right)-g\left(\mathbf{y}_{(J)}\right)\right]}}{\lambda_J\left(\mathbf{y}_{(J)}, f\left(\mathbf{y}_{(J)}\right)\right)} \mu\left(\mathbb{R}\times \mathrm{d}\mathbf{y}_{(J)}\right)$$

and expanding λ_J around $\mathbf{0}$, and remarking that $\lambda_J(\mathbf{0}) = \left\|\lambda^{(i)}\right\|$:

$$\mathbb{P}_{\theta^*}\left(\sum_{i=1}^{n} \mathbf{X}_k^{(i)} \in n\Gamma_n\right) = n^{(J-2)/2} \cdot e^{-n\cdot\Lambda^{(i),*}} \cdot \frac{\phi_{\mathbf{0},\mathbf{H}_{\Lambda^{(i)}}\left(\lambda^{(i)}\right)}(\mathbf{0})}{\left\|\lambda^{(i)}\right\|} \cdot (1+o(1))$$
$$\cdot \int_{\Lambda} e^{n\cdot\left\|\lambda^{(i)}\right\|\cdot\left[f\left(\mathbf{y}_{(J)}\right)-g\left(\mathbf{y}_{(J)}\right)\right]} \mu\left(\mathbb{R}\times \mathrm{d}\mathbf{y}_{(J)}\right).$$

□

Remark 2.9. *This formula can be used in calculations, once we realize that $g\left(\mathbf{y}_{(J)}\right)$ can be substituted by its development in Taylor's series and that $f\left(\mathbf{y}_{(J)}\right)$ is, in the present case, positively homogeneous of degree 1.*

2.3 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 Poor and Verdù (1995) and to the *Bayesian estimator* of Vajda (1971a, b). 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 2.4.1 for a definition) is the mode of the posterior distribution and is given by:

$$\begin{aligned}\check{\theta}^n &\triangleq \arg\max_{\theta\in\Theta} \ln \mathbb{P}\{\theta\,|\mathbf{Y}\}\\ &= \arg\max_{\theta\in\Theta}\left[\frac{1}{n}\sum_{i=1}^{n}\ln f_{Y_i}(y_i;\theta) + \frac{\ln\pi(\theta)}{n}\right].\end{aligned} \tag{2.9}$$

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 $(\Omega,\mathcal{A})$ indexed by a parameter θ ranging over a parameter space $\Theta = \{\theta_0,\theta_1,...,\theta_J\}$, of finite cardinality. Let $(\mathfrak{Y},\mathcal{Y})$ a measurable space and μ a positive $\sigma-$finite measure defined on $(\mathfrak{Y},\mathcal{Y})$

such that, for every $\theta \in \Theta$, $\mathbb{P}_\theta$ 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 $\mathbb{P}_{\theta^*} \in \mathcal{P}$, for θ^* not necessarily belonging to Θ.

The following two assumptions are standard requirements to show, through Jensen's inequality, that the likelihood function is asymptotically maximized at θ_0 only.

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

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

where $\mathbb{E}_{\theta^*}$ denotes the expectation taken under the true probability measure $\mathbb{P}_{\theta^*}$.

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

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

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

Proposition 2.1 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, a standard Wald's argument (see e.g. Lemma 2.2 in Newey and McFadden, 1994, p. 2124) shows that $\mathbb{E}_{\theta_0} \ln f_Y(Y;\theta_j)$ is maximized for $\theta_j = \theta_0$.

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 $\mathsf{L}_n(\theta)$ and $\mathbf{X}_k$ and $\mathbf{X}_k^{(i)}$ by the loglikelihood processes:

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

Also $M(\lambda)$ and $M^{(i)}(\lambda)$ are consequently defined. Then, for $J = 1$, equation (2.5) hold under Assumptions Hom, L^1-Dom and Id. Proposition 2.3 holds under Assumptions Hom, L^1-Dom, Id and Min. However, if $\theta^* = \theta_0$, then the solution to $0 = \left(\Lambda^{(1)}\right)'(\eta)$ exists, is positive and belongs to $]0,1[$. Moreover, Propositions 2.4 and 2.5 hold when Assumption Obj is substituted by Assumptions Hom, L^1-Dom and Id.

When the model is correctly specified (that is $\theta^* = \theta_0$), 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:

$$M^{(i)}(\lambda) \triangleq \mathbb{E}_{\theta_0}\left[e^{\sum_{j=0,\ldots,J,j\neq i}\lambda_j\cdot[\ln f_Y(Y;\theta_i)-\ln f_Y(Y;\theta_j)]}\right]$$
$$= \int f_Y(y;\theta_i)^{\sum_{j=0,\ldots,J,j\neq i}\lambda_j} \cdot \prod_{j=1,\ldots,J,j\neq i} f_Y(y;\theta_j)^{-\lambda_j}\cdot f_Y(y;\theta_0)^{1-\lambda_0}\,\mu(\mathrm{d}y). \tag{2.10}$$

Therefore, in this case, the moment generating function $M^{(i)}(\lambda)$ reduces to the so-called Hellinger transform $\mathsf{H}_\gamma(\theta_0,...,\theta_J)$ (see Le Cam and Yang, 1990, p. 27) for a certain linear transformation of λ in γ:

$$\mathsf{H}_\gamma(\theta_0,...,\theta_J) \triangleq \int \prod_{j=0}^{J}\left[\mathbb{P}_{\theta_j}(\mathrm{d}y)\right]^{\gamma_j} = \int\left[\prod_{j=0}^{J} f_Y(y;\theta_j)^{\gamma_j}\right]\mu(\mathrm{d}y), \qquad \sum_{j=0}^{J}\gamma_j = 1.$$

Moreover, due to its convexity, $\mathsf{H}_\gamma(\theta_0,...,\theta_J)$ 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., Jacod and Shiryaev, 1987, p. 192, Shiryaev, 1995, p. 363):

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

the *Hellinger arc* (see Genon-Catalot and Picard, 1993), that is the set of measures defined by:

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

the *error function of order* $u \in]0,1[$ (Aczél and Daróczy, 1975, p. 208):

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

the $\gamma-$*deviation of order* $\gamma\in[-1,2]$ (Amari, 1985) on the space of finite positive measures:

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

and the *Chernoff index* (Chernoff, 1952, pp. 500, 507, Chernoff, 1956, p. 17, Cover and Thomas, 1991, p. 314)

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

Proposition 2.6 holds if Assumption Obj is substituted by Assumptions Hom, $L^1-\mathsf{Dom}$ and Id. However, the fact that $\mathrm{int}\left(\mathbb{R}_+^J\cap\mathcal{S}^{(i)}\right)\neq\varnothing$ can be proved

showing that $\mathbf{0} \in \operatorname{int} \mathcal{S}^{(i)}$. This is equivalent to the existence, for $j = 1, ..., J, j \neq i$, of two sets A_j^* and A_j^{**} of positive $\mu-$measure and included in the support of Y, such that for $y_j^* \in A_j^*$ and $y_j^{**} \in A_j^{**}$:

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

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 2.7 holds if Assumption Obj is substituted by Assumptions Hom, L^1-Dom and Id.

In order to derive the asymptotic distribution of Bayes estimators, we consider equation (2.9) and we let $\ln \pi^{(i)} \triangleq \left[\ln \frac{\pi(\theta_i)}{\pi(\theta_j)}\right]_{j=0,\ldots,J,j\neq i}$. Then, we can write:

$$\begin{aligned}\mathbb{P}_{\theta_0}\left(\check{\theta}^n = \theta_i\right) &= \mathbb{P}_{\theta_0}\left(\sum_{k=1}^n \mathbf{X}_k^{(i)} + \ln \pi^{(i)} \in \operatorname{int} \mathbb{R}_+^J\right) \\ &= \mathbb{P}_{\theta_0}\left(\sum_{k=1}^n \mathbf{X}_k^{(i)} \in \prod_{j=0,\ldots,J,j\neq i}\left(\ln \frac{\pi\left(\theta_i\right)}{\pi\left(\theta_j\right)}, +\infty\right)\right),\end{aligned}$$

and we can use the previous large deviations formulas, simply changing the set over which the inf is taken. However, care is needed since this large deviations result for $\mathbb{P}_{\theta_0}\left(\check{\theta}^n = \theta_i\right)$ only holds under the hypothesis:

$$\mathbb{E}_{\theta_0} \mathbf{X}_k^{(i)} + \frac{1}{n} \cdot \ln \pi^{(i)} \in \operatorname{int}\left(\mathbb{R}_+^J\right)^c.$$

The similarity of these formulas with the corresponding ones for a Neyman-Pearson test (in the case $J = 1$) 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.

2.4 Optimality and Efficiency of the MLE

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 $\theta_0 \in \Theta$: this will be reflected in the probabilities and expectation that will be evaluated for $\theta^* = \theta_0$. Indeed, efficiency statement for misspecified models are quite difficult to interpret.

Since the variance seen as a measure of efficiency seems to be well suited only in cases in which the limiting distribution is Gaussian, in Section 2.4.1, we present some alternative risk functions; then we derive in Section 2.4.2 some lower bounds for these risk functions and we prove in Section 2.4.3 some optimality and efficiency results for the Bayes and ML estimators.

2.4.1 Risk Functions

An interesting problem concerns the choice of a measure of efficiency for the MLE in discrete parameter models: in his seminal paper, Hammersley (1950) derives a generalization of Cramér-Rao inequality for the variance that is also valid when the parameter space is countable. The same inequality has been rediscovered, in slightly more generality, by Chapman and Robbins (1951) and by Blyth and Roberts (1972). Therefore, we will consider the following cost and risk functions:

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

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 Hammersley (1950), we consider also a different cost function $\mathcal{C}_2(\theta, \theta_0)$:

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

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

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

Moreover, Khan (1973) 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. However, in many cases, a more general loss function can be considered, as suggested in Gouriéroux and Monfort (1995, Vol. 1, p. 51) for multiple tests:

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

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

weighted probability of missclassification:

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

The $a_j\left(\theta_0\right)$'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 $\mathcal{R}_2\left(\tilde{\theta}^n, \theta_0\right)$ as:

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

If $\pi\left(\theta_j\right) = \left(J+1\right)^{-1}$ we define $\mathbb{P}_e \triangleq r_2\left(\tilde{\theta}^n, \pi\right)$ as the *average probability of error*. Remark that this is indeed the measure of error used by Vajda (1971a, b).

2.4.2 Information Inequalities

This Section contains a number of lower bounds for the previously introduced risk functions and in particular for the risk function $\mathcal{R}_2\left(\tilde{\theta}^n, \theta_0\right)$ corresponding to the zero-one loss. In the specific case of discrete parameters, these generalize and unify the lower bounds proposed by Hammersley (1950), Chapman and Robbins (1951), Kester and Kallenberg (1986) and Hall (1989). 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:

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

First of all, a lower bound is proved and then, we obtain a minimax version of the same result. Then, from these results, we derive lower bounds for the MSE, for the weighted probability of missclassification and for the Bayes risk.

A Lower Bound for the Probability of Missclassification The Proposition of this Section is intended to play the role of Cramér-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 (1972) (see also Le Cam, 1973, for

a related technique). Here, Stein's Lemma is applied as in Theorem 9.2 in Ibragimov and Has'minskii (1981, p. 96), taking into account the fact that the parameter space is made up of more than two points. Moreover, the equivalent representation (2.12) is obtained.

Proposition 2.9. *Under Assumptions* Hom *and* Id*, for a strongly consistent estimator* $\tilde{\theta}^n$*:*

$$\begin{aligned}\mathcal{R}_2\left(\tilde{\theta}^n,\theta_0\right) &= \mathbb{P}_{\theta_0}\left(\tilde{\theta}^n \neq \theta_0\right)\\ &\geq \sup_{\theta_1\in\Theta\setminus\{\theta_0\}} \exp\left\{n\cdot\mathbb{E}_{\theta_1}\ln\left(\frac{f_Y(Y;\theta_0)}{f_Y(Y;\theta_1)}\right)+o(n)\right\} \qquad (2.11)\\ &= \sup_{\theta_1\in\Theta\setminus\{\theta_0\}}\sup_{u>1}\left[\mathbb{E}_{\theta_0}\left(\frac{f_Y(Y;\theta_1)}{f_Y(Y;\theta_0)}\right)^u\right]^{-\frac{n+o(n)}{u-1}}. \qquad (2.12)\end{aligned}$$

Proof. First of all, we prove (2.11). We suppose that $\int \ln\frac{f_Y(y;\theta_1)}{f_Y(y;\theta_0)} f_Y(y;\theta_1)\,\mu(\mathrm{d}y) < \infty$, otherwise the inequality is trivial. Then, for any $\theta_1 \in \Theta\setminus\{\theta_0\}$, we apply Lemma 3.4.7 in Dembo and Zeitouni (1998, p. 94) with $\alpha_n = \mathbb{P}_{\theta_1}\left\{\hat{\theta}^n \neq \theta_1\right\}$ and $\beta_n = \mathbb{P}_{\theta_0}\left\{\hat{\theta}^n \neq \theta_0\right\}$; since $\hat{\theta}^n$ is strongly consistent, α_n is ultimately less than any $\varepsilon > 0$ and the bound holds.

Now, we show the equivalence of (2.11) and (2.12). We start showing that

$$\exp\left\{n\cdot\mathbb{E}_{\theta_1}\ln\left(\frac{f_Y(Y;\theta_0)}{f_Y(Y;\theta_1)}\right)\right\} \geq \frac{1}{\left[\mathbb{E}_{\theta_0}\left(\frac{f_Y(Y;\theta_1)}{f_Y(Y;\theta_0)}\right)^u\right]^{\frac{n}{u-1}}}. \qquad (2.13)$$

Indeed, for $v > 0$, we have:

$$\mathbb{E}_{\theta_1}\ln\left(\frac{f_Y(Y;\theta_1)}{f_Y(Y;\theta_0)}\right) \leq \frac{1}{v}\cdot\ln\mathbb{E}_{\theta_1}\left(\frac{f_Y(Y;\theta_1)}{f_Y(Y;\theta_0)}\right)^v = \frac{1}{v}\cdot\ln\mathbb{E}_{\theta_0}\left(\frac{f_Y(Y;\theta_1)}{f_Y(Y;\theta_0)}\right)^{v+1}$$

and this yields the result for $u = v+1 > 1$. This implies that if $\mathbb{E}_{\theta_1}\ln\left(\frac{f_Y(Y;\theta_1)}{f_Y(Y;\theta_0)}\right) = \infty$, then the equality holds.

Now we prove the converse of equation (2.13). Setting $v \triangleq u-1$, we get

$$\mathbb{E}_{\theta_0}\left(\frac{f_Y(Y;\theta_1)}{f_Y(Y;\theta_0)}\right)^u = \mathbb{E}_{\theta_1}\left(\frac{f_Y(Y;\theta_1)}{f_Y(Y;\theta_0)}\right)^v = \mathbb{E}_{\theta_1}\exp\left\{v\cdot\ln\left(\frac{f_Y(Y;\theta_1)}{f_Y(Y;\theta_0)}\right)\right\}$$

and we remark that since

$$\sup_{u>1}\mathbb{E}_{\theta_0}\left(\frac{f_Y(Y;\theta_1)}{f_Y(Y;\theta_0)}\right)^{\frac{n}{u-1}} = \sup_{v>0}\mathbb{E}_{\theta_0}\left(\frac{f_Y(Y;\theta_1)}{f_Y(Y;\theta_0)}\right)^{\frac{n}{v}}, \qquad (2.14)$$

as long as $n \to \infty$, the value v^* at which the supremum is attained converges to 0. Finally, we try to estimate $\mathbb{E}_{\theta_1}\left(\frac{f_Y(Y;\theta_1)}{f_Y(Y;\theta_0)}\right)^v$. Since

$$O\left[v^2\cdot\left|\ln^2\left(\frac{f_Y(y;\theta_1)}{f_Y(y;\theta_0)}\right)\right|\right] = \left|\exp\left\{v\cdot\ln\left(\frac{f_Y(y;\theta_1)}{f_Y(y;\theta_0)}\right)\right\} - 1 - v\cdot\ln\left(\frac{f_Y(y;\theta_1)}{f_Y(y;\theta_0)}\right)\right|,$$

then

$$
\begin{aligned}
& O\left[v^2 \cdot \mathbb{E}_{\theta_1}\left|\ln^2\left(\frac{f_Y(Y;\theta_1)}{f_Y(Y;\theta_0)}\right)\right|\right] \\
& = \left|\mathbb{E}_{\theta_1}\exp\left\{v \cdot \ln\left(\frac{f_Y(Y;\theta_1)}{f_Y(Y;\theta_0)}\right)\right\} - 1 - v \cdot \mathbb{E}_{\theta_1}\ln\left(\frac{f_Y(Y;\theta_1)}{f_Y(Y;\theta_0)}\right)\right|.
\end{aligned}
$$

Therefore:

$$
\begin{aligned}
& \mathbb{E}_{\theta_1}\exp\left\{v \cdot \ln\left(\frac{f_Y(Y;\theta_1)}{f_Y(Y;\theta_0)}\right)\right\} \\
& \leq 1 + v \cdot \mathbb{E}_{\theta_1}\ln\left(\frac{f_Y(Y;\theta_1)}{f_Y(Y;\theta_0)}\right) + O\left[v^2 \cdot \mathbb{E}_{\theta_1}\left|\ln^2\left(\frac{f_Y(Y;\theta_1)}{f_Y(Y;\theta_0)}\right)\right|\right] \\
& \leq \exp\left\{v \cdot \mathbb{E}_{\theta_1}\ln\left(\frac{f_Y(Y;\theta_1)}{f_Y(Y;\theta_0)}\right) \cdot (1 + o(1))\right\}.
\end{aligned}
$$

This establishes the result. □

Remark 2.10. ***(i)*** *Remark that this inequality only holds for consistent estimators, while the one of Proposition 2.10 holds for any estimator.*
(ii) *Taking $u = 2$ in (2.12) yields a lower bound similar to the one outlined by Hammersley (1950) and Chapman and Robbins (1951) for the variance, and to the minimax bound of Hall (1989). Remark that the Cramér-Rao lower bound can be derived from the previous ones under certain hypotheses of regularity (see Chapman and Robbins, 1951, for a proof).*
(iii) *The* inaccuracy rate *considered by Kester and Kallenberg (1986) can be defined as (see also Bahadur, 1960, and Sievers, 1978, for related concepts):*

$$
\begin{aligned}
e\left(\varepsilon, \theta_0, \tilde{\theta}^n\right) & \triangleq -\liminf_{n\to\infty}\frac{1}{n} \cdot \ln \mathbb{P}_{\theta_0}\left(\left\|\tilde{\theta}^n - \theta_0\right\| > \varepsilon\right) \\
& = -\liminf_{n\to\infty}\frac{1}{n} \cdot \ln \mathbb{P}_{\theta_0}\left(\tilde{\theta}^n \neq \theta_0\right) = -\liminf_{n\to\infty}\frac{1}{n} \cdot \ln \mathcal{R}_2\left(\tilde{\theta}^n, \theta_0\right),
\end{aligned}
$$

for any ε small enough ($\varepsilon < \min_{\theta_1 \in \Theta \setminus \{\theta_0\}} \|\theta_1 - \theta_0\|$), this leads to the following upper bound:

$$
\begin{aligned}
e\left(\varepsilon, \theta_0, \tilde{\theta}^n\right) & \leq \inf_{\theta_1 \in \Theta \setminus \{\theta_0\}} \mathbb{E}_{\theta_1}\ln\left(\frac{f_Y(Y;\theta_1)}{f_Y(Y;\theta_0)}\right) \\
& = \inf_{\theta_1 \in \Theta \setminus \{\theta_0\}} \inf_{u>1} \frac{1}{u-1} \cdot \ln\left[\mathbb{E}_{\theta_0}\left(\frac{f_Y(Y;\theta_1)}{f_Y(Y;\theta_0)}\right)^u\right].
\end{aligned}
$$

Remark in particular the presence of the Kullback-Leibler divergence and of the error function of order u quoted in Section 2.3. Since the bound of Proposition 1.1 of Kester and Kallenberg (1986) coincides with the one implied by equation (2.11) and the attainment of this limit, that is Chapman-Robbins optimality, coincides with IR*−optimality (see Kester and Kallenberg, 1986, p. 649), this enables us to use their results in order to prove that our* MLE *does not attain the Chapman-Robbins lower bound under the risk function $\mathcal{R}_2$ (Proposition 2.11).*

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

$$\begin{aligned}\sup_{\theta_0\in\Theta}\mathcal{R}_2\left(\tilde{\theta}^n,\theta_0\right) &= \sup_{\theta_0\in\Theta}\mathbb{P}_{\theta_0}\left(\tilde{\theta}^n\neq\theta_0\right) \geq \sup_{\theta_1\in\Theta\setminus\{\theta_0\}}\sup_{\theta_0\in\Theta}\exp\left\{-n\cdot\Lambda^*(0)+o_{\inf}(n)\right\}\\ &= \sup_{\theta_1\in\Theta\setminus\{\theta_0\}}\sup_{\theta_0\in\Theta}\left[\inf_{1>u>0}\int f_Y(y;\theta_1)^u\cdot f_Y(y;\theta_0)^{1-u}\,\mu(\mathrm{d}y)\right]^n\cdot e^{o_{\inf}(n)}.\end{aligned} \tag{2.15}$$

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

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

for an arbitrary couple of different alternatives θ_0 and θ_1 in Θ. Then we can use Chernoff's Bound (Dembo and Zeitouni, 1998, p. 93): the second expression derives from the equality $\Lambda^*(0)=-\inf_{\lambda\in\mathbb{R}}\Lambda(\lambda)$. □

Remark 2.11. ***(i)*** *Following Hall (1989, Theorem 3.1, p. 50) and using Hölder inequality instead of Cauchy-Schwarz inequality, we could obtain the information inequality:*

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

more similar to the one of Proposition 2.9. Especially using the first representation, it is not difficult to see that the lower bound of Proposition 2.10 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 Bahadur, 1960, and Korostelev and Leonov, 1994) analogous to Chernoff's Bound, thus providing a minimax version of the result in Remark 2.10* ***(iii)****.*
(iii) *Other methods to derive similar minimax inequalities are Assouad's Lemma and Fano's Inequality: however, in the present case they do not allow us to obtain tight bounds. 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 varies. Clearly, this cannot be done in the present case.*

Lower Bounds for the MSE The missclassification probability has often an exponential limit behavior (see Bahadur, 1960, Lemma 5.2, p. 245, Fu, 1982, p. 762), in the sense that there exists ρ 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, this result does not contrast with $\sqrt{n}-$consistency. Here we show that 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 $\mathcal{R}_1\left(\tilde{\theta}^n, \theta_0\right)$. The MSE of a generic estimator $\tilde{\theta}^n$ can be shown to be:

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

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

$$\mathbb{E}_{\theta_0}\left(\tilde{\theta}^n - \theta_0\right)^2 = K_n \cdot \mathbb{P}_{\theta_0}\left(\tilde{\theta}^n \neq \theta_0\right), \tag{2.16}$$

where K_n is a function of n belonging to the set $\left[\inf_{\theta_1 \in \Theta \setminus \{\theta_0\}} \left|\theta_1 - \theta_0\right|^2,\ \sup_{\theta_1 \in \Theta \setminus \{\theta_0\}} \left|\theta_1 - \theta_0\right|^2\right]$.

Moreover, using equation (2.16), the lower bound of Proposition 2.9 can be expressed as:

$$\begin{aligned}
\mathsf{MSE}\left(\tilde{\theta}^n\right) &= K_n \cdot \mathbb{P}_{\theta_0}\left(\tilde{\theta}^n \neq \theta_0\right) \cdot (1 + o(1)) \\
&\geq \inf_{\theta_1 \in \Theta \setminus \{\theta_0\}} \left|\theta_1 - \theta_0\right|^2 \cdot \sup_{\theta_1 \in \Theta \setminus \{\theta_0\}} \exp\left\{n \cdot \mathbb{E}_{\theta_1} \ln\left(\frac{f_Y(Y;\theta_0)}{f_Y(Y;\theta_1)}\right) + o_{\inf}(n)\right\} \\
&= \inf_{\theta_1 \in \Theta \setminus \{\theta_0\}} \left|\theta_1 - \theta_0\right|^2 \cdot \sup_{\theta_1 \in \Theta \setminus \{\theta_0\}} \sup_{u>1} \frac{1}{\left[\mathbb{E}_{\theta_0}\left(\frac{f_Y(Y;\theta_1)}{f_Y(Y;\theta_0)}\right)^u\right]^{\frac{n + o_{\inf}(n)}{u-1}}}
\end{aligned}$$

where $o_{\inf}(n)$ is a function such that $\liminf_{n\to\infty} \frac{o_{\inf}(n)}{n} = 0$, while the minimax lower bound of Proposition 2.10 can be written as (see also Hall, 1989, p. 51 for a comparison):

$$\begin{aligned}
&\sup_{\theta_0 \in \Theta}\left[\mathsf{MSE}\left(\tilde{\theta}^n\right)\right]^{1/2} \geq \sup_{\theta_1 \in \Theta \setminus \{\theta_0\}} \sup_{\theta_0 \in \Theta} \sup_{u>1} \frac{\left|\theta_0 - \theta_1\right|}{2} \cdot \exp\left\{-n \cdot \Lambda^*(\mathbf{0}) + o(n)\right\} \\
&= \sup_{\theta_1 \in \Theta \setminus \{\theta_0\}} \sup_{\theta_0 \in \Theta} \frac{\left|\theta_0 - \theta_1\right|}{2} \cdot \left[\inf_{1>u>0} \int f_Y(y;\theta_1)^u \cdot f_Y(y;\theta_0)^{1-u}\, \mu(\mathrm{d}y)\right]^{-n} \cdot e^{o(n)}.
\end{aligned}$$

Lower Bounds for the Weighted Probability of Missclassification It is simple to give a lower bound to the weighted probability of missclassification $\mathcal{R}_3\left(\tilde{\theta}^n, \theta_0\right) \triangleq \sum_{j=1}^{J} a_j\left(\theta_0\right) \cdot \mathbb{P}_{\theta_0}\left\{\tilde{\theta}^n = \theta_j\right\}$ under Assumption Hom, using the inequalities:

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

Indeed, these imply:

$$\max_{\theta_j \in \Theta} \ln a_j\left(\theta_0\right) + \ln \mathcal{R}_2\left(\tilde{\theta}^n, \theta_0\right) \geq \ln \mathcal{R}_3\left(\tilde{\theta}^n, \theta_0\right) \geq \min_{\theta_j \in \Theta} \ln a_j\left(\theta_0\right) + \ln \mathcal{R}_2\left(\tilde{\theta}^n, \theta_0\right) - \ln J;$$

and

$$\lim_{n \to \infty} \frac{1}{n} \ln \mathcal{R}_3\left(\tilde{\theta}^n, \theta_0\right) = \lim_{n \to \infty} \frac{1}{n} \ln \mathcal{R}_2\left(\tilde{\theta}^n, \theta_0\right).$$

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:

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

Under Assumptions Hom and Bay, the lower bound in logarithmic asymptotics can be obtained from the following inequalities:

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

leading to

$$\lim_{n \to \infty} \frac{1}{n} \ln r_2\left(\tilde{\theta}^n, \pi\right) = \lim_{n \to \infty} \frac{1}{n} \ln \max_{\theta_0 \in \Theta} \mathcal{R}_2\left(\tilde{\theta}^n, \theta_0\right). \tag{2.17}$$

Then, from Proposition 2.10 we obtain:

$$\begin{aligned}\liminf_{n \to \infty} \frac{1}{n} \ln r_2\left(\tilde{\theta}^n, \pi\right) &\geq -\inf_{\theta_1 \in \Theta \setminus \{\theta_0\}} \inf_{\theta_0 \in \Theta} \Lambda^*(0) \\ &= \max_{\theta_1 \in \Theta \setminus \{\theta_0\}} \max_{\theta_0 \in \Theta} \inf_{1 > u > 0} \left[\ln \int f_Y\left(y; \theta_1\right)^u \cdot f_Y\left(y; \theta_0\right)^{1-u} \mu(\mathrm{d}y)\right].\end{aligned}$$

Clearly this bound is independent of the prior distribution π 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 for a different risk function (see e.g. Gill and Levit, 1995).

2.4.3 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 2.1. *The estimator $\tilde{\theta}^n$ is* efficient with respect to *(wrt)* $\tilde{\Theta}$ and *wrt* $\mathcal{R}$ at θ_0 *if:*

$$\mathcal{R}\left(\tilde{\theta}^n, \theta_0\right) \leq \mathcal{R}\left(\bar{\theta}^n, \theta_0\right), \qquad \forall \bar{\theta}^n \in \tilde{\Theta}. \tag{2.18}$$

The estimator $\tilde{\theta}^n$ is superefficient *wrt* $\tilde{\Theta}$ and *wrt* $\mathcal{R}$ at θ_0 *if there is a value of θ_0 such that:*

$$\mathcal{R}\left(\tilde{\theta}^n, \theta_0\right) < \mathcal{R}\left(\bar{\theta}^n, \theta_0\right), \qquad \forall \bar{\theta}^n \in \tilde{\Theta}.$$

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

$$\sup_{\theta_0 \in \Theta} \mathcal{R}\left(\tilde{\theta}^n, \theta_0\right) \leq \sup_{\theta_0 \in \Theta} \mathcal{R}\left(\bar{\theta}^n, \theta_0\right), \qquad \forall \bar{\theta}^n \in \tilde{\Theta}. \tag{2.19}$$

The estimator $\tilde{\theta}^n$ is asymptotically CR−efficient *wrt* $\mathcal{R}$ at θ_0 *if it attains the Chapman-Robbins lower bound of Section 2.4.2 at θ_0 (say* CR$-\mathcal{R}(\theta_0)$*) in the asymptotic form:*

$$0 = \liminf_{n \to \infty} \frac{1}{n}\left[\ln \mathrm{CR}{-}\mathcal{R}(\theta_0) - \ln \mathcal{R}\left(\tilde{\theta}^n, \theta_0\right)\right].$$

The estimator $\tilde{\theta}^n$ is asymptotically CR−superefficient *wrt* $\mathcal{R}$ at θ_0 *if it has a risk function smaller than the Chapman-Robbins lower bound of Section 2.4.2 at θ_0 in the asymptotic form:*

$$0 < \liminf_{n \to \infty} \frac{1}{n}\left[\ln \mathrm{CR}{-}\mathcal{R}(\theta_0) - \ln \mathcal{R}\left(\tilde{\theta}^n, \theta_0\right)\right].$$

The estimator $\tilde{\theta}^n$ is asymptotically minimax CR−efficient *wrt* $\mathcal{R}$ *if it attains the minimax Chapman-Robbins lower bound of Section 2.4.2 (say* CR$-\mathcal{R}_{\max}$*) in the asymptotic form:*

$$0 = \liminf_{n \to \infty} \frac{1}{n}\left[\ln \mathrm{CR}{-}\mathcal{R}_{\max} - \ln \sup_{\theta_0 \in \Theta} \mathcal{R}\left(\tilde{\theta}^n, \theta_0\right)\right].$$

Remark 2.12. *As in Remark 2.10* ***(iii)****, 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 2.1. 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 2.1. *Let us consider the estimation of the mean of a Gaussian distribution $\mathcal{N}\left(\alpha, \sigma^2\right)$ whose variance is known: we suppose that the parameter space is $\{-\alpha, \alpha\}$, where α and σ^2 are 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:

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

Proposition 2.3 allows also for recovering the right convergence rate. Indeed, we have:

$$\Lambda^{(1)}(\lambda) = -\lambda(1-\lambda) \cdot \tfrac{2\alpha^2}{\sigma^2}, \qquad \eta = \tfrac{1}{2},$$
$$\left(\Lambda^{(1)}\right)'(0) = -\tfrac{2\alpha^2}{\sigma^2}, \qquad \left(\Lambda^{(1)}\right)''(0) = \tfrac{2\alpha^2}{\sigma^2};$$

therefore, we get:

$$\mathbb{P}_{\theta_0}\left(\hat{\theta}^n \neq \mu_0\right) = \mathbb{P}_{\theta_0}\left(\hat{\theta}^n = -\alpha\right) = \frac{e^{-\frac{n\alpha^2}{2\sigma^2}}}{\sqrt{2\pi n}} \frac{\sigma}{\alpha} \cdot (1 + o(1)).$$

On the other hand, the lower bound of Proposition 2.9 yields:

$$\mathbb{P}_{\theta_0}\left(\hat{\theta}^n \neq \theta_0\right) \geq \exp\left\{-\frac{2n \cdot \alpha^2}{\sigma^2} \cdot (1 + o(1))\right\},$$

and the lower bound of Proposition 2.10 yields:

$$\sup_{\theta_0 \in \{-\alpha, \alpha\}} \mathbb{P}_{\theta_0}\left(\hat{\theta}^n \neq \theta_0\right) \geq \exp\left\{-\frac{n \cdot \alpha^2}{2\sigma^2} \cdot (1 + o_{\inf}(1))\right\}.$$

Therefore, the MLE attains the minimax lower bound but not the classical one. In the following, we will show that estimators can be more pointwise efficient than the MLE; consider the estimator defined by:

$$\tilde{\theta}^n = \begin{cases} \theta_0 & \text{if } \mathsf{L}_n(\theta_0) \geq \mathsf{L}_n(\theta_1) + k_n, \\ \theta_1 & \text{else.} \end{cases}$$

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

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

If we take $k_n \triangleq k \cdot n$, we have consistency if:

$$2\left(\frac{\alpha}{\sigma}\right)^2 > k > -2\left(\frac{\alpha}{\sigma}\right)^2. \tag{2.20}$$

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

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

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. Indeed, $\mathbb{P}_{\theta_0}\left(\tilde{\theta}^n \neq \theta_0\right)$ 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 (2.20). If this estimator is pointwise more efficient than the MLE under θ_0, then under θ_1, this estimator has risk $\mathcal{R}_2\left(\tilde{\theta}^n, \theta_1\right)$ given by:

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

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

Optimality wrt the Information Inequalities In this Subsection, we will show that the MLE does not attain the Chapman-Robbins lower bound in the

[3] It is enough to take k such that:

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

form of Proposition 2.9 but that it attains the minimax form of Proposition 2.10.

Therefore, the situation described in Example 2.1 is general, for it is possible to show that the MLE is generally inefficient with respect to the lower bounds exposed in Proposition 2.9: since we have shown in Remark 2.10 **(iii)** that IR−optimality and CR−efficiency are equivalent in this setting, we use the method developed by Kester and Kallenberg (1986) to assess IR−optimality; therefore, in the proof of the following Proposition, KK stands for this paper.

Proposition 2.11. *Under Assumptions Hom, L^1-Dom and Id:*
(a) *the MLE is not asymptotically* CR*−efficient wrt* $\mathcal{R}_2$ *at* θ_0*;*
(b) *the MLE is asymptotically minimax* CR*−efficient wrt* $\mathcal{R}_2$*.*

Proof. In order to prove the first statement, we apply Lemma 2.4 (KK, 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

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

does not belong to $\mathcal{P}$. Therefore, from Lemma 2.4 **(iii)** in KK, there exists $\theta_1', \theta_2' \in \Theta$ such that equation (2.12) in KK holds and, as a consequence of Lemma 2.4 **(i)** in KK, 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':

$$\liminf_{n\to\infty} \frac{1}{n} \ln \mathbb{P}_{\theta_1'} \left\{ \left| \hat{\theta}^n - \theta_1' \right| > \varepsilon \right\} > \sup_{\theta \in \Theta, \left|\theta - \theta_1'\right| > \varepsilon} \mathbb{E}_{\theta} \ln \left(\frac{f_Y(Y;\theta_1')}{f_Y(Y;\theta)} \right),$$

and this implies that the Chapman-Robbins bound is not attained at θ_1'.
The second statement follows easily from the results of Kanaya and Han (1995, Theorem 2) on $\lim_{n\to\infty} \frac{1}{n} \ln r_2\left(\tilde{\theta}^n, \pi\right)$, using equation (2.17). Indeed, the MLE attains the lower bound (2.15) and is therefore asymptotically minimax efficient. □

Remark 2.13. ***(i)*** *Indeed, to obtain part* ***(a)*** *of the Proposition, the hypothesis of homogeneity of the probability measures is not even necessary, but it is used in order to derive part* ***(b)****: it can be removed along the lines of KK.*
(ii) *Using the asymptotic equalities of Section 2.4.2 for* $\mathcal{R}_1$ *and* $\mathcal{R}_3$*, it is trivial to extend the results of this Proposition to these risk functions.*

Optimality of the Bayes and ML 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 2.12. *Under Hom, L^1-Dom, Id and Bay, the Bayes risk $r_2\left(\tilde{\theta}^n,\pi\right)$ (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.

Proposition 2.13. *Under 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 $\mathbb{P}_e$.*

Proof. 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 $\tilde{\theta}^n$ coincides with the MLE $\hat{\theta}^n$. Therefore, respectively from Lemma 2.10 and Proposition 6.3 in Robert (1992), $\hat{\theta}^n$ is minimax and admissible (see also Property 2.2 in Gouriéroux and Monfort, 1995, Vol. 1, p. 60). The fact that the MLE minimizes the average probability of error derives from Proposition 2.12. □

Remark 2.14. *If we change the loss function, the result does not necessarily transpose. As an example, using the asymptotic equalities of Section 2.4.2 for $\mathcal{R}_1$ and $\mathcal{R}_3$, it is simple to show that with respect to the risk functions $\mathcal{R}_1\left(\tilde{\theta}^n,\theta_0\right)$ and $\mathcal{R}_3\left(\tilde{\theta}^n,\theta_0\right)$, the estimator is only asymptotically minimax efficient (see Khan, 1973, for a similar result in a related setup).*

The Evil of Superefficiency Since it was discovered by Hodges, the problem of superefficiency has been dealt with extensively in regular statistical problems (see e.g., Le Cam, 1953 and van der Vaart, 1997). However, these proofs do not transpose to discrete parameter estimation problems, since they are based on the equivalence of 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. Therefore, the following Proposition yields a solution to the superefficiency problem and shows that no estimator can have a convergence rate faster than the one prescribed by 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 $\theta_1 \in \Theta$.

Proposition 2.14. *Under Assumptions Hom, L^1-Dom and Id:*
(i) *no estimator $\tilde{\theta}^n$ is superefficient wrt the MLE $\hat{\theta}^n$ and wrt $\mathcal{R}_2$ for any $\theta_0 \in \Theta$, but an estimator can be superefficient wrt the MLE $\hat{\theta}^n$ and wrt $\mathcal{R}_2$ at $\theta_0 \in \Theta$;*
(ii) *no estimator $\tilde{\theta}^n$ is asymptotically* CR*−superefficient wrt $\mathcal{R}_2$ at $\theta_0 \in \Theta$.*

Proof. Under the Assumptions, the MLE is strongly consistent and the Chapman-Robbins lower bound holds.

Then, the first statement is equivalent to the fact that, if the estimator $\tilde{\theta}^n$ is superefficient wrt the MLE $\hat{\theta}^n$ and wrt $\mathcal{R}_2$ at $\theta_0 \in \Theta$, there is a point $\theta_1 \in \Theta$ such that $\tilde{\theta}^n$ is inefficient wrt the MLE $\hat{\theta}^n$ and wrt $\mathcal{R}_2$ at θ_1. We use the fact that the MLE minimizes the probability of error $\mathbb{P}_e$; therefore, if $\mathbb{P}_{\theta_0}\left(\tilde{\theta}^n \neq \theta_0\right) < \mathbb{P}_{\theta_0}\left(\hat{\theta}^n \neq \theta_0\right)$, since when $\pi(\theta_j) = (J+1)^{-1}$ we have $r\left(\tilde{\theta}^n, \pi\right) \geq r\left(\hat{\theta}^n, \pi\right)$, the desired consequence follows.

As concerns the second statement, it is enough to follow the proof of Proposition 2.9 and to reason by contradiction. □

Remark 2.15. *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.*

3 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 3.1 recalls the asymptotics of standard $\sqrt{n}-$*consistent and asymptotically normal* ($\sqrt{n}-$CAN) estimation theory. Section 3.2 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 3.3 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. Testing properties of this estimator are investigated.

3.1 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 $f = (\mathrm{d}\mathbb{P}./\mathrm{d}\mu) : \Theta \to \mathbb{R}$ 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 Newey and McFadden (1994) or Gouriéroux and Monfort (1995, Vol. 2, Ch. 24).

3.2 Discretized Continuous Estimators

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

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

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

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

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

this is a *classical minimum distance* (CMD) *estimator* (see Newey and McFadden, 1994, p. 2116), in the sense that it minimizes the distance between the reduced form parameter estimator, say $\breve{\theta}^n$, and the structural parameter $\theta_i \in \Theta$. Therefore:

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

Using the asymptotic approximation $\Phi(-t) \sim \left(1 + O\left(t^{-2}\right)\right) \cdot \phi(t)/t$, the probability of missclassification can be obtained as:

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

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 White, 1982, p. 5).

This estimator can be generalized to the case of a distance $d(\cdot,\cdot)$ (instead of the norm $\|\cdot\|$), such that, for any $\theta_j \in \Theta \backslash \{\theta_0\}$ 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:

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

therefore:

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

The probability of missclassification is:

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

3.3 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 $\hat{\theta}^{n,\mathsf{D}}$, and of the continuous one, say $\hat{\theta}^{n,\mathsf{C}}$:

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

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

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

from the previous results (see also Ryu, 1999), it is clear that there exist $C_1 > 0$, $\rho \in]0,1[$ and $R > 0$ such that:

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

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:

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

$\hat{\theta}^n$ 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 one 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:

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

References

[1] Aczél, J. and Daróczy, Z. (1975), *On Measures of Information and Their Characterizations*, Mathematics in Science and Engineering 115, Academic Press, New York.

[2] Amari, S. (1985), *Differential-geometrical Methods of Statistics*, Springer Verlag, New York.

[3] Andersen, P.K. and Gill, R.D. (1982), Cox's regression model for counting processes: A large sample study, *Annals of Statistics*, **10**, pp. 1100-1120.

[4] Bahadur, R.R. (1960), On the asymptotic efficiency of tests and estimates, *Sankhyā*, **22**, pp. 229-252.

[5] Bahadur, R.R. and Rao, R.R. (1960), On deviations of the sample mean, *Annals of Mathematical Statistics*, **31**, pp. 1015-1027.

[6] Baram, Y. and Sandell, N.R. Jr. (1978a), An information theoretic approach to dynamical systems modeling and identification, *IEEE Transactions on Automatic Control*, **23**, pp. 61-66.

[7] Baram, Y. and Sandell, N.R. Jr. (1978b), Consistent estimation on finite parameter sets with application to linear systems identification, *IEEE Transactions on Automatic Control*, **23**, pp. 451-454.

[8] Barron, A.R. (1985), The strong ergodic theorem for densities: Generalized Shannon-McMillan-Breiman theorem, *Annals of Probability*, **13**, pp. 1292-1303.

[9] Bhattacharya, R.N. and Rao, R.R. (1976), *Normal Approximations and Asymptotic Expansions*, John Wiley and Sons, New York.

[10] Blackwell, D. and Hodges, J.L. (1959), The probability in the extreme tail of a convolution, *Annals of Mathematical Statistics*, **30**, pp. 1113-1120.

[11] Blyth, C.R. and Roberts, D.M. (1972), On inequalities of Cramér-Rao type and admissibility proofs, pp. 17-30 in *Sixth Berkeley Symposium*, University of California Press.

[12] Caines, P.E. (1975), A note on the consistency of maximum likelihood estimates for finite families of stochastic processes, *Annals of Statistics*, **3**, pp. 539-546.

[13] Chamberlain, G. (2000), Econometric applications of maxmin expected utility, *Journal of Applied Econometrics*, **15**, pp. 625-644.

[14] Chapman, D.G. and Robbins, H. (1951), Minimum variance estimation without regularity assumptions, *Annals of Mathematical Statistics*, **22**, pp. 581-586.

[15] Chernoff, H. (1952), A measure of asymptotic efficiency for tests of a hypothesis based on the sum of observations, *Annals of Mathematical Statistics*, **23**, pp. 493-507.

[16] CHERNOFF, H. (1956), Large-sample theory: Parametric theory, *Annals of Mathematical Statistics*, **27**, pp. 1-22.

[17] CHOIRAT, C., HESS, C. and SERI, R. (2001), A functional version of Birkhoff ergodic theorem for a normal integrand: A variational approach, *Working Paper, Université Paris 9 Dauphine.*

[18] CLÉMENT, E. (1995), Modélisation statistique en finance et estimation de processus de diffusion, *PhD Thesis, Université Paris 9 Dauphine.*

[19] COVER, T.M. and THOMAS, J.A. (1991), *Elements of Information Theory*, John Wiley and Sons, New York.

[20] CRAMÉR, H. (1946), *Mathematical Methods of Statistics*, Princeton University Press, Princeton.

[21] DEMBO, A. and ZEITOUNI, O. (1998), *Large Deviations Techniques and Applications*, 2nd Edition, Springer Verlag, Berlin.

[22] FARRELL, R.H. (1972), On the best obtainable asymptotic rates of convergence in estimation of a density at a point, *Annals of Mathematical Statistics*, **43**, pp. 170-180.

[23] FLINN, C. and HECKMAN, J. (1982), New methods for analyzing structural models of labor force dynamics, *Journal of Econometrics*, **18**, pp. 115-168.

[24] FU, J.C. (1982), Large sample point estimation: A large deviation theory approach, *Annals of Statistics*, **10**, pp. 762-771.

[25] FUTSCHIK, A. and PFLUG, G. (1995), Confidence sets for discrete stochastic optimization, *Annals of Operations Research*, **56**, pp. 95-108.

[26] GENON-CATALOT, V. and PICARD, D. (1993), *Eléments de Statistique Asymptotique*, Springer Verlag, Paris.

[27] GHOSH, M. and MEEDEN, G. (1978), Admissibility of the MLE of the normal integer mean, *Sankhyā. Series B*, **40**, pp. 1-10.

[28] GILL, R.D. and LEVIT, B.Y. (1995), Applications of the van Trees inequality: a Bayesian Cramér-Rao bound, *Bernoulli*, **1**, pp. 59-79.

[29] GOURIÉROUX, C. and MONFORT, A. (1995), *Statistics and Econometric Models*, Cambridge University Press.

[30] GRENANDER, U. (1981), *Abstract Inference*, Springer Verlag, New York.

[31] HALL, P. (1982), On Starr and Vardi's estimates of the number of transmission sources, *Journal of Applied Probability*, **19**, pp. 52-63.

[32] HALL, P. (1989), On convergence rates in nonparametric problems, *International Statistical Review*, **57**, pp. 45-58.

[33] HALL, P. (1992), *The Bootstrap and Edgeworth Expansion*, Springer Series in Statistics, Springer Verlag, New York.

[34] HAMMERSLEY, J.M. (1950), On estimating restricted parameters, *Journal of the Royal Statistical Society. Series B*, **12**, pp. 192-240.

[35] HASSIBI, A. and BOYD, S. (1996), Integer parameter estimation in linear models with applications to GPS, in *Proceedings of the IEEE Conference on Decision and Control*, Kobe, Japan.

[36] HASSIBI, A. and BOYD, S. (1998), Integer parameter estimation in linear models with applications to GPS, *IEEE Transactions on Signal Processing*, **46**, pp. 2938-2952.

[37] HAWKES, R.M. and MOORE, J.B. (1976a), Performance bounds for adaptive estimation, *Proceedings of the IEEE*, **64**, pp. 523–527.

[38] HAWKES, R.M. and MOORE, J.B. (1976b), An upper bound for the mean-square error for Bayesian parameter estimators, *IEEE Transactions on Information Theory*, **22**, pp. 610–615.

[39] HAWKES, R.M. and MOORE, J.B. (1976c), Performance of Bayesian parameter estimators for linear signal models, *IEEE Transactions on Automatic Control*, **21**, pp. 523-527.

[40] HERO, A. (1971), Signal detection and classification, in *Digital Signal Processing (DSP) Handbook*, edited by V.K. Madisetti and D.B. Williams, CRC Press/IEEE Press.

[41] HUBER, P.J. (1973), Robust regression: Asymptotics, conjectures and Monte Carlo, *Annals of Statistics*, **1**, pp. 799-821.

[42] IBRAGIMOV, I.A. and HAS'MINSKII, R.Z. (1981), *Statistical Estimation. Asymptotic Theory*, Springer Verlag, New York.

[43] ILTIS, M. (1995), Sharp asymptotics of large deviations in $\mathbb{R}^d$, *Journal of Theoretical Probability*, **8**, pp. 501-522.

[44] ILTIS, M. (2000), Sharp asymptotics of large deviations for general state-space Markov-additive chains in $\mathbb{R}^d$, *Statistics and Probability Letters*, **4**, pp. 365-380.

[45] JACOD, J. and SHIRYAEV, A.N. (1987), *Limit Theorems for Stochastic Processes*, Springer Verlag, Berlin.

[46] KANAYA, F. and HAN, T.S. (1995), The asymptotics of posterior entropy and error probability for Bayesian estimation, *IEEE Transactions on Information Theory*, **41**, pp. 1988-1992.

[47] KARLIN, S. (1958), Admissibility for estimation with quadratic loss, *Annals of Mathematical Statistics*, **29**, pp. 406-436.

[48] Kester, A.D.M. and Kallenberg, W.C.M. (1986), Large deviations of estimators, *Annals of Statistics*, **14**, pp. 648-664.

[49] Khan, R.A. (1973), On some properties of Hammersley's estimator of an integer mean, *Annals of Statistics*, **1**, pp. 756-762.

[50] Khan, R.A. (1998), A note on Hammersley's estimator of an integer mean, *Journal of Statistical Planning and Inference*, **88**, pp. 37-45.

[51] Kim, J. and Pollard, D. (1990), Cube root asymptotics, *Annals of Statistics*, **18**, pp. 191-219.

[52] Kleywegt, A.J., Shapiro, A. and Homem-de-Mello, T. (1999), The sample average approximation method for stochastic discrete optimization, *Working paper*.

[53] Korostelev, A. and Leonov, S. (1994), Minimax Bahadur efficiency for small confidence intervals, *Working Paper*.

[54] Lainiotis, D.G. (1969), A class of upper bounds on the probability of error for multihypotheses pattern recognition, *IEEE Transactions on Information Theory*, **15**, pp. 730–731.

[55] Lainiotis, D.G. (1969), On a general relationship between estimation, detection, and the Bhattacharyya coefficient, *IEEE Transactions on Information Theory*, **15**, pp. 504-505.

[56] Le Cam, L.M. (1953), On some asymptotic properties of maximum likelihood estimators and related Bayes' estimates, *University of California Publications in Statistics*, **1**, pp. 277-330.

[57] Le Cam, L.M. (1969), *Théorie Asymptotique de la Décision Statistique*, Les Presses de l'Université de Montréal, Montréal.

[58] Le Cam, L.M. (1973), Convergence of estimates under dimensionality restrictions, *Annals of Statistics*, **1**, pp. 38-53.

[59] Le Cam, L.M. and Yang, G.L. (1990), *Asymptotics in Statistics. Some Basic Concepts*, Springer Verlag, New York.

[60] Lindsay, B.G. and Roeder, K. (1987), A unified treatment of integer parameter models, *Journal of the American Statistical Association*, **82**, pp. 758-764.

[61] Liporace, L.A. (1971), Variance of Bayes estimates, *IEEE Transactions on Information Theory*, **17**, pp. 665–669.

[62] Lukacs, E. and Laha, R.G. (1964), *Applications of Characteristic Functions*, Griffin's Statistical Monographs and Courses **14**, Charles Griffin and Co., London.

[63] MANSKI, C.F. (1975), Maximum score estimation of the stochastic utility model of choice, *Journal of Econometrics*, **3**, pp. 205-228.

[64] MANSKI, C.F. (1985), Semiparametric analysis of discrete response: Asymptotic properties of the maximum score estimator, *Journal of Econometrics*, **27**, pp. 313-333.

[65] MCCABE, G.P. JR. (1972), Sequential estimation of a Poisson integer mean, *Annals of Mathematical Statistics*, **43**, pp. 803-813.

[66] MEEDEN, G. and GHOSH, M. (1981), Admissibility in finite problems, *Annals of Statistics*, **9**, pp. 846-852.

[67] NAFIE, M. and TEWFIK, A.H. (1998), Reduced complexity $M-$ary hypotheses testing in wireless communications, Paper presented at *1998 IEEE International Conference on Acoustics, Speech and Signal Processing*, May 12-15, Seattle, USA.

[68] NEWEY, W.K. and MCFADDEN, D.L. (1994), Large sample estimation and hypothesis testing, Chapter 36 in *Handbook of Econometrics, Volume IV*, edited by R.F. Engle and D.L. McFadden, Elsevier Science.

[69] NEY, P.E. (1983), Dominating points and asymptotics of large deviations for random walk on $\mathbb{R}^d$, *Annals of Probability*, **11**, pp. 158-167.

[70] NEY, P.E. (1984), Convexity and large deviations, *Annals of Probability*, **12**, pp. 903-906.

[71] NEY, P.E. and ROBINSON, S.M. (1995), Polyhedral approximation of convex sets with an application to large deviation probability theory, *Journal of Convex Analysis*, **2**, pp. 229-240.

[72] NEYMAN, J. (1949), On the problem of estimating the number of schools of fish, *University of California Publications in Statistics*, **1**, pp. 21-36.

[73] PHILLIPS, P.C.B. (1980), Finite sample theory and the distributions of alternative estimators of the marginal propensity to consume, *Review of Economic Studies*, **XLVII**, pp. 183-224.

[74] POOR, H.V. and VERDÙ, S. (1995), A lower bound on the probability of error in multihypothesis testing, *IEEE Transactions on Information Theory*, **41**, pp. 1992-1994.

[75] ROBERT, C.P. (1992), *L'Analyse Statistique Bayesienne*, Economica, Paris.

[76] ROBSON, D.S. (1958), Admissible and minimax integer-valued estimators of an integer-valued parameters, *Annals of Mathematical Statistics*, **29**, pp. 801-812.

[77] Rothenberg, T.J. (1971), Identification in parametric models, *Econometrica*, **39**, pp. 577-592.

[78] Rozovsky, L.V. (1986), Asymptotic expansions for probabilities of large deviations, *Probability Theory and Related Fields*, **73**, pp. 299-318.

[79] Ryu, K. (1993), Structural duration analysis of management data, *Journal of Econometrics*, **57**, pp. 91-115.

[80] Ryu, K. (1999), Econometric analysis of mixed parameter models, *Journal of Economic Theory and Econometrics*, **5**, pp. 113-124.

[81] Shiryaev, A.N. (1995), *Probability*, 2nd Edition, Springer Verlag, Berlin.

[82] Sievers, G.L. (1978), Estimates of location: A large deviation comparison, *Annals of Statistics*, **6**, pp. 610-618.

[83] Stark, A.E. (1975), Some estimators of the integer-valued parameters of a Poisson variate, *Journal of the American Statistical Association*, **70**, pp. 685-689.

[84] Starr, N. (1974), Optimal and adaptive stopping based on capture times, *Journal of Applied Probability*, **11**, pp. 294-301.

[85] Vajda, I. (1967), Rate of convergence of the information in a sample concerning a parameter, *Czechoslovak Mathematical Journal*, **17**, pp. 225-231.

[86] Vajda, I. (1971a), A discrete theory of search I), *Aplikace Matematiky*, **16**, pp. 241-255.

[87] Vajda, I. (1971b), A discrete theory of search II), *Aplikace Matematiky*, **16**, pp. 319-335.

[88] van der Vaart, A.W. (1997), Superefficiency, pp. 397-410 in *Festschrift for Lucien Le Cam*, Springer Verlag, New York.

[89] van der Vlerk, M.H. (2001), *Stochastic Integer Programming Bibliography*, World Wide Web, `http://mally.eco.rug.nl/biblio/stoprog.html`.

[90] Vardi, Y. (1980), On a stopping time of Starr and its use in estimating the number of transmission sources, *Journal of Applied Probability*, **17**, pp. 235-242.

[91] White, H. (1982), Maximum likelihood estimation of misspecified models, *Econometrica*, **50**, pp. 1-26.

[92] Wong, W.H. (1986), Theory of partial likelihood, *Annals of Statistics*, **14**, pp. 88-123.