

LABORATOIRE



INFORMATIQUE, SIGNAUX ET SYSTÈMES  
DE SOPHIA ANTIPOLIS  
UMR 6070

# PENALIZED AND RESPONSE-ADAPTIVE OPTIMAL DESIGNS WITH APPLICATION TO DOSE-FINDING

*Luc Pronzato*

*Equipe SYSTEMES*

Rapport de recherche  
ISRN I3S/RR-2008-16-FR

Septembre 2008

---

RÉSUMÉ :

On considère la planification adaptative d'une expérience optimale avec une contrainte de coût, en régression nonlinéaire et pour un modèle de type Bernoulli, avec une application aux essais cliniques. La convergence forte et la normalité asymptotique des estimateurs est démontrée pour une planification sur un domaine expérimental fini, lorsque le coût moyen est imposé (et le plan d'expérience converge vers le plan optimum sous contrainte de coût) et lorsque l'objectif est de minimiser le coût. Un exemple avec un modèle binaire bivarié est présenté.

MOTS CLÉS :

Plans d'expériences adaptatifs; Plans d'expériences optimaux; Planification d'expériences pénalisée; Planification d'expériences avec contraintes; Essais cliniques; Modèle binaire bivarié; Convergence; Normalité asymptotique

---

ABSTRACT:

Adaptive optimal design with a cost constraint is considered, both for nonlinear regression and Bernoulli type experiments, with application in clinical trials. The strong consistency and asymptotic normality of the estimators is proved for designs over a finite space, both when the cost level is fixed, and the adaptive design converges to an optimum constrained design, and when the objective is to minimize the cost. An example with a bivariate binary model is given.

KEY WORDS :

Adaptive design; Optimal experimental design; Penalized experimental design; Constrained experimental design; Dose finding; Bivariate binary model; Consistency; Asymptotic normality

# Penalized and response-adaptive optimal designs with application to dose-finding\*

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`

September 8, 2008

**Abstract** Adaptive optimal design with a cost constraint is considered, both for nonlinear regression and Bernoulli type experiments, with application in clinical trials. The strong consistency and asymptotic normality of the estimators is proved for designs over a finite space, both when the cost level is fixed, and the adaptive design converges to an optimum constrained design, and when the objective is to minimize the cost. An example with a bivariate binary model is given.

**Key words:** Adaptive design; Optimal experimental design; Penalized experimental design; Constrained experimental design; Dose finding; Bivariate binary model; Consistency; Asymptotic normality

---

\*This work was partly accomplished while the author was invited at the Isaac Newton Institute for Mathematical Sciences, Cambridge, UK. The support of the Newton Institute and of CNRS are gratefully acknowledged.

# 1 Introduction and motivations

In the papers [5, 6] the authors make a stimulating step towards a clear formalization of the necessary compromise between individual and collective ethics in experimental design for dose-finding studies. The idea is to use a penalty function that accounts for poor efficacy and for toxicity, and to construct a penalized  $D$ -optimal design (or cost-constrained  $D$ -optimal design, the cost being defined through the penalty function, see [9, Chap. 4]). Using a parametric model for the dose/efficacy-toxicity responses (Gumbel or Cox model as in [5] or a bivariate probit model as in [6]), the Fisher information matrix can be calculated and optimal designs can be obtained through (now standard) algorithmic methods. This shows a neat advantage over more standard approaches. Indeed,  $D$ -optimal design alone (and its extensions to Bayesian, minimax and adaptive variants) favors collective ethics in the sense that future patients will benefit from the information collected in the trials, but it neglects individual ethics in the sense that the patients enrolled in the trials may receive doses with low efficacy or high toxicity. In contrast, dose-finding approaches based on up-and-down [20] or bandit methods [14] focus on individual ethics by determining the best dose associated with some predefined criterion (e.g., the probability of efficacy and no toxicity) but may result in a poor learning from the trial and thus in poorly adapted dosage regimen for future individuals. By a suitable tuning of the penalty function, the approach used in [5, 6] makes a clear compromise between the efficient treatment of individuals in the trial (by preventing the use of doses with low efficacy or high toxicity) and the precise estimation of the model parameters (accompanied with measures of statistical accuracy), to be used for making efficient decisions for future treatments. As such, it has a great potential in combining traditional phase I and phase II clinical trials into a single one, thereby accelerating the drug development process.

The aim of the present paper is to build on this approach and improve it in two directions.

First, the penalized optimal design problem formulated in [5, 6] does not allow flexibility in setting the balance between the information gained (in terms of precision of parameter estimation) and the cost of the experiment (in terms of poor success for the patients enrolled in the trial). Here we shall promote another formulation for optimal design under a cost constraint, for which a scalar tuning parameter

sets the compromise between information and cost. We show that, for suitable penalty functions, by increasing the weight set on the penalty one guarantees that all doses in the experiment have a small cost (and concentrate around the optimal dose when this one is unique). This permits to avoid the extreme doses generally suggested by experimental design for parameter estimation.

Second, whereas [5, 6] advocate the use of *adaptive* experimental design, the convergence of the procedure (strong consistency of the estimator of the model parameters and convergence of the design to the optimal one) is left as an open issue. This difficulty is usually overcome by considering an initial experiment (non adaptive) that grows in size when the total number of observations increases. Although this number is often severely limited in practise, we think it would be reassuring to know that *for a given initial experiment*, adaptive design guarantees suitable asymptotic convergence properties. Using simple arguments, we show that this is indeed the case when *the design space is finite*, which forms a rather natural assumption in the context of clinical trials. Our results concern penalized  $D$ -optimal design but also cover the case, more classical, of fully adaptive  $D$ -optimal design. The asymptotic distribution of the parameter estimator is shown to be normal, with variance-covariance matrix given by inverse of the usual information matrix, similarly to the non-adaptive case. Moreover, we show that, for suitable penalty functions, when the weight given to the cost for bad treatment increases with the number of patients enrolled, the doses allocated converge to the optimal one while the parameters are still estimated consistently.

The results presented are of rather general applicability and, although this work has been motivated by considerations in the context of clinical trials, we show that they also cover the case of least-squares estimation in nonlinear regression models, for which there exist even less consistency results than for maximum likelihood estimation when the design is adaptive. In particular, the results of Sect. 3 and 4 form a major improvement over those in [26] where only linear regression models were considered.

Penalized optimal design for a fixed value of the model parameters (locally optimal design) is considered in Sect. 2. We show in particular that a scalar penalty coefficient can be used to make a suitable compromise between learning (gaining information) and optimization (targeting the optimal dose), and

eventually force the design points to concentrate around the optimal one. Adaptive penalized  $D$ -optimal design is introduced at the end of this section, its asymptotic properties are investigated in the rest of the paper under the assumption that the design space is finite. Sect. 3 concerns the case where the penalty coefficient is bounded. We show that when both the design and the coefficient are adapted to the current estimated value of the model parameters, one guarantees the strong consistency and asymptotic normality of the parameter estimator and the strong convergence of the design to the optimal one. Both least-squares estimation in nonlinear regression and maximum-likelihood estimation in Bernoulli-type experiments are considered. In Sect. 4, the penalty coefficient is allowed to grow to infinity. We show that when the increase is not too fast, the strong consistency and asymptotic normality of the parameter estimator are preserved, while at the same time the design asymptotically concentrates around points of minimum cost. Some simulation results with a bivariate dose-response model are presented in Sect. 5. Finally, Sect. 6 concludes and suggests several directions for further developments. The proofs of lemmas and theorems are collected in an Appendix.

## 2 Penalized $D$ -optimal design

### 2.1 Models, designs and constraints

Let  $\mathcal{X}$ , a compact subset of  $\mathbb{R}^d$ , denote the admissible domain for the experimental variables  $x$  (design points) and  $\theta \in \mathbb{R}^p$  denote the ( $p$ -dimensional) vector of parameter of interest in a parametric model generating the log-likelihood  $l(Y, x; \theta)$  for the observation  $Y$  at the design point  $x$ . We shall always suppose that  $\theta \in \Theta$ , a compact subset of  $\mathbb{R}^p$ . For  $N$  independent observations  $\mathbf{Y} = (Y_1, \dots, Y_N)$  at  $\mathbf{X} = (x_1, \dots, x_N)$  the log-likelihood at  $\theta$  is  $l(\mathbf{Y}, \mathbf{X}; \theta) = \sum_{i=1}^N l(Y_i, x_i; \theta)$ . Let  $\mathbf{M}(X, \theta)$  denote the corresponding Fisher information matrix,

$$\mathbf{M}(X, \theta) = -\mathbb{E}_\theta \left\{ \frac{\partial^2 l(\mathbf{Y}, \mathbf{X}; \theta)}{\partial \theta \partial \theta^\top} \right\} = \sum_{i=1}^N \mu(x_i, \theta).$$

When  $N(x_i)$  denotes the number of observations made at  $x = x_i$ , we get the following normalized information matrix per observation

$$\mathbf{M}(\xi, \theta) = \frac{1}{N} \mathbf{M}(X, \theta) = \sum_{i=1}^K \frac{N(x_i)}{N} \mu(x_i, \theta),$$

where  $K$  is the number of distinct design points and  $\xi$  is the design measure (a probability measure on  $\mathcal{X}$ ) that puts mass  $N(x_i)/N$  at  $x_i$ . Following the usual approximate design approach, we shall relax the constraints on design measures and consider  $\xi$  as any element of  $\Xi$ , the set of probability measures on  $\mathcal{X}$ , so that

$$\mathbf{M}(\xi, \theta) = \int_{\mathcal{X}} \mu(x, \theta) \xi(dx).$$

In a regression model with independent and homoscedastic observations satisfying  $\mathbb{E}_{\theta}(Y|x, \theta) = \eta(x, \theta)$ , with  $\eta(x, \theta)$  differentiable with respect to  $\theta$  for any  $x$ , we have

$$\mu(x, \theta) = \mathcal{I} \frac{\partial \eta(x, \theta)}{\partial \theta} \frac{\partial \eta(x, \theta)}{\partial \theta^{\top}} \quad (1)$$

with  $\mathcal{I} = \int [\varphi'(t)/\varphi(t)]^2 \varphi(t) dt$  the Fisher information for location, where  $\varphi(\cdot)$  is the probability density function of the observation errors and  $\varphi'(\cdot)$  its derivative.

In a dose-response problem with single response  $Y \in \{0, 1\}$  (efficacy or toxicity response at the dose  $x$  for instance) and  $\text{Prob}\{Y = 1|x, \theta\} = \pi(x, \theta)$  we have

$$l(Y, x; \theta) = Y \log[\pi(x, \theta)] + (1 - Y) \log[1 - \pi(x, \theta)] \quad (2)$$

so that, assuming  $\pi(x, \theta)$  differentiable with respect to  $\theta$  for any  $x$ ,

$$\mu(x, \theta) = \frac{\partial \pi(x, \theta)}{\partial \theta} \frac{\partial \pi(x, \theta)}{\partial \theta^{\top}} \frac{1}{\pi(x, \theta)[1 - \pi(x, \theta)]}. \quad (3)$$

Bivariate extensions, where both efficacy and toxicity responses are observed at a dose  $x$ , are considered in [5] (Gumbel and Cox models) and [6] (bivariate probit model). See also Example 2 below and Sect. 5. Besides a few additional technical difficulties, the main difference with the single response case is the fact that  $\mu(x, \theta)$  may have rank larger than one, so that less than  $p$  support points in  $\xi$  may suffice to estimate  $\theta$  consistently. The same situation occurs for regression models when  $\dim(\eta) > 1$  so that (1) may have rank larger than one. We shall always assume that  $\mu(x, \theta)$  is bounded on  $\mathcal{X}$ .

In its now traditional form, local optimal design consists in determining a measure  $\xi^*$  that maximizes a concave function  $\Psi(\cdot)$  of the Fisher information matrix  $\mathbf{M}(\xi, \theta)$  for a given value of  $\theta$ . We assume that  $\Psi(\cdot)$  is monotone for the Loewner ordering (therefore,  $\Psi(a\mathbf{M})$  is a non-decreasing function of  $a \in \mathbb{R}^+$  for any non-negative definite matrix  $\mathbf{M}$ ) and shall pay special attention to local  $D$ -optimal design, for which  $\Psi(\mathbf{M}) = \log \det \mathbf{M}$ . The extension to other global optimality criteria, such as  $[\text{trace}(\mathbf{M}^{-1})]^{-1}$  for instance, can be obtained by following a similar route. The denomination ‘local’ in local optimal design comes from the fact that in nonlinear situations  $\mathbf{M}(\xi, \theta)$  depends on  $\theta$  and the optimal design thus depends on the value  $\theta$  that is chosen. Extensions to minimax and Bayesian (or average optimal designs) are also considered in [6] (see also [23, 8, 29, 30, 1]). A natural and widely used approach to face the problem of dependency into the unknown value of  $\theta$  is adaptive design. In its simplest form (one-step adaptive locally optimal) the design points  $x_1, x_2, \dots, x_N, x_{N+1}, \dots$  associated with a sequence of observations are chosen sequentially, the determination of the point  $x_{N+1}$  being based on the value  $\hat{\theta}^N$  estimated from the  $N$  previous observations. The asymptotic properties of estimators and designs obtained via this procedure will be investigated in Sect. 3 and 4.

In many circumstances, besides the optimality criterion  $\psi(\xi) = \Psi[\mathbf{M}(\xi, \theta)]$ , it is desirable to introduce a constraint of the form  $\Phi(\xi, \theta) \leq C$  for the design measure. In dose-finding problems, the introduction of such a constraint allows one to take individual ethical concerns into account. For instance, when both the efficacy and toxicity responses are observed, one can relate  $\Phi(\xi, \theta)$  to the probability of success (efficacy and no toxicity) for a given dose, as done in [5, 6]. See also Example 2. We suppose that the cost function  $\Phi(\xi, \theta)$  is linear in  $\xi$ , that is

$$\Phi(\xi, \theta) = \int_{\mathcal{X}} \phi(x, \theta) \xi(dx),$$

and that  $\phi(x, \theta)$  is bounded on  $\mathcal{X}$ . The extension to nonlinear constraints is considered, e.g., in [4] and [9, Chap. 4]. Also, we shall restrict our attention to the case where a single — scalar — constraint is present, some of the issues caused by the presence of several constraints are addressed in the same references. See also Sect. 6. Matrices are denoted by bold capital letters and we denote by  $\|\mathbf{A}\|$  the usual norm of  $\mathbf{A}$ ,  $\|\mathbf{A}\| = [\text{trace}(\mathbf{A}^\top \mathbf{A})]^{1/2} = (\sum_{i,j} \{\mathbf{A}\}_{i,j}^2)^{1/2}$ .

## 2.2 Two equivalent formulations for maximizing information per cost-unit

Suppose that  $\phi(x, \theta) > 0$  for all  $x \in \mathcal{X}$ . The approach used in [5, 6] formulates the problem as follows.

**Problem  $P_1(\theta)$ :** maximize the total information for  $N$  observations, that is, maximize  $\Psi[N\mathbf{M}(\xi, \theta)]$  with respect to  $N$  and  $\xi \in \Xi$  satisfying the total cost constraint

$$N\Phi(\xi, \theta) = N \int_{\mathcal{X}} \phi(x, \theta) \xi(dx) \leq C.$$

For any  $\xi$ , the optimal value of  $N$  is  $N^*(\xi) = C/\Phi(\xi, \theta)$  so that an optimal measure  $\xi^* \in \Xi$  for  $P_1(\theta)$  maximizes  $\Psi(C\mathbf{M}(\xi, \theta)/\Phi(\xi, \theta))$ .

Denote now  $\nu = N\xi$ , which is a measure on  $\mathcal{X}$  not normalized to 1; we have  $\int_{\mathcal{X}} \nu(dx) = N$  which becomes a free variable.  $P_1(\theta)$  is then equivalent to the maximization of  $\Psi[\mathbf{M}(\nu, \theta)]$  under the constraint  $\Phi(\nu, \theta) \leq C$ . The constraint is saturated at the optimum, i.e.  $\Phi(\nu^*, \theta) = C$ , which we can thus set as an active constraint. Imposing  $\Phi(\nu, \theta) = C$  and defining  $\xi'(dx) = \nu(dx)\phi(x, \theta)/C$  we obtain  $\int_{\mathcal{X}} \xi'(dx) = 1$  and

$$\mathbf{M}(\nu, \theta) = \int_{\mathcal{X}} \mu(x, \theta) \nu(dx) = \int_{\mathcal{X}} C \frac{\mu(x, \theta)}{\phi(x, \theta)} \xi'(dx) = \mathbf{M}'(\xi', \theta).$$

The constraint design problem  $P_1(\theta)$  is thus equivalent to a standard unconstrained one, with  $\mu(x, \theta)$  simply replaced by  $C\mu(x, \theta)/\phi(x, \theta)$ . Call  $P_2(\theta)$  this problem.

**Problem  $P_2(\theta)$ :** maximize  $\Psi[\mathbf{M}'(\xi, \theta)]$  with respect to  $\xi \in \Xi$ .

The equivalence between  $P_1(\theta)$  and  $P_2(\theta)$  is further evidenced by considering the necessary and sufficient conditions for optimality expressed by the Equivalence Theorem, see [19] for  $D$ -optimality and, e.g., [31] for a general formulation. For  $P_1(\theta)$  with  $\psi(\xi) = \log \det[\mathbf{M}(\xi, \theta)/\Phi(\xi, \theta)]$ , the measure  $\xi^*$  is optimal if and only if

$$\forall x \in \mathcal{X}, \text{ trace}[\mu(x, \theta)\mathbf{M}^{-1}(\xi^*, \theta)] \leq p \frac{\phi(x, \theta)}{\Phi(\xi^*, \theta)} \quad (4)$$

(note that the condition does not depend on the normalization constant  $\int_{\mathcal{X}} \xi^*(dx)$ ). For  $P_2(\theta)$  with

$\psi(\xi') = \log \det \mathbf{M}'(\xi', \theta)$ ,  $\xi'^*$  is optimal in  $\Xi$  if and only if

$$\forall x \in \mathcal{X}, C \text{ trace} \left[ \frac{\mu(x, \theta)}{\phi(x, \theta)} \mathbf{M}'^{-1}(\xi'^*, \theta) \right] \leq p \quad (5)$$

(note that  $\mathbf{M}'$  is proportional to  $C$  which thus cancels out). The two conditions are equivalent: (5) can be written as (4) by setting  $\xi^*(dx) = C \xi'^*(dx) / \phi(x, \theta)$ .

One should note that the value of  $C$  plays no role in the definition of optimal designs for  $P_1(\theta)$ ,  $P_2(\theta)$ . For the dose-response problem considered in [5, 6] this has the important consequence that the prohibition of excessively low (with poor efficacy) or high (with high toxicity) doses can only be obtained by an ad-hoc modification of the penalty function  $\phi(x, \theta)$ . Indeed, this is the only way to modify the optimal design and hopefully to change its support. This can be contrasted with the solution of the constrained design problem that we present in the next section and then consider in all the rest of the paper.

### 2.3 Maximizing information per observation under a cost constraint

A direct formulation of the optimal design problem under constraint is as follows.

**Problem  $P_3(\theta)$ :** maximize  $\Psi[\mathbf{M}(\xi, \theta)]$  with respect to  $\xi \in \Xi$  under the constraint  $\Phi(\xi, \theta) \leq C$ .

We say that a design measure  $\xi \in \Xi$  is  $\theta$ -admissible if  $\Phi(\xi, \theta) \leq C$  and we suppose that a strictly  $\theta$ -admissible measure exists in  $\Xi$  ( $\Phi(\xi, \theta) < C$  for some  $\xi \in \Xi$ ). A necessary and sufficient condition for the optimality of a  $\theta$ -admissible  $\xi^* \in \Xi$  is the existence of  $\lambda^* \geq 0$  such that  $\lambda^* [C - \Phi(\xi^*, \theta)] = 0$  with  $\xi^* = \xi^*(\lambda^*)$  maximizing the design criterion  $\mathcal{L}_\theta(\xi, \lambda^*) = \Psi[\mathbf{M}(\xi, \theta)] + \lambda^* [C - \Phi(\xi, \theta)]$  (the Lagrangian of the problem) with respect to  $\xi \in \Xi$ . Moreover, the Lagrange coefficient  $\lambda^*$  minimizes  $\mathcal{L}_\theta[\xi^*(\lambda), \lambda]$  with respect to  $\lambda \in \mathbb{R}^+$ . When  $\Psi(\cdot) = \log \det(\cdot)$ , the necessary and sufficient condition for the optimality of a  $\theta$ -admissible  $\xi^* \in \Xi$  becomes

$$\exists \lambda^* \geq 0 \text{ such that } \begin{cases} \lambda^* [C - \Phi(\xi^*, \theta)] = 0 \\ \forall x \in \mathcal{X}, \text{ trace}[\mu(x, \theta) \mathbf{M}^{-1}(\xi^*, \theta)] \leq p + \lambda^* [\phi(x, \theta) - \Phi(\xi^*, \theta)]. \end{cases} \quad (6)$$

In practice,  $\xi^*$  can be determined by maximizing

$$H_\theta(\xi, \lambda) = \Psi[\mathbf{M}(\xi, \theta)] - \lambda \Phi(\xi, \theta) \quad (7)$$

for an increasing sequence  $(\lambda_i)$  of Lagrange coefficients  $\lambda$ , starting at  $\lambda_0 = 0$  and stopping at the first  $\lambda_i$  such that the associated optimal design  $\xi^*$  satisfies  $\Phi(\xi^*, \theta) \leq C$ , see, e.g., [24] (for  $C$  large enough, the unconstrained optimal design for  $\Psi(\cdot)$  is optimal for the constrained problem). The parameter  $\lambda$  can thus be used to set the tradeoff between the maximization of  $\Psi[\mathbf{M}(\xi, \theta)]$  (gaining information) and minimization of  $\Phi(\xi, \theta)$  (reducing cost). Notice that maximizing  $H_\theta(\xi, \lambda)$  for  $\lambda \geq 0$  is equivalent to maximizing  $(1 - \gamma) \Psi[\mathbf{M}(\xi, \theta)] + \gamma[-\Phi(\xi, \theta)]$  with  $\gamma = \lambda/(1 + \lambda) \in [0, 1)$ . Also, when  $\Psi(\mathbf{M}) = \log \det \mathbf{M}$ , there is an obvious relation between the maximization of (7) and the solution of a problem in the form  $P_1(\theta)$ . Indeed, one can write  $H_\theta(\xi, \lambda) = \log \det[\mathbf{M}(\xi, \theta)/\Phi'(\xi, \theta)]$  with  $\Phi'(\xi, \theta) = \exp[(\lambda/p)\Phi(\xi, \theta)]$  (which, however, is not linear in  $\xi$ ).

Similarly to the case of unconstrained optimal design with a strictly concave criterion, the optimal matrix  $\mathbf{M}(\xi^*, \theta)$  is unique when the function  $\Psi(\cdot)$  is strictly concave on the set of positive definite matrices. Indeed, suppose that there exist two optimal designs  $\xi_1^*, \xi_2^*$  in  $\Xi$  for  $P_3(\theta)$  such that  $\mathbf{M}(\xi_1^*, \theta) \neq \mathbf{M}(\xi_2^*, \theta)$ . The optimality of  $\xi_1^*$  and  $\xi_2^*$  implies  $\Psi[\mathbf{M}(\xi_1^*, \theta)] = \Psi[\mathbf{M}(\xi_2^*, \theta)]$  and  $\Phi(\xi_1^*, \theta) = \Phi(\xi_2^*, \theta) = C$ . Therefore the Lagrange multipliers  $\lambda$  of both solutions coincide, see (6), and  $\xi_1^*, \xi_2^*$  maximize (7) for this  $\lambda$ , that is,  $\max_{\xi \in \Xi} H_\theta(\xi, \lambda) = H_\theta(\xi_1^*, \lambda) = H_\theta(\xi_2^*, \lambda)$ . Take now any  $\alpha$  in  $(0, 1)$  and consider  $\xi = (1 - \alpha)\xi_1^* + \alpha\xi_2^* \in \Xi$ . From the strict concavity of  $\Psi(\cdot)$  and the linearity in  $\xi$  of  $\Phi(\xi, \theta)$ , we obtain  $H_\theta(\xi, \lambda) > (1 - \alpha)H_\theta(\xi_1^*, \lambda) + \alpha H_\theta(\xi_2^*, \lambda) = H_\theta(\xi_1^*, \lambda) = H_\theta(\xi_2^*, \lambda)$ , which contradicts the optimality of  $\xi_1^*, \xi_2^*$ . The optimal information matrix is thus unique (but the optimal design measure  $\xi^*$  is not necessarily).

Let  $\xi^*(\lambda)$  denote an optimal design for  $H_\theta(\xi, \lambda)$  given by (7). One can easily check that both  $\Psi\{\mathbf{M}[\xi^*(\lambda), \theta]\}$  and  $\Phi[\xi^*(\lambda), \theta]$  are non-increasing functions of  $\lambda$ . Three questions (at least) naturally arise concerning the tradeoff between maximization of  $\Psi[\mathbf{M}(\xi, \theta)]$  and minimization of  $\Phi(\xi, \theta)$ .

- (i) How fast does the cost  $\Phi(\xi, \theta)$  decrease when  $\lambda$  increases?
- (ii) How big is the loss of information (decrease of  $\Psi[\mathbf{M}(\xi, \theta)]$ ) when  $\Phi(\xi, \theta)$  decreases?

- (iii) Can we force the costs  $\phi(\hat{x}_i, \theta)$  to decrease for all support points  $\hat{x}_i$  of the optimal design  $\xi^*(\lambda)$  by increasing  $\lambda$ ?

Suppose that  $\mu(x, \theta)$  and  $\phi(x, \theta)$  are continuous in  $x \in \mathcal{X}$ , with  $\mathcal{X}$  a compact subset of  $\mathbb{R}^d$ , and define

$$\phi_\theta^* = \min_{x \in \mathcal{X}} \phi(x, \theta), \quad x^* = x^*(\theta) = \arg \min_{x \in \mathcal{X}} \phi(x, \theta). \quad (8)$$

We do not assume for the moment that  $x^*$  is unique (but at least one such point exists in  $\mathcal{X}$ ). For any  $\xi \in \Xi$ , we also define

$$\Delta_\theta(\xi) = \Phi(\xi, \theta) - \phi_\theta^*.$$

Only the case of  $D$ -optimality is considered and we denote by  $\xi_D^*$  a  $D$ -optimal design that maximizes  $\log \det \mathbf{M}(\xi, \theta)$  with respect to  $\xi \in \Xi$ . We assume that  $\Delta_\theta(\xi_D^*) > 0$  (otherwise  $\xi_D^*$  maximizes  $\log \det \mathbf{M}(\xi, \theta) - \lambda \Phi(\xi, \theta)$  for any  $\lambda \geq 0$ ) and that  $\log \det \mathbf{M}(\xi_D^*, \theta) > 0$  (otherwise  $\mathbf{M}(\xi, \theta)$  is singular for any  $\xi \in \Xi$ ). We then have the following results concerning the three questions above. The proof is given in Appendix.

**Proposition 1** *Let  $\xi^* = \xi^*(\lambda)$  be an optimal design that maximises  $H_\theta(\xi, \lambda)$  given by (7) with respect to  $\xi \in \Xi$ , with  $\Psi(\mathbf{M}) = \log \det \mathbf{M}$ . It satisfies*

(i)  $\Delta_\theta(\xi^*) \leq p/\lambda$ ;

(ii) for any  $\xi$  such that  $\Delta_\theta(\xi) > 0$ , any  $a > 0$  and any  $\lambda \geq a/\Delta_\theta(\xi)$ ,

$$\log \det \mathbf{M}(\xi^*, \theta) \geq \log \det \mathbf{M}(\xi, \theta) + p \log \{a/[\lambda \Delta_\theta(\xi)]\} - a, \quad (9)$$

moreover,  $\lambda_{\min}[\mathbf{M}(\xi^*, \theta)] > \delta/(p + \lambda[\bar{\phi}_\theta - \phi_\theta^*])$  for some positive constant  $\delta$ , with  $\bar{\phi}_\theta = \max_{x \in \mathcal{X}} \phi(x, \theta)$ ;

(iii) any support point  $\hat{x}$  of  $\xi^*$  satisfies

$$\phi(\hat{x}, \theta) - \phi_\theta^* \leq 2\Delta_\theta(\xi_\lambda) \text{trace}[\mu(\hat{x}, \theta)\mathbf{M}^{-1}(\xi_\lambda, \theta)], \quad \forall \xi_\lambda \in \Xi \text{ such that } \Delta_\theta(\xi_\lambda) \geq p/\lambda. \quad (10)$$

Property (i) shows the guaranteed cost-reduction obtained when  $\lambda$  is increased and (ii) shows that  $\lambda_{\min}[\mathbf{M}(\xi^*, \theta)]$  decreases not faster than  $1/\lambda$ . A similar property will be obtained in Sect. 4 for adaptive design with an increasing sequence of penalty coefficients  $\lambda_k$ ; it is a central argument for obtaining

the strong consistency and asymptotic normality of estimators, see Theorem 4 and Corollary 2. Notice that taking  $\xi = \xi_D^*$  in (9) ensures  $\det \mathbf{M}(\xi^*, \theta) \geq \det[\mathbf{M}(\xi_D^*, \theta) \exp(-a/p)]$  for  $\lambda = a/\Delta_\theta(\xi_D^*)$  and  $\log \det \mathbf{M}(\xi^*, \theta) \geq \log \det \mathbf{M}(\xi_D^*, \theta) + p \log[\Delta_\theta(\xi^*)/\Delta_\theta(\xi_D^*)]$  for any  $\lambda \geq 0$  (take  $a = \lambda \Delta_\theta(\xi^*)$ ).

Property (iii) shows that, for suitable penalty functions, the support of an optimal design for  $P_3(\theta)$  depends on  $C$  or, equivalently, that the support of an optimal design for (7) depends on  $\lambda$ . When  $x^*$  is unique, (iii) implies that if

there exist designs  $\xi_\lambda \in \Xi$  such that  $\Delta_\theta(\xi_\lambda) \geq p/\lambda$  and

$$\forall \epsilon > 0, \limsup_{\lambda \rightarrow \infty} \sup_{\|x - x^*\| > \epsilon} \frac{2\Delta_\theta(\xi_\lambda) \text{trace}[\mu(x, \theta)\mathbf{M}^{-1}[\xi_\lambda, \theta]]}{\phi(x, \theta) - \phi_\theta^*} < 1, \quad (11)$$

then the supporting points of  $\xi^*$  converge to  $x^*$  as  $\lambda \rightarrow \infty$ . The choice of suitable designs  $\xi_\lambda$  is central for testing if (11) is satisfied. Designs formed by points neighboring  $x^*$  are good candidates, see Examples 1 and 2. Notice that, when  $\text{rank}[\mu(x, \theta)] < p$ , for (11) to be satisfied the  $\xi_\lambda$ 's must necessarily have support points that approach  $x^*$  as  $\lambda \rightarrow \infty$ . Indeed, suppose that it is not the case. It means that there exists  $\gamma > 0$  such that, for all  $\lambda$  larger than some  $\lambda_0$ , the support points  $x_\lambda^{(i)}$  of  $\xi_\lambda$  satisfy  $\|x_\lambda^{(i)} - x^*\| > \gamma$ . Replace  $\mathcal{X}$  by  $\mathcal{X}' = \mathcal{X} \setminus \mathcal{B}(x^*, \gamma) \cup \{x^*\}$ , that is, remove the ball  $\mathcal{B}(x^*, \gamma) = \{x : \|x - x^*\| \leq \gamma\}$  from  $\mathcal{X}$  but keep  $x^*$ . Then,  $\xi_\lambda$  is a design measure on  $\mathcal{X}'$  for  $\lambda > \lambda_0$  and (11) would indicate that the optimal design  $\xi^*$  on  $\mathcal{X}^*$  is the delta measure  $\delta_{x^*}^*$ , which is impossible since the optimal information matrix must have full rank. The same is true if the designs  $\xi_\lambda$  have one support point at  $x^*$  and the others outside the ball  $\mathcal{B}(x^*, \gamma)$  for  $\lambda$  larger than some  $\lambda_0$ . Finally, note that the support points of  $\xi^*(\lambda')$  for  $\lambda' > \lambda$  must also satisfy (10) for the same  $\xi_\lambda$  (since  $\Delta_\theta(\xi_\lambda) > p/\lambda'$ ).

For dose-response problems, the property (11) has the important consequence that excessively high or low doses can be prohibited by choosing  $C$  small enough or, equivalently,  $\lambda$  large enough. Its effectiveness very much depends on the choice of the penalty function, and in particular on its local behavior around  $x^*$ . Contrary to what intuition might suggest, it requires the cost function  $\phi(\cdot, \theta)$  to be sufficiently flat around  $x^*$ . Indeed, in that case a design  $\xi$  supported in the neighborhood of  $x^*$  can at the same time have a small cost  $\Phi(\xi, \theta)$  and be dispersed enough to carry significant information through  $\log \det \mathbf{M}(\xi, \theta)$ . This

is illustrated by examples below. The first one is simple enough to make the optimal designs calculable explicitly.

## 2.4 Examples

### Example 1

We take

$$\mu(x) = \begin{pmatrix} 1 \\ x \\ x^2 \end{pmatrix} \begin{pmatrix} 1 & x & x^2 \end{pmatrix} = \begin{pmatrix} 1 & x & x^2 \\ x & x^2 & x^3 \\ x^2 & x^3 & x^4 \end{pmatrix},$$

$\mathcal{X} = [-1, 1]$ ,  $\Psi(\cdot) = \log \det(\cdot)$  and consider different penalty functions  $\phi(x)$ , symmetric with respect to  $x = 0$  (nothing depends on  $\theta$  in this example and we thus omit the dependence in  $\theta$  in the notation). For all cases considered, the optimal designs  $\xi^*$  are symmetric and take the form

$$\xi^* = \begin{Bmatrix} -z & 0 & z \\ \frac{1-\alpha}{2} & \alpha & \frac{1-\alpha}{2} \end{Bmatrix}, \quad (12)$$

where the first row gives the support points and the second their respective weights. This gives  $\det \mathbf{M}(\xi^*) = \alpha(1-\alpha)^2 z^6$ . The  $D$ -optimal design  $\xi_D^*$  corresponds to  $z = 1$  and  $\alpha = 1/3$ .

For  $\phi(x) = 1 + x^{2q}$ ,  $q$  integer, the optimal designs for problems  $P_1$  and  $P_2$  correspond to  $\alpha = \min\{q/[3(q-1)], 1/2\}$  and  $z = \min\{[3/(2q-3)]^{1/2q}, 1\}$  (note that  $z < 1$  only for  $q \geq 4$ ). The costs  $\phi(x) = 1 + x^2 + x^4$  and  $\phi(x) = 1/(1-x^2)$  respectively give the optimal designs defined by  $\alpha = 3/5$ ,  $z = 1$  and  $\alpha = 5/9$ ,  $z = \sqrt{3/5}$ .

The optimal designs  $\xi^*$  for  $P_3$  obtained for various choices for  $\phi(\cdot)$  are given in Table 1, together with the optimal value  $\lambda^*(C)$  of the Lagrange coefficient associated with  $C$ . When there is no solution, it means that  $\lambda^*(C) = \infty$ . When there is one, then  $\Phi(\xi^*) = C$ .

In order to check if the support of  $\xi^*$  concentrates around  $x^* = 0$  when  $\lambda$  increases (without computing

$\xi^*$ ), we use (10) with the design

$$\xi_\lambda = \xi_\lambda(\gamma) = \begin{Bmatrix} -\gamma & 0 & \gamma \\ 1/3 & 1/3 & 1/3 \end{Bmatrix}.$$

For  $\phi(x) = x^{2q}$  we get  $\Delta(\xi_\lambda) = 2\gamma^{2q}/3$  and (10) then gives  $\hat{x}^{2q} \leq [4\gamma^{2q}/3] \text{trace}[\mu(\hat{x})\mathbf{M}^{-1}(\xi_\lambda)]$ . Noticing that  $\text{trace}[\mu(\gamma t)\mathbf{M}^{-1}(\xi_\lambda)] = P(t) = 3(1 - 3/2t^2 + 3/2t^4)$  independently of  $\gamma$  (a property of  $D$ -optimal design for polynomial regression), we obtain that a support point  $\hat{x}$  of  $\xi^*$  must satisfy  $t^{2q} \leq 4P(t)/3$ , with  $t = \hat{x}/\gamma$ . For  $q = 3$  we obtain  $t \leq [1 + (1 + \beta^{1/3})^2]^{1/2}/\beta^{1/6}$  with  $\beta = 4 + 2\sqrt{2}$ , that is  $t \lesssim 2.2252$ . For  $q = 4$ , we get  $t \leq \sqrt{2}$ . Since we need to have  $\Delta(\xi_\lambda) \geq p/\lambda = 3/\lambda$ , we take  $\gamma = [9/(2\lambda)]^{1/(2q)}$  (which corresponds to  $\alpha = 1$  and  $\tilde{\xi}_\lambda = \xi_\lambda$  in the proof of Proposition 1-(ii)). It gives  $\hat{x} \leq \hat{x}_{\max}$  with  $\hat{x}_{\max} \simeq 2.860\lambda^{-1/6}$  for  $q = 3$  and  $\hat{x}_{\max} = \sqrt{2}[9/(2\lambda)]^{1/8} \simeq 1.707\lambda^{-1/8}$  for  $q = 4$ . When  $q = 1, 2$  all  $t$  are admissible and  $\hat{x}_{\max} = \infty$ . We obtain in a similar way  $\hat{x}_{\max} = \infty$  for  $\phi(x) = 1 + x^2 + x^4$ .

Next case illustrates that it is the local behavior of  $\phi(x)$  around the minimum  $x^*$  that influences the support of  $\xi^*$  when  $\lambda$  tends to infinity. Take  $\phi(x) = 1/(1 - x^2)$ , which tends to infinity for  $x$  tending to  $\pm 1$  but is equal to  $1 + x^2 + x^4 + \mathcal{O}(x^6)$  around  $x^* = 0$ . Condition (10) then writes  $\phi(\gamma t) - \phi(0) = \gamma^2 t^2 / (1 - \gamma^2 t^2) \leq 2\Delta(\xi_\lambda)P(t) = (4/3)\gamma^2 P(t)/(1 - \gamma^2)$ . The bound obtained for  $t$  now depends on  $\gamma$ ; the best bound (minimum) for  $\hat{x}$  is  $\hat{x}_{\max} \simeq 0.9649$ , obtained at  $\gamma \simeq 0.7385$ , and  $\Delta(\xi_\lambda) \geq p/\lambda$  imposes  $\lambda \gtrsim 3.7516$ . Therefore, we only learn from (10) that the support of  $\xi^*$  is included in  $[-0.9649, 0.9649]$  for  $\lambda$  large enough. This is consistent with the behavior of the support points  $-z, z$  of  $\xi^*$  as  $\lambda$  tends to infinity, which do not converge to zero ( $\lim_{\lambda \rightarrow \infty} z = 1/\sqrt{3}$ , see Table 1).  $\square$

Next example is taken from [5] and concerns a problem with bivariate binary responses.

**Example 2** : *Cox model for efficacy-toxicity response.*

For  $Y$  (respectively  $Z$ ) the binary indicator of efficiency (resp. of toxicity) at dose  $x$  for a model with parameters  $\theta$ , we write  $\text{Prob}\{Y = y, Z = z|x, \theta\} = \pi_{yz}(x, \theta)$ ,  $Y, y, Z, z \in \{0, 1\}$ , with

$$\begin{aligned} \pi_{11}(x, \theta) &= \frac{e^{a_{11} + b_{11}x}}{1 + e^{a_{01} + b_{01}x} + e^{a_{10} + b_{10}x} + e^{a_{11} + b_{11}x}} \\ \pi_{10}(x, \theta) &= \frac{e^{a_{10} + b_{10}x}}{1 + e^{a_{01} + b_{01}x} + e^{a_{10} + b_{10}x} + e^{a_{11} + b_{11}x}} \end{aligned}$$

$\phi(x)$	$C$	$\lambda^*(C)$	$\alpha$	$z$
$1 + x^2$	$5/3 \leq C$	0	1/3	1
	$1 < C \leq 5/3$	$\frac{5-3C}{(C-1)(2-C)}$	$2 - C$	1
	$C \leq 1$	$\infty$	— no solution —	
$1 + x^4$	$5/3 \leq C$	0	1/3	1
	$4/3 \leq C \leq 5/3$	$\frac{5-3C}{(C-1)(2-C)}$	$2 - C$	1
	$1 < C \leq 4/3$	$3/[2(C-1)]$	2/3	$[3(C-1)]^{1/4}$
	$C \leq 1$	$\infty$	— no solution —	
$1 + x^6$	$5/3 \leq C$	0	1/3	1
	$3/2 \leq C \leq 5/3$	$\frac{5-3C}{(C-1)(2-C)}$	$2 - C$	1
	$1 < C \leq 3/2$	$1/(C-1)$	1/2	$[2(C-1)]^{1/6}$
	$C \leq 1$	$\infty$	— no solution —	
$1 + x^8$	$5/3 \leq C$	0	1/3	1
	$14/9 \leq C \leq 5/3$	$\frac{5-3C}{(C-1)(2-C)}$	$2 - C$	1
	$1 < C \leq 14/9$	$3/[4(C-1)]$	4/9	$[9(C-1)/5]^{1/8}$
	$C \leq 1$	$\infty$	— no solution —	
$1 + x^2 + x^4$	$7/3 \leq C$	0	1/3	1
	$1 < C \leq 7/3$	$\frac{7-3C}{(C-1)(3-C)}$	$(3-C)/2$	1
	$C \leq 1$	$\infty$	— no solution —	
$1/(1-x^2)$	$1 < C$	$\frac{2}{C(C-1)}$	$\frac{C}{3C-2}$	$\frac{(3C-2)^{1/2}}{(3C)^{1/2}}$
	$C \leq 1$	$\infty$	— no solution —	

Table 1: Optimal designs  $\xi^*$  for problem  $P_3$  in Example 1, see (12).

$$\begin{aligned}\pi_{01}(x, \theta) &= \frac{e^{a_{01}+b_{01}x}}{1 + e^{a_{01}+b_{01}x} + e^{a_{10}+b_{10}x} + e^{a_{11}+b_{11}x}} \\ \pi_{00}(x, \theta) &= (1 + e^{a_{01}+b_{01}x} + e^{a_{10}+b_{10}x} + e^{a_{11}+b_{11}x})^{-1}\end{aligned}$$

and  $\theta = (a_{11}, b_{11}, a_{10}, b_{10}, a_{01}, b_{01})^\top$ . The log-likelihood function of a single observation  $(Y, Z)$  at dose  $x$  is then  $l(Y, Z, x; \theta) = YZ \log \pi_{11}(x, \theta) + Y(1 - Z) \log \pi_{10}(x, \theta) + (1 - Y)Z \log \pi_{01}(x, \theta) + (1 - Y)(1 - Z) \log \pi_{00}(x, \theta)$  and elementary calculations show that the contribution to the Fisher information matrix is

$$\mu(x, \theta) = \frac{\partial \mathbf{p}^\top(x, \theta)}{\partial \theta} (\mathbf{P}^{-1}(x, \theta) + [1 - \pi_{11}(x, \theta) - \pi_{10}(x, \theta) - \pi_{01}(x, \theta)]^{-1} \mathbf{1} \mathbf{1}^\top) \frac{\partial \mathbf{p}(x, \theta)}{\partial \theta^\top}$$

where  $\mathbf{p}(x, \theta) = [\pi_{11}(x, \theta), \pi_{10}(x, \theta), \pi_{01}(x, \theta)]^\top$ ,  $\mathbf{P}(x, \theta) = \text{diag}\{\mathbf{p}(x, \theta)\}$  and  $\mathbf{1} = (1, 1, 1)^\top$ . Note that  $\mu(x, \theta)$  is generally of rank 3. As in [5], we take  $\theta = (3, 3, 4, 2, 0, 1)^\top$  and  $\mathcal{X}$  the finite set  $\{x^{(1)}, \dots, x^{(11)}\}$  where the doses  $x^{(i)}$  are equally spaced in the interval  $[-3, 3]$ . We choose a cost function related to the probability  $\pi_{10}(x, \theta)$  of efficacy and no toxicity and take

$$\phi(x, \theta) = \{\pi_{10}^{-1}(x, \theta) - [\max_x \pi_{10}(x, \theta)]^{-1}\}^2. \quad (13)$$

The Optimal Safe Dose (OSD) minimizing  $\phi(x, \theta)$ , is  $x^{(5)} = -0.6$ . Using the condition (10) with the test designs  $\xi_{\lambda,1}$ , giving weights  $(1 - \alpha)/2, \alpha$  and  $(1 - \alpha)/2$  at  $x^{(4)}, x^{(5)}$  and  $x^{(6)}$  respectively, and  $\xi_{\lambda,2}$ , giving weights  $1 - \beta$  and  $\beta$  at  $x^{(4)}$  and  $x^{(5)}$  respectively, we obtain that the support of  $\xi^*(\lambda)$  is included in  $\{x^{(4)}, x^{(5)}, x^{(6)}\}$  for  $\alpha \gtrsim 0.9993$  and  $\beta \gtrsim 0.4508$ , showing that the optimal designs concentrate on three doses around the optimal one when  $\lambda$  is large enough. The  $D$ -optimal design (corresponding to  $\lambda = 0$ ), is supported on  $x^{(1)}, x^{(4)}, x^{(5)}$  and  $x^{(10)}$ , with associated weights 0.3318, 0.3721, 0.1259 and 0.1701. Figure 1 presents the optimal designs  $\xi^*(\lambda)$  for  $\lambda$  varying between 0 and 100 along the horizontal axis. The weight associated with each  $x^{(i)}$  on the vertical axis is proportional to the thickness of the plot.

The figure indicates that for  $\lambda \gtrsim 75$  the optimum designs are supported on  $x^{(4)}$  and  $x^{(6)}$  only, with weights approximately 1/2 each, that is, all patients in the trial receive a dose close to the optimal one,  $x^{(5)}$ . However, none receives the optimal dose. The situations changes for larger values of  $\lambda$ , see Figure 2 where the optimal design is supported on  $\{x^{(4)}, x^{(5)}, x^{(6)}\}$  for  $\lambda \gtrsim 160$  and the weight of the optimal dose  $x^{(5)}$  increases with  $\lambda$ . Consider now the cost function  $\phi(x, \theta) = \pi_{10}^{-1}(x, \theta)$ , which is less flat than

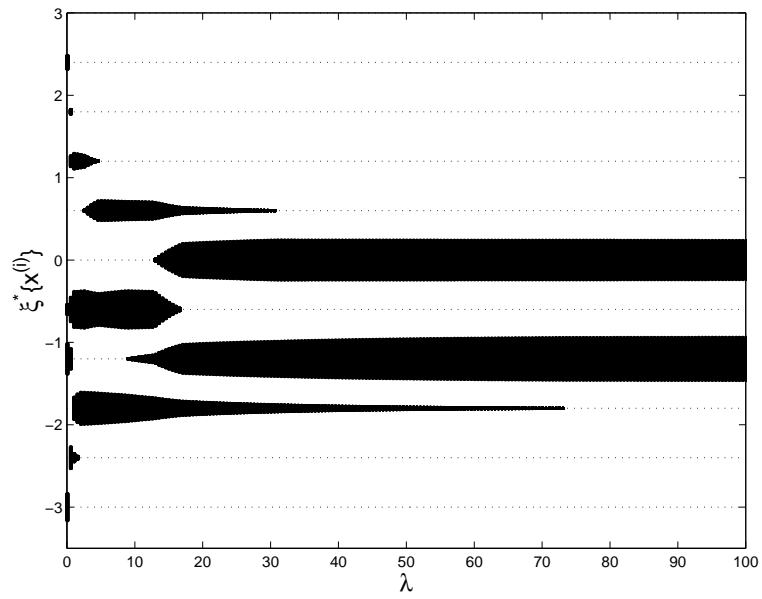


Figure 1: Optimal designs  $\xi^*(\lambda)$  as function of  $\lambda \in [0, 100]$  in Example 2 with the cost-function (13): each horizontal dotted line corresponds to a point in  $\mathcal{X}$ , the thickness of the plot indicates the associated weight.

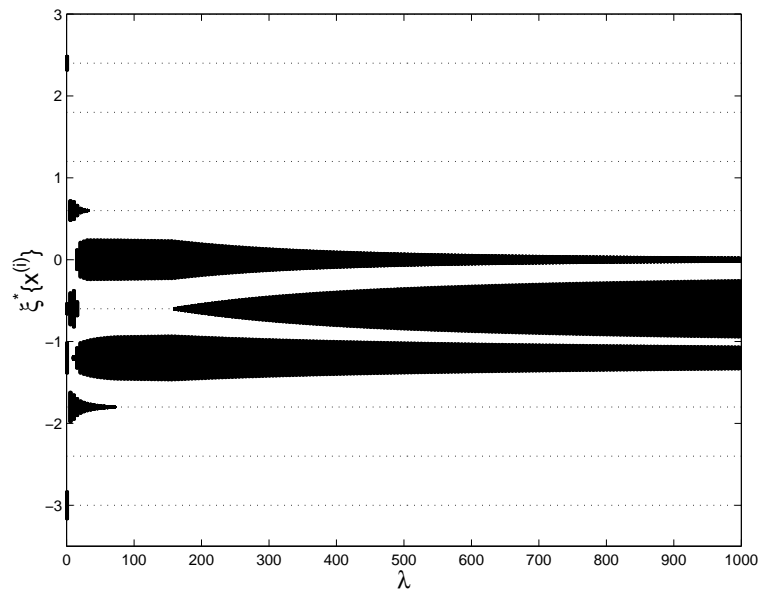


Figure 2: Same as Figure 1, but for  $\lambda \in [0, 1000]$ .

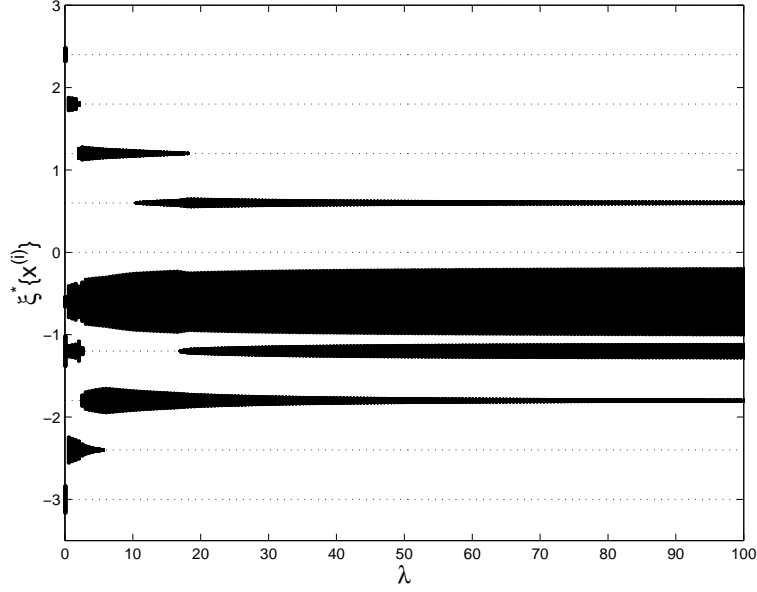


Figure 3: Same as Figure 1, but for the cost function  $\pi_{10}^{-1}(x, \theta)$ .

(13) around the optimum. The support points are then less concentrated around the optimal dose  $x^{(5)}$ , compare Figure 3 with Figure 1.  $\square$

In a nonlinear situation, like in Example 2, where  $\mathbf{M}(\xi, \theta)$  and  $\Phi(\xi, \theta)$  depend on  $\theta$ , robustness with respect to misspecifications of  $\theta$  can be achieved by considering average-optimal design. Problem  $P_3(\theta)$  is then transformed into: maximize  $\mathbb{E}_\theta\{\Psi[\mathbf{M}(\xi, \theta)]\}$  with respect to  $\xi \in \Xi$  under the constraint  $\mathbb{E}_\theta\{\Phi(\xi, \theta)\} \leq C$ , where the expectation  $\mathbb{E}_\theta$  is calculated for some prior probability measure  $\pi$  for  $\theta$ . For  $D$ -optimality, the optimality condition (6) becomes

$$\exists \lambda^* \geq 0 \text{ such that } \begin{cases} \lambda^* [C - \mathbb{E}_\theta\{\Phi(\xi^*, \theta)\}] = 0 \\ \forall x \in \mathcal{X}, \mathbb{E}_\theta\{\text{trace}[\mu(x, \theta)\mathbf{M}^{-1}(\xi^*, \theta)]\} \leq p + \lambda^* \mathbb{E}_\theta\{\phi(x, \theta) - \Phi(\xi^*, \theta)\}. \end{cases} \quad (14)$$

Apart from additional numerical cost (which remains reasonable when  $\pi$  is a discrete measure with a limited number of support points), the introduction of a prior probability for  $\theta$  does not raise any special difficulty. This is used in [13], with  $\phi(x, \theta) = \mathbb{I}_{[Q_R(\theta), \infty)}(x)$ , where  $\mathbb{I}_{\mathcal{A}}(x)$  is the indicator function of the set  $\mathcal{A}$  (1 if  $x \in \mathcal{A}$ , 0 otherwise) and  $Q_R(\theta)$  is a quantile of the probability of toxicity, parameterized by  $\theta$ ,

corresponding to the maximum acceptable probability of toxicity (note that  $\mathbb{E}_\theta\{\phi(x, \theta)\} = \pi\{Q_R(\theta) \leq x\}$ , the prior probability that  $x$  exceeds  $Q_R$ ).

Another, rather common, approach for facing the issue of dependence of the optimum design in  $\theta$  is to design the experiment sequentially. By alternating between estimation based on previous observations and determination of the next design point where to observe, one may hope that the empirical design measure will progressively adapt to the correct (true) value of the model parameters. Adaptive design is briefly introduced in the next section, the asymptotic properties of designs and estimators obtained in this way are considered into details in the rest of the paper. An example (continuation of Example 2) is presented in Sect. 5.

## 2.5 Adaptive penalized $D$ -optimal design

In fully-adaptive  $D$ -optimal design, next design point after  $N$  observations is taken as

$$x_{N+1} = \arg \max_{x \in \mathcal{X}} \text{trace}[\mu(x, \hat{\theta}^N) \mathbf{M}^{-1}(\xi_N, \hat{\theta}^N)], \quad (15)$$

where  $\hat{\theta}^N \in \Theta \subset \mathbb{R}^p$  is the current estimated value for  $\theta$  (based on  $x_1, Y_1, \dots, x_N, Y_N$ ) and  $\xi_N = (1/N) \sum_{i=1}^N \delta_{x_i}$  (with  $\delta_z$  the delta measure that puts mass 1 at  $z$ ) is the current empirical design measure (leaving aside some initialisation issues: we simply assume that  $x_1, \dots, x_p$  are such that  $\mathbf{M}(\xi_p, \theta)$  is nonsingular for any  $\theta \in \Theta$ ). Note that (15) can only be considered as an algorithm for choosing design points, in the sense that  $\mathbf{M}(\xi_N, \theta)$  is not the Fisher information matrix for parameters  $\theta$  due to the sequential construction of the design (we shall see, however, in Sect. 3 that when  $\mathcal{X}$  is finite, from the same repeated sampling principle as in [36], one can still use  $\mathbf{M}(\xi_N, \theta)$  to characterize the precision of the estimation of  $\theta$  as  $N \rightarrow \infty$ ).

For constrained  $D$ -optimal design we take  $x_{N+1}$  as

$$x_{N+1} = \arg \max_{x \in \mathcal{X}} \left\{ \text{trace}[\mu(x, \hat{\theta}^N) \mathbf{M}^{-1}(\xi_N, \hat{\theta}^N)] - \lambda_N \phi(x, \hat{\theta}^N) \right\}. \quad (16)$$

Since (15) can be considered as a special case of (16), only the latter will be considered in the following.

This means that the results in Sect. 3 also cover the case of classical (unconstrained) sequential  $D$ -

optimal design (15) for which  $\lambda_N = 0$  for all  $N$ . This also includes the constrained problems  $P_1(\theta)$ ,  $P_2(\theta)$  considered in [5, 6], which can be formulated as standard  $D$ -optimal design problems, see Sect. 2.2. One can notice the similarity between (16) and the construction used in [26] for optimizing a parametric function, the parameters of which being estimated by least-squares in a linear regression model.

When  $\hat{\theta}^N$  is frozen to a fixed value  $\theta$  and  $\lambda_N$  is constant, the iterations (15) and (16) correspond to one step of a steepest-ascent vertex-direction algorithm with step-length  $1/N$  at step  $N$ . Convergence to an optimal design measure is proved in [38] for iterations given by (15) and in [26] for (16) (using a general argument developed in [37]).

The fact that  $\hat{\theta}^N$  is estimated in adaptive design makes the proof of convergence a much more complicated issue for which few results are available: [10, 36, 25] concern a particular example with least-squares estimation; [16] is specific of Bayesian estimation by posterior mean and does not use a fully sequential design of the form (15); [21] and [3] require the introduction of a subsequence of non-adaptive design points to ensure consistency of the estimator and [2] requires that the size of the initial experiment (non-adaptive) grows with the increase in size of the total experiment. Intuitively, the almost sure convergence of  $\hat{\theta}^N$  to some  $\hat{\theta}^\infty$  would be enough to imply the convergence of  $\xi_N$  to an optimal design measure for  $\hat{\theta}^\infty$  (this will be shown in Theorem 3) and, conversely, convergence of  $\xi_N$  to a design  $\xi_\infty$  such that  $\mathbf{M}(\xi_\infty, \theta)$  is non-singular for any  $\theta$  would be enough in general to make an estimator consistent. It is thus the interplay between estimation and design iterations (which implies that each design point depends on previous observations) that creates difficulties. As shown in the next sections, those difficulties disappear when  $\mathcal{X}$  is a finite set (notice that the assumption that  $\mathcal{X}$  is finite is seldom limitative since practical considerations often impose such a restriction on possible choices for the design points; this can be contrasted with the much less natural assumption that would consist in considering the feasible parameter set as finite).

Two situations will be considered concerning the choice of the sequence  $(\lambda_N)$  in (16), respectively in Sect. 3 and 4. In the first one, the objective is to obtain an optimal design for problem  $P_3(\theta)$ . We shall then adapt  $\lambda_N$  to  $\hat{\theta}^N$  and take  $\lambda_N = \lambda_N^* = \lambda^*(\hat{\theta}^N)$ , the optimal Lagrange coefficient for  $P_3(\hat{\theta}^N)$ . The

second situation corresponds to the case where  $(\lambda_N)$  forms an increasing sequence, which gives more and more importance to the constraint in the construction of the design. When  $\phi(x, \theta)$  has a single minimum, by letting the Lagrange coefficient  $\lambda_N$  increase with  $N$  one may hope to be able to force the design to concentrate at the minimizer of  $\phi$  associated with the true value of  $\theta$  (that is, for clinical trials, to focus more and more on individual ethics by allocating treatments with increasing efficiency).

The results presented in the next sections rely on simple arguments based on three ideas. First, the sequence  $(\hat{\theta}^N)$  in (16) is taken as *any sequence* of vectors in  $\Theta$ . The asymptotic design properties obtained within this framework will thus also apply when  $\hat{\theta}^N$  corresponds to some estimator of  $\theta$  in  $\Theta$ . Second, when  $\mathcal{X}$  is finite we obtain a lower bound on the sampling rate of a subset of points of  $\mathcal{X}$  associated with a nonsingular information matrix. Third, we show that this bound guarantees the strong consistency of the estimator of  $\theta$ , both for least-squares estimation in nonlinear regression and maximum-likelihood estimation for Bernoulli trials. With a few additional technicalities, this yields almost sure convergence results for the adaptive designs constructed via (16).

### 3 Asymptotic properties of adaptive design with bounded penalty

The results in this section and the next one apply to a wide range of situations and we try to keep the presentation general enough. However, to avoid unnecessary complications we only treat the univariate case where  $\mu(x, \theta)$  has rank one, and write

$$\mu(x, \theta) = \mathbf{f}_\theta(x) \mathbf{f}_\theta^\top(x)$$

with  $\mathbf{f}_\theta(x)$  a  $p$ -dimensional vector, as it is the case in (1, 3). The extension to the multivariate case does not raise particular difficulties. We shall use the following assumptions on the design space  $\mathcal{X}$ , vector  $\mathbf{f}(x, \theta)$ , constraint function  $\phi(x, \theta)$  and Lagrange coefficients  $\lambda_N$ .

**H <sub>$\mathcal{X}$</sub>** -(i): The design space  $\mathcal{X}$  is finite,  $\mathcal{X} = \{x^{(1)}, x^{(2)}, \dots, x^{(K)}\}$ .

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

**H <sub>$\phi$</sub>** -(i):  $0 \leq \phi(x, \theta) < \bar{\phi}$ ,  $\forall x \in \mathcal{X}$  and  $\theta \in \Theta$ .

$\mathbf{H}_\lambda\text{-(i)}$ :  $0 \leq \lambda_N < \bar{\lambda} < \infty, \quad \forall N$ .

When  $\lambda_N = \lambda_N^* = \lambda^*(\hat{\theta}^N)$ , the optimal Lagrange coefficient for problem  $P_3(\hat{\theta}^N)$ , and  $\hat{\theta}^N \in \Theta$ , the following condition guarantees that  $\mathbf{H}_\lambda\text{-(i)}$  is satisfied.

$\mathbf{H}_\lambda\text{-(i')}$ : There exists  $C' < C$  such that  $\forall \theta \in \Theta, \exists \hat{\xi}(\theta) \in \Xi$  with  $\Phi[\hat{\xi}(\theta), \theta] < C'$  and  $\mathbf{M}[\hat{\xi}(\theta), \theta]$  has full rank.

We first obtain a lower bound on the sampling rate of nonsingular designs, which will be the cornerstone for proving the consistency and asymptotic normality of estimators. The proof is given in Appendix.

**Lemma 1** *Let  $(\hat{\theta}^N)$  be an arbitrary sequence in  $\Theta$  used to generate design points according to (16), with an initialisation such that  $\mathbf{M}(\xi_N, \theta)$  is non-singular for all  $\theta$  in  $\Theta$  and all  $N \geq p$ . Let  $r_{N,i} = r_N(x^{(i)})$  denote the number of times  $x^{(i)}$  appears in the sequence  $x_1, \dots, x_N$ ,  $i = 1, \dots, K$ , and consider the associated order statistics  $r_{N,1:K} \geq r_{N,2:K} \geq \dots \geq r_{N,K:K}$ . Define*

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

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

### 3.1 Consistency of estimators

#### 3.1.1 Least-squares estimation in nonlinear regression

Consider a regression model with observations

$$Y_i = Y(x_i) = \eta(x_i, \bar{\theta}) + \varepsilon_i, \tag{17}$$

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  $\mathbb{E}\{\varepsilon_1^2\} = \sigma^2 < \infty$  (with  $\sigma = 1$  without loss of generality). We assume that the model response  $\eta(x, \theta)$  is differentiable with respect to  $\theta \in \text{int}(\Theta)$  for any  $x \in \mathcal{X}$ .

Denote

$$S_N(\theta) = \sum_{k=1}^N [Y(x_k) - \eta(x_k, \theta)]^2 \quad (18)$$

and let  $\hat{\theta}_{LS}^N = \arg \min_{\theta \in \Theta} S_N(\theta)$  be the least-squares (LS) estimator of  $\theta$ . The contribution of the design point  $x$  to the information matrix is then  $\mu(x, \theta) = \mathbf{f}_\theta(x) \mathbf{f}_\theta^\top(x)$  with

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

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, see (1).

Define

$$D_N(\theta, \bar{\theta}) = \sum_{k=1}^N [\eta(x_k, \theta) - \eta(x_k, \bar{\theta})]^2. \quad (19)$$

Next theorem shows that the consistency of the LS estimator is a consequence of  $D_N(\theta, \bar{\theta})$  tending to infinity fast enough for  $\|\theta - \bar{\theta}\| \geq \delta > 0$ . The fact that the design space  $\mathcal{X}$  is finite makes the minimum rate of increase of  $D_N(\theta, \bar{\theta})$  required for consistency quite slow. The result is valid whether the  $x_k$ 's are non-random constants or are generated via a sequential design algorithm such as (16).

**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 (19) satisfies*

$$\text{for all } \delta > 0, \left[ \inf_{\|\theta - \bar{\theta}\| \geq \delta} D_N(\theta, \bar{\theta}) \right] / (\log \log N) \rightarrow \infty, \quad N \rightarrow \infty, \quad (20)$$

*then  $\hat{\theta}_{LS}^N \xrightarrow{\text{a.s.}} \bar{\theta}$  as  $N \rightarrow \infty$  (almost sure convergence). The result remains valid for  $(x_i)$  a random sequence on  $\mathcal{X}$  finite when (20) holds almost surely.*

The proof is given in [27] and is based on Lemma 1 in [35]. Note that the condition (20) is much less restrictive than the classical one for strong consistency of LS estimation in nonlinear regression ( $D_N(\theta, \bar{\theta}) = \mathcal{O}(N)$  for  $\theta \neq \bar{\theta}$ ), see [18]. It is also less restrictive than the condition obtained in [21] for sequential design.

Consider the following identifiability assumption on the regression model (17).

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

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

For any sequence  $(\hat{\theta}^N)$  used in (16), the conditions of Lemma 1 ensure the existence of  $N_1$  and  $\alpha > 0$  such that  $r_{N,j;K} > \alpha N$  for all  $N > N_1$  and all  $j = 1, \dots, p$ . Under the additional assumption  $\mathbf{H}_{\mathcal{X}}$ -(iii) we thus obtain that  $D_N(\theta, \bar{\theta})$  given by (19) satisfies  $\inf_{\|\theta - \bar{\theta}\| \geq \delta} D_N(\theta, \bar{\theta}) > \alpha N \epsilon(\delta)$ ,  $N > N_1$ . Therefore,  $\hat{\theta}_{LS}^N \xrightarrow{\text{a.s.}} \bar{\theta}$  ( $N \rightarrow \infty$ ) from Theorem 1. Since this holds for any sequence  $(\hat{\theta}^N)$  in  $\Theta$ , it is true in particular when  $\hat{\theta}_{LS}^N$  is substituted for  $\hat{\theta}^N$  in (16).

### 3.1.2 Maximum-likelihood estimation in Bernoulli trials

Consider now the case of dose-response experiments with

$$Y \in \{0, 1\}, \quad \text{with } \text{Prob}\{Y = 1 | x_i, \theta\} = \pi(x_i, \theta). \quad (21)$$

We suppose that  $\Theta$  is a compact subset of  $\mathbb{R}^p$ , that  $\bar{\theta}$ , the ‘true’ value of  $\theta$  that generates the observations, lies in the interior of  $\Theta$ , and that  $\pi(x, \theta) \in (0, 1)$  for any  $\theta \in \Theta$  and  $x \in \mathcal{X}$ . We also assume that  $\pi(x, \theta)$  is differentiable with respect to  $\theta \in \text{int}(\Theta)$  for any  $x \in \mathcal{X}$ .

The log-likelihood for the observation  $Y$  at the design point  $x$  is given by (2) and the contribution of the point  $x$  to the Fisher information matrix is (3), which we can write  $\mu(x, \theta) = \mathbf{f}_\theta(x) \mathbf{f}_\theta^\top(x)$  with

$$\mathbf{f}_\theta(x) = \frac{1}{\sqrt{\pi(x, \theta)[1 - \pi(x, \theta)]}} \frac{\partial \pi(x, \theta)}{\partial \theta}. \quad (22)$$

Let  $\hat{\theta}_{ML}^N$  denote the Maximum-Likelihood (ML) estimator of  $\theta$ ,  $\hat{\theta}_{ML}^N = \arg \max_{\theta \in \Theta} L_N(\theta)$ , with  $L_N(\theta) = \sum_{i=1}^N l(Y_i, x_i; \theta)$ , see (2). It satisfies the following.

**Lemma 2** *If for any  $\delta > 0$*

$$\liminf_{N \rightarrow \infty} \inf_{\|\theta - \bar{\theta}\| \geq \delta} [L_N(\bar{\theta}) - L_N(\theta)] > 0 \quad \text{almost surely,}$$

*then  $\hat{\theta}_{ML}^N \xrightarrow{\text{a.s.}} \bar{\theta}$  as  $N \rightarrow \infty$ .*

The proof is identical to that of Lemma 1 in [35]. We then obtain a property similar to Theorem 1, see [27].

**Theorem 2** *Let  $(x_i)$  be a non-random design sequence on a finite set  $\mathcal{X}$ . Assume that*

$$D_N(\theta, \bar{\theta}) = \sum_{i=1}^N \pi(x_i, \bar{\theta}) \log \left[ \frac{\pi(x_i, \bar{\theta})}{\pi(x_i, \theta)} \right] + [1 - \pi(x_i, \bar{\theta})] \log \left[ \frac{1 - \pi(x_i, \bar{\theta})}{1 - \pi(x_i, \theta)} \right] \quad (23)$$

*satisfies (20). Then,  $\hat{\theta}_{ML}^N \xrightarrow{\text{a.s.}} \bar{\theta}$  as  $N \rightarrow \infty$  in the model (21). The same is true for  $(x_i)$  a random sequence such that (20) holds almost surely.*

Consider the following identifiability assumption for the Bernoulli model.

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

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

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  $a \in (0, 1)$ ,  $g(a, b) > 2(a - b)^2$  with  $g(a, a) = 0$ , so that each term of the sum (23) is positive. Also,  $\mathbf{H}_{\mathcal{X}}$ -(iii)' implies that

$$\inf_{\|\theta - \bar{\theta}\| \geq \delta} \sum_{j=1}^p \pi(x^{(i_j)}, \bar{\theta}) \log \left[ \frac{\pi(x^{(i_j)}, \bar{\theta})}{\pi(x^{(i_j)}, \theta)} \right] + [1 - \pi(x^{(i_j)}, \bar{\theta})] \log \left[ \frac{1 - \pi(x^{(i_j)}, \bar{\theta})}{1 - \pi(x^{(i_j)}, \theta)} \right] > \epsilon(\delta) > 0.$$

for any  $\delta > 0$  and any subset  $\{i_1, \dots, i_p\}$  of  $\{1, \dots, K\}$ . Similarly to the case of LS estimation in nonlinear regression, but using now Theorem 2 instead of Theorem 1, we thus obtain that under the conditions of Lemma 1 and with the additional assumption  $\mathbf{H}_{\mathcal{X}}$ -(iii)' the ML estimator satisfies  $\hat{\theta}_{ML}^N \xrightarrow{\text{a.s.}} \bar{\theta}$ ,  $N \rightarrow \infty$ , when  $\hat{\theta}_{ML}^N$  is substituted for  $\hat{\theta}^N$  in (16).

### 3.2 Asymptotic optimality of adaptive penalized $D$ -optimal design

We consider the adaptive design algorithm (16) with  $\lambda_N = \lambda_N^* = \lambda^*(\hat{\theta}^N)$ , the optimal Lagrange coefficient for  $P_3(\hat{\theta}^N)$ . Define

$$H_{\bar{\theta}}^* = \max_{\xi \in \Xi} \left\{ \log \det \mathbf{M}(\xi, \bar{\theta}) + \lambda^*(\bar{\theta}) [C - \Phi(\xi, \bar{\theta})] \right\}. \quad (24)$$

By asymptotic optimality we mean that the empirical design measure  $\xi_N$  is such that  $\hat{\theta}^N \xrightarrow{\text{a.s.}} \bar{\theta}$  and

$$H_{\bar{\theta}}(\xi_N) = \log \det \mathbf{M}(\xi_N, \bar{\theta}) + \lambda^*(\bar{\theta}) [C - \Phi(\xi_N, \bar{\theta})] \xrightarrow{\text{a.s.}} H_{\bar{\theta}}^*, \quad N \rightarrow \infty. \quad (25)$$

Since  $\mathbf{M}(\xi^*, \bar{\theta})$  is unique, see Sect. 2.3, (25) is equivalent to  $\mathbf{M}(\xi_N, \bar{\theta}) \xrightarrow{\text{a.s.}} \mathbf{M}[\xi^*(\bar{\theta}), \bar{\theta}]$ ,  $N \rightarrow \infty$ , that is,  $\xi_N$  tends to be optimal for  $P_3(\bar{\theta})$ . We state this property below as a theorem (the proof is given in Appendix). The following assumptions are used.

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

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

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

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

**Theorem 3** *Suppose that in the regression model (17) (respectively, in the Bernoulli model (21)) the design points for  $N > p$  are generated sequentially according to (16), where  $\lambda_N = \lambda^*(\hat{\theta}^N)$ , the optimal Lagrange coefficient for  $P_3(\hat{\theta}^N)$ , with  $\hat{\theta}^N = \hat{\theta}_{LS}^N$  (respectively,  $\hat{\theta}^N = \hat{\theta}_{ML}^N$ ). Suppose, moreover, that the first  $p$  design points are such that the information matrix is nonsingular for any  $\theta \in \Theta$ . Then, under  $H_{\mathcal{X}}$ -(i),  $H_{\mathcal{X}}$ -(ii),  $H_{\mathcal{X}}$ -(iii) (respectively,  $H_{\mathcal{X}}$ -(iii')),  $H_{\mathcal{X}}$ -(iv),  $H_{\lambda}$ -(i'),  $H_{\phi}$ -(i),  $H_{\phi}$ -(ii) and  $H_f$ -(i) we have  $\hat{\theta}_{LS}^N \xrightarrow{\text{a.s.}} \bar{\theta}$  (respectively,  $\hat{\theta}_{ML}^N \xrightarrow{\text{a.s.}} \bar{\theta}$ ) and  $\mathbf{M}(\xi_N, \bar{\theta}) \xrightarrow{\text{a.s.}} \mathbf{M}[\xi^*(\bar{\theta}), \bar{\theta}]$ ,  $N \rightarrow \infty$ , with  $\xi^*(\bar{\theta})$  an optimal design for  $P_3(\bar{\theta})$ .*

One may notice that Theorem 3 also applies in the case where the sequence  $(\lambda_N)$  is not adapted to  $\hat{\theta}^N$  but is simply controlled so as to satisfy a suitable compromise between designing for precise estimation of  $\theta$  and cost-minimization, and satisfies  $\lambda_N \xrightarrow{\text{a.s.}} \lambda$ ,  $N \rightarrow \infty$ , for some  $\lambda$ .

### 3.3 Asymptotic normality of estimators

Under a fixed design (penalized  $D$ -optimal for instance) the information matrix can be considered as a large sample approximation for the variance-covariance matrix of the estimator, thus allowing straightforward statistical inference from the trial. The situation is more complicated for adaptive designs and has

been intensively discussed in the literature. Intuitively, the usual asymptotic normality of  $\hat{\theta}^N$  ( $N \rightarrow \infty$ ) should hold when the sequence  $(x_N)$  is such that  $\hat{\theta}^N$  is strongly consistent and  $\mathbf{M}(\xi_N, \bar{\theta})$  converges to a nonsingular matrix. The theorem below shows that this is indeed the case in the present situation. One may refer e.g. to [32, 12, 11, 39] for statistical inference in dose-finding problems when using up-and-down [20] or bandit methods [14], randomized Pólya-urn [7] etc. In contrast with those approaches, a guaranteed level of precision for the estimation of the model parameters can easily be imposed by choosing the targeted cost  $C$  in Problem  $P_3(\theta)$  or the value of  $\lambda$  in (7), see Proposition 1. Our result is a corollary of the lemma below (its proof is given in Appendix), which uses the following regularity assumption for the model.

**H<sub>f</sub>-(ii):** For all  $x$  in  $\mathcal{X}$ , the components of  $\mathbf{f}_\theta(x)$  are continuously differentiable with respect to  $\theta$  in some open neighborhood of  $\bar{\theta}$ .

**Lemma 3** *Assume that H<sub>f</sub>-(ii) is satisfied, that the design points belong to a finite set, see H<sub>X</sub>-(i), and are such that*

$$\liminf_{N \rightarrow \infty} \tau_N \lambda_{\min}[\mathbf{M}(\xi_N, \bar{\theta})] > \Lambda > 0 \text{ a.s.}$$

for some sequence  $(\tau_N)$  satisfying  $\lim_{N \rightarrow \infty} \tau_N/N^{1/4} = 0$ . Then, if  $\hat{\theta}_{ML}^N \xrightarrow{\text{a.s.}} \bar{\theta}$  ( $N \rightarrow \infty$ ) in the Bernoulli model (21), we also have

$$\sqrt{N} \mathbf{M}^{1/2}(\xi_N, \hat{\theta}_{ML}^N)(\hat{\theta}_{ML}^N - \bar{\theta}) \xrightarrow{d} \omega \sim \mathcal{N}(\mathbf{0}, \mathbf{I}), \quad N \rightarrow \infty. \quad (26)$$

The same is true in the regression model (17): when  $\hat{\theta}_{LS}^N \xrightarrow{\text{a.s.}} \bar{\theta}$  we also have

$$\sqrt{N} \mathbf{M}^{1/2}(\xi_N, \hat{\theta}_{LS}^N)(\hat{\theta}_{LS}^N - \bar{\theta}) \xrightarrow{d} \omega \sim \mathcal{N}(\mathbf{0}, \mathbf{I}), \quad N \rightarrow \infty, \quad (27)$$

under the additional condition  $\lim_{N \rightarrow \infty} \tau_N N^{-\delta/(2+\delta)} = 0$  for some  $\delta$  such that  $\mathbb{E}\{|\varepsilon_1|^{2+\delta}\} < \infty$ .

Note that when  $\lim_{N \rightarrow \infty} \tau_N/N^{1/4} = 0$ , the condition  $\lim_{N \rightarrow \infty} \tau_N N^{-\delta/(2+\delta)} = 0$  for regression models is not restrictive when moments  $\mathbb{E}\{|\varepsilon_1|^\alpha\}$  exist for  $\alpha > 8/3$ . One may also notice that, compared to [35], we do not require that  $\tau_N \mathbf{M}(\xi_N, \bar{\theta})$  tends to some positive definite matrix, and compared to [22, 21] we do not require the existence of a non-random matrix  $\mathbf{C}_N$  such that  $\mathbf{C}_N \mathbf{M}^{1/2}(\xi_N, \bar{\theta}) \xrightarrow{P} \mathbf{I}$ ,  $N \rightarrow \infty$ . On

the other hand, we need that  $\lambda_{\min}[\mathbf{M}(\xi_N, \bar{\theta})]$  decreases more slowly than  $N^{-1/4}$ . The lemma applies to more general designs than (15, 16). In particular, adaptive rules on a finite design space that have a non degenerate limiting distribution  $\xi_\infty$  (such that  $\det \mathbf{M}(\xi_\infty, \theta) \neq 0$ ) satisfy the conditions of the lemma with  $\tau_N \equiv 1$ . This is the case in particular for up-and-down methods in clinical trials, see Sect. 5.

Under the conditions of Theorem 3, there exist  $N_0$  and  $\alpha > 0$  such that, for all  $N > N_0$  we have  $\lambda_{\min}[\mathbf{M}(\xi_N, \bar{\theta})] > \alpha \bar{\gamma}$ , with  $\bar{\gamma}$  as in  $H_{\mathcal{X}}\text{-}(iv)$ . We can thus take  $\tau_N \equiv 1$  in Lemma 3 and obtain the following property, which indicates that it is legitimate (asymptotically) to characterize the precision of the estimation by the inverse information matrix  $\mathbf{M}^{-1}(\xi_N, \hat{\theta}^N)$  when using the adaptive design scheme (16) on a finite design set  $\mathcal{X}$ .

**Corollary 1** *Under the conditions of Theorem 3, and assuming that, moreover,  $H_f\text{-}(ii)$  is satisfied, the ML estimator in the model (21) satisfies (26) and the LS estimator in the model (17) satisfies (27).*

## 4 Asymptotic properties of adaptive design with increasing penalty

We consider now the case where the sequence  $(\lambda_N)$  of Lagrange coefficients in (16) is unbounded and satisfies

$\mathbf{H}_\lambda\text{-}(ii)$ :  $(\lambda_N)$  is a non-decreasing positive sequence and  $\lim_{N \rightarrow \infty} \lambda_N = \infty$ .

Replacing  $H_\lambda\text{-}(i)$  by  $H_\lambda\text{-}(ii)$  in the assumptions of Sect. 3, we obtain the following lower bound on the sampling rate of nonsingular designs. The proof is given in Appendix.

**Lemma 4** *Let  $(\hat{\theta}^N)$  be an arbitrary sequence in  $\Theta$  used to generate design points according to (16), with an initialisation such that  $\mathbf{M}(\xi_N, \theta)$  is non-singular for all  $\theta$  in  $\Theta$  and all  $N \geq p$ . Let  $r_{N,j:K}$  be defined as in Lemma 1,  $j = 1, \dots, K$ , and define*

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

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

## 4.1 Consistency of estimators

As in Sect. 3.1, we consider the consistency of the LS and ML estimators in regression and dose-response experiments respectively, but now in the case where  $(\lambda_N)$  is an unbounded increasing sequence of penalty coefficients. We show that strong consistency is ensured provided that this sequence tends to infinity slowly enough.

### 4.1.1 Least-squares estimation in nonlinear regression

For any sequence  $(\hat{\theta}^N)$  used in (16), the conditions of Lemma 4 ensure the existence of  $N_1$  and  $\alpha > 0$  such that  $r_{N,j;K} > \alpha N/\lambda_N$  for all  $N > N_1$  and all  $j = 1, \dots, p$ . Under the additional assumption  $\mathbf{H}_{\mathcal{X}}\text{-(iii)}$  we thus obtain that  $D_N(\theta, \bar{\theta})$  given by (19) satisfies

$$\frac{1}{\log \log N} \inf_{\|\theta - \bar{\theta}\| \geq \delta} D_N(\theta, \bar{\theta}) > \frac{\alpha N \epsilon(\delta)}{\lambda_N \log \log N}, \quad N > N_1.$$

Therefore, if  $(\lambda_N \log \log N)/N \rightarrow 0$  when  $N \rightarrow \infty$ ,  $\hat{\theta}_{LS}^N \xrightarrow{\text{a.s.}} \bar{\theta}$  from Theorem 1. Since this holds for any sequence  $(\hat{\theta}^N)$  in  $\Theta$ , it is true in particular when  $\hat{\theta}_{LS}^N$  is substituted for  $\hat{\theta}^N$  in (16).

### 4.1.2 Maximum-likelihood estimation in Bernoulli trials

The situation is similar to previous one. Using now Theorem 2 instead of Theorem 1, we obtain that under the conditions of Lemma 4 and with the additional assumption  $\mathbf{H}_{\mathcal{X}}\text{-(iii)'}$  the ML estimator satisfies  $\hat{\theta}_{ML}^N \xrightarrow{\text{a.s.}} \bar{\theta}$ ,  $N \rightarrow \infty$ , when  $\hat{\theta}_{ML}^N$  is substituted for  $\hat{\theta}^N$  in (16) with  $(\lambda_N \log \log N)/N \rightarrow 0$  when  $N \rightarrow \infty$ .

## 4.2 Convergence to minimum-cost design and asymptotic normality

The following theorem shows that using the following assumptions

$\mathbf{H}_{\lambda}\text{-(iii)}$ : the sequence  $(\lambda_N)$  is such that  $\lambda_N/N$  is non-increasing with  $(\lambda_N \log \log N)/N \rightarrow 0$ ,  $N \rightarrow \infty$ ;

$\mathbf{H}_{\phi}\text{-(iii)}$ :  $\phi(x, \bar{\theta})$  has a unique global minimizer  $x^*$ :  $\forall \beta > 0, \exists \epsilon > 0$  such that  $\phi(x, \bar{\theta}) < \phi(x^*, \bar{\theta}) + \epsilon$  implies  $\|x - x^*\| < \beta$ ;

in complement of  $H_\lambda$ -(ii), the adaptive design algorithm (16) is such that  $(x_N)$  tends to accumulate at the point of minimum cost for  $\bar{\theta}$ . The proof is given in Appendix.

**Theorem 4** *Suppose that in the regression model (17) (respectively, in the Bernoulli model (21)) the design points for  $N > p$  are generated sequentially according to (16), where  $\lambda_N$  satisfies  $H_\lambda$ -(ii) and  $H_\lambda$ -(iii). Suppose, moreover, that the first  $p$  design points are such that the information matrix is nonsingular for any  $\theta \in \Theta$ . Then, under  $H_{\mathcal{X}}$ -(i),  $H_{\mathcal{X}}$ -(ii),  $H_{\mathcal{X}}$ -(iii) (respectively,  $H_{\mathcal{X}}$ -(iii')),  $H_{\mathcal{X}}$ -(iv),  $H_\phi$ -(i),  $H_\phi$ -(ii), and  $H_f$ -(i) we have  $\hat{\theta}_{LS}^N \xrightarrow{\text{a.s.}} \bar{\theta}$  (respectively,  $\hat{\theta}_{ML}^N \xrightarrow{\text{a.s.}} \bar{\theta}$ ) and*

$$\Phi(\xi_N, \bar{\theta}) \xrightarrow{\text{a.s.}} \phi_\theta^* = \min_{x \in \mathcal{X}} \phi(x, \bar{\theta}), \quad N \rightarrow \infty. \quad (28)$$

If, moreover,  $H_\phi$ -(iii) is satisfied, then

$$\xi_N \xrightarrow{\text{w}} \delta_{x^*} \text{ almost surely, } N \rightarrow \infty, \quad (29)$$

with  $\xrightarrow{\text{w}}$  denoting the weak convergence of probability measures and  $\delta_{x^*}$  the delta measure at  $x^* = \arg \min_{x \in \mathcal{X}} \phi(x, \bar{\theta})$ .

The property (29) does not imply that the  $x_k$ 's generated by (16) converge to  $x^*$ . However, when  $\mathcal{X}$  is obtained by the discretization of a compact set  $\mathcal{X}'$  and (11) is satisfied at  $\theta = \bar{\theta}$  for design measures on  $\mathcal{X}'$ , then the points  $x_N$  will gather around  $x^*$  as  $N \rightarrow \infty$ , see Example 2. Convergence results similar to those in Theorem 4 are obtained in [26] for LS estimation in a *linear* regression model, without the assumption that  $\mathcal{X}$  is finite, but under more restrictive conditions than  $H_\lambda$ -(ii),  $H_\lambda$ -(iii) on the growth rate of the sequence  $(\lambda_N)$ .

Under the conditions of Theorem 4, there exist  $N_0$  and  $\alpha > 0$  such that, for all  $N > N_0$ ,  $\lambda_{\min}[\mathbf{M}(\xi_N, \bar{\theta})] > \alpha \bar{\gamma} / \lambda_N$ , with  $\bar{\gamma}$  as in  $H_{\mathcal{X}}$ -(iv). We use again Lemma 3, with now  $\tau_N = \lambda_N$  and obtain the following.

**Corollary 2** *Under the conditions of Theorem 4 and assuming that, moreover,  $H_f$ -(ii) is satisfied and  $\lim_{N \rightarrow \infty} \lambda_N / N^{1/4} = 0$ , the ML estimator in the Bernoulli model (21) satisfies (26). Also, the LS estimator in the regression model (17) satisfies (27) under the additional assumption  $\lim_{N \rightarrow \infty} \lambda_N N^{-\delta/(2+\delta)} = 0$  for some  $\delta$  such that  $\mathbb{E}\{|\varepsilon_1|^{2+\delta}\} < \infty$ .*

## 5 Example

As an illustration of the behavior of the adaptive scheme (16), we continue Example 2 and present some simulation results (using the value  $\theta = (3, 3, 4, 2, 0, 2)^\top$ ). For comparison, we use the up-and-down rule of [17] (which is also considered in [5]), defined by

$$x_{N+1} = \begin{cases} \max\{x^{(i_N-1)}, x^{(1)}\} & \text{if } Z_N = 1, \\ x^{(i_N)} & \text{if } Y_N = 1 \text{ and } Z_N = 0, \\ \min\{x^{(i_N+1)}, x^{(11)}\} & \text{if } Y_N = 0 \text{ and } Z_N = 0, \end{cases} \quad (30)$$

where the index  $i_N \in \{1, \dots, 11\}$  is defined by  $x^{(i_N)} = x_N$  and  $(Y_N, Z_N)$  denotes the observation for  $x_N$ .

The stationary allocation distribution  $\xi_{u\&d}$  is log-concave, see [17], and is approximately given by

$$\xi_{u\&d}(\theta) \simeq \begin{pmatrix} x^{(1)} & x^{(2)} & x^{(3)} & x^{(4)} & x^{(5)} & x^{(6)} & x^{(7)} & x^{(8)} \\ 1.70 \cdot 10^{-3} & 2.12 \cdot 10^{-2} & 0.146 & 0.426 & 0.345 & 5.88 \cdot 10^{-2} & 1.90 \cdot 10^{-3} & 1.13 \cdot 10^{-5} \end{pmatrix}$$

(the total weight on  $x^{(9)}, x^{(10)}, x^{(11)}$  is less than  $10^{-7}$ ). Note that the mode is at  $x^{(4)}$ , one dose below the OSD  $x^{(5)}$ .

We consider trials on 36 patients, organized in a similar way as in [5]: the allocation for the first 10 patients uses the up-and-down rule above, starting with the lowest dose  $x^{(1)}$ ; after the 10th patient, the up-and-down rule is still used until the first observed toxicity ( $Z_N = 1$ ); we then switch to the adaptive design rule (16), with the restriction that we do not allow allocation at a dose one step higher than the maximum level tested so far (following recommendations for practical implementation, see [5]). The parameters are estimated by maximum likelihood (the log-likelihood  $\sum_i l(Y_i, Z_i, x_i; \theta)$  being regularized by the addition of the term  $0.01 \|\theta\|^2$ , which amounts to maximum *a posteriori* estimation with the normal prior  $\mathcal{N}(\mathbf{0}, 50\mathbf{I})$ ). We use the cost function  $\phi(x, \theta) = \pi_{10}^{-1}(x, \theta)$ , with minimum value at the OSD  $x^{(5)}$ ,  $\phi(x^{(5)}, \theta) \simeq 1.2961$ .

Figure 4 shows the progress of a typical trial with  $\lambda_N \equiv 2$ . The symbols indicate the values of the observations at the given points:  $\triangle$  for  $(Y = 0, Z = 0)$ ,  $\triangleright$  for  $(Y = 1, Z = 0)$ ,  $\diamond$  for  $(Y = 1, Z = 1)$  and  $\nabla$  for  $(Y = 0, Z = 1)$ . The up-and-down rule is used until  $N = 15$  where toxicity is observed. The next dose should have been  $x^{(4)}$  but the adaptive design rule (16) selects  $x^{(6)}$  instead.

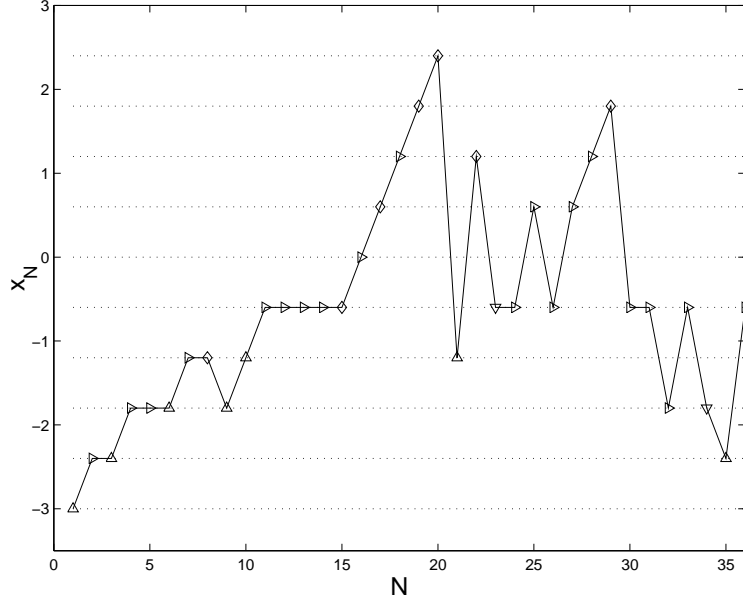


Figure 4: Graphical presentation of a trial with  $\phi(x, \theta) = \pi_{10}^{-1}(x, \theta)$  and  $\lambda_N \equiv 2$  in (16);  $\Delta$  is for  $(Y = 0, Z = 0)$ ,  $\triangleright$  for  $(Y = 1, Z = 0)$ ,  $\diamond$  for  $(Y = 1, Z = 1)$  and  $\nabla$  for  $(Y = 0, Z = 1)$ .

We perform 1,000 independent repetitions of similar trials, using three different adaptive rules: (i) the up-and-down rule (30) used along the whole trial, for the 36 patients; (ii) the up-and-down rule followed by adaptive  $D$ -optimal design (15); and (iii) the up-and-down rule followed by adaptive penalized  $D$ -optimal design (16), with  $\phi(x, \theta) = \pi_{10}^{-1}(x, \theta)$  and  $\lambda_N \equiv 2$ . Since the choice  $\lambda_N \equiv 2$  may seem arbitrary, we also considered the situation (iv) where  $\lambda_N$  is adapted to the estimated value of  $\theta$ . To reduce the computational cost, we only adapt  $\lambda_N$  once in the trial, at the value  $N_s$  when we abandon the up-and-down rule. The value  $\lambda_N = \lambda_{N_s}^*$ ,  $N = N_s, \dots, 36$ , is chosen as the solution for  $\lambda$  of  $\Phi[\xi^*(\lambda), \hat{\theta}_{ML}^{N_s}] = C_\gamma = (1 + \gamma) \min_{x \in \mathcal{X}} \phi(x, \hat{\theta}_{ML}^{N_s})$ ; we take  $\gamma = 0.52$  because it yields  $\lambda = 2$  when  $\theta$  is substituted for  $\hat{\theta}_{ML}^{N_s}$  (it corresponds to allowing an average reduction of about 34% for the probability of success compared to  $\max_{x \in \mathcal{X}} \pi_{10}(x, \hat{\theta}_{ML}^{N_s})$ ). The solution for  $\lambda$  is easily obtained by dichotomy, since we know that the solution satisfies  $0 < \lambda \leq (1 + 1/\gamma) p/C_\gamma$ , see Proposition 1-(i), and  $\Phi[\xi^*(\lambda), \theta]$  decreases when  $\lambda$  increases. Table 2 summarizes the results in terms of the following performance measures:  $\Phi(\xi_{36}, \theta)$ , the total cost of the experiment;  $\det^{-1/6}[\mathbf{M}(\xi_{36}, \theta)]$ , which indicates the precision of the estimation of  $\theta$ ;  $\widehat{x}_{\{t=i\}}^*$ , the number

design rule	$\Phi(\xi, \theta)$	$\psi(\xi, \theta)$	$\widehat{x}_{\{t<4\}}^*$	$\widehat{x}_{\{t=4\}}^*$	$\widehat{x}_{\{t=5\}}^*$	$\widehat{x}_{\{t=6\}}^*$	$\widehat{x}_{\{t>6\}}^*$	$\#x^{(11)}$
(i): (30)	1.87	28.02	20	386	369	86	139	0
$\xi_{u\&d}(\theta)$	1.47	29.4						
(ii): (30)–(15)	3.16	17.23	0	198	705	78	19	1782
$\xi_D^*(\theta)$	4.45	14.99						
(iii): (30)–(16), $\lambda_N \equiv 2$	2.25	19.22	3	231	693	59	14	586
$\xi^*(\lambda = 2, \theta)$	1.97	17.00						
(iv): (30)–(16), $\lambda_N^*$	2.38	18.78	0	223	682	70	25	833

Table 2: Performance measures of different adaptive designs for 36 patients (sample mean over 1,000 repetitions for  $\Phi(\xi, \theta)$  and  $\psi(\xi, \theta) = \det^{-1/6}[\mathbf{M}(\xi, \theta)]$ ;  $\widehat{x}_{\mathcal{A}}^*$  is the number of times  $\arg \max_{x \in \mathcal{X}} \pi_{10}(x, \hat{\theta}_{ML}^{36}) \in \mathcal{A}$  and  $\#x^{(11)}$  is the number of times the highest dose  $x^{(11)}$  has been used, over the 1,000 repetitions).

of times the estimated OSD at the end of the trial, that is,  $\arg \max_{x \in \mathcal{X}} \pi_{10}(x, \hat{\theta}_{ML}^{36})$ , coincided with  $x^{(i)}$ , for  $i = 4, 5, 6$ , and  $\widehat{x}_{\{t<4\}}^*$  (resp.  $\widehat{x}_{\{t>6\}}^*$ ), the number of times the estimated OSD was smaller than  $x^{(4)}$  (resp. larger than  $x^{(6)}$ ); finally  $\#x^{(11)}$ , the number of times the highest dose  $x^{(11)}$  has been used in the trials. The values of  $\Phi(\xi, \theta)$  and  $\psi(\xi, \theta)$  for the designs  $\xi_{u\&d}$ ,  $\xi_D^*$  and  $\xi^*(\lambda = 2)$  computed at the true value of  $\theta$  are also indicated.

Table 2 reveals that the up-and-down rule (30) is very cautious: its associated cost  $\Phi(\xi_{36}, \theta)$  is low, the extreme dose  $x^{(11)}$  has never been used over the 1,000 repetitions. On the other hand, it fails at providing a precise estimation of the model parameters, and the OSD is estimated at values higher than  $x^{(6)}$  in more than 15% of the cases. Parameter estimation is naturally more precise with adaptive  $D$ -optimal design, and the true OSD  $x^{(5)}$  is recognized in more than 70% of the cases. However, this successful behavior in terms of collective ethics is obtained at the price of having more than 5% of patients receiving a dose as high as  $x^{(11)}$ ; also, the associated value of  $\Phi(\xi_{36}, \theta)$  is rather high. The adaptive penalized  $D$ -optimal design appears to make a good compromise between the two strategies: the value of  $\Phi(\xi_{36}, \theta)$  is close to that of the up-and-down rule, the value of  $\det^{-1/6}[\mathbf{M}(\xi_{36}, \theta)]$  is close to that obtained for adaptive

$D$ -optimal design. It recognized  $x^{(5)}$  as the OSD in about 70% of the cases and only 1.6% of the patients received the dose  $x^{(11)}$  when  $\lambda_N \equiv 2$  (2.3% when  $\lambda_N$  is adapted). Of course, other choices of  $\lambda$  would set other compromises. Note that the estimation of the OSD is more cautious for adaptive penalized design than for the up-and-down rule, in the sense that its estimation at  $x^{(6)}$  or a higher dose occurs much less frequently.

In order to limit more severely the number of patients that receive very high doses, we finally consider a compromise strategy that implements a smoother (and less arbitrary) transition between up-and-down and adaptive penalized  $D$ -optimal design. To better illustrate the potential interest of letting  $\lambda_N$  increase in (16), we consider longer trials, with  $N_T = 240$  patients enrolled. Define  $x^{**}(\theta) = \arg \min_{x \in \mathbb{R}} \phi(x, \theta)$  and  $h(x, \theta) = \partial \phi(x, \theta) / \partial x|_{x=x^{**}(\theta)}$ . From the implicit function theorem,

$$\nabla_{\theta} x^{**}(\theta) = \frac{dx^{**}(\theta)}{d\theta} = - \left[ \frac{\partial h(x, \theta)}{\partial x} \Big|_{x=x^{**}(\theta)} \right]^{-1} \frac{\partial h(x, \theta)}{\partial \theta} \Big|_{x=x^{**}(\theta)}$$

and, when using the up-and-down rule (30) the estimator  $\hat{\theta}_{ML}^N$  asymptotically satisfies

$$\sqrt{N} V_N^{-1/2} [x^{**}(\hat{\theta}_{ML}^N) - x^{**}(\bar{\theta})] \xrightarrow{d} z \sim \mathcal{N}(0, 1), \quad N \rightarrow \infty,$$

where  $V_N = [\nabla_{\theta} x^{**}(\hat{\theta}_{ML}^N)]^{\top} \mathbf{M}^{-1}(\xi_N, \hat{\theta}_{ML}^N) [\nabla_{\theta} x^{**}(\hat{\theta}_{ML}^N)]$ . Based on that, we decide to switch from the up-and-down rule to the adaptive one when  $\sqrt{V_N/N} < x^{(2)} - x^{(1)}$ , the interval between two consecutive doses. If  $N_s$  is the index of the patient for which the rule changes, we take  $\lambda_{N_s}$  as the solution for  $\lambda$  of  $\Phi[\xi^*(\lambda), \hat{\theta}_{ML}^{N_s}] = C_{\gamma} = (1 + \gamma) \min_{x \in \mathcal{X}} \phi(x, \hat{\theta}_{ML}^{N_s})$  with  $\gamma = 0.5$  (thus targeting 33% of decrease with respect to the maximum of  $\pi_{10}(x, \hat{\theta}_{ML}^{N_s})$ ). The value of  $\lambda_{N_T}$  at the end of the trial is chosen as the solution for  $\lambda$  of the same equation with  $\gamma = 0.1$  (allowing only 9% of decrease with respect to the maximum of  $\pi_{10}(x, \hat{\theta}_{ML}^{N_s})$ ). In between  $\lambda_N$  increases at a logarithmic rate, that is,  $\lambda_N = \lambda_{N_s} [1 + a \log(N/N_s)]$ ,  $N = N_s, \dots, N_T$ , with  $a = (\lambda_{N_T}/\lambda_{N_s} - 1) / \log(N_T/N_s)$ . When uncertainty on the OSD is large, that is when  $\sqrt{V_N/N} > [x^{(2)} - x^{(1)}] / 2$ , we also restrict the allocations at high doses by adapting the design space, taken as  $\mathcal{X}_N = \{x^{(1)}, \dots, x^{(i_N)}\}$  at step  $N$ : the maximum dose  $x^{(i_N)}$  allowed in (16) is never more than one step higher than previous dose and is smaller than previous dose if toxicity was observed. The results obtained for 150 repetitions of the experiment are summarized in Table 3 (the results obtained

	$\Phi(\xi, \theta)$	$\psi(\xi, \theta)$	$\widehat{x}_{\{t<4\}}^*$	$\widehat{x}_{\{t=4\}}^*$	$\widehat{x}_{\{t=5\}}^*$	$\widehat{x}_{\{t=6\}}^*$	$\widehat{x}_{\{t>6\}}^*$	$\#x^{(11)}$
(30)	1.54	29.04	0	21	116	11	2	0
(30)–(16), $\lambda_N \nearrow$	1.52	27.87	0	13	135	1	1	40

Table 3: Performance measures of adaptive design (16) with increasing  $\lambda_N$  for 240 patients (sample mean over 150 repetitions for  $\Phi(\xi, \theta)$  and  $\psi(\xi, \theta) = \det^{-1/6}[\mathbf{M}(\xi, \theta)]$ ;  $\widehat{x}_{\mathcal{A}}^*$  is the number of times  $\arg \max_{x \in \mathcal{X}} \pi_{10}(x, \hat{\theta}_{ML}^{200}) \in \mathcal{A}$  and  $\#x^{(11)}$  is the number of times the highest dose  $x^{(11)}$  has been used, over the 150 repetitions).

when the up-and-down rule (30) is used for the 240 patients are also indicated). One may notice the precise estimation of the OSD for the adaptive penalized design compared to the up-and-down rule (it even does slightly better than the up-and-down rule both in terms of  $\Phi(\xi, \theta)$  and  $\det^{-1/6}[\mathbf{M}(\xi, \theta)]$ ). At the same time, only about 0.11% of the patients received the maximal dose  $x^{(11)}$ . Also note that, from the results of Sect. 4.2, instead of setting  $N_T = 240$  in advance, the decision to stop the trial could be based on the value of  $V_N$ .  $\square$

## 6 Conclusion and further developments

We have shown that constrained optimal design can be formulated in a way that allows a clear compromise between gaining information and minimizing a cost. We have also shown that the optimal design can be constructed sequentially and proved the strong consistency and asymptotic normality of the estimator of the model parameters in such adaptive designs. The dose-finding example with bivariate binary responses has illustrated the potential of adaptive penalized  $D$ -optimal design to set compromises between individual and collective ethics. Further developments and numerical studies are required to define suitable rules for selecting cost functions and for choosing the value (or the sequence of values) for the penalty coefficients  $\lambda_N$ . Relating  $\lambda_N$  to the precision of the estimation of the OSD is a possible option to investigate in dose-finding problems. We mention below some extensions of this work, some straightforward, others more challenging, and indicate a motivating objective concerning the design of non-stationary dose-finding

experiments preserving individual ethics.

**Bayesian estimation** The extension of the results presented to Bayesian estimators should not raise particular difficulties, especially since consistency is usually easier to obtained than for LS or ML estimation using martingales properties, see, e.g., [16]. A straightforward modification of (16) is to replace  $\mathbf{M}(\xi_N, \hat{\theta}^N)$  by  $[\mathbf{M}^{-1}(\xi_N, \hat{\theta}^N) + \Omega/N]^{-1}$ , with  $\Omega$  the prior covariance matrix for the model parameters.

**Multiple constraints** The results obtained in Sect. 3 and 4 easily generalize to the presence of several constraints, that is, when problem  $P_3(\theta)$  is transformed into:

$$\text{maximize } \log \det \mathbf{M}(\xi, \theta) \text{ with respect to } \xi \in \Xi \text{ under the constraints } \Phi(\xi, \theta) \leq C_i, \quad i = 1, \dots, m.$$

A Lagrange coefficient is then associated with each constraint and the design algorithm (16) becomes

$$x_{N+1} = \arg \max_{x \in \mathcal{X}} \left\{ \text{trace}[\mu(x, \hat{\theta}^N) \mathbf{M}^{-1}(\xi_N, \hat{\theta}^N)] - \sum_{i=1}^m \lambda_N^{(i)} \phi_i(x, \hat{\theta}^N) \right\}.$$

When the  $\lambda_N^{(i)}$ 's are controlled to increase to infinity, define  $\rho_N^{(j)} = \lambda_N^{(j)} / \sum_i \lambda_N^{(i)}$  and suppose that a limit  $\bar{\rho}_j$  exists for each  $\rho_N^{(j)}$ ,  $j = 1, \dots, m$ . Then, if all cost functions  $\phi_i(\cdot, \theta)$  are bounded on  $\mathcal{X}$ , the asymptotic behaviors of the design and estimators are the same as in Sect. 4 for  $\lambda_N = \sum_i \lambda_N^{(i)}$  and  $\phi(x, \theta) = \sum_j \bar{\rho}_j \phi_j(x, \theta)$ . Also, the developments of Sect. 3 remain valid when the  $\lambda_N^{(i)}$ 's are kept constant, or when they are adapted to  $\hat{\theta}^N$ , that is, when they correspond to the optimal coefficients in the solution of  $P_3(\hat{\theta}^N)$ . Note, however, that the presence of several constraints makes this optimal solution more difficult to determine (it can no longer be obtained by solving a series of unconstrained problems with an increasing sequence of coefficients, contrary to the single constraint case).

**Finite horizon** The results of Sect. 4 indicate that, when  $\lambda_N$  increases to infinity at suitable speed, the design converges to the delta measure located at the optimum. In clinical trials, this means that more and more patients receive doses close to the optimal one (and none receives extreme doses if the penalty function satisfies (11)). This is an asymptotic result, however, and approaching the optimal solution over a finite horizon is a challenging task. In the case of LS estimation in linear regression, design strategies are

suggested in [28] that are shown to be close to the optimum control (stochastic dynamic programming) solution. It is then tempting to replace the Bernoulli model by a regression type model (observation  $Y_k$  at design point  $x_k$  has mean value  $\pi(x_k, \theta)$  and variance  $\pi(x_k, \theta)[1 - \pi(x_k, \theta)]$ ), as suggested in [34], with a linear parametrization, and then try to apply the finite-horizon results of [28].

**Non stationary clinical trials** Strong consistency of the estimator is obtained in Sect. 4 when the penalty coefficient  $\lambda_N$  in (16) tends to infinity. Moreover, when the growth of  $\lambda_N$  is not too fast, the estimator is asymptotically normal. This means that, although the design is non-stationary in the sense that patients enrolled in the trial receive better and better treatments, the information collected at the end of the experiment can be used to set future treatments. In particular, the minimum effective and maximum tolerated doses can be estimated and confidence intervals can be given. At the same time, the fact that patients receive unequal treatments in the trial raises ethical issues: there is no randomization, a new patient tends to receive a better treatment than patients previously treated (since when  $\lambda_N$  increases, the design points tend to get closer to the optimal dose). This emphasizes the importance of constructing a fair rule for choosing the increasing sequence  $(\lambda_N)$ . Trying to give equal probabilities of success at patients  $P_N$  and  $P_{N+1}$  when patient  $P_N$  is treated first seems to be an honest ambition, and the increase of  $\lambda_N$  could then be used to compensate for the late treatment of patient  $P_{N+1}$ . This requires a model for the evolution of the probability of success as a function of the delay in treatment, to be combined with a suitable characterization of the improvement of treatment that can be expected when increasing  $\lambda_N$ .

## Appendix

*Proof of Proposition 1.*

- (i) Since  $\xi^*$  is optimal we have for all  $x \in \mathcal{X}$ ,  $\text{trace}[\mu(x, \theta)\mathbf{M}^{-1}(\xi^*, \theta)] \leq p + \lambda [\phi(x, \theta) - \Phi(\xi^*, \theta)]$ , see
- (6). This is true in particular at a  $x^*$  defined by (8) and  $\text{trace}[\mu(x^*, \theta)\mathbf{M}^{-1}(\xi^*, \theta)] \geq 0$  gives the result.
- (ii) For any  $a > 0$ , take  $\lambda \geq a/\Delta_\theta(\xi)$  and define  $\tilde{\xi} = (1 - \alpha)\xi + \alpha\delta_{x^*}$  with  $\delta_{x^*}$  the delta measure at

a point  $x^*$  satisfying (8) and  $\alpha = 1 - a/[\lambda\Delta_\theta(\xi)]$ . This gives  $\Phi(\tilde{\xi}, \theta) - \phi_\theta^* = a/\lambda$  and  $\log \det \mathbf{M}(\tilde{\xi}, \theta) \geq p \log(1 - \alpha) + \log \det \mathbf{M}(\xi, \theta)$ . Therefore,

$$\begin{aligned} \log \det \mathbf{M}(\xi^*, \theta) - \lambda[\Phi(\xi^*, \theta) - \phi_\theta^*] &\geq \log \det \mathbf{M}(\tilde{\xi}, \theta) - \lambda[\Phi(\tilde{\xi}, \theta) - \phi_\theta^*] \\ &\geq p \log a - a - p \log \Delta_\theta(\xi) + \log \det \mathbf{M}(\xi, \theta) - p \log \lambda. \end{aligned}$$

Since  $\Phi(\xi^*, \theta) \geq \phi_\theta^*$ , the result follows.

When  $\phi(x, \theta)$  is bounded by  $\bar{\phi}_\theta$ , the optimality of  $\xi^*$  implies that for all  $x \in \mathcal{X}$ ,  $\text{trace}[\mu(x, \theta)\mathbf{M}^{-1}(\xi^*, \theta)] \leq B = p + \lambda(\bar{\phi}_\theta - \phi_\theta^*)$ . Write  $\mu(x, \theta) = \mathbf{F}_\theta^\top(x)\mathbf{F}_\theta(x)$  with  $\mathbf{F}_\theta^\top(x) = [\mathbf{f}_{1,\theta}(x), \dots, \mathbf{f}_{m,\theta}(x)]$  and  $\mathbf{f}_{i,\theta}(x)$  a  $p$ -dimensional vector,  $i = 1, \dots, m$ . From the inequality  $\text{trace}[\mu(x, \theta)\mathbf{M}^{-1}(\xi^*, \theta)] \leq B$  we obtain that  $\mathbf{f}_{i,\theta}^\top(x)\mathbf{M}^{-1}(\xi^*, \theta)\mathbf{f}_{i,\theta}(x) \leq B$ ,  $i = 1, \dots, m$ . We have

$$\lambda_{\min}[\mathbf{M}(\xi^*, \theta)] = \lambda_{\max}^{-1}[\mathbf{M}^{-1}(\xi^*, \theta)] = \left[ \max_{\|\mathbf{u}\|=1} \mathbf{u}^\top \mathbf{M}^{-1}(\xi^*, \theta) \mathbf{u} \right]^{-1}.$$

Consider the optimization problem defined by: maximize  $\mathbf{u}^\top \mathbf{A}^\top \mathbf{A} \mathbf{u}$  with respect to  $\mathbf{A}$  and  $\mathbf{u}$  respectively in  $\mathbb{R}^{n \times p}$  and  $\mathbb{R}^p$ ,  $n \leq p$ , subject to the constraints  $\|\mathbf{u}\| = 1$  and  $\mathbf{f}_{i,\theta}^\top(x)\mathbf{A}^\top \mathbf{A} \mathbf{f}_{i,\theta}(x) \leq B$ ,  $\forall x \in \mathcal{X}$  and  $\forall i = 1, \dots, m$ . The optimal solution is obtained for  $\mathbf{A} = \mathbf{v}^\top \in \mathbb{R}^p$  such that  $|\mathbf{f}_{i,\theta}^\top(x)\mathbf{v}| \leq \sqrt{B}$ ,  $\forall x \in \mathcal{X}$ ,  $\forall i = 1, \dots, m$ , and  $b = \mathbf{v}^\top \mathbf{v}$  is maximal. For  $x$  varying in  $\mathcal{X}$  the  $\mathbf{f}_{i,\theta}(x)$ 's span  $\mathbb{R}^p$  (since a nonsingular information matrix exists). Therefore, there exists a positive constant  $\delta$  such that the optimal value for  $b$  is bounded by  $B/\delta$ , so that  $\lambda_{\min}[\mathbf{M}(\xi^*, \theta)] > \delta/B$ .

(iii) Since  $\text{trace}[\mu(x, \theta)\mathbf{M}^{-1}(\xi^*, \theta)] \leq p + \lambda[\phi(x, \theta) - \Phi(\xi^*, \theta)]$  for all  $x \in \mathcal{X}$  when  $\xi^*$  is optimal, and  $\int_{\mathcal{X}} \{\text{trace}[\mu(x, \theta)\mathbf{M}^{-1}(\xi^*, \theta)] - \lambda\phi(x, \theta)\} \xi^*(dx) = p - \lambda\Phi(\xi^*, \theta)$ , we have  $\text{trace}[\mu(\hat{x}, \theta)\mathbf{M}^{-1}(\xi^*, \theta)] = p + \lambda[\phi(\hat{x}, \theta) - \Phi(\xi^*, \theta)]$  at any  $\hat{x}$  support point of  $\xi^*$ . Suppose that  $\lambda$  is large enough so that there exists a design  $\xi_\lambda \in \Xi$  satisfying  $\Delta_\theta(\xi_\lambda) \geq p/\lambda$ . We proceed as in (ii) and construct a design  $\tilde{\xi}_\lambda = (1 - \alpha)\xi_\lambda + \alpha\delta_{x^*}$  with  $\alpha = 1 - p/[\lambda\Delta_\theta(\xi_\lambda)]$  so that  $\Phi(\tilde{\xi}_\lambda, \theta) - \phi_\theta^* = p/\lambda$ . With the same notation as in (ii), we can write  $\text{trace}[\mu(\hat{x}, \theta)\mathbf{M}^{-1}(\tilde{\xi}_\lambda, \theta)] = \sum_{i=1}^m \mathbf{f}_{i,\theta}^\top(\hat{x})\mathbf{M}^{-1}(\tilde{\xi}_\lambda, \theta)\mathbf{f}_{i,\theta}(\hat{x})$ . We then follow the same approach as in [15] and define  $\mathbf{H}(\xi^*, \xi, \theta) = \mathbf{M}^{1/2}(\xi^*, \theta)\mathbf{M}^{-1}(\xi, \theta)\mathbf{M}^{1/2}(\xi^*, \theta)$ . We obtain

$$\text{trace}[\mu(\hat{x}, \theta)\mathbf{M}^{-1}(\tilde{\xi}_\lambda, \theta)] = \sum_{i=1}^m \mathbf{f}_{i,\theta}^\top(\hat{x})\mathbf{M}^{-1/2}(\xi^*, \theta) \mathbf{H}(\xi^*, \tilde{\xi}_\lambda, \theta)\mathbf{M}^{-1/2}(\xi^*, \theta)\mathbf{f}_{i,\theta}(\hat{x})$$

$$\begin{aligned}
&\geq \lambda_{\min}[\mathbf{H}(\xi^*, \tilde{\xi}_\lambda, \theta)] \text{trace}[\mu(\hat{x}, \theta) \mathbf{M}^{-1}(\xi^*, \theta)] \\
&= \lambda_{\min}[\mathbf{H}(\xi^*, \tilde{\xi}_\lambda, \theta)] \{p + \lambda [\phi(\hat{x}, \theta) - \Phi(\xi^*, \theta)]\} \\
&\geq \lambda_{\min}[\mathbf{H}(\xi^*, \tilde{\xi}_\lambda, \theta)] \left\{ p + \lambda [\phi(\hat{x}, \theta) - \Phi(\tilde{\xi}_\lambda, \theta)] \right\} \\
&= \lambda_{\min}[\mathbf{H}(\xi^*, \tilde{\xi}_\lambda, \theta)] \lambda [\phi(\hat{x}, \theta) - \phi_\theta^*]
\end{aligned}$$

where we used the property  $\Delta_\theta(\xi^*) \leq p/\lambda = \Delta_\theta(\tilde{\xi}_\lambda)$ , see (i). Therefore,

$$\text{trace}[\mu(\hat{x}, \theta) \mathbf{M}^{-1}(\xi_\lambda, \theta)] \geq (1 - \alpha) \text{trace}[\mu(\hat{x}, \theta) \mathbf{M}^{-1}(\tilde{\xi}_\lambda, \theta)] \geq \frac{p \lambda_{\min}[\mathbf{H}(\xi^*, \tilde{\xi}_\lambda, \theta)] [\phi(\hat{x}, \theta) - \phi_\theta^*]}{\Delta_\theta(\xi_\lambda)}. \quad (31)$$

The last step consists in deriving a lower bound on  $\lambda_{\min}[\mathbf{H}(\xi^*, \tilde{\xi}_\lambda, \theta)]$  that does not depend on  $\xi^*$ . We have

$$\text{trace} \left[ \mathbf{H}^{-1}(\xi^*, \tilde{\xi}_\lambda, \theta) \right] = \int_{\mathcal{X}} \text{trace} [\mu(x, \theta) \mathbf{M}^{-1}(\xi^*, \theta)] \tilde{\xi}_\lambda(dx)$$

and thus from the optimality of  $\xi^*$ ,

$$\begin{aligned}
\text{trace} \left[ \mathbf{H}^{-1}(\xi^*, \tilde{\xi}_\lambda, \theta) \right] &\leq p + \lambda \int_{\mathcal{X}} [\phi(x, \theta) - \Phi(\xi^*, \theta)] \tilde{\xi}_\lambda(dx) \\
&= p + \lambda \left[ \Phi(\tilde{\xi}_\lambda, \theta) - \Phi(\xi^*, \theta) \right] \leq p + \lambda \left[ \Phi(\tilde{\xi}_\lambda, \theta) - \phi_\theta^* \right] = 2p.
\end{aligned}$$

Therefore,  $\lambda_{\min}[\mathbf{H}(\xi^*, \tilde{\xi}_\lambda, \theta)] \geq 1/(2p)$ , which, together with (31), concludes the proof.  $\blacksquare$

*Proof of Lemma 1.* 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

$$\begin{aligned}
\mathbf{M}(\xi_N, \theta) &= \frac{1}{N} \sum_{k=1}^N \mathbf{f}_\theta(x_k) \mathbf{f}_\theta^\top(x_k) \\
&= \frac{1}{N} \sum_{i=1}^{q^*} r_{N,i:K} \mathbf{f}_\theta(x^{(i_N)}) \mathbf{f}_\theta^\top(x^{(i_N)}) + \frac{1}{N} \sum_{x_k \notin \mathcal{X}_N(q^*)} \mathbf{f}_\theta(x_k) \mathbf{f}_\theta^\top(x_k), \quad (32)
\end{aligned}$$

where  $i_N$  is the index (depending on  $N$ ) of a design point appearing  $r_{N,i:K}$  times in  $x_1, \dots, x_N$  and  $\mathcal{X}_N(q^*) = \{x^{(1_N)}, \dots, x^{(q^*_N)}\}$  is the set of such points for  $i \leq q^*$ . Let  $\mathbf{M}_N(\theta)$  denote the first matrix on the right-hand side of (32). For any  $x^{(i_N)} \in \mathcal{X}_N(q^*)$  we have

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

with  $\mathbf{M}_N^-(\theta)$  any  $g$ -inverse of  $\mathbf{M}_N(\theta)$ . Therefore, from the definition of  $q^*$ , there exists  $N_1$  such that

$$\text{for all } i \leq q^*, N > N_1 \text{ and } \theta \in \Theta, \mathbf{f}_\theta^\top(x^{(i_N)})\mathbf{M}^{-1}(\xi_N, \theta)\mathbf{f}_\theta(x^{(i_N)}) - \lambda_N \phi(x^{(i_N)}, \theta) \leq \frac{1}{\alpha}. \quad (33)$$

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

Define

$$\mathbf{M}_N^{(1)}(\theta) = \sum_{i=1}^{q^*} \mathbf{f}_\theta(x^{(i_N)})\mathbf{f}_\theta^\top(x^{(i_N)}), \quad \mathbf{M}_N^{(2)}(\theta) = \sum_{i=1}^K \mathbf{f}_\theta(x^{(i_N)})\mathbf{f}_\theta^\top(x^{(i_N)}). \quad (34)$$

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

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

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

$$\mathbf{u}^\top \mathbf{M}^{-1}(\xi_N, \theta) \mathbf{u} \geq \frac{1}{\beta_N} \mathbf{u}^\top [\mathbf{M}_N^{(2)}(\theta)]^{-1} [\mathbf{I} - \mathbf{P}_N(\theta)] \mathbf{u} \quad (35)$$

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

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

Note that the right-hand side of (35) is zero when  $\mathbf{u} \in \mathcal{M}[\mathbf{M}_N^{(1)}(\theta)]$  (i.e., when  $\mathbf{u} = \mathbf{M}_N^{(1)}(\theta)\mathbf{v}$  for some  $\mathbf{v} \in \mathbb{R}^p$ ). When  $\mathbf{u} = \mathbf{f}_\theta(x^{(i_N)})$  for some  $i \in \{q^* + 1, \dots, K\}$  we can construct a lower bound for this term, of the form  $A/\beta_N$  with  $A$  constant. Indeed, from (34) and  $\text{H}_{\mathcal{X}}\text{-}(ii)$ ,

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

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

$$\max_{i=q^*+1, \dots, K} [\mathbf{z}^\top \mathbf{f}_\theta(x^{(i_N)})]^2 > \frac{\gamma}{K - q^*} \|\mathbf{z}\|^2. \quad (36)$$

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

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

and thus from (36),

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

Let  $i_N^*$  denote the argument of the maximum on the left-hand side, for which we have,  $\mathbf{z}_{\theta, i_N^*}^\top \mathbf{f}_\theta(x^{(i_N^*)}) = \mathbf{z}_{\theta, i_N^*}^\top \mathbf{M}_N^{(2)}(\theta) \mathbf{z}_{\theta, i_N^*} \leq K L \|\mathbf{z}_{\theta, i_N^*}\|^2$  with  $L = \max_{x \in \mathcal{X}, \theta \in \Theta} \|\mathbf{f}_\theta(x)\|^2$ . This finally gives: for all  $\theta \in \Theta$ ,  $\max_{i=q^*+1, \dots, K} \mathbf{f}_\theta^\top(x^{(i_N)}) [\mathbf{M}_N^{(2)}(\theta)]^{-1} [\mathbf{I} - \mathbf{P}_N(\theta)] \mathbf{f}_\theta(x^{(i_N)}) > \gamma/[LK(K - q^*)]$ , and thus, from (35) and  $\text{H}_\phi$ -(i),

$$\text{for all } \theta \in \Theta, \quad \max_{i=q^*+1, \dots, K} \left\{ \mathbf{f}_\theta^\top(x^{(i_N)}) \mathbf{M}^{-1}(\xi_N, \theta) \mathbf{f}_\theta(x^{(i_N)}) - \lambda_N \phi(x^{(i_N)}, \theta) \right\} > \frac{1}{\beta_N} \frac{\gamma}{LK(K - q^*)} - \bar{\lambda} \bar{\phi}. \quad (37)$$

Together with (33), it gives:  $N > N_1$  and  $\beta_N < \beta^* = \gamma\alpha/[LK(K - q^*)(1 + \alpha\bar{\lambda}\bar{\phi})] \Rightarrow x_{N+1} \notin \mathcal{X}_N(q^*)$  in the sequence (16). 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  $N > N_1$ ,  $(\sum_{i=1}^{q^*} r_{N, i:K})/N > q^*\alpha$ , and therefore  $\beta_N^* < (1 - q^*\alpha)/(K - q^*)$ . By construction,  $\beta_N < \beta^*$  and  $N > N_1$  imply

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

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 than  $\beta^*$ . Suppose that the threshold  $\beta^*$  is crossed downwards at  $N_2 > N_1$ , i.e.,  $\beta_{N_2-1} \geq \beta^*$  and  $\beta_{N_2} < \beta^*$ . This implies  $\beta_{N_2} = \beta_{N_2-1}(N_2 - 1)/N_2$  and thus  $\beta^*(N_2 - 1)/N_2 \leq \beta_{N_2} < \beta^*$ ,

so that  $\beta_{N_2}$  tends to  $\beta^*$  when  $N_2 \rightarrow \infty$ . We thus obtain  $\liminf_{N \rightarrow \infty} \beta_N \geq \underline{\beta} = \beta^*/(K - q^*)$ , showing that  $q^* \geq p$ , which concludes the proof.  $\blacksquare$

*Proof of Theorem 3.* We have already seen that the conditions of Lemma 1 with  $\text{H}_{\mathcal{X}}\text{-(iii)}$  (respectively,  $\text{H}_{\mathcal{X}}\text{-(iii)'}$ ) imply  $\hat{\theta}_{LS}^N \xrightarrow{\text{a.s.}} \bar{\theta}$  (respectively,  $\hat{\theta}_{ML}^N \xrightarrow{\text{a.s.}} \bar{\theta}$ ) as  $N \rightarrow \infty$ . All what we need to obtain the asymptotic optimality of  $\xi_N$  is thus a continuity property of the form: for all  $\epsilon > 0$ , there exists  $\beta > 0$  such that  $[\|\hat{\theta}^N - \bar{\theta}\| < \beta \text{ for all } N \text{ larger than some } N_0]$  implies  $\liminf_{N \rightarrow \infty} H_{\bar{\theta}}(\xi_N) > H_{\bar{\theta}}^* - \epsilon$ . We show that this is indeed true under the additional assumptions  $\text{H}_{\mathcal{X}}\text{-(iv)}$ ,  $\text{H}_{\phi}\text{-(ii)}$  and  $\text{H}_f\text{-(i)}$ .

First note that  $\lambda^*(\theta) = \arg \min_{\lambda \geq 0} \max_{\xi \in \Xi} \{\log \det \mathbf{M}(\xi, \theta) + \lambda [C - \Phi(\xi, \theta)]\}$  is continuous in  $\theta$  as the argument of the minimum of a convex function (given by the maximum of convex functions) that is continuous in  $\theta$ . Therefore,  $\forall \epsilon_0, \exists \beta_1$  such that  $\|\theta - \bar{\theta}\| < \beta_1 \Rightarrow |\lambda^*(\theta) - \lambda^*(\bar{\theta})| < \epsilon_0$ . Also,  $\text{H}_f\text{-(i)}$ ,  $\text{H}_{\phi}\text{-(ii)}$  and  $\text{H}_{\mathcal{X}}\text{-(i)}$  imply:  $\forall \epsilon_0, \exists \beta_2$  such that  $\|\theta - \bar{\theta}\| < \beta_2 \Rightarrow \max_{x \in \mathcal{X}} \|\mathbf{f}_{\theta}(x) - \mathbf{f}_{\bar{\theta}}(x)\| < \epsilon_0$  and  $\exists \beta_3$  such that  $\|\theta - \bar{\theta}\| < \beta_3 \Rightarrow \max_{x \in \mathcal{X}} |\phi(x, \theta) - \phi(x, \bar{\theta})| < \epsilon_0$ .

From Lemma 1, there exists  $N_1$  and  $\alpha > 0$  such that for all  $N > N_1$ ,  $r_{N,j:K} > \alpha N$ ,  $j = 1, \dots, q^*$ , with  $q^* > p$ , and thus from  $\text{H}_{\mathcal{X}}\text{-(iv)}$ ,  $\lambda_{\min}[\mathbf{M}(\xi_N, \bar{\theta})] > \alpha \bar{\gamma}$ . Direct calculations then give

$$\max_{x \in \mathcal{X}} \max_{\|\theta - \bar{\theta}\| < \beta_2} |\mathbf{f}_{\theta}^{\top}(x) \mathbf{M}^{-1}(\xi_N, \theta) \mathbf{f}_{\theta}(x) - \mathbf{f}_{\bar{\theta}}^{\top}(x) \mathbf{M}^{-1}(\xi_N, \bar{\theta}) \mathbf{f}_{\bar{\theta}}(x)| < B \epsilon_0$$

for  $\epsilon_0$  small enough and  $N > N_1$ , with  $B$  depending on  $\alpha, \bar{\gamma}$  and  $\bar{f} = \max_{x \in \mathcal{X}} \|\mathbf{f}_{\bar{\theta}}(x)\|$ .

Therefore, for all  $\epsilon_0$  small enough, we can take  $\beta = \min\{\beta_1, \beta_2, \beta_3\}$  and obtain, for all  $N > N_1$  and  $\|\hat{\theta}^N - \bar{\theta}\| < \beta$  in (16),

$$\begin{aligned} & \mathbf{f}_{\bar{\theta}}^{\top}(x_{N+1}) \mathbf{M}^{-1}(\xi_N, \theta) \mathbf{f}_{\bar{\theta}}(x_{N+1}) - \lambda^*(\bar{\theta}) \phi(x_{N+1}, \bar{\theta}) \\ & > \mathbf{f}_{\hat{\theta}^N}^{\top}(x_{N+1}) \mathbf{M}^{-1}(\xi_N, \hat{\theta}^N) \mathbf{f}_{\hat{\theta}^N}(x_{N+1}) - \lambda^*(\hat{\theta}^N) \phi(x_{N+1}, \hat{\theta}^N) - \epsilon_0 (B + \bar{\lambda} + \bar{\phi}) - \epsilon_0^2 \\ & = \max_{x \in \mathcal{X}} \left[ \mathbf{f}_{\hat{\theta}^N}^{\top}(x) \mathbf{M}^{-1}(\xi_N, \hat{\theta}^N) \mathbf{f}_{\hat{\theta}^N}(x) - \lambda^*(\hat{\theta}^N) \phi(x, \hat{\theta}^N) \right] - \epsilon_0 (B + \bar{\lambda} + \bar{\phi}) - \epsilon_0^2 \\ & > \max_{x \in \mathcal{X}} \left[ \mathbf{f}_{\bar{\theta}}^{\top}(x) \mathbf{M}^{-1}(\xi_N, \bar{\theta}) \mathbf{f}_{\bar{\theta}}(x) - \lambda^*(\bar{\theta}) \phi(x, \bar{\theta}) \right] - 2\epsilon_0 (B + \bar{\lambda} + \bar{\phi}). \end{aligned}$$

For a given  $\epsilon$ , take  $\epsilon_0 = \epsilon / (B + \bar{\lambda} + \bar{\phi})$  and  $\beta$  as above, and suppose that there exists  $\delta > 0$  such that

$$H_{\bar{\theta}}(\xi_N) < H_{\bar{\theta}}^* - \delta - \epsilon \tag{38}$$

for all  $N$  larger than some  $N_2$ , with  $H_{\bar{\theta}}(\xi_N)$  and  $H_{\bar{\theta}}^*$  respectively given by (25) and (24). From the concavity of the design criterion, this implies

$$\max_{x \in \mathcal{X}} [\mathbf{f}_{\bar{\theta}}^\top(x) \mathbf{M}^{-1}(\xi_N, \bar{\theta}) \mathbf{f}_{\bar{\theta}}(x) - \lambda^*(\bar{\theta}) \phi(x, \bar{\theta})] > p - \lambda^*(\bar{\theta}) \Phi(\xi_N, \bar{\theta}) + \epsilon + \delta$$

and thus

$$\mathbf{f}_{\bar{\theta}}^\top(x_{N+1}) \mathbf{M}^{-1}(\xi_N, \bar{\theta}) \mathbf{f}_{\bar{\theta}}(x_{N+1}) - \lambda^*(\bar{\theta}) \phi(x_{N+1}, \bar{\theta}) > p - \lambda^*(\bar{\theta}) \Phi(\xi_N, \bar{\theta}) + \delta \quad (39)$$

for all  $N > \max\{N_1, N_2, N_0\}$  when  $\|\hat{\theta}^N - \bar{\theta}\| < \beta$  for all  $N > N_0$ . Direct calculations give

$$\begin{aligned} H_{\bar{\theta}}(\xi_{N+1}) - H_{\bar{\theta}}(\xi_N) &= \log \left[ 1 + \frac{\mathbf{f}_{\bar{\theta}}^\top(x_{N+1}) \mathbf{M}^{-1}(\xi_N, \bar{\theta}) \mathbf{f}_{\bar{\theta}}(x_{N+1})}{N} \right] - p \log \left( 1 + \frac{1}{N} \right) \\ &\quad + \frac{\lambda^*(\bar{\theta})}{N+1} [\Phi(\xi_N, \bar{\theta}) - \phi(x_{N+1}, \bar{\theta})] \end{aligned} \quad (40)$$

and thus from (39),

$$\begin{aligned} H_{\bar{\theta}}(\xi_{N+1}) - H_{\bar{\theta}}(\xi_N) &> \log \left[ 1 + \frac{\mathbf{f}_{\bar{\theta}}^\top(x_{N+1}) \mathbf{M}^{-1}(\xi_N, \bar{\theta}) \mathbf{f}_{\bar{\theta}}(x_{N+1})}{N} \right] - p \log \left( 1 + \frac{1}{N} \right) \\ &\quad + \frac{p + \delta - \mathbf{f}_{\bar{\theta}}^\top(x_{N+1}) \mathbf{M}^{-1}(\xi_N, \bar{\theta}) \mathbf{f}_{\bar{\theta}}(x_{N+1})}{N+1}. \end{aligned}$$

Since  $\mathbf{f}_{\bar{\theta}}^\top(x_{N+1}) \mathbf{M}^{-1}(\xi_N, \bar{\theta}) \mathbf{f}_{\bar{\theta}}(x_{N+1})$  is bounded by  $\bar{f}^2/(\alpha\bar{\gamma})$ , we obtain that  $H_{\bar{\theta}}(\xi_{N+1}) - H_{\bar{\theta}}(\xi_N) > \delta/(2N)$  for  $N$  large enough, which implies that  $H_{\bar{\theta}}(\xi_N) \rightarrow \infty$  as  $N \rightarrow \infty$ , in contradiction with  $H_{\phi}$ -(i),  $H_{\lambda}$ -(i') and  $H_f$ -(i), see the definition of  $H_{\bar{\theta}}(\xi)$  in (25).

Therefore, for any  $\epsilon > 0$ ,  $\|\hat{\theta}^N - \bar{\theta}\| < \beta$  for all  $N > N_0$  implies the existence of a subsequence  $\xi_{N_t}$  such that  $\limsup_{t \rightarrow \infty} H_{\bar{\theta}}(\xi_{N_t}) \geq H_{\bar{\theta}}^* - \epsilon$ . From (40), for all  $\delta > 0$  there exists  $N_3$  such that for all  $N > N_3$ ,  $H_{\bar{\theta}}(\xi_{N+1}) > H_{\bar{\theta}}(\xi_N) - \delta$ . Also, from the developments just above, there exists  $N_4$  such that for all  $N > N_4$ , (38) implies  $H_{\bar{\theta}}(\xi_{N+1}) > H_{\bar{\theta}}(\xi_N)$ . Take any  $N_t > \max(N_0, N_3, N_4)$  satisfying  $H_{\bar{\theta}}(\xi_{N_t}) > H_{\bar{\theta}}^* - \epsilon - \delta$ , we obtain  $H_{\bar{\theta}}(\xi_N) > H_{\bar{\theta}}^* - \epsilon - 2\delta$ , for all  $N > N_t$ . Since  $\delta$  is arbitrary,  $\liminf_{N \rightarrow \infty} H_{\bar{\theta}}(\xi_N) \geq H_{\bar{\theta}}^* - \epsilon$ , which concludes the proof.  $\blacksquare$

*Proof of Lemma 3.* We consider the case of LS estimation in the regression model (17). The proof is similar (but simpler) for ML estimation in the model (21) and we shall only indicate the adaptations that are required.

A first-order series development of the gradient of the LS criterion (18) around  $\bar{\theta}$  gives

$$\nabla_{\theta} S_N(\hat{\theta}_{LS}^N) = \mathbf{0} = \nabla_{\theta} S_N(\bar{\theta}) + \nabla_{\theta}^2 S_N(\bar{\theta}^N)(\hat{\theta}_{LS}^N - \bar{\theta}), \quad (41)$$

where  $\nabla_{\theta} S_N(\bar{\theta}) = -2 \sum_{k=1}^N \varepsilon_k \mathbf{f}_{\bar{\theta}}(x_k)$ ,  $\nabla_{\theta}^2 S_N(\theta)$  is the (Hessian) matrix of second-order derivatives of  $S_N(\theta)$ , given by

$$\nabla_{\theta}^2 S_N(\theta) = 2N \mathbf{M}(\xi_N, \theta) - 2 \sum_{k=1}^N [Y_k - \eta(x_k, \theta)] \frac{\partial \mathbf{f}_{\theta}(x_k)}{\partial \theta^{\top}},$$

and  $\bar{\theta}^N = (1 - \gamma_N) \bar{\theta} + \gamma_N \hat{\theta}_{LS}^N$ ,  $\gamma_N \in (0, 1)$ , with  $\bar{\theta}^N$  measurable, see [18].

We first try to bound  $\|\hat{\theta}_{LS}^N - \bar{\theta}\|$ . We have  $(1/N) \nabla_{\theta}^2 S_N(\theta) = 2\mathbf{M}(\xi_N, \theta) + 2\mathbf{A}(\xi_N, \theta, \bar{\theta}) - 2\mathbf{B}_N(\xi_N, \theta)$ ,

with

$$\mathbf{A}(\xi_N, \theta, \bar{\theta}) = \frac{1}{N} \sum_{k=1}^N [\eta(x_k, \theta) - \eta(x_k, \bar{\theta})] \frac{\partial \mathbf{f}_{\theta}(x_k)}{\partial \theta^{\top}} \quad \text{and} \quad \mathbf{B}_N(\xi_N, \theta) = \frac{1}{N} \sum_{k=1}^N \varepsilon_k \frac{\partial \mathbf{f}_{\theta}(x_k)}{\partial \theta^{\top}}.$$

From  $\mathbf{H}_f$ -(ii), there exist  $A_1 > 0$  and  $A_2 > 0$  such that  $\sup_{x \in \mathcal{X}} \sup_{\|\theta - \bar{\theta}\| \leq \delta} |\eta(x, \theta) - \eta(x, \bar{\theta})| < A_1 \delta$  and  $\sup_{x \in \mathcal{X}} \sup_{\|\theta - \bar{\theta}\| \leq \delta} \|\partial \mathbf{f}_{\theta}(x) / \partial \theta^{\top}\| < A_2$ . This implies  $\lim_{\delta \rightarrow 0} \sup_{\|\theta - \bar{\theta}\| \leq \delta} \|\mathbf{A}(\xi_N, \theta, \bar{\theta})\| = 0$ . We also have

$$\sup_{\|\theta - \bar{\theta}\| \leq \delta} \|\mathbf{B}(\xi_N, \theta)\| < \sum_{x \in \mathcal{X}} \frac{\left| \sum_{k=1, x_k=x}^N \varepsilon_k \right|}{r_N(x)} A_2 \frac{r_N(x)}{N} \quad (42)$$

where  $r_N(x)$  denotes the number of times  $x$  appears in the sequence  $x_1, \dots, x_N$ . Either  $r_N(x)$  is bounded or  $r_N(x)$  tends to infinity (but remains smaller than  $N$ ), in any case  $\lim_{N \rightarrow \infty} \sup_{\|\theta - \bar{\theta}\| \leq \delta} \|\mathbf{B}(\xi_N, \theta, \bar{\theta})\| = 0$  a.s. We have similarly  $\mathbf{M}(\xi_N, \theta) - \mathbf{M}(\xi_N, \bar{\theta}) = \sum_{x \in \mathcal{X}} [r_N(x)/N] [\mathbf{f}_{\theta}(x) \mathbf{f}_{\theta}^{\top}(x) - \mathbf{f}_{\bar{\theta}}(x) \mathbf{f}_{\bar{\theta}}^{\top}(x)]$ , and therefore  $\lim_{\delta \rightarrow 0} \sup_{\|\theta - \bar{\theta}\| \leq \delta} \|\mathbf{M}(\xi_N, \theta) - \mathbf{M}(\xi_N, \bar{\theta})\| = 0$ . Since  $\hat{\theta}_{LS}^N \xrightarrow{\text{a.s.}} \bar{\theta}$ ,  $N \rightarrow \infty$ , we obtain that there exists a.s.  $N_0$  such that for all  $N > N_0$ ,  $\lambda_{\min}[\nabla_{\theta}^2 S_N(\bar{\theta}^N)/N] > \Lambda/\tau_N$ . From the series development (41), it implies  $\|\hat{\theta}_{LS}^N - \bar{\theta}\| < [\tau_N/(\Lambda\sqrt{N})] \|\nabla_{\theta} S_N(\bar{\theta})\|/\sqrt{N}$ . We have  $\nabla_{\theta} S_N(\bar{\theta})/\sqrt{N} = -2 \sum_{x \in \mathcal{X}} \mathbf{v}_N(x)$ , where  $\mathbf{v}_N(x) = \zeta_N(x) \alpha_N(x) \mathbf{f}_{\bar{\theta}}(x)$  with  $\zeta_N(x) = (\sum_{k=1, x_k=x}^N \varepsilon_k) / \sqrt{r_N(x)}$  and  $\alpha_N(x) = \sqrt{r_N(x)/N} < 1$ , therefore  $\|\nabla_{\theta} S_N(\bar{\theta})\|/\sqrt{N} < 2\bar{f} \sum_{x \in \mathcal{X}} |\zeta_N(x)|$  with  $\bar{f} = \max_{x \in \mathcal{X}} \|\mathbf{f}_{\bar{\theta}}(x)\|$  and  $\zeta_N(x) = \mathcal{O}_p(1)$  (that is,  $\zeta_N(x)$  is bounded in probability) for all  $x$ . We thus obtain  $\|\hat{\theta}_{LS}^N - \bar{\theta}\| < (2\bar{f}\tau_N\omega_N)/(\Lambda\sqrt{N})$  for  $N > N_0$  with  $\omega_N = \sum_{x \in \mathcal{X}} |\zeta_N(x)| = \mathcal{O}_p(1)$ .

Now, for  $N$  large enough we have  $\|\mathbf{M}^{-1}(\xi_N, \bar{\theta}) \mathbf{A}(\xi_N, \theta, \bar{\theta})\| < (2\tau_N/\Lambda) \|\mathbf{A}(\xi_N, \theta, \bar{\theta})\|$  and therefore  $\|\mathbf{M}^{-1}(\xi_N, \bar{\theta}) \mathbf{A}(\xi_N, \bar{\theta}^N, \bar{\theta})\| < (4A_1 A_2 \bar{f} / \Lambda^2) (\tau_N^2 \omega_N / \sqrt{N}) \xrightarrow{P} 0$ ,  $N \rightarrow \infty$  (since  $\tau_N^2 / \sqrt{N} \rightarrow 0$  and  $\omega_N =$

$\mathcal{O}_p(1)$ ). Also,  $\|\mathbf{M}^{-1}(\xi_N, \bar{\theta})\mathbf{B}(\xi_N, \theta)\| < (2\tau_N/\Lambda)\|\mathbf{B}(\xi_N, \theta)\|$  for  $N$  large enough, which implies that  $\sup_{\|\theta - \bar{\theta}\| \leq \delta} \|\mathbf{M}^{-1}(\xi_N, \bar{\theta})\mathbf{B}(\xi_N, \theta)\| < (2A_2/\Lambda)(\tau_N\omega_N/\sqrt{N})$ , see (42), and  $\|\mathbf{M}^{-1}(\xi_N, \bar{\theta})\mathbf{B}(\xi_N, \bar{\theta}^N)\| \xrightarrow{P} 0$ ,  $N \rightarrow \infty$ . Moreover,  $\|\mathbf{M}^{-1}(\xi_N, \bar{\theta})\mathbf{M}(\xi_N, \theta) - \mathbf{I}\| = \|\mathbf{M}^{-1}(\xi_N, \bar{\theta})[\mathbf{M}(\xi_N, \theta) - \mathbf{M}(\xi_N, \bar{\theta})]\|$  and, for  $N$  large enough,  $\|\mathbf{M}^{-1}(\xi_N, \bar{\theta})\mathbf{M}(\xi_N, \theta) - \mathbf{I}\| < (2\tau_N/\Lambda) \sup_{x \in \mathcal{X}} \|\mathbf{f}_\theta(x)\mathbf{f}_\theta^\top(x) - \mathbf{f}_{\bar{\theta}}(x)\mathbf{f}_{\bar{\theta}}^\top(x)\|$ . Since  $\|\mathbf{f}_\theta(x)\mathbf{f}_\theta^\top(x) - \mathbf{f}_{\bar{\theta}}(x)\mathbf{f}_{\bar{\theta}}^\top(x)\|^2 = \|\mathbf{f}_\theta(x) - \mathbf{f}_{\bar{\theta}}(x)\|^4 + 2\{[\mathbf{f}_\theta(x) - \mathbf{f}_{\bar{\theta}}(x)]^\top \mathbf{f}_{\bar{\theta}}(x)\}^2 + 4\|\mathbf{f}_\theta(x) - \mathbf{f}_{\bar{\theta}}(x)\|^2[\mathbf{f}_\theta(x) - \mathbf{f}_{\bar{\theta}}(x)]^\top \mathbf{f}_{\bar{\theta}}(x)$  and  $\sup_{x \in \mathcal{X}} \sup_{\|\theta - \bar{\theta}\| \leq \delta} \|\mathbf{f}_\theta(x) - \mathbf{f}_{\bar{\theta}}(x)\| < A_2\delta$ , there exists  $A_3 > 0$  such that for  $\delta$  small enough we have  $\sup_{x \in \mathcal{X}} \sup_{\|\theta - \bar{\theta}\| \leq \delta} \|\mathbf{f}_\theta(x)\mathbf{f}_\theta^\top(x) - \mathbf{f}_{\bar{\theta}}(x)\mathbf{f}_{\bar{\theta}}^\top(x)\| < A_3\delta$ . We thus obtain  $\|\mathbf{M}^{-1}(\xi_N, \bar{\theta})\mathbf{M}(\xi_N, \bar{\theta}^N) - \mathbf{I}\| < (4A_3\bar{f}/\Lambda^2)(\tau_N^2\omega_N/\sqrt{N}) \xrightarrow{P} 0$ ,  $N \rightarrow \infty$  and finally  $\|\mathbf{M}^{-1/2}(\xi_N, \bar{\theta})[\nabla_\theta^2 S_N(\bar{\theta}^N)/N]\mathbf{M}^{-1/2}(\xi_N, \bar{\theta}) - 2\mathbf{I}\| \xrightarrow{P} 0$ , so that (41) gives

$$[2 + o_p(1)]\sqrt{N}\mathbf{M}^{1/2}(\xi_N, \bar{\theta})(\hat{\theta}_{LS}^N - \bar{\theta}) = -\mathbf{M}^{-1/2}(\xi_N, \bar{\theta})\nabla_\theta S_N(\bar{\theta})/\sqrt{N}, \quad N \rightarrow \infty. \quad (43)$$

We show now that  $\mathbf{M}^{-1/2}(\xi_N, \bar{\theta})\nabla_\theta S_N(\bar{\theta})/\sqrt{N}$  is asymptotically normal.

Denote by  $\mathcal{F}_k$  the  $\sigma$ -field generated by  $\{Y_1, \dots, Y_k\}$ . Notice that from the adaptive construction of the design,  $x_k$  is  $\mathcal{F}_{k-1}$ -measurable. Take any  $\mathbf{u} \in \mathbb{R}^p$  with  $\|\mathbf{u}\| = 1$  and consider

$$R_N = -\frac{1}{2\sqrt{N}}\mathbf{u}^\top \mathbf{M}^{-1/2}(\xi_N, \bar{\theta})\nabla_\theta S_N(\bar{\theta}) = \sum_{k=1}^N \zeta_{Nk}$$

where  $\zeta_{Nk} = \varepsilon_k z_{Nk}/\sqrt{N}$  with  $z_{Nk} = \mathbf{u}^\top \mathbf{M}^{-1/2}(\xi_N, \bar{\theta})\mathbf{f}_{\bar{\theta}}(x_k)$ . Using [33, Th. 1, p. 541], to prove that  $R_N \xrightarrow{d} r \sim \mathcal{N}(0, s^2)$  it is enough to show that, for any  $\gamma \in (0, 1)$ ,

$$\sum_{k=1}^N \text{Prob}\{|\zeta_{Nk}| > \gamma | \mathcal{F}_{k-1}\} \xrightarrow{P} 0 \quad (44)$$

$$\sum_{k=1}^N \mathbb{E}\{\zeta_{Nk} \mathbb{I}(|\zeta_{Nk}| \leq 1) | \mathcal{F}_{k-1}\} \xrightarrow{P} 0 \quad (45)$$

$$\sum_{k=1}^N \text{Var}\{\zeta_{Nk} \mathbb{I}(|\zeta_{Nk}| \leq \gamma) | \mathcal{F}_{k-1}\} \xrightarrow{P} s^2 \quad (46)$$

as  $N \rightarrow \infty$ , with  $\mathbb{I}(\cdot)$  the indicator function. First consider (44). Define  $t_{Nk} = \text{Prob}\{|\zeta_{Nk}| > \gamma | \mathcal{F}_{k-1}\} = \int_{|\varepsilon| > \gamma\sqrt{N}/|z_{Nk}|} dF(\varepsilon)$  with  $F(\cdot)$  the probability distribution of the errors in the model (17). Since  $|z_{Nk}| \leq \bar{f}\lambda_{\min}^{-1/2}\mathbf{M}(\xi_N, \bar{\theta}) < \bar{f}\sqrt{2\tau_N}/\sqrt{\Lambda}$  for  $N$  large enough (a.s.), with  $\bar{f} = \max_{x \in \mathcal{X}} \|\mathbf{f}_{\bar{\theta}}(x)\|$ , we get  $t_{Nk} < \int_{|\varepsilon| > \gamma\rho\sqrt{N/\tau_N}} dF(\varepsilon) < (\gamma\rho)^{-(2+\delta)}(\tau_N/N)^{1+\delta/2} \int_{|\varepsilon| > \gamma\rho\sqrt{N/\tau_N}} |\varepsilon|^{2+\delta} dF(\varepsilon)$ , where  $\rho = \sqrt{\Lambda}/(\bar{f}\sqrt{2})$ . Since

$\lim_{N \rightarrow \infty} \tau_N N^{-\delta/(2+\delta)} = 0$  and  $\mathbb{E}\{|\varepsilon_1|^{2+\delta}\} < \infty$ ,  $Nt_{Nk} \rightarrow 0$  as  $N \rightarrow \infty$  and (44) is satisfied. Consider (45) and define  $t'_{Nk} = \mathbb{E}\{\zeta_{Nk} \mathbb{I}(|\zeta_{Nk}| \leq 1) | \mathcal{F}_{k-1}\} = (z_{Nk}/\sqrt{N}) \int_{|\varepsilon| \leq \sqrt{N}/|z_{Nk}|} \varepsilon dF(\varepsilon)$ , so that  $|t'_{Nk}| = (|z_{Nk}|/\sqrt{N}) \left| \int_{|\varepsilon| > \sqrt{N}/|z_{Nk}|} \varepsilon dF(\varepsilon) \right|$  (since  $\mathbb{E}\{\varepsilon_1\} = 0$ ). Therefore, for  $N$  large enough,  $|t'_{Nk}| \leq (1/\rho) \sqrt{\tau_N/N} \int_{|\varepsilon| > \rho\sqrt{N/\tau_N}} |\varepsilon| dF(\varepsilon) < \rho^{-(2+\delta)} (\tau_N/N)^{1+\delta/2} \int_{|\varepsilon| > \rho\sqrt{N/\tau_N}} |\varepsilon|^{2+\delta} dF(\varepsilon)$ , which implies that  $N|t'_{Nk}| \rightarrow 0$  as  $N \rightarrow \infty$  and (45) is satisfied. We proceed in a similar way for (46) and define  $t''_{Nk} = \text{Var}\{\zeta_{Nk} \mathbb{I}(|\zeta_{Nk}| \leq \gamma) | \mathcal{F}_{k-1}\} = (z_{Nk}^2/N) \left[ \int_{|\varepsilon| \leq \gamma\sqrt{N}/|z_{Nk}|} \varepsilon^2 dF(\varepsilon) - \left( \int_{|\varepsilon| \leq \gamma\sqrt{N}/|z_{Nk}|} \varepsilon dF(\varepsilon) \right)^2 \right]$ . Since  $\mathbb{E}\{\varepsilon_1^2\} = 1$ , we obtain  $\sum_k t''_{Nk} = (1/N) \sum_k z_{Nk}^2 + Q_N$  where, for  $N$  large enough,  $Q_N$  satisfies  $|Q_N| < \tau_N/(N\rho^2) \sum_k \left[ \int_{|\varepsilon| > \gamma\sqrt{N}/|z_{Nk}|} \varepsilon^2 dF(\varepsilon) + \left( \int_{|\varepsilon| > \gamma\sqrt{N}/|z_{Nk}|} |\varepsilon| dF(\varepsilon) \right)^2 \right]$  (using  $\mathbb{E}\{\varepsilon_1\} = 0$ ). Therefore, for  $N$  large enough,  $|Q_N| < (\tau_N/\rho^2) \left[ \int_{|\varepsilon| > \gamma\rho\sqrt{N/\tau_N}} \varepsilon^2 dF(\varepsilon) + \left( \int_{|\varepsilon| > \gamma\rho\sqrt{N/\tau_N}} |\varepsilon| dF(\varepsilon) \right)^2 \right] < \gamma^{-\delta} \rho^{-(2+\delta)} (\tau_N^{1+\delta/2}/N^{\delta/2}) \left[ \int_{|\varepsilon| > \gamma\rho\sqrt{N/\tau_N}} |\varepsilon|^{2+\delta} dF(\varepsilon) + \left( \int_{|\varepsilon| > \gamma\rho\sqrt{N/\tau_N}} |\varepsilon|^{1+\delta/2} dF(\varepsilon) \right)^2 \right] \rightarrow 0$  as  $N \rightarrow \infty$ . Moreover,  $(1/N) \sum_k z_{Nk}^2 = \mathbf{u}^\top \mathbf{u} = 1$  and we have thus proved that  $R_N \xrightarrow{d} r \sim \mathcal{N}(0, 1)$  as  $N \rightarrow \infty$ . Since this is true for any  $\mathbf{u}$ , we have  $\mathbf{M}^{-1/2}(\xi_N, \bar{\theta}) \nabla_\theta S_N(\bar{\theta})/\sqrt{N} \xrightarrow{d} \mathbf{v} \sim \mathcal{N}(\mathbf{0}, 4\mathbf{I})$ , and therefore from (43),  $\sqrt{N} \mathbf{M}^{1/2}(\xi_N, \bar{\theta}) (\hat{\theta}_{LS}^N - \bar{\theta}) \xrightarrow{d} \mathbf{w} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ ,  $N \rightarrow \infty$ . Since  $\|\mathbf{M}^{-1}(\xi_N, \bar{\theta}) \mathbf{M}(\xi_N, \hat{\theta}_{LS}^N) - \mathbf{I}\| \xrightarrow{p} 0$  as  $N \rightarrow \infty$ , we obtain the announced result (27).

Consider now ML estimation in the model (21). We substitute  $L_N(\theta)$  for  $S_N(\theta)$  is the series development (41) (with the difference that  $L_N(\theta)$  is now *maximum* at  $\hat{\theta}_{ML}^N$ ). Denoting  $\pi_k = \pi(x_k, \theta)$  and  $\bar{\pi}_k = \pi(x_k, \bar{\theta})$ , we have  $L_N(\theta) = \sum_k Y_k \log \pi_k + (1 - Y_k) \log(1 - \pi_k)$ ,  $\nabla_\theta L_N(\theta) = \sum_k (Y_k - \pi_k) / \sqrt{\pi_k(1 - \pi_k)} \mathbf{f}_\theta(x_k)$  and  $\nabla_\theta^2 L_N(\theta)/N = -\mathbf{M}(\xi_N, \theta) + \mathbf{A}_N(\xi_N, \theta, \bar{\theta}) + \mathbf{B}_N(\xi_N, \theta, \bar{\theta})$ , where  $\mathbf{f}_\theta(x)$  is given by (22),

$$\mathbf{A}_N(\xi_N, \theta, \bar{\theta}) = \frac{1}{N} \sum_{k=1}^N \frac{\bar{\pi}_k - \pi_k}{\sqrt{\pi_k(1 - \pi_k)}} \mathbf{Q}_k(x_k, \theta) \quad \text{and} \quad \mathbf{B}_N(\xi_N, \theta, \bar{\theta}) = \frac{1}{N} \sum_{k=1}^N Z_k \frac{\sqrt{\bar{\pi}_k(1 - \bar{\pi}_k)}}{\sqrt{\pi_k(1 - \pi_k)}} \mathbf{Q}_k(x_k, \theta),$$

with

$$\mathbf{Q}_k(x_k, \theta) = \frac{1}{\sqrt{\pi_k(1 - \pi_k)}} \left[ \frac{\partial^2 \pi_k}{\partial \theta \partial \theta^\top} + (2\pi_k - 1) \mathbf{f}_\theta(x_k) \mathbf{f}_\theta^\top(x_k) \right] \quad \text{and} \quad Z_k = \frac{Y_k - \bar{\pi}_k}{\sqrt{\bar{\pi}_k(1 - \bar{\pi}_k)}}.$$

Notice that  $\mathbb{E}\{Z_k\} = 0$  and  $\mathbb{E}\{Z_k^2\} = 1$ . Similarly to the case of LS estimation, we obtain that exists a.s.  $N_0$  such that, for all  $N > N_0$ ,  $\lambda_{\min}[-\nabla_\theta^2 L_N(\hat{\theta}^N)/N] > 1/(2\tau_N)$ , where  $\hat{\theta}^N$  is now a point between  $\bar{\theta}$  and  $\hat{\theta}_{ML}^N$ . This implies that  $\|\hat{\theta}_{ML}^N - \bar{\theta}\| < c\tau_N\omega_N/\sqrt{N}$  for some constant  $c$  and  $\omega_N = \mathcal{O}_p(1)$ . We consider then  $\|\mathbf{M}^{-1}(\xi_N, \bar{\theta}) \nabla_\theta^2 L_N(\theta)/N + \mathbf{I}\|$  and obtain, since  $\lim_{N \rightarrow \infty} \tau_N/N^{1/4} = 0$ ,

$\|\mathbf{M}^{-1/2}(\xi_N, \bar{\theta})[\nabla_{\bar{\theta}}^2 L_N(\bar{\theta}^N)/N]\mathbf{M}^{-1/2}(\xi_N, \bar{\theta}) + \mathbf{I}\| \xrightarrow{P} 0, N \rightarrow \infty$ . This gives  $[1 + o_p(1)]\sqrt{N}\mathbf{M}^{1/2}(\xi_N, \bar{\theta})(\hat{\theta}_{ML}^N - \bar{\theta}) = \mathbf{M}^{-1/2}(\xi_N, \bar{\theta})\nabla_{\theta} L_N(\bar{\theta})/\sqrt{N}, N \rightarrow \infty$ . We consider finally  $R_N = (1/\sqrt{N})\mathbf{u}^\top \mathbf{M}^{-1/2}(\xi_N, \bar{\theta})\nabla_{\theta} L_N(\bar{\theta})$  for some vector  $\mathbf{u}$  with norm 1, and write  $R_N = \sum_k \zeta_{N,k}$  with now  $\zeta_{N,k} = Z_k \mathbf{u}^\top \mathbf{M}^{-1/2}(\xi_N, \bar{\theta})\mathbf{f}_{\bar{\theta}}(x_k)/\sqrt{N}$ . The properties (44, 45, 46) directly follow from the fact that  $Z_k$  is bounded,  $|Z_k| < \max_{x \in \mathcal{X}} \max\{([1 - \pi(x, \bar{\theta})]/\pi(x, \bar{\theta}))^{1/2}, (\pi(x, \bar{\theta})/[1 - \pi(x, \bar{\theta})])^{1/2}\}$  for all  $k$ . The rest of the proof is similar to the case of LS estimation.  $\blacksquare$

*Proof of Lemma 4.* First note that  $\liminf_{N \rightarrow \infty} r_{N,1:K}/N > 1/K$  since  $\mathcal{X}$  is finite, so that  $q^* \geq 1$ . Suppose that  $p \geq 2$  and  $q^* < p$ , we show that we arrive at a contradiction. The proof follows the same lines as for Lemma 1.

The property (33) is replaced by

$$\text{for all } i \leq q^*, N > N_1 \text{ and } \theta \in \Theta, \mathbf{f}_{\theta}^\top(x^{(i_N)})\mathbf{M}^{-1}(\xi_N, \theta)\mathbf{f}_{\theta}(x^{(i_N)}) - \lambda_N \phi(x^{(i_N)}, \theta) \leq \frac{\lambda_N}{\alpha}. \quad (47)$$

Define  $\rho_N$  as  $\rho_N = \lambda_N \beta_N$ , with  $\beta_N = r_{N,(q^*+1):K}/N$  as in Lemma 1. We show that  $\liminf_{N \rightarrow \infty} \rho_N \geq \underline{\rho}$  for some  $\underline{\rho} > 0$ , which contradicts the definition of  $q^*$ .

The property (37) becomes

$$\text{for all } \theta \in \Theta, \max_{i=q^*+1, \dots, K} \left\{ \mathbf{f}_{\theta}^\top(x^{(i_N)})\mathbf{M}^{-1}(\xi_N, \theta)\mathbf{f}_{\theta}(x^{(i_N)}) - \lambda_N \phi(x^{(i_N)}, \theta) \right\} > \frac{1}{\beta_N} \frac{\gamma}{LK(K - q^*)} - \lambda_N \bar{\phi}.$$

Together with (47), it gives  $N > N_1$  and  $\rho_N < \rho^* = \gamma\alpha/[LK(K - q^*)(1 + \alpha\bar{\phi})] \Rightarrow x_{N+1} \notin \mathcal{X}_N(q^*)$ . Define

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

so that  $\rho_N \geq \rho_N^* \geq \rho_N/(K - q^*)$ . Also, when  $N > N_1$ ,  $\sum_{i=1}^{q^*} r_{N,i:K} > q^*\alpha N/\lambda_N$ , so that  $\sum_{i=q^*+1}^K r_{N,i:K} < N(1 - q^*\alpha/\lambda_N)$  and  $\rho_N^* < \lambda_N(1 - q^*\alpha/\lambda_N)/(K - q^*)$ . By construction,  $\rho_N < \rho^*$  and  $N > N_1$  imply

$$\begin{aligned} \rho_{N+1} &\geq \rho_{N+1}^* = \lambda_{N+1} \frac{N\rho_N^*(K - q^*)/\lambda_N + 1}{(K - q^*)(N + 1)} = \frac{\lambda_{N+1}}{\lambda_N} \rho_N^* + \frac{\lambda_{N+1}}{N + 1} \left( \frac{1}{K - q^*} - \frac{\rho_N^*}{\lambda_N} \right) \\ &> \frac{\lambda_{N+1}}{\lambda_N} \left[ \rho_N^* + \frac{1}{N + 1} \frac{q^*\alpha}{K - q^*} \right] \geq \rho_N^* + \frac{1}{N + 1} \frac{q^*\alpha}{K - q^*} \geq \frac{\rho_N}{K - q^*} + \frac{1}{N + 1} \frac{q^*\alpha}{K - q^*}. \end{aligned}$$

The end of the proof is strictly identical to that of Lemma 1. By induction, the lower bound on  $\rho_{N+k}$  increases with  $k$ ,

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

until  $\rho_{N+k}$  becomes larger than  $\rho^*$  and  $\liminf_{N \rightarrow \infty} \rho_N > \underline{\rho} = \rho^*/(K - q^*)$ , which shows that  $q^* \geq p$  and concludes the proof.  $\blacksquare$

*Proof of Theorem 4.* We have already seen that the conditions of Lemma 4 with  $H_{\mathcal{X}}\text{-}(iii)$  (respectively,  $H_{\mathcal{X}}\text{-}(iii')$ ) imply  $\hat{\theta}_{LS}^N \xrightarrow{\text{a.s.}} \bar{\theta}$  (respectively,  $\hat{\theta}_{ML}^N \xrightarrow{\text{a.s.}} \bar{\theta}$ ) as  $N \rightarrow \infty$ . Therefore, what we first need to show is that for all  $\delta > 0$ , there exist some  $N_0 > 0$  and  $\beta > 0$  such that  $\|\hat{\theta}^N - \bar{\theta}\| < \beta$  for all  $N > N_0$  implies  $\limsup_{N \rightarrow \infty} \Phi(\xi_N, \bar{\theta}) < \phi_{\bar{\theta}}^* + \delta$ . This will prove (28).

As in the proof of Theorem 3,  $H_f\text{-}(i)$ ,  $H_{\phi}\text{-}(ii)$  and  $H_{\mathcal{X}}\text{-}(i)$  imply:  $\forall \epsilon_0, \exists \beta_1, \beta_2$  such that  $\|\theta - \bar{\theta}\| < \beta_1 \Rightarrow \max_{x \in \mathcal{X}} \|\mathbf{f}_{\theta}(x) - \mathbf{f}_{\bar{\theta}}(x)\| < \epsilon_0$  and  $\|\theta - \bar{\theta}\| < \beta_2 \Rightarrow \max_{x \in \mathcal{X}} |\phi(x, \theta) - \phi(x, \bar{\theta})| < \epsilon_0$ . Also, from Lemma 4, there exists  $N_1$  and  $\alpha > 0$  such that for all  $N > N_1$ ,  $r_{N,j;K} > \alpha N / \lambda_N$ ,  $j = 1, \dots, q^*$ , with  $q^* > p$ , and thus from  $H_{\mathcal{X}}\text{-}(iv)$ ,  $\lambda_{\min}[\mathbf{M}(\xi_N, \bar{\theta})] > \alpha \bar{\gamma} / \lambda_N$ . Direct calculations give

$$\max_{x \in \mathcal{X}} \max_{\|\theta - \bar{\theta}\| < \beta_1} |\mathbf{f}_{\theta}^{\top}(x) \mathbf{M}^{-1}(\xi_N, \theta) \mathbf{f}_{\theta}(x) - \mathbf{f}_{\bar{\theta}}^{\top}(x) \mathbf{M}^{-1}(\xi_N, \bar{\theta}) \mathbf{f}_{\bar{\theta}}(x)| < B \lambda_N \epsilon_0$$

for  $\epsilon_0$  small enough and  $N > N_1$ , with  $B$  depending on  $\alpha, \bar{\gamma}$  and  $\bar{f} = \max_{x \in \mathcal{X}} \|\mathbf{f}_{\bar{\theta}}(x)\|$ .

Therefore, for all  $\epsilon_0$  small enough, we can take  $\beta = \min\{\beta_1, \beta_2\}$  and obtain, for all  $N > N_1$  and  $\|\hat{\theta}^N - \bar{\theta}\| < \beta$  in (16),

$$\begin{aligned} & \mathbf{f}_{\bar{\theta}}^{\top}(x_{N+1}) \mathbf{M}^{-1}(\xi_N, \theta) \mathbf{f}_{\bar{\theta}}(x_{N+1}) - \lambda_N \phi(x_{N+1}, \bar{\theta}) \\ & > \mathbf{f}_{\hat{\theta}^N}^{\top}(x_{N+1}) \mathbf{M}^{-1}(\xi_N, \hat{\theta}^N) \mathbf{f}_{\hat{\theta}^N}(x_{N+1}) - \lambda_N \phi(x_{N+1}, \hat{\theta}^N) - \epsilon_0 (B + 1) \lambda_N \\ & = \max_{x \in \mathcal{X}} \left[ \mathbf{f}_{\hat{\theta}^N}^{\top}(x) \mathbf{M}^{-1}(\xi_N, \hat{\theta}^N) \mathbf{f}_{\hat{\theta}^N}(x) - \lambda_N \phi(x, \hat{\theta}^N) \right] - \epsilon_0 (B + 1) \lambda_N \\ & > \max_{x \in \mathcal{X}} \left[ \mathbf{f}_{\bar{\theta}}^{\top}(x) \mathbf{M}^{-1}(\xi_N, \bar{\theta}) \mathbf{f}_{\bar{\theta}}(x) - \lambda_N \phi(x, \bar{\theta}) \right] - 2\epsilon_0 (B + 1) \lambda_N. \end{aligned} \quad (48)$$

Suppose now that exists  $\delta > 0$  such that  $\liminf_{N \rightarrow \infty} \Phi(\xi_N, \bar{\theta}) > \phi_{\bar{\theta}}^* + \delta$ , that is, there exists  $N_2$  such that

$$\Phi(\xi_N, \bar{\theta}) > \phi_{\bar{\theta}}^* + \delta \quad (49)$$

for all  $N > N_2$ . This implies

$$\mathbf{f}_{\bar{\theta}}^{\top}(x^*) \mathbf{M}^{-1}(\xi_N, \bar{\theta}) \mathbf{f}_{\bar{\theta}}(x^*) - \lambda_N \phi(x^*, \bar{\theta}) > -\lambda_N \phi(x^*, \bar{\theta}) > \lambda_N [\delta - \Phi(\xi_N, \bar{\theta})], \quad N > N_2,$$

with  $x^*$  such that  $\phi(x^*, \bar{\theta}) = \min_{x \in \mathcal{X}} \phi(x, \bar{\theta})$ , and therefore,

$$\max_{x \in \mathcal{X}} [\mathbf{f}_{\bar{\theta}}^\top(x) \mathbf{M}^{-1}(\xi_N, \bar{\theta}) \mathbf{f}_{\bar{\theta}}(x) - \lambda_N \phi(x, \bar{\theta})] > \lambda_N [\delta - \Phi(\xi_N, \bar{\theta})], \quad N > N_2.$$

For  $\epsilon = \delta/2$ , take  $\epsilon_0 = \epsilon/[2(B+1)]$  and  $\beta$  as above and define

$$G_{\bar{\theta}}(\xi_N, \lambda_N) = \frac{1}{\lambda_N} \log \det \mathbf{M}(\xi_N, \bar{\theta}) + [C - \Phi(\xi_N, \bar{\theta})]. \quad (50)$$

From (40, 48), we obtain

$$\begin{aligned} G_{\bar{\theta}}(\xi_{N+1}, \lambda_N) - G_{\bar{\theta}}(\xi_N, \lambda_N) &> \frac{1}{\lambda_N} \log \left[ 1 + \frac{\mathbf{f}_{\bar{\theta}}^\top(x_{N+1}) \mathbf{M}^{-1}(\xi_N, \bar{\theta}) \mathbf{f}_{\bar{\theta}}(x_{N+1})}{N} \right] - \frac{p}{\lambda_N} \log \left( 1 + \frac{1}{N} \right) \\ &\quad + \frac{\delta}{2(N+1)} - \frac{1}{N+1} \frac{\mathbf{f}_{\bar{\theta}}^\top(x_{N+1}) \mathbf{M}^{-1}(\xi_N, \bar{\theta}) \mathbf{f}_{\bar{\theta}}(x_{N+1})}{\lambda_N}, \end{aligned}$$

for  $\|\hat{\theta}^N - \bar{\theta}\| < \beta$  and  $N > \max\{N_1, N_2\}$ . Since  $\lambda_{\min}[\mathbf{M}(\xi_N, \bar{\theta})] > \alpha\bar{\gamma}/\lambda_N$  for  $N > N_1$ , we also have  $\mathbf{f}_{\bar{\theta}}^\top(x_{N+1}) \mathbf{M}^{-1}(\xi_N, \bar{\theta}) \mathbf{f}_{\bar{\theta}}(x_{N+1}) < \bar{f}^2 \lambda_N / (\alpha\bar{\gamma})$ , with  $\bar{f} = \max_{x \in \mathcal{X}} \|\mathbf{f}_{\bar{\theta}}(x)\|$ . Since  $\lambda_N \rightarrow \infty$  from H $_{\lambda}$ -(ii) and  $\lambda_N/N \rightarrow 0$  from H $_{\lambda}$ -(iii) when  $N \rightarrow \infty$ , for any constant  $D < 1$  there exists  $N_3$  such that for all  $N > N_3$ ,

$$\log \left[ 1 + \frac{\mathbf{f}_{\bar{\theta}}^\top(x_{N+1}) \mathbf{M}^{-1}(\xi_N, \bar{\theta}) \mathbf{f}_{\bar{\theta}}(x_{N+1})}{N} \right] > D \frac{\mathbf{f}_{\bar{\theta}}^\top(x_{N+1}) \mathbf{M}^{-1}(\xi_N, \bar{\theta}) \mathbf{f}_{\bar{\theta}}(x_{N+1})}{N+1}$$

and

$$\frac{p}{\lambda_N} \log \left( 1 + \frac{1}{N} \right) < \frac{\delta}{4(N+1)}.$$

This gives

$$\begin{aligned} G_{\bar{\theta}}(\xi_{N+1}, \lambda_N) - G_{\bar{\theta}}(\xi_N, \lambda_N) &> \frac{\delta}{4(N+1)} + (D-1) \frac{\mathbf{f}_{\bar{\theta}}^\top(x_{N+1}) \mathbf{M}^{-1}(\xi_N, \bar{\theta}) \mathbf{f}_{\bar{\theta}}(x_{N+1})}{(N+1)\lambda_N} \\ &> \frac{1}{N+1} \left[ \frac{\delta}{4} + (D-1) \frac{\bar{f}^2}{\alpha\bar{\gamma}} \right] > \frac{\delta}{8(N+1)} \end{aligned}$$

for  $N > \max\{N_1, N_2, N_3\}$  and  $\|\hat{\theta}^N - \bar{\theta}\| < \beta$  when choosing  $D > 1 - \delta\alpha\bar{\gamma}/(8\bar{f}^2)$ . Consider now

$$G_{\bar{\theta}}(\xi_{N+1}, \lambda_{N+1}) - G_{\bar{\theta}}(\xi_N, \lambda_N) = \left[ \frac{1}{\lambda_{N+1}} - \frac{1}{\lambda_N} \right] \log \det \mathbf{M}(\xi_{N+1}, \bar{\theta}) + G_{\bar{\theta}}(\xi_N, \lambda_{N+1}) - G_{\bar{\theta}}(\xi_N, \lambda_N).$$

For  $N$  large enough and  $\|\hat{\theta}^N - \bar{\theta}\| < \beta$ , it satisfies

$$G_{\bar{\theta}}(\xi_{N+1}, \lambda_{N+1}) - G_{\bar{\theta}}(\xi_N, \lambda_N) > \left[ \frac{1}{\lambda_{N+1}} - \frac{1}{\lambda_N} \right] \log \det \mathbf{M}(\xi_{N+1}, \bar{\theta}) + \frac{\delta}{8(N+1)}.$$

Denote  $L^* = \max_{\xi \in \Xi} \log \det \mathbf{M}(\xi, \bar{\theta})$ . Since  $\lambda_N$  is non-decreasing from  $H_\lambda$ -(ii) and  $\lambda_N/N$  is non-increasing from  $H_\lambda$ -(iii), we get

$$G_{\bar{\theta}}(\xi_{N+1}, \lambda_{N+1}) - G_{\bar{\theta}}(\xi_N, \lambda_N) > \frac{\delta}{8(N+1)} - \left[ \frac{1}{\lambda_N} - \frac{1}{\lambda_{N+1}} \right] L^* \geq \frac{\delta}{8(N+1)} - \frac{L^*}{\lambda_N(N+1)} > \frac{\delta}{16(N+1)}$$

for  $N$  large enough, in contradiction with the fact that  $G_{\bar{\theta}}(\xi, \lambda)$  is bounded for  $\lambda > 1$ , see (50).

Therefore, for all  $\delta > 0$  we can find a  $\beta$  such that  $\|\hat{\theta}^N - \bar{\theta}\| < \beta$  for all  $N > N_0$  implies that there exists a subsequence  $\xi_{N_t}$  such that  $\liminf_{t \rightarrow \infty} \Phi(\xi_{N_t}, \bar{\theta}) < \phi_{\bar{\theta}}^* + \delta$ . First note that  $H_\phi$ -(i) implies  $\Phi(\xi_{N+1}, \bar{\theta}) - \Phi(\xi_N, \bar{\theta}) = 1/(N+1) [\phi(x_{N+1}, \bar{\theta}) - \Phi(\xi_N, \bar{\theta})] < \bar{\phi}/(N+1)$  so that there exists  $N_4$  such that  $\Phi(\xi_{N+1}, \bar{\theta}) < \Phi(\xi_N, \bar{\theta}) + \delta/2$  for all  $N > N_4$ . Also, from the developments above, there exists  $N_5$  such that for  $N > N_5$ ,  $\|\hat{\theta}^N - \bar{\theta}\| < \beta$  and  $\Phi(\xi_N, \bar{\theta}) > \phi_{\bar{\theta}}^* + \delta$  imply  $G_{\bar{\theta}}(\xi_{N+1}, \lambda_{N+1}) > G_{\bar{\theta}}(\xi_N, \lambda_N)$ . Moreover, since  $\lambda_{\min}[\mathbf{M}(\xi_N, \bar{\theta})] > \alpha\bar{\gamma}/\lambda_N$  for all  $N > N_1$ , we have  $p \log(\alpha\bar{\gamma}) - p \log \lambda_N < \log \det \mathbf{M}(\xi_N, \bar{\theta}) < L^*$  and thus

$$\frac{|\log \det \mathbf{M}(\xi_N, \bar{\theta})|}{\lambda_N} < \frac{\delta}{4} \quad (51)$$

for  $N$  larger than some  $N_6$ . Take any  $N_t > \max(N_0, N_4, N_5, N_6)$  and such that  $\Phi(\xi_{N_t}, \bar{\theta}) \leq \phi_{\bar{\theta}}^* + \delta$ . We show that  $\Phi(\xi_N, \bar{\theta}) < \phi_{\bar{\theta}}^* + 2\delta$  for all  $N > N_t$  until the next  $N'$  such that  $\Phi(\xi_{N'}, \bar{\theta}) \leq \phi_{\bar{\theta}}^* + \delta$ . We have  $\Phi(\xi_{N_t+1}, \bar{\theta}) < \phi_{\bar{\theta}}^* + (3/2)\delta$ . If  $\Phi(\xi_{N_t+1}, \bar{\theta}) \leq \phi_{\bar{\theta}}^* + \delta$  we take  $N' = N_t + 1$ . Suppose that  $\Phi(\xi_{N_t+1}, \bar{\theta}) > \phi_{\bar{\theta}}^* + \delta$ . Then,  $G_{\bar{\theta}}(\xi_{N_t+2}, \lambda_{N_t+2}) > G_{\bar{\theta}}(\xi_{N_t+1}, \lambda_{N_t+1})$  and, from the definition of  $G_{\bar{\theta}}(\xi, \lambda)$  and (51),  $\Phi(\xi_{N_t+2}, \bar{\theta}) < \Phi(\xi_{N_t+1}, \bar{\theta}) + \delta/2 < \phi_{\bar{\theta}}^* + 2\delta$ . If  $\Phi(\xi_{N_t+2}, \bar{\theta}) \leq \phi_{\bar{\theta}}^* + \delta$  we take  $N' = N_t + 2$ . Otherwise, we have  $G_{\bar{\theta}}(\xi_{N_t+3}, \lambda_{N_t+3}) > G_{\bar{\theta}}(\xi_{N_t+2}, \lambda_{N_t+2}) > G_{\bar{\theta}}(\xi_{N_t+1}, \lambda_{N_t+1})$  and thus  $\Phi(\xi_{N_t+3}, \bar{\theta}) < \Phi(\xi_{N_t+1}, \bar{\theta}) + \delta/2 < \phi_{\bar{\theta}}^* + 2\delta$ . By induction,  $\Phi(\xi_N, \bar{\theta}) < \phi_{\bar{\theta}}^* + 2\delta$  for all  $N > N_t$ . This concludes the proof of (28).

Define  $B_N(\beta) = \{x_i : \|x_i - x^*\| > \beta, i = 1, \dots, N\}$  and denote by  $b_N(\beta)$  the number of elements of  $B_N(\beta)$ . Assume that there exists  $\beta > 0$  such that  $\limsup_{N \rightarrow \infty} b_N(\beta)/N > \gamma > 0$ . From  $H_\phi$ -(iii), this implies that there exists  $\epsilon > 0$  such that  $\limsup_{N \rightarrow \infty} \Phi(\xi_N, \bar{\theta}) - \phi_{\bar{\theta}}^* > \gamma\epsilon$ , which contradicts (28). Therefore, for all  $\beta > 0$ ,  $\lim_{N \rightarrow \infty} b_N(\beta)/N = 0$  which gives (29).  $\blacksquare$

## References

- [1] K. Chaloner and K. Larntz. Optimal Bayesian design applied to logistic regression experiments. *Journal of Statistical Planning and Inference*, 21:191–208, 1989.
- [2] P. Chaudhuri and P.A. Mykland. Nonlinear experiments: optimal design and inference based likelihood. *Journal of the American Statistical Association*, 88(422):538–546, 1993.
- [3] P. Chaudhuri and P.A. Mykland. On efficiently designing of nonlinear experiments. *Statistica Sinica*, 5:421–440, 1995.
- [4] D. Cook and V.V. Fedorov. Constrained optimization of experimental design (invited discussion paper). *Statistics*, 26:129–178, 1995.
- [5] V. Dragalin and V. Fedorov. Adaptive designs for dose-finding based on efficacy-toxicity response. *Journal of Statistical Planning and Inference*, 136:1800–1823, 2006.
- [6] V. Dragalin, V. Fedorov, and Y. Wu. Adaptive designs for selecting drug combinations based on efficacy-toxicity response. *Journal of Statistical Planning and Inference*, 138:352–373, 2008.
- [7] S. Durham, N. Flournoy, and W. Li. A sequential design for maximizing the probability of a favorable response. *Can. Journal Statist.*, 26:479–495, 1998.
- [8] V.V. Fedorov. Convex design theory. *Math. Operationsforsch. Statist., Ser. Statistics*, 11(3):403–413, 1980.
- [9] V.V. Fedorov and P. Hackl. *Model-Oriented Design of Experiments*. Springer, Berlin, 1997.
- [10] I. Ford and S.D. Silvey. A sequentially constructed design for estimating a nonlinear parametric function. *Biometrika*, 67(2):381–388, 1980.
- [11] A. Giovagnoli. Asymptotic properties of biased coin designs for treatment allocation. In A. Di Bucchianico, H. Läuter, and H.P. Wynn, editors, *mODa'7 – Advances in Model-Oriented Design*

- and Analysis, *Proceedings of the 7th Int. Workshop, Heeze (Netherlands)*, pages 81–88, Heidelberg, June 2004. Physica Verlag.
- [12] A. Giovagnoli and N. Pintacuda. Properties of frequency distributions induced by general ‘up-and-down’ methods for estimating quantiles. *Journal of Statistical Planning and Inference*, 74:51–63, 1998.
- [13] L.M. Haines, I. Perevozskaya, and W.F. Rosenberger. Bayesian optimal designs in Phase I clinical trials. *Biometrics*, 59:591–600, 2003.
- [14] J. Hardwick and Q. Stout. Optimizing a unimodal response function for binary variables. In A. Atkinson, B. Bogacka, and A. Zhigljavsky, editors, *Optimum Design 2000*, chapter 18, pages 195–210. Kluwer, Dordrecht, 2001.
- [15] R. Harman and L. Pronzato. Improvements on removing non-optimal support points in D-optimum design algorithms. *Statistics & Probability Letters*, 77:90–94, 2007.
- [16] I. Hu. On sequential designs in nonlinear problems. *Biometrika*, 85(2):496–503, 1998.
- [17] A. Ivanova. A new dose-finding design for bivariate outcomes. *Biometrics*, 59:1001–1007, 2003.
- [18] R.I. Jennrich. Asymptotic properties of nonlinear least squares estimation. *Annals of Math. Stat.*, 40:633–643, 1969.
- [19] J. Kiefer and J. Wolfowitz. The equivalence of two extremum problems. *Canadian Journal of Mathematics*, 12:363–366, 1960.
- [20] E.E. Kpamegan and N. Flournoy. An optimizing up-and-down design. In A. Atkinson, B. Bogacka, and A. Zhigljavsky, editors, *Optimum Design 2000*, chapter 19, pages 211–224. Kluwer, Dordrecht, 2001.
- [21] T.L. Lai. Asymptotic properties of nonlinear least squares estimates in stochastic regression models. *Annals of Statistics*, 22(4):1917–1930, 1994.

- [22] T.L. Lai and C.Z. Wei. Least squares estimates in stochastic regression models with applications to identification and control of dynamic systems. *Annals of Statistics*, 10(1):154–166, 1982.
- [23] V.B. Melas. Optimal designs for exponential regressions. *Math. Operationsforsch. und Statist., Ser. Statistics*, 9:753–768, 1978.
- [24] J. Mikulecká. On a hybrid experimental design. *Kybernetika*, 19(1):1–14, 1983.
- [25] W.G. Müller and B.M. Pötscher. Batch sequential design for a nonlinear estimation problem. In V.V. Fedorov, W.G. Müller, and I.N. Vuchkov, editors, *Model-Oriented Data Analysis II, Proceedings 2nd IIASA Workshop, St Kyrik (Bulgaria), May 1990*, pages 77–87. Physica Verlag, Heidelberg, 1992.
- [26] L. Pronzato. Adaptive optimisation and  $D$ -optimum experimental design. *Annals of Statistics*, 28(6):1743–1761, 2000.
- [27] L. Pronzato. One-step ahead adaptive  $D$ -optimal design on a finite design space is asymptotically optimal. Technical Report I3S/RR-2008-14-FR, 22 pages, Laboratoire I3S, CNRS–Université de Nice-Sophia Antipolis, 06903 Sophia Antipolis, France, 2008. <http://www.i3s.unice.fr/~mh/RR/rapports.html>, submitted.
- [28] L. Pronzato and E. Thierry. Sequential experimental design and response optimisation. *Statistical Methods and Applications*, 11(3):277–292, 2003.
- [29] L. Pronzato and E. Walter. Robust experiment design via stochastic approximation. *Mathematical Biosciences*, 75:103–120, 1985.
- [30] L. Pronzato and E. Walter. Robust experiment design via maximin optimization. *Mathematical Biosciences*, 89:161–176, 1988.
- [31] F. Pukelsheim. *Optimal Experimental Design*. Wiley, New York, 1993.
- [32] W.F. Rosenberger, N. Flournoy, and S.D. Durham. Asymptotic normality of maximum likelihood estimators for multiparameter response-driven designs. *Journal of Statistical Planning and Inference*, 60:69–76, 1997.

- [33] A.N. Shiryaev. *Probability*. Springer, Berlin, 1996.
- [34] J. Volaufová, S.M. Redmann, Jr., and L. LaMotte. Is logistic regression the best for binomial response? In *Tatra Mountains Mathematical Publications*, volume 26, pages 237–246, 2003.
- [35] C.F.J. Wu. Asymptotic theory of nonlinear least squares estimation. *Annals of Statistics*, 9(3):501–513, 1981.
- [36] C.F.J. Wu. Asymptotic inference from sequential design in a nonlinear situation. *Biometrika*, 72(3):553–558, 1985.
- [37] C.F.J. Wu and H.P. Wynn. The convergence of general step-length algorithms for regular optimum design criteria. *Annals of Statistics*, 6(6):1273–1285, 1978.
- [38] H.P. Wynn. The sequential generation of  $D$ -optimum experimental designs. *Annals of Math. Stat.*, 41:1655–1664, 1970.
- [39] Y. Yi and X. Wang. Asymptotically efficient estimation in response adaptive trials. *Journal of Statistical Planning and Inference*, 138:2899–2905, 2008.