cemmap

centre for microdata methods and practice

Maximal uniform convergence rates in parametric estimation problems

Walter Beckert
Daniel L. McFadden

The Institute for Fiscal Studies
Department of Economics, UCL

cemmap working paper CWP28/07

E·S·R·C ECONOMIC & SOCIAL RESEARCH COUNCIL

An ESRC Research Centre

Maximal Uniform Convergence Rates in Parametric Estimation Problems

Walter Beckert and Daniel L. McFadden*

23 October 2007

Abstract

This paper considers parametric estimation problems with independent, identically, non-regularly distributed data. It focuses on rate-efficiency, in the sense of maximal possible convergence rates of stochastically bounded estimators, as an optimality criterion, largely unexplored in parametric estimation. Under mild conditions, the Hellinger metric, defined on the space of parametric probability measures, is shown to be an essentially universally applicable tool to determine maximal possible convergence rates. These rates are shown to be attainable in general classes of parametric estimation problems.

JEL Classification: C13, C16

Keywords: parametric estimators, uniform convergence, Hellinger distance, Locally Asymptotically Quadratic (LAQ) Families

Correspondence: w.beckert@bbk.ac.uk, Walter Beckert, School of Economics, Mathematics and Statistics, Birkbeck College, Malet Street, London WC1E 7HX, UK.

*The first author is at Birkbeck College, University of London, and the Institute for Fiscal Studies; the second author is Morris Cox Professor of Economics at UC Berkeley. We are indebted to Masafumi Akahira, Richard Blundell, Andrew Chesher, David Donoho, Hide Ichimura, Oliver Linton and two anonymous referees for helpful comments and discussions.

1 Introduction

A general aim in large-sample statistical inference is to construct estimators that are efficient in the sense that they converge to the target of estimation at the maximum possible rate, and minimize asymptotic dispersion at this rate. Best convergence rates are an important concern in non-parametric estimation problems, where they depend on the dimensionality of the data and the degree of smoothness of the estimation target (e.g., Stone (1980, 1982); Rao (1982); Hall (1989); Donoho and Liu (1992))[1]. In most regular parametric estimation problems, $n^{-\frac{1}{2}}$ is the best uniform convergence rate in a random sample of size n, independent of the dimension of the data or the degree of smoothness of the probability law beyond that required for regularity, and one can concentrate on finding estimators that achieve the Cramér-Rao lower bound for the asymptotic covariance matrix. However, there are exceptions even in some textbook parametric models - the triangular density $f(y;\alpha,\beta) = 2\frac{y-\alpha}{(\beta-\alpha)^2}$, $\alpha \leq y \leq \beta$, has a best rate of $(n\log(n))^{-\frac{1}{2}}$ for α and a best rate of n^{-1} for β, and the quadratic density $f(y;\theta) = 3y(2\theta - y)/4\theta^3$, $0 \leq y \leq 2\theta$, has a best rate of $(n\log(n))^{-\frac{1}{2}}$ for θ.[2]

This paper characterizes optimal uniform convergence rates for non-regular parametric estimation problems, utilizing the concepts of Hellinger distance between probability densities and a Hellinger rate derived from this metric. Our results provide a bridge between econometric textbook analysis of asymptotic efficiency in parametric and semi-parametric estimation and the general treatments of non-parametric convergence rates in the statistics literature. Non-regular parametric estimation problems are not common in econometrics, but they have been receiving increasing attention in the applied literature on auctions (Paarsch (1992)) and the literature on threshold regression models (Chan (1993); Chan and Tsay (1998); Hansen (2000); Seo and Linton (2005)). Hirano and Porter (2003) consider efficient estimation in a class of non-regular models - whose limit experiments are not locally asymptotically normal, but can be approximated by locally shifted max-

[1]For example, Stone (1980) establishes that the best rate in terms of minimizing mean integrated squared error for estimation in a sample of size n of a positive density of dimension m that is continuously differentiable of degree $k \geq 0$ with k-th derivative Lipschitz is $n^{(k+1)/(2k+m+2)}$.

[2]These densities have the regularity properties that they are positive and differentiable to all orders on the interior of their support, but their log likelihoods behave badly at the boundaries of the support. They are shown in Appendix B to be locally asymptotically quadratic, but not locally asymptotically normal, and to admit estimators that attain the best rates. That demonstration illustrates the value of being able to first determine the best rate for a problem, and then to use this rate to test whether the problem is locally asymptotically quadratic.

imum likelihood estimators that achieve asymptotic efficiency; see LeCam (1972, 1986). The econometric literature on efficiency bounds in semi-parametric estimation include Cosslett (1987), Horowitz (1993), Klein and Spady (1993), and Kahn and Tamer (2007).

Our analysis draws upon an extensive literature on convergence rate bounds, particularly methods used by Ibragimov and Has'minskii (1981) for estimation of location parameters and by Donoho and Liu (1991a) for non-parametric density estimation. We address four limitations of the Ibragimov and Has'minskii analysis. First, we obtain convergence rates from properties of the parameterized problem, rather than generic properties of parametric spaces of densities, facilitating application. Second, we avoid a restrictive assumption that estimators be integrable, so that our analysis encompasses locally asymptotically quadratic (LAQ) problems that typically can only be shown to be stochastically bounded (see LeCam (1986); LeCam and Yang (2000); Hajek (1970)). Third, we relax a Hölder assumption on the Hellinger distance to allow cases where the best convergence rate is not necessarily a power of sample size (LeCam and Yang (2000); Prakasa Rao (1968)). Fourth, we are explicit about identification requirements.

A result closely related to this paper is due to Akahira (1991) and Akahira and Takeuchi (1995). These authors show for the case of location parameters in general non-regular models that a maximum bound on the convergence rate of parametric estimators can be deduced from the absolute variation metric, which in turn can be bounded by functions of the Hellinger metric. Our paper can be viewed as an extension of their results to a wider class of parametric estimation problems. Another related result is the analysis of Hellinger distance as a metric for convergence in the context of maximum likelihood (ML) estimation. Van de Geer (1993, 2000) establishes rates of Hellinger consistency of ML estimators under entropy conditions, drawing on the theory of empirical processes (Pollard (1984, 1989)). Entropy-based rates of Hellinger consistency are not guaranteed to be best, however, since entropy, as a measure of the complexity of the set of densities to which the target density belongs, provides an upper bound on squared Hellinger distance and, hence, not a sharp bound on the best possible rate.[3] Moreover, the invoked entropy conditions embed a uniform envelope or dominance condition on the set of densities. This excludes some interesting non-regular cases from the analysis.

The analysis in this paper employs arguments based on the Hellinger distance. The rate at which the distance between two parameter values converges to zero such that the Hellinger distance of an i.i.d. sample is bounded away from zero and one in the limit, henceforth referred to as the Hellinger rate, plays a central role in this analysis. After

[3]See, for example, Van de Geer (2000), example 7.4.6., and Birgé and Massart (1993).

reviewing the definition and main properties of Hellinger distance in Section 2, Section 3 of the paper establishes the existence of unique equivalence classes of Hellinger rates under mild conditions on the data generating process, and it gives necessary and sufficient conditions under which the Hellinger rate does not depend on the parameter value to be estimated. Section 4 connects Hellinger rates to convergence rates in parametric estimation. It establishes that, in the sense of Stone (1980), any attainable rate converges no faster, and no bounding rate converges less fast than the Hellinger rate, and that, in fact, the Hellinger rate constitutes a maximal bounding rate.[4] It also identifies classes of parametric estimation problems in which estimators exist that achieve this bound. Section 5 concludes.

2 Hellinger Distance: Definition and Some Properties

Let $(Y, \mathbf{B}, \mu)$ be a real-valued, σ-finite measure space with Borel σ-field $\mathbf{B}$ and Lebesgue measure μ. Denote by $\{F(y;\theta), \theta \in \Theta\}$ a parametric family of probability measures on $\mathbf{B}$, where the parameter space Θ is an open bounded subset of Euclidean space. In what follows, the scalar case $\Theta \subset \mathbb{R}$ will be considered. Our analysis of the scalar case will also apply when a target of estimation is scalar after re-parametrization of a vector parameter problem[5]. Suppose further that $F(y;\theta)$ is absolutely continuous with respect to Lebesgue measure, and $f(y;\theta)$ is the Radon-Nikodym derivative of $F(y;\theta)$.

Let $h^2(\theta,\theta') = \frac{1}{2}\int_y \left(\sqrt{f(y;\theta)} - \sqrt{f(y;\theta')}\right)^2 dy$ denote the squared Hellinger distance of the parametric densities $f(y;\theta)$ and $f(y;\theta')$, $\theta, \theta' \in \Theta$. Let $H_n^2(\theta,\theta')$ denote the squared Hellinger distance of the densities, evaluated at θ and θ', respectively, of an i.i.d. sample $\{y_i, i = 1,\ldots,n\}$.

The Hellinger metric is of interest because it enjoys a number of convenient properties.

1. Let $\rho(\theta,\theta') = \int_y \sqrt{f(y;\theta)f(y;\theta')}dy$ denote the *affinity* between the densities $f(y;\theta)$ and $f(y;\theta')$; see Matusita (1955). Expanding the square in the definition of Hellinger distance,

$$h^2(\theta,\theta') = 1 - \rho(\theta,\theta').$$

[4]See the following section for a formal definition of attainable, bounding and maximal bounding rate.

[5]Our scalar analysis extends directly to the vector case when all parameters converge at the same rate. In the case with different rates for each vector component, our analysis applies to a linear combination of a vector of parameters, with a rate determined by the least rapidly converging component.

2. For i.i.d. data,

$$\begin{aligned} H_n^2(\theta,\theta') &= \frac{1}{2}\int_{y_1}\cdots\int_{y_n}\left(\sqrt{\prod_{i=1}^n f(y_i;\theta)}-\sqrt{\prod_{i=1}^n f(y_i;\theta')}\right)^2 dy_1\cdots dy_n \\ &= 1-\int_{y_1}\cdots\int_{y_n}\left(\sqrt{\prod_{i=1}^n f(y_i;\theta)f(y_i;\theta')}\right)dy_1\cdots dy_n \\ &= 1-\rho(\theta,\theta')^n \in [0,1]. \end{aligned}$$

Hence, the squared Hellinger distance for i.i.d. data involves a product of affinities.[6]

3. An identification condition that $\rho(\theta_n,\theta_0)\to 1$ only if $\theta_n\to\theta_0$ implies that the Hellinger distance is a metric on the space of root densities, and that $\lim_n H_n^2(\theta,\theta')=1$ for $\theta\neq\theta'$.[7] Since the affinity is related to the sample log-likelihood ratio

$$\Lambda_n(\theta,\theta_0)=\log\left(\prod_{i=1}^n\frac{f(y_i;\theta)}{f(y_i;\theta_0)}\right)$$

by

$$\mathbb{E}\left[\exp\left(\frac{1}{2}\Lambda_n(\theta,\theta_0)\right)\right]=\int\exp\left(\frac{1}{2}\Lambda_n(\theta,\theta_0)\right)\prod_{i=1}^n f(y_i;\theta_0)dy_i=\rho(\theta,\theta_0)^n,$$

this identification condition can be stated equivalently as follows: If θ_n does not converge in probability to θ_0, then $\operatorname{plim}\inf_{n\to\infty}\Lambda_n(\theta_n,\theta_0)=-\infty$.[8]

4. Let σ_n, $n=1,2,\cdots$ be a decreasing sequence of positive scalars with $\lim_n\sigma_n=0$. In a harmless abuse of terminology, σ_n will be called a *convergence rate*. For example, $\sigma_n=n^{-\frac{1}{2}}$ is the convergence rate encountered in regular parametric estimation problems.

[6]Akahira and Takeuchi (1991) define an information measure based on Hellinger affinity, $I_n(\theta,\theta')=-8\log\rho(\theta,\theta')^n$. This measure is interpreted as the information between the product measures of the i.i.d. sample, parameterized by θ and θ', respectively.

[7]Nonnegativity, symmetry and reflexivity are obvious, identity of indiscernibles follows from the identification definition, and the triangle inequality is the same as in the case of the L^2 norm. Among the most frequently used measures of divergence on the space of densities is the Kullback-Leibler divergence; it is not a distance because it is not symmetric. Hellinger distance and Kullback-Leibler divergence are related by

$$H_n^2(\theta,\theta')\leq 1-\exp\left(-\frac{1}{2}KL_n(\theta,\theta')\right).$$

Therefore, convergence of the Kullback-Leibler divergence implies convergence of the Hellinger distance, but not vice versa.

[8]The equivalence is supported by the following argument. Suppose that $\operatorname{plim}\inf_{n\to\infty}\Lambda_n(\theta_n,\theta_0)>-\infty$. Then, with a probability that remains bounded positive, the integrand in the expectation above is bounded positive, implying that $\rho(\theta_n,\theta_0)^n$ is bounded positive, which requires that $\rho(\theta_n,\theta_0)\to 1$.

If convergence rates σ_n and σ'_n satisfy $\liminf_{n\to\infty} \sigma_n/\sigma'_n > 0$, we write $\sigma_n \succeq \sigma'_n$ and say that σ'_n is *at least as fast as* σ_n. If $\sigma_n \succeq \sigma'_n$ and $\sigma'_n \succeq \sigma''_n$, then $\liminf_{n\to\infty} \sigma_n/\sigma''_n \geq (\liminf_{n\to\infty} \sigma_n/\sigma'_n)(\liminf_{n\to\infty} \sigma'_n/\sigma''_n) > 0$, implying $\sigma_n \succeq \sigma''_n$. Then, $\succeq$ is a partial order on convergence rates. If $\sigma_n \succeq \sigma'_n$, but not $\sigma'_n \succeq \sigma_n$, we write $\sigma_n \succ \sigma'_n$ and say that σ'_n is faster than σ_n. We say that a convergence rate σ_n is a *speed limit* on a family of convergence rates D if $\sigma'_n \succeq \sigma_n$ for all $\sigma'_n \in D$. We say σ_n is *maximal* if there exists no convergence rate $\sigma''_n \succ \sigma_n$ that is also a speed limit for D.

The notion of rates of convergence employed in this paper builds on Stone (1980). Let T_n denote a sequence of estimators of $\theta \in \Theta$ that are functionals of an i.i.d. sample of size n drawn from $f(y;\theta)$.

Definition: A convergence rate σ_n is *attainable* if there exists a sequence of estimators T_n whose deviations from the target θ, scaled by σ_n^{-1}, are uniformly stochastically bounded; i.e.

$$\lim_{M\to\infty} \limsup_{n\to\infty} \sup_{\theta\in\Theta} P_\theta(\sigma_n^{-1}|T_n - \theta| > M) = 0.$$

A convergence rate is *bounding* if for every sequence of estimators T_n, deviations from the target θ, scaled by σ_n^{-1}, fail to converge uniformly in probability to zero; i.e.

$$\lim_{M\to 0+} \limsup_{n\to\infty} \sup_{\theta\in\Theta} P_\theta(\sigma_n^{-1}|T_n - \theta| > M) > 0.$$

A convergence rate is *optimal* or *best* if it is both attainable and bounding. We show later (in Lemma 6) that a bounding convergence rate is a speed limit on the family of attainable convergence rates, so that any attainable rate that achieves a speed limit must be optimal, and the achieved speed limit must be maximal.

The main result of the paper exploits the fact that in the i.i.d. case, the limit of the squared Hellinger distance $H_n^2(\theta, T_n)$ and uniform stochastic boundedness of $\sigma_n^{-1}|T_n - \theta|$ can be related via the Cauchy-Schwartz inequality.

3 Hellinger Rates

3.1 Theory

Having reviewed the Hellinger metric and its main properties, this section uses this metric to define Hellinger rates. A sequence of lemmas illuminates conditions under which Hellinger rates exist, form unique equivalence classes and enjoy certain uniformity and invariance properties.

We use the following definition of a Hellinger rate in the i.i.d. case:

Definition: A positive, non-increasing sequence $\delta_n(\theta_0)$ with $\lim_n \delta_n(\theta_0) = 0$ is called a *Hellinger rate* at $\theta_0 \in \Theta$ if

$$0 < \liminf_n H_n^2(\theta_0 \pm \beta\delta_n(\theta_0), \theta_0) \leq \limsup_n H_n^2(\theta_0 \pm \beta\delta_n(\theta_0), \theta_0) < 1 \quad \text{for some } \beta > 0.$$

An immediate consequence of this definition is that any positive scalar multiple of a Hellinger rate is again a Hellinger rate. The following result provides a useful characterization of Hellinger rates:

Lemma 1: *A non-increasing sequence* $\delta_n(\theta_0)$ *with* $\lim_n \delta_n(\theta_0) = 0$ *is a Hellinger rate at* $\theta_0 \in \Theta$ *if and only if*

$$0 < \liminf_n nh^2(\theta_0 \pm \beta\delta_n(\theta_0), \theta_0) \leq \limsup_n nh^2(\theta_0 \pm \beta\delta_n(\theta_0), \theta_0) < \infty \quad \text{for some } \beta > 0.$$

The proofs of this and all subsequent results are given in Appendix A of the paper.

In our analysis of the existence and properties of Hellinger rates, we employ the following definitions:

Definition: A *bimodulus of continuity* of the squared Hellinger distance at $\theta \in \Theta$ is a positive function $\lambda(\delta; \theta)$ defined for $0 < \delta < \nu$, for some $\nu > 0$, coupled with a scalar $\kappa \in (0, 1)$, such that $\lambda(\delta; \theta)$ is increasing in δ with $\lim_{\delta \to 0} \lambda(\delta; \theta) = 0$ and

$$\max\{\kappa\lambda(|\tau|; \theta), \lambda(\kappa|\tau|; \theta)\} \leq h^2(\theta + \tau, \theta) \leq \min\{\lambda(|\tau|; \theta)/\kappa, \lambda(|\tau|/\kappa; \theta)\} \text{ for all } |\tau| < \nu.$$

Definition: A *bimodulus rate* at $\theta \in \Theta$ is a decreasing sequence $\delta_n(\theta)$ implicitly defined by $n\lambda(\delta_n(\theta), \theta) = 1$.

Remark: A bimodulus of continuity bounds the curvature of the Hellinger distance above and below. It extends the conventional definition of modulus of continuity, which gives an upper bound on curvature, and generalizes the bi-Lipschitz property that characterizes isomorphisms of Lipschitz maps. If h^2 is locally quadratic, $h^2(\theta + \tau, \theta) = a(\theta)\tau^2 + o(\tau^2)$ with $a(\theta) > 0$, then $\lambda(\delta; \theta) = a(\theta)\delta^2$ is a bimodulus. If the bimodulus satisfies $\lambda(\delta; \theta) = C(\theta)\delta^\alpha$ with $\alpha > 0$, then $\lambda(\delta; \theta)$ is said to be of a α-Hölder class at θ. In such power cases, the bimodulus is proportional to the square of the inverse of the modulus of continuity $\omega(\epsilon) = \sup\{\| T(f(\cdot; \theta_0)) - T(f(\cdot; \theta)) \| : h(\theta_0, \theta) \leq \epsilon\}$, $\epsilon > 0$, of the functional T over the class $\{f(\cdot; \theta); \theta \in \Theta\}$ in the parametric case $T(f(\cdot; \theta)) = \theta$, as employed in Donoho and Liu (1991a, 1991b).

The requirement in the definition of a bimodulus that $h^2(\theta + \tau, \theta)$ be bounded by both scalar multiples of the function $\lambda(|\tau|; \theta)$ and by this function evaluated at scalar multiples of its first argument is not restrictive for members of the α-Hölder class, where $\kappa\lambda(|\tau|; \theta)$ and $\lambda(\kappa|\tau|; \theta)$ are scalar multiples. However, it restricts candidate bimodulus functions in

cases where $h^2(\theta+\tau,\theta)$ increases at a very rapidly decreasing rate with $|\tau|$ or a very slowly increasing rate with $|\tau|$. An example of the first case is the function $\lambda(|\tau|;\theta) = |\log(|\tau|)|^{-\alpha}$, $\alpha > 0$, which approaches zero less rapidly than $|\tau|^\beta$ for any $\beta > 0$ as $|\tau| \to 0$, and the constraint $\kappa\lambda(|\tau|;\theta) \leq h^2(\theta+\tau,\theta)$ is binding; and an example of the second case is the function $\lambda(|\tau|;\theta) = \exp(-\alpha/|\tau|)$, $\alpha > 0$, which approaches zero more rapidly than $|\tau|^\beta$ for any $\beta > 0$, and the constraint $\lambda(\kappa|\tau|;\theta) \leq h^2(\theta+\tau,\theta)$ is binding. An obvious case where h^2 fails to have a bimodulus with the defined properties is $h^2(\theta+\tau,\theta) = |\tau|^{\alpha'}$ for $\tau > 0$ and $h^2(\theta+\tau,\theta) = |\tau|^{\alpha''}$ for $\tau < 0$, with $0 < \alpha', \alpha'' < 1$ and $\alpha' \neq \alpha''$. Another failure is $h^2(\theta+\tau,\theta) = \exp(-\alpha'/|\tau|)$ for $\tau > 0$ and $h^2(\theta+\tau,\theta) = \exp(-\alpha''/|\tau|)$ for $\tau < 0$ with $\alpha' < \alpha''$, where the candidate function $\lambda(|\tau|;\theta) = \exp(-\alpha'/|\tau|)$ has $\lambda(|\tau|\alpha'/\alpha'';\theta) \leq h^2(\theta+\tau,\theta) \leq \lambda(|\tau|;\theta)$, but fails to satisfy $\kappa\lambda(|\tau|;\theta) \leq h^2(\theta+\tau,\theta)$ for any $\kappa > 0$.

There exists a family of probability measures associated with any candidate bimodulus function $\lambda(\delta;\theta)$. Consider the family $f(y;\theta)$ of uniform densities on the interval with end points 0 and $\text{sgn}(\theta)(1-\lambda(\delta;0))^2$. The affinity between $f(y;\theta)$ and $f(y;0)$ is $\rho(\theta,0) = 1-\lambda(|\theta|;0)$, implying $h^2(\theta,0) = \lambda(|\theta|;0)$. This example illustrates the sensitivity of the Hellinger distance to the choice of parametrization of a family of probability measures. In particular, if a parametric family of probability measures $\{f_\Theta(y;\theta) : \theta \in \Theta\}$ has a bimodulus $\lambda_\Theta(|\tau|;\theta_0)$ at θ_0, then the parameter transformation $\gamma = \psi(\theta-\theta_0) = \text{sgn}(\theta-\theta_0)\lambda_\Theta(|\theta-\theta_0|;\theta_0)^{\frac{1}{2}}$ defined on $\Gamma = \psi(\Theta)$ yields a parametric family of densities $f_\Gamma(y;\gamma) = f_\Theta(y;\psi(\gamma))$ with the quadratic bimodulus $\lambda_\Gamma(|\tau|;0) = |\tau|^2$ at $\gamma = 0$.

A useful class of parametric densities with bounds that translate into a bimodulus for Hellinger distance is described in the following assumption:

A0: $f^{\frac{1}{2}}(y;\theta+\tau)f^{\frac{1}{2}}(y;\theta)$ has an expansion for $A = \{y : f(y,\theta) > 0\}$ and τ in some interval $(-\nu,\nu)$ of the form $f^{\frac{1}{2}}(y;\theta+\tau)f^{\frac{1}{2}}(y;\theta) = f(y;\theta) + q_0(y,\theta,\tau) - q_1(y,\theta,\tau)$, where $q_0(y,\theta,\tau)$ and $q_1(y,\theta,\tau)$ are integrable with $\int_A q_0(y,\theta,\tau)dy = 0$ for all τ and $\int_A q_1(y,\theta,\tau)dy = C(\theta)(|\tau|^\alpha + o(|\tau|^\alpha))$, with $C(\theta) > 0$ and $\alpha > 0$.

If **A0** holds, then $\lambda(\delta;\theta) = C(\theta)\delta^\alpha$ is a Hölder-class bimodulus. If the log density meets classical regularity conditions, as in Example 1 below, then **A0** is satisfied with bimodulus $\lambda(\delta;\theta) = C(\theta)\delta^2$.

To establish the existence of Hellinger rates, the following assumptions will be maintained:

A1: $(Y,\mathbf{B},\mu)$ is a real-valued σ-finite measure space with Borel σ-field $\mathbf{B}$ and Lebesgue measure μ, $\Theta \subset \mathbb{R}$ is an open bounded subset of Euclidean space, $\{F(y;\theta) : \theta \in \Theta\}$ is a family of probability measures that are absolutely continuous with respect to μ,

and $f(y;\theta)$ is the Radon-Nikodym derivative of $F(y;\theta)$.

A2: y_i, for $i = 1, \cdots, n$, is i.i.d. with probability measure $F(y;\theta)$, $\theta \in \Theta$.

A3: (*Identification*) For $\theta_n, \theta_0 \in \Theta$, $\rho(\theta_n, \theta_0) \to 1$ only if $\theta_n \to \theta_0$.

A4: The squared Hellinger distance $h^2(\theta_0, \theta)$ for $\{f(y;\theta), \theta \in \Theta\}$ has bimodulus $\lambda(\delta;\theta)$ at $\theta \in \Theta$; i.e. there exist $\kappa \in (0,1)$ and $\nu > 0$, such that $\lambda(\delta;\theta)$ is finite and increasing in δ with $\lim_{\delta\to 0} \lambda(\delta;\theta) = 0$, and

$$\max\{\kappa\lambda(|\tau|;\theta), \lambda(\kappa|\tau|;\theta)\} \leq h^2(\theta+\tau,\theta) \leq \min\{\lambda(|\tau|;\theta)/\kappa, \lambda(|\tau|/\kappa;\theta)\} \text{ for all } |\tau| < \nu.$$

The following result establishes the existence of Hellinger rates.

Lemma 2: *Under **A1-A4** with bimodulus $\lambda(\delta;\theta_0)$, there exists, for each $\theta_0 \in \Theta$, a bimodulus rate $\delta_n(\theta_0)$ satisfying*

$$1 - \exp(-\kappa) \leq \liminf_n H_n^2(\theta_0 \pm \delta_n(\theta_0), \theta_0) \leq \limsup_n H_n^2(\theta_0 \pm \delta_n(\theta_0), \theta_0) \leq 1 - \exp(-1/\kappa),$$

so that $\delta_n(\theta_0)$ is also a Hellinger rate.

Definition: Two Hellinger rates δ_n' and δ_n'' are *rate equivalent*, denoted $\delta_n' \sim \delta_n''$, if $\delta_n'' \succeq \delta_n'$ and $\delta_n' \succeq \delta_n''$; i.e., each converges at least as fast as the other.

Since $\succeq$ is a partial order, rate-equivalence is an equivalence relation.

The following result establishes that Hellinger rates form unique equivalence classes.

Lemma 3: *If **A1-A4** hold, $\lambda'(\delta;\theta_0)$ and $\lambda''(\delta;\theta_0)$ are bimodulus functions for $h^2(\theta_0+\tau,\theta_0)$, and $\delta_n'(\theta_0)$ and $\delta_n''(\theta_0)$ are respective bimodulus rates at θ_0, then there exists a constant $K > 1$, independent of n, such that $1/K \leq \lambda'(\delta;\theta_0)/\lambda''(\delta;\theta_0) \leq K$ for all $0 < \delta < \nu$, and $\delta_n'(\theta_0)$ and $\delta_n''(\theta_0)$ are rate equivalent, satisfying $1/K \leq \delta_n'(\theta_0)/\delta_n''(\theta_0) \leq K$. Every Hellinger rate at θ_0 is rate equivalent to a bimodulus rate at θ_0, implying that the equivalence class of rate equivalent Hellinger rates at θ_0 is unique.*

To determine the Hellinger rate, Hellinger distance and/or Hellinger affinity need to be calculated. Hence, in order to characterize general properties of Hellinger rates, it seems sensible to deduce them from further conditions on Hellinger distance or affinity and to check in applications whether these conditions are met. The following result provides a necessary and sufficient condition on the bimodulus for the Hellinger rate to be uniform on Θ.

Lemma 4: *Suppose **A1-A4** hold. A necessary and sufficient condition (H) for a Hellinger rate to be uniform on Θ is*

$$0 < \liminf_{\delta\to 0} \lambda(\delta;\theta)/\lambda(\delta;\theta') \leq \limsup_{\delta\to 0} \lambda(\delta;\theta)/\lambda(\delta;\theta') < \infty \text{ for any } \theta, \theta' \in \Theta.$$

This result covers many cases of interest, in particular the case of location and scale parameters, as the following Corollary to Lemma 4 establishes, but also certain cases of shape parameters, as illustrated in the next section.

Corollary 1: *Suppose* **A1-A4** *and condition H hold, and (i) θ is a location parameter, or (ii)* $\exp(\theta)$ *is a scale parameter. Then, the Hellinger rate does not depend on the value of θ.*

Remark: Notice that a Hölder continuity assumption on the Hellinger distance, as in Ibragimov and Has'minskii (1981), of the form

$$h^2(\theta, \theta + |\tau|) \leq \mathbb{E}\left[K(y)\right] |\tau|^\beta \text{ for some } \beta > 0$$

yields

$$\rho(\theta, \theta + |\tau|) \geq 1 - \mathbb{E}\left[K(y)\right] |\tau|^\beta$$

establishing, by the argument of Lemma 1, that $n^{-1/\beta}$ is a lower bound on the Hellinger rate, but not that this bound is achieved uniformly. Thus, the Hölder continuity assumption is not nested by Lemma 4.

The final result in this section shows that the Hellinger rate is invariant under transformations of the random variable that do not depend on the parameter of interest.

Lemma 5: *Suppose that* **A1-A4** *hold. Consider an increasing differentiable transformation $Z = g(Y)$ of the random variable Y which does not depend on θ. Let $\mathcal{D}_Y(\delta_n(\theta))$ and $\mathcal{D}_Z(\delta_n(\theta))$ denote the Hellinger rate equivalence classes based on the random variables Y and Z, respectively. Then, $\mathcal{D}_Y(\delta_n(\theta)) = \mathcal{D}_Z(\delta_n(\theta))$ for all θ.*

3.2 Examples

Example 1: (Regular Case) Suppose for each $\theta \in \Theta$ and for all y, $f^{\frac{1}{2}}(y;\theta)$ is twice continuously differentiable in some neighborhood of $\tau = 0$. Then a Taylor's expansion gives

$$f^{\frac{1}{2}}(y;\theta+\tau) f^{\frac{1}{2}}(y;\theta) = f(y;\theta) + q_0(y,\theta)\tau - q_1(y,\theta)\tau^2 + R(y,\theta,\tau),$$

where

$$\begin{aligned} q_0(y,\theta) &= \frac{\partial}{\partial\theta} f(y;\theta)/2, \\ q_1(y,\theta) &= (\frac{\partial}{\partial\theta^2} f(y;\theta))^2/8f(y;\theta) - \frac{\partial^2}{\partial\theta^2} f(y;\theta)/4, \\ R(y,\theta,\tau) &= (q_1(y,\theta') - q_1(y,\theta))\tau^2, \end{aligned}$$

with θ' a point on the line segment between θ and $\theta+\tau$. Assume that there is an integrable function $g(y,\theta) \geq 0$ that dominates $\frac{\partial}{\partial\theta} f(y;\theta')$, $(\frac{\partial}{\partial\theta} f(y;\theta))^2/f(y;\theta')$, and $\frac{\partial^2}{\partial\theta^2} f(y;\theta')$ for

θ' in a neighborhood of θ; that $g(y,\theta)\cdot o(\tau^2)$ dominates $R(y,\theta',\tau)$ in a neighborhood of θ and of $\tau = 0$; and that $\int q_1(y,\theta)dy = \mathbb{E}\left[\frac{\partial}{\partial\theta}f(y;\theta)/f(y;\theta)\right]^2/8 \equiv C(\theta) > 0$. Then, $|R(y,\theta,\tau)| \leq \int g(y,\theta)dy \cdot o(\tau^2)$. Defining $q_0(y,\theta,\tau) = \frac{\partial}{\partial\theta}f(y;\theta)\tau/2 + \frac{\partial^2}{\partial\theta^2}f(y;\theta)/4$ and $q_1(y,\theta,\tau) = (\frac{\partial}{\partial\theta}f(y;\theta))^2\tau^2/8f(y;\theta) + R(y,\theta,\tau)$, one then has $\int q_1(y,\theta,\tau)dy = C(\theta)(\tau^2 + o(\tau^2))$, so that **A0** is satisfied and $C(\theta)\tau^2$ is a bimodulus. Lemma 1 and 2 then imply that $\delta_n = n^{-\frac{1}{2}}$ is a bimodulus and Hellinger convergence rate, so that all limit points of $H_n^2(\theta \pm \delta_n\beta,\theta)$ are contained in the interior of the unit interval for $0 < \beta < \infty$. This can also be established directly. Integrating the Taylor's expansion term-by-term, $h^2(\theta+\tau,\theta) = C(\theta)(\tau^2 + o(\tau^2))$. Then, for β_n any bounded sequence with limit β,

$$\begin{aligned} H_n^2(\theta+\delta_n\beta_n,\theta) &= 1-(1-h^2(\theta+\delta_n\beta_n,\theta))^n = 1-(1-C(\theta)(\beta_n^2 + n\cdot o(\beta_n^2/n))/n)^n \\ &\to 1-\exp(-C(\theta)\beta^2). \end{aligned}$$

Remark: In this regular case, maximum likelihood estimators are locally asymptotically normal (LAN) and achieve the Hellinger rate $\delta_n = n^{-\frac{1}{2}}$. This is a special case of locally asymptotically quadratic (LAQ) problems treated in Proposition 2 below.

Example 2: (Nonregular cases) Consider the generalized gamma density

$$g(z;\alpha,\beta,\gamma) = z^{\alpha-1}\exp(-z^\beta/\gamma)\beta/\gamma^{\alpha/\beta}\Gamma(\alpha/\beta), \quad z,\alpha,\beta,\gamma,> 0.$$

This density has moments $E[Z^k] = \gamma^{k/\beta}\Gamma((\alpha+k)/\beta)/\Gamma(\alpha/\beta)$, for $k > -\alpha$. Now form the bilateral generalized gamma density about a location parameter θ,

$$f(y;\theta,\alpha,\beta,\gamma) = g(|y-\theta|;\alpha,\beta,\gamma)/2 = |y-\theta|^{\alpha-1}\exp(-|y-\theta|^\beta/\gamma)\beta/2\gamma^{\alpha/\beta}\Gamma(\alpha/\beta).$$

This density has $E[(Y-\theta)^k] = 0$ for k odd, and $E[(Y-\theta)^k] = \gamma^{k/\beta}\Gamma((\alpha+k)/\beta)/\Gamma(\alpha/\beta)$ for k even. The square root of this density is in the same class of functions,

$$f(y;\theta,\alpha,\beta,\gamma)^{\frac{1}{2}} = Cf(y;\theta,(\alpha+1)/2,\beta,2\gamma) = C'|y-\theta|^{\frac{\alpha-1}{2}}\exp(-|y-\theta|^\beta/2\gamma),$$

where $C = \beta^{\frac{1}{2}}2^{\frac{\alpha+1}{2\beta}}\gamma^{\frac{1}{2\beta}}\Gamma((\alpha+1)/2\beta)/\Gamma(\alpha/\beta)^{\frac{1}{2}}$ and $C' = [\beta/2\gamma^{\alpha/\beta}\Gamma(\alpha/\beta)]^{\frac{1}{2}}$. The bilateral generalized gamma includes various cases where conventional regularity conditions leading to $\sqrt{n}$-LAN behavior of maximum likelihood estimators of the location parameter θ are violated. We confine attention to the non-regular cases (1) $\alpha < 1$ and $\beta \leq 1$, in which the density has a pole at $y = \theta$, and (2) $\alpha = 1$ and $\beta < 1$, in which it has a cusp at $y = \theta$. Lemma 4 implies that Hellinger rates in these cases do not depend on the value of θ. Table 1 summarizes the results from an analysis of this example; detailed calculations are provided in Appendix B.2.[9]

α	β	bimodulus	Hellinger rate
$\alpha < 1$	any	$O(\theta^{\alpha})$	$n^{-\frac{1}{\alpha}}$
$\alpha = 1$	$\beta < \frac{1}{2}$	$O(\theta^{2\beta+1})$	$n^{-\frac{1}{2\beta+1}}$
$\alpha = 1$	$\beta = \frac{1}{2}$	$O(\lvert\theta^2 \log(\theta)\rvert)$	$(n \log(n))^{-\frac{1}{2}}$
$\alpha = 1$	$\beta > \frac{1}{2}$	$O(\theta^2)$	$n^{-\frac{1}{2}}$

Table 1: Non-regular bilateral generalized gamma density

Example 3: (shape parameters) Suppose $z \sim u[-1/2, 1/2]$. Consider the transformation $\psi(z, \lambda) = |z|^{\lambda}\text{sgn}(z)$, parameterized by $\lambda \in (0, 1)$, and let $y = \psi(z, 1/\lambda)$. Then, $f(y; \lambda) = \lambda|y|^{\lambda-1}$ with support $[-(1/2)^{\lambda}, (1/2)^{\lambda}]$. For small $|\tau|$ the Hellinger affinity in this example is

$$\begin{aligned} \rho(\lambda, \lambda + |\tau|) &= \int_{-(1/2)^{\frac{1}{\lambda}}}^{(1/2)^{\frac{1}{\lambda}}} \lambda^{\frac{1}{2}} |y|^{\frac{\lambda-1}{2}} (\lambda + |\tau|)^{\frac{1}{2}} |y|^{\frac{\lambda+|\tau|-1}{2}} dy \\ &= \frac{\sqrt{1 + \frac{|\tau|}{\lambda}}}{1 + \frac{|\tau|}{2\lambda}} \left(\frac{1}{2}\right)^{\frac{|\tau|}{2\lambda}}. \end{aligned}$$

Hence, the squared Hellinger distance of an i.i.d. sample is

$$H_n^2\left(\lambda, \lambda + \frac{1}{n}\right) = 1 - \rho\left(\lambda, \lambda + \frac{1}{n}\right)^n = 1 - \frac{\sqrt{1 + \frac{|1/n|}{\lambda}}}{1 + \frac{|1/n|}{2\lambda}} \left(\frac{1}{2}\right)^{\frac{1}{2\lambda}} \to 1 - \left(\frac{1}{2}\right)^{\frac{1}{2\lambda}} \in (0, 1),$$

as $n \to \infty$. In this non-regular example, there is no need to bound the Hellinger distance to establish its convergence to an interior limit for the uniform Hellinger rate of $\frac{1}{n}$. This example is just a special case of a bilateral generalized gamma, for which $\alpha = \lambda$ - the parameter of interest in this example - and $\beta = 0$, $\gamma = 1$ and $\theta = 0$. Note that this is an instance of a multi-parameter problem in which the Hellinger rates differ across parameters.

Another example involving a shape parameter is the density $f(y; \theta) = (1 - y/\theta)^{-\frac{1}{2}}/2\theta$, for $0 \leq y \leq \theta$. The corresponding log likelihood ratio does not have an LAQ expansion.[10]

[9] In cases where the bimodulus does not exist, e.g. the generalized bilateral gamma with different parameters on either side of θ, one can carry out analyses with the respective one-sided bounding functions and, with careful attention regarding uniformity, determine the relevant rate as the minimum of the two resulting rates. In this case, estimators converging at this rate will be stochastically bounded on one side of its support, but not on the other.

[10] Here, as in Example 2 with $\alpha < 1$, the second order term in the expansion of the log likelihood ratio is negative, rather than positive as required by the LAQ definition.

It can be shown that the bimodulus is $\lambda(\delta;\theta) = O(\delta^{\frac{1}{2}})$, so that the Hellinger rate is n^{-2}. An estimator of θ that attains this rate is the maximal order statistic $y_{(n)}$.

4 Maximal Uniform Convergence Rates

This section establishes that, under assumptions **A1-A4** on the data generating process, Hellinger rates are maximal bounding convergence rates.

To provide some intuition for the arguments provided in this section, an illustrative example, summarized in Appendix B.3, may be useful. It suggests that if an estimator T_n attains the Hellinger rate δ_n (i.e. $\delta_n^{-1}(T_n - \theta)$ is stochastically bounded), then the random Hellinger distance $H_n^2(T_n, \theta)$ has a non-degenerate distribution on $[0, 1]$. This contrasts with the case of an estimator T_n' that converges at a rate $\sigma_n \succ \delta_n$, which induces a degenerate limiting distribution of $H_n^2(T_n', \theta)$, placing all probability mass at 1. The main result of this section, in Proposition 1, establishes that the Hellinger rate is a maximal speed limit when **A1-A4** hold and the Hellinger rate is uniform on Θ. Then, an estimator that attains the Hellinger rate uniformly is rate-efficient, and in the terminology of Stone (1980) achieves an optimal or best convergence rate.

The following lemma establishes that bounding rates, defined in Section 2, are speed limits on attainable rates, and it gives a criterion for best rates. Given this result, a useful approach to obtaining rate-efficient estimators is to find a maximal bounding convergence rate, and look for estimators that attain this rate.

Lemma 6: *If σ_n is an attainable convergence rate, and σ_n' is a bounding convergence rate, then $\sigma_n \succeq \sigma_n'$. Hence, bounding convergence rates are speed limits for attainable rates, and a convergence rate σ_n that is both attainable and bounding is best in the sense that there is no faster attainable rate and σ_n is a maximal speed limit.*

The uniformity in Θ of the conditions for attainable and bounding convergence rates is essential. There exist non-uniform "super-convergent" estimators, a variant on Hodges' super-efficient estimators. Suppose a sequence of estimators T_n attains a maximal bounding rate σ_n. Given $\theta_0 \in \Theta$ and a convergence rate σ_n' that is faster than σ_n, define a second sequence of estimators $T_n' = (\sigma_n'/\sigma_n)T_n = (1 - \sigma_n'/\sigma_n)\theta_0$ if $|T_n - \theta_0| < \sigma_n'$ and $T_n' = T_n$ otherwise. At θ_0, this estimator achieves the super-convergent rate σ_n'; i.e. $(T_n' - \theta_0)/\sigma_n' = (T_n - \theta_0)/\sigma_n$ is stochastically bounded.

The Proposition 1 below relates Hellinger rates to maximal bounding rates. It uses the following auxiliary result which uses the Cauchy-Schwartz inequality to bound the squared Hellinger distance.

Lemma 7: *Suppose A1-A3 hold. Let* $\theta_n \in \Theta$, *and* $B_n \in \bigotimes_{i \leq n} \boldsymbol{B}$. *Then, the squared Hellinger distance for i.i.d. data satisfies*

$$H_n^2(\theta_n, \theta) \geq \frac{1}{2}\left(P(B_n;\theta_n)^{\frac{1}{2}} - P(B_n;\theta)^{\frac{1}{2}}\right)^2,$$

where $P(B_n;\theta) \equiv \int_{B_n} \prod_{i=1}^n f(y_i;\theta)dy_i$, *and analogously for* $P(B_n;\theta_n)$.

Proposition 1: *Suppose* **A1-A4** *and condition H hold so that the Hellinger rate* δ_n *is uniform. Then,* δ_n *is a bounding rate and a maximal speed limit on attainable rates, satisfying* $\sigma_n \succeq \delta_n \succeq \sigma_n'$ *for each attainable rate* σ_n *and each bounding rate* σ_n'. *For every* $\theta \in \Theta$ *and some* $\beta > 0$, $\liminf_{n\to\infty} H_n^2(\theta + \beta\sigma_n, \theta) > 0$ *for every attainable rate* σ_n, *while* $\limsup_{n\to\infty} H_n^2(\theta + \sigma_n'\beta, \theta) < 1$ *for every bounding rate* σ_n'.

Comment: The proposition implies that the Hellinger rate constitutes a minimum speed for bounding rates, as well as a speed limit on attainable rates. Thus, an estimator that achieves the Hellinger rate is rate-efficient. It also establishes a necessary condition for convergence rates σ_n to be attainable. One interpretation that can be given to this condition is by its contrapositive: At a rate $\check{\delta}_n$ faster than the Hellinger rate δ_n, $\liminf_n H_n^2(\theta, \theta + \beta\check{\delta}_n) = 0$ for some $\beta > 0$, and hence Proposition 1 leads to the conclusion that no uniformly $\check{\delta}_n^{-1}$-stochastically bounded estimator can exist. Hence, Proposition 1 implies that the Hellinger rate is an upper bound on attainable rates under assumptions **A1-A4** and condition H. This bound may or may not be tight, depending on whether an estimator exists that attains this bound. In light of the construction of Hellinger rates by means of the bimodulus, rate-efficient estimators have $(T_n - \theta_0)/\lambda^{-1}(1/n;\theta_0)$ asymptotically stochastically bounded and non-degenerate. Note that the bimodulus is proportional to the square of the inverse of the modulus of continuity $\omega(\epsilon;\theta_0) = \sup\{|\theta - \theta_0| : h(\theta_0, \theta) \leq \epsilon\}$, $\epsilon > 0$, i.e.

$$\lambda(\omega(\epsilon;\theta_0);\theta_0) = C\epsilon^2, \text{ or } \lambda(\delta;\theta_0) = C(\omega^{-1}(\delta;\theta_0))^2,$$

where C is a positive constant. In fact, all rate derivations in this paper can be obtained by substituting $C(\omega^{-1}(\delta;\theta_0))^2$. Donoho and Liu (1991b) show that, for linear functionals T over a convex class $\{f(\cdot;\theta);\theta \in \Theta\}$, the implied rate $\omega(n^{-\frac{1}{2}})$ is attainable under quite general conditions.[11] The remainder of this section illustrates attainability in some other cases.

Example 2 (continued): Prakasa Rao (1968) shows that, for $\alpha = 1$ and $0 < \beta < 1/2$, the maximum likelihood estimator for the location parameter θ converges at the (inverse)

[11] Donoho and Liu (1991b) also treat some nonlinear cases: estimating the rate of decay and the mode of a density, and robust nonparametric regression.

Hellinger rate $n^{\frac{1}{1+2\beta}}$; i.e. the Hellinger rate forms a tight bound on attainable rates in this case.

Remark: Akahira (1991) and Akahira and Takeuchi (1995) provide a related result for the special case of location parameter families. For $\mathbf{y} = (y_1, \ldots, y_n)'$, they use the absolute variation metric (L^1 norm)[12]

$$d_n(\theta, \theta') = \int_y |f(\mathbf{y}; \theta) - f(\mathbf{y}; \theta')| dy,$$

and show that, if a δ_n^{-1} consistent estimator exists, then, for each $\theta \in \Theta$ and every $\epsilon > 0$, there exists a positive number t_0 such that, for any $t \geq t_0$,

$$\liminf_{n \to \infty} d_n(\theta, \theta - t\delta_n) \geq 2 - \epsilon.$$

Akahira and Takeuchi (1995) show (Lemma 3.5.1) that, for any $\theta, \theta' \in \Theta$,

$$2H_n^2(\theta, \theta') \leq d_n(\theta, \theta') \leq 2\sqrt{2H_n^2(\theta, \theta')},$$

which implies that

$$\frac{1}{8} d_n^2(\theta, \theta') \leq H_n^2(\theta, \theta') \leq \frac{1}{2} d_n(\theta, \theta').$$

Hence, convergence in the Hellinger metric is equivalent to convergence in the absolute variation metric.

Proposition 1 applies in particular to parametric families that are locally asymptotically quadratic (LAQ), in the sense of LeCam (1980) and LeCam and Yang (2000).

Definition: The family of densities $\{f(y; \theta), \theta \in \Theta\}$, Θ open and bounded, is *locally asymptotically quadratic (LAQ)* at $\theta_0 \in \Theta$ at a rate $\delta_n > 0$ satisfying $\delta_n \to 0$, if for any $M > 0$, the log likelihood ratio $\Lambda_n(\theta_0 + \delta_n t, \theta_0)$ satisfies

$$\sup_{|t| \leq M} \left| \Lambda_n(\theta_0 + \delta_n t, \theta_0) - \delta_n S_n(\theta_0) t + \frac{1}{2} \delta_n^2 K_n(\theta_0) t^2 \right| = o_p(1),$$

where $\delta_n S_n(\theta_0)$ is stochastically bounded and non-degenerate (i.e. it does not converge in probability to a constant), and $\delta_n^2 K_n(\theta_0)$ is asymptotically almost surely positive definite (i.e., given $\epsilon > 0$, there exists $\kappa > 0$ such that $\liminf_{n \to \infty} P(1/\kappa \leq \delta_n^2 K_n(\theta_0) \leq \kappa; \theta_0) > 1 - \epsilon$).

A LAQ family is *locally asymptotically normal (LAN)* if in addition $\delta_n = n^{-\frac{1}{2}}$, $\delta_n S_n(\theta_0)$ is asymptotically normal, and $\delta_n^2 K_n(\theta_0)$ converges in probability to a constant. The regular case given in Example 1 above is LAN. Appendix B.1 shows that the triangular

[12] See also Hoeffding and Wolfowitz (1958) for a discussion of the properties of this metric.

and quadratic densities given in the introduction are LAQ, but not LAN. The bilateral generalized gamma family in Example 2 with $\alpha < 1$ fails to be LAQ.

The next proposition establishes for a family that is LAQ at a uniform Hellinger rate that approximate maximum likelihood estimators attain this rate, and gives conditions under which an estimator obtained in one step from an initially consistent estimator attains this rate.

Proposition 2: *Assume **A1-A4** and condition H, so that the Hellinger rate δ_n is uniform. Assume for each $\theta_0 \in \Theta$ that the log likelihood ratio is LAQ at the rate δ_n. Then:*

(1) The infeasible estimator $\theta_{n0} = \theta_0 + K_n(\theta_0)^{-1} S_n(\theta_0)$ achieves the Hellinger rate.

(2) Assume the property (M) that θ_{nm} is a sequence of approximate maximum likelihood estimators satisfying $P\left(\sup_{\theta \in \Theta} \Lambda_n(\theta, \theta_{nm}) > \gamma_n; \theta_0\right) \leq \zeta_n$, where γ_n is a positive sequence satisfying $\gamma_n \to 0$, and ζ_n is a positive sequence satisfying $\sum_{n=1}^{\infty} \zeta_n < +\infty$. Then, θ_{nm} converges almost surely to θ_0 and is asymptotically equivalent to θ_{n0}, i.e. $\delta_n^{-1}(\theta_{nm} - \theta_{n0}) = o_p(1)$, so that it attains the Hellinger rate.

(3) Assume the property (S) that $\delta_n^2 K_n(\theta)$ satisfies a stochastic Hölder condition in a neighborhood of each $\theta_0 \in \Theta$ that bounds the error in the LAQ approximation to the log likelihood ratio; i.e., given $\epsilon > 0$, there exist a neighborhood Θ' of θ_0 and scalars $M > 0$ and $\psi > 0$ such that

$$\liminf_{n \to \infty} P\left(\left|K_n(\theta) - K_n(\theta')\right| \leq \delta_n^{-2} M |\theta - \theta'|^{\psi} \text{ for all } \theta, \theta' \in \Theta'; \theta_0\right) > 1 - \epsilon,$$

and

$$\liminf_{n \to \infty} P\left(\left|\Lambda_n(\theta, \theta') - S_n(\theta')(\theta - \theta') + K_n(\theta')(\theta - \theta')^2/2\right| \leq \delta_n^{-2} M |\theta - \theta'|^{2+\psi} \text{ for all } \theta, \theta' \in \Theta'; \theta_0\right) > 1 - \epsilon.$$

Assume that there exists an initially consistent estimator θ_{n1} for θ_0 that attains a convergence rate δ_n' satisfying $\delta_n^{-1}(\delta_n')^{1+\psi} = o(1)$. Then, the one-step estimator $\theta_{n2} = \theta_{n1} + K_n(\theta_{n1})^{-1} S_n(\theta_{n1})$ achieves the Hellinger rate, and is asymptotically equivalent to θ_{n0}.

Comment: If a maximum likelihood estimator is achieved at a finite log likelihood ratio almost surely for a family of densities with the LAQ property, then property (M) in result (2) holds for this estimator. More generally, if the log likelihood function has a finite supremum almost surely, then (M) admits estimators that come within γ_n of achieving this supremum. Result (2) continues to hold if the log likelihood has an infinite supremum,

but θ_{nm} can be selected so that with probability one, eventually $\Lambda_n(\theta_{nm}, \theta_0) \geq -\gamma_n \to 0$. The Hölder property (S) in (3) implies that $\delta_n^2 K_n(\theta)$ is stochastically equicontinuous in a neighborhood of θ_0. The assumption $\delta_n^{-1}(\delta_n')^{1+\psi} \to 0$ is satisfied, for example, if $\delta_n' = n^{-\frac{1}{2}}$ and $\delta_n = (n \log(n))^{-\frac{1}{2}}$, as in the case of the α parameter in the triangular density and the quadratic density given in the introduction. Appendix B.1 shows that these densities are LAQ, and satisfy the conditions of (3), so that there exist one-step estimators for these families that achieve the Hellinger rate.

5 Conclusions

This paper considers rate efficiency in parametric estimation as a criterion to judge the quality of estimators, next to other efficiency criteria, such as e.g. the Cramér Rao bound, within a given class of estimators converging at a specific rate, e.g. $\sqrt{n}$. It addresses the question of what convergence rates parametric estimators can attain in parametric estimation problems with i.i.d. data. The Hellinger metric is proposed as a very convenient tool to identify the Hellinger rate as an upper bound on attainable rates and thereby as a benchmark or gold standard for rate-efficiency. The paper also identifies classes of parametric estimation problems in which this bound is tight, i.e. in which the Hellinger rate is the maximal attainable rate.

This work deals only with scalar parameters of interest, or with parameter vectors whose components converge at the same rate. Future work might deal with cases like Examples 2 and 3, in which different components of a parameter vector converge at different rates, and the rates of convergence of one depend on the other; and with the case of dependent data, where convergence rates may depend on the value of the parameter of interest.[13]

A Proofs

A.1 Lemma 1

We use the elementary analytic result that for any real sequence α_n,

$$\exp(-\limsup_n \alpha_n) = \liminf_n (1 - \alpha_n/n)^n \leq \limsup_n (1 - \alpha_n/n)^n = \exp(-\liminf_n \alpha_n).$$

[13] An example is, for instance, the case of the parameter of an autoregressive process of order 1. In the unit root case, estimators converge at rate T, while otherwise they converge at rate $\sqrt{T}$, where T is the sample size.

Defining $\alpha_n = nh^2(\theta_0 \pm \beta\delta_n(\theta_0), \theta_0)$, one has

$$H_n^2(\theta_0 \pm \beta\delta_n(\theta_0), \theta_0) = 1 - \rho(\theta_0 \pm \beta\delta_n(\theta_0), \theta_0)^n = 1 - (1 - \alpha_n/n)^n.$$

Then, $\delta_n(\theta_0)$ is a Hellinger rate at θ_0 if and only if

$$0 < \liminf_n H_n^2(\theta_0 \pm \beta\delta_n(\theta_0), \theta_0) = \exp(-\limsup_n nh^2(\theta_0 \pm \beta\delta_n(\theta_0), \theta_0)),$$

and

$$1 > \limsup_n H_n^2(\theta_0 \pm \beta\delta_n(\theta_0), \theta_0) = \exp(-\liminf_n nh^2(\theta_0 \pm \beta\delta_n(\theta_0), \theta_0)),$$

proving the result. □

A.2 Lemma 2

The rate $\delta_n(\theta_0)$ solves the equation $n\lambda(\delta_n(\theta_0); \theta_0) = 1$. Then $\lim_n \delta_n(\theta_0) = 0$, implying $\delta_n(\theta_0) \leq \nu$ and $\theta_n = \theta_0 \pm \delta_n(\theta_0) \in \Theta$ for n sufficiently large. From the definition of a bimodulus at θ_0, $\kappa\lambda(\delta_n(\theta_0); \theta_0) \leq h^2(\theta_n, \theta_0) \leq \lambda(\delta_n(\theta_0); \theta_0)/\kappa$, and hence

$$[1 - n\lambda(\delta_n(\theta_0); \theta_0)/\kappa n]^n \leq [1 - h^2(\theta_n, \theta_0)]^n \leq [1 - \kappa n\lambda(\delta_n(\theta_0); \theta_0)]^n.$$

Taking the limit in n of this expression using the analytic result in the proof of Lemma 1 implies that the limit points of $H_n^2(\theta_n, \theta_0) = 1 - [1 - h^2(\theta_n, \theta_0)]^n$ are bracketed by $1 - \exp(-\kappa)$ and $1 - \exp(-1/\kappa)$. □

A.3 Lemma 3

The bimodulus rates $\delta_n'(\theta_0)$ and $\delta_n''(\theta_0)$ satisfy $n\lambda'(\delta_n'; \theta_0) = n\lambda''(\delta_n''; \theta_0) = 1$ for the bimodulus functions λ' and λ''. The bimodulus inequalities imply $\lambda'(\kappa'\delta_n''(\theta_0); \theta_0) \leq \lambda''(\delta_n''(\theta_0); \theta_0) = 1/n = \lambda'(\delta_n'(\theta_0); \theta_0)$, and hence $\kappa'\delta_n''(\theta_0) \leq \delta_n'(\theta_0)$. Similarly, one has $\lambda''(\kappa''\delta_n'(\theta_0); \theta_0) \leq \lambda'(\delta_n'(\theta_0); \theta_0) = 1/n = \lambda''(\delta_n''(\theta_0); \theta_0)$, implying $\kappa''\delta_n'(\theta_0) \leq \delta_n''(\theta_0)$. Therefore, $\kappa' \leq \delta_n'(\theta_0)/\delta_n''(\theta_0) \leq 1/\kappa''$. Then, $\delta_n'(\theta_0)$ and $\delta_n''(\theta_0)$ are rate equivalent. The bimodulus inequalities also imply $\kappa'\lambda'(\delta; \theta_0) \leq \lambda''(\delta; \theta_0)/\kappa''$ and $\kappa''\lambda''(\delta; \theta_0) \leq \lambda'(\delta; \theta_0)/\kappa'$, so that $\kappa'\kappa'' \leq \lambda'(\delta; \theta_0)/\lambda''(\delta; \theta_0) \leq 1/\kappa'\kappa''$. So, $K = 1/\kappa'\kappa'' > 1$.

If, at θ_0, $\delta_n'(\theta_0)$ is a Hellinger rate, and $\delta_n(\theta_0)$ is the bimodulus rate for a bimodulus $\lambda(\delta; \theta_0)$, then for some $\beta > 0$, $n\lambda(\kappa\beta\delta_n'(\theta_0); \theta_0) \leq nh^2(\theta_0 \pm \beta\delta_n'(\theta_0); \theta_0) \leq n\lambda(\beta\delta_n'(\theta_0)/\kappa; \theta_0)$. Then, Lemma 1 implies that there are positive constants α and γ such that the inequalities $n\lambda(\kappa\beta\delta_n'(\theta_0); \theta_0) \leq \gamma$ and $n\lambda(\beta\delta_n'(\theta_0)/\kappa; \theta_0) \geq \alpha$ hold for n sufficiently large. Then, there exist scale factors $\zeta, \eta > 0$ such that $n\lambda(\zeta\kappa\beta\delta_n'(\theta_0); \theta_0) \leq 1 = n\lambda(\delta_n(\theta_0); \theta_0)$, implying $\delta_n'(\theta_0) \leq \delta_n(\theta_0)/\zeta\kappa\beta$, and $n\lambda(\delta_n(\theta_0); \theta_0) = 1 \geq n\lambda(\eta\beta\delta_n'(\theta_0)/\kappa; \theta_0)$, implying

$\delta'_n(\theta_0) \geq \delta_n(\theta_0)\kappa/\eta\beta$. Then, $\delta'_n(\theta_0)$ and $\delta_n(\theta_0)$ are rate equivalent, and the bimodulus rate for any bimodulus defines a unique equivalence class. □

A.4 Lemma 4

From Lemma 1, δ_n is a uniform Hellinger rate if and only if, for some $\beta > 0$ and any $\theta \in \Theta$,

$$0 < \alpha(\theta) \equiv \liminf_n nh^2(\theta \pm \beta\delta_n, \theta) \leq \limsup_n nh^2(\theta \pm \beta\delta_n, \theta) \equiv \gamma(\theta) < \infty.$$

The bimodulus at θ satisfies $\kappa n\lambda(\delta;\theta) \leq nh^2(\theta \pm \delta, \theta) \leq n\lambda(\delta;\theta)/\kappa$. If δ_n is a uniform Hellinger rate, then, for any $\theta \in \Theta$, $\kappa\alpha(\theta) \leq n\lambda(\beta\delta_n;\theta) \leq \gamma(\theta)/\kappa$. Hence, $\kappa^2\alpha(\theta)/\gamma(\theta) \leq \lambda(\beta\delta_n;\theta)/\lambda(\beta\delta_n;\theta') \leq \gamma(\theta)/\kappa^2\alpha(\theta)$, implying (H).

Alternately, if (H) holds, and $\delta_n(\theta)$ is the bimodulus rate at θ, and hence by Lemma 3 a Hellinger rate, one has

$$\begin{aligned}
(\kappa^2\alpha(\theta')/\gamma(\theta))\kappa n\lambda(\beta\delta_n(\theta);\theta) &\leq \kappa n\lambda(\beta\delta_n(\theta);\theta') \\
&\leq nh^2(\theta' \pm \beta\delta_n(\theta), \theta') \\
&\leq n\lambda(\beta\delta_n(\theta);\theta')/\kappa \\
&\leq (\gamma(\theta')/\kappa^2\alpha(\theta))n\lambda(\beta\delta(\theta);\theta)/\kappa.
\end{aligned}$$

Then, by Lemma 2, $\delta_n(\theta)$ is a Hellinger rate (with common β) for all $\theta' \in \Theta$. □

A.5 Corollary 1

Case (i) follows from the transformation $y' = y - \theta$, yielding

$$\rho(\theta + \delta, \theta) = \int f^{\frac{1}{2}}(y - \theta - \delta)f^{\frac{1}{2}}(y - \theta)dy = \int f^{\frac{1}{2}}(y' - \delta)f^{\frac{1}{2}}(y')dy' = \rho(\delta, 0).$$

Case (ii) follows from the transformation $y' = y\exp(-\theta)$, yielding

$$\begin{aligned}
\rho(\theta + \delta, \theta) &= \int f^{\frac{1}{2}}(y\exp(-\theta - \delta))f^{\frac{1}{2}}(y\exp(-\theta))dy\exp(-\theta - \delta/2) \\
&= \exp(-\delta/2)\int f^{\frac{1}{2}}(y'\exp(-\delta))f^{\frac{1}{2}}(y')dy' \\
&= \rho(\delta, 0).
\end{aligned}$$

□

A.6 Lemma 5

Since $f_y(y;\theta) = f_Z(g(y);\theta)g'(y)$, where $g'(y)$ denotes the derivative of $g(y)$,

$$\begin{aligned}
\rho_Y(\theta+\delta,\theta) &= \int f^{\frac{1}{2}}(y;\theta+\delta) f_Y^{\frac{1}{2}}(y;\theta) dy \\
&= \int f_Z(g(y);\theta+\delta) f_Z(g(y);\theta) g'(y) dy \\
&= \int f_Z(z;\theta+\delta) f_Z(z;\theta) dz \\
&= \rho_Z(\theta+\delta,\theta).
\end{aligned}$$

Then, the Hellinger distance at $\theta+\delta$ and θ is invariant under differentiable one-to-one transformations of the random variable, and consequently the Hellinger rates are also invariant. □

A.7 Lemma 6

By the definitions of attainable and bounding rates, given $\epsilon > 0$, there exist n', M and M' positive such that for $n > n'$,

$$\sup_{\theta\in\Theta} P_\theta(\sigma_n^{-1}|T_n-\theta| > M) < \epsilon \text{ and } \sup_{\theta\in\Theta} P_\theta(\sigma_n^{'-1}|T_n-\theta| > M') > \epsilon.$$

But the first condition implies $\sup_{\theta\in\Theta} P_\theta(\sigma_n^{'-1}|T_n-\theta| > M\sigma_n/\sigma_n') < \epsilon$ for $n > n'$, and hence $M\sigma_n/\sigma_n' > M'$, implying $\sigma_n/\sigma_n' > M'/M$. □

A.8 Lemma 7

For $B_n \in \bigotimes_{i\le n} \mathbf{B}$ and $\theta_n \in \Theta$,

$$\begin{aligned}
H_n^2(\theta_n,\theta) &= \frac{1}{2}\int_{y_1}\cdots\int_{y_n}\left(\sqrt{\prod_{i=1}^n f(y_i;\theta_n)} - \sqrt{\prod_{i=1}^n f(y_i;\theta)}\right)^2 dy_1\cdots dy_n \\
&\ge \frac{1}{2}\int_{B_n}\left(\sqrt{\prod_{i=1}^n f(y_i;\theta_n)} - \sqrt{\prod_{i=1}^n f(y_i;\theta)}\right)^2 dy_1\cdots dy_n \\
&= \frac{1}{2}P_{\theta_n}(B_n) + \frac{1}{2}P_\theta(B_n) - \int_{B_n}\sqrt{\prod_{i=1}^n f(y_i;\theta_n) f(y_i;\theta)}\, dy_1\cdots dy_n.
\end{aligned}$$

By the Cauchy-Schwartz inequality,

$$\begin{aligned}\int_{B_n}\sqrt{\prod_{i=1}^{n}f(y_i;\theta_n)f(y_i;\theta)}dy_1\cdots dy_n &\leq \left[\int_{B_n}\prod_{i=1}^{n}f(y_i;\theta_n)dy_i\right]^{\frac{1}{2}}\left[\int_{B_n}\prod_{i=1}^{n}f(y_i;\theta)dy_i\right]^{\frac{1}{2}}\\ &= P(B_n;\theta_n)^{\frac{1}{2}}P(B_n;\theta)^{\frac{1}{2}}.\end{aligned}$$

Hence,

$$H_n^2(\theta_n,\theta)\geq\frac{1}{2}P(B_n;\theta_n)+\frac{1}{2}P(B_n;\theta)-P(B_n;\theta_n)^{\frac{1}{2}}P(B_n;\theta)^{\frac{1}{2}}=\frac{1}{2}\left(P(B_n;\theta_n)^{\frac{1}{2}}-P(B_n;\theta)^{\frac{1}{2}}\right)^2$$

□

A.9 Proposition 1

The sequence of sample spaces with their respective product σ-fields that contain events such as $B_{nM}=\{\mathbf{y}_n\in Y^n:\sigma_n^{-1}|T_n(\mathbf{y}_n)-\theta|<M\}$, where $\mathbf{y}_n=(y_1,\cdots,y_n)$, are all embedded in the infinite product space Y^∞ with its product σ-field $\mathbf{B}_\infty$, the σ-field generated by all cylinders of the form $Y^{N''}\times\bigotimes_{t\in N'}C_t$, where N' is a finite subset of the positive integers, N'' is its complement, and $C_t\in\mathbf{B}$. By the Kolmogorov extension theorem, there is a unique probability $P(\cdot;\theta)$ on Y^∞ that extends all the sample probabilities $P_n(\cdot;\theta)$; i.e. rewrite

$$B_{nM}(\theta)=\{\mathbf{y}\in Y^\infty:\sigma_n^{-1}|T_n(\mathbf{y}_n)-\theta|<M\},$$

where $\mathbf{y}=(y_1,y_2,\cdots)$, and

$$P_n(\{\mathbf{y}_n\in Y^n:\sigma_n^{-1}|T_n(\mathbf{y}_n)-\theta|<M;\theta)=P(B_{nM}(\theta);\theta).$$

For σ_n an attainable rate, given $0<\epsilon<1/3$, there exists a sequence of estimators $T_n(\mathbf{y}_n)$ and a constant $M>0$ such that for all $n>n_0$ and all $\theta\in\Theta$,

$$P(B_{nM}(\theta);\theta)>1-\epsilon.$$

Define $\theta_{n+}=\theta+2\sigma_nM$ and $\theta_{n-}=\theta-2\sigma_nM$. Then,

$$\begin{aligned}B_{nM}(\theta) &= \{\mathbf{y}\in Y^\infty:-M<\sigma_n^{-1}(T_n(\mathbf{y}_n)-\theta)<M\}\\ &= \{\mathbf{y}\in Y^\infty:-3M<\sigma_n^{-1}(T_n(\mathbf{y}_n)-\theta_{n+})<-M\}\subseteq Y^\infty\setminus B_{nM}(\theta_{n+})\\ &= \{\mathbf{y}\in Y^\infty:M<\sigma_n^{-1}(T_n(\mathbf{y}_n)-\theta_{n-})<3M\}\subseteq Y^\infty\setminus B_{nM}(\theta_{n-}).\end{aligned}$$

Then, $P(B_{nM}(\theta);\theta_{n+})\leq\epsilon$ and $P(B_{nM}(\theta);\theta_{n-})\leq\epsilon$, implying

$$\begin{aligned}P(B_{nM}(\theta);\theta)-P(B_{nM}(\theta);\theta_{n+}) &> 1-2\epsilon>\epsilon,\\ P(B_{nM}(\theta);\theta)-P(B_{nM}(\theta);\theta_{n-}) &> 1-2\epsilon>\epsilon.\end{aligned}$$

Therefore, by Lemma 7, for all $n > n_0$ and all $\theta \in \Theta$,

$$\begin{aligned} H_n^2(\theta + 2\sigma_n M; \theta) &\geq \left(P(B_{nM}(\theta); \theta)^{\frac{1}{2}} - P(B_{nM}(\theta); \theta_{n+})^{\frac{1}{2}}\right)^2 \\ &\geq \min_{0 \leq p \leq \epsilon} \left((p+\epsilon)^{\frac{1}{2}} - p^{\frac{1}{2}}\right)^2 \geq \epsilon/6, \\ H_n^2(\theta - 2\sigma_n M; \theta) &\geq \left(P(B_{nM}(\theta); \theta)^{\frac{1}{2}} - P(B_{nM}(\theta); \theta_{n-})^{\frac{1}{2}}\right)^2 \\ &\geq \min_{0 \leq p \leq \epsilon} \left((p+\epsilon)^{\frac{1}{2}} - p^{\frac{1}{2}}\right)^2 \geq \epsilon/6. \end{aligned}$$

From the bimodulus inequalities, this implies

$$\begin{aligned} \epsilon/6 &\leq H_n^2(\theta \pm 2\sigma_n M, \theta) = 1 - \rho(\theta \pm 2\sigma_n M, \theta)^n \\ &= 1 - (1 - h^2(\theta \pm 2\sigma_n M, \theta))^n \\ &\leq 1 - (1 - n\lambda(2\sigma_n M, \theta)/\kappa n)^n. \end{aligned}$$

Therefore,

$$\epsilon/6 \leq \lim_n H_n^2(\theta \pm 2\sigma_n M, \theta) \leq 1 - \exp(-\liminf_n n\lambda(2\sigma_n M, \theta)/\kappa),$$

implying that $\liminf_n n\lambda(2\sigma_n M, \theta) > 0$ for all $\theta \in \Theta$. Since the Hellinger rates $\delta_n \sim \delta_n'$ for rate equivalent bimodulus rates δ_n' satisfying $n\lambda(\delta_n', \theta) = 1$, $\sigma_n' \prec \delta_n'$ implies $\liminf_n nh^2(\theta \pm 2\sigma_n' M, \theta) = 0$ by Lemmas 1 and 3. This implies that $\liminf_n n\lambda(2\sigma_n' M, \theta) = 0$ for all $\theta \in \Theta$. Hence, $\sigma_n \succeq \delta_n' \sim \delta_n$, i.e. $\sigma_n \succeq \delta_n$ and the Hellinger rate is a speed limit on attainable rates.

Since every bounding rate σ_n' is declining at least as rapidly as any attainable rate, $\delta_n \succeq \sigma_n'$ and, hence, $\limsup_n H_n^2(\theta + \sigma_n' M, \theta) < 1$. By the definition of Hellinger rate,

$$\lim_{M \to 0+} \liminf_{n \to \infty} \sup_\theta \rho(\theta, \theta + \delta_n M)^n = 1,$$

and therefore

$$\lim_{M \to 0+} \limsup_{n \to \infty} \sup_\theta H_n^2(\theta, \theta + \delta_n M) = 0.$$

Then, the above arguments imply that

$$\begin{aligned} \lim_{M \to 0+} \liminf_{n \to \infty} \sup_\theta P(B_{nM}(\theta); \theta) &= \lim_{M \to 0+} \liminf_{n \to \infty} \sup_\theta P(B_{nM}(\theta); \theta_{n+}) \\ &= \lim_{M \to 0+} \liminf_{n \to \infty} \sup_\theta P(B_{nM}(\theta); \theta_{n-}). \end{aligned}$$

Since, as $M > 0$ tends to zero, $\liminf_n \sup_\theta P(B_{nM}(\theta); \theta)$ is non-increasing, while, on the other hand, $\liminf_n \sup_\theta P(B_{nM}(\theta); \theta_{n+})$ and $\liminf_n \sup_\theta P(B_{nM}(\theta); \theta_{n-})$ are non-decreasing, the last set of equalities implies that

$$0 < \lim_{M \to 0+} \liminf_{n \to \infty} \sup_\theta P(B_{nM}(\theta); \theta) < 1.$$

Hence, δ_n is itself a bounding rate. It is also attainable and, therefore, a maximal bounding rate. □

A.10 Proposition 2

(1) The LAQ conditions that $\delta_n S_n(\theta_0)$ is stochastically bounded and non-degenerate and that $\delta_n^2 K_n(\theta_0)$ is asymptotically almost surely positive definite imply by construction that for any $\epsilon > 0$, $\delta_n^{-1}(\theta_{n0} - \theta_0)$ is bracketed by the stochastically bounded expressions $\delta_n S_n(\theta_0)/\kappa$ and $\delta_n S_n(\theta_0)\kappa$ with probability at least $1-\epsilon$, and is therefore stochastically bounded.

(2) By the Borel-Cantelli Lemma, the conditions $P\left(\sup_{\theta\in\Theta} \Lambda_n(\theta, \theta_{nm}) > \gamma_n; \theta_0\right) \leq \zeta_n$ and $\sum_{n=1}^{\infty} \zeta_n < +\infty$ imply that, with probability one, eventually $\Lambda_n(\theta_{nm}, \theta_0) \geq -\gamma_n$. In this event,

$$\rho(\theta_{nm}, \theta_0)^n = \mathbb{E}\left[\exp\left(\frac{1}{2}\Lambda_n(\theta, \theta_0)\right)\Big|_{\theta=\theta_{nm}}\right] \geq \exp\left(-\frac{1}{2}\gamma_n\right) \to 1,$$

and **A3** implies that with probability one, θ_{nm} converges to θ_0. In this event, the LAQ expansion implies

$$\begin{aligned}
\gamma_n &\geq -\Lambda_n(\theta_{nm}, \theta_{n0}) = -\Lambda_n(\theta_{nm}, \theta_0) + \Lambda_n(\theta_{n0}, \theta_0) \\
&= -\delta_n S_n(\theta_0)\delta_n^{-1}(\theta_{nm} - \theta_{n0}) + \frac{1}{2}\delta_n^2 K_n(\theta_0)\delta_n^{-2}(\theta_{nm} - \theta_0)^2 \\
&\quad -\frac{1}{2}\delta_n^2 K_n(\theta_0)\delta_n^{-2}(\theta_{n0} - \theta_0)^2 + o_p(1) \\
&= -\delta_n S_n(\theta_0)\delta_n^{-1}(\theta_{nm} - \theta_{n0}) + \frac{1}{2}\delta_n^2 K_n(\theta_0)\delta_n^{-2}(\theta_{nm} - \theta_{n0})^2 \\
&\quad +\delta_n^2 K_n(\theta_0)\delta_n^{-2}(\theta_{n0} - \theta_0)(\theta_{nm} - \theta_{n0}) + o_p(1) \\
&= \frac{1}{2}\delta_n^2 K_n(\theta_0)\delta_n^{-2}(\theta_{nm} - \theta_{n0})^2 + o_p(1).
\end{aligned}$$

Given $\epsilon > 0$, there exists a constant $\kappa > 1$ such that $\delta_n^2 K_n(\theta_0) > 1/\kappa$, implying that $\delta_n^{-2}(\theta_{nm} - \theta_{n0})^2 \leq \kappa(\gamma_n + o_p(1))$, with probability at least $1 - \epsilon$. Then, it follows that $\delta_n^{-1}(\theta_{nm} - \theta_{n0}) = o_p(1)$.

(3) Let $\epsilon > 0$, the neighborhood Θ' of θ_0, and the scalars $M > 0$ and $\psi > 0$ be given as in assumption (S). Then, for $\theta, \theta' \in \Theta'$, one has with probability at least $1 - 2\epsilon$,

$$\Lambda_n(\theta, \theta') = S_n(\theta')(\theta - \theta') - \frac{1}{2}K_n(\theta')(\theta - \theta')^2 + \alpha_1\delta_n^{-2}M|\theta - \theta'|^{2+\psi},$$

and

$$\Lambda_n(\theta', \theta) = S_n(\theta)(\theta' - \theta) - \frac{1}{2}K_n(\theta)(\theta - \theta')^2 + \alpha_2\delta_n^{-2}M|\theta - \theta'|^{2+\psi},$$

for some $\alpha_1, \alpha_2 \in [-1,1]$. Adding these expressions for $\theta \neq \theta'$ and using the Hölder condition that $K_n(\theta) = K_n(\theta') + \alpha_3 \delta_n^{-2} M|\theta - \theta'|^\psi$ for some $\alpha_3 \in [-1,1]$ yields

$$\begin{aligned}(\star) \quad S_n(\theta) &= S_n(\theta') - K_n(\theta')(\theta - \theta') + \alpha_4 \delta_n^{-2} M|\theta - \theta'|^{1+\psi} \\ &= K_n(\theta')(\theta' + K_n(\theta')^{-1} S_n(\theta') - \theta) + \alpha_4 \delta_n^{-2} M|\theta - \theta'|^{1+\psi}\end{aligned}$$

for some $\alpha_4 \in [-5/2, 5/2]$. Taking $\theta = \theta_{n0}$ and $\theta' = \theta_0$ in $(\star)$ yields

$$\delta_n S_n(\theta_{n0}) = \alpha_4 \delta_n^\psi M|\delta_n^{-1}(\theta_{n0} - \theta_0)|^{1+\psi} = o_p(1),$$

since $\delta_n^{-1}(\theta_{n0} - \theta_0)$ is stochastically bounded by result (1) and $\delta_n^\psi \to 0$. Next, taking $\theta = \theta_{n0}$ and $\theta' = \theta_{n1}$ in $(\star)$ yields

$$\begin{aligned}\delta_n S_n(\theta_{n0}) &= \delta_n^2 K_n(\theta_{n1}) \delta_n^{-1}(\theta_{n1} + K_n(\theta_{n1})^{-1} S_n(\theta_{n1}) - \theta_{n0}) + \alpha_4 \delta_n^{-1} M|\theta_{n0} - \theta_{n1}|^{1+\psi} \\ &= \delta_n^2 K_n(\theta_{n1}) \delta_n^{-1}(\theta_{n2} - \theta_{n0}) + \alpha_4 \delta_n^{-1}(\delta_n')^{1+\psi} M|(\delta_n')^{-1}(\theta_{n0} - \theta_{n1})|^{1+\psi}.\end{aligned}$$

But $\delta_n S_n(\theta_{n0}) = o_p(1)$, $\delta_n^{-1}(\delta_n')^{1+\psi} \to 0$ by assumption, and

$$|(\delta_n')^{-1}(\theta_{n0} - \theta_{n1})| \leq |(\delta_n')^{-1}(\theta_{n0} - \theta_0)| + |(\delta_n')^{-1}(\theta_{n1} - \theta_0)| = O_p(1),$$

since $(\delta_n')^{-1}(\theta_{n1} - \theta_0)$ is stochastically bounded by assumption, and $|(\delta_n')^{-1}(\theta_{n0} - \theta_0)| \leq |\delta_n^{-1}(\theta_{n0} - \theta_0)|$, which is stochastically bounded by result (1). Then, with probability at least $1-2\epsilon$, all terms in the expression above other than $\delta_n^2 K_n(\theta_{n1}) \delta_n^{-1}(\theta_{n2} - \theta_{n0})$ are $o_p(1)$. Further, $(\delta_n')^{-1}(\theta_{n1} - \theta_0) = O_p(1)$, and the Hölder condition on $\delta_n^2 K_n(\theta)$ implies that with probability at least $1 - \epsilon$, $\delta_n^2 K_n(\theta) \geq 1/2\kappa > 0$. Hence, with probability at least $1 - 3\epsilon$, $\delta_n^{-1}(\theta_{n2} - \theta_{n0}) = o_p(1)$. Since ϵ can be made as small as one pleases, this proves that θ_{n2} and θ_{n0} are asymptotically equivalent. □

B Miscellaneous Minor Results

B.1 Details on the Triangular and Quadratic Densities

(i) Triangular Density

Let $y_{(1)}$ and $y_{(n)}$ denote the extreme value statistics from an i.i.d. sample of size n.

Then, for $c > 0$,

$$\begin{aligned}
\Pr\left(y_{(1)} > \alpha + cn^{-\frac{1}{2}}\right) &= \left(1 - F(\alpha + cn^{-\frac{1}{2}}; \alpha, \beta)\right)^n = \left(1 - \frac{c^2}{n(\beta-\alpha)^2}\right)^n \\
&\to \exp\left(-\frac{c^2}{(\beta-\alpha)^2}\right), \\
\Pr\left(y_{(n)} < b - cn^{-1}\right) &= F\left(b - cn^{-1}; \alpha, \beta\right)^n = \left(1 - \frac{c}{n(\beta-\alpha)}\right)^{2n} \\
&\to \exp\left(-\frac{2c}{\beta-\alpha}\right),
\end{aligned}$$

so the parameters α and β can be estimated by $y_{(1)}$ and $y_{(n)}$, respectively, at the respective rates $n^{-\frac{1}{2}}$ and n^{-1}.

The Hellinger distance between $f(y; \alpha, \beta)$ and $f(y; \alpha', \beta')$ with $\alpha' \leq \alpha$ and $\beta' \geq \beta$ is

$$\begin{aligned}
h^2(\alpha, \beta, \alpha', \beta') &= 1 - \int_\alpha^\beta f(y; \alpha, \beta)^{\frac{1}{2}} f(y; \alpha', \beta')^{\frac{1}{2}} dy \\
&= 1 - \frac{2}{(\beta-\alpha)(\beta'-\alpha')} \int_\alpha^\beta (y-\alpha)^{\frac{1}{2}} (y-\alpha')^{\frac{1}{2}} dy.
\end{aligned}$$

When $\alpha = \alpha'$, this simplifies to $h^2(\beta, \beta') = (\beta' - \beta)/(\beta' - \alpha)$. The sample Hellinger distance between β and $\beta' = \beta + cn^{-1}$ the satisfies

$$\begin{aligned}
H^2(\beta, \beta + cn^{-1}) &= 1 - \left(\frac{\beta' - \alpha - cn^{-1}}{\beta' - \alpha}\right)^n \\
&\to 1 - \exp(c/(\beta' - \alpha)) \in (0, 1).
\end{aligned}$$

This establishes that the estimator $y_{(n)}$ for β is rate optimal at rate n^{-1}.

When $\beta = \beta'$, let $\Delta = \alpha - \alpha' > 0$ and $\alpha'' = (\alpha + \alpha')/2$. Then, $\alpha = \alpha'' + \Delta/2$, and $\alpha' = \alpha'' - \Delta/2$. Using the inequalities $1 - z \leq (1-z)^{\frac{1}{2}} \leq 1 - z/2$ for $0 \leq z \leq 1$,

$$\begin{aligned}
1 - h^2(\alpha, \alpha') &= \frac{2}{(\beta-\alpha)(\beta-\alpha')} \int_\alpha^\beta (y-\alpha)^{\frac{1}{2}} (y-\alpha')^{\frac{1}{2}} dy \\
&= \frac{2}{(\beta-\alpha'')^2 - \Delta^2/4} \int_\alpha^\beta ((y-\alpha'')^2 - \Delta^2/4)^{\frac{1}{2}} dy \\
&= \frac{2}{(\beta-\alpha'')^2 - \Delta^2/4} \int_\alpha^\beta (y-\alpha'')(1 - \gamma(y)\Delta^2/4(y-\alpha'')^2) dy,
\end{aligned}$$

for some $\gamma(y) \in (\frac{1}{2}, 1)$. But $2\int_\alpha^\beta (y - \alpha'')dy = (\beta - \alpha'')^2 - \Delta^2/4$. Hence, h^2 is bracketed by

$$\frac{2}{(\beta-\alpha'')^2 - \Delta^2/4} \int_\alpha^\beta (\Delta^2/8(y-\alpha''))dy = \frac{\Delta^2}{4(\beta-\alpha'')^2 - \Delta^2} (\log(\beta - \alpha'') - \log(\Delta/2))$$

and

$$\frac{2}{(\beta-\alpha'')^2-\Delta^2/4}\int_\alpha^\beta(\Delta^2/4(y-\alpha''))dy=\frac{2\Delta^2}{4(\beta-\alpha'')^2-\Delta^2}(\log(\beta-\alpha'')-\log(\Delta/2)).$$

Then, the bimodulus function $\lambda(\alpha'+\Delta,\alpha')=\Delta^2|\log(\Delta)|/4(\beta-\alpha'')^2$ satisfies

$$\frac{1}{4}\lambda(\alpha'+\Delta,\alpha')\leq h^2(\alpha'+\Delta,\alpha')\leq 4\lambda(\alpha'+\Delta,\alpha')$$

for Δ sufficiently small. The rate $\delta_n=4(\beta-\alpha')(2n\log(n))^{-\frac{1}{2}}$ satisfies

$$n\lambda(\alpha'+\delta_n,\alpha')=1+(\log\log n)/(\log n)+O(1/(\log n))\to 1,$$

and $(n\log(n))^{-\frac{1}{2}}$ is therefore a Hellinger rate.

The limiting distribution of the estimator $\alpha_{(1)}=y_{(1)}$ of α induces an exponential asymptotic distribution of the statistic $T_{(1)}=n^{\frac{1}{2}}(\alpha_{(1)}-\alpha)$ which has a density given by $2t\exp(-t^2/(\beta-\alpha)^2)/(\beta-\alpha)^2$, with moments $\mathbb{E}[T_{(1)}^k]=(\beta-\alpha)^k\Gamma(k/2+1)$. It does not attain the best rate. In comparison, the maximum likelihood estimator (MLE) for α is the solution $\hat{\alpha}_{MLE}$ of $\sum_{i=1}^n\left(\frac{1}{y_i-\hat{\alpha}_{MLE}}-\frac{2}{\beta-\hat{\alpha}_{MLE}}\right)=0$. The MLE does not satisfy conventional regularity conditions. A Taylor's expansion of the log-likelihood ratio $\Lambda_n(\alpha+\delta,\alpha)$ gives

$$\Lambda_n(\alpha+\delta,\alpha)=S_n(\alpha)\delta-\frac{1}{2}K_n(\alpha)\delta^2+o(\delta^2),$$

where $S_n(\alpha)=\sum_{i=1}^n[2/(\beta-\alpha)-1/(y_i-\alpha)]$ and $K_n(\alpha)=\sum_{i=1}^n[1/(y_i-\alpha)^2-2/(\beta-\alpha)^2]$. One has $\mathbb{E}[S_n(\alpha)]=0$, but the expectations of $S_n(\alpha)^2$ and $K_n(\alpha)$ do not exist. Therefore, $n^{-\frac{1}{2}}S_n(\alpha)$ is not stochastically bounded and the triangular density does not belong to the LAN family. This suggests that the MLE attains the best rate $\delta_n=(n\log(n))^{-\frac{1}{2}}$. To see this, let $M_n>0$ and note

(1)

$$\begin{aligned}P(y_i-\alpha>M_n^{-1}\text{ for }i=1,\cdots,n)&=P(y_{(1)}-\alpha>M_n^{-1})\\&=[1-(nM_n^{-2})/n(\beta-\alpha)^2]^n\\&\to\exp(-\lim_n nM_n^{-2}/(\beta-\alpha)^2).\end{aligned}$$

Then, the probability of the event $y_{(1)}-\alpha>M_n^{-1}$ goes to one if $nM_n^{-2}\to 0$ as $n\to\infty$.

(2) Let $Z=1_{\{y-\alpha>M_n^{-1}\}}/(y-\alpha)$. Then,

$$\begin{aligned}\mathbb{E}[Z]&=2(\beta-\alpha-M_n^{-1})/(\beta-\alpha)^2,\\\mathbb{E}[Z^2]&=2[\log(\beta-\alpha)+\log(M_n)]/(\beta-\alpha)^2.\end{aligned}$$

Then, Chebyshev's inequality implies

$$P\left(\left|\delta_n \sum_{i=1}^{n}(Z_i - \mathbb{E}[Z_i])\right| > \epsilon\right) < n\delta_n^2 \mathbb{E}[Z^2]/\epsilon^2$$
$$= 2[\log(\beta-\alpha) + \log(M_n)]/\epsilon^2(\beta-\alpha)^2 \log(n).$$

This is uniformly small for ϵ large if $\log(M_n)/\log(n)$ is bounded.

(3) $\mathbb{E}[1/(y-\alpha)] - \mathbb{E}[Z] = 2M_n^{-1}/(\beta-\alpha)^2$. Then, $\delta_n \sum_{i=1}^{n}(\mathbb{E}[1/(y-\alpha)] - \mathbb{E}[Z]) = 2n\delta_n M_n^{-1}/(\beta-\alpha)^2$. This remains bounded if $(n/\log(n))^{\frac{1}{2}} M_n^{-1}$ remains bounded.

Together, (1)-(3) imply $\delta_n S_n(\alpha)$ stochastically bounded. Collecting the requirements, $nM_n^{-2} \to 0$, $\log(M_n)/\log(n)$ bounded, and $(n/\log(n))^{\frac{1}{2}} M_n^{-1}$ bounded. All are satisfied if $M_n = n^{\frac{1}{2}}(\log(n))^{\gamma}$, for any $\gamma > 0$.

The next steps obtain the properties of $K_n(\alpha)$. The M_n need not be the same as above, but the first condition $nM_n^{-2} \to 0$ must still hold:

(4) Let $W = 1_{\{y-\alpha > M_n^{-1}\}}/(y-\alpha)^2$. Then,

$$\mathbb{E}[W] = \mathbb{E}[Z^2] = 2[\log(\beta-\alpha) + \log(M_n)]/(\beta-\alpha)^2,$$

and

$$\mathbb{E}[W^2] = 2[1/(\beta-\alpha)^2 + M_n^2]/(\beta-\alpha)^2.$$

Again by Chebyshev's inequality,

$$P\left(\left|\delta_n^2 \sum_{i=1}^{n}(W_i - \mathbb{E}[W_i])\right| > \epsilon\right) < n\delta_n^4 \mathbb{E}[W^2]/\epsilon^2,$$

and this is uniformly small for ϵ large if $M_n^2/n(\log(n))^2$ remains bounded.

(5) If $n\delta_n^2 \log(M_n)$ is bounded, then $\delta_n^2 \sum_{i=1}^{n} \mathbb{E}[W]$ is bounded, and $\delta_n^2 \sum_{i=1}^{n} W_i$ converges in probability to $\lim_n n\delta_n^2 \log(M_n)$. This is sufficient to establish that $\delta_n^2 K_n(\alpha)$ converges in probability to this limit. To assure that the limit points of $\delta_n^2 K_n(\alpha)$ are positive and finite, one then needs $n\delta_n^2 \log(M_n)$ to have a positive finite limit.

Collecting requirements, $nM_n^{-2} \to 0$, $M_n^2/n(\log(n))^2$ bounded, and $n\delta_n^2 \log(M_n)$ with a positive finite limit. All are satisfied if $M_n = n^{\frac{1}{2}}(\log(n))^{\gamma}$ if $\gamma \leq 1$. Then, it suffices for the proof to take $M_n = (n\log(n))^{\frac{1}{2}}$ throughout.

From $\frac{\partial W}{\partial \alpha} = 2 \cdot 1_{\{y-\alpha > M_n^{-1}\}}/(y-\alpha)^3$, one has

$$\mathbb{E}\left[\frac{\partial W}{\partial \alpha}\right] = 4[1/(\beta-\alpha) + M_n]/(\beta-\alpha)^2.$$

Then, $n^{-\frac{1}{2}}\delta_n^2 n\mathbb{E}[\partial W/\partial\alpha] = (\log(n))^{\gamma-1}$ when $M_n = n^{\frac{1}{2}}(\log(n))^{\gamma}$. It follows that, starting from a consistent estimator of α that converges at a $n^{-\frac{1}{2}}$ rate, e.g. $y_{(1)}$, the argument of Proposition 3 applies to establish that the one-step estimator is asymptotically equivalent to the infeasible one-step estimator and attains the best rate $(n\log(n))^{-\frac{1}{2}}$.

(ii) Quadratic Density

For the quadratic density $f(y;\theta) = 3y(2\theta - y)/4\theta^3$, $0 \leq y \leq 2\theta$, the affinity between $f(y;1)$ and $f(y;1+\delta)$, for $\delta > 0$, is

$$\begin{aligned}
\rho(1+\delta,1) &= \int_0^2 (3y/4(1+\delta)^{\frac{3}{2}})((2+2\delta-y)(2-y))^{\frac{1}{2}}dy \\
&= \int_0^2 (3y/4(1+\delta)^{\frac{3}{2}})((2+\delta-y+\delta)(2+\delta-y-\delta))^{\frac{1}{2}}dy \\
&= \int_0^2 (3y/4(1+\delta)^{\frac{3}{2}})(2+\delta-y)(1-\delta^2/(2+\delta-y)^2)^{\frac{1}{2}}dy \\
&= \int_0^2 (3y/4(1+\delta)^{\frac{3}{2}})(2+\delta-y)dy - \gamma\delta^2\int_0^2 (3y/4(1+\delta)^{\frac{3}{2}})/(2+\delta-y)dy \\
&\quad \text{for some } \gamma \in [1/2,1], \\
&= (3/4(1+\delta)^{\frac{3}{2}})((2+\delta)2 - 8/3) \\
&\quad -\gamma\delta^2\int_0^2 (3/4(1+\delta)^{\frac{3}{2}})(-1+(2+\delta)/(2+\delta-y))dy \\
&= (1+3\delta/2)/(1+\delta)^{\frac{3}{2}} \\
&\quad -\gamma\delta^2(3/4(1+\delta)^{\frac{3}{2}})(-2-(2+\delta)\log(\delta)+(2+\delta)\log(2+\delta)) \\
&= (1-3\delta/2+\gamma'\delta^2)((1+3\delta/2)+\gamma\delta^2((3/2)\log(\delta)+3/2+O(\delta))) \\
&\quad \text{for some } \gamma' \in [5/16,15/4], \\
&= 1-9\delta^2/4+\gamma'\delta^2-3\gamma\delta^2/2+3\gamma\delta^2\log(\delta)/2+O(\delta^3) \\
&= 1+O(\delta^2\log(\delta))+O(\delta^2),
\end{aligned}$$

and the Hellinger distance $h^2(1+\delta,1) = O(\delta^2\log(\delta))$. The convergence rate $\delta_n = (n\log(n)^{-\frac{1}{2}}$ satisfies

$$\begin{aligned}
nh^2(1+\delta_n,1) &= O\left((\log(n))^{-1}\left(-\frac{1}{2}\log(n)-\frac{1}{2}\log\log(n)\right)\right) \\
&= O\left(\frac{1}{2}+(\log\log(n))/\log(n)\right) \\
&\to \text{const.},
\end{aligned}$$

so δ_n is the Hellinger rate.

The log likelihood is

$$L = n\log(3/4) - 3n\log(\theta) + \sum_{i=1}^{n}(\log(y_i)+\log(2\theta-y_i)).$$

The first-order condition for a MLE θ_n is

$$0 = \sum_{i=1}^{n}(2/(2\theta_n - y_i) - 3/\theta_n) = \sum_{i=1}^{n}(3y_i - 4\theta_n)/(2\theta_n - y_i).$$

Then, the MLE solves iteratively,

$$\theta_n = (3/4)\sum_{i=1}^{n} y_{(i)}w_{ni} / \sum_{i=1}^{n} w_{ni},$$

where $w_{ni} = 1/(2\theta_n - y_{(i)})$. The log likelihood is continuous and bounded above for $\theta > y_{(n)}/2$, and approaches $-\infty$ as $\theta \to y_{(n)}/2$. Then, the MLE exists.

The following steps, mimicking those for the triangular density, establish that the quadratic family is LAQ.

Define $S_n(\theta) = \frac{\partial}{\partial\theta}\Lambda_n(\theta,\theta_0) = \sum_{i=1}^{n}(2/(2\theta - y_i) - 3/\theta)$ and $K_n(\theta) = -\frac{\partial^2}{\partial\theta^2}\Lambda_n(\theta,\theta_0) = \sum_{i=1}^{n}[4/(2\theta - y_i)^2 - 3/\theta^2)$. Furthermore, define $M_n = (n\log(n))^{\frac{1}{2}}$, as well as $S_n^\star(\theta) = \sum_{i=1}^{n}(2/(\max\{2\theta - y_i, M_n^{-1}\}) - 3/\theta)$ and $K_n^\star(\theta) = \sum_{i=1}^{n}(4/(\max\{2\theta - y_i, M_n^{-1}\})^2 - 3/\theta^2)$. In parallel with the triangular density case, $P(2\theta_0 - y_{(n)} > M_n^{-1}) \to \exp(-c\lim_n nM_n^{-2}) = 0$. Then, $S_n(\theta)$ and $K_n(\theta)$ have the needed LAQ properties if $S_n^\star(\theta)$ and $K_n^\star(\theta)$ do. But in a neighborhood of 2θ, the density $f(y;\theta)$ behaves like the term $3(2\theta - y)/2\theta^2$, which except for scale is the triangular density. Then, the calculations for that density establish that $\delta_n S_n^\star(\theta)$ is stochastically bounded and $\delta_n^2 K_n^\star(\theta)$ converges in probability to a positive definite limit.

B.2 Calculations for Example 2

Consider the squared Hellinger distance between $f(y;0,\alpha,\beta,\gamma)$ and $f(y;\theta,\alpha,\beta,\gamma)$ for small, positive θ, defined as

$$h^2(\theta,0) = \frac{1}{2}\int_{-\infty}^{+\infty}\eta(y;\theta)^2 dy,$$

where

$$\begin{aligned}
\eta(y;\theta) &= f^{\frac{1}{2}}(y;\theta,\alpha,\beta,\gamma) - f^{\frac{1}{2}}(y;0,\alpha,\beta,\gamma) \\
&= C'\left(|y-\theta|^{\frac{\alpha-1}{2}}\exp(-|y-\theta|^\beta/2\gamma) - |y|^{\frac{\alpha-1}{2}}\exp(-|y|^\beta/2\gamma)\right) \\
&= R(y;\theta) + S(y;\theta),
\end{aligned}$$

and

$$\begin{aligned}
R(y;\theta) &= C'(|y-\theta|^{\frac{\alpha-1}{2}} - |y|^{\frac{\alpha-1}{2}})\exp(-|y|^\beta/2\gamma), \\
S(y;\theta) &= C'|y-\theta|^{\frac{\alpha-1}{2}}\left(\exp(-|y-\theta|^\beta/2\gamma) - \exp(-|y|^\beta/2\gamma)\right).
\end{aligned}$$

Note that $R(y;\theta)$ drops out if $\alpha = 1$ (case (2)). Notice also that, in this case,

$$\begin{aligned} h^2(\theta,0) &\propto \theta^2 \int_{-\infty}^{+\infty} \left[\frac{\partial}{\partial\theta}\log f(y;\theta,\alpha,\beta,\gamma)\right]_{\theta=0}^2 f(y;0,\alpha,\beta,\gamma)dy \\ &= \theta^2 \int_{-\infty}^{+\infty} \beta^2|y|^{2\beta-2}\exp(-|y|^\beta/\gamma)dy \\ &\propto \theta^2 \int_{-\infty}^{+\infty} |z|^{1-1/\beta}\exp(-|z|)dz, \end{aligned}$$

which is proportional to a gamma function and converges if $\beta > \frac{1}{2}$. Therefore, in case (2) the Hellinger rate is $n^{-\frac{1}{2}}$ when $\beta > \frac{1}{2}$. Hereafter, we therefore concentrate on cases for which $\beta \leq \frac{1}{2}$.

Since $\eta(y;\theta)^2$ is symmetric about $\theta/2$, the decomposition $h^2(\theta,0) = 2(A+B+C+D)$ holds, with

$$\begin{aligned} A &= \int_{-\infty}^{-1} \eta(y;\theta)^2 dy, \\ B &= \int_{-1}^{-\theta/2} \eta(y;\theta)^2 dy, \\ C &= \int_{-\theta/2}^{0} \eta(y;\theta)^2 dy, \\ D &= \int_{0}^{\theta/2} \eta(y;\theta)^2 dy. \end{aligned}$$

Note that $C \geq D$. It is straightforward to show that $A = O(\theta^2)$ in all cases. We will derive lower and upper bounds on B as well as a lower bound on D and an upper bound on C. In doing so, we will show that in case (1) the contribution due to $R(y;\theta)$ dominates, in the sense of exhibiting the fastest convergence to zero when θ approaches zero, while in case (2) the contribution of $S(y;\theta)$ dominates. Note also that, since $R(y;\theta)$ and $S(y;\theta)$ are both negative on $(-\infty, \theta/2]$ when $\alpha < 1$, the cross terms that emerge when completing the square can be ignored in the derivation of the lower bounds in this case.

We will employ the inequalities[14]

$$\begin{aligned} (c-1)[a^2/c - b^2] &\leq (a-b)^2 \leq 2(a^2+b^2) \text{ for } c > 1, \\ \theta\kappa|y-\theta|^{-\kappa-1} &\leq |y|^{-\kappa} - |y-\theta|^{-\kappa} \leq \theta\kappa|y|^{-\kappa-1} \text{ for } y < 0 < \theta \text{ and } \kappa > 0. \end{aligned}$$

[14]The first inequality comes from $0 \leq (c^{-\frac{1}{2}}a - c^{\frac{1}{2}}b)^2 = (a-b)^2 - [(1-1/c)a^2 - (c-1)b^2]$ and from $2(a^2+b^2) - (a-b)^2 = (a+b)^2 \geq 0$. The remaining inequality comes from the theorem of the mean for convex functions.

In case (1), the contribution of $R(y;\theta)$ to the lower bound on D is

$$\begin{aligned}
\int_0^{\theta/2} R(y;\theta)^2 dy &\geq (C')^2 \exp\left(-|\theta/2|^\beta/\gamma\right) \int_0^{\theta/2} \left[|y-\theta|^{\frac{\alpha-1}{2}} - |y|^{\frac{\alpha-1}{2}}\right]^2 dy \\
&\geq (C')^2 \exp\left(-|\theta/2|^\beta/\gamma\right)(c-1) \int_0^{\theta/2} \left[|y|^{\alpha-1}/c - |y-\theta|^{\alpha-1}\right] dy \\
&= (C')^2 \exp\left(-|\theta/2|^\beta/\gamma\right)(c-1)\theta^\alpha \frac{1}{\alpha}\left[2^{-\alpha}/c + 2^{-\alpha} - 1\right].
\end{aligned}$$

For $\alpha < 1$ and $c < 1/(2^\alpha - 1)$, the term in square brackets is positive, so that $\int_0^{\theta/2} R(y;\theta)^2 dy = O(\theta^\alpha)$. The contribution of $S(y;\theta)$ to the lower bound on D is

$$\begin{aligned}
\int_0^{\theta/2} S(y;\theta)^2 dy &= \int_0^{\theta/2} (C')^2 |y-\theta|^{\alpha-1} \left(\exp(-|y-\theta|^\beta/2\gamma) - \exp(-|y|^\beta/2\gamma)\right)^2 dy \\
&\propto \int_0^{\theta/2} \theta^2 |y-\theta|^{\alpha-1} (\beta/2\gamma)^2 |y-\theta|^{2\beta-2} \exp(-|y|^\beta/\gamma) dy \\
&\geq (\beta/2\gamma)^2 \theta^2 \exp\left(-|\theta/2|^\beta/\gamma\right) \int_0^{\theta/2} |y-\theta|^{\alpha+2\beta-3} dy \\
&\propto (\beta/2\gamma)^2 \theta^2 \exp\left(-|\theta/2|^\beta/\gamma\right) \theta^{\alpha+2\beta-2}/|\alpha+2\beta-2| \\
&= (\beta/2\gamma)^2 \exp\left(-|\theta/2|^\beta/\gamma\right) \theta^{\alpha+2\beta}/|\alpha+2\beta-2|.
\end{aligned}$$

Note that in case (2), this term is $O(\theta^{2\beta+1})$ if $\beta < 1/2$, and approaches a term of order $O(\theta^2 \log(\theta))$ when β tends to $1/2$. In case (1), this term converges to zero at a slower rate than the contribution due to $R(y;\theta)$.

Turning to term C, $C \leq 2\int_{-\theta/2}^0 R(y;\theta)^2 dy + 2\int_{-\theta/2}^0 S(y;\theta)^2 dy$. The contribution of $R(y;\theta)$ to the upper bound is

$$\begin{aligned}
\int_{-\theta/2}^0 R(y;\theta)^2 dy &\leq (C')^2 \int_{-\theta/2}^2 \left[|y-\theta|^{\frac{\alpha-1}{2}} - |y|^{\frac{\alpha-1}{2}}\right]^2 dy \\
&\leq 2(C')^2 \int_{-\theta/2}^0 \left[|y-\theta|^{\alpha-1} + |y|^{\alpha-1}\right] dy \\
&\propto \theta^\alpha.
\end{aligned}$$

The contribution of $S(y;\theta)$ to the upper bound on C is

$$\begin{aligned}
\int_{-\theta/2}^0 S(y;\theta)^2 dy &= \int_{-\theta/2}^0 (C')^2 |y-\theta|^{\alpha-1} \left(\exp(-|y-\theta|^\beta/2\gamma) - \exp(-|y|^\beta/2\gamma)\right)^2 dy \\
&\propto \theta^2 \int_{-\theta/2}^0 |y-\theta|^{\alpha+2\beta-3} \exp(-|y|^\beta/\gamma) dy \\
&\leq \theta^2 \int_{-\theta/2}^0 |y-\theta|^{\alpha+2\beta-3} dy \\
&\propto \theta^{\alpha+2\beta}/|\alpha+2\beta-2|.
\end{aligned}$$

Again, this term is $O(\theta^{2\beta+1})$ if $\beta < 1/2$, and approaches a term of order $O(\theta^2 \log(\theta))$ when β tends to $1/2$. Also, in case (1) this term converges to zero at a slower rate than the contribution due to $R(y;\theta)$.

Finally, with regard to term B, $B \leq 2\int_{-1}^{-\theta/2} R(y;\theta)^2 dy + 2\int_{-1}^{-\theta/2} S(y;\theta)^2 dy$. Towards an upper bound, the contribution of $R(y;\theta)$ is

$$\begin{aligned}\int_{-1}^{\theta/2} R(y;\theta)^2 dy &\leq (C'(1-\alpha)/2)^2\theta^2 \int_{-1}^{-\theta/2} |y|^{\alpha-3} dy \\ &= (C'(1-\alpha)/2)^2\theta^2[(\theta/2)^{\alpha-2} - 1]/(2-\alpha) \\ &= O(\theta^\alpha).\end{aligned}$$

The contribution of $S(y;\theta)$ is

$$\begin{aligned}\int_{-1}^{-\theta/2} S(y;\theta)^2 dy &= \int_{-1}^{-\theta/2} (C')^2 |y-\theta|^{\alpha-1} \left(\exp(-|y-\theta|^\beta/2\gamma) - \exp(-|y|^\beta/2\gamma)\right)^2 dy \\ &\propto \theta^2 \int_{-1}^{-\theta/2} |y-\theta|^{\alpha+2\beta-3} \exp(-|y|^\beta/\gamma) dy \\ &\leq \theta^2 \exp\left(-|\theta/2|^\beta/\gamma\right) \int_{-1}^{-\theta/2} |y-\theta|^{\alpha+2\beta-3} dy \\ &\propto \theta^{\alpha+2\beta}/|\alpha+2\beta-2|,\end{aligned}$$

Lower bounds can be derived in an analogous fashion, as above,

$$\begin{aligned}\int_{-1}^{-\theta/2} R(y;\theta)^2 dy &\geq (C')^2 \exp(-1/\gamma)(c-1)\theta^\alpha \left[2^{-\alpha}/c + 2^{-\alpha} - 1\right] = O(\theta^\alpha), \\ \int_{-1}^{-\theta/2} S(y;\theta)^2 dy &\geq \theta^2 \exp\left(-1/\gamma\right) \int_{-1} -\theta/2 |y-\theta|^{\alpha+2\beta-3} dy \propto \theta^{\alpha+2\beta}/|\alpha+2\beta-2|.\end{aligned}$$

The same comments apply as above: The contributions due to $S(y;\theta)$ are $O(\theta^{2\beta+1})$ if $\beta < 1/2$, and approach terms of order $O(\theta^2 \log(\theta))$ when β tends to $1/2$. Also, in case (1) these contributions converge to zero at a slower rate than the contributions due to $R(y;\theta)$.

The bimodulus rates follow immediately from the respective bimodulus functions and, by Lemma 3, are equivalent to the Hellinger rates reported in Table 1. For the case $\alpha = 1, \beta = \frac{1}{2}$, the bimodulus function $\lambda(\theta, 0)$ is proportional to $\theta^2|\log(\theta)|$, so that the rate $\delta_n \sim (n\log(n))^{-\frac{1}{2}}$ satisfies $n\lambda(\delta_n, 0) = 1 + (\log\log(n))/(\log(n)) + O(1/(\log(n))) \to 1$, and $(n\log(n))^{-\frac{1}{2}}$ is therefore the Hellinger rate in this case.

B.3 An Illustrative Example of Maximal Uniform Convergence Rates

Consider the location parameter example $y_i \sim$ i.i.d. $u[\theta - \frac{1}{2}, \theta + \frac{1}{2}]$. As in Corollary 1(i), $\bar{\rho}(\tau) \equiv \rho(\theta, \theta + |\tau|) = \rho(\theta', \theta' + |\tau|)$ for any $\theta, \theta' \in \Theta$ and $\tau \in \mathbb{R}$. It is easy to show that $\bar{\rho}(|\tau|) = 1 - |\tau|$, and so $H_n^2(\theta, \theta + |\tau|) = 1 - (1 - |\tau|)^n$ for all θ. Then, $\delta_n = \frac{1}{n}$.

Consider two estimators for θ in this example: (i) $\hat{\theta}_n = \frac{1}{2}(y_{(1)} + y_{(n)})$, where $y_{(1)}$ ($y_{(n)}$) denotes the minimum (maximum) of the sample $\{y_i, i = 1, \ldots, n\}$; and (ii) $\bar{\theta}_n = \bar{y}_n$. It is well-known[15] that

$$\begin{aligned} \text{var}(\hat{\theta}_n) &= \frac{1}{2(n+1)(n+2)}, \\ \text{var}(\bar{\theta}_n) &= \frac{1}{12n}, \end{aligned}$$

i.e. $\hat{\theta}_n$ converges at the Hellinger rate n, while $\bar{\theta}_n$ converges at the slower rate $\sqrt{n}$. W.l.o.g., let $\theta = 0$.

Let $\epsilon \in (0, 1)$. Then,

$$\Pr(H_n^2(0, |\tau|) > 1 - \epsilon) = \Pr(1 - (1 - |\tau|)^n > 1 - \epsilon) = \Pr(|\tau| > 1 - \epsilon^{\frac{1}{n}}).$$

Note that $1 - \epsilon^{\frac{1}{n}} = -\log(\epsilon)\epsilon^{\frac{\alpha}{n}}\frac{1}{n} \to 0$, as $n \to \infty$ for $\epsilon \in (0, 1)$ and some $\alpha \in (0, 1)$. Consider the estimator $\bar{\theta}_n$. Its asymptotic distribution is $\sqrt{n}\bar{\theta}_n \xrightarrow{d} N(0, 1/12)$. Hence,

$$\Pr(|\bar{\theta}_n| > 1 - \epsilon^{\frac{1}{n}}) \sim 2\Phi\left(\frac{\sqrt{n}(\epsilon^{\frac{1}{n}} - 1)}{\sqrt{12}}\right) \to 2\Phi(0) = 1 \text{ as } n \to \infty,$$

where $\Phi(\cdot)$ denotes the cumulative distribution function of a standard normally distributed random variable.

Next, consider $\hat{\theta}_n$. For any $\epsilon \in (0, 1)$, by Chebyshev's inequality

$$\begin{aligned} \Pr(|\hat{\theta}_n| > 1 - \epsilon^{\frac{1}{n}}) &\leq \left[2(n+1)(n+2)(1 - \epsilon^{\frac{1}{n}})^2\right]^{-1} \\ &= \frac{n^2}{2(n+1)(n+2)} \frac{1}{\epsilon^{\frac{2\alpha}{n}}(\log(\epsilon))^2} \text{ for some } \alpha \in (0, 1) \\ &\to \frac{1}{2(\log(\epsilon))^2} \text{ as } n \to \infty. \end{aligned}$$

This limit can be made as small as desired by letting ϵ approach zero.

[15] Cp., e.g., David (1970)

References

[1] Akahira, M. (1975): "Asymptotic Theory for Estimation of Location in Non-regular Cases, I: Order of Convergence of Consistent Estimators", *Stat. Appl. Res., JUSE*, **22(1)**, 8-26

[2] Akahira, M. (1991): "The amount of information and the bound for the order of consistency for a location parameter family of densitites", *Symposia Gaussiana*, Conf. B, Mammitzsch and Schneeweiss, eds.; Berlin: Gryuter & Co.

[3] Akahira, M. and K. Takeuchi (1991): "A definition of information amount applicable to non-regular cases", *J. Comput. Inform.*, **2**, 71-92

[4] Akahira, M. and K. Takeuchi (1995): *Non-Regular Statistical Estimation*, New York: Spinger Verlag

[5] Birgé, L. and P. Massart (1993): "Rates of Convergence for Minimum Contrast Estimators", *Probability Theory and Related Fields*, **97**, 113-150

[6] Chan, K.S. (1993): "Consistency and Limiting Distribution of the Least Squares Estimator of a Threshold Autoregression Model", *The Annals of Statistics*, **21**, 520-533

[7] Chan, K.S. and R.S. Tsay (1998): "Limiting properties of the least squares estimator of a continuous threshold autoregressive model", *Biometrika*, **85(2)**, 413-426

[8] Cosslett, S.R. (1987): "Efficiency Bounds for Distribution-Free Estimators of the Binary Choice and Censored Regression Models", *Econometrica*, **55(3)**, 559-585

[9] David, H.A. (1997): *Order Statistics*, 2nd edition, New York: Wiley

[10] Donoho, D.L. and R.C. Liu (1991a): "Geometrizing Rates of Convergence, III", *The Annals of Statistics*, **19(2)**, 668-701

[11] Donoho, D.L. and R.C. Liu (1991b): "Geometrizing Rates of Convergence, II", *The Annals of Statistics*, **19(2)**, 633-667

[12] Durrett, R. (1996): *Probability: theory and examples*, 2nd ed., Belmont, CA: Duxbury Press

[13] Hajek, J. (1970): "A characterization of the limiting distributions of regular estimates", *Z. Wahrsch. Verw. Gebiete*, **14**, 323-330

[14] Hall, P. (1989): "On Convergence Rates in Nonparametric Problems", *International Statistical Review / Revue Internationale de Statistique*, **57(1)**, 45-58

[15] Hansen, B. (2000): "Sample splitting and threshold estimation", *Econometrica*, **68**, 575-603

[16] Hirano, K. and J.R. Porter (2003): "Efficiency in Asymptotic Shift Experiments", mimeo, University of Miami and Harvard University

[17] Hoeffding, W. and J. Wolfowitz (1958): "Distinguishability of sets of distributions", *Ann. Math. Statist.*, **3**, 700-718

[18] Horowitz, J.L. (1993): "Optimal Rates of Convergence of Parameter Estimators in the Binary Response Model with Weak Distributional Assumptions", *Econometric Theory*, **9(1)**, 1-18

[19] Ibragimov, I.A. and R.Z. Has'minskii (1981): *Statistical Estimation*, New York: Springer

[20] Kakutani, S. (1948): "On Equivalence of Infinite Product Measures", *The Annals of Mathematics*, **49**, 214-224

[21] Klein, R.W. and R.H. Spady (1993): "An Efficient Semiparametric Estimator for Binary Response Models", *Econometrica*, **61(2)**, 387-421

[22] LeCam, L.M. (1970): "On the assumptions used to prove asymptotic normality of maximum likelihood estimators", *Annals of Mathematical Statistics*, **41**, 802-828

[23] LeCam, L.M. (1972): "Limits of Experiments", in: *Proceedings of the Sixth Berkeley Symposium of Mathematical Statistics*, **1**, 245-261

[24] LeCam, L.M. (1986): *Asymptotic Methods in Statistical Decision Theory*, New York: Springer

[25] LeCam, L.M. and G.L. Yang (2000): *Asymptotics in Statistics - Some Basic Concepts*, New York: Springer

[26] Matusita, K. (1955): "Decision rules based on the distance for problems of fit, two samples and estimation", *Annals of Mathematical Statistics*, **26**, 631-640

[27] McFadden, D.L., Beckert, W. and A. Eymann (2001): "Statistical Simulation", mimeo, U.C. Berkeley

[28] Newey, W.K. (1991): "Uniform Convergence in Probability and Stochastic Equicontinuity", *Econometrica*, **59(4)**, 1161-1167

[29] Paarsch, H.J. (1992): "A Comparison of Estimators for Empirical Models of Auctions", University of Western Ontario working paper No. 9210

[30] Pollard, D. (1984): *Convergence of Stochastic Processes*, New York: Springer

[31] Pollard, D. (1989): "Asymptotics via Empirical Processes", *Statistical Science*, **4(4)**, 341-354

[32] Pollard, D. (1997): "Another Look at Differentiability in Quadratic Mean", in: Pollard, D., Torgersen, E. and G. Yang (eds.), *Festschrift for Lucien LeCam*, New York: Springer

[33] Prakasa Rao, B.L.S. (1968): "Estimation of the location of the cusp of a continuous density", *The Annals of Mathematical Statistics*, **39**, 76-87

[34] Prakasa Rao, B.L.S. (2003): "Estimation of Cusp in Nonregular Nonlinear Regression Models", mimeo, Indian Statistical Institute

[35] Ruud, P.A. (2000): *Classical Econometric Theory*, Oxford: Oxford University Press

[36] Seo, M. and O. Linton (2005): "A Smoothed Least Squares Estimator For The Threshold Regression Model", mimeo, London School of Economics

[37] Stone, C.J. (1980): "Optimal Rates of Convergence for Nonparametric Estimators", *The Annals of Statistics*, **8(6)**, 1348-1360

[38] Stone, C.J. (1982): "Optimal Global Rates of Convergence for Nonparametric Regression", *The Annals of Statistics*, **10(4)**, 1040-1053

[39] Van de Geer, S. (1993): "Hellinger-Consistency of Certain Nonparametric Maximum Likelihood Estimators", *The Annals of Statistics*, **21(1)**, 14-44

[40] Van de Geer, S. (2000): *Empirical Processes in M-Estimation*, Cambridge: Cambridge University Press