



HAL
open science

Asymptotic normality of nonlinear least squares under singular experimental designs

Andrej Pazman, Luc Pronzato

► **To cite this version:**

Andrej Pazman, Luc Pronzato. Asymptotic normality of nonlinear least squares under singular experimental designs. Luc Pronzato, Anatoly Zhigljavsky. Optimal Design and Related Areas in Optimization and Statistics, Springer, pp.167-191, 2009, Springer Optimization and its Applications, 10.1007/978-0-387-79936-0_8. hal-00358761

HAL Id: hal-00358761

<https://hal.science/hal-00358761>

Submitted on 4 Feb 2009

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Asymptotic normality of nonlinear least squares under singular experimental designs

A. Pázman and L. Pronzato

Summary. We study the consistency and asymptotic normality of the LS estimator of a function $h(\theta)$ of the parameters θ in a nonlinear regression model with observations $y_i = \eta(x_i, \theta) + \varepsilon_i$, $i = 1, 2, \dots$ and independent errors ε_i . Optimum experimental design for the estimation of $h(\theta)$ frequently yields singular information matrices, which corresponds to the situation considered here. The difficulties caused by such singular designs are illustrated by a simple example: depending on the true value of the model parameters and on the type of convergence of the sequence of design points x_1, x_2, \dots to the limiting singular design measure ξ , the convergence of the estimator of $h(\theta)$ may be slower than $1/\sqrt{n}$, and, when convergence is at a rate of $1/\sqrt{n}$ and the estimator is asymptotically normal, its asymptotic variance may differ from that obtained for the limiting design ξ (which we call *irregular asymptotic normality* of the estimator). For that reason we focus our attention on two types of design sequences: those that converge strongly to a discrete measure and those that correspond to sampling randomly from ξ . We then give assumptions on the limiting expectation surface of the model and on the estimated function h which, for the designs considered, are sufficient to ensure the regular asymptotic normality of the LS estimator of $h(\theta)$.

1.1 Introduction

Although the singularity of regression models has been somewhat neglected in the literature on parameter estimation and experimental design, the difficulties it induces for statistical inference have been noticed and investigated in several domains of applications, see, e.g., Stoica (2001) for signal processing, Hero *et al.* (1996) for image processing (emission tomography) or Sjöberg *et al.* (1995) for general black-box modelling of dynamical systems. There, the singularity comes from an over-parametrization and the attention is directed to the Cramèr-Rao bound. The motivation of this paper is different and comes from optimum design theory. In nonsingular nonlinear models, a standard way to design an optimum experiment for parameter estimation is to optimize a criterion based on the asymptotic normality of the estimator. This is justified by the fact that the asymptotic variance of the estimator depends on the limiting design measure (it is the inverse of the associated information matrix), but not on the way this measure is approached by the sequence of

design points. Can we use a similar approach also for singular models? It is the purpose of this paper to present easily interpreted conditions which allow this. It is important since optimum designs may produce singular models. Indeed, when the interest is in a function $h(\theta)$ of the parameters θ of the model, or more simply in a subset of θ , the information matrix at the optimum design is often singular, see, e.g., Silvey (1980, p. 58) and Example 1 below for the case of c -optimality. We then say that *the design is singular*.

Singular designs cause no special difficulty in linear regression (which partly explains why singularity issues have been disregarded in the design literature), as briefly shown hereafter: a linear combination $\mathbf{c}^\top \theta$ of the parameters is either estimable or not, depending on the direction of \mathbf{c} .

1.1.1 Singular designs in linear models

The standard set-up for an optimum experimental design problem in a linear model is as follows. Given a set $\mathcal{X} \subset \mathbb{R}^k$, the design space, a design (an *exact design*, or a *design of size N*) is a choice x_1, \dots, x_N of points from \mathcal{X} . According to this design, we observe N random variables $y(x_1), \dots, y(x_N)$ modelled by

$$y(x_i) = \mathbf{f}^\top(x_i)\bar{\theta} + \varepsilon_i, \quad i = 1, \dots, N \quad (1.1)$$

where the errors ε_i are independent and $\mathbb{E}(\varepsilon_i) = 0$, $\text{Var}(\varepsilon_i) = \sigma^2$ for all i . Here, the vectors $\mathbf{f}(x_i) \in \mathbb{R}^p$ are known, $\sigma^2 \in \mathbb{R}^+$ is unknown and $\bar{\theta} \in \mathbb{R}^p$ is the unknown true value of the model parameters θ . We emphasize that throughout the paper the choice of design points x_1, \dots, x_N is independent of the observed variables $y(x_i)$ (that is, the design is *not sequential*). If the information matrix

$$\mathbf{M}(x_1, \dots, x_N) = \sum_{i=1}^N \mathbf{f}(x_i)\mathbf{f}^\top(x_i)$$

is nonsingular, then the least squares estimator (LSE) of θ ,

$$\hat{\theta}^N \in \arg \min_{\theta} \sum_{i=1}^N [y(x_i) - \mathbf{f}^\top(x_i)\theta]^2 \quad (1.2)$$

is unique and its variance is

$$\text{Var}(\hat{\theta}^N) = \sigma^2 \mathbf{M}^{-1}(x_1, \dots, x_N).$$

On the other hand, if $\mathbf{M}(x_1, \dots, x_N)$ is singular then $\hat{\theta}^N$ is not defined uniquely. However, $\mathbf{c}^\top \hat{\theta}^N$ does not depend on the choice of the solution $\hat{\theta}^N$ of (1.2) if and only if $\mathbf{c} \in \text{Range}[\mathbf{M}(x_1, \dots, x_N)]$. Then

$$\text{Var}(\mathbf{c}^\top \hat{\theta}^N) = \sigma^2 \mathbf{c}^\top \mathbf{M}^{-}(x_1, \dots, x_N) \mathbf{c} \quad (1.3)$$

where the choice of the g -inverse matrix \mathbf{M}^- is arbitrary. This last expression can be used as a criterion (the criterion of c -optimality) for an optimal choice of the N -point design x_1, \dots, x_N , and the design minimizing this criterion may be singular, see Silvey (1980); Pázman (1980) for some properties and Wu (1980, 1983) for a detailed investigation of the consistency of $\mathbf{c}^\top \hat{\theta}^N$ when $N \rightarrow \infty$.

1.1.2 Designs in nonlinear models

Consider now the same set-up with, however, one noticeable difference, namely that the observations are modelled by

$$y(x_i) = \eta(x_i, \bar{\theta}) + \varepsilon_i, \quad i = 1, \dots, N$$

where the function $\eta(\cdot, \cdot) : \mathcal{X} \times \Theta \rightarrow \mathbb{R}$ is nonlinear in the parameters $\theta \in \Theta \subset \mathbb{R}^p$ with unknown true value $\bar{\theta}$. One can often suppose, at least approximately, that the parameter space Θ is bounded and closed, hence compact. Also, it is standard in experimental design to take \mathcal{X} as a compact subset of \mathbb{R}^k . Besides these classical assumptions we shall also assume that Θ has no isolated points, i.e., that $\Theta \subset \overline{\text{int}(\Theta)}$ (the closure of the interior of Θ), and that $\Theta \subset \Theta_0$, with Θ_0 and open subset of \mathbb{R}^p such that for any $x \in \mathcal{X}$ the function $\eta(\theta, x)$ is defined and two times continuously differentiable on Θ_0 .

The LSE of θ , defined by

$$\hat{\theta}^N \in \arg \min_{\theta \in \Theta} \sum_{i=1}^N [y(x_i) - \eta(x_i, \theta)]^2,$$

is a random vector with rather complicated statistical properties even in seemingly simple situations. In the case of normal errors, optimum design can be based on fairly accurate approximations of the distribution of $\hat{\theta}^N$ for small N , see Pázman and Pronzato (1992); Pronzato and Pázman (1994). However, this is technically difficult and the standard approach (for large N) is to use the much simpler asymptotic normal approximation of the distribution of $\hat{\theta}^N$. Under suitable conditions on the sequence of design points x_1, x_2, \dots and on the model response $\eta(x, \theta)$, see Jennrich (1969); Gallant (1987); Ivanov (1997), $\hat{\theta}^N$ is proved to be strongly consistent ($\lim_{N \rightarrow \infty} \hat{\theta}^N = \bar{\theta}$ a.s.) and to converge in distribution to a normal random vector

$$\sqrt{N}(\hat{\theta}^N - \bar{\theta}) \xrightarrow{d} \nu \sim \mathcal{N}(0, \sigma^2 \mathbf{M}_\infty^{-1}(\bar{\theta})), \quad N \rightarrow \infty. \quad (1.4)$$

Here, the limit information matrix $\mathbf{M}_\infty(\theta)$, defined by

$$\mathbf{M}_\infty(\theta) = \lim_{N \rightarrow \infty} \frac{1}{N} \sum_{i=1}^N \mathbf{f}_\theta(x_i) \mathbf{f}_\theta^\top(x_i),$$

with the notation

$$\mathbf{f}_\theta(x) = \frac{\partial \eta(x, \theta)}{\partial \theta}, \quad (1.5)$$

is supposed to exist and to be nonsingular at $\bar{\theta}$. We denote by ξ the probability measure (called the *design measure*, or simply the design, a concept introduced by Kiefer and Wolfowitz (1959)), that corresponds to the limit of relative frequencies of the sequence $x_1, x_2 \dots$ when it exists. In that case, we can write

$$\mathbf{M}_\infty(\theta) = \mathbf{M}(\xi, \theta) = \int_{\mathcal{X}} \mathbf{f}_\theta(x) \mathbf{f}_\theta^\top(x) \xi(dx) \quad (1.6)$$

and $\sigma^2 \mathbf{M}^{-1}(\xi, \bar{\theta})$ forms an approximation of $N \text{Var}_{\bar{\theta}}(\hat{\theta}^N)$ for large N . This is basic in experimental design theory; for instance D -optimum design corresponds to maximizing $\det \mathbf{M}(\xi, \bar{\theta})$ with respect to ξ , see Fedorov (1972); Silvey (1980); Pázman (1986); Atkinson and Donev (1992).

On the other hand, when $\mathbf{M}(\xi, \theta)$ is singular the estimator $\hat{\theta}^N$ may not be asymptotically normal, or even not consistent and not uniquely defined. Still, it might seem possible to base the design on a generalization of the c -optimality criterion used for linear models, of the form

$$\Phi[\mathbf{M}(x_1, \dots, x_N; \theta)] = \text{Var}_\theta[h(\hat{\theta}^N)] + \mathcal{O}(1/N^2)$$

for a suitable choice of the function $h(\cdot)$, where Var_θ denotes the variance conditional to θ being the true value of the model parameters. In the regular case, this is justified by the delta-method, see Lehmann and Casella (1998, p. 61): under some regularity conditions on $h(\theta)$, from the asymptotic normality (1.4) of $\hat{\theta}^N$ one obtains the approximation

$$N \text{Var}_\theta[h(\hat{\theta}^N)] = \sigma^2 \frac{\partial h(\theta)}{\partial \theta^\top} \mathbf{M}^{-1}(\xi, \theta) \frac{\partial h(\theta)}{\partial \theta} + \mathcal{O}(1/N), \quad (1.7)$$

an expression similar to that used for c -optimality in linear models, see (1.3). Also, the estimator $h(\hat{\theta}^N)$ is asymptotically normal,

$$\sqrt{N}[h(\hat{\theta}^N) - h(\bar{\theta})] \xrightarrow{d} \nu \sim \mathcal{N}\left(0, \sigma^2 \left[\frac{\partial h(\theta)}{\partial \theta^\top} \mathbf{M}^{-1}(\xi, \theta) \frac{\partial h(\theta)}{\partial \theta} \right]_{\bar{\theta}}\right), \quad N \rightarrow \infty,$$

which we shall call *regular asymptotic normality* (with \mathbf{M}^{-1} replaced by a g -inverse when \mathbf{M} is singular, see Sect. 1.5).

The question is for which functions $h(\cdot)$, and under which conditions on the model $\eta(\cdot, \cdot)$ and design sequence $x_1, x_2 \dots$ a formula similar to (1.7) is justified and regular asymptotic normality also holds in the singular case? We shall show that essentially three types of conditions must be fulfilled:

- i) conditions on the convergence of the design sequence $x_1, x_2 \dots$ to the design measure ξ , as discussed in Sect. 1.2;
- ii) conditions on $\bar{\theta}$, the true value of the model parameters θ , in relation to the geometry of the model (Sect. 1.4);

iii) conditions on the function $h(\cdot)$.

The consistency of $\hat{\theta}^N$ and $h(\hat{\theta}^N)$ is considered in Sect. 1.3, based on the conditions of Sect. 1.2 on the design sequence. The regular asymptotic normality of $h(\hat{\theta}^N)$ is investigated in Sect. 1.5. The extension to a multidimensional function of interest $H(\theta)$ is considered in Sect. 1.6.

1.2 The convergence of the design sequence to a design measure

The investigation of the asymptotic properties of the estimator in a regression model requires the specification of the asymptotic behavior of the design sequence $(x_i)_i$. Rather than formulating conditions in terms of finite tail products, as in the seminal paper of Jennrich (1969), we shall state conditions in terms of limiting design measures. To each truncated subsequence x_1, \dots, x_N , we associate the empirical design measure ξ_N and its cumulative distribution function (c.d.f.)

$$\mathbb{F}_{\xi_N}(x) = \sum_{i=1, x_i \leq x}^N \frac{1}{N}$$

(where the inequality $x_i \leq x$ must be understood componentwise). The sequence $(\xi_N)_N$ is said to converge weakly to a limit design measure ξ with the c.d.f. \mathbb{F}_ξ if $\lim_{N \rightarrow \infty} \mathbb{F}_{\xi_N}(x) = \mathbb{F}_\xi(x)$ at every continuity point of \mathbb{F}_ξ (which corresponds to the weak convergence of probability measures, see, e.g., Billingsley (1971), Shiryaev (1996, p. 314)). However, even in linear models, weak convergence is not enough to ensure regular asymptotic normality of the estimator when the limiting design is singular. This is illustrated by the example below.

Example 1. Consider the linear regression model (1.1) with $p = 2$ and $\mathbf{f}(x) = (x \ x^2)^\top$. The true value $\bar{\theta}$ of the model parameters $\theta = (\theta_1 \ \theta_2)^\top$ is assumed to satisfy $\bar{\theta}_1 \geq 0, \bar{\theta}_2 < 0$. The errors ε_i are i.i.d., with zero mean and variance 1. We are interested in the estimation of the point x where $\eta(x, \theta) = \mathbf{f}^\top(x)\theta$ is maximum, that is, $h = h(\theta) = -\theta_1/(2\theta_2)$, with $h \geq 0$ and

$$\frac{\partial h(\theta)}{\partial \theta} = -\frac{1}{2\theta_2} \begin{pmatrix} 1 \\ 2h \end{pmatrix}.$$

Let θ^* be a prior guess for θ with $\theta_1^* \geq 0, \theta_2^* < 0, h_* = -\theta_1^*/(2\theta_2^*)$ denote the corresponding prior guess for h and define $x_* = 2h_*$. The c -optimum design ξ_* supported in $\mathcal{X} = [0, 1]$ that minimizes $[\partial h(\theta)/\partial \theta^\top \mathbf{M}^-(\xi) \partial h(\theta)/\partial \theta]_{\theta^*}$ is easily computed from Elfving's Theorem (1952), and is given by

$$\xi_* = \begin{cases} \gamma_* \delta_{\sqrt{2}-1} + (1 - \gamma_*) \delta_1 & \text{if } 0 \leq x_* \leq \sqrt{2} - 1 \text{ or } 1 \leq x_* \\ \delta_{x_*} & \text{otherwise} \end{cases} \quad (1.8)$$

with δ_x the delta measure that puts weight 1 at x and

$$\gamma_* = \frac{\sqrt{2}}{2} \frac{1 - x_*}{2(\sqrt{2} - 1) - x_*}.$$

Here we suppose that the prior guess θ^* is such that $\sqrt{2} - 1 < x_* \leq 1$ so that the c -optimum design ξ_* puts mass 1 at x_* . When ξ_* is used, $\theta_1 x_* + \theta_2 x_*^2$ is estimable since $\mathbf{u}_* = (x_* \ x_*^2)^\top$ is in the range of

$$\mathbf{M}(\xi_*) = \begin{pmatrix} x_*^2 & x_*^3 \\ x_*^3 & x_*^4 \end{pmatrix}.$$

The variance of $\mathbf{u}_*^\top \hat{\theta}^N$ for ξ_* , which we denote $\text{Var}(\mathbf{u}_*^\top \hat{\theta}^N | \xi_*)$, satisfies

$$N \text{Var}(\mathbf{u}_*^\top \hat{\theta}^N | \xi_*) = \mathbf{u}_*^\top \mathbf{M}^-(\xi_*) \mathbf{u}_* = 1,$$

with \mathbf{M}^- any g -inverse of \mathbf{M} .

We consider design sequences that converge to ξ_* and investigate the case when convergence is weak. Suppose that the design points satisfy

$$x_i = \begin{cases} x_* & \text{if } i = 2k - 1 \\ x_* + (1/k)^\alpha & \text{if } i = 2k \end{cases} \quad (1.9)$$

for some $\alpha \geq 0$, $i = 1, 2, \dots$, with $\sqrt{2} - 1 < x_* \leq 1$. From Corollary 1 of Wu (1980), one can show that the LSE $\hat{\theta}^N$ is strongly consistent when $\alpha \leq 1/2$, which we suppose in the rest of the example (see Pázman and Pronzato (2006) for details when $\alpha = 1/4$).

The variance of $\mathbf{u}_*^\top \hat{\theta}^N$, with $\mathbf{u}_* = (x_* \ x_*^2)^\top$, for the design ξ_N satisfies $N \text{Var}(\mathbf{u}_*^\top \hat{\theta}^N | \xi_N) = \mathbf{u}_*^\top \mathbf{M}^{-1}(\xi_N) \mathbf{u}_*$ with,

$$\mathbf{M}(\xi_N) = \begin{pmatrix} \mu_2(N) & \mu_3(N) \\ \mu_3(N) & \mu_4(N) \end{pmatrix}$$

and, for $N = 2M$, $\mu_i(N) = x_*^i/2 + (1/N) \sum_{k=1}^M [x_* + (1/k)^\alpha]^i$, $i = 2, 3, 4$. We then obtain

$$\lim_{N \rightarrow \infty} N \text{Var}(\mathbf{u}_*^\top \hat{\theta}^N | \xi_N) = V(\alpha) = \frac{2(1 - \alpha)^2}{\alpha^2 + (1 - \alpha)^2}, \quad (1.10)$$

which is monotonically decreasing in α for α varying between 0 and $1/2$, with $V(0) = 2$ and $V(1/2) = 1$. For any $\alpha \in [0, 1/2)$ we thus have

$$\lim_{N \rightarrow \infty} N \text{Var}(\mathbf{u}_*^\top \hat{\theta}^N | \xi_N) = V(\alpha) > N \text{Var}(\mathbf{u}_*^\top \hat{\theta}^N | \xi_*) = 1,$$

that is, the limiting variance for ξ_N is always larger than the variance for the limiting design ξ_* (this is due to the discontinuity of the function $\mathbf{M}(\xi) \mapsto N \text{Var}(\mathbf{u}_*^\top \hat{\theta}^N | \xi)$ at $\mathbf{M}(\xi_*)$, see Pázman (1980, 1986 p. 67)).

Moreover, we can easily show that Lindeberg's condition is satisfied for any linear combination of θ , see e.g. Shiryaev (1996), and, for any $\mathbf{u} \neq 0$

$$\sqrt{N} \frac{\mathbf{u}^\top (\hat{\theta}^N - \bar{\theta})}{(\mathbf{u}^\top \mathbf{M}^{-1}(\xi_N) \mathbf{u})^{1/2}} \xrightarrow{d} \zeta \sim \mathcal{N}(0, 1), \quad N \rightarrow \infty.$$

Since $\mathbf{u}_*^\top \mathbf{M}^{-1}(\xi_N) \mathbf{u}_*$ tends to $V(\alpha)$,

$$\sqrt{N} \mathbf{u}_*^\top (\hat{\theta}^N - \bar{\theta}) \xrightarrow{d} \zeta_* \sim \mathcal{N}(0, V(\alpha)). \quad N \rightarrow \infty,$$

On the other hand, $\mathbf{u}^\top \mathbf{M}^{-1}(\xi_N) \mathbf{u}$ grows as $N^{2\alpha}$ for \mathbf{u} not parallel to \mathbf{u}_* . For instance, for $\mathbf{u} = \mathbf{u}_0 = (1 \ 0)^\top$ we obtain

$$N^{1/2-\alpha} \mathbf{u}_0^\top (\hat{\theta}^N - \bar{\theta}) \xrightarrow{d} \zeta_0 \sim \mathcal{N}(0, W(\alpha)), \quad N \rightarrow \infty,$$

with

$$W(\alpha) = 2^{2(1-\alpha)} \frac{(1-2\alpha)(1-\alpha)^2}{\alpha^2 + (1-\alpha)^2}. \quad (1.11)$$

Come back now to the estimation of $h(\theta) = -\theta_1/(2\theta_2)$.

When $\theta_1 + x_* \bar{\theta}_2 \neq 0$, that is when $h(\bar{\theta}) \neq h_*$, which corresponds to the typical situation, we have

$$h(\hat{\theta}^N) = h(\bar{\theta}) + (\hat{\theta}^N - \bar{\theta})^\top \left[\frac{\partial h(\theta)}{\partial \theta} \Big|_{\bar{\theta}} + o_p(1) \right]$$

with $\partial h(\theta)/\partial \theta|_{\bar{\theta}} = -1/(2\bar{\theta}_2)[1 \ 2h(\bar{\theta})]^\top$ not parallel to \mathbf{u}_* , and

$$N^{1/2-\alpha} [h(\hat{\theta}^N) - h(\bar{\theta})] \xrightarrow{d} \zeta \sim \mathcal{N}(0, v_{\bar{\theta}}), \quad N \rightarrow \infty,$$

with $v_{\bar{\theta}} = W(\alpha)[x_* - 2h(\bar{\theta})]^2/(4\bar{\theta}_2^2 x_*^2)$ where $W(\alpha)$ is given by (1.11); $h(\hat{\theta}^N)$ is thus asymptotically normal but converges as $N^{\alpha-1/2}$.

In the particular situation where the prior guess h_* coincides with the true value $h(\bar{\theta})$, $\bar{\theta}_1 + x_* \bar{\theta}_2 = 0$ and we write

$$\begin{aligned} h(\hat{\theta}^N) &= h(\bar{\theta}) + (\hat{\theta}^N - \bar{\theta})^\top \frac{\partial h(\theta)}{\partial \theta} \Big|_{\bar{\theta}} \\ &\quad + \frac{1}{2} (\hat{\theta}^N - \bar{\theta})^\top \left[\frac{\partial^2 h(\theta)}{\partial \theta \partial \theta^\top} \Big|_{\bar{\theta}} + o_p(1) \right] (\hat{\theta}^N - \bar{\theta}) \end{aligned} \quad (1.12)$$

with

$$\frac{\partial h(\theta)}{\partial \theta} \Big|_{\bar{\theta}} = -\frac{1}{2\bar{\theta}_2 x_*} \mathbf{u}_* \quad \text{and} \quad \frac{\partial^2 h(\theta)}{\partial \theta \partial \theta^\top} \Big|_{\bar{\theta}} = \frac{1}{2\bar{\theta}_2^2} \begin{pmatrix} 0 & 1 \\ 1 & 2x_* \end{pmatrix}.$$

Define $\delta_N = \hat{\theta}^N - \bar{\theta}$ and $E_N = 2\bar{\theta}_2^2 \delta_N^\top \partial^2 h(\theta)/(\partial \theta \partial \theta^\top)|_{\bar{\theta}} \delta_N$. The eigenvector decomposition of $\partial^2 h(\theta)/(\partial \theta \partial \theta^\top)|_{\bar{\theta}}$ gives

$$E_N = \beta [(\mathbf{v}_1^\top \delta_N)^2 - (\mathbf{v}_2^\top \delta_N)^2]$$

with $\mathbf{v}_{1,2} = (1 \ x_* \pm \sqrt{1+x_*^2})$ and $\beta = (x_* + \sqrt{1+x_*^2})/[2(1+x_*^2+x_*\sqrt{1+x_*^2})]$. We then obtain

$$N^{1/2-\alpha} \mathbf{v}_{1,2}^\top \delta_N \xrightarrow{d} \zeta_{1,2} \sim \mathcal{N}(0, [1+1/x_*^2]W(\alpha)), \quad N \rightarrow \infty.$$

From (1.12), the limiting distribution of $h(\hat{\theta}^N)$ is not normal when $\alpha \geq 1/4$. When $\alpha < 1/4$ we have $\sqrt{N} E_N = o_p(1)$, $N \rightarrow \infty$, and (1.12) implies

$$\sqrt{N}[h(\hat{\theta}^N) - h(\bar{\theta})] \xrightarrow{d} \zeta \sim \mathcal{N}(0, V(\alpha)/(4\bar{\theta}_2^2 x_*^2)), \quad N \rightarrow \infty,$$

with $V(\alpha)$ given by (1.10). Note that the limiting variance is larger than

$$[\partial h(\theta)/\partial \theta^\top \mathbf{M}^-(\xi_*) \partial h(\theta)/\partial \theta]_{\bar{\theta}} = 1/(4\bar{\theta}_2^2 x_*^2).$$

To summarize, the estimation of $h(\theta)$ requires $\alpha \leq 1/2$ in the design (1.9), $h(\hat{\theta}^N)$ is then generally asymptotically normal but converges as slowly as $N^{\alpha-1/2}$. In the special case where the design is optimum for the true value $h(\bar{\theta})$, $h(\hat{\theta}^N)$ is not asymptotically normal when $1/4 \leq \alpha \leq 1/2$; it is asymptotically normal for $\alpha < 1/4$ and converges as $1/\sqrt{N}$, but the limiting variance differs from that computed from the limiting optimum design ξ_* . \square

Regular asymptotic normality may thus fail to hold when the design sequence converges weakly to a singular design. Stronger types of convergence are required and we shall consider two situations that arise quite naturally.

The first one concerns the case where *the limiting design ξ is discrete*. Since, from Caratheodory's Theorem, optimum designs can be written as discrete probability measures (see e.g. (Fedorov, 1972; Silvey, 1980)), sequences of design points that converge to discrete measures are of special interest. In that case, we shall require *strong convergence* (or convergence in variation, see Shiryaev (1996, p. 360)) of the empirical measure ξ_N .

Definition 1. *Let ξ be a discrete probability measure on \mathcal{X} , with finite support*

$$\mathcal{S}_\xi = \{x \in \mathcal{X} : \xi(\{x\}) > 0\}.$$

We say that the design sequence $(x_i)_i$ converges strongly to ξ when

$$\lim_{N \rightarrow \infty} \xi_N(\{x\}) = \xi(\{x\}) \quad \text{for any } x \in \mathcal{X}.$$

In the second situation the limiting design measure ξ is not necessary discrete and ξ_N converges weakly to ξ , but we require that *the design sequence is a random sample from ξ* .

Definition 2. *Let ξ be a probability measure on \mathcal{X} . We say that the design sequence $(x_i)_i$ is a randomized design with measure ξ if the points $x_i \in \mathcal{X}$ are independently sampled according to the probability measure ξ .*

The following example shows, however, that strong convergence of the design sequence is not enough to ensure the regular asymptotic normality of the estimator and further conditions on the model are required. They will be presented in Sect. 1.4 and 1.5.

Example 2. This is a slight extension of the example considered in (Pázman and Pronzato, 2006). We consider the same linear regression model as in Example 1, but now the design is such that $N - m$ observations are taken at $x = x_* = 2h_* \in (0, 1]$, with $h_* = -\theta_1^*/(2\theta_2^*)$ a prior guess for the location of the maximum of the function $\theta_1 x + \theta_2 x^2$, and m observations are taken at $x = z \in (0, 1]$, $z \neq x_*$. We shall suppose that either m is fixed¹ or $m \rightarrow \infty$ with $m/N \rightarrow 0$ as N tends to infinity. In both cases the sequence $(x_i)_i$ converges strongly to δ_{x_*} as $N \rightarrow \infty$, in the sense of Definition 1. Note that $\delta_{x_*} = \xi_*$, the c -optimum design measure for $h(\theta_*)$, when $\sqrt{2} - 1 < x_* \leq 1$, see (1.8).

The LSE $\hat{\theta}^N$ is given by

$$\hat{\theta}^N = \bar{\theta} + \frac{1}{x_* z (x_* - z)} \left[\frac{\beta_m}{\sqrt{m}} \begin{pmatrix} x_*^2 \\ -x_* \end{pmatrix} + \frac{\gamma_{N-m}}{\sqrt{N-m}} \begin{pmatrix} -z^2 \\ z \end{pmatrix} \right] \quad (1.13)$$

where $\beta_m = (1/\sqrt{m}) \sum_{x_i=z} \varepsilon_i$ and $\gamma_{N-m} = (1/\sqrt{N-m}) \sum_{x_i=x_*} \varepsilon_i$ are independent random variables that tend to be distributed $\mathcal{N}(0, 1)$ as $m \rightarrow \infty$ and $N - m \rightarrow \infty$. Obviously, $\hat{\theta}^N$ is consistent if and only if $m \rightarrow \infty$. However, $h(\hat{\theta}^N)$ is also consistent when m is finite provided that $\bar{\theta}_1 + x_* \bar{\theta}_2 = 0$. Indeed, for m finite we have

$$\hat{\theta}^N \xrightarrow{\text{a.s.}} \hat{\theta}^\# = \bar{\theta} + \frac{1}{z(x_* - z)} \frac{\beta_m}{\sqrt{m}} \begin{pmatrix} x_* \\ -1 \end{pmatrix}, \quad N \rightarrow \infty,$$

and $h(\hat{\theta}^\#) = -\hat{\theta}_1^\#/(2\hat{\theta}_2^\#) = x_*/2 = h(\bar{\theta})$. Also,

$$\sqrt{N}[h(\hat{\theta}^N) - h(\bar{\theta})] = \sqrt{N}[h(\hat{\theta}^N) - h(\hat{\theta}^\#)] = \sqrt{N}(\hat{\theta}^N - \hat{\theta}^\#)^\top \left[\frac{\partial h(\theta)}{\partial \theta} \Big|_{\hat{\theta}^\#} + o_p(1) \right]$$

with $\partial h(\theta)/\partial \theta|_{\hat{\theta}^\#} = -1/(2\hat{\theta}_2^\#)[1 \quad x_*]^\top$ and

$$\sqrt{N}(\hat{\theta}^N - \hat{\theta}^\#) = \frac{\sqrt{N}}{x_*(x_* - z)} \frac{\gamma_{N-m}}{\sqrt{N-m}} \begin{pmatrix} -z \\ 1 \end{pmatrix}.$$

Therefore, $\sqrt{N}[h(\hat{\theta}^N) - h(\bar{\theta})] \xrightarrow{d} \nu/(2\zeta)$ with $\nu \sim \mathcal{N}(0, 1/x_*^2)$ and $\zeta \sim \mathcal{N}(\bar{\theta}_2, 1/[mz^2(x_* - z)^2])$, and $h(\hat{\theta}^N)$ is not asymptotically normal.

¹ Taking only a finite number of observations at another place than x_* might seem an odd strategy; note, however, that the algorithm of Wynn (1972) for the minimization of $[\partial h(\theta)/\partial \theta^\top \mathbf{M}^-(\xi) \partial h(\theta)/\partial \theta]_{\theta^*}$ generates such a sequence of design points when the design space is $\mathcal{X} = [-1, 1]$, see (Pázman and Pronzato, 2006), or when \mathcal{X} is a finite set containing x_* .

Suppose now that $m = m(N) \rightarrow \infty$ with $m/N \rightarrow 0$ as $N \rightarrow \infty$. If $\bar{\theta}_1 + x_*\bar{\theta}_2 \neq 0$ we can write

$$\sqrt{m}[h(\hat{\theta}^N) - h(\bar{\theta})] = \sqrt{m}(\hat{\theta}^N - \bar{\theta})^\top \left[\frac{\partial h(\theta)}{\partial \theta} \Big|_{\bar{\theta}} + o_p(1) \right]$$

and using (1.13) we get

$$\sqrt{m}[h(\hat{\theta}^N) - h(\bar{\theta})] \xrightarrow{d} \zeta \sim \mathcal{N} \left(0, \frac{(\bar{\theta}_1 + x_*\bar{\theta}_2)^2}{4\bar{\theta}_2^4 z^2 (x_* - z)^2} \right), \quad N \rightarrow \infty.$$

$h(\hat{\theta}^N)$ thus converges as $1/\sqrt{m}$ and is asymptotically normal with a limiting variance depending on z . If $\bar{\theta}_1 + x_*\bar{\theta}_2 = 0$,

$$\sqrt{N}[h(\hat{\theta}^N) - h(\bar{\theta})] = \sqrt{N} \left(-\frac{\hat{\theta}_1^N}{2\hat{\theta}_2^N} - \frac{x_*}{2} \right) = -\frac{\sqrt{N}}{2} \frac{\gamma_{N-m}}{\sqrt{N-m}} \frac{1}{x_*\hat{\theta}_2^N}$$

and

$$\sqrt{N}[h(\hat{\theta}^N) - h(\bar{\theta})] \xrightarrow{d} \zeta \sim \mathcal{N}(0, 1/(4\bar{\theta}_2^2 x_*^2)).$$

This is the only situation within Examples 1 and 2 where regular asymptotic normality holds: $h(\hat{\theta}^N)$ converges as $1/\sqrt{N}$, is asymptotically normal and has a limiting variance that can be computed from the limiting design ξ_* , that is, which coincides with $[\partial h(\theta)/\partial \theta^\top \mathbf{M}^-(\xi_*) \partial h(\theta)/\partial \theta]_{\bar{\theta}}$. Note that assuming that $\bar{\theta}_1 + x_*\bar{\theta}_2 = 0$ amounts to assuming that the prior guess $h_* = x_*/2$ coincides with the true location of the maximum of the model response, which is rather unrealistic. \square

1.3 Consistency of estimators

The (strong) consistency results presented below are based on the following two lemmas which respectively concern designs satisfying the conditions of Definitions 1 and 2. The proofs are given in Appendix.

Lemma 1. *Let the sequence $(x_i)_i$ converge strongly to a discrete design ξ in the sense of Definition 1. Assume that $a(x, \theta)$ is a bounded function on $\mathcal{X} \times \Theta$ and that $(\alpha_i)_i$ is an i.i.d. sequence of random variables having finite mean and variance. Then*

$$\lim_{N \rightarrow \infty} \frac{1}{N} \sum_{k=1}^N a(x_k, \theta) \alpha_k = \mathbb{E}\{\alpha_1\} \sum_{x \in \mathcal{S}_\xi} a(x, \theta) \xi(\{x\})$$

a.s. with respect to $\alpha_1, \alpha_2 \dots$ and uniformly on Θ .

Lemma 2. *Let $(z_i)_i$ be a sequence of i.i.d. random vectors from \mathbb{R}^r and $a(z, \theta)$ be a Borel measurable real function on $\mathbb{R}^r \times \Theta$, continuous in $\theta \in \Theta$ for any z , with Θ a compact subset of \mathbb{R}^p . Assume that*

$$\mathbb{E} \left\{ \max_{\theta \in \Theta} |a(z_1, \theta)| \right\} < \infty,$$

then $\mathbb{E} \{a(z_1, \theta)\}$ is continuous in $\theta \in \Theta$ and

$$\lim_{N \rightarrow \infty} \frac{1}{N} \sum_{i=1}^N a(z_i, \theta) = \mathbb{E} \{a(z_1, \theta)\}$$

a.s. and uniformly on Θ .

Next result (Theorem 1) is quite standard and concerns the set of possible limiting points of the sequence of LSE of θ . We shall denote $y^N = (y(x_1), \dots, y(x_N))$, $y = (y(x_i))_i$ the sequence of observations $y(x_i) = \eta(x_i, \bar{\theta}) + \varepsilon_i$, where $\bar{\theta}$ is the unknown true value of the model parameters, and

$$\begin{aligned} J_N(\theta, y^N) &= \frac{1}{N} \sum_{i=1}^N [y(x_i) - \eta(x_i, \theta)]^2, \\ J(\theta) &= \int_{\mathcal{X}} [\eta(x, \bar{\theta}) - \eta(x, \theta)]^2 \xi(dx) + \sigma^2. \end{aligned} \quad (1.14)$$

We shall assume the following.

ASSUMPTION A1: Θ is a compact subset of \mathbb{R}^p and $\eta(x, \theta)$ is bounded on $\mathcal{X} \times \Theta$ and continuous in θ for any $x \in \mathcal{X}$, with $\mathcal{X} \subset \mathbb{R}^k$.

Theorem 1. *Let the sequence $(x_i)_i$ either converge to ξ in the sense of Definition 1, or be generated by ξ according to Definition 2. Then, under A1, with probability one all limit points of the sequence $(\hat{\theta}^N(y^N))_N$ of LSE,*

$$\hat{\theta}^N(y^N) \in \arg \min_{\theta \in \Theta} J_N(\theta, y^N), \quad (1.15)$$

are elements of the set

$$\Theta^\# = \arg \min_{\theta \in \Theta} J(\theta).$$

Proof. Lemmas 1 and 2 allow to proceed essentially in a standard way. We can write

$$\begin{aligned} \frac{1}{N} \sum_{k=1}^N [y(x_k) - \eta(x_k, \theta)]^2 &= \frac{1}{N} \sum_{k=1}^N \varepsilon_k^2 + \frac{2}{N} \sum_{k=1}^N [\eta(x_k, \bar{\theta}) - \eta(x_k, \theta)] \varepsilon_k \\ &\quad + \frac{1}{N} \sum_{k=1}^N [\eta(x_k, \bar{\theta}) - \eta(x_k, \theta)]^2, \end{aligned}$$

with $\varepsilon_k = y(x_k) - \eta(x_k, \bar{\theta})$. The first term of the right-hand side converges a.s. to σ^2 by the strong law of large numbers. As required in Lemma 1, $\eta(x, \bar{\theta}) - \eta(x, \theta)$ is bounded on $\mathcal{X} \times \Theta$, hence

$$\begin{aligned} \lim_{N \rightarrow \infty} \frac{1}{N} \sum_{k=1}^N [\eta(x_k, \bar{\theta}) - \eta(x_k, \theta)] \varepsilon_k &= \sum_{x \in \mathcal{S}_\xi} [\eta(x, \bar{\theta}) - \eta(x, \theta)] \xi(\{x\}) \mathbb{E} \{\varepsilon_1\} \\ &= 0 \end{aligned}$$

a.s. and uniformly on Θ . Similarly, taking $z = (x, \varepsilon)$ and $a(z, \theta) = [\eta(x, \bar{\theta}) - \eta(x, \theta)]\varepsilon$ in Lemma 2, we obtain

$$\lim_{N \rightarrow \infty} \frac{1}{N} \sum_{k=1}^N [\eta(x_k, \bar{\theta}) - \eta(x_k, \theta)] \varepsilon_k = \mathbb{E}_\xi \{ \eta(x, \bar{\theta}) - \eta(x, \theta) \} \mathbb{E} \{ \varepsilon_1 \} = 0$$

a.s. and uniformly on Θ . By the same arguments we also obtain

$$\lim_{N \rightarrow \infty} \frac{1}{N} \sum_{k=1}^N [\eta(x_k, \bar{\theta}) - \eta(x_k, \theta)]^2 = \int_{\mathcal{X}} [\eta(x, \bar{\theta}) - \eta(x, \theta)]^2 \xi(dx)$$

a.s. and uniformly on Θ . We have thus proved that the sequence $(J_N(\theta, y^N))_N$ converges a.s. and uniformly on Θ to $J(\theta)$.

Let $\theta^\# = \theta^\#(y)$ be a limit point of $(\hat{\theta}^N(y^N))_N$ (which exists since Θ is compact). There exists a subsequence $(\hat{\theta}^{N_t})_t$ of this sequence that converges to $\theta^\#$. From the definitions of $J_N(\theta, y^N)$ and $\hat{\theta}^N(y^N)$, we can write

$$J_{N_t}(\bar{\theta}, y^{N_t}) = \frac{1}{N_t} \sum_{k=1}^{N_t} \varepsilon_k^2 \geq J_{N_t}(\hat{\theta}^{N_t}, y^{N_t}).$$

The left-hand side converges to σ^2 a.s. as $t \rightarrow \infty$ while the right-hand side converges a.s. to $J(\theta^\#) = \int_{\mathcal{X}} [\eta(x, \bar{\theta}) - \eta(x, \theta^\#)]^2 \xi(dx) + \sigma^2$ (since $J_N(\theta, y^N)$ converges a.s. and uniformly in θ to $J(\theta)$). This implies that $\int_{\mathcal{X}} [\eta(x, \bar{\theta}) - \eta(x, \theta^\#)]^2 \xi(dx) = 0$, i.e. that $\theta^\# \in \Theta^\#$. \blacksquare

One may notice that Theorem 1 does not mean that $\hat{\theta}^N$ can take any value in $\Theta^\#$. For instance, in Example 2 we have $\Theta^\# = \{\theta : \theta_1 + x_*\theta_2 = \bar{\theta}_1 + x_*\bar{\theta}_2\}$ although $\hat{\theta}^N$ converges a.s. to $\bar{\theta}$ when the number m of observations at $z \neq x_*$ tends to infinity, see also Wu (1981)².

Next result concerns the consistency of $h(\hat{\theta}^N)$. We shall require the following.

² This reference presents sufficient conditions for consistency that are weaker than those considered here, in the sense that they do not require the design sequence $(x_i)_i$ to obey Definitions 1 or 2. However, this may lead to irregular asymptotic normality, as shown in Examples 1 and 2, and is not considered in this paper.

ASSUMPTION A2: *The function $h(\theta)$ is continuous in $\theta \in \Theta$ and such that*

$$\int_{\mathcal{X}} [\eta(x, \theta) - \eta(x, \bar{\theta})]^2 \xi(dx) = 0 \Rightarrow h(\theta) = h(\bar{\theta}).$$

Theorem 2 (Consistency of $h(\hat{\theta}^N)$). *Under the assumptions of Theorem 1 we suppose that $h(\cdot)$ satisfies A2. Let $(\hat{\theta}^N(y^N))_N$ be any sequence defined by (1.15). Then*

$$\lim_{N \rightarrow \infty} h[\hat{\theta}^N(y^N)] = h(\bar{\theta}) \quad \text{a.s.}$$

Proof. As in the proof of Theorem 1 we extract from $(\hat{\theta}^N(y^N))_N$ a subsequence $(\hat{\theta}^{N_t})_t$ converging to $\theta^\# = \theta^\#(y) \in \Theta^\#$. From the continuity of $h(\cdot)$, $\lim_{t \rightarrow \infty} h(\hat{\theta}^{N_t}) = h(\theta^\#)$, and, from Theorem 1 and A2, $h(\theta^\#) = h(\bar{\theta})$ a.s. Therefore, every converging subsequence of $(h[\hat{\theta}^N(y^N)])_N$ has a.s. the same limit $h(\bar{\theta})$, which is the limit of the whole sequence $(h[\hat{\theta}^N(y^N)])_N$. ■

1.4 On the geometry of the model under the design measure ξ

We shall assume the following in the rest of the paper.

ASSUMPTION A3: $\bar{\theta} \in \text{int}(\Theta)$ and, for any $x \in \mathcal{X}$, $\eta(x, \theta)$ is two times continuously differentiable with respect to $\theta \in \Theta_0$; these first two derivatives are bounded on $\mathcal{X} \times \Theta$.

Let ξ be a design measure, and $\mathcal{L}_2(\xi)$ be the Hilbert space of real valued functions ϕ on \mathcal{X} which are square integrable, i.e.,

$$\int_{\mathcal{X}} \phi^2(x) \xi(dx) < \infty.$$

Two functions ϕ and ϕ^* are equivalent in $\mathcal{L}_2(\xi)$ if $\int_{\mathcal{X}} [\phi(x) - \phi^*(x)]^2 \xi(dx) = 0$, that is, $\phi(x) = \phi^*(x)$ ξ -a.s., which we shall denote

$$\phi \stackrel{\xi}{\equiv} \phi^*.$$

Note that for a discrete measure ξ the equivalence $\phi \stackrel{\xi}{\equiv} \phi^*$ means that $\phi(x) = \phi^*(x)$ on the support of ξ . The elements of $\mathcal{L}_2(\xi)$ are thus classes of equivalent functions rather than functions; by the sentence “the function ϕ belongs to a subset \mathcal{A} of $\mathcal{L}_2(\xi)$ ” we mean that the whole class containing ϕ belongs to \mathcal{A} , which we shall denote $\phi \stackrel{\xi}{\in} \mathcal{A}$. The notation $\mathcal{A} \stackrel{\xi}{\subset} \mathcal{B}$ is defined similarly. The inner product and the norm in $\mathcal{L}_2(\xi)$ are respectively defined by

$$\langle \phi, \phi^* \rangle_\xi = \int_{\mathcal{X}} \phi(x) \phi^*(x) \xi(dx) \quad \text{and} \quad \|\phi\|_\xi = \left[\int_{\mathcal{X}} \phi^2(x) \xi(dx) \right]^{1/2}.$$

Under A3, the functions $\eta(\cdot, \theta)$ and

$$\{\mathbf{f}_\theta\}_i(\cdot) = \partial \eta(\cdot, \theta) / \partial \theta_i, \quad i = 1, \dots, p, \quad (1.16)$$

belong to $\mathcal{L}_2(\xi)$. We shall denote

$$\theta \stackrel{\xi}{\sim} \theta^*$$

when the values θ and θ^* satisfy $\eta(\cdot, \theta) \stackrel{\xi}{=} \eta(\cdot, \theta^*)$. Notice that $\theta \stackrel{\xi}{\sim} \theta^*$ does not imply that $\{\mathbf{f}_\theta\}_i \stackrel{\xi}{=} \{\mathbf{f}_{\theta^*}\}_i$ (this is the object of Lemma 3 below). Also note that $\theta \stackrel{\xi}{\sim} \bar{\theta}$ is equivalent to $\theta \in \Theta^\#$ as defined in Theorem 1.

For any $\theta \in \Theta_0$ we define an operator P_θ which acts upon any $\phi \in \mathcal{L}_2(\xi)$ as follows,

$$(P_\theta \phi)(x') = \mathbf{f}_\theta^\top(x') \mathbf{M}^+(\xi, \theta) \int_{\mathcal{X}} \mathbf{f}_\theta(x) \phi(x) \xi(dx)$$

where \mathbf{M}^+ denotes the Moore-Penrose g -inverse of \mathbf{M} , $\mathbf{M}(\xi, \theta)$ is defined in (1.6) and $\mathbf{f}_\theta(x)$ in (1.5). We keep the same notation $P_\theta \phi$ when ϕ is vector valued with components in $\mathcal{L}_2(\xi)$. We denote

$$\mathcal{L}_\theta = \{ \alpha^\top \mathbf{f}_\theta(\cdot) : \alpha \in \mathbb{R}^p \}.$$

From $\mathbf{M} \mathbf{M}^+ \mathbf{M} = \mathbf{M}$ and $\mathbf{M}^+ \mathbf{M} \mathbf{M}^+ = \mathbf{M}^+$ we obtain

$$\begin{aligned} \phi \in \mathcal{L}_2(\xi) &\Rightarrow P_\theta \phi \stackrel{\xi}{\in} \mathcal{L}_\theta, \\ \phi \in \mathcal{L}_\theta &\Rightarrow P_\theta \phi \stackrel{\xi}{=} \phi, \\ \phi, \phi^* \in \mathcal{L}_2(\xi) &\Rightarrow \langle \phi, P_\theta \phi^* \rangle_\xi = \langle P_\theta \phi, \phi^* \rangle_\xi, \end{aligned}$$

hence P_θ is the orthogonal projector onto \mathcal{L}_θ .

We shall need the following technical assumptions on the geometry of the model.

ASSUMPTION A4: For any point $\theta^* \stackrel{\xi}{\sim} \bar{\theta}$ there exists a neighborhood $\mathcal{V}(\theta^*)$ such that

$$\forall \theta \in \mathcal{V}(\theta^*), \quad \text{rank}[\mathbf{M}(\xi, \theta)] = \text{rank}[\mathbf{M}(\xi, \theta^*)].$$

ASSUMPTION A5: Define

$$\mathcal{S}_\epsilon = \left\{ \theta \in \text{int}(\Theta) : \|\eta(\cdot, \theta) - \eta(\cdot, \bar{\theta})\|_\xi^2 < \epsilon \right\}.$$

There exists $\epsilon > 0$ such that for every $\theta^\#, \theta^* \in \mathcal{S}_\epsilon$ we have

$$\left[\frac{\partial}{\partial \theta} \|\eta(\cdot, \theta) - \eta(\cdot, \theta^\#)\|_\xi^2 \right]_{\theta=\theta^*} = 0 \Rightarrow \theta^\# \stackrel{\xi}{\sim} \theta^*.$$

The assumptions A4 and A5 admit a straightforward geometrical and statistical interpretation in the case where the measure ξ is discrete. Let $\{x^{(1)}, \dots, x^{(k)}\}$ denote the support of ξ and define

$$\eta(\theta) = (\eta(x^{(1)}, \theta), \dots, \eta(x^{(k)}, \theta))^\top.$$

The set $\mathcal{E} = \{\eta(\theta) : \theta \in \Theta\}$ is then the expectation surface of the model under the design ξ and $\theta^* \stackrel{\xi}{\sim} \bar{\theta}$ is equivalent to $\eta(\theta^*) = \eta(\bar{\theta})$. When A4 is not satisfied, it means that the surface \mathcal{E} possesses edges, and the point $\eta(\bar{\theta})$ belongs to such an edge (although $\bar{\theta} \in \text{int}(\Theta)$). When A5 is not satisfied, it means that the surface \mathcal{E} intersects itself at the point $\eta(\bar{\theta})$ and therefore there are points $\eta(\theta)$ arbitrarily close to $\eta(\bar{\theta})$ with θ far from $\bar{\theta}$. In any of such circumstances asymptotic normality of the least squares estimator does not hold (from the geometrical interpretation of least squares estimation as the projection of the vector of observations onto the expectation surface \mathcal{E}).

The geometrical assumptions above yield the following.

Lemma 3. *Under A4 and A5, $\bar{\alpha} \stackrel{\xi}{\sim} \bar{\theta}$ implies*

$$\forall \phi, \phi^* \in \mathcal{L}_2(\xi), \quad \langle \phi, P_{\bar{\alpha}} \phi^* \rangle_\xi = \langle \phi, P_{\bar{\theta}} \phi^* \rangle_\xi.$$

The proof is given in Appendix.

One may notice that Assumptions A4 and A5 are only used to prove

$$\bar{\alpha} \stackrel{\xi}{\sim} \bar{\theta} \Rightarrow \mathcal{L}_{\bar{\alpha}} \stackrel{\xi}{=} \mathcal{L}_{\bar{\theta}},$$

to be used in Theorem 3, see the proof of Lemma 3. This result could also be obtained in a different way. Indeed, from the properties of the projector $P_{\bar{\theta}}$ we have

$$\begin{aligned} \mathcal{L}_{\bar{\alpha}} \stackrel{\xi}{\subset} \mathcal{L}_{\bar{\theta}} &\Leftrightarrow P_{\bar{\theta}} \{\mathbf{f}_{\bar{\alpha}}\}_i \stackrel{\xi}{=} \{\mathbf{f}_{\bar{\alpha}}\}_i, \quad i = 1, \dots, p, \\ &\Leftrightarrow \|P_{\bar{\theta}} \{\mathbf{f}_{\bar{\alpha}}\}_i - \{\mathbf{f}_{\bar{\alpha}}\}_i\|_\xi^2 = 0, \quad i = 1, \dots, p, \\ &\Leftrightarrow \{\mathbf{M}(\xi, \bar{\alpha})\}_{ii} - \{\mathbf{N}(\xi, \bar{\alpha}, \bar{\theta}) \mathbf{M}^+(\xi, \bar{\theta}) \mathbf{N}(\xi, \bar{\theta}, \bar{\alpha})\}_{ii} = 0, \quad i = 1, \dots, p, \end{aligned}$$

where we used the notation

$$\mathbf{N}(\xi, \bar{\alpha}, \bar{\theta}) = \int_{\mathcal{X}} \mathbf{f}_{\bar{\alpha}}(x) \mathbf{f}_{\bar{\theta}}^\top(x) \xi(dx).$$

Therefore, in Theorem 3 below instead of Assumptions A4 and A5 we can use equivalently the following, which may be easier to check although it does not have a geometrical interpretation.

ASSUMPTION B1: $\bar{\alpha} \stackrel{\xi}{\approx} \bar{\theta}$ implies

$$\begin{aligned} \{\mathbf{M}(\xi, \bar{\alpha})\}_{ii} &= \{\mathbf{N}(\xi, \bar{\alpha}, \bar{\theta})\mathbf{M}^+(\xi, \bar{\theta})\mathbf{N}(\xi, \bar{\theta}, \bar{\alpha})\}_{ii}, \\ \{\mathbf{M}(\xi, \bar{\theta})\}_{ii} &= \{\mathbf{N}(\xi, \bar{\theta}, \bar{\alpha})\mathbf{M}^+(\xi, \bar{\alpha})\mathbf{N}(\xi, \bar{\alpha}, \bar{\theta})\}_{ii}, \end{aligned}$$

for every $i = 1, \dots, p$.

1.5 The regular asymptotic normality of $h(\hat{\theta}^N)$

Definition 3. We say that $h(\hat{\theta}^N)$ satisfies the property of regular asymptotic normality when $\sqrt{N}[h(\hat{\theta}^N) - h(\bar{\theta})]$ converges in distribution to a variable distributed $\mathcal{N}(0, [\partial h(\theta)/\partial \theta^\top \mathbf{M}^-(\xi, \theta) \partial h(\theta)/\partial \theta]_{\bar{\theta}})$.

Examples 1 and 2 demonstrate that for a singular ξ regular asymptotic normality does not hold in general when $h(\cdot)$ is a nonlinear function of θ . We thus introduce an assumption on $h(\cdot)$ in addition to A2.

ASSUMPTION A6: The function $h(\cdot)$ is defined and has a continuous nonzero vector of derivatives $\partial h(\theta)/\partial \theta$ on Θ_0 . Moreover, for any $\theta \stackrel{\xi}{\approx} \bar{\theta}$ there exists a linear mapping A_θ from $\mathcal{L}_2(\xi)$ to \mathbb{R} (a continuous linear functional on $\mathcal{L}_2(\xi)$), such that $A_\theta = A_{\bar{\theta}}$ and that

$$\frac{\partial h(\theta)}{\partial \theta_i} = A_\theta [\{\mathbf{f}_\theta\}_i], \quad i = 1, \dots, p,$$

where $\{\mathbf{f}_\theta\}_i$ is defined by (1.16).

This receives a simple interpretation when ξ is a discrete design measure with support $\{x^{(1)}, \dots, x^{(k)}\}$. Suppose that Assumption A2 holds for every $\theta \in \Theta$. Then A6 is equivalent to the assumption that there exists a function Ψ , with continuous gradient, such that $h(\theta) = \Psi[\eta(\theta)]$, with $\eta(\theta) = (\eta(x^{(1)}, \theta), \dots, \eta(x^{(k)}, \theta))^\top$, and we obtain

$$\frac{\partial h(\theta)}{\partial \theta^\top} = \frac{\partial \Psi(t)}{\partial t^\top} \Big|_{t=\eta(\theta)} \frac{\partial \eta(\theta)}{\partial \theta^\top}.$$

A6 thus holds for every $\bar{\theta} \in \text{int}(\Theta)$ with $A_\theta = \partial \Psi(t)/\partial t^\top|_{t=\eta(\theta)}$.

It is useful to discuss A2 and A6 in the context of Example 2. There the limiting design is $\xi_* = \delta(x_*)$, the measure that puts mass one at x_* . Therefore, $\theta \stackrel{\xi_*}{\approx} \bar{\theta} \Leftrightarrow \theta_1 + x_*\theta_2 = \bar{\theta}_1 + x_*\bar{\theta}_2$. It follows that $\theta \stackrel{\xi_*}{\approx} \bar{\theta} \Rightarrow h(\theta) = h(\bar{\theta})$ only if $\bar{\theta}_1 + x_*\bar{\theta}_2 = 0$, and this is the only case where A2 holds. We have seen in Example 2 that regular asymptotic normality does not hold when $\bar{\theta}_1 + x_*\bar{\theta}_2 \neq 0$, hence the importance of A2. Consider now the derivative of $h(\theta)$. We have $\partial h(\theta)/\partial \theta = -1/(2\theta_2)[1 \quad -\theta_1/\theta_2]^\top$ and $\partial \eta(x_*, \theta)/\partial \theta = (x_* \quad x_*^2)^\top$. Therefore, even if $\bar{\theta}_1 + x_*\bar{\theta}_2 = 0$, we obtain $\theta \stackrel{\xi_*}{\approx} \bar{\theta} \Rightarrow \partial h(\theta)/\partial \theta = -1/(2\theta_2)[1 \quad x_*]^\top =$

$-1/(2x_*\theta_2)\partial\eta(x_*,\theta)/\partial\theta$, and A6 does not hold if $\theta \neq \bar{\theta}$. When m is fixed in Example 2, $\Theta^\# = \operatorname{argmin} J(\theta)$, with $J(\theta)$ given by (1.14), contains other points than $\bar{\theta}$; then A6 does not hold and there is no regular asymptotic normality for $h(\hat{\theta}^N)$. On the contrary, when $m \rightarrow \infty$ $\Theta^\# = \{\bar{\theta}\}$ and A6 holds trivially: this is the only situation in the example where regular asymptotic normality holds.

When ξ is a continuous design measure, an example where A6 holds is when

$$h(\theta) = \Psi[h_1(\theta), \dots, h_k(\theta)]$$

with Ψ a continuously differentiable function of k variables, and with

$$h_i(\theta) = \int_{\mathcal{X}} g_i[\eta(x, \theta), x] \xi(\mathrm{d}x), \quad i = 1, \dots, k,$$

for some functions $g_i(t, x)$ differentiable with respect to t for any x in the support of ξ . Then, supposing that we can interchange the order of derivatives and integrals, we obtain

$$\frac{\partial h(\theta)}{\partial \theta_i} = \sum_{j=1}^k \left[\frac{\partial \Psi(v)}{\partial v_j} \right]_{v_j=h_j(\theta)} \int_{\mathcal{X}} \left[\frac{\partial g_j(t, x)}{\partial t} \right]_{t=\eta(x, \theta)} \{\mathbf{f}_\theta\}_i(x) \xi(\mathrm{d}x),$$

and for any $\phi \in \mathcal{L}_2(\xi)$

$$A_\theta(\phi) = \sum_{j=1}^k \left[\frac{\partial \Psi(v)}{\partial v_j} \right]_{v_j=h_j(\theta)} \int_{\mathcal{X}} \left[\frac{\partial g_j(t, x)}{\partial t} \right]_{t=\eta(x, \theta)} \phi(x) \xi(\mathrm{d}x)$$

so that A6 holds when A2 is satisfied.

We can now formulate the main result of the paper, concerning regular asymptotic normality.

Theorem 3. *Let the sequence $(x_i)_i$ either converge to ξ in the sense of Definition 1, or be generated by ξ according to Definition 2. Suppose that the model response $\eta(x, \theta)$ satisfies assumptions A1, A3 and that the function of interest $h(\theta)$ satisfies A2, A6. Let $(\hat{\theta}^N(y^N))_N$ be a sequence of solutions of (1.15). Then, under A4, A5, or under B1, regular asymptotic normality holds: the sequence $\sqrt{N} \{h[\hat{\theta}^N(y^N)] - h(\bar{\theta})\}$ converges in distribution as $N \rightarrow \infty$ to a random variable distributed*

$$\mathcal{N} \left(0, \sigma^2 \left[\frac{\partial h(\theta)}{\partial \theta^\top} \mathbf{M}^-(\xi, \theta) \frac{\partial h(\theta)}{\partial \theta} \right]_{\theta=\bar{\theta}} \right)$$

where the choice of the g -inverse is arbitrary.

Proof. From the properties of P_θ we have for every $\theta \stackrel{\xi}{\sim} \bar{\theta}$

$$\frac{\partial h(\theta)}{\partial \theta_i} = A_\theta [P_\theta \{\mathbf{f}_\theta\}_i] = \mathbf{u}^\top(\theta) \{\mathbf{M}(\xi, \theta)\}_{.i}, \quad (1.17)$$

where

$$u_j(\theta) = A_\theta [\mathbf{f}_\theta^\top \{\mathbf{M}^+(\xi, \theta)\}_{.j}].$$

From Lemma 3, we have

$$\theta^* \stackrel{\xi}{\sim} \bar{\theta} \Rightarrow \mathbf{u}^\top(\theta^*) \mathbf{f}_{\theta^*} \stackrel{\xi}{\cong} \mathbf{u}^\top(\bar{\theta}) \mathbf{f}_{\bar{\theta}}. \quad (1.18)$$

Indeed, for any ϕ in $\mathcal{L}_2(\xi)$ we can write

$$\langle \mathbf{u}^\top(\theta^*) \mathbf{f}_{\theta^*}, \phi \rangle_\xi = \sum_{j=1}^p A_{\theta^*} [\mathbf{f}_{\theta^*}^\top \{\mathbf{M}^+(\xi, \theta^*)\}_{.j}] \int_{\mathcal{X}} \{\mathbf{f}_{\theta^*}\}_j(x) \phi(x) \xi(\mathrm{d}x).$$

From the linearity of A_{θ^*} this is equal to

$$\begin{aligned} A_{\theta^*} \left[\sum_{j=1}^p \mathbf{f}_{\theta^*}^\top \{\mathbf{M}^+(\xi, \theta^*)\}_{.j} \int_{\mathcal{X}} \{\mathbf{f}_{\theta^*}\}_j(x) \phi(x) \xi(\mathrm{d}x) \right] &= A_{\theta^*} [P_{\theta^*} \phi] \\ &= A_{\bar{\theta}} [P_{\bar{\theta}} \phi] = \langle \mathbf{u}^\top(\bar{\theta}) \mathbf{f}_{\bar{\theta}}, \phi \rangle_\xi \end{aligned}$$

so that $\mathbf{u}^\top(\theta^*) \mathbf{f}_{\theta^*} \stackrel{\xi}{\cong} \mathbf{u}^\top(\bar{\theta}) \mathbf{f}_{\bar{\theta}}$.

Let $(\hat{\theta}^N(y^N))_N$ be a sequence of solutions of (1.15) and $(\hat{\theta}^{N_t})_t$ be a subsequence converging to a limit point $\theta^\# = \theta^\#(y) \stackrel{\xi}{\sim} \bar{\theta}$, see Theorem 1. By the Taylor formula we have

$$\begin{aligned} 0 &= \frac{\partial J_{N_t}(\theta, y^{N_t})}{\partial \theta} \Big|_{\hat{\theta}^{N_t}} \\ &= \frac{\partial J_{N_t}(\theta, y^{N_t})}{\partial \theta} \Big|_{\theta^\#} + \left[\frac{\partial^2 J_{N_t}(\theta, y^{N_t})}{\partial \theta \partial \theta^\top} \right]_{\beta_t} (\hat{\theta}^{N_t} - \theta^\#), \end{aligned} \quad (1.19)$$

where $\beta_t = \beta_t(y)$ lies on the segment joining $\hat{\theta}^{N_t}$ with $\theta^\#(y)$. (Notice that $\lim_{t \rightarrow \infty} \beta_t(y) = \theta^\#(y)$ a.s.) Now,

$$\begin{aligned} -\sqrt{N_t} \mathbf{u}^\top(\theta^\#) \left[\frac{\partial J_{N_t}(\theta, y^{N_t})}{\partial \theta} \right]_{\theta^\#} &= \\ \frac{2}{\sqrt{N_t}} \sum_{i=1}^{N_t} \mathbf{u}^\top(\theta^\#) \left[\frac{\partial \eta(x_i, \theta)}{\partial \theta} \right]_{\theta^\#} [y(x_i) - \eta(x_i, \theta^\#)]. \end{aligned} \quad (1.20)$$

Consider first the case where $(x_i)_i$ converges strongly to a discrete design in the sense of Definition 1. We decompose the sum on the right-hand side of (1.20) into two sums: the first one corresponds to indices i such that $x_i \notin S_\xi$, the support of ξ , the second one is for $x_i \in S_\xi$. The first sum then tends to zero

in probability. For the second, we use (1.18) and the fact that $\eta[x_i, \theta^\#(y)] = \eta(x_i, \bar{\theta})$ for $x_i \in S_\xi$ to obtain

$$\frac{2}{\sqrt{N_t}} \sum_{i=1, x_i \in S_\xi}^{N_t} \mathbf{u}^\top(\bar{\theta}) \mathbf{f}_{\bar{\theta}}(x_i) \varepsilon_i,$$

which, by the central limit theorem, converges in distribution to $\mathcal{N}(0, 4D)$ with

$$D = \sigma^2 \mathbf{u}^\top(\bar{\theta}) \mathbf{M}(\xi, \bar{\theta}) \mathbf{u}(\bar{\theta}). \quad (1.21)$$

Consider alternatively the case of a randomized design in the sense of Definition 2. For almost every sequence of errors $\varepsilon_1, \varepsilon_2 \dots$ we have

$$\xi(\{x : \eta(x, \bar{\theta}) = \eta[x, \theta^\#(y)]\}) = 1,$$

see Theorem 1, and

$$\xi(\{x : \mathbf{u}^\top[\theta^\#(y)] \mathbf{f}_{\theta^\#(y)}(x) = \mathbf{u}^\top(\bar{\theta}) \mathbf{f}_{\bar{\theta}}(x)\}) = 1,$$

see (1.18). Hence (1.20) implies that for each t

$$-\sqrt{N_t} \mathbf{u}^\top(\theta^\#) \left[\frac{\partial J_{N_t}(\theta, y^{N_t})}{\partial \theta} \right]_{\theta^\#} = \frac{2}{\sqrt{N_t}} \sum_{i=1}^{N_t} \mathbf{u}^\top(\bar{\theta}) \mathbf{f}_{\bar{\theta}}(x_i) \varepsilon_i, \quad \xi\text{-a.s.}$$

From the independence of x_i and ε_i , $\mathbb{E}\{\varepsilon_i\} = 0$ and $\text{Var}\{\varepsilon_i\} = \sigma^2$ we obtain, again by the central limit theorem, that the last sum converges in distribution to a variable distributed $\mathcal{N}(0, 4D)$.

Therefore, for both types of designs we have

$$-\sqrt{N_t} \mathbf{u}^\top(\theta^\#) \left[\frac{\partial J_{N_t}(\theta, y^{N_t})}{\partial \theta} \right]_{\theta^\#} \xrightarrow{d} \nu \sim \mathcal{N}(0, 4D), \quad t \rightarrow \infty, \quad (1.22)$$

with D given by (1.21).

Moreover, Lemmas 1 and 2 imply that $\partial^2 J_{N_t}(\theta, y^{N_t}) / \partial \theta \partial \theta^\top$ converges a.s. and uniformly in θ to

$$-2 \int_{\mathcal{X}} [\eta(x, \bar{\theta}) - \eta(x, \theta)] \frac{\partial^2 \eta(x, \theta)}{\partial \theta \partial \theta^\top} \xi(dx) + 2\mathbf{M}(\xi, \theta)$$

as $t \rightarrow \infty$. Therefore, for both types of designs we have

$$\begin{aligned} \mathbf{u}^\top(\theta^\#) \left[\frac{\partial^2 J_{N_t}(\theta, y^{N_t})}{\partial \theta \partial \theta^\top} \right]_{\beta_t(y)} &\xrightarrow{\text{a.s.}} 2\mathbf{u}^\top(\theta^\#) \mathbf{M}(\xi, \theta^\#) \\ &= 2 \left[\frac{\partial h(\theta)}{\partial \theta^\top} \right]_{\theta^\#}, \quad t \rightarrow \infty. \end{aligned}$$

Hence from (1.19) and (1.22) we obtain

$$2\sqrt{N_t} \left[\frac{\partial h(\theta)}{\partial \theta^\top} \right]_{\theta^\#} (\hat{\theta}^{N_t} - \theta^\#) \xrightarrow{d} \nu \sim \mathcal{N}(0, 4D), \quad t \rightarrow \infty. \quad (1.23)$$

Applying the Taylor formula again, we obtain

$$\begin{aligned} \sqrt{N_t} [h(\hat{\theta}^{N_t}) - h(\bar{\theta})] &= \sqrt{N_t} [h(\hat{\theta}^{N_t}) - h(\theta^\#)] \\ &= \sqrt{N_t} \left[\frac{\partial h(\theta)}{\partial \theta^\top} \right]_{\delta_t(y)} (\hat{\theta}^{N_t} - \theta^\#), \quad t \rightarrow \infty, \end{aligned}$$

where $\delta_t(y)$ is on the segment connecting $\hat{\theta}^{N_t}$ with $\theta^\#$, hence $\delta_t(y) \xrightarrow{\text{a.s.}} \theta^\#$ and $[\partial h(\theta)/\partial \theta^\top]_{\delta_t(y)} \xrightarrow{\text{a.s.}} [\partial h(\theta)/\partial \theta^\top]_{\theta^\#}$, $t \rightarrow \infty$. Finally, we obtain from (1.23)

$$\sqrt{N_t} [h(\hat{\theta}^{N_t}) - h(\bar{\theta})] \xrightarrow{d} \frac{\nu}{2} \sim \mathcal{N}(0, D), \quad t \rightarrow \infty,$$

and, according to (1.17), $[\partial h(\theta)/\partial \theta]_{\bar{\theta}}$ is in the range of $\mathbf{M}(\xi, \bar{\theta})$, which gives

$$D = \sigma^2 \mathbf{u}^\top(\bar{\theta}) \mathbf{M}(\xi, \bar{\theta}) \mathbf{u}(\bar{\theta}) = \sigma^2 \left[\frac{\partial h(\theta)}{\partial \theta^\top} \mathbf{M}^-(\xi, \theta) \frac{\partial h(\theta)}{\partial \theta} \right]_{\bar{\theta}}$$

for any choice of the g -inverse $\mathbf{M}^-(\xi, \bar{\theta})$. ■

1.6 Estimation of a multidimensional function $H(\theta)$

Let $H(\theta) = [h_1(\theta), \dots, h_q(\theta)]^\top$ be a q -dimensional function defined on Θ . We shall assume the following.

ASSUMPTION A2*: *The functions $h_i(\theta)$ are continuous in Θ , $i = 1, \dots, q$, and such that*

$$\theta \xrightarrow{\xi} \bar{\theta} \Rightarrow H(\theta) = H(\bar{\theta}).$$

We then have the following straightforward extension of Theorem 2.

Theorem 4. *Under the assumptions of Theorem 2, but with A2* replacing A2, we have*

$$\lim_{N \rightarrow \infty} H[\hat{\theta}^N(y^N)] = H(\bar{\theta}) \quad \text{a.s.}$$

for $(\hat{\theta}^N(y^N))_N$ any sequence defined by (1.15).

Consider now the following assumption; its substitution for A6 in Theorem 3 gives Theorem 5 below.

ASSUMPTION A6*: *The vector function $H(\theta)$ has a continuous Jacobian $\partial H(\theta)/\partial \theta^\top$ on Θ_0 . Moreover, for each $\theta \xrightarrow{\xi} \bar{\theta}$ there exists a continuous linear mapping B_θ from $L_2(\xi)$ to \mathbb{R}^q such that $B_\theta = B_{\bar{\theta}}$ and that*

$$\frac{\partial H(\theta)}{\partial \theta_i} = B_\theta [\{\mathbf{f}_\theta\}_i], \quad i = 1, \dots, p,$$

where $\{\mathbf{f}_\theta\}_i$ is given by (1.16).

Theorem 5. Under the assumptions of Theorem 3, but with A2* and A6* replacing A2 and A6, for $(\hat{\theta}^N(y^N))_N$ a sequence of solutions of (1.15), $\sqrt{N} \left\{ H[\hat{\theta}^N(y^N)] - H(\bar{\theta}) \right\}$ converges in distribution as $N \rightarrow \infty$ to a random vector distributed normally

$$\mathcal{N} \left(0, \sigma^2 \left[\frac{\partial H(\theta)}{\partial \theta^\top} \mathbf{M}^-(\xi, \theta) \frac{\partial H^\top(\theta)}{\partial \theta} \right]_{\bar{\theta}} \right)$$

where the choice of the g-inverse is arbitrary.

Proof. Take any $\mathbf{c} \in \mathbb{R}^q$, and define $h_c(\theta) = \mathbf{c}^\top H(\theta)$. Evidently $h_c(\theta)$ satisfies the assumptions of Theorem 3, and $\sqrt{N} \left\{ h_c[\hat{\theta}^N(y^N)] - h_c(\bar{\theta}) \right\}$ converges in distribution as $N \rightarrow \infty$ to a random variable distributed

$$\mathcal{N} \left(0, \sigma^2 \mathbf{c}^\top \left[\frac{\partial H(\theta)}{\partial \theta^\top} \mathbf{M}^-(\xi, \theta) \frac{\partial H^\top(\theta)}{\partial \theta} \right]_{\bar{\theta}} \mathbf{c} \right).$$

■

Appendix. Proofs of Lemmas 1-3

Proof of Lemma 1. We can write

$$\begin{aligned} & \left| \frac{1}{N} \sum_{k=1}^N a(x_k, \theta) \alpha_k - \mathbb{E}\{\alpha_1\} \sum_{x \in S_\xi} a(x, \theta) \xi(\{x\}) \right| \\ & \leq \left| \frac{1}{N} \sum_{k=1, x_k \notin S_\xi}^N a(x_k, \theta) \alpha_k \right| \\ & \quad + \sum_{x \in S_\xi} \sup_{\theta \in \Theta} |a(x, \theta)| \left| \frac{N(x)}{N} \left(\frac{1}{N(x)} \sum_{k=1, x_k=x}^N \alpha_k \right) - \mathbb{E}\{\alpha_1\} \xi(\{x\}) \right|, \end{aligned}$$

where $N(x)/N$ is the relative frequency of the point x in the sequence x_1, x_2, \dots, x_N . The last sum for $x \in S_\xi$ tends to zero a.s. and uniformly on Θ , since $N(x)/N$ tends to $\xi(\{x\})$, and $[1/N(x)] \sum_{k=1, x_k=x}^N \alpha_k$ converges a.s. to $\mathbb{E}\{\alpha_1\}$. The first sum on the right-hand side is bounded by

$$\sup_{x \in \mathcal{X}, \theta \in \Theta} |a(x, \theta)| \frac{N(\mathcal{X} \setminus S_\xi)}{N} \frac{1}{N(\mathcal{X} \setminus S_\xi)} \sum_{k=1, x_k \in \mathcal{X} \setminus S_\xi} \alpha_k.$$

This expression tends a.s. to zero, since $N(\mathcal{X} \setminus S_\xi)/N$ tends to zero, and the law of large numbers applies for the remaining part in case $N(\mathcal{X} \setminus S_\xi) \rightarrow \infty$. \blacksquare

Proof of Lemma 2. We use a construction similar to that in (Bierens, 1994, p. 43). Take some fixed $\theta^1 \in \Theta$ and consider the set

$$\mathcal{B}(\theta^1, \delta) = \{\theta \in \Theta : \|\theta - \theta^1\| \leq \delta\}.$$

Define $\bar{a}_\delta(z)$ and $\underline{a}_\delta(z)$ as the maximum and the minimum of $a(z, \theta)$ over the set $\mathcal{B}(\theta^1, \delta)$.

The expectations $\mathbb{E}\{\underline{a}_\delta(z)\}$ and $\mathbb{E}\{\bar{a}_\delta(z)\}$ are bounded by

$$\mathbb{E}\{\max_{\theta \in \Theta} |a(z, \theta)|\} < \infty.$$

Also, $\bar{a}_\delta(z) - \underline{a}_\delta(z)$ is an increasing function of δ . Hence, we can interchange the order of the limit and expectation in the following expression

$$\lim_{\delta \searrow 0} [\mathbb{E}\{\bar{a}_\delta(z)\} - \mathbb{E}\{\underline{a}_\delta(z)\}] = \mathbb{E}\left\{\lim_{\delta \searrow 0} [\bar{a}_\delta(z) - \underline{a}_\delta(z)]\right\} = 0,$$

which proves the continuity of $\mathbb{E}\{a(z, \theta)\}$ at θ^1 and implies

$$\forall \beta > 0, \exists \delta(\beta) > 0 \text{ such that } \left| \mathbb{E}\{\bar{a}_{\delta(\beta)}(z)\} - \mathbb{E}\{\underline{a}_{\delta(\beta)}(z)\} \right| < \frac{\beta}{2}.$$

Hence we can write for every $\theta \in \mathcal{B}(\theta^1, \delta(\beta))$

$$\begin{aligned} \frac{1}{N} \sum_k \underline{a}_{\delta(\beta)}(z_k) - \mathbb{E}\{\underline{a}_{\delta(\beta)}(z)\} - \frac{\beta}{2} &\leq \frac{1}{N} \sum_k \underline{a}_{\delta(\beta)}(z_k) - \mathbb{E}\{\bar{a}_{\delta(\beta)}(z)\} \\ &\leq \frac{1}{N} \sum_k a(z_k, \theta) - \mathbb{E}\{a(z, \theta)\} \\ &\leq \frac{1}{N} \sum_k \bar{a}_{\delta(\beta)}(z_k) - \mathbb{E}\{\underline{a}_{\delta(\beta)}(z)\} \\ &\leq \frac{1}{N} \sum_k \bar{a}_{\delta(\beta)}(z_k) - \mathbb{E}\{\bar{a}_{\delta(\beta)}(z)\} + \frac{\beta}{2}. \end{aligned}$$

From the strong law of large numbers, we have that $\forall \gamma > 0, \exists N_1(\beta, \gamma)$ such that

$$\begin{aligned} \text{Prob} \left\{ \forall N > N_1(\beta, \gamma), \left| \frac{1}{N} \sum_k \bar{a}_{\delta(\beta)}(z_k) - \mathbb{E}\{\bar{a}_{\delta(\beta)}(z)\} \right| < \frac{\beta}{2} \right\} &> 1 - \frac{\gamma}{2}, \\ \text{Prob} \left\{ \forall N > N_1(\beta, \gamma), \left| \frac{1}{N} \sum_k \underline{a}_{\delta(\beta)}(z_k) - \mathbb{E}\{\underline{a}_{\delta(\beta)}(z)\} \right| < \frac{\beta}{2} \right\} &> 1 - \frac{\gamma}{2}. \end{aligned}$$

Combining with previous inequalities, we obtain

$$\text{Prob} \left\{ \forall N > N_1(\beta, \gamma), \max_{\theta \in \mathcal{B}(\theta^i, \delta(\beta))} \left| \frac{1}{N} \sum_k a(z_k, \theta) - \mathbb{E}\{a(z, \theta)\} \right| < \beta \right\} > 1 - \gamma.$$

It only remains to cover Θ with a finite numbers of sets $\mathcal{B}(\theta^i, \delta(\beta))$, $i = 1, \dots, n(\beta)$, which is always possible from the compactness assumption. For any $\alpha > 0, \beta > 0$, take $\gamma = \alpha/n(\beta)$, $N(\beta) = \max_i N_i(\beta, \gamma)$. We obtain

$$\text{Prob} \left\{ \forall N > N(\beta), \max_{\theta \in \Theta} \left| \frac{1}{N} \sum_k a(z_k, \theta) - \mathbb{E}\{a(z, \theta)\} \right| < \beta \right\} > 1 - \alpha,$$

which completes the proof. \blacksquare

Proof of Lemma 3. Since P_θ is the orthogonal projector onto \mathcal{L}_θ it is sufficient to prove that $\bar{\alpha} \stackrel{\xi}{\sim} \bar{\theta}$ implies that any element of $\mathcal{L}_{\bar{\alpha}}$ is in $\mathcal{L}_{\bar{\theta}}$.

From $\{\mathbf{f}_{\bar{\theta}}\}_1, \dots, \{\mathbf{f}_{\bar{\theta}}\}_p$ we choose r functions that form a linear basis of $\mathcal{L}_{\bar{\theta}}$. Without any loss of generality we can suppose that they are the first r ones. Decompose θ into $\theta = (\beta, \gamma)$, where β corresponds to the first r components of θ and γ to the $p - r$ remaining ones. Define similarly $\bar{\theta} = (\bar{\beta}, \bar{\gamma})$. From A4, the components of $\partial\eta[x, (\beta, \gamma)]/\partial\gamma$ are linear combinations of components of $\partial\eta[x, (\beta, \gamma)]/\partial\beta$ not only for $\theta = \bar{\theta}$ but also for θ in some neighborhood of $\bar{\theta}$.

Define the following mapping G from \mathbb{R}^{r+p} to \mathbb{R}^r by

$$G(\beta, \alpha) = \int_{\mathcal{X}} \frac{\partial\eta[x, (\beta, \bar{\gamma})]}{\partial\beta} \{\eta[x, (\beta, \bar{\gamma})] - \eta(x, \alpha)\} \xi(dx).$$

From $\bar{\alpha} \stackrel{\xi}{\sim} \bar{\theta}$ we obtain $G(\bar{\beta}, \bar{\alpha}) = 0$. The matrix

$$\frac{\partial G(\beta, \alpha)}{\partial\beta^\top} \Big|_{\bar{\beta}, \bar{\alpha}} = \int_{\mathcal{X}} \frac{\partial\eta[x, (\beta, \bar{\gamma})]}{\partial\beta} \frac{\partial\eta[x, (\beta, \bar{\gamma})]}{\partial\beta^\top} \Big|_{\bar{\beta}} \xi(dx)$$

is a nonsingular $r \times r$ submatrix of $\mathbf{M}(\xi, \bar{\theta})$, with $\text{rank}[\mathbf{M}(\xi, \theta)] = r$ for θ in a neighborhood of $\bar{\theta}$. From the Implicit Function Theorem, see Spivak (1965, Th. 2-12, p. 41), there exist neighborhoods $\mathcal{V}(\bar{\alpha})$, $\mathcal{W}(\bar{\beta})$ and a differentiable mapping $\psi : \mathcal{V}(\bar{\alpha}) \rightarrow \mathcal{W}(\bar{\beta})$ such that $\psi(\bar{\alpha}) = \bar{\beta}$ and that $\alpha \in \mathcal{V}(\bar{\alpha})$ implies $G[\psi(\alpha), \alpha] = 0$. It follows that

$$\begin{aligned} & \frac{\partial}{\partial\beta} \int_{\mathcal{X}} \{\eta[x, (\beta, \bar{\gamma})] - \eta(x, \alpha)\}^2_{\beta=\psi(\alpha)} \xi(dx) \\ &= 2 \int_{\mathcal{X}} \left[\frac{\partial\eta[x, (\beta, \bar{\gamma})]}{\partial\beta} \right]_{\beta=\psi(\alpha)} \{\eta[x, (\psi(\alpha), \bar{\gamma})] - \eta(x, \alpha)\} \xi(dx) = 0. \end{aligned} \quad (1.24)$$

Since the components of $\partial\eta[x, (\beta, \gamma)]/\partial\gamma$ are linear combinations of the components of $\partial\eta[x, (\beta, \gamma)]/\partial\beta$ for any $\theta = (\beta, \gamma)$ in some neighborhood of $\bar{\theta}$, we obtain from (1.24)

$$\begin{aligned} & \frac{\partial}{\partial \gamma} \int_{\mathcal{X}} \{\eta[x, (\beta, \gamma)] - \eta(x, \alpha)\}_{\beta=\psi(\alpha), \gamma=\bar{\gamma}}^2 \xi(dx) = \\ & 2 \int_{\mathcal{X}} \left[\frac{\partial \eta[x, (\beta, \gamma)]}{\partial \gamma} \right]_{\beta=\psi(\alpha), \gamma=\bar{\gamma}} \{\eta[x, (\psi(\alpha), \bar{\gamma})] - \eta(x, \alpha)\} \xi(dx) = 0. \end{aligned}$$

Combining with (1.24) we obtain that

$$\left\{ \frac{\partial}{\partial \theta} \int_{\mathcal{X}} [\eta(x, \theta) - \eta(x, \alpha)]^2 \xi(dx) \right\}_{\theta=[\psi(\alpha), \bar{\gamma}]} = 0$$

for all α belonging to some neighborhood $\mathcal{U}(\bar{\alpha})$. We can make $\mathcal{U}(\bar{\alpha})$ small enough to satisfy the inequality $\|\eta[x, (\psi(\alpha), \bar{\gamma})] - \eta(x, \bar{\theta})\|_{\xi}^2 < \epsilon$ required in A5. It follows that $(\psi(\alpha), \bar{\gamma}) \stackrel{\xi}{\approx} \alpha$, that is, $\eta(\cdot, \alpha) \stackrel{\xi}{\approx} \eta[\cdot, (\psi(\alpha), \bar{\gamma})]$ for all α in a neighborhood of $\bar{\alpha}$. By taking derivatives we then obtain

$$\begin{aligned} \left[\frac{\partial \eta(\cdot, \alpha)}{\partial \alpha^{\top}} \right]_{\bar{\alpha}} & \stackrel{\xi}{\approx} \left[\frac{\partial \eta[\cdot, (\psi(\alpha), \bar{\gamma})]}{\partial \alpha^{\top}} \right]_{\bar{\alpha}} \\ & \stackrel{\xi}{\approx} \left[\frac{\partial \eta[\cdot, (\beta, \bar{\gamma})]}{\partial \beta^{\top}} \right]_{(\psi(\bar{\alpha}), \bar{\gamma})} \left[\frac{\partial \psi(\alpha)}{\partial \alpha^{\top}} \right]_{\bar{\alpha}}, \end{aligned}$$

that is, $\mathcal{L}_{\bar{\alpha}} \stackrel{\xi}{\subset} \mathcal{L}_{(\psi(\bar{\alpha}), \bar{\gamma})} = \mathcal{L}_{\bar{\theta}}$.

By interchanging $\bar{\alpha}$ with $\bar{\theta}$ we obtain $\mathcal{L}_{\bar{\theta}} \stackrel{\xi}{\subset} \mathcal{L}_{\bar{\alpha}}$. ■

Acknowledgments

The research of the first author has been supported by the VEGA grant No. 1/3016/06. The work of the second author was partially supported by the IST Programme of the European Community, under the PASCAL Network of Excellence, IST-2002-506778. This publication only reflects the authors's view.

References

- Atkinson, A. and Donev, A. (1992). *Optimum Experimental Design*. Oxford University Press.
- Bierens, H. (1994). *Topics in Advanced Econometrics*. Cambridge University Press, Cambridge.
- Billingsley, P. (1971). *Weak Convergence of Measures: Applications in Probability*. SIAM, Philadelphia.
- Elfving, G. (1952). Optimum allocation in linear regression. *The Annals of Math. Stat.*, **23**, 255–262.
- Fedorov, V. (1972). *Theory of Optimal Experiments*. Academic Press, New York.
- Gallant, A. (1987). *Nonlinear Statistical Models*. Wiley, New York.

- Hero, A., Fessler, J., and Usman, M. (1996). Exploring estimator bias-variance tradeoffs using the uniform CR bound. *IEEE Transactions on Signal Processing*, **44**, 2026–2041.
- Ivanov, A. (1997). *Asymptotic Theory of Nonlinear Regression*. Kluwer, Dordrecht.
- Jennrich, R. (1969). Asymptotic properties of nonlinear least squares estimation. *The Annals of Math. Stat.*, **40**, 633–643.
- Kiefer, J. and Wolfowitz, J. (1959). Optimum designs in regression problems. *The Annals of Math. Stat.*, **30**, 271–294.
- Lehmann, E. and Casella, G. (1998). *Theory of Point Estimation*. Springer, Heidelberg.
- Pázman, A. (1980). Singular experimental designs. *Math. Operationsforsch. Statist., Ser. Statistics*, **16**, 137–149.
- Pázman, A. (1986). *Foundations of Optimum Experimental Design*. Reidel (Kluwer group), Dordrecht (co-pub. VEDA, Bratislava).
- Pázman, A. and Pronzato, L. (1992). Nonlinear experimental design based on the distribution of estimators. *Journal of Statistical Planning and Inference*, **33**, 385–402.
- Pázman, A. and Pronzato, L. (2006). On the irregular behavior of LS estimators for asymptotically singular designs. *Statistics & Probability Letters*, **76**, 1089–1096.
- Pronzato, L. and Pázman, A. (1994). Second-order approximation of the entropy in nonlinear least-squares estimation. *Kybernetika*, **30**(2), 187–198. *Erratum* 32(1):104, 1996.
- Shiryayev, A. (1996). *Probability*. Springer, Berlin.
- Silvey, S. (1980). *Optimal Design*. Chapman & Hall, London.
- Sjöberg, J., Zhang, Q., Ljung, L., Benveniste, A., Delyon, B., Glorennec, P.-Y., Hjalmarsson, H., and Juditsky, A. (1995). Nonlinear black-box modeling in system identification: a unified overview. *Automatica*, **31**(12), 1691–1724.
- Spivak, M. (1965). *Calculus on Manifolds. A Modern Approach to Classical Theorems of Advanced Calculus*. W. A. Benjamin, Inc., New York.
- Stoica, P. (2001). Parameter estimation problems with singular information matrices. *IEEE Transactions on Signal Processing*, **49**, 87–90.
- Wu, C.-F. (1980). Characterizing the consistent directions of least squares estimates. *The Annals of Statistics*, **8**(4), 789–801.
- Wu, C.-F. (1981). Asymptotic theory of nonlinear least squares estimation. *The Annals of Statistics*, **9**(3), 501–513.
- Wu, C.-F. (1983). Further results on the consistent directions of least squares estimators. *The Annals of Statistics*, **11**(4), 1257–1262.
- Wynn, H. (1972). Results in the theory and construction of D -optimum experimental designs. *Journal of the Royal Statistical Society B*, **34**, 133–147.