# One-step ahead adaptive D-optimal design on a finite design space is asymptotically optimal

Luc Pronzato

# One-step ahead adaptive $D$-optimal design on a finite design space is asymptotically optimal

Luc Pronzato

Laboratoire I3S, CNRS/Université de Nice-Sophia Antipolis

Bât Euclide, Les Algorithmes

2000 route des lucioles, BP 121

06903 Sophia Antipolis cedex, France

pronzato@i3s.unice.fr

June 19, 2009

**Abstract** We study the consistency of parameter estimators in adaptive designs generated by a one-step ahead $D$-optimal algorithm. We show that when the design space is finite, under mild conditions the least-squares estimator in a nonlinear regression model is strongly consistent and the information matrix evaluated at the current estimated value of the parameters strongly converges to the $D$-optimal matrix for the unknown true value of the parameters. A similar property is shown to hold for maximum-likelihood estimation in Bernoulli trials (dose-response experiments). Some examples are presented.

**Key words:** adaptive design, consistency, $D$-optimal design, sequential design

# 1 Introduction: motivation and problem statement

We consider experimental design for a parametric model for which $N$ independent observations $Y_1, \ldots, Y_N$ yield the information matrix

$$\mathrm{M}(X_1^N, \theta) = \mathrm{M}(x_1, \ldots, x_N, \theta) = \sum_{i=1}^{N} \mu(x_i, \theta)\,,$$

where $x_i \in \mathcal{X} \subset \mathbb{R}^d$ is the $i$-th design point, characterizing the experimental conditions for the $i$-th observation, and $\theta$ is the $p$-dimensional vector of model parameters to be estimated. Two situations will be considered in more detail, namely nonlinear regression and Bernoulli trials. When $r_N(x_i)$ denotes the number of (repetitions of) observations made at $x = x_i$, the normalized information matrix per observation can be written as $\mathrm{M}(\xi_N, \theta) = (1/N)\,\mathrm{M}(X_1^N, \theta) = \sum_{i=1}^{K} [r_N(x_i)/N]\,\mu(x_i, \theta)$, where $K$ is the number of distinct design points and $\xi_N$ is the design measure (a probability measure on $\mathcal{X}$) that puts mass $r_N(x_i)/N$ at $x_i$. Following the usual approximate-design approach, we shall relax the constraints on design measures by considering $\xi$ as any element of $\Xi$, the set of probability measures on $\mathcal{X}$, and write $\mathrm{M}(\xi, \theta) = \int_{\mathcal{X}} \mu(x, \theta)\,\xi(dx)$. We shall denote $\lambda_{\min}(\mathrm{M})$ and $\lambda_{\max}(\mathrm{M})$ the minimum and maximum eigenvalues of the matrix M.

Local $D$-optimal design consists of determining a measure $\xi_D^*$ that maximizes $\log\det \mathrm{M}(\xi, \theta)$ for a given value of $\theta$. The denomination 'local' comes from the fact that in nonlinear situations $\mathrm{M}(\xi, \theta)$ depends on $\theta$, and the optimal $\xi_D^*$ for estimating $\theta$ thus depends on the value $\theta$ to be estimated. Minimax-optimal and average-optimal (also called bayesian-optimal) designs can be used to avoid the dependency of $\xi_D^*$ in $\theta$. However, in practice these approaches only replace the choice of a prior nominal value (for local design) by that of a prior admissible set (minimax design) or a prior distribution for $\theta$ (bayesian design), see, e.g., Melas (1978); Fedorov (1980); Pronzato and Walter (1985, 1988); Chaloner and Larntz (1989). See also Pázman and Pronzato (2007) for an approach based on quantile and probability-level criteria. Another rather common and intuitively appealing approach consists of making the design adaptive. The design points $x_1, x_2 \ldots, x_k, x_{k+1}, \ldots$ associated with a sequence of observations are then chosen sequentially, the determination of the point $x_{k+1}$ being based on the value $\hat{\theta}^k$ estimated from the $k$

previous observations (by least-squares, maximum likelihood or a bayesian method). The motivation is that alternating estimations based on previous observations with determinations of the next design points where to observe may hopefully force the empirical design measure to progressively adapt to the correct (true) value of the model parameters. It is the purpose of this paper to show that under suitable conditions, in particular when the $x_k$'s are restricted to belong to a finite set, the estimator $\hat{\theta}^k$ is strongly consistent and the corresponding adaptive design is asymptotically optimal. Although the condition that the design space $\mathcal{X}$ is finite is often imposed by practical considerations, it can be perceived here as a restriction compared to results in the literature obtained under more general conditions. It is important to notice, however, that those results are based on different types of designs or estimators. Hu (1998) considers Bayesian estimation by posterior mean; Lai (1994) and Chaudhuri and Mykland (1995) require the introduction of a subsequence of non-adaptive design points to ensure consistency of the estimator, see Example 2 of Sect. 5; Chaudhuri and Mykland (1993) require that the size of the initial experiment (non-adaptive) should be allowed to grow with the increase in size of the total experiment. No such conditions are required here and the design is fully adaptive.

We only consider the case of adaptive $D$-optimal design, see Sect. 2, but the results can presumably be extended to design approaches based on other global criteria (such that the information matrix has full rank at the optimum). Sect. 3 concerns adaptive design for least-squares estimation in nonlinear regression, whereas Sect. 4 is about design for maximum-likelihood estimation in Bernoulli-trial experiments. Sect. 5 provides a few illustrative examples. The proofs of lemmas and theorems are collected in an appendix.

## 2 One-step ahead adaptive $D$-optimal design

Consider the criterion $\phi(\xi) = \log\det \mathrm{M}(\xi,\theta)$. For any $\xi \in \Xi$ such that $\mathrm{M}(\xi,\theta)$ is non-singular, the directional derivative $F_\phi(\xi,\nu) = \lim_{\alpha\to 0^+}\{\phi[(1-\alpha)\xi+\alpha\nu]-\phi(\xi)\}/\alpha$ is given by $F_\phi(\xi,\nu) = \mathrm{tr}[\mathrm{M}(\nu,\theta)\mathrm{M}^{-1}(\xi,\theta)]-p$, with $p = \dim(\theta)$. The measure $\nu^* \in \Xi$ that maximizes $F_\phi(\xi,\nu)$ is then $\nu^* = \delta_{x^*}$, with $x^* = \arg\max_{x\in\mathcal{X}} \mathrm{tr}[\mu(x,\theta)\mathrm{M}^{-1}(\xi,\theta)]$ and $\delta_z$ the delta measure that puts mass 1 at $z$. Moreover, the celebrated

Kiefer and Wolfowitz (1960) equivalence theorem states that $\xi_D^* \in \Xi$ is locally $D$-optimal (it maximizes $\log\det \mathrm{M}(\xi,\theta)$) if and only if $F_\phi(\xi_D^*, \delta_{x^*}) = 0$, that is, $\max_{x\in\mathcal{X}} \mathrm{tr}[\mu(x,\theta)\mathrm{M}^{-1}(\xi_D^*,\theta)] = p$. Based on these considerations, the $k$-th iteration of a vertex-direction algorithm for local $D$-optimal design transforms the current design measure $\xi_k$ into $\xi_{k+1} = (1-\alpha_k)\xi_k + \alpha_k \delta_{x_{k+1}}$, with $x_{k+1} = \arg\max_{x\in\mathcal{X}} \mathrm{tr}[\mu(x,\theta)\mathrm{M}^{-1}(\xi_k,\theta)]$ the support point that gives the steepest-ascent direction and $\alpha_k$ some suitably chosen step-length, see, e.g., Fedorov (1972); Atkinson and Donev (1992).

The idea is similar in adaptive design and, after $k$ observations, one-step ahead adaptive $D$-optimal design chooses the next design point $x_{k+1}$ as

$$x_{k+1} = \arg\max_{x\in\mathcal{X}} \mathrm{tr}[\mu(x,\hat\theta^k)\mathrm{M}^{-1}(\xi_k,\hat\theta^k)]\,, \tag{1}$$

where $\hat\theta^k \in \mathbb{R}^p$ is the current estimated value for $\theta$, based on $x_1, Y_1, \ldots, x_k, Y_k$, and $\xi_k = (1/k)\sum_{i=1}^k \delta_{x_i}$ is the current empirical design measure. We leave aside initialisation issues and throughout the paper we assume that the first $p$ design points $x_1, \ldots, x_p$ are such that $\mathrm{M}(\xi_p,\theta)$ is non-singular for any $\theta$ (and $\mathrm{M}(\xi_k,\theta)$ is thus non-singular for any $k \geq p$).

**Remark 1** *Note that (1) can only be considered as an algorithm for choosing design points, in the sense that* $\mathrm{M}(\xi_k,\theta)$ *is not the information matrix for parameters* $\theta$ *due to the sequential construction of the design. It is common, however, to still use* $\mathrm{M}(\xi_k,\hat\theta^k)$ *as a characterization of the precision of the estimation in a sequential context, see Ford and Silvey (1980); Ford et al. (1985); Wu (1985) for a justification. The difficulty disappears in a bayesian context where* $\mathrm{M}(\xi_k,\hat\theta^k)$ *is used in an approximation of the posterior covariance matrix of the parameters, see, e.g., Chaloner and Verdinelli (1995). From the same repeated-sampling principle as that used by Wu (1985), the characterization of the precision of the estimation through* $\mathrm{M}(\xi_k,\hat\theta^k)$ *is also justified asymptotically* ($k \to \infty$) *when the admissible set* $\mathcal{X}$ *for the* $x_k$*'s is finite, see Sect. 3. See also Rosenberger et al. (1997) for maximum-likelihood estimation in a more general context.*

When $\hat\theta^k$ is frozen to a fixed value $\theta$, the iteration (1) corresponds to one step of a steepest-ascent vertex-direction algorithm, with step-length $1/(k+1)$ at step $k$ since $\mathrm{M}(\xi_{k+1},\theta) = [1-1/(k+1)]\mathrm{M}(\xi_k,\theta) +$

$[1/(k+1)]\mathrm{M}(\delta_{x_{k+1}}, \theta)$. Convergence to an optimal design measure is proved in Wynn (1970). The fact that $\hat{\theta}^k$ is estimated in adaptive design makes the proof of convergence a much more complicated issue for which few results are available, see e.g. Ford and Silvey (1980); Wu (1985); Müller and Pötscher (1992) for least-squares estimation and Hu (1998) for bayesian estimation. The idea that the almost sure convergence of $\hat{\theta}^k$ to some $\hat{\theta}^\infty$ would imply the convergence of $\xi_k$ to a $D$-optimal design measure for $\hat{\theta}^\infty$ is rather well admitted (it follows from Lemma 3 given in Sect. 3). Conversely, the convergence of $\xi_k$ to a design $\xi_\infty$ such that $\mathrm{M}(\xi_\infty, \theta)$ is non-singular for any $\theta$ would be enough in general to make an estimator consistent. It is clearly the interplay between estimation and design iterations that generates difficulties. As shown below, those difficulties disappear when $\mathcal{X}$ is a finite set. Notice that the assumption that $\mathcal{X}$ is finite is seldom limiting since practical considerations often impose such a restriction on possible choices for the $x_k$'s. This can be contrasted with the much less natural assumption that would consist in considering the feasible parameter set as finite. Although the interest of studying asymptotic properties of designs and estimators in contexts where the number of observations is usually limited might seem questionable, we think it is reassuring to know that, at least in the idealized framework of a known model with independent observations, the iterations (1) ensure suitable convergence properties. The results apply to a wide range of situations, but we focuss here on least-squares estimation in nonlinear regression and maximum-likelihood estimation for Bernoulli trials, which share many common aspects. The former case is considered in the next section, the modifications required for the latter are presented in Sect. 4.

# 3 Least-squares estimation in nonlinear regression

We first consider the case of a regression model with observations

$$Y_i = Y(x_i) = \eta(x_i, \bar{\theta}) + \varepsilon_i \,, \tag{2}$$

with $\bar{\theta}$ in the interior of $\Theta$, a compact subset of $\mathbb{R}^p$, $x_i \in \mathcal{X} \subset \mathbb{R}^d$, and $\{\varepsilon_i\}$ a sequence of independently and identically distributed random variables with $\mathbb{E}\{\varepsilon_1\} = 0$ and $\mathbb{E}\{\varepsilon_1^2\} = \sigma^2 < \infty$. We denote

$$S_N(\theta) = \sum_{k=1}^{N} [Y(x_k) - \eta(x_k, \theta)]^2$$

and $\hat{\theta}_{LS}^N$ the least-squares estimator minimizing $S_N(\theta)$, that is, $\hat{\theta}_{LS}^N = \arg\min_{\theta\in\Theta} S_N(\theta)$. We restrict our attention to ordinary least-squares and stationary errors and assume, without any further loss of generality, that $\sigma^2 = 1$. Assuming that $\eta(x,\theta)$ is differentiable with respect to $\theta$ for any $x$, the contribution of the design point $x$ to the information matrix is then $\mu(x,\theta) = \mathrm{f}_\theta(x)\,\mathrm{f}_\theta^\top(x)$, where we denote

$$\mathrm{f}_\theta(x) = \frac{\partial\eta(x,\theta)}{\partial\theta}\,.$$

The results can easily be extended to non stationary errors and weighted least-squares. In the case of maximum-likelihood estimation, the contribution of $x$ to the Fisher information matrix only differs by a multiplicative constant and is given by $\mu(x,\theta) = \mathcal{I}\,\mathrm{f}_\theta(x)\,\mathrm{f}_\theta^\top(x)$, with $\mathcal{I}$ the Fisher information for location: $\mathcal{I} = \int \left[\varphi'(t)/\varphi(t)\right]^2 \varphi(t)\,dt$, with $\varphi(\cdot)$ the probability density function of $\varepsilon_1$ and $\varphi'(\cdot)$ its derivative.

We shall need the following lemma, see Wu (1981, p. 504).

**Lemma 1** *If for any $\delta > 0$*

$$\liminf_{N\to\infty} \inf_{\|\theta-\bar{\theta}\|\geq\delta} [S_N(\theta) - S_N(\bar{\theta})] > 0 \quad \textit{almost surely}\,, \tag{3}$$

*then $\hat{\theta}_{LS}^N \overset{\text{a.s.}}{\to} \bar{\theta}$ as $N \to \infty$ (almost sure convergence). If for any $\delta > 0$*

$$\Pr\left\{ \inf_{\|\theta-\bar{\theta}\|\geq\delta} [S_N(\theta) - S_N(\bar{\theta})] > 0 \right\} \to 1\,,\ N \to \infty\,, \tag{4}$$

*then $\hat{\theta}_{LS}^N \overset{\text{p}}{\to} \bar{\theta}$ as $N \to \infty$ (convergence in probability).*

One can then show that the convergence of the least-squares estimator is a consequence of

$$D_N(\theta,\bar{\theta}) = \sum_{k=1}^{N} [\eta(x_k,\theta) - \eta(x_k,\bar{\theta})]^2 \tag{5}$$

tending to infinity fast enough for $\|\theta - \bar{\theta}\| \geq \delta > 0$. When the design space $\mathcal{X}$ for the $x_k$'s is finite, the rate of increase required for the strong consistency of $\hat{\theta}_{LS}^N$ ($\hat{\theta}_{LS}^N \overset{\text{a.s.}}{\to} \bar{\theta}$, $N \to \infty$) is quite slow, and the result is much stronger than what is needed to obtain strong consistency under the adaptive design (1) with $\hat{\theta}^k = \hat{\theta}_{LS}^k$. However, we think the result is interesting *per se* and state it as a theorem.

**Theorem 1** *Let* $\{x_i\}$ *be a non-random design sequence* on a finite set $\mathcal{X}$. *If* $D_N(\theta,\bar{\theta})$ *given by (5) satisfies*

$$\textit{for all}\ \ \delta>0\,,\ \left[\inf_{\|\theta-\bar{\theta}\|\geq\delta} D_N(\theta,\bar{\theta})\right]/(\log\log N)\to\infty\,,\ N\to\infty\,, \tag{6}$$

*then* $\hat{\theta}_{LS}^N \overset{\text{a.s.}}{\to} \bar{\theta}$ *as* $N\to\infty$. *This result remains valid for* $\{x_i\}$ *a random sequence on* $\mathcal{X}$ *finite when (6) holds almost surely. If* $D_N(\theta,\bar{\theta})$ *simply satisfies*

$$\textit{for all}\ \ \delta>0\,,\ \inf_{\|\theta-\bar{\theta}\|\geq\delta} D_N(\theta,\bar{\theta})\to\infty\,,\ N\to\infty\,, \tag{7}$$

*then* $\hat{\theta}_{LS}^N \overset{\text{P}}{\to} \bar{\theta}$, $N\to\infty$. *This result remains valid for* $\{x_i\}$ *a random sequence on* $\mathcal{X}$ *finite when (7) holds in probability.*

**Remark 2** *The condition*

$$\textit{for all}\ \ \theta\neq\bar{\theta}\,,\ D_N(\theta,\bar{\theta})=\sum_{k=1}^{N}[\eta(x_k,\theta)-\eta(x_k,\bar{\theta})]^2\to\infty\ \textit{as}\ N\to\infty\,,$$

*is sufficient for the strong consistency of* $\hat{\theta}_{LS}^N$ *when the parameter set* $\Theta$ *is finite, see Wu (1981). From Theorem 1, when* $\mathcal{X}$ *is finite this condition is also sufficient for the weak consistency of* $\hat{\theta}_{LS}^N$ *without restriction on* $\Theta$. *It is proved in (Wu, 1981) to be necessary for the existence of a weakly consistent estimator of* $\bar{\theta}$ *in the regression model (2) when the errors* $\varepsilon_i$ *are independent with a distribution having a density* $\varphi(\cdot)$ *positive almost everywhere and absolutely continuous with respect to the Lebesgue measure with finite Fisher information for location. Notice that a classical condition for strong consistency of least-squares estimation in nonlinear regression is* $D_N(\theta,\bar{\theta})=\mathcal{O}(N)$ *for* $\theta\neq\bar{\theta}$, *see e.g. Jennrich (1969), which is much stronger than (6).*

A major interest of Theorem 1 is that it does require the $x_k$'s to be non-random constants and also applies for sequential design.

**Remark 3** *In the context of sequential design, it is interesting to compare the results of the theorem with those obtained without the assumption of a finite design space* $\mathcal{X}$. *For linear regression, the condition (6) takes the form*

$$\log\log N = o\{\lambda_{\min}[N\mathrm{M}(\xi_N)]\}\,.$$

*Noticing that* $\lambda_{\max}[N\mathrm{M}(\xi_N)] = \mathcal{O}(N)$*, we thus get a condition much weaker than the sufficient condition*

$$\{\log(\lambda_{\max}[N\mathrm{M}(\xi_N)])\}^{1+\alpha} = o\{\lambda_{\min}[N\mathrm{M}(\xi_N)]\} \text{ for some } \alpha > 0\,,$$

*derived by Lai and Wei (1982) for the strong convergence of the least-squares estimator in a linear regression model under a sequential design. Also, in nonlinear regression, (6) is much less restrictive than the condition obtained by Lai (1994) for the strong consistency of the least-squares estimator under a sequential design; indeed, his proof, based on properties of Hilbert space-valued martingales, requires a condition that gives for linear regression*

$$\lambda_{\max}[N\mathrm{M}(\xi_N)] = \mathcal{O}\{(\lambda_{\min}[N\mathrm{M}(\xi_N)])^{\rho}\} \text{ for some } \rho \in (1,2)\,.$$

We shall make the following assumptions on the model (2), the parameter space $\Theta$ and experimental domain $\mathcal{X}$.

$\mathbf{H}_f$: For all $x$ in $\mathcal{X}$, $\mathrm{f}_\theta(x)$ is a continuous function of $\theta$ in the interior of $\Theta$.

$\mathbf{H}_{\mathcal{X}}$-(i): The design space $\mathcal{X}$ is finite, $\mathcal{X} = \{x^{(1)}, x^{(2)}, \ldots, x^{(K)}\}$.

$\mathbf{H}_{\mathcal{X}}$-(ii): $\inf_{\theta\in\Theta} \lambda_{\min}\left[\sum_{i=1}^{K} \mathrm{f}_\theta(x^{(i)})\mathrm{f}_\theta^\top(x^{(i)})\right] > \gamma > 0$.

$\mathbf{H}_{\mathcal{X}}$-(iii): The regression model (2) satisfies the following identifiability condition. For all $\delta > 0$ there exists $\epsilon(\delta) > 0$ such that for any subset $\{i_1, \ldots, i_p\}$ of distinct elements of $\{1, \ldots, K\}$,

$$\inf_{\|\theta-\bar\theta\|\geq\delta} \sum_{j=1}^{p} [\eta(x^{(i_j)}, \theta) - \eta(x^{(i_j)}, \bar\theta)]^2 > \epsilon(\delta)\,.$$

$\mathbf{H}_{\mathcal{X}}$-(iv): For any subset $\{i_1, \ldots, i_p\}$ of distinct elements of $\{1, \ldots, K\}$,

$$\lambda_{\min}\left[\sum_{j=1}^{p} \mathrm{f}_{\bar\theta}(x^{(i_j)})\mathrm{f}_{\bar\theta}^\top(x^{(i_j)})\right] \geq \bar\gamma > 0\,.$$

$\mathrm{H}_{\mathcal{X}}$-(iv) and $\mathrm{H}_f$ imply that $\mathrm{H}_{\mathcal{X}}$-(ii) is satisfied when $\Theta$ corresponds to some neighborhood of $\bar\theta$. As such $\mathrm{H}_{\mathcal{X}}$-(iii) is a global identifiability condition, which can be violated in some trivial examples (take for instance $p = 1$ and $\eta(x, \theta) = x\theta(1-\theta)$, so that $\eta(x, 1-\bar\theta) = \eta(x, \bar\theta)$ for all $x$). When this happens, it indicates a difficulty in the LS estimation problem, in the sense that the LS estimator $\hat\theta_{LS}^N$ may not be

unique if only $p$ design points are used in the experiment. Notice, however, that only values of $\theta \in \Theta$ have to be considered. This difficulty may thus disappear when $\Theta$ is small enough (note that $\mathrm{H}_{\mathcal{X}}$-(iv) can be considered as a local version of $\mathrm{H}_{\mathcal{X}}$-(iii) for $\theta$ close to $\bar{\theta}$). Finally, when $\mathrm{H}_f$ and $\mathrm{H}_{\mathcal{X}}$-(i) are satisfied, the maximum eigenvalue $\lambda_{\max}[\mathrm{M}(\xi,\theta)]$ of any information matrix $\mathrm{M}(\xi,\theta)$, $\xi \in \Xi$, is bounded by $L = \max_{x\in\mathcal{X},\,\theta\in\Theta} \|\mathrm{f}_\theta(x)\|^2$. Therefore, $\lambda_{\min}[\mathrm{M}(\xi,\theta)] \geq \det \mathrm{M}(\xi,\theta)/L^{p-1}$ and $\mathrm{H}_{\mathcal{X}}$-(ii), $\mathrm{H}_{\mathcal{X}}$-(iv) can be replaced by similar conditions involving the determinants of the matrices instead of their minimum eigenvalues. These assumptions will be discussed in Section 5 in the light of a series of examples showing that they are not very restrictive.

In order to avoid the difficulties raised by the interplay between estimation and design in (1) when $\hat{\theta}^k_{LS}$ is substituted for $\hat{\theta}^k$, we first state a uniform result on the number of design points receiving a weight bounded away from zero, by considering $(\hat{\theta}^k)$ in (1) as *any sequence* taking values in $\Theta$. We then have the following.

**Lemma 2** *Let $(\hat{\theta}^k)$ be an arbitrary sequence in $\Theta$ used to generate design points according to (1) for $k \geq p$, with an initialisation such that $\mathrm{M}(\xi_p,\theta)$ is non-singular for all $\theta$ in $\Theta$. Let $r_{N,i} = r_N(x^{(i)})$ denote the number of times $x^{(i)}$ appears in the sequence $x_1,\ldots,x_N$, $i = 1,\ldots,K$, and consider the associated order statistics $r_{N,1:K} \geq r_{N,2:K} \geq \cdots \geq r_{N,K:K}$. Define*

$$q^* = \max\{j : \text{ there exists } \alpha > 0 \text{ such that } \liminf_{N\to\infty} r_{N,j:K}/N > \alpha\}\,.$$

*Then, $H_{\mathcal{X}}$-(i) and $H_{\mathcal{X}}$-(ii) imply $q^* \geq p$. When the sequence $(\hat{\theta}^k)$ is random, the statement holds with probability one.*

Consider now a regression model satisfying $\mathrm{H}_{\mathcal{X}}$-(i-iii). Lemma 2 implies that there exist $N_0$ and $\alpha > 0$ such that $r_{N,j:K} > \alpha N$ for all $N > N_0$ and all $j = 1,\ldots,p$, and $\mathrm{H}_{\mathcal{X}}$-(iii) thus implies that $D_N(\theta,\bar{\theta})$ given by (5) satisfies $\inf_{\|\theta-\bar{\theta}\|\geq\delta} D_N(\theta,\bar{\theta}) > \alpha N\epsilon(\delta)$, $N > N_0$. Therefore, $\hat{\theta}^N_{LS} \overset{\text{a.s.}}{\to} \bar{\theta}$ $(N \to \infty)$ from Theorem 1. This holds for any sequence $(\hat{\theta}^k)$ in $\Theta$ and thus in particular when $\hat{\theta}^k_{LS}$ is substituted for $\hat{\theta}^k$ in (1). The last step before stating the main result of the paper concerns the following continuity argument.

**Lemma 3** *Under the conditions of Lemma 2 and the additional assumptions $H_f$ and $H_{\mathcal{X}}$-(iv) we have the continuity property: for all $\epsilon > 0$ there exists $\beta > 0$ such that*

$$\|\hat{\theta}^k - \bar{\theta}\| < \beta \text{ for all } k \text{ larger than some } K_0$$

*implies* $\liminf_{k\to\infty} \log\det \mathrm{M}(\xi_k, \bar{\theta}) \geq \log\det \mathrm{M}[\xi_D^*(\bar{\theta}), \bar{\theta}] - \epsilon$, *with $\xi_D^*(\bar{\theta})$ a D-optimal design measure for $\bar{\theta}$.*

We finally obtain the following.

**Theorem 2** *Suppose that in the regression model (2) the design points are generated sequentially according to (1) for $k \geq p$ with the least-squares estimator $\hat{\theta}_{LS}^k$ substituted for $\hat{\theta}^k$, and that the first $p$ design points are such that the information matrix $\mathrm{M}(x_1, \ldots, x_p, \theta)$ is non-singular for any $\theta$ in $\Theta$. Then, under $H_f$ and $H_{\mathcal{X}}$-(i-iv) we have $\hat{\theta}_{LS}^k \overset{\text{a.s.}}{\to} \bar{\theta}$ and $\mathrm{M}(\xi_k, \bar{\theta}) \overset{\text{a.s.}}{\to} \mathrm{M}[\xi_D^*(\bar{\theta}), \bar{\theta}]$, $k \to \infty$, with $\xi_D^*(\bar{\theta})$ a D-optimal design measure for $\theta = \bar{\theta}$.*

The proof directly follows from Lemma 2, Theorem 1 and Lemma 3.

**Remark 4** *One may notice that the property $q^* \geq p$ in Lemma 2 is in fact a property on the rank of the matrix $\mathrm{M}_N^{(1)}(\theta)$, see (17). Under the condition*

$$\text{for all } \xi \in \Xi\,,\ \mathrm{rank}[\mathrm{M}(\xi, \theta)] = r(\xi) \text{ is independent of } \theta \in \Theta\,,$$

*Theorem 2 remains valid when the assumptions $H_{\mathcal{X}}$-(iii-iv) are replaced by: for all $\delta > 0$ there exists $\epsilon(\delta) > 0$ such that, for any subset $\{i_1, \ldots, i_m\}$ of $\{1, \ldots, K\}$ such that* $\mathrm{rank}\left[\sum_{j=1}^m \mathrm{f}_{\bar{\theta}}(x^{(i_j)})\mathrm{f}_{\bar{\theta}}^\top(x^{(i_j)})\right] = p$, *we have* $\inf_{\|\theta-\bar{\theta}\|\geq\delta} \sum_{j=1}^m [\eta(x^{(i_j)}, \theta) - \eta(x^{(i_j)}, \bar{\theta})]^2 > \epsilon(\delta)$ *and* $\lambda_{\min}\left[\sum_{j=1}^m \mathrm{f}_{\bar{\theta}}(x^{(i_j)})\mathrm{f}_{\bar{\theta}}^\top(x^{(i_j)})\right] \geq \bar{\gamma} > 0$.

We conclude the section by a justification of the use $\mathrm{M}(\xi_k, \hat{\theta}_{LS}^k)$ as an asymptotic characterization of the precision of the estimation ($k \to \infty$). Complementing $\mathrm{H}_f$ by the assumption that $\eta(x, \theta)$ is two times continuously differentiable for $\theta$ in some open neighborhood of $\bar{\theta}$ for any $x$ in $\mathcal{X}$, we can easily obtain the following. A first-order series expansion of the components of the gradient $\nabla_\theta S_N(\theta)$ around $\bar{\theta}$ gives

$$\{\nabla_\theta S_N(\hat{\theta}_{LS}^N)\}_i = 0 = \{\nabla_\theta S_N(\bar{\theta})\}_i + \{\nabla_\theta^2 S_N(\tilde{\theta}_i^N)(\hat{\theta}_{LS}^N - \bar{\theta})\}_i\,,\ i = 1, \ldots, p\,,$$

where $\nabla^2_\theta S_N(\theta)$ is the (Hessian) matrix of second-order derivatives of $S_N(\theta)$ and $\tilde{\theta}_i^N = (1-\gamma_{N,i})\bar{\theta} + \gamma_{N,i}\hat{\theta}_{LS}^N$, $\gamma_{N,i} \in (0,1)$, with $\tilde{\theta}_i^N$ measurable, see Jennrich (1969). Using the fact that $\mathcal{X}$ is finite, straightforward calculations then give $\lim_{\delta\to 0}\sup_{\|\theta-\bar{\theta}\|\le\delta}\|\nabla^2_\theta S_N(\theta)/N - 2\mathrm{M}[\xi_D^*(\bar{\theta}),\bar{\theta}]\| \overset{\text{a.s.}}{\to} 0$, $N\to\infty$, and therefore, under the conditions of Theorem 2, $\nabla^2_\theta S_N(\tilde{\theta}_i^N)/N \overset{\text{a.s.}}{\to} 2\mathrm{M}[\xi_D^*(\bar{\theta}),\bar{\theta}]$ when $N\to\infty$. Also, $N^{-1/2}\,\nabla_\theta S_N(\bar{\theta}) = -2\sum_x \mathrm{v}_N(x)$, where $\mathrm{v}_N(x) = \zeta_N(x)\alpha_N(x)\mathrm{f}_{\bar{\theta}}(x)$ with $\zeta_N(x) = (\sum_{k=1,\,x_k=x}^N \varepsilon_k)/\sqrt{r_N(x)}$ and $\alpha_N(x) = \sqrt{r_N(x)/N}$. For $x$ such that $r_N(x)$ tends to infinity, $\zeta_N(x)$ tends to be distributed as a standard normal random variable, $\zeta_N(x) \overset{\text{d}}{\to} \zeta(x) \sim \mathcal{N}(0,1)$, $N\to\infty$; moreover, $\zeta(x^{(i)})$ and $\zeta(x^{(j)})$ are independent for $x^{(i)} \neq x^{(j)}$. From Theorem 2, $\sum_x \alpha_N^2(x)\mathrm{f}_{\bar{\theta}}(x)\mathrm{f}_{\bar{\theta}}^\top(x) \overset{\text{a.s.}}{\to} \mathrm{M}[\xi_D^*(\bar{\theta}),\bar{\theta}]$, $N\to\infty$. Finally from the series expansion above, we obtain $\sqrt{N}(\hat{\theta}_{LS}^N - \bar{\theta}) \overset{\text{d}}{\to} \mathrm{z} \sim \mathcal{N}(0, \mathrm{M}^{-1}[\xi_D^*(\bar{\theta}),\bar{\theta}])$, and therefore $[N\,\mathrm{M}(\xi_N,\hat{\theta}_{LS}^N)]^{1/2}(\hat{\theta}_{LS}^N - \bar{\theta}) \overset{\text{d}}{\to} \omega \sim \mathcal{N}(0,\mathrm{I})$, $N\to\infty$, which justifies the use of the information matrix to characterize the precision of the estimation under the adaptive design scheme (1) when the set $\mathcal{X}$ is finite.

The results above obtained for least-squares estimation in nonlinear regression can be extended to maximum-likelihood estimation in Bernoulli-trial experiments. This is considered in the next section.

# 4 Maximum-likelihood estimation in Bernoulli trials

Consider a Bernoulli-trial experiment (a dose-response problem for instance) with single response $Y$ equal to 0 or 1 (efficacy or toxicity response at the dose $x$) and $\Pr\{Y=1|\theta,x\} = \pi(x,\theta)$. The log-likelihood for the observation $Y$ at the design point $x$ is

$$l(\theta|Y,x) = Y\log[\pi(x,\theta)] + (1-Y)\log[1-\pi(x,\theta)]\,. \tag{8}$$

Suppose that $\pi(x,\theta)$ is differentiable with respect to $\theta$ for any $x$ and denote

$$\mathrm{f}_\theta(x) = \frac{\partial\pi(x,\theta)}{\partial\theta}\,\frac{1}{\sqrt{\pi(x,\theta)[1-\pi(x,\theta)]}}$$

so that the contribution of the point $x$ to the Fisher information matrix is $\mu(x,\theta) = \mathrm{f}_\theta(x)\mathrm{f}_\theta^\top(x)$. Multivariate extensions (e.g., where both efficacy and toxicity responses are observed at a dose $x$) could be considered similarly; see, e.g., Dragalin and Fedorov (2006) for Gumbel and Cox models. In terms of

design, the main difference with the single response case is the fact that $\mu(x,\theta)$ may have rank two, so that less than $p$ support points in $\xi$ may suffice to estimate $\theta$ consistently. Note that the same situation occurs for the regression model (2) when $\dim(\eta) > 1$ (it also happens when the variance function is not homogeneous and depends on unknown parameters of the model, see, e.g., Downing et al. (2001); Pázman and Pronzato (2004)).

We suppose that $\pi(x,\theta) \in (0,1)$ for any $\theta \in \Theta$ and $x \in \mathcal{X}$ and denote by $\theta_{ML}^N$ the maximum-likelihood estimator $\hat{\theta}_{ML}^N = \arg\max_{\theta\in\Theta} L_N(\theta)$, with $L_N(\theta) = \sum_{i=1}^N l(\theta|Y_i, x_i)$, see (8). We also suppose that $\Theta$ is a compact subset of $\mathbb{R}^p$ and that $\bar{\theta}$, the 'true' value of $\theta$ that generates the observations, lies in the interior of $\Theta$. We have the following equivalent of Lemma 1 for this context of binary trials.

**Lemma 4** *If for any $\delta > 0$*

$$\liminf_{N\to\infty} \inf_{\|\theta-\bar{\theta}\|\geq\delta} [L_N(\bar{\theta}) - L_N(\theta)] > 0 \quad \textit{almost surely (resp. in probability)}\,,$$

*then $\hat{\theta}_{ML}^N \overset{\text{a.s.}}{\to} \bar{\theta}$ (resp. $\hat{\theta}_{ML}^N \overset{\text{p}}{\to} \bar{\theta}$) as $N \to \infty$.*

The proof is identical to that of Lemma 1, see Wu (1981). We then obtain a property similar to Theorem 1.

**Theorem 3** *Let $\{x_i\}$ be a non-random design sequence* on a finite set $\mathcal{X}$. *Assume that*

$$D_N(\theta,\bar{\theta}) = \sum_{i=1}^N \pi(x_i,\bar{\theta}) \log\left[\frac{\pi(x_i,\bar{\theta})}{\pi(x_i,\theta)}\right] + [1-\pi(x_i,\bar{\theta})] \log\left[\frac{1-\pi(x_i,\bar{\theta})}{1-\pi(x_i,\theta)}\right] \tag{9}$$

*satisfies (6). Then, $\hat{\theta}_{ML}^N \overset{\text{a.s.}}{\to} \bar{\theta}$ as $N \to \infty$. The same is true for a random sequence on a finite $\mathcal{X}$ when (6) holds almost surely. If $D_N(\theta,\bar{\theta})$ simply satisfies (7), or $\inf_{\|\theta-\bar{\theta}\|\geq\delta} D_N(\theta,\bar{\theta}) \overset{\text{p}}{\to} \infty$ as $N \to \infty$ for all $\delta > 0$, then $\hat{\theta}_{ML}^N \overset{\text{p}}{\to} \bar{\theta}$.*

Lemmas 2 and 3 are still valid and Theorem 2 still applies, with $\text{H}_{\mathcal{X}}$-(iii) replaced by.

$\mathbf{H}_{\mathcal{X}}$-(iii'): For all $\delta > 0$ there exists $\epsilon(\delta) > 0$ such that for any subset $\{i_1,\ldots,i_p\}$ of distinct elements of $\{1,\ldots,K\}$,

$$\inf_{\|\theta-\bar{\theta}\|\geq\delta} \sum_{j=1}^p \pi(x^{(i_j)},\bar{\theta}) \log\left[\frac{\pi(x^{(i_j)},\bar{\theta})}{\pi(x^{(i_j)},\theta)}\right] + [1-\pi(x^{(i_j)},\bar{\theta})] \log\left[\frac{1-\pi(x^{(i_j)},\bar{\theta})}{1-\pi(x^{(i_j)},\theta)}\right] > \epsilon(\delta)\,.$$

Finally, similarly to the case of least-squares estimation in nonlinear regression, $\mathrm{M}(\xi_k, \hat{\theta}^k_{ML})$ can be used as an asymptotic characterization of the precision of the estimation ($k \to \infty$) under the adaptive scheme (1) when $\mathcal{X}$ is finite and $\pi(x,\theta)$ is two times continuously differentiable for $\theta$ in some open neighborhood of $\bar{\theta}$ for any $x$ in $\mathcal{X}$.

**Remark 5** *Defining* $g(a,b) = a\log(a/b)+(1-a)\log[(1-a)/(1-b)]$, $a,b \in (0,1)$, *we can easily check that, for any fixed* $a \in (0,1)$, $g(a,b) \geq 2(a-b)^2$ *with* $g(a,a) = 0$, *so that each term of the sum (9) is positive. Indeed, define* $h(a,b) = g(a,b) - 2(a-b)^2$; *we have* $\partial h(a,b)/\partial b = (2b-1)^2(b-a)/[b(1-b)]$ *so that, as a function of* $b$, $h(a,b)$ *monotonically decreases for* $0 < b < a$ *and monotonically increases for* $a < b < 1$. *Also, Theorem 3 is satisfied when* $D_N(\theta,\bar{\theta})$ *is replaced by* $D'_N(\theta,\bar{\theta}) = \sum_{i=1}^N \left[\pi(x_i,\bar{\theta}) - \pi(x_i,\theta)\right]^2$, *and a sufficient condition for* $H_{\mathcal{X}}$*-(iii') is* $\inf_{\|\theta-\bar{\theta}\|\geq\delta} \sum_{j=1}^p \left[\pi(x^{(i_j)},\bar{\theta}) - \pi(x^{(i_j)},\theta)\right]^2 > \epsilon(\delta) > 0$ *for any* $\delta > 0$ *and any subset* $\{i_1,\ldots,i_p\}$ *of* $\{1,\ldots,K\}$.

# 5 Examples

We consider the same four models as in (Hu, 1998). The response function, or probability of success, is continuously differentiable with respect to $\theta$ for each of them so that $\mathrm{H}_f$ is satisfied.

**Example 1:** Dilution method for estimating the density of an organism. Suppose that $\pi(x,\theta) = \Pr(Y = 1|\theta,x) = \exp(-\theta x)$, with $Y \in \{0,1\}$, $x \in \mathbb{R}^+$, $\theta \in \Theta = [a,b]$, $0 < a < b$, and $\bar{\theta}$, the unknown true value of $\theta$ satisfying $a < \bar{\theta} < b$. The Fisher information matrix at $\theta$ for $\beta = \log\theta$ and a design measure $\xi$ is then $\mathrm{M}(\xi,\beta) = \int_{\mathcal{X}} \theta^2 x^2/[\exp(\theta x) - 1]\,\xi(dx)$. The $D$-optimal measure is $\xi^*_D(\theta) = \delta_{x^*/\theta}$ with $x^*$ satisfying $(2-x^*)\exp(x^*) = 2$, that is, $x^* \simeq 1.5936$. Following this optimal design consideration, Hu (1998) suggests the adaptive construction $x_{k+1} = x^*/\hat{\theta}^k$ with $\hat{\theta}^k$ an estimate of $\theta$ based on $x_1, Y_1, \ldots, x_k, Y_k$. When $\hat{\theta}^k$ is the Bayes estimator $\hat{\theta}^k_B = \mathbb{E}\{\theta|\mathcal{F}_k\}$, with $\mathcal{F}_k$ the $\sigma$-field generated by the observations $Y_1,\ldots,Y_k$, his results imply the almost sure convergence of $\hat{\theta}^k_B$ to $\bar{\theta}$ and of the empirical measure $\xi_k$ to $\xi^*_D(\bar{\theta})$.

One can easily check that $\mathrm{H}_{\mathcal{X}}$(ii-iv) are satisfied for $\theta \in \Theta$ when $\mathcal{X} \subset \mathbb{R}^+$ is finite and does not contain 0 (indeed, $a < \theta < b$ implies $[\pi(x,\theta) - \pi(x,\bar{\theta})]^2 \geq \exp(-2bx)(\theta-\bar{\theta})^2\{1-\exp[(a-b)x]\}^2/(b-a)^2$ and

$\mu(x,\beta) = \theta^2 x^2/[\exp(\theta x) - 1] > 0)$. Since $p = 1$ in this example, $q^*$ of Lemma 2 satisfies $q^* \geq p$ and *any* adaptive design ensures that the maximum-likelihood estimator is strongly consistent when $\mathcal{X}$ is finite and $0 \notin \mathcal{X}$. In particular, when $\mathcal{X}$ corresponds to the discretization of some set $\mathcal{X}'$, taking $x_{k+1}$ as the point in $\mathcal{X}$ closest to $x^*/\hat{\theta}^k_{ML}$ ensures $\hat{\theta}^k_{ML} \overset{\text{a.s.}}{\to} \bar{\theta}$ and the empirical design measure $\xi_k$ converges a.s. to a design that can be made arbitrarily close to the $D$-optimal design on $\mathcal{X}'$ for $\bar{\theta}$ when the discretization of $\mathcal{X}'$ is fine enough. Note that a similar consistency result can be obtained in general for *one-parameter models*.

**Example 2:** Michaelis-Menten regression. Suppose that $Y_i = \bar{\theta}_1 x_i/(\bar{\theta}_2 + x_i) + \varepsilon_i$, with $(\varepsilon_i)$ satisfying the assumptions of Sect. 3, $\Theta = [L_1, U_1] \times [L_2, U_2]$, $0 < L_i < \bar{\theta}_i < U_i$, $i = 1, 2$. When $\mathcal{X} = (0, \bar{x}]$, the $D$-optimal measure for $\theta$ on $\mathcal{X}$ is

$$\xi^*_D(\theta) = (1/2)\,\delta_{x^*_1(\theta)} + (1/2)\,\delta_{x^*_2} \tag{10}$$

with $x^*_1(\theta) = \theta_2 \bar{x}/(2\theta_2 + \bar{x}) < x^*_2 = \bar{x}$. Lai (1994) suggests the following design sequence

$$\begin{cases} x_k = x^*_1(\hat{\theta}^{k-1}_{LS}) & \text{if } k \text{ is even and } k \notin \{k_1, k_2, \ldots\} \\ x_k = \bar{x} & \text{if } k \text{ is odd and } k \notin \{k_1, k_2, \ldots\} \\ c/(1 + \log k) & \text{if } k \in \{k_1, k_2, \ldots\} \end{cases}$$

where $k_i \sim i^\alpha$ as $i \to \infty$, for some $c > 0$ and $1 < \alpha < 2$, in order to obtain the strong convergence of $\hat{\theta}^k_{LS}$, see also Remark 3. Hu (1998) shows that the introduction of the perturbations $x_k = c/(1 + \log k)$ if $k \in \{k_1, k_2, \ldots\}$ is not necessary when using the Bayes estimator $\hat{\theta}^{k-1}_B = \mathbb{E}\{\theta | \mathcal{F}_{k-1}\}$, and that the sequence

$$\begin{cases} x_k = x^*_1(\hat{\theta}^{k-1}_B) & \text{if } k \text{ is even} \\ x_k = \bar{x} & \text{if } k \text{ is odd} \end{cases} \tag{11}$$

ensures $\hat{\theta}^k_B \overset{\text{a.s.}}{\to} \bar{\theta}$, $k \to \infty$.

Suppose now that $\mathcal{X}$ is finite, with $0 < \min(\mathcal{X}) < \max(\mathcal{X}) = \bar{x}$. One can easily check that $\mathrm{H}_\mathcal{X}$-(iii) is satisfied for $\theta \in \Theta$. Indeed, $\eta(x, \theta) = \eta(x, \bar{\theta})$ and $\eta(z, \theta) = \eta(z, \bar{\theta})$ for $\theta, \bar{\theta} \in \Theta$, $x > 0$, $z > 0$ and $x \neq z$ imply $\theta = \bar{\theta}$. Also, $\det[\mathrm{f}_\theta(x)\mathrm{f}^\top_\theta(x) + \mathrm{f}_\theta(z)\mathrm{f}^\top_\theta(z)] = x^2 z^2 \theta_1^2 (x - z)^2/[\theta_2 + x)^4(\theta_2 + z)^4]$ so that $\mathrm{H}_\mathcal{X}$-(ii), $\mathrm{H}_\mathcal{X}$-(iv) are satisfied and Theorem 2 applies. This means in particular that we do not need to know the

form (10) of the $D$-optimal design and can directly generate the design through (1) with $\hat{\theta}^k = \hat{\theta}^k_{LS}$, the least-squares estimator which is much easier to obtain than the Bayes estimator $\hat{\theta}^k_B$. Moreover, numerical simulations indicate that when $\hat{\theta}^k$ is frozen to some given value $\theta$, the convergence to the $D$-optimal design for $\theta$ is generally faster and more regular for (1) than for (11).

**Example 3:** First-order exponential regression. Suppose that $Y_i = \bar{\theta}_1 \exp(-\bar{\theta}_2 x_i) + \varepsilon_i$, with $(\varepsilon_i)$ satisfying the assumptions of Sect. 3 and $\Theta$ as in Example 2. Take $\mathcal{X}$ finite with $\min(\mathcal{X}) = \underline{x} \geq 0$. The $D$-optimal design measure is then $\xi^*_D(\theta) = (1/2)\,\delta_{\underline{x}} + (1/2)\,\delta_{\underline{x}+1/\theta_2}$. One can easily check that $\mathrm{H}_{\mathcal{X}}$-(iii) is satisfied for $\theta \in \Theta$ ($\eta(x,\theta) = \eta(x,\bar{\theta})$ and $\eta(z,\theta) = \eta(z,\bar{\theta})$ for $x \neq z$ imply $\theta = \bar{\theta}$); $\mathrm{H}_{\mathcal{X}}$-(ii), $\mathrm{H}_{\mathcal{X}}$-(iv) are satisfied too since $\det[\mathrm{f}_\theta(x)\mathrm{f}_\theta^\top(x) + \mathrm{f}_\theta(z)\mathrm{f}_\theta^\top(z)] = \theta_1^2(x-z)^2\exp[-2\theta_2(x+z)]$, and Theorem 2 applies again.

**Example 4:** Binary logistic regression. Take $\pi(x,\theta) = \exp(\theta_1+\theta_2 x)/[1+\exp(\theta_1+\theta_2 x)]$ in Sect. 4, with $x \in \mathcal{X}$ finite and $\theta \in \Theta$ compact. Using Remark 5 one can easily show that $\mathrm{H}_{\mathcal{X}}$-(iii') is satisfied. Indeed, $\pi(x,\theta) = \pi(x,\bar{\theta})$ is equivalent to $\theta_1 + \theta_2 x = \bar{\theta}_1 + \bar{\theta}_2 x$, so that $\pi(x,\theta) = \pi(x,\bar{\theta})$ and $\pi(x,\theta) = \pi(x,\bar{\theta})$ with $x \neq z$ imply $\theta = \bar{\theta}$. Also, $\det[\mathrm{f}_\theta(x)\mathrm{f}_\theta^\top(x) + \mathrm{f}_\theta(z)\mathrm{f}_\theta^\top(z)] = (x-z)^2\exp[2\theta_1 + \theta_2(x+z)]/\{[1+\exp(\theta_1 + \theta_2 x)][1+\exp(\theta_1+\theta_2 z)]\}^2$ so that $\mathrm{H}_{\mathcal{X}}$-(ii) and $\mathrm{H}_{\mathcal{X}}$-(iv) are satisfied. Therefore, Theorem 2 applies and the almost sure convergence of the maximum-likelihood estimator $\hat{\theta}^k_{ML}$ to the true parameter value $\bar{\theta}$ and of $\xi_k$ to a $D$-optimal design for $\bar{\theta}$ is guaranteed when using (1) with $\hat{\theta}^k_{ML}$ substituted for $\hat{\theta}^k$.

# Appendix: proofs

*Proof of Theorem 1.* The proof is based on Lemma 1. We have

$$\begin{aligned} S_N(\theta) - S_N(\bar{\theta}) &= D_N(\theta,\bar{\theta})\left[1 + 2\,\frac{\sum_{x\in\mathcal{X}}\left(\sum_{k=1,\,x_k=x}^N \varepsilon_k\right)[\eta(x,\bar{\theta}) - \eta(x,\theta)]}{D_N(\theta,\bar{\theta})}\right] \\ &\geq D_N(\theta,\bar{\theta})\left[1 - 2\,\frac{\sum_{x\in\mathcal{X}}\left|\sum_{k=1,\,x_k=x}^N \varepsilon_k\right||\eta(x,\bar{\theta}) - \eta(x,\theta)|}{D_N(\theta,\bar{\theta})}\right]. \end{aligned}$$

From Lemma 1, under the condition (6) it suffices to prove that

$$\sup_{\|\theta-\bar{\theta}\|\geq\delta} \frac{\sum_{x\in\mathcal{X}} \left|\sum_{k=1,\, x_k=x}^{N} \varepsilon_k\right| |\eta(x,\bar{\theta})-\eta(x,\theta)|}{D_N(\theta,\bar{\theta})} \xrightarrow{\text{a.s.}} 0 \tag{12}$$

for any $\delta > 0$ to obtain the strong consistency of $\hat{\theta}_{LS}^N$. Since $D_N(\theta,\bar{\theta}) \to \infty$ and $\mathcal{X}$ is finite, only the design points such that $r_N(x) \to \infty$ have to be considered, where $r_N(x)$ denotes the number of times $x$ appears in the sequence $x_1, \ldots, x_N$. Define $\beta(n) = \sqrt{n \log\log n}$. From the law of the iterated logarithm, we have

$$\text{for all} \quad x \in \mathcal{X}\,, \quad \limsup_{r_N(x)\to\infty} \left| \frac{1}{\beta[r_N(x)]} \sum_{k=1,\, x_k=x}^{N} \varepsilon_k \right| = \sigma\sqrt{2}\,, \quad \text{almost surely}\,, \tag{13}$$

see e.g. Shiryaev (1996, p. 397). Next, for any $x \in \mathcal{X}$,

$$\frac{\beta[r_N(x)]|\eta(x,\bar{\theta})-\eta(x,\theta)|}{D_N(\theta,\bar{\theta})} \leq \frac{[\log\log r_N(x)]^{1/2}}{D_N^{1/2}(\theta,\bar{\theta})}\,,$$

which, together with (6) and (13), gives (12). When the sequence $\{x_i\}$ is random, the only modification consists in working conditionally on the event (6).

When $\inf_{\|\theta-\bar{\theta}\|\geq\delta} D_N(\theta,\bar{\theta}) \to \infty$ as $N \to \infty$, we only need to prove that

$$\sup_{\|\theta-\bar{\theta}\|\geq\delta} \frac{\sum_{x\in\mathcal{X}} \left|\sum_{k=1,\, x_k=x}^{N} \varepsilon_k\right| |\eta(x,\bar{\theta})-\eta(x,\theta)|}{D_N(\theta,\bar{\theta})} \xrightarrow{\text{p}} 0 \tag{14}$$

for any $\delta > 0$ to obtain the weak consistency of $\hat{\theta}_{LS}^N$. We proceed as above and only consider the design points such that $r_N(x) \to \infty$, with now $\beta(n) = \sqrt{n}$. From the central limit theorem, for any $x \in \mathcal{X}$, $\left(\sum_{k=1,\, x_k=x}^{N} \varepsilon_k\right) / \sqrt{r_N(x)} \xrightarrow{\text{d}} \zeta_x \sim \mathcal{N}(0,\sigma^2)$ as $r_N(x) \to \infty$ and is thus bounded in probability. Also, for any $x \in \mathcal{X}$, $\sqrt{r_N(x)}|\eta(x,\bar{\theta})-\eta(x,\theta)|/D_N(\theta,\bar{\theta}) \leq D_N^{-1/2}(\theta,\bar{\theta})$, so that (7) implies (14). ■

*Proof of Lemma 2.* First note that $q^* \geq 1$ since $\mathcal{X}$ is finite. Suppose that $p \geq 2$ and $q^* < p$. We show that this leads to a contradiction.

For any $N$ we can write

$$\begin{aligned} \mathrm{M}(\xi_N,\theta) &= \frac{1}{N}\sum_{k=1}^{N} \mathrm{f}_\theta(x_k)\mathrm{f}_\theta^\top(x_k) \\ &= \frac{1}{N}\sum_{i=1}^{q^*} r_{N,i:K}\, \mathrm{f}_\theta(x^{(i_N)})\mathrm{f}_\theta^\top(x^{(i_N)}) + \frac{1}{N}\sum_{x_k\notin\mathcal{X}_N(q^*)} \mathrm{f}_\theta(x_k)\mathrm{f}_\theta^\top(x_k)\,, \end{aligned} \tag{15}$$

where $i_N$ is the index (depending on $N$) of a design point appearing $r_{N,i:K}$ times in $x_1, \ldots, x_N$ and $\mathcal{X}_N(q^*) = \{x^{(1_N)}, \ldots, x^{(q^*_N)}\}$ is the set of such points for $i \leq q^*$. Let $\mathrm{M}_N(\theta)$ denote the first matrix on the right-hand side of (15). For any $x^{(i_N)} \in \mathcal{X}_N(q^*)$ we have

$$\mathrm{f}_\theta^\top(x^{(i_N)})\mathrm{M}^{-1}(\xi_N,\theta)\mathrm{f}_\theta(x^{(i_N)}) \leq \mathrm{f}_\theta^\top(x^{(i_N)})\mathrm{M}_N^-(\theta)\mathrm{f}_\theta(x^{(i_N)}) = \frac{N}{r_{N,i:K}}\,,$$

with $\mathrm{M}_N^-(\theta)$ any $g$-inverse of $\mathrm{M}_N(\theta)$. Therefore, from the definition of $q^*$, there exists $N_0$ such that

$$\text{for all } \; i \leq q^*\,,\; N > N_0 \text{ and } \theta \in \Theta\,,\; \mathrm{f}_\theta^\top(x^{(i_N)})\mathrm{M}^{-1}(\xi_N,\theta)\mathrm{f}_\theta(x^{(i_N)}) \leq \frac{1}{\alpha}\,. \tag{16}$$

Let $\beta_N = r_{N,(q^*+1):K}/N$. Showing that $\liminf_{N\to\infty} \beta_N \geq \underline{\beta}$ for some $\underline{\beta} > 0$ will contradict the definition of $q^*$.

Define

$$\mathrm{M}_N^{(1)}(\theta) = \sum_{i=1}^{q^*} \mathrm{f}_\theta(x^{(i_N)})\mathrm{f}_\theta^\top(x^{(i_N)})\,, \quad \mathrm{M}_N^{(2)}(\theta) = \sum_{i=1}^{K} \mathrm{f}_\theta(x^{(i_N)})\mathrm{f}_\theta^\top(x^{(i_N)})\,. \tag{17}$$

We have $(1-\beta_N)\,\mathrm{M}_N^{(1)}(\theta) + \beta_N\,\mathrm{M}_N^{(2)}(\theta) - \mathrm{M}(\xi_N,\theta) \in \mathbb{M}^{\geq}$, where $\mathbb{M}^{\geq}$ is the set of symmetric nonnegative definite $p \times p$ matrices. For any $\mathrm{u} \in \mathbb{R}^p$,

$$\begin{aligned}
\mathrm{u}^\top\mathrm{M}^{-1}(\xi_N,\theta)\mathrm{u} \quad &\geq \quad \mathrm{u}^\top[(1-\beta_N)\,\mathrm{M}_N^{(1)}(\theta) + \beta_N\,\mathrm{M}_N^{(2)}(\theta)]^{-1}\mathrm{u} \\
&= \max_{\mathrm{z}\in\mathbb{R}^p} 2\mathrm{z}^\top\mathrm{u} - \mathrm{z}^\top[(1-\beta_N)\,\mathrm{M}_N^{(1)}(\theta) + \beta_N\,\mathrm{M}_N^{(2)}(\theta)]\mathrm{z} \\
&\geq \max_{\mathrm{z}\in\mathcal{N}[\mathrm{M}_N^{(1)}(\theta)]} 2\mathrm{z}^\top\mathrm{u} - \mathrm{z}^\top[(1-\beta_N)\,\mathrm{M}_N^{(1)}(\theta) + \beta_N\,\mathrm{M}_N^{(2)}(\theta)]\mathrm{z}
\end{aligned}$$

with $\mathcal{N}(\mathrm{M}) = \{\mathrm{v} : \mathrm{Mv} = 0\}$ the null-space of the matrix M. Direct calculations then give

$$\mathrm{u}^\top\mathrm{M}^{-1}(\xi_N,\theta)\mathrm{u} \geq \frac{1}{\beta_N}\,\mathrm{u}^\top[\mathrm{M}_N^{(2)}(\theta)]^{-1}\,[\mathrm{I} - \mathrm{P}_N(\theta)]\,\mathrm{u} \tag{18}$$

with I the $p$-dimensional identity matrix and $\mathrm{P}_N(\theta)$ the projector

$$\mathrm{P}_N(\theta) = \mathrm{M}_N^{(1)}(\theta)\left[\mathrm{M}_N^{(1)}(\theta)[\mathrm{M}_N^{(2)}(\theta)]^{-1}\mathrm{M}_N^{(1)}(\theta)\right]^-\mathrm{M}_N^{(1)}(\theta)[\mathrm{M}_N^{(2)}(\theta)]^{-1}\,.$$

Note that the right-hand side of (18) is zero when $\mathrm{u} \in \mathcal{M}[\mathrm{M}_N^{(1)}(\theta)]$ (i.e., when $\mathrm{u} = \mathrm{M}_N^{(1)}(\theta)\mathrm{v}$ for some $\mathrm{v} \in \mathbb{R}^p$). When $\mathrm{u} = \mathrm{f}_\theta(x^{(i_N)})$ for some $i \in \{q^*+1, \ldots, K\}$ we can construct a lower bound for this term, of the form $C/\beta_N$ with $C$ constant. Indeed, from (17) and $\mathrm{H}_\mathcal{X}$-(ii),

$$\text{for all } \; \theta \in \Theta \text{ and } \mathrm{v} \in \mathbb{R}^p\,,\; \mathrm{v}^\top\left[\mathrm{M}_N^{(1)}(\theta) + \sum_{i=q^*+1}^{K} \mathrm{f}_\theta(x^{(i_N)})\mathrm{f}_\theta^\top(x^{(i_N)})\right]\mathrm{v} > \gamma\,\|\mathrm{v}\|^2$$

so that for all $\theta \in \Theta$ and $\mathrm{z} \in \mathcal{N}[\mathrm{M}_N^{(1)}(\theta)]$,

$$\max_{i=q^*+1,\ldots,K} [\mathrm{z}^\top \mathrm{f}_\theta(x^{(i_N)})]^2 > \frac{\gamma}{K-q^*} \|\mathrm{z}\|^2 . \tag{19}$$

Take $\mathrm{z} = \mathrm{z}_{\theta,i_N} = [\mathrm{M}_N^{(2)}(\theta)]^{-1}[\mathrm{I} - \mathrm{P}_N(\theta)]\mathrm{f}_\theta(x^{(i_N)})$ for some $i \in \{q^*+1,\ldots,K\}$, so that $\mathrm{z}_{\theta,i_N} \in \mathcal{N}[\mathrm{M}_N^{(1)}(\theta)]$ and

$$\mathrm{f}_\theta^\top(x^{(i_N)})[\mathrm{M}_N^{(2)}(\theta)]^{-1}\left[\mathrm{I} - \mathrm{P}_N(\theta)\right]\mathrm{f}_\theta(x^{(i_N)}) = \mathrm{z}_{\theta,i_N}^\top \mathrm{f}_\theta(x^{(i_N)}) = \mathrm{z}_{\theta,i_N}^\top \mathrm{M}_N^{(2)}(\theta)\mathrm{z}_{\theta,i_N} .$$

We obtain

$$\begin{aligned}\max_{i=q^*+1,\ldots,K} \mathrm{f}_\theta^\top(x^{(i_N)})[\mathrm{M}_N^{(2)}(\theta)]^{-1}\left[\mathrm{I} - \mathrm{P}_N(\theta)\right]\mathrm{f}_\theta(x^{(i_N)}) &= \max_{i,j=q^*+1,\ldots,K} \mathrm{z}_{\theta,i_N}^\top \mathrm{M}_N^{(2)}(\theta)\mathrm{z}_{\theta,j_N} \\ &= \max_{i,j=q^*+1,\ldots,K} \mathrm{z}_{\theta,i_N}^\top \mathrm{f}_\theta(x^{(j_N)}) ,\end{aligned}$$

and thus from (19),

$$\text{for all } \theta \in \Theta , \max_{i=q^*+1,\ldots,K} \mathrm{f}_\theta^\top(x^{(i_N)})[\mathrm{M}_N^{(2)}(\theta)]^{-1}\left[\mathrm{I} - \mathrm{P}_N(\theta)\right]\mathrm{f}_\theta(x^{(i_N)}) > \left(\frac{\gamma}{K-q^*}\right)^{1/2} \max_{i=q^*+1,\ldots,K} \|\mathrm{z}_{\theta,i_N^*}\| .$$

Let $i_N^*$ denote the argument of the maximum on the left-hand side. We have,

$$\mathrm{z}_{\theta,i_N^*}^\top \mathrm{f}_\theta(x^{(i_N^*)}) = \mathrm{z}_{\theta,i_N^*}^\top \mathrm{M}_N^{(2)}(\theta)\mathrm{z}_{\theta,i_N^*} \leq K\, L\, \|\mathrm{z}_{\theta,i_N^*}\|^2$$

with $L = \max_{x\in\mathcal{X},\,\theta\in\Theta} \|\mathrm{f}_\theta(x)\|^2$, so that finally

$$\text{for all } \theta \in \Theta , \max_{i=q^*+1,\ldots,K} \mathrm{f}_\theta^\top(x^{(i_N)})[\mathrm{M}_N^{(2)}(\theta)]^{-1}\left[\mathrm{I} - \mathrm{P}_N(\theta)\right]\mathrm{f}_\theta(x^{(i_N)}) > \frac{\gamma}{LK(K-q^*)} .$$

To summarize, from (16, 18), there exists $N_0$ such that for any $N > N_0$ and for all $\theta \in \Theta$,

$$\begin{aligned}&\max_{i=1,\ldots,q^*} \mathrm{f}_\theta^\top(x^{(i_N)})\mathrm{M}^{-1}(\xi_N,\theta)\mathrm{f}_\theta(x^{(i_N)}) \leq \frac{1}{\alpha} \\ &\max_{i=q^*+1,\ldots,K} \mathrm{f}_\theta^\top(x^{(i_N)})\mathrm{M}^{-1}(\xi_N,\theta)\mathrm{f}_\theta(x^{(i_N)}) \geq \frac{1}{\beta_N}\,\frac{\gamma}{LK(K-q^*)} .\end{aligned}$$

Therefore, for $N > N_0$, $\beta_N < \beta^* = \gamma\alpha/[LK(K-q^*)]$ implies $x_{N+1} \in \{x^{((q^*+1)_N)},\ldots,x^{(K_N)}\}$ in the sequence (1). Define

$$\beta_N^* = \frac{\sum_{i=q^*+1}^K r_{N,i:K}}{(K-q^*)N} ,$$

so that $\beta_N \geq \beta_N^* \geq \beta_N/(K-q^*)$. Also, when $N > N_0$, $(\sum_{i=1}^{q^*} r_{N,i:K})/N > q^*\alpha$, so that $\beta_N^* < (1-q^*\alpha)/(K-q^*)$. By construction, $\beta_N < \beta^*$ and $N > N_0$ then give

$$\begin{aligned}\beta_{N+1} \geq \beta_{N+1}^* = \frac{N\beta_N^*(K-q^*)+1}{(K-q^*)(N+1)} &= \beta_N^* + \frac{1}{N+1}\left(\frac{1}{K-q^*} - \beta_N^*\right) \\ &> \beta_N^* + \frac{1}{N+1}\,\frac{q^*\alpha}{K-q^*} \geq \frac{\beta_N}{K-q^*} + \frac{1}{N+1}\,\frac{q^*\alpha}{K-q^*}\,.\end{aligned}$$

By induction, this lower bound on $\beta_{N+k}$ increases with $k$,

$$\beta_{N+k} > \frac{\beta_N}{K-q^*} + \frac{q^*\alpha}{K-q^*}\sum_{i=1}^{k}\frac{1}{N+i}\,,$$

until $\beta_{N+k}$ becomes larger that $\beta^*$. Suppose that the threshold $\beta^*$ is crossed downwards at $N_1 > N_0$, i.e., $\beta_{N_1-1} \geq \beta^*$ and $\beta_{N_1} < \beta^*$. This implies $\beta_{N_1} = \beta_{N_1-1}(N_1-1)/N_1$ and thus $\beta^*(N_1-1)/N_1 \leq \beta_{N_1} < \beta^*$, so that $\beta_{N_1}$ tends to $\beta^*$ when $N_1 \to \infty$.

We thus obtain $\liminf_{N\to\infty} \beta_N \geq \underline{\beta} = \beta^*/(K-q^*)$, showing that $q^* \geq p$, which concludes the proof.

■

*Proof of Lemma 3.* With the same notations as in Lemma 2, there exists $N_0$ and $\alpha > 0$ such that

$$\text{for all}\;\; N > N_0\,,\;\; r_{N,j:K} > \alpha N\,,\;\; j = 1,\ldots,q^* \geq p$$

and thus, from $\mathrm{H}_{\mathcal{X}}$-(iv), $\lambda_{\min}[\mathrm{M}(\xi_N,\bar{\theta})] > \alpha\bar{\gamma}$. From $\mathrm{H}_f$ and $\mathrm{H}_{\mathcal{X}}$-(i),

$$\text{for all}\;\; \epsilon_1 > 0\,,\;\; \text{there exists}\;\; \beta_1\;\; \text{such that}\;\; \|\theta-\bar{\theta}\| \leq \beta_1 \Rightarrow \max_{x\in\mathcal{X}} \|\mathrm{f}_\theta(x) - \mathrm{f}_{\bar{\theta}}(x)\| < \epsilon_1\,.$$

Direct calculations then give $\max_{x\in\mathcal{X}} \max_{\|\theta-\bar{\theta}\|\leq\beta_1} |\mathrm{f}_\theta^\top(x)\mathrm{M}^{-1}(\xi_N,\theta)\mathrm{f}_\theta(x) - \mathrm{f}_{\bar{\theta}}^\top(x)\mathrm{M}^{-1}(\xi_N,\bar{\theta})\mathrm{f}_{\bar{\theta}}(x)| < C\epsilon_1$ for $\epsilon_1$ small enough and $N > N_0$, with $C$ a constant depending on $\alpha$, $\bar{\gamma}$ and $\bar{f} = \max_{x\in\mathcal{X}} \|f_{\bar{\theta}}(x)\|$. Therefore, for all $\epsilon > 0$, there exists $\beta > 0$ such that for all $N > N_0$, $\|\hat{\theta}^N - \bar{\theta}\| < \beta$ in the algorithm (1) implies

$$\begin{aligned}\mathrm{f}_{\bar{\theta}}^\top(x_{N+1})\mathrm{M}^{-1}(\xi_N,\bar{\theta})\mathrm{f}_{\bar{\theta}}(x_{N+1}) &> \mathrm{f}_{\hat{\theta}^N}^\top(x_{N+1})\mathrm{M}^{-1}(\xi_N,\hat{\theta}^N)\mathrm{f}_{\hat{\theta}^N}(x_{N+1}) - \frac{\epsilon}{2} \\ &= \max_{x\in\mathcal{X}} \mathrm{f}_{\hat{\theta}^N}^\top(x)\mathrm{M}^{-1}(\xi_N,\hat{\theta}^N)\mathrm{f}_{\hat{\theta}^N}(x) - \frac{\epsilon}{2} \\ &> \max_{x\in\mathcal{X}} \mathrm{f}_{\bar{\theta}}^\top(x)\mathrm{M}^{-1}(\xi_N,\bar{\theta})\mathrm{f}_{\bar{\theta}}(x) - \epsilon\,.\end{aligned}$$

The rest of the proof only exploits the consequences of this property.

Take $\epsilon$, $\beta$ as above and suppose that there exists $\delta > 0$ such that

$$\log\det \mathrm{M}(\xi_N, \bar{\theta}) < \Psi^* - \epsilon - \delta \tag{20}$$

for all $N$ larger than some $N_1$, where $\Psi^* = \log\det \mathrm{M}(\xi_D^*, \bar{\theta})$ with $\xi_D^* = \xi_D^*(\bar{\theta})$ a $D$-optimal design measure for $\bar{\theta}$. By concavity, this implies $\max_{x\in\mathcal{X}} \mathrm{f}_{\bar{\theta}}^\top(x)\mathrm{M}^{-1}(\xi_N, \bar{\theta})\mathrm{f}_{\bar{\theta}}(x) > p + \epsilon + \delta$ for all $N > N_1$, and thus

$$\mathrm{f}_{\bar{\theta}}^\top(x_{N+1})\mathrm{M}^{-1}(\xi_N, \bar{\theta})\mathrm{f}_{\bar{\theta}}(x_{N+1}) > p + \delta\,, \text{ for all } \; N > \max(N_0, N_1, K_0)$$

when $\|\hat{\theta}^k - \bar{\theta}\| < \beta$ for all $k > K_0$. Direct calculations give

$$\log\det \mathrm{M}(\xi_{N+1}, \bar{\theta}) - \log\det \mathrm{M}(\xi_N, \bar{\theta}) = \log\left[1 + \frac{\mathrm{f}_{\bar{\theta}}^\top(x_{N+1})\mathrm{M}^{-1}(\xi_N, \bar{\theta})\mathrm{f}_{\bar{\theta}}(x_{N+1})}{N}\right] - p\log\left(1 + \frac{1}{N}\right) \tag{21}$$

and thus,

$$\log\det \mathrm{M}(\xi_{N+1}, \bar{\theta}) - \log\det \mathrm{M}(\xi_N, \bar{\theta}) \geq \log\left(1 + \frac{p+\delta}{N}\right) - p\log\left(1 + \frac{1}{N}\right) \geq \frac{\delta}{2N}$$

for $N$ large enough. This implies $\log\det \mathrm{M}(\xi_N, \bar{\theta}) \to \infty$ as $N \to \infty$, which is in contradiction with $\mathrm{H}_f$. Therefore, there exists a subsequence $\xi_{N_t}$ such that $\limsup_{t\to\infty} \log\det \mathrm{M}(\xi_{N_t}, \bar{\theta}) \geq \Psi^* - \epsilon$. From (21),

$$\text{for all } \; \delta > 0\,, \;\; \text{there exists } \; N_1 \text{ such that for all } \; N > N_1 \;\; \log\det \mathrm{M}(\xi_{N+1}, \bar{\theta}) > \log\det \mathrm{M}(\xi_N, \bar{\theta}) - \delta\,.$$

Also, from the developments just above, there exists $N_2$ such that for all $N > N_2$, (20) implies

$$\log\det \mathrm{M}(\xi_{N+1}, \bar{\theta}) > \log\det \mathrm{M}(\xi_N, \bar{\theta})\,.$$

Take any $N_t > \max(N_1, N_2)$ satisfying $\log\det \mathrm{M}(\xi_{N_t}, \bar{\theta}) > \Psi^* - \epsilon - \delta$, we obtain

$$\log\det \mathrm{M}(\xi_N, \bar{\theta}) > \Psi^* - \epsilon - 2\delta\,, \text{ for all } \; N > N_t\,.$$

Since $\delta$ is arbitrary, $\liminf_{N\to\infty} \log\det \mathrm{M}(\xi_N, \bar{\theta}) \geq \Psi^* - \epsilon$. ■

*Proof of Theorem 3.* We can write

$$\begin{aligned} L_N(\bar{\theta}) - L_N(\theta) &= D_N(\theta, \bar{\theta}) \left[1 + \frac{\sum_{x\in\mathcal{X}} \left(\sum_{k=1,\, x_k = x}^N [Y_k - \pi(x, \bar{\theta})]\right) \log\frac{\pi(x,\bar{\theta})[1-\pi(x,\theta)]}{\pi(x,\theta)[1-\pi(x,\bar{\theta})]}}{D_N(\theta, \bar{\theta})}\right] \\ &\geq D_N(\theta, \bar{\theta}) \left[1 - \frac{\sum_{x\in\mathcal{X}} \left|\sum_{k=1,\, x_k = x}^N [Y_k - \pi(x, \bar{\theta})]\right| \left|\log\frac{\pi(x,\bar{\theta})[1-\pi(x,\theta)]}{\pi(x,\theta)[1-\pi(x,\bar{\theta})]}\right|}{D_N(\theta, \bar{\theta})}\right]. \end{aligned}$$

From Lemma 4, under the condition (6) it suffices to prove that

$$\sup_{\|\theta-\bar\theta\|\geq\delta} \frac{\sum_{x\in\mathcal{X}} \left|\sum_{k=1,\, x_k=x}^{N} [Y_k - \pi(x,\bar\theta)]\right| \left|\log \frac{\pi(x,\bar\theta)[1-\pi(x,\theta)]}{\pi(x,\theta)[1-\pi(x,\bar\theta)]}\right|}{D_N(\theta,\bar\theta)} \stackrel{\text{a.s.}}{\to} 0$$

for any $\delta > 0$ to obtain the strong consistency of $\hat\theta_{LS}^N$. Similarly to the case of Theorem 1, since $D_N(\theta,\bar\theta) \to \infty$ only the design points such that $r_N(x) \to \infty$ have to be considered, and, from the law of the iterated logarithm,

$$\text{for all } \; x \in \mathcal{X}\,, \; \limsup_{r_N(x)\to\infty} \left| \frac{1}{\beta[r_N(x)]} \sum_{k=1,\, x_k=x}^{N} [Y_k - \pi(x,\bar\theta)] \right| = \sqrt{2\pi(x,\bar\theta)[1-\pi(x,\bar\theta)]}\,, \quad \text{almost surely}\,,$$

with $\beta(n) = \sqrt{n \log\log n}$. The rest of the proof for almost sure convergence is as for Theorem 1. We have

$$\frac{\beta[r_N(x)] \left|\log \frac{\pi(x,\bar\theta)[1-\pi(x,\theta)]}{\pi(x,\theta)[1-\pi(x,\bar\theta)]}\right| \sqrt{\pi(x,\bar\theta)[1-\pi(x,\bar\theta)]}}{D_N(\theta,\bar\theta)} \leq \frac{(\log\log r_N(x))^{1/2}}{D_N^{1/2}(\theta,\bar\theta)}\, \rho[\pi(x,\bar\theta),\pi(x,\theta)]$$

where

$$\rho(a,b) = \frac{\left|\log \frac{a(1-b)}{b(1-a)}\right| \sqrt{a(1-a)}}{\left[a\, \log(a/b) + (1-a)\, \log\left(\frac{1-a}{1-b}\right)\right]^{1/2}}\,, \; a,b \in (0,1)\,,$$

which, for any fixed $a \in (0,1)$, tends to infinity for $b$ tending to 0 or 1 and has a unique minimum in $(0,1)$ at $b = 1-a$. Since $0 < \pi(x,\theta) < 1$ for all $x$ and all $\theta$ in $\Theta$ compact, $\rho[\pi(x,\bar\theta),\pi(x,\theta)]$ is bounded and (6) gives the result.

When $\inf_{\|\theta-\bar\theta\|\geq\delta} D_N(\theta,\bar\theta) \to \infty$ as $N \to \infty$, we only need to prove that

$$\sup_{\|\theta-\bar\theta\|\geq\delta} \frac{\sum_{x\in\mathcal{X}} \left|\sum_{k=1,\, x_k=x}^{N} [Y_k - \pi(x,\bar\theta)]\right| \left|\log \frac{\pi(x,\bar\theta)[1-\pi(x,\theta)]}{\pi(x,\theta)[1-\pi(x,\bar\theta)]}\right|}{D_N(\theta,\bar\theta)} \stackrel{\text{p}}{\to} 0$$

for any $\delta > 0$ to obtain the weak consistency of $\hat\theta_{LS}^N$. We proceed as above and only consider the design points such that $r_N(x) \to \infty$, with now $\beta(n) = \sqrt{n}$. From the central limit theorem, for any $x \in \mathcal{X}$, $\left(\sum_{k=1,\, x_k=x}^{N} [Y_k - \pi(x,\bar\theta)]\right) / \sqrt{r_N(x)} \stackrel{\text{d}}{\to} \zeta_x \sim \mathcal{N}(0, \pi(x,\bar\theta)[1-\pi(x,\bar\theta)])$ as $r_N(x) \to \infty$ and is thus bounded in probability. The rest of the proof is as above, again using the fact that $\rho[\pi(x,\bar\theta),\pi(x,\theta)]$ is bounded. ■

# References


Atkinson, A., Donev, A., 1992. Optimum Experimental Design. Oxford University Press.

Chaloner, K., Larntz, K., 1989. Optimal Bayesian design applied to logistic regression experiments. Journal of Statistical Planning and Inference 21, 191–208.

Chaloner, K., Verdinelli, I., 1995. Bayesian experimental design: a review. Statistical Science 10 (3), 273–304.

Chaudhuri, P., Mykland, P., 1993. Nonlinear experiments: optimal design and inference based likelihood. Journal of the American Statistical Association 88 (422), 538–546.

Chaudhuri, P., Mykland, P., 1995. On efficiently designing of nonlinear experiments. Statistica Sinica 5, 421–440.

Downing, D., Fedorov, V., Leonov, S., 2001. Extracting information from the variance function: optimal design. In: Atkinson, A., Hackl, P., Müller, W. (Eds.), mODa6 – Advances in Model–Oriented Design and Analysis. Physica Verlag, Heidelberg, pp. 45–52.

Dragalin, V., Fedorov, V., 2006. Adaptive designs for dose-finding based on efficacy-toxicity response. Journal of Statistical Planning and Inference 136, 1800–1823.

Fedorov, V., 1972. Theory of Optimal Experiments. Academic Press, New York.

Fedorov, V., 1980. Convex design theory. Math. Operationsforsch. Statist., Ser. Statistics 11 (3), 403–413.

Ford, I., Silvey, S., 1980. A sequentially constructed design for estimating a nonlinear parametric function. Biometrika 67 (2), 381–388.

Ford, I., Titterington, D., Wu, C., 1985. Inference and sequential design. Biometrika 72 (3), 545–551.

Hu, I., 1998. On sequential designs in nonlinear problems. Biometrika 85 (2), 496–503.

Jennrich, R., 1969. Asymptotic properties of nonlinear least squares estimation. Annals of Math. Stat. 40, 633–643.

Kiefer, J., Wolfowitz, J., 1960. The equivalence of two extremum problems. Canadian Journal of Mathematics 12, 363–366.

Lai, T., 1994. Asymptotic properties of nonlinear least squares estimates in stochastic regression models. Annals of Statistics 22 (4), 1917–1930.

Lai, T., Wei, C., 1982. Least squares estimates in stochastic regression models with applications to identification and control of dynamic systems. Annals of Statistics 10 (1), 154–166.

Melas, V., 1978. Optimal designs for exponential regressions. Math. Operationsforsch. und Statist., Ser. Statistics 9, 753–768.

Müller, W., Pötscher, B., 1992. Batch sequential design for a nonlinear estimation problem. In: Fedorov, V., Müller, W., Vuchkov, I. (Eds.), Model-Oriented Data Analysis II, Proceedings 2nd IIASA Workshop, St Kyrik (Bulgaria), May 1990. Physica Verlag, Heidelberg, pp. 77–87.

Pázman, A., Pronzato, L., 2004. Simultaneous choice of design and estimator in nonlinear regression with parameterized variance. In: Di Bucchianico, A., Läuter, H., Wynn, H. (Eds.), mODa'7 – Advances in Model–Oriented Design and Analysis, Proceedings of the 7th Int. Workshop, Heeze (Netherlands). Physica Verlag, Heidelberg, pp. 117–124.

Pázman, A., Pronzato, L., 2007. Quantile and probability-level criteria for nonlinear experimental design. In: López-Fidalgo, J., Rodríguez-Díaz, J., Torsney, B. (Eds.), mODa'8 – Advances in Model–Oriented Design and Analysis, Proceedings of the 8th Int. Workshop, Almagro (Spain). Physica Verlag, Heidelberg, pp. 157–164.

Pronzato, L., Walter, E., 1985. Robust experiment design via stochastic approximation. Mathematical Biosciences 75, 103–120.

Pronzato, L., Walter, E., 1988. Robust experiment design via maximin optimization. Mathematical Biosciences 89, 161–176.

Rosenberger, W., Flournoy, N., Durham, S., 1997. Asymptotic normality of maximum likelihood estimators for multiparameter response-driven designs. Journal of Statistical Planning and Inference 60, 69–76.

Shiryaev, A., 1996. Probability. Springer, Berlin.

Wu, C., 1981. Asymptotic theory of nonlinear least squares estimation. Annals of Statistics 9 (3), 501–513.

Wu, C., 1985. Asymptotic inference from sequential design in a nonlinear situation. Biometrika 72 (3), 553–558.

Wynn, H., 1970. The sequential generation of $D$-optimum experimental designs. Annals of Math. Stat. 41, 1655–1664.