Statistical Science
2012, Vol. 27, No. 2, 278–293
DOI: 10.1214/11-STS371

Estimation in Discrete Parameter Models

Christine Choirat and 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.

Key words and phrases: Discrete parameter space, detection, large deviations, information inequalities, efficiency, superefficiency.

1. INTRODUCTION

Sometimes, especially in applications dealing with signal processing and biology, theory provides us with some additional information allowing us to restrict the parameter space to a finite number of points; in these cases, we speak of *discrete parameter models*. Statistical inference when the parameter space is reduced to a lattice was first considered by Hammersley [33] in a seminal paper. 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 [33], page 192); this case was further developed by Khan [46–49]. The Poisson case also met some attention in the literature and was dealt with by Hammersley ([33], page 199) and others [61, 75].

Previous works have shown that the rate of convergence of m-estimators is often exponential [33, 80, 82, 83]. General treatments of admissibility and related topics are in [28, 38, 62, 73] (see also the book [9]); special cases have been dealt with in [44] (page 424, for the case of a translation integral parameter and of integral data under the quadratic loss), [29, 33, 46–49] (for the case of the Gaussian distribution) and [11] (for the case of the discrete uniform distribution). Other papers dealing with optimality in discrete parameter spaces are [27, 78, 79, 81, 84]. Optimality of estimation under a discrete parameter space was also considered by Vajda [80, 82, 83] in a nonorthodox setting inspired by Rényi's theory of random search. Other aspects that have been studied are Bayesian encompassing [24], construction of confidence intervals ([19], pages 224–225), comparison of statistical experiments ([77], [56], Section 2.2), sufficiency and minimal sufficiency [54] and best prediction [76]. Moreover, in the estimation of complex statistical models (see [31], [18], Chapter 4) and in the calculation of efficiency rates (see [1, 15, 56]), approximating a general parameter space by a sequence of finite sets has proved to be a valuable tool. A few papers showed the practical importance of discrete parameter models in signal processing, automatic control and information theory and derived some bounds on the performance of the estimators (see [3–6, 34–36, 52, 53, 58]). More recently, the topic has received new interest in the information theory literature (see [43, 69], and the review paper [37]), in stochastic integer programming (see [25, 50, 86]), and in geodesy (see, e.g., [76], Section 5).

However, no general formula for the convergence rate has ever been obtained, no optimality proof under generic conditions has been provided and no general discussion of efficiency and superefficiency in discrete parameter models has appeared in the literature. In the present paper, we provide a full answer to these problems in the case of discrete parameter models for sam-

Christine Choirat is Associate Professor, Department of Economics, School of Economics and Business Management, Universidad de Navarra, Edificio de Bibliotecas (Entrada Este), 31080 Pamplona, Spain (e-mail: cchoirat@unav.es). Raffaello Seri is Assistant Professor, Dipartimento di Economia, Università degli Studi dell'Insubria, Via Monte Generoso 71, 21100 Varese, Italy (e-mail: raffaello.seri@uninsubria.it).

ples of i.i.d. (independent and identically distributed) random variables. Therefore, after introducing some examples of discrete parameter models in Section 2, in Section 3 we investigate the properties of a class of m-estimators. In particular, in Section 3.1, we derive some conditions for strong consistency; then, in Section 3.2, we calculate an asymptotic approximation of the distribution of the estimator and we establish its convergence rate. These results are specialized to the case of the maximum likelihood estimator (MLE) and extended to Bayes estimators in Section 3.3. In Section 4, we derive upper bounds for the convergence rate in the standard and in the minimax contexts, and we discuss the relations between information inequalities, efficiency and superefficiency. In particular, we prove that estimators of discrete parameters have uncommon efficiency properties. Indeed, under the zero–one loss function, no estimator is efficient in the class of consistent estimators for any value of $\theta_0 \in \Theta$ (θ_0 being here the true value of the parameter) and no estimator attains the information inequality we derive. But the MLE still has some appealing properties since it is minimax efficient and attains the minimax information inequality bound.

2. EXAMPLES OF DISCRETE PARAMETER MODELS

The following examples are intended to show the relevance of discrete parameter spaces in applied and theoretical statistics. In particular, they show that the results in the following sections solve some long-standing problems in statistics, optimization, information theory and signal processing.

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 a Euclidean or of a more abstract space.

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

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

In this case the parameter space is discrete and the maximum likelihood estimator can be shown to be $\hat{\theta}^n = \mathrm{ni}[\frac{\ln(n_0/n)}{\ln k}]$ where $\mathrm{ni}[x]$ is the integer nearest to x (see [33], page 236).

Example 2 (Exponential family restricted to a lattice). Consider a random variable X distributed according to an exponential family where the natural parameter θ is restricted to a lattice $\{\theta_0 + \varepsilon \cdot N, N \in \mathbb{N}^k\}$, for fixed θ_0 and ε (see [57], page 759). The case of a Gaussian distribution has been considered in [33] (page 192) and [46, 48], the Poisson case in [33] (page 199), [61, 75]. In particular, [33] uses the Gaussian model to estimate the molecular weight of insulin, assumed to be an integer (however, see the remarks of Tweedie in the discussion of the same paper).

Example 3 (Stochastic discrete optimization). We consider the optimization problem of the form $\min_{x \in S} g(x)$, where $g(x) = \mathbb{E}G(x, W)$ is an integral functional, $\mathbb{E}$ is the mean under probability $\mathbb{P}$, $G(x, w)$ is a real-valued function of two variables x and w, W is a random variable having probability distribution $\mathbb{P}$ and S is a finite set. We approximate this problem through the sample average function $\hat{g}_n(x) \triangleq \frac{1}{n}\sum_{i=1}^n G(x, W_i)$ and the associated problem $\min_{x \in S} \hat{g}_n(x)$. See [50] for some theoretical results and a discussion of the stochastic knapsack problem and [86] for an up-to-date bibliography.

Example 4 (Approximate inference). In many applied cases, the requirement that the true model generating the data corresponds to a point belonging to the parameter space appears to be too strong and unlikely. Moreover, the objective is often to recover a model reproducing some stylized facts from the original data. In these cases, approximation of a continuous parameter space with a finite number of points allows for obtaining such a model under weaker assumptions. This situation arises, for example, in signal processing and automatic control applications [4–6, 34–36] and is reminiscent of some related statistical techniques, such as the *discretization* device of Le Cam ([56], Section 6.3), or the *sieve estimation* of Grenander ([31]; see also [26], Remark 5).

Example 5 (M-ary hypotheses testing and related fields). In information theory, discrete parameter models are quite common, and their estimation is a generalization of binary hypothesis testing that goes under the names of *M-ary hypotheses* (or *multihypothesis*) *testing*, *classification* or *detection* (see the examples in [63]). Consider a received waveform $r(t)$ described by the equation $r(t) = m(t) + \sigma n(t)$ for $t \geq 0$, where $m(t)$ is a deterministic signal, $n(t)$ is an additive Gaussian white noise and σ is the noise intensity. The set of possible signals is restricted to a finite number of alternatives, say $\{m_0(t), \dots, m_J(t)\}$: the chosen signal is usually the one that maximizes the log-likelihood of the sample, or an alternative criterion function. For example, if the log-likelihood of the process based on the observation window $[0, T]$ is used, we have

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

Much more complex cases can be dealt with; see [37] for an introduction.

3. *m*-ESTIMATORS IN DISCRETE PARAMETER MODELS

In this section, we consider an estimator obtained by maximizing an objective function of the form

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

in what follows, we allow for misspecification. Note that the expression m-estimator stands for *maximum likelihood type estimator*, in the spirit of Huber [39], and not for *maximum* (or *extremum*) *estimator* (see, e.g., [64], page 2114).

3.1 Consistency of *m*-Estimators

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

A1. The data $(Y_i)_{i=1}^n$ are realizations of i.i.d. $(\mathfrak{Y}, \mathcal{Y})$-valued random variables having probability measure $\mathbb{P}_0$.

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

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

The function q is $\mathcal{Y}$-measurable for each $\theta \in \Theta$ and satisfies the L^1-domination condition $\mathbb{E}_0|\ln q(Y;\theta)| < +\infty$ for every $\theta \in \Theta$, where $\mathbb{E}_0$ denotes the expectation taken under the true probability measure $\mathbb{P}_0$.

Moreover, θ_0 is the point of Θ maximizing $\mathbb{E}_0 \ln q(Y;\theta)$ and θ_0 is globally identified (see [64], Section 2.2).

REMARK 1. (i) The assumption of a finite parameter space seems restrictive with respect to the more general assumption of Θ being countable (see, e.g., [33]). However, A1 is compatible with the convex hull of Θ being compact, as in standard asymptotic theory. Indeed, the cases analyzed in [33] have convex likelihood functions and this is a well-known substitute for compactness of Θ (see [64], page 2133; see [17], for consistency with neither convexity nor compactness). Moreover, the restriction to finite parameter spaces seems to be necessary to derive the asymptotic approximation to the distribution of m-estimators.

(ii) The relative position of the points of Θ is unimportant and the choice of θ_0 as the maximizer is arbitrary and is made only for practical purposes. Note that θ_0 has no link with $\mathbb{P}_0$ apart from being the pseudo-true value of $\ln q$ with respect to $\mathbb{P}_0$ on the parameter space Θ (see, e.g., [30], Volume 1, page 14).

PROPOSITION 1. *Under Assumption* A1, *the* m*-estimator* $\hat{\theta}^n$ *is a* $\mathbb{P}_0$*-strongly consistent estimator of* θ_0 *and is* $\mathcal{Y}^{\otimes n}$*-measurable.*

REMARK 2. A similar result of consistency for discrete parameter spaces has been provided by [74] (page 446), by [13, 14] (pages 325–333), by [8] (pages 1293–1294) as an application of the Shannon–McMillan–Breiman Theorem of information theory, by [87] (Section 2.1) as a preliminary result of his work on partial likelihood, and by [60] (page 96, Section 7.1.6).

3.2 Distribution of the *m*-Estimator

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

To simplify the following discussion, we introduce the processes:

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

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

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

The only approaches that have been successful in our experience are large deviations (in logarithmic and exact form) and saddlepoint approximations. Note that we could have defined the probability in (2) as $\mathbb{P}_0(Q_n(\theta_i) \geq Q_n(\theta_j), \forall j \neq i)$ or through any other combination of equality and inequality signs; this introduces some arbitrariness in the distribution of $\hat{\theta}^n$. However, we will give some conditions (see Proposition 2) under which this difference is asymptotically irrelevant.

Section 3.2.1 introduces definitions and assumptions and discusses a preliminary result. In Section 3.2.2 we derive some results on the asymptotic behavior of $\mathbb{P}_0(\hat{\theta}^n = \theta_i)$ using large deviations principles (LDP). Then, we provide some refinements of the previous expressions using the theory of exact asymptotics for large deviations, with special reference to the case $J = 1$. At last, Section 3.2.3 derives saddlepoint approximations for probabilities of the form (2).

3.2.1 *Definitions, assumptions and preliminary results.* As concerns the distribution of the m-estimator $\hat{\theta}^n$, we shall need some concepts and functions derived from large deviations theory (see [21]); we recall that the processes $Q_n(\theta_j)$, $\mathbf{X}_k$ and $\mathbf{X}_k^{(i)}$ have been introduced in (1). Then, for $i = 0,\dots,J$, we define the moment generating functions

$$\begin{aligned} M^{(i)}(\boldsymbol{\lambda}) &\triangleq \mathbb{E}_0\big[e^{\sum_{j=0,\dots,J,j\neq i} \lambda_j \cdot [\ln q(Y;\theta_i) - \ln q(Y;\theta_j)]}\big] \\ &= \mathbb{E}_0\big[e^{\boldsymbol{\lambda}^\mathsf{T} \mathbf{X}^{(i)}}\big], \end{aligned}$$

the logarithmic moment generating functions

$$\begin{aligned} \Lambda^{(i)}(\boldsymbol{\lambda}) &\triangleq \ln M^{(i)}(\boldsymbol{\lambda}) \\ &= \ln \mathbb{E}_0\big[e^{\sum_{j=0,\dots,J,j\neq i} \lambda_j \cdot [\ln q(Y;\theta_i) - \ln q(Y;\theta_j)]}\big] \\ &= \ln \mathbb{E}_0\big[e^{\boldsymbol{\lambda}^\mathsf{T} \mathbf{X}^{(i)}}\big], \end{aligned}$$

and the Cramér transforms

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

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

The following assumptions will be used to approximate the distribution of $\hat{\theta}^n$.

A2. There exists a $\delta > 0$ such that, for any $\eta \in (-\delta, \delta)$, we have

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

REMARK 3. In what follows, this assumption could be replaced by a condition as in [68] (Assumptions H1 and H2).

A3. $\Lambda^{(i)}(\boldsymbol{\lambda})$ is *steep*, that is, $\lim_{n\to\infty} \|\frac{\partial \Lambda^{(i)}(\mathbf{x})}{\partial \mathbf{x}}\| = \infty$ whenever $\{\mathbf{x}_n\}_n$ is a sequence in $\operatorname{int}(\mathcal{D}_{\Lambda^{(i)}})$ converging to a boundary point of $\operatorname{int}\mathcal{D}_{\Lambda^{(i)}}$.

REMARK 4. Under Assumptions A1, A2 and A3, $\Lambda^{(i)}(\cdot)$ is *essentially smooth* (see, e.g., [21], page 44). A sufficient condition for A3 and essential smoothness is openness of $\mathcal{D}_{\Lambda^{(i)}}$ (see [66], page 905, and [40], pages 505–506).

A4. $\operatorname{int}(\mathbb{R}_+^J \cap \mathcal{S}^{(i)}) \neq \varnothing$, where $\mathcal{S}^{(i)}$ is the closure of the convex hull of the support of the law of $\mathbf{X}^{(i)}$.

We will also need the following lemma showing the equivalence between Assumption A2 and the so-called *Cramér condition* $\mathbf{0} \in \operatorname{int}(\mathcal{D}_{\Lambda^{(i)}})$, for any $i = 0,\dots,J$.

LEMMA 1. *Under Assumption* A1, *the following conditions are equivalent*:

(i) *Assumption* A2 *holds*;
(ii) $\mathbf{0} \in \operatorname{int}(\mathcal{D}_{\Lambda^{(i)}})$, *for any* $i = 0,\dots,J$.

As concerns the saddlepoint approximation of Section 3.2.3, we need the following assumption:

A5. The inequality

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

holds for $\mathbf{u}\in\operatorname{int}(\mathcal{D}_{\Lambda^{(i)}})$, $\delta>0$ and $c<|\mathbf{t}|<C\cdot n^{(s-3)/2}$ (ι denotes the imaginary unit).

3.2.2 *Large deviations asymptotics.* In this section we consider large deviations asymptotics. We note that, in what follows, $\operatorname{int}(\mathbb{R}_+^J)^c$ stands for $\operatorname{int}\{[(\mathbb{R}_+)^J]^c\}$.

PROPOSITION 2. (i) *For* $i=1,\dots,J$, *under Assumption* A1, *the following result holds*:

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

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

(ii) *Under Assumptions* A1 *and* A2:

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

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

(iii) *Under Assumptions* A1, A2, A3 *and* A4:

$$\mathbb{P}_0(\hat\theta^n=\theta_i)=\exp\Bigl\{-(n+o(n))\cdot\inf_{\mathbf{y}\in\operatorname{int}(\mathbb{R}_+^J)}\Lambda^{(i),*}(\mathbf{y})\Bigr\}$$
$$=\exp\Bigl\{-(n+o(n))\cdot\inf_{\mathbf{y}\in\mathbb{R}_+^J}\Lambda^{(i),*}(\mathbf{y})\Bigr\}.$$

PROPOSITION 3. *Under Assumption* A1, *the following inequality holds*:

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

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

Under Assumptions A1 *and* A2:

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

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

REMARK 5. The proposition allows us to obtain an upper bound on the bias of the m-estimator, $\mathsf{Bias}(\hat\theta^n)\leq\sup_{j\neq0}|\theta_j-\theta_0|\cdot\mathbb{P}_0(\hat\theta^n\neq\theta_0)$.

A better description of the asymptotic behavior of the probability $\mathbb{P}_0(\hat\theta^n=\theta_i)$ could be obtained, under some additional conditions, from the study of the neighborhood of the contact point between the set $(\mathbb{R}_+)^J$ and the level sets of the Cramér transform $\Lambda^{(i),*}(\cdot)$. We leave the topic for future work. Here we just remark the following brackets on the convergence rate.

PROPOSITION 4. *Under Assumptions* A1, A2, A3 *and* A4, *for sufficiently large* n, *the following result holds*:

$$c_1\frac{e^{-n\cdot\inf_{\mathbf{y}\in\mathbb{R}_+^J}\Lambda^{(i),*}(\mathbf{y})}}{n^{J/2}}\leq\mathbb{P}_0(\hat\theta^n=\theta_i)\leq c_2\frac{e^{-n\cdot\inf_{\mathbf{y}\in\mathbb{R}_+^J}\Lambda^{(i),*}(\mathbf{y})}}{n^{1/2}}$$

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

When $J=1$, a more precise convergence rate can be obtained under the following assumption:

A6. When $J=1$, there is a positive value $\mu\in\operatorname{int}(\mathcal{D}_{\Lambda^{(1)}})$ such that $\frac{\partial\Lambda^{(1)}(\lambda)}{\partial\lambda}|_{\lambda=\mu}=0$. Moreover, the law of $\ln\frac{q(Y;\theta_1)}{q(Y;\theta_0)}$ is nonlattice (see [21], page 110).

PROPOSITION 5. *Under Assumptions* A1, A2, A3, A4 *and* A6, *with* $\Theta=\{\theta_0,\theta_1\}$ *and* $J=1$, *we have*

$$\mathbb{P}_0(\hat\theta^n=\theta_1)=\mathbb{P}_0(\hat\theta^n\neq\theta_0)$$
$$=\frac{e^{n\cdot\Lambda^{(1)}(\mu)}}{\mu\cdot\sqrt{\Lambda^{(1),\prime\prime}(\mu)2\pi n}}\cdot\bigl(1+o(1)\bigr)$$
$$=\frac{e^{-n\cdot\Lambda^{(1),*}(0)}}{(\Lambda^{(1),*})^{\prime}(0)}\cdot\sqrt{\frac{(\Lambda^{(1),*})^{\prime\prime}(0)}{2\pi n}}\cdot\bigl(1+o(1)\bigr).$$

REMARK 6. A refinement of the previous asymptotic rates can be obtained using results in [2, 10].

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

THEOREM 1. *Under Assumptions* A1, A2 *and* A5, *for* $i\neq0$, *it is possible to choose* $\mathbf{u}$ *such that, for every*

$\mathbf{v} \in [(\operatorname{int}\mathbb{R}^J_+) \ominus \frac{\partial\Lambda^{(i)}(\mathbf{u})}{\partial\mathbf{u}}]$, $\mathbf{u}^\mathsf{T}\mathbf{v} \geq 0$ *and*

$$\mathbb{P}_0(\hat{\theta}^n = \theta_i) = \exp\biggl(n\biggl[\Lambda^{(i)}(\mathbf{u}) - \mathbf{u}\cdot\frac{\partial\Lambda^{(i)}(\mathbf{u})}{\partial\mathbf{u}}\biggr]\biggr) \cdot \bigl[e_{s-3}\bigl(\mathbf{u}, \operatorname{int}\mathbb{R}^J_+ \ominus \mathbb{E}_0\mathbf{X}^{(i)}\bigr) + \delta\bigl(\mathbf{u}, \operatorname{int}\mathbb{R}^J_+ \ominus \mathbb{E}_0\mathbf{X}^{(i)}\bigr)\bigr],$$

where

$$e_{s-3}\bigl(\mathbf{u}, \operatorname{int}\mathbb{R}^J_+ \ominus \mathbb{E}_0\mathbf{X}^{(i)}\bigr) = \int_{\operatorname{int}\mathbb{R}^J_+ \ominus \frac{\partial\Lambda^{(i)}(\mathbf{u})}{\partial\mathbf{u}}} \frac{\exp(-n\mathbf{u}\cdot\mathbf{y} - n\|\mathbf{y}^*\|^2/2)}{(2\pi/n)^{J/2}\Delta^{1/2}} \cdot \Biggl[1 + \sum_{i=1}^{s-3} n^{-i/2} Q_{i\mathbf{u}}\bigl(\sqrt{n}\mathbf{y}^*\bigr)\Biggr]\,\mathrm{d}\mathbf{y},$$

$$Q_{\ell\mathbf{u}}(\mathbf{x}) = \sum_{m=1}^{\ell} \frac{1}{m!} \sum\nolimits^{*} \sum\nolimits^{**} \biggl(\frac{\kappa_{\nu_1 n}\cdots\kappa_{\nu_m n}}{\nu_1!\cdots\nu_m!}\biggr) \cdot H_{I_1}(x_1)\cdots H_{I_d}(x_d),$$

$$\bigl|\delta\bigl(\mathbf{u}, \operatorname{int}\mathbb{R}^J_+ \ominus \mathbb{E}_0\mathbf{X}^{(i)}\bigr)\bigr| \leq C\cdot n^{-(s-2)/2}$$

and $\mathbf{V} = \frac{\partial^2\Lambda^{(i)}(\mathbf{u})}{\partial\mathbf{u}^2}$, $\mathbf{y}^* = \mathbf{V}^{-1/2}\mathbf{y}$, $\|\mathbf{y}^*\|^2 = \mathbf{y}^*\cdot\mathbf{y}^* = \mathbf{y}^\mathsf{T}\mathbf{V}^{-1}\mathbf{y}$, $\Delta = |\mathbf{V}|$, H_m *is the usual Hermite–Chebyshev polynomial of degree* m, $\sum^*$ *denotes the sum over all* m*-tuples of positive integers* $(j_1, \dots, j_m)$ *satisfying* $j_1 + \cdots + j_m = \ell$, $\sum^{**}$ *denotes the sum over all* m*-tuples* $(\nu_1, \dots, \nu_m)$ *with* $\nu_i = (\nu_{1i}, \dots, \nu_{di})$, *satisfying* $(\nu_{1i} + \cdots + \nu_{di} = j_i + 2, i = 1, \dots, m)$, *and* $I_h = \nu_{h1} + \cdots + \nu_{hm}, h = 1, \dots, d$. *Note that* $Q_{\ell\mathbf{u}}$ *depends on* $\mathbf{u}$ *through the cumulants calculated at* $\mathbf{u}$.

REMARK 7. The main question that this theorem leaves open is the choice of the point $\mathbf{u}$. Usually this point is chosen as a solution $\hat{\mathbf{u}}$ of $\mathbf{m}(\hat{\mathbf{u}}) = \hat{\mathbf{x}}$; this corresponds to a saddlepoint in $\kappa(\mathbf{u})$. [20] (Section 6) and [59] (page 480) give some conditions for $J = 1$; [41] (page 23) and [7] (page 153) give conditions for general J. [42] suggests that the most common solution is to choose $\hat{\mathbf{x}}$ and $\hat{\mathbf{u}}$ ($\hat{\mathbf{x}}$ belonging to the boundary of $[\operatorname{int}\mathbb{R}^J_+ \ominus \mathbb{E}_0\mathbf{X}^{(i)}]$ and $\hat{\mathbf{u}}$ solving $\mathbf{m}(\hat{\mathbf{u}}) = \hat{\mathbf{x}}$), such that for every $\mathbf{v} \in [\operatorname{int}\mathbb{R}^J_+ \ominus \frac{\partial\Lambda^{(i)}(\mathbf{u})}{\partial\mathbf{u}}]$, $\hat{\mathbf{u}}^\mathsf{T}\mathbf{v} \geq 0$. This is the same as a dominating point in [65–67]; therefore, A2, A3 and A4, for sufficiently large n, imply the existence of this point for any i.

3.3 The MLE and Bayes Estimators in Discrete Parameter Models

In this section, we show how the previous results can be applied to the MLE and Bayes estimators under the zero–one loss function. The MLE is defined by

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

This corresponds to the *minimum-error-probability estimate* of [69] and to the *Bayesian estimator* of [82, 83]. On the other hand, using the prior densities given by $\pi(\theta)$ for $\theta \in \Theta$, the posterior densities of the Bayesian estimator are given by

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

The Bayes estimator relative to zero–one loss $\check{\theta}^n$ (see Section 4.3 for a definition) is the mode of the posterior distribution and is given by

$$\check{\theta}^n \triangleq \arg\max_{\theta\in\Theta} \ln\mathbb{P}\{\theta|\mathbf{Y}\} = \arg\max_{\theta\in\Theta} \Biggl[\frac{1}{n}\sum_{i=1}^n \ln f_{Y_i}(y_i; \theta) + \frac{\ln\pi(\theta)}{n}\Biggr]. \tag{3}$$

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

Assumption A1 can be replaced by the following ones (where Assumptions A8 and A9 entail that the likelihood function is asymptotically maximized at θ_0 only):

A7. 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 $\Theta = \{\theta_0, \theta_1, \dots, \theta_J\}$, of finite cardinality. Let $(\mathfrak{Y}, \mathcal{Y})$ be a measurable space and μ a positive σ-finite measure defined on $(\mathfrak{Y}, \mathcal{Y})$ such that, for every $\theta \in \Theta$, $\mathbb{P}_\theta$ 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 i.i.d. realizations from the probability measure $\mathbb{P}_0$.

A8. The log density satisfies the L^1-domination condition $\mathbb{E}_0|\ln f_Y(Y; \theta_i)| < +\infty$, for $\theta_i \in \Theta$, where $\mathbb{E}_0$ denotes the expectation taken under the true probability measure $\mathbb{P}_0$.

A9. θ_0 is the point of Θ maximizing $\mathbb{E}_0 \ln f_Y(Y; \theta)$ and is globally identified.

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

A10. The prior distribution verifies $\pi(\theta) > 0$ for any $\theta \in \Theta$.

Proposition 1 holds for the MLE under Assumptions A7, A8 and A9, while for Bayes estimators A10 is required, too. Note that, under correct specification (i.e., when the true parameter value belongs to Θ), a standard Wald's argument (see, e.g., Lemma 2.2 in [64], page 2124) shows that $\mathbb{E}_{\theta_0} \ln f_Y(Y;\theta)$ is maximized for $\theta = \theta_0$.

As concerns the 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 log-likelihood function $\mathsf{L}_n(\theta)$, and $\mathbf{X}_k$ and $\mathbf{X}_k^{(i)}$ by the log-likelihood processes:

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

Also $M(\boldsymbol{\lambda})$ and $M^{(i)}(\boldsymbol{\lambda})$ are consequently defined. Propositions 2 and 3 hold when Assumption A1 is replaced by Assumptions A7, A8 and A9.

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

$$\begin{aligned} M^{(i)}(\boldsymbol{\lambda}) &\triangleq \mathbb{E}_{\theta_0}\bigl[e^{\sum_{j=0,\dots,J,j\neq i} \lambda_j \cdot [\ln f_Y(Y;\theta_i) - \ln f_Y(Y;\theta_j)]}\bigr] \\ &= \int f_Y(y;\theta_i)^{\sum_{j=0,\dots,J,j\neq i} \lambda_j} \\ &\qquad \cdot \prod_{j=1,\dots,J,j\neq i} f_Y(y;\theta_j)^{-\lambda_j} \\ &\qquad\qquad \cdot f_Y(y;\theta_0)^{1-\lambda_0} \mu(\mathrm{d}y). \end{aligned} \tag{4}$$

Therefore, in this case, the moment generating function $M^{(i)}(\boldsymbol{\lambda})$ reduces to the so-called Hellinger transform $\mathsf{H}_{\boldsymbol{\gamma}}(\theta_0,\dots,\theta_J)$ (see [56], page 43) for a certain linear transformation of $\boldsymbol{\lambda}$ in $\boldsymbol{\gamma}$:

$$\begin{aligned} &\mathsf{H}_{\boldsymbol{\gamma}}(\theta_0,\dots,\theta_J) \\ &\quad \triangleq \int \prod_{j=0}^{J} [\mathbb{P}_{\theta_j}(\mathrm{d}y)]^{\gamma_j} \\ &\quad = \int \Biggl[\prod_{j=0}^{J} f_Y(y;\theta_j)^{\gamma_j}\Biggr] \mu(\mathrm{d}y), \qquad \sum_{j=0}^{J} \gamma_j = 1. \end{aligned}$$

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

Proposition 4 holds if Assumption A1 is replaced by Assumptions A7, A8 and A9, and if A2 and A3 hold true. However, Assumption A4 is unnecessary; indeed, the fact that $\operatorname{int}(\mathbb{R}_+^J \cap \mathcal{S}^{(i)}) \neq \varnothing$ can be proved showing that $\mathbf{0} \in \operatorname{int}(\mathcal{S}^{(i)})$. This is equivalent to the existence, for $j = 1,\dots,J$, $j \neq i$, of two sets A_j^* and A_j^{**} of positive μ-measure and included in the support of Y such that, for $y_j^* \in A_j^*$ and $y_j^{**} \in A_j^{**}$, $f_Y(y_j^*;\theta_i) > f_Y(y_j^*;\theta_j)$ and $f_Y(y_j^{**};\theta_i) < f_Y(y_j^{**};\theta_j)$. This follows easily noting that these densities have to integrate to 1, are almost surely (a.s.) different according to Assumption A9 and have the same support according to Assumption A7.

In order to derive the distribution of Bayes estimators, we consider Equation (3) and we let $\ln \boldsymbol{\pi}^{(i)} \triangleq [\ln \frac{\pi(\theta_i)}{\pi(\theta_j)}]_{j=0,\dots,J,j\neq i}$. Then, we can write

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

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

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

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

4. OPTIMALITY AND EFFICIENCY

In this section, we are interested in the problem of efficiency, with special reference to maximum likelihood and Bayes estimators. In what follows, we will suppose that the true parameter value belongs to Θ; this will be reflected in the probabilities that will be written as $\mathbb{P}_0 = \mathbb{P}_{\theta_0}$. Indeed, efficiency statements for misspecified models are quite difficult to interpret.

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 continuous case, the two concepts almost coincide (despite superefficiency). However, in the discrete 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.

An interesting problem concerns the choice of a measure of efficiency for the MLE in discrete parameter models: in his seminal paper, Hammersley [33] 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 derived, in slightly more generality, in [12, 16]. However, this choice is well-suited only in cases in which the MSE is a good measure of risk, for example, if the limiting distribution of the normalized estimator is normal. Following the discussion by Lindley in [33], we consider a different cost function $\mathcal{C}_1(\theta, \theta_0)$, whose risk function is given by the probability of missclassification:

$$\mathcal{C}_1(\tilde{\theta}^n, \theta_0) = \mathsf{1}_{\{\tilde{\theta}^n \neq \theta_0\}},$$
$$\mathcal{R}_1(\tilde{\theta}^n, \theta_0) = \mathbb{P}_{\theta_0}(\tilde{\theta}^n \neq \theta_0).$$

We also 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 $\mathcal{R}_1(\tilde{\theta}^n, \theta_0)$ as

$$r_1(\tilde{\theta}^n, \pi) = \sum_{j=0}^{J} \pi(\theta_j) \cdot \mathbb{P}_{\theta_j}(\tilde{\theta}^n \neq \theta_j).$$

If $\pi(\theta_j) = (J+1)^{-1}$ we define $\mathbb{P}_e \triangleq r_1(\tilde{\theta}^n, \pi)$ as the *average probability of error*. Note that this is indeed the measure of error used by [82, 83].

Using the risk function $\mathcal{R}_1$, in Section 4.1 we derive some information inequalities and we prove in Section 4.2 some optimality and efficiency results for Bayes and ML estimators. In Section 4.3 we briefly deal with alternative risk functions.

4.1 Information Inequalities

This section contains lower bounds for the previously introduced risk function $\mathcal{R}_1$. In the specific case of discrete parameters, these generalize and unify the lower bounds proposed in [16, 32, 33, 45].

In the following, first of all, a lower bound is proved and then a minimax version of the same result is obtained. When needed, we will refer to the former as *Chapman–Robbins lower bound* (and to the related efficiency concept as *Chapman–Robbins efficiency*) since it recalls the lower bound proposed by these two authors in their 1951 paper, and to the latter as *minimax Chapman–Robbins lower bound*. Then, from these results, we derive a lower bound for the Bayes risk.

4.1.1 *Lower bounds for the risk function $\mathcal{R}_1$.* 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. Moreover, a version of the same bound for estimators respecting (6) is provided; this corresponds to a similar result proposed in [23]

PROPOSITION 6. *Under Assumptions* A7 *and* A9, *for a strongly consistent estimator $\tilde{\theta}^n$*:

$$\text{(5)}\quad \begin{aligned}&\lim_{n\to\infty} \frac{1}{n} \ln \mathcal{R}_1(\tilde{\theta}^n, \theta_0)\\ &\quad\geq \sup_{\theta_1 \in \Theta \setminus \{\theta_0\}} \mathbb{E}_{\theta_1} \ln\left(\frac{f_Y(Y;\theta_0)}{f_Y(Y;\theta_1)}\right).\end{aligned}$$

On the other hand, if

$$\text{(6)}\quad \limsup_{n\to\infty} \mathbb{P}_{\theta_j}\{\tilde{\theta}^n \neq \theta_j\} < 1,$$

then

$$\liminf_{n\to\infty} \frac{1}{n} \ln \mathcal{R}_1(\tilde{\theta}^n, \theta_0) \geq \sup_{\theta_1 \in \Theta \setminus \{\theta_0\}} \mathbb{E}_{\theta_1} \ln\left(\frac{f_Y(Y;\theta_0)}{f_Y(Y;\theta_1)}\right).$$

REMARK 8. (i) Note that this inequality only holds for estimators that are consistent or respect condition (6), while the one of Proposition 7 holds for any estimator.

(ii) Proposition 6 provides an upper bound for the *inaccuracy rate* of [45]:

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

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

4.1.2 *Minimax lower bounds for the risk function $\mathcal{R}_1$.* The following result is a minimax lower bound on the probability of misclassification. It is based on the Neyman–Pearson Lemma and Chernoff's Bound.

PROPOSITION 7. *Under Assumptions* A7 *and* A9, *for any estimator $\tilde{\theta}^n$*:

$$\begin{aligned}&\liminf_{n\to\infty} \frac{1}{n} \ln \sup_{\theta_0 \in \Theta} \mathcal{R}_1(\tilde{\theta}^n, \theta_0)\\ \text{(7)}\quad &\geq \sup_{\theta_1 \in \Theta \setminus \{\theta_0\}} \sup_{\theta_0 \in \Theta} \ln\left[\inf_{1>u>0} \int f_Y(y;\theta_1)^u\right.\\ &\qquad\left.\cdot f_Y(y;\theta_0)^{1-u} \mu(\mathrm{d}y)\right].\end{aligned}$$

REMARK 9. (i) The previous proposition provides an expression for the *minimax Bahadur risk* (also called *(minimax) rate of inaccuracy*; see [1, 51]) analogous to Chernoff's Bound, thus providing a minimax version of Remark 8(ii).

(ii) Other methods to derive similar minimax inequalities are Fano's Inequality and Assouad's Lemma (see [56], page 220); however, in the present case they do not allow us to obtain tight bounds, since the usual application of these methods relies on the approximation of the parameter space with a finite set of points Θ whose cardinality increases with n. Clearly, this cannot be done in the present case.

(iii) Using Lemma 5.2 in [70], it is possible to show that the minimax bound is larger than the classical one.

(iv) Under Assumption A10, the Bayes risk r_1 under the risk function $\mathcal{R}_1$ and the prior π respects the equality

$$(8)\quad \lim_{n\to\infty}\frac{1}{n}\ln r_1(\tilde{\theta}^n,\pi)=\lim_{n\to\infty}\frac{1}{n}\ln\max_{\theta_0\in\Theta}\mathcal{R}_1(\tilde{\theta}^n,\theta_0).$$

Then, Proposition 7 holds also for the Bayes risk: clearly this bound is independent of the prior distribution π (provided it is strictly positive, i.e., A10 holds) and also holds for the probability of error $\mathbb{P}_e$. This inequality can be seen as an asymptotic version of the van Trees inequality for a different risk function.

4.2 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 $\mathcal{R}=\mathcal{R}(\bar{\theta}^n,\theta_0)$ the risk function of the estimator $\bar{\theta}^n$ evaluated at θ_0, and by $\tilde{\Theta}$ a class of estimators.

DEFINITION 1. The estimator $\bar{\theta}^n$ is *efficient with respect to* (w.r.t.) $\tilde{\Theta}$ *and* w.r.t. $\mathcal{R}$ *at* θ_0 if

$$(9)\qquad \mathcal{R}(\bar{\theta}^n,\theta_0)\le\mathcal{R}(\tilde{\theta}^n,\theta_0)\quad\forall\tilde{\theta}^n\in\tilde{\Theta}.$$

The estimator $\bar{\theta}^n$ is *minimax efficient* w.r.t. $\tilde{\Theta}$ *and* w.r.t. $\mathcal{R}$ if

$$(10)\quad \sup_{\theta_0\in\Theta}\mathcal{R}(\bar{\theta}^n,\theta_0)\le\sup_{\theta_0\in\Theta}\mathcal{R}(\tilde{\theta}^n,\theta_0)\quad\forall\tilde{\theta}^n\in\tilde{\Theta}.$$

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

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

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

The estimator $\bar{\theta}^n$ is *asymptotically* CR-*efficient* w.r.t. $\mathcal{R}$ *at* θ_0 if it attains the Chapman–Robbins lower bound of Proposition 6 at θ_0 [say CR$-\mathcal{R}(\theta_0)$] in the asymptotic form:

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

The estimator $\bar{\theta}^n$ is *asymptotically minimax* CR-*efficient* w.r.t. $\mathcal{R}$ if it attains the minimax Chapman–Robbins lower bound of Proposition 7 (say CR$-\mathcal{R}_{\max}$) in the asymptotic form:

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

The estimator $\bar{\theta}^n$ is *asymptotically* CR-*superefficient* w.r.t. $\mathcal{R}$ if

$$\liminf_{n\to\infty}\frac{1}{n}\ln\mathcal{R}(\bar{\theta}^n,\theta_0)\le\ln\mathrm{CR}-\mathcal{R}(\theta_0)$$

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

REMARK 10. As in Remark 8(ii), it is easy to see that IR-optimality and CR-efficiency w.r.t. $\mathcal{R}_1$ coincide.

The efficiency landscape offered by discrete parameter models will be illustrated by Example 6. This shows that, even in the simplest case, that is, 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 6. Let us consider the estimation of the mean of a Gaussian distribution whose variance σ^2 is known: we suppose that the true mean is α, while the parameter space is $\{-\alpha,\alpha\}$, where α is known. The maximum likelihood estimator $\hat{\theta}^n$ takes the value $-\alpha$ if the sample mean takes on its value in $(-\infty,0)$ and α if it falls in $[0,+\infty)$ (the position of 0 is a convention). Therefore:

$$\begin{aligned}\mathbb{P}_{\theta_0}(\hat{\theta}^n\neq\theta_0)&=\mathbb{P}_{\theta_0}(\hat{\theta}^n=-\alpha)\\&=\int_{-\infty}^{0}\frac{e^{-(\bar{y}-\alpha)^2/(2\sigma^2/n)}}{\sqrt{2\pi\sigma^2/n}}\,\mathrm{d}\bar{y}\end{aligned}$$

$$
\begin{aligned}
&= \int_{-\infty}^{-\sqrt{n}\alpha/\sigma} \frac{e^{-t^2/2}}{\sqrt{2\pi}}\,\mathrm{d}t \\
&= \Phi\left(-\frac{\sqrt{n}\alpha}{\sigma}\right) \\
&= \frac{e^{-n\alpha^2/(2\sigma^2)}}{\sqrt{2\pi n}}\frac{\sigma}{\alpha}\cdot\left(1+O\left(\frac{1}{n}\right)\right),
\end{aligned}
$$

where we have used Problem 1 on page 193 in [22]. Proposition 5 allows also for recovering the right convergence rate. Indeed, we have

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

On the other hand, the lower bound of Proposition 6 yields

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

and the lower bound of Proposition 7 yields

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

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

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

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

When $k=0$, $\tilde{\theta}^n(k)$ coincides with the MLE $\hat{\theta}^n$. Then, the behavior of the estimator is characterized by the probabilities:

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

We have (weak) consistency if

$$
2\left(\frac{\alpha}{\sigma}\right)^2 > k > -2\left(\frac{\alpha}{\sigma}\right)^2. \tag{11}
$$

The risk $\mathcal{R}_1(\tilde{\theta}^n(k),\theta_0)$ under θ_0 is then

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

this can be made smaller than the probability of error of the MLE simply taking $k>0$, thus implying that the MLE is not pointwise efficient.

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

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

and this is never true under (11). If this estimator is pointwise more efficient than the MLE under θ_0, then its risk under θ_1 is given by

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

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

4.2.1 *Optimality w.r.t. classes of estimators.* In the following section, we show some optimality properties of Bayes and ML estimators. We start with an important and well-known fact.

PROPOSITION 8. *Under* A7, A8, A9 *and* A10, *the Bayes risk* $r_1(\tilde{\theta}^n,\pi)$ (*under the zero–one loss function*) *associated with a prior distribution* π *is strictly minimized by the posterior mode corresponding to the prior* π, *for any finite* n.

The following proposition shows that the MLE is admissible and minimax efficient under the zero–one loss and minimizes the average probability of error. It implies 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 $\theta_1 \in \Theta$. As a result, estimators can be more efficient than minimax efficient ones only on portions of the parameter space, but are then strictly less efficient elsewhere.

PROPOSITION 9. *Under Assumptions* A7, A8 *and* A9, *the* MLE *is admissible and minimax efficient w.r.t. the class of all estimators and w.r.t.* $\mathcal{R}_1$ *and minimizes the average probability of error* $\mathbb{P}_e$.

4.2.2 *Optimality w.r.t. the information inequalities.* In this subsection, we will show that the MLE does not attain the Chapman–Robbins lower bound in the form of Proposition 6 but that it attains the minimax form of Proposition 7 and that efficiency and minimax efficiency are generally incompatible.

Therefore, the situation described in Example 6 is general, for it is possible to show that the MLE is generally inefficient with respect to the lower bounds exposed in Proposition 6.

PROPOSITION 10. *Under Assumptions* A7, A8 *and* A9:

(i) *the* MLE *is not asymptotically* CR-*efficient w.r.t.* $\mathcal{R}_1$ *at* θ_0;
(ii) *the* MLE *is asymptotically minimax* CR-*efficient w.r.t.* $\mathcal{R}_1$;
(iii) *an estimator that is asymptotically* CR-*efficient w.r.t.* $\mathcal{R}_1$ *at* θ_0 *is not asymptotically minimax* CR-*efficient w.r.t.* $\mathcal{R}_1$.

REMARK 11. The assumption of homogeneity of the probability measures, necessary to derive (ii), can be removed in the proof of (i) along the lines of [45].

4.2.3 *The evil of superefficiency.* Ever since it was discovered by Hodges, the problem of superefficiency has been dealt with extensively in regular statistical problems (see, e.g., [55, 85]). However, these proofs do not transpose to discrete parameter estimation problems, since they are mostly based on the equivalence of prior probability measures with the Lebesgue measure and on properties of Bayes estimators that do not hold in this case. Moreover, the discussion of the previous sections has shown that, in discrete parameter problems, CR-efficiency and efficiency with respect to a class of estimators do not coincide. The following proposition yields a solution to the superefficiency problem.

PROPOSITION 11. *Under Assumptions* A7, A8 *and* A9:

(i) *no estimator* $\tilde{\theta}^n$ *is asymptotically* CR-*superefficient w.r.t.* $\mathcal{R}_1$ *at* $\theta_0 \in \Theta$;
(ii) *no estimator* $\tilde{\theta}^n$ *is superefficient w.r.t. the* MLE *and* $\mathcal{R}_1$.

4.3 Alternative Risk Functions

Now we consider in what measure the previous results transpose when changing the risk function. Following [33], we first consider the quadratic cost function and the corresponding risk function:

$$\mathcal{C}_2(\tilde{\theta}^n, \theta_0) = (\tilde{\theta}^n - \theta_0)^2,$$
$$\mathcal{R}_2(\tilde{\theta}^n, \theta_0) = \mathsf{MSE}(\tilde{\theta}^n).$$

The cost function $\mathcal{C}_1$ has the drawback of weighting in the same way points of the parameter space that lie at different distances with respect to the true value θ_0. In many cases, a more general loss function can be considered, as suggested in [30] (Volume 1, page 51) for multiple tests:

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

where $a_j(\theta_0) > 0$ for $j = 1, \ldots, J$ can be tuned in order to give more or less weight to different points of the parameter space. The risk function is therefore given by the weighted probability of misclassification $\mathcal{R}_3(\tilde{\theta}^n, \theta_0) = \sum_{j=1}^J a_j(\theta_0) \cdot \mathbb{P}_{\theta_0}\{\tilde{\theta}^n = \theta_j\}$.

It is trivial to remark that

$$\begin{aligned} &\lim_{n\to\infty} \frac{1}{n} \ln \mathcal{R}_2(\tilde{\theta}^n, \theta_0) \\ &\quad = \lim_{n\to\infty} \frac{1}{n} \ln \mathbb{P}_{\theta_0}(\tilde{\theta}^n \neq \theta_0), \\ &\liminf_{n\to\infty} \frac{1}{n} \ln \sup_{\theta_0\in\Theta} \mathcal{R}_2(\tilde{\theta}^n, \theta_0) \\ &\quad = \liminf_{n\to\infty} \frac{1}{n} \ln \sup_{\theta_0\in\Theta} \mathbb{P}_{\theta_0}(\tilde{\theta}^n \neq \theta_0), \end{aligned}$$

and the lower bounds of Propositions 6 and 7 hold also in this case. The same equalities hold also for $\mathcal{R}_3$. As a result, Proposition 10 and Proposition 11(i) apply also to these risk functions.

On the other hand, as concerns Proposition 9 and Proposition 11(ii), it is simple to show that with respect to the risk functions $\mathcal{R}_2(\tilde{\theta}^n, \theta_0)$ and $\mathcal{R}_3(\tilde{\theta}^n, \theta_0)$, the results hold only asymptotically (see [46], for asymptotic minimax efficiency of the estimator of the integral mean of a Gaussian sample with known variance).

5. PROOFS

PROOF OF PROPOSITION 1. Under A1, Kolmogorov's SLLN implies that $\mathbb{P}_0$-a.s. $\frac{1}{n}\sum_{i=1}^n \ln q(Y_i;\theta_j) \to \mathbb{E}_0 \ln q(Y;\theta_j)$, and for $\mathbb{P}_0$-a.s. any sequence of realizations, $\hat{\theta}^n$ converges to θ_0. Measurability follows from the fact that the following set belongs to $\mathcal{Y}^{\otimes n}$:

$$\begin{aligned} &\left\{\omega \in \Omega \middle| \sup_{\theta\in\Theta} \frac{1}{n} \sum_{i=1}^n \ln q(y_i;\theta) \leq t\right\} \\ &\quad = \bigcap_{\theta_j\in\Theta} \left\{\omega \in \Omega \middle| \frac{1}{n} \sum_{i=1}^n \ln q(y_i;\theta_j) \leq t\right\}. \qquad \square \end{aligned}$$

PROOF OF LEMMA 1. Clearly (ii) implies A2 for a certain $\eta > 0$. On the other hand, suppose that A2 holds; then, applying recursively Hölder inequality:

$$\begin{aligned} \Lambda^{(i)}(\boldsymbol{\lambda}) &\triangleq \ln \mathbb{E}_0\left[\prod_{j=0,\ldots,J,j\neq i} \left(\frac{q(Y;\theta_i)}{q(Y;\theta_j)}\right)^{\lambda_j}\right] \\ &\leq \sum_{j=0,\ldots,J,j\neq i} \frac{1}{J} \cdot \ln \mathbb{E}_0\left[\left(\frac{q(Y;\theta_i)}{q(Y;\theta_j)}\right)^{J\cdot\lambda_j}\right] \end{aligned}$$

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

PROOF OF PROPOSITION 2. The first two results are straightforward applications of Cramér's Theorem in $\mathbb{R}^d$ (see, e.g., [21], Corollary 6.1.6, page 253). Indeed, it is known that the lower bound holds without any supplementary assumption, while the upper bound requires a Cramér condition $\mathbf{0} \in \operatorname{int}(\mathcal{D}_{\Lambda^{(i)}})$; indeed, from Lemma 1, this is equivalent to Assumption A2. Then, a full LDP holds:

$$\begin{aligned}&\liminf_{n\to\infty} \frac{1}{n} \ln \mathbb{P}_0(\hat{\theta}^n = \theta_i)\\ &\quad \geq - \inf_{\mathbf{y}\in \operatorname{int}\mathbb{R}_+^J} \sup_{\boldsymbol{\lambda}\in\mathbb{R}^J} \{\langle \mathbf{y}, \boldsymbol{\lambda}\rangle - \Lambda^{(i)}(\boldsymbol{\lambda})\},\\ &\limsup_{n\to\infty} \frac{1}{n} \ln \mathbb{P}_0(\hat{\theta}^n = \theta_i)\\ &\quad \leq - \inf_{\mathbf{y}\in \mathbb{R}_+^J} \sup_{\boldsymbol{\lambda}\in\mathbb{R}^J} \{\langle \mathbf{y}, \boldsymbol{\lambda}\rangle - \Lambda^{(i)}(\boldsymbol{\lambda})\}.\end{aligned}$$

In order to prove the final result, we have to show that $\mathbb{R}_+^J$ is a $\Lambda^{(i),*}$-*continuity set*, that is, $\inf_{\mathbf{y}\in\operatorname{int}\mathbb{R}_+^J} \Lambda^{(i),*}(\mathbf{y}) = \inf_{\mathbf{y}\in\mathbb{R}_+^J} \Lambda^{(i),*}(\mathbf{y})$. It is enough to apply part (ii) in Lemma on page 903 of [66]. □

PROOF OF PROPOSITION 3. First of all, we note that $\mathbb{P}_0(\hat{\theta}^n \neq \theta_0) = \mathbb{P}_0(\sum_{k=1}^n \mathbf{X}_k \in \operatorname{int}(\mathbb{R}_+^J)^c)$. Therefore, we can apply large deviations principles, with the candidate rate function $\Lambda^*(\mathbf{y})$; this is a strictly convex function on $\operatorname{int}\mathcal{D}_{\Lambda^*}$ globally minimized at

$$\mathbf{y}' = [\mathbb{E}_0(\ln q(Y;\theta_0) - \ln q(Y;\theta_j))]_{j=1,\dots,J}.$$

By Assumption A1, $\mathbf{y}'$ is finite and belongs to $\operatorname{int}\mathbb{R}_+^J$. From the strict convexity of the level sets of $\Lambda^*(\mathbf{y})$, the set $\arg\inf_{\mathbf{y}\in\operatorname{int}(\mathbb{R}_+^J)^c} \Lambda^*(\mathbf{y})$ has at most finite cardinality H. Moreover, since large deviations theory allows us to ignore the part of $\operatorname{int}(\mathbb{R}_+^J)^c$ where $\Lambda^*(\mathbf{y}) \geq \varepsilon + \inf_{\mathbf{y}\in\operatorname{int}(\mathbb{R}_+^J)^c} \Lambda^*(\mathbf{y})$, we can replace $(\mathbb{R}_+^J)^c$ with a collection of H disjoint sets, say Γ_h, $h = 1,\dots,H$, each of them containing in its interior one and only one of the points of $\arg\inf_{\mathbf{y}\in\operatorname{int}(\mathbb{R}_+^J)^c} \Lambda^*(\mathbf{y})$ (see [40], page 508):

$$\begin{aligned}&\mathbb{P}_0\Biggl(\sum_{k=1}^n \mathbf{X}_k \in \operatorname{int}(\mathbb{R}_+^J)^c\Biggr)\\ &\quad = (1+o(1)) \cdot \mathbb{P}_0\Biggl(\sum_{k=1}^n \mathbf{X}_k \in \operatorname{int} \bigcup_{h=1}^H \Gamma_h\Biggr) \qquad (12)\\ &\quad = (1+o(1)) \cdot \sum_{h=1}^H \mathbb{P}_0\Biggl(\sum_{k=1}^n \mathbf{X}_k \in \operatorname{int}\Gamma_h\Biggr).\end{aligned}$$

As before, the bounds derive from Cramér's Theorem in $\mathbb{R}^d$. Noting that the contribution of any Γ_h is the same and recalling (12), we get the results. □

PROOF OF PROPOSITION 4. The assumptions of the theorem on page 904 of [66] are easily verified. This shows that a unique dominating point $\mathbf{y}^{(i)}$ exists and implies, through Proposition on page 161 of [65] (according to the "Remarks on the hypotheses" in [66], page 905, the "lattice" conditions are not necessary), that the stated bracketing of $\mathbb{P}_0(\hat{\theta}^n = \theta_i)$ holds. □

PROOF OF PROPOSITION 5. Under Assumptions A1, A2, A3 and A4, according to Proposition 2(iii) we have $\mathbb{P}_0\{Q_n(\theta_1) \geq Q_n(\theta_0)\} = \mathbb{P}_0\{Q_n(\theta_1) > Q_n(\theta_0)\} \cdot (1+o(1))$ and we can study the behavior of

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

Assumption A8 implies that the conditions of Theorem 3.7.4 in [21] (page 110) are verified, in particular the existence of a positive $\mu \in \operatorname{int}(\mathcal{D}_{\Lambda^{(1)}})$ solution to the equation $0 = (\Lambda^{(1)})'(\mu)$. From Lemma 2.2.5(c) in [21], this implies $\Lambda^{(1)}(\mu) = -\Lambda^{(1),*}(0)$, and the result follows. □

PROOF OF THEOREM 1. We note that the function $\kappa(\cdot)$ in [42] (page 1117) is given by

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

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

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

From (2), we have

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

Now we verify Assumptions (S.1)–(S.4) of [42]. Assumption (S.1) is implied by A2. Assumptions (S.2) and (S.3) hold since the random vectors are i.i.d. and nontrivial. At last, (S.4) is implied by A5 (see, e.g., [72], page 735). Since $\mathbb{E}_0 \mathbf{X}^{(i)}$ is strictly negative by A1, $\operatorname{int} \mathbb{R}_+^J \ominus \mathbb{E}_0 \mathbf{X}^{(i)}$ does not contain $\mathbf{0}$ and, according to Theorem 1 in [42] (page 1118), the result of the theorem follows. □

PROOF OF PROPOSITION 6. First of all, we prove (5). 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 [21] (page 94) with $\alpha_n = \mathbb{P}_{\theta_1}\{\tilde{\theta}^n \neq \theta_1\}$ and $\beta_n = \mathbb{P}_{\theta_0}\{\tilde{\theta}^n \neq \theta_0\}$; since $\tilde{\theta}^n$ is strongly consistent, α_n is ultimately less than any $\varepsilon > 0$ and the bound holds.

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

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

Therefore, we have

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

This implies:

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

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

PROOF OF PROPOSITION 7. From the Neyman–Pearson Lemma, we have

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

for an arbitrary couple of different alternatives θ_0 and θ_1 in Θ. Then we can use Chernoff's Bound ([21], page 93); the final expression derives from the equality $\Lambda^*(0) = -\inf_{\lambda\in\mathbb{R}} \Lambda(\lambda)$. □

PROOF OF PROPOSITION 9. In order to prove that the MLE is admissible and minimax we use the Bayesian method. Using the prior densities given by $\pi(\theta_k) = (J+1)^{-1}$, the Bayes estimator relative to zero–one loss $\check{\theta}^n$ coincides with the MLE $\hat{\theta}^n$. Therefore, respectively from Lemma 2.10 and Proposition 6.3 in [71], $\hat{\theta}^n$ is minimax and admissible. The fact that the MLE minimizes the average probability of error derives from Proposition 8. □

PROOF OF PROPOSITION 10. (i) In order to prove the first statement, we apply Lemma 2.4 in [45] (page 653). Clearly $\mathcal{P}$ is closed in total variation, since it is finite, and is not exponentially convex; indeed, under Assumption A7, 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 [45], there exist $\theta_1', \theta_2' \in \Theta$ such that Equation (2.12) in [45] holds and, as a consequence of Lemma 2.4(i) in [45], 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':

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

and this implies that the Chapman–Robbins bound is not attained at θ_1'.

(ii) The second statement follows easily from the results of [43] (Theorem 2) on $\lim_{n\to\infty} \frac{1}{n} \ln r_1(\tilde{\theta}^n, \pi)$, using Equation (8). Indeed, the MLE attains the lower bound (7) and is therefore asymptotically minimax efficient.

(iii) If the estimator is asymptotically CR-efficient w.r.t. $\mathcal{R}_1$ at θ_0, this means that at θ_0 it is more efficient than the MLE and therefore it has to be less efficient elsewhere (since from Proposition 9 the MLE minimizes the probability of error). Therefore, it cannot be minimax CR-efficient. □

PROOF OF PROPOSITION 11. For (i) it is enough to follow the proof of Proposition 6 and to reason by contradiction, while (ii) is simply another way of stating Proposition 9. □

ACKNOWLEDGMENTS

The authors would like to thank Lucien Birgé, Mehmet Caner, Jean-Pierre Florens, Christian Gouriéroux, Christian Hess, Marc Hoffmann, Pierre Jacob, Søren Johansen, Rasul A. Khan, Oliver B. Linton, Christian P. Robert, Keunkwan Ryu, Igor Vajda and the participants to seminars at Université Paris 9 Dauphine, CREST and Institut Henri Poincaré, to ESEM 2001 in Lausanne, to XXXIVèmes Journées de Statistique 2002 in Bruxelles, to BS/IMSC 2004 in Barcelona, and to ESEM 2004 in Madrid. All the remaining errors are our responsibility.

REFERENCES

[1] BAHADUR, R. R. (1960). On the asymptotic efficiency of tests and estimates. *Sankhyā* **22** 229–252. MR0293767

[2] BAHADUR, R. R. and RANGA RAO, R. (1960). On deviations of the sample mean. *Ann. Math. Statist.* **31** 1015–1027. MR0117775

[3] BARAM, Y. (1978). A sufficient condition for consistent discrimination between stationary Gaussian models. *IEEE Trans. Automat. Control* **23** 958–960.

[4] BARAM, Y. and SANDELL, N. R. JR. (1977). An information theoretic approach to dynamical systems modeling and identification. In *Proceedings of the* 1977 *IEEE Conference on Decision and Control* (*New Orleans*, *La.*, 1977), *Vol.* 1 1113–1118. Inst. Electrical Electron. Engrs., New York. MR0504256

[5] BARAM, Y. and SANDELL, N. R. JR. (1978). Consistent estimation on finite parameter sets with application to linear systems identification. *IEEE Trans. Automat. Control* **23** 451–454. MR0496912

[6] BARAM, Y. and SANDELL, N. R. JR. (1978). An information theoretic approach to dynamical systems modeling and identification. *IEEE Trans. Automat. Control* **AC-23** 61–66. MR0490387

[7] BARNDORFF-NIELSEN, O. (1978). *Information and Exponential Families in Statistical Theory*. Wiley, Chichester. MR0489333

[8] BARRON, A. R. (1985). The strong ergodic theorem for densities: Generalized Shannon–McMillan–Breiman theorem. *Ann. Probab.* **13** 1292–1303. MR0806226

[9] BERGER, J. O. (1993). *Statistical Decision Theory and Bayesian Analysis*. Springer, New York. MR1234489

[10] BLACKWELL, D. and HODGES, J. L. JR. (1959). The probability in the extreme tail of a convolution. *Ann. Math. Statist.* **30** 1113–1120. MR0112197

[11] BLYTH, C. R. (1974). Necessary and sufficient conditions for inequalities of Cramér–Rao type. *Ann. Statist.* **2** 464–473. MR0356333

[12] BLYTH, C. R. and ROBERTS, D. M. (1972). On inequalitites of Cramér–Rao type and admissibility proofs. In *Proceedings of the Sixth Berkeley Symposium on Mathematical Statistics and Probability* (*Univ. California*, *Berkeley*, *Calif.*, 1970/1971), *Vol. I*: *Theory of Statistics* 17–30. Univ. California Press, Berkeley, CA. MR0415822

[13] CAINES, P. E. (1975). A note on the consistency of maximum likelihood estimates for finite families of stochastic processes. *Ann. Statist.* **3** 539–546. MR0368255

[14] CAINES, P. E. (1988). *Linear Stochastic Systems*. Wiley, New York. MR0944080

[15] CHAMBERLAIN, G. (2000). Econometric applications of maxmin expected utility. *J. Appl. Econometrics* **15** 625–644.

[16] CHAPMAN, D. G. and ROBBINS, H. (1951). Minimum variance estimation without regularity assumptions. *Ann. Math. Statist.* **22** 581–586. MR0044084

[17] CHOIRAT, C., HESS, C. and SERI, R. (2003). A functional version of the Birkhoff ergodic theorem for a normal integrand: A variational approach. *Ann. Probab.* **31** 63–92. MR1959786

[18] CLÉMENT, E. (1995). Modélisation statistique en finance et estimation de processus de diffusion. Ph.D. thesis, Université Paris 9 Dauphine.

[19] COX, D. R. and HINKLEY, D. V. (1974). *Theoretical Statistics*. Chapman & Hall, London. MR0370837

[20] DANIELS, H. E. (1954). Saddlepoint approximations in statistics. *Ann. Math. Statist.* **25** 631–650. MR0066602

[21] DEMBO, A. and ZEITOUNI, O. (1998). *Large Deviations Techniques and Applications*, 2nd ed. *Applications of Mathematics* (*New York*) **38**. Springer, New York. MR1619036

[22] FELLER, W. (1968). *An Introduction to Probability Theory*, *Vol.* 1, 3rd ed. Wiley, New York, NY.

[23] FINESSO, L., LIU, C.-C. and NARAYAN, P. (1996). The optimal error exponent for Markov order estimation. *IEEE Trans. Inform. Theory* **42** 1488–1497. MR1426225

[24] FLORENS, J. P. and RICHARD, J. F. (1989). Encompassing in finite parametric spaces. Discussion Paper 89-03. Institute of Statistics and Decision Sciences, Duke University.

[25] FUTSCHIK, A. and PFLUG, G. (1995). Confidence sets for discrete stochastic optimization. *Ann. Oper. Res.* **56** 95–108. MR1339787

[26] GEMAN, S. and HWANG, C.-R. (1982). Nonparametric maximum likelihood estimation by the method of sieves. *Ann. Statist.* **10** 401–414. MR0653512

[27] GERŠANOV, A. M. (1979). Optimal estimation of a discrete parameter. *Teor. Veroyatnost. i Primenen.* **24** 220–224. MR0522259

[28] GERŠANOV, A. M. and ŠAMRONI, S. K. (1976). Randomized estimation in problems with a discrete parameter space. *Teor. Verojatnost. i Primenen.* **21** 195–200. MR0411027

[29] GHOSH, M. and MEEDEN, G. (1978). Admissibility of the mle of the normal integer mean. *Sankhyā Ser. B* **40** 1–10. MR0588734

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

[31] GRENANDER, U. (1981). *Abstract Inference*. Wiley, New York. MR0599175

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

[33] HAMMERSLEY, J. M. (1950). On estimating restricted parameters (with discussion). *J. Roy. Statist. Soc. Ser. B* **12** 192–240. MR0040631

[34] HAWKES, R. M. and MOORE, J. B. (1976). Performance bounds for adaptive estimation. *Proc. IEEE* **64** 1143–1150. MR0429280

[35] HAWKES, R. M. and MOORE, J. B. (1976). Performance of Bayesian parameter estimators for linear signal models. *IEEE Trans. Automat. Control* **AC-21** 523–527. MR0429279

[36] HAWKES, R. M. and MOORE, J. B. (1976). An upper bound on the mean-square error for Bayesian parameter estimators. *IEEE Trans. Inform. Theory* **IT-22** 610–615. MR0416715

[37] HERO, A. E. (1999). Signal detection and classification. In *Digital Signal Processing Handbook* (V. K. Madisetti and D. B. Williams, eds.) Chapter 13. CRC Press, Boca Raton, FL.

[38] HSUAN, F. C. (1979). A stepwise Bayesian procedure. *Ann. Statist.* **7** 860–868. MR0532249

[39] HUBER, P. J. (1972). The 1972 Wald lecture. Robust statistics: A review. *Ann. Math. Statist.* **43** 1041–1067. MR0314180

[40] ILTIS, M. (1995). Sharp asymptotics of large deviations in $\mathbf{R}^d$. *J. Theoret. Probab.* **8** 501–522. MR1340824

[41] JENSEN, J. L. (1995). *Saddlepoint Approximations*. *Oxford Statistical Science Series* **16**. Oxford Univ. Press, New York. MR1354837

[42] JING, B.-Y. and ROBINSON, J. (1994). Saddlepoint approximations for marginal and conditional probabilities of transformed variables. *Ann. Statist.* **22** 1115–1132. MR1311967

[43] KANAYA, F. and HAN, T. S. (1995). The asymptotics of posterior entropy and error probability for Bayesian estimation. *IEEE Trans. Inform. Theory* **41** 1988–1992. MR1385590

[44] KARLIN, S. (1958). Admissibility for estimation with quadratic loss. *Ann. Math. Statist.* **29** 406–436. MR0124101

[45] KESTER, A. D. M. and KALLENBERG, W. C. M. (1986). Large deviations of estimators. *Ann. Statist.* **14** 648–664. MR0840520

[46] KHAN, R. A. (1973). On some properties of Hammersley's estimator of an integer mean. *Ann. Statist.* **1** 756–762. MR0334350

[47] KHAN, R. A. (1978). A note on the admissibility of Hammersley's estimator of an integer mean. *Canad. J. Statist.* **6** 113–119. MR0521655

[48] KHAN, R. A. (2000). A note on Hammersley's estimator of an integer mean. *J. Statist. Plann. Inference* **88** 37–45. MR1767557

[49] KHAN, R. A. (2003). A note on Hammersley's inequality for estimating the normal integer mean. *Int. J. Math. Math. Sci.* **34** 2147–2156.

[50] KLEYWEGT, A. J., SHAPIRO, A. and HOMEM-DE MELLO, T. (2001/02). The sample average approximation method for stochastic discrete optimization. *SIAM J. Optim.* **12** 479–502. MR1885572

[51] KOROSTELEV, A. P. and LEONOV, S. L. (1996). Minimax efficiency in the sense of Bahadur for small confidence levels. *Problemy Peredachi Informatsii* **32** 3–15. MR1441518

[52] LAINIOTIS, D. G. (1969). A class of upper bounds on probability of error for multi-hypothesis pattern recognition. *IEEE Trans. Information Theory* **IT-15** 730–731. MR0276006

[53] LAINIOTIS, D. G. (1969). On a general relationship between estimation, detection, and the Bhattacharyya coefficient. *IEEE Trans. Inform. Theory* **IT-15** 504–505. MR0246692

[54] LAMOTTE, L. R. (2008). Sufficiency in finite parameter and sample spaces. *Amer. Statist.* **62** 211–215. MR2526138

[55] LE CAM, L. (1953). On some asymptotic properties of maximum likelihood estimates and related Bayes' estimates. *Univ. California Publ. Statist.* **1** 277–329. MR0054913

[56] LE CAM, L. and YANG, G. L. (2000). *Asymptotics in Statistics*: *Some Basic Concepts*, 2nd ed. Springer, New York. MR1784901

[57] LINDSAY, B. G. and ROEDER, K. (1987). A unified treatment of integer parameter models. *J. Amer. Statist. Assoc.* **82** 758–764. MR0909980

[58] LIPORACE, L. A. (1971). Variance of Bayes estimates. *IEEE Trans. Inform. Theory* **IT-17** 665–669. MR0339392

[59] LUGANNANI, R. and RICE, S. (1980). Saddle point approximation for the distribution of the sum of independent random variables. *Adv. in Appl. Probab.* **12** 475–490. MR0569438

[60] MANSKI, C. F. (1988). *Analog Estimation Methods in Econometrics*. Chapman & Hall, New York. MR0996421

[61] MCCABE, G. P. JR. (1972). Sequential estimation of a Poisson integer mean. *Ann. Math. Statist.* **43** 803–813. MR0301875

[62] MEEDEN, G. and GHOSH, M. (1981). Admissibility in finite problems. *Ann. Statist.* **9** 846–852. MR0619287

[63] NAFIE, M. and TEWFIK, A. (1998). Reduced complexity M-ary hypotheses testing in wireless communications. In *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing, Seattle, Washington*, 1998, *Vol.* 6 3209–3212. Inst. Electrical Electron. Engrs., New York.

[64] NEWEY, W. K. and MCFADDEN, D. (1994). Large sample estimation and hypothesis testing. In *Handbook of Econometrics*, *Vol*. *IV*. *Handbooks in Econom*. **2** 2111–2245. North-Holland, Amsterdam. MR1315971

[65] NEY, P. (1983). Dominating points and the asymptotics of large deviations for random walk on $\mathbf{R}^d$. *Ann*. *Probab*. **11** 158–167. MR0682806

[66] NEY, P. (1984). Convexity and large deviations. *Ann*. *Probab*. **12** 903–906. MR0744245

[67] NEY, P. (1999). Notes on dominating points and large deviations. *Resenhas* **4** 79–91. MR1712848

[68] NEY, P. E. and ROBINSON, S. M. (1995). Polyhedral approximation of convex sets with an application to large deviation probability theory. *J*. *Convex Anal*. **2** 229–240. MR1363371

[69] POOR, H. V. and VERDÚ, S. (1995). A lower bound on the probability of error in multihypothesis testing. *IEEE Trans*. *Inform*. *Theory* **41** 1992–1994. MR1385591

[70] PUHALSKII, A. and SPOKOINY, V. (1998). On large-deviation efficiency in statistical inference. *Bernoulli* **4** 203–272. MR1632979

[71] ROBERT, C. P. (1994). *The Bayesian Choice*. Springer, New York. MR1313727

[72] ROBINSON, J., HÖGLUND, T., HOLST, L. and QUINE, M. P. (1990). On approximating probabilities for small and large deviations in $\mathbf{R}^d$. *Ann*. *Probab*. **18** 727–753. MR1055431

[73] ROBSON, D. S. (1958). Admissible and minimax integer-valued estimators of an integer-valued parameter. *Ann*. *Math*. *Statist*. **29** 801–812. MR0096339

[74] SILVEY, S. D. (1961). A note on maximum-likelihood in the case of dependent random variables. *J*. *Roy*. *Statist*. *Soc*. *Ser*. *B* **23** 444–452. MR0138158

[75] STARK, A. E. (1975). Some estimators of the integer-valued parameter of a Poisson variate. *J*. *Amer*. *Statist*. *Assoc*. **70** 685–689. MR0395009

[76] TEUNISSEN, P. J. G. (2007). Best prediction in linear models with mixed integer/real unknowns: Theory and application. *J*. *Geod*. **81** 759–780.

[77] TORGERSEN, E. N. (1970). Comparison of experiments when the paramenter space is finite. *Z*. *Wahrsch*. *Verw*. *Gebiete* **16** 219–249. MR0283909

[78] VAJDA, I. (1967). On the statistical decision problems with discrete parameter space. *Kybernetika* (*Prague*) **3** 110–126. MR0215428

[79] VAJDA, I. (1967). On the statistical decision problems with finite parameter space. *Kybernetika* (*Prague*) **3** 451–466. MR0223009

[80] VAJDA, I. (1967). Rate of convergence of the information in a sample concerning a parameter. *Czechoslovak Math*. *J*. **17** **(92)** 225–231. MR0215435

[81] VAJDA, I. (1968). On the convergence of information contained in a sequence of observations. In *Proc*. *Colloquium on Information Theory* (*Debrecen*, 1967), *Vol*. *II* 489–501. János Bolyai Math. Soc., Budapest. MR0258525

[82] VAJDA, I. (1971). A discrete theory of search. I. *Apl*. *Mat*. **16** 241–255. MR0294045

[83] VAJDA, I. (1971). A discrete theory of search. II. *Apl*. *Mat*. **16** 319–335. MR0295483

[84] VAJDA, I. (1974). On the convergence of Bayes empirical decision functions. In *Proceedings of the Prague Symposium on Asymptotic Statistics* (*Charles Univ*., *Prague*, 1973), *Vol*. *II* 413–425. Charles Univ., Prague. MR0383596

[85] VAN DER VAART, A. W. (1997). Superefficiency. In *Festschrift for Lucien Le Cam* 397–410. Springer, New York. MR1462961

[86] VAN DER VLERK, M. H. (1996–2007). Stochastic integer programming bibliography. Available at http://www.eco.rug.nl/mally/biblio/sip.html.

[87] WONG, W. H. (1986). Theory of partial likelihood. *Ann*. *Statist*. **14** 88–123. MR0829557