Sequential online subsampling for thinning experimental designs

Luc Pronzato* and HaiYing Wang†

August 5, 2020

Abstract

We consider a design problem where experimental conditions (design points X_i) are presented in the form of a sequence of i.i.d. random variables, generated with an unknown probability measure μ, and only a given proportion $\alpha \in (0,1)$ can be selected. The objective is to select good candidates X_i on the fly and maximize a concave function Φ of the corresponding information matrix. The optimal solution corresponds to the construction of an optimal bounded design measure $\xi^*_\alpha \le \mu/\alpha$, with the difficulty that μ is unknown and ξ^*_α must be constructed online. The construction proposed relies on the definition of a threshold τ on the directional derivative of Φ at the current information matrix, the value of τ being fixed by a certain quantile of the distribution of this directional derivative. Combination with recursive quantile estimation yields a nonlinear two-time-scale stochastic approximation method. It can be applied to very long design sequences since only the current information matrix and estimated quantile need to be stored. Convergence to an optimum design is proved. Various illustrative examples are presented.

Keywords: Active learning, data thinning, design of experiments, sequential design, subsampling
AMS subject classifications: 62K05, 62L05, 62L20, 68Q32

1 Introduction

Consider a rather general parameter estimation problem in a model with independent observations $Y_i = Y_i(x_i)$ conditionally on the experimental variables x_i, with x_i in some set $\mathscr{X}$. Suppose that for any $x \in \mathscr{X}$ there exists a measurable set $\mathscr{Y}_x \in \mathbb{R}$ and a σ-finite measure μ_x on $\mathscr{Y}_x$ such that $Y(x)$ has the density $\varphi_{x,\overline{\boldsymbol{\theta}}}$ with respect to μ_x, with $\overline{\boldsymbol{\theta}}$ the true value of the model parameters $\boldsymbol{\theta}$ to be estimated, $\boldsymbol{\theta} \in \mathbb{R}^p$. In particular, this covers the case of regression models, with μ_x the Lebesgue measure on $\mathscr{Y}_x = \mathbb{R}$ and $Y(x) = \eta(\overline{\boldsymbol{\theta}}, x) + \varepsilon(x)$, where the $\varepsilon(x_i)$ are independently distributed with zero mean and known variance σ_i^2 (or unknown but constant variance σ^2), and the case of generalized linear models, with $\varphi_{x,\overline{\boldsymbol{\theta}}}$ in the exponential family and logistic regression as a special case. Denoting by $\hat{\boldsymbol{\theta}}^n$ the estimated value of $\boldsymbol{\theta}$ from data (X_i, Y_i), $i = 1, \ldots, n$, under rather weak conditions on the x_i and $\varphi_{x,\overline{\boldsymbol{\theta}}}$, see below, we have

$$\sqrt{n}(\hat{\boldsymbol{\theta}}^n - \overline{\boldsymbol{\theta}}) \xrightarrow{\mathrm{d}} \mathscr{N}(\mathbf{0}, \mathbf{M}^{-1}(\xi, \overline{\boldsymbol{\theta}})) \ \text{ as } n \to \infty\,, \tag{1.1}$$

*CNRS, Université Côte d'Azur, I3S, France, Luc.Pronzato@cnrs.fr (corresponding author)
†Department of Statistics, University of Connecticut, USA, haiying.wang@uconn.edu

where $\mathbf{M}(\xi, \boldsymbol{\theta})$ denotes the (normalized) Fisher information matrix for parameters $\boldsymbol{\theta}$ and (asymptotic) design ξ (that is, a probability measure on $\mathscr{X}$),

$$\begin{aligned}\mathbf{M}(\xi, \boldsymbol{\theta}) &= \lim_{n\to\infty} \frac{1}{n} \mathsf{E}_{x_1,\ldots,x_n,\boldsymbol{\theta}} \left\{ \sum_{i=1}^{n} \frac{\partial \log \varphi_{x,\boldsymbol{\theta}}(Y_i)}{\partial \boldsymbol{\theta}} \sum_{j=1}^{n} \frac{\partial \log \varphi_{x,\boldsymbol{\theta}}(Y_j)}{\partial \boldsymbol{\theta}^\top} \right\} \\ &= \int_{\mathscr{X}} \left[\int_{\mathscr{Y}_x} \frac{\partial \log \varphi_{x,\boldsymbol{\theta}}(y)}{\partial \boldsymbol{\theta}} \frac{\partial \log \varphi_{x,\boldsymbol{\theta}}(y)}{\partial \boldsymbol{\theta}^\top} \varphi_{x,\boldsymbol{\theta}}(y)\, \mu_x(\mathrm{d}y) \right] \xi(\mathrm{d}x)\,.\end{aligned}$$

This is true in particular for randomized designs such that the x_i are independently sampled from ξ, and for asymptotically discrete designs, such that ξ is a discrete measure on $\mathscr{X}$ and the empirical design measure $\xi_n = \sum_{i=1}^{n} \delta_{x_i}$ converges strongly to ξ; see Pronzato and Pázman (2013). The former case corresponds to the situation considered here. The choice of μ_x is somewhat arbitrary, provided that $\int_{\mathscr{Y}_x} \varphi_{x,\overline{\boldsymbol{\theta}}}(y)\, \mu_x(\mathrm{d}y) = 1$ for all x, and we shall assume that $\mu_x(\mathrm{d}y) \equiv 1$. We can then write

$$\mathbf{M}(\xi, \boldsymbol{\theta}) = \int_{\mathscr{X}} \mathscr{M}(x, \boldsymbol{\theta})\, \xi(\mathrm{d}x)\,, \text{ where } \mathscr{M}(x, \boldsymbol{\theta}) = \int_{\mathscr{Y}_x} \frac{\partial \log \varphi_{x,\overline{\boldsymbol{\theta}}}(y)}{\partial \boldsymbol{\theta}} \frac{\partial \log \varphi_{x,\overline{\boldsymbol{\theta}}}(y)}{\partial \boldsymbol{\theta}^\top} \varphi_{x,\overline{\boldsymbol{\theta}}}(y)\, \mathrm{d}y$$

denotes the elementary information matrix at x.

Taking motivation from (1.1), optimal experimental design (approximate theory) aims at choosing a measure ξ that minimizes a scalar function of the asymptotic covariance matrix $\mathbf{M}^{-1}(\xi, \overline{\boldsymbol{\theta}})$ of $\hat{\boldsymbol{\theta}}^n$, or equivalently, that maximizes a function Φ of $\mathbf{M}(\xi, \overline{\boldsymbol{\theta}})$. For a nonlinear model $\mathscr{M}(x, \boldsymbol{\theta})$ and $\mathbf{M}(\xi, \boldsymbol{\theta})$ depend on the model parameters $\boldsymbol{\theta}$. Since $\overline{\boldsymbol{\theta}}$ is unknown, the standard approach is local, and consists in constructing an optimal design for a nominal value $\boldsymbol{\theta}_0$ of $\boldsymbol{\theta}$. This is the point of view we shall adopt here — although sequential estimation of $\boldsymbol{\theta}$ is possible, see Section 6. When $\boldsymbol{\theta}$ is fixed at some $\boldsymbol{\theta}_0$, there is fundamentally no difference with experimental design in a linear model for which $\mathscr{M}(x, \boldsymbol{\theta})$ and $\mathbf{M}(\xi, \boldsymbol{\theta})$ do not depend on $\boldsymbol{\theta}$. For example, in the linear regression model

$$Y(X_i) = \mathbf{f}^\top(X_i)\overline{\boldsymbol{\theta}} + \varepsilon_i\,,$$

where the errors ε_i are independent and identically distributed (i.i.d.), with a density φ_ε with respect to the Lebesgue measure having finite Fisher information for location $I_\varepsilon = \int_{\mathbb{R}} \left\{ [\varphi'_\varepsilon(t)]^2 / \varphi_\varepsilon(t) \right\} \mathrm{d}t < \infty$ ($I_\varepsilon = 1/\sigma^2$ for normal errors $\mathscr{N}(0, \sigma^2)$), then $\mathscr{M}(x) = I_\varepsilon\, \mathbf{f}(x)\mathbf{f}^\top(x)$, $\mathbf{M}(\xi) = I_\varepsilon \int_{\mathscr{X}} \mathbf{f}(x)\mathbf{f}^\top(x)\, \xi(\mathrm{d}x)$. Polynomial regression provides typical examples of such a situation and will be used for illustration in Section 4. The construction of an optimal design measure ξ^* maximizing $\Phi[\mathbf{M}(\xi, \theta_0)]$ usually relies on the application of a specialized algorithm to a discretization of the design space $\mathscr{X}$; see, e.g., Pronzato and Pázman (2013, Chap. 9).

With the rapid development of connected sensors and the pervasive usage of computers, there exist more and more situations where extraordinary amounts of massive data (X_i, Y_i), $i = 1, \ldots, N$, are available to construct models. When N is very large, using all the data to construct $\hat{\boldsymbol{\theta}}^N$ is then unfeasible, and selecting the most informative subset through the construction of an n-point optimal design, $n \ll N$, over the discrete set $\mathscr{X}_N = \{X_i, i = 1, \ldots, N\}$ is also not feasible. The objective of this paper is to present a method to explore $\mathscr{X}_N$ sequentially and select a proportion $n = \lfloor \alpha N \rfloor$ of the N data points to be used to estimate $\boldsymbol{\theta}$. Each candidate X_i is considered only once, which allows very large datasets to be processed: when the X_i are i.i.d. and are received sequentially, they can be selected on the fly which makes the method applicable to data streaming; when N data points are available simultaneously, a random permutation allows $\mathscr{X}_N$ to be processed as an i.i.d. sequence. When N is too large for the storage capacity and the i.i.d. assumption is not tenable, interleaving or scrambling techniques can be used. Since de-scrambling is not necessary here (the

objective is only to randomize the sequence), a simple random selection in a fixed size buffer may be sufficient; an example is presented in Section 4.3.

The method is based on the construction of an optimal bounded design measure and draws on the paper (Pronzato, 2006). In that paper, the sequential selection of the X_i relies on a threshold set on the directional derivative of the design criterion, given by the $(1-\alpha)$-quantile of the distribution of this derivative. At stage k, all previous X_i, $i = 1, \ldots, k$, are used for the estimation of the quantile C_k that defines the threshold for the possible selection of the candidate X_{k+1}. In the present paper, we combine this approach with the recursive estimation of C_k, following (Tierney, 1983): as a result, the construction is fully sequential and only requires to record the current value of the information matrix $\mathbf{M}_k$ and of the estimated quantile $\widehat{C}_k$ of the distribution of the directional derivative. It relies on a reinterpretation of the approach in (Pronzato, 2006) as a stochastic approximation method for the solution of the necessary and sufficient optimality conditions for a bounded design measure, which we combine with another stochastic approximation method for quantile estimation to obtain a two-time-scale stochastic approximation scheme.

The paper is organized as follows. Section 2 introduces the notation and assumptions and recalls main results on optimal bounded design measures. Section 3 presents our subsampling algorithm based on a two-time-scale stochastic approximation procedure and contains the main result of the paper. Several illustrative examples are presented in Section 4. We are not aware of any other method for thinning experimental designs that is applicable to data streaming; nevertheless, in Section 5 we compare our algorithm with an exchange method and with the IBOSS algorithm of Wang et al. (2019) in the case where the N design points are available and can be processed simultaneously. Section 6 concludes and suggests a few directions for further developments. A series of technical results are provided in the Appendix.

2 Optimal bounded design measures

2.1 Notation and assumptions

Suppose that X is distributed with the probability measure μ on $\mathscr{X} \subseteq \mathbb{R}^d$, a subset of $\mathbb{R}^d$ with nonempty interior, with $d \geq 1$. For any $\xi \in \mathscr{P}^+(\mathscr{X})$, the set of positive measure ξ on $\mathscr{X}$ (not necessarily of mass one), we denote $\mathbf{M}(\xi) = \int_{\mathscr{X}} \mathscr{M}(x)\, \xi(\mathrm{d}x)$ where, for all x in $\mathscr{X}$, $\mathscr{M}(x) \in \mathbb{M}^{\geq}$, the set (cone) of symmetric non-negative definite $p \times p$ matrices. We assume that $p > 1$ in the rest of the paper (the optimal selection of information in the case $p = 1$ forms a variant of the secretary problem for which an asymptotically optimal solution can be derived, see Albright and Derman (1972); Pronzato (2001)).

We denote by $\Phi : \mathbb{M}^{\geq} \to \mathbb{R} \cup \{-\infty\}$ the design criterion we wish to maximize, and by $\lambda_{\min}(\mathbf{M})$ and $\lambda_{\max}(\mathbf{M})$ the minimum and maximum eigenvalues of $\mathbf{M}$, respectively; we shall use the ℓ_2 norm for vectors and Frobenius norm for matrices, $\|\mathbf{M}\| = \mathrm{trace}^{1/2}[\mathbf{M}\mathbf{M}^\top]$; all vectors are column vectors. For any $t \in \mathbb{R}$, we denote $[t]^+ = \max\{t, 0\}$ and, for any $t \in \mathbb{R}^+$, $\lfloor t \rfloor$ denotes the largest integer smaller than t. For $0 \leq \ell \leq L$ we denote by $\mathbb{M}^{\geq}_{\ell,L}$ the (convex) set defined by

$$\mathbb{M}^{\geq}_{\ell,L} = \{\mathbf{M} \in \mathbb{M}^{\geq} : \ell < \lambda_{\min}(\mathbf{M}) \text{ and } \lambda_{\max}(\mathbf{M}) < L\}\,,$$

and by $\mathbb{M}^{>}$ the open cone of symmetric positive definite $p \times p$ matrices. We make the following assumptions on Φ.

$\mathbf{H}_\Phi$ Φ is strictly concave on $\mathbb{M}^{>}$, linearly differentiable and increasing for Loewner ordering; its gradient $\nabla_\Phi(\mathbf{M})$ is well defined in $\mathbb{M}^{\geq}$ for any $\mathbf{M} \in \mathbb{M}^{>}$ and satisfies $\|\nabla_\Phi(\mathbf{M})\| < A(\ell)$ and

$\lambda_{\min}[\nabla_\Phi(\mathbf{M})] > a(L)$ for any $\mathbf{M} \in \mathbb{M}^{\geq}_{\ell,L}$, for some $a(L) > 0$ and $A(\ell) < \infty$; moreover, ∇_Φ satisfies the following Lipschitz condition: for all $\mathbf{M}_1$ and $\mathbf{M}_2$ in $\mathbb{M}^{\geq}$ such that $\lambda_{\min}(\mathbf{M}_i) > \ell > 0$, $i = 1, 2$, there exists $K_\ell < \infty$ such that $\|\nabla_\Phi(\mathbf{M}_2) - \nabla_\Phi(\mathbf{M}_1)\| < K_\ell \|\mathbf{M}_2 - \mathbf{M}_1\|$.

The criterion $\Phi_0(\mathbf{M}) = \log\det(\mathbf{M})$ and criteria $\Phi_q(\mathbf{M}) = -\operatorname{trace}(\mathbf{M}^{-q})$, $q \in (-1, \infty)$, $q \neq 0$, with $\Phi_q(\mathbf{M}) = -\infty$ if $\mathbf{M}$ is singular, which are often used in optimal design (in particular with q a positive integer) satisfy H_Φ; see, e.g., Pukelsheim (1993, Chap. 6). Their gradients are $\nabla_{\Phi_0}(\mathbf{M}) = \mathbf{M}^{-1}$ and $\nabla_{\Phi_q}(\mathbf{M}) = q\,\mathbf{M}^{-(q+1)}$, $q \neq 0$; the constants $a(L)$ and $A(\ell)$ are respectively given by $a(L) = 1/L$, $A(\ell) = \sqrt{p}/\ell$ for Φ_0 and $a(L) = q/L^{q+1}$, $A(\ell) = q\sqrt{p}/\ell^{q+1}$ for Φ_q. The Lispchitz condition follows from the fact that the criteria are twice differentiable on $\mathbb{M}^{>}$. The positively homogeneous versions $\Phi_0^+(\mathbf{M}) = \det^{1/p}(\mathbf{M})$ and $\Phi_q^+(\mathbf{M}) = [(1/p)\operatorname{trace}(\mathbf{M}^{-q})]^{-1/q}$, which satisfy $\Phi^+(a\mathbf{M}) = a\,\Phi^+(\mathbf{M})$ for any $a > 0$ and any $\mathbf{M} \in \mathbb{M}^{\geq}$, and $\Phi^+(\mathbf{I}_p) = 1$, with $\mathbf{I}_p$ the $p \times p$ identity matrix, could be considered too; see Pukelsheim (1993, Chaps. 5, 6). The strict concavity of Φ implies that, for any convex subset $\widehat{\mathbb{M}}$ of $\mathbb{M}^{>}$, there exists a unique matrix $\mathbf{M}^*$ maximizing $\Phi(\mathbf{M})$ with respect to $\mathbf{M} \in \widehat{\mathbb{M}}$.

We denote by $F_\Phi(\mathbf{M}, \mathbf{M}')$ the directional derivative of Φ at $\mathbf{M}$ in the direction $\mathbf{M}'$,

$$F_\Phi(\mathbf{M}, \mathbf{M}') = \lim_{\gamma \to 0^+} \frac{\Phi[(1-\gamma)\mathbf{M} + \gamma\mathbf{M}'] - \Phi(\mathbf{M})}{\gamma} = \operatorname{trace}[\nabla_\Phi(\mathbf{M})(\mathbf{M}' - \mathbf{M})]\,,$$

and we make the following assumptions on μ and $\mathscr{M}$.

H_μ μ has a bounded positive density φ with respect to the Lebesgue measure on every open subset of $\mathscr{X}$.

$\mathrm{H}_\mathscr{M}$ (i) $\mathscr{M}$ is continuous on $\mathscr{X}$ and satisfies $\int_\mathscr{X} \|\mathscr{M}(x)\|^2\,\mu(\mathrm{d}x) < B < \infty$;

(ii) for any $\mathscr{X}_\epsilon \subset \mathscr{X}$ of measure $\mu(\mathscr{X}_\epsilon) = \epsilon > 0$, $\lambda_{\min}\left\{\int_{\mathscr{X}_\epsilon} \mathscr{M}(x)\,\mu(\mathrm{d}x)\right\} > \ell_\epsilon$ for some $\ell_\epsilon > 0$.

Since all the designs considered will be formed by points sampled from μ, we shall confound $\mathscr{X}$ with the support of μ: $\mathscr{X} = \{x \in \mathbb{R}^d : \mu(\mathscr{B}_d(x, \epsilon)) > 0 \quad \forall \epsilon > 0\}$, with $\mathscr{B}_d(x, \epsilon)$ the open ball with center x and radius ϵ. Notice that $\mathrm{H}_\mathscr{M}$-(i) implies that $\lambda_{\max}[\mathbf{M}(\mu)] < \sqrt{B}$ and $\|\mathbf{M}(\mu)\| < \sqrt{p\,B}$.

Our sequential selection procedure will rely on the estimation of the $(1-\alpha)$-quantile $C_{1-\alpha}(\mathbf{M})$ of the distribution $F_\mathbf{M}(z)$ of the directional derivative $Z_\mathbf{M}(X) = F_\Phi[\mathbf{M}, \mathscr{M}(X)]$ when $X \sim \mu$, and we shall assume that $\mathrm{H}_{\mu,\mathscr{M}}$ below is satisfied. It implies in particular that $C_{1-\alpha}(\mathbf{M})$ is uniquely defined by $F_\mathbf{M}(C_{\mathbf{M},1-\alpha}) = 1 - \alpha$.

$\mathrm{H}_{\mu,\mathscr{M}}$ For all $\mathbf{M} \in \mathbb{M}^{\geq}_{\ell,L}$, $F_\mathbf{M}$ has a uniformly bounded density $\varphi_\mathbf{M}$; moreover, for any $\alpha \in (0,1)$, there exists $\epsilon_{\ell,L} > 0$ such that $\varphi_\mathbf{M}[C_{1-\alpha}(\mathbf{M})] > \epsilon_{\ell,L}$ and $\varphi_\mathbf{M}$ is continuous at $C_{1-\alpha}(\mathbf{M})$.

$\mathrm{H}_{\mu,\mathscr{M}}$ is overrestricting (we only need the existence and boundedness of $\varphi_\mathbf{M}$, and its positiveness and continuity at $C_{1-\alpha}(\mathbf{M})$), but is satisfied is many common situations; see Section 4 for examples. Let us emphasize that H_μ and $\mathrm{H}_\mathscr{M}$ are not enough to guarantee the existence of a density $\varphi_\mathbf{M}$, since $\operatorname{trace}[\nabla_\Phi(\mathbf{M})\mathscr{M}(x)]$ may remain constant over subsets of $\mathscr{X}$ having positive measure. Assuming the existence of $\varphi_\mathbf{M}$ and the continuity of φ on $\mathscr{X}$ is also insufficient, since $\varphi_\mathbf{M}$ is generally not continuous when $Z_\mathbf{M}(x)$ is not differentiable in x, and $\varphi_\mathbf{M}$ is not necessarily bounded.

2.2 Optimal design

As mentioned in introduction, when the cardinality of $\mathscr{X}_N$ is very large, one may wish to select only n candidates X_i among the N available, a fraction $n = \lfloor \alpha N \rfloor$ say, with $\alpha \in (0,1)$. For any $n \leq N$, we denote by $\mathbf{M}^*_{n,N}$ a design matrix (non necessarily unique) obtained by selecting n points optimally within $\mathscr{X}_N$; that is, $\mathbf{M}^*_{n,N}$ gives the maximum of $\Phi(\mathbf{M}_n)$ with respect to $\mathbf{M}_n = (1/n)\sum_{j=1}^n \mathscr{M}(X_{i_j})$, where the X_{i_j} are n distinct points in $\mathscr{X}_N$. Note that this forms a difficult combinatorial problem, unfeasible for large n and N. If one assumes that the X_i are i.i.d., with μ their probability measure on $\mathscr{X}$, for large N the optimal selection of $n = \lfloor \alpha N \rfloor$ points amounts at constructing an optimal bounded design measure ξ^*_α, such that $\Phi[\mathbf{M}(\xi^*_\alpha)]$ is maximum and $\xi_\alpha \leq \mu/\alpha$ (in the sense $\xi_\alpha(\mathcal{A}) \leq \mu(\mathcal{A})/\alpha$ for any μ-measurable set $\mathcal{A}$, which makes ξ_α absolutely continuous with respect to μ). Indeed, Lemma A.1 in Appendix A indicates that $\limsup_{N\to\infty} \Phi(\mathbf{M}^*_{\lfloor \alpha N \rfloor, N}) = \Phi[\mathbf{M}(\xi^*_\alpha)]$. Also, under H_Φ, $\mathsf{E}\{\Phi(\mathbf{M}^*_{n,N})\} \leq \Phi[\mathbf{M}(\xi^*_{n/N})]$ for all $N \geq n > 0$; see Pronzato (2006, Lemma 3).

A key result is that, when all subsets of $\mathscr{X}$ with constant $Z_{\mathbf{M}}(x)$ have zero measure, $Z_{\mathbf{M}(\xi^*_\alpha)}(x) = F_\Phi[\mathbf{M}(\xi^*_\alpha), \mathscr{M}(x)]$ separates two sets $\mathscr{X}^*_\alpha$ and $\mathscr{X} \setminus \mathscr{X}^*_\alpha$, with $F_\Phi[\mathbf{M}(\xi^*_\alpha), \mathscr{M}(x)] \geq C^*_{1-\alpha}$ and $\xi^*_\alpha = \mu/\alpha$ on $\mathscr{X}^*_\alpha$, and $F_\Phi[\mathbf{M}(\xi^*_\alpha), \mathscr{M}(x)] \leq C^*_{1-\alpha}$ and $\xi^*_\alpha = 0$ on $\mathscr{X} \setminus \mathscr{X}^*_\alpha$, for some constant $C^*_{1-\alpha}$; moreover, $\int_{\mathscr{X}} F_\Phi[\mathbf{M}(\xi^*_\alpha), \mathscr{M}(x)]\, \xi^*_\alpha(\mathrm{d}x) = \int_{\mathscr{X}^*_\alpha} F_\Phi[\mathbf{M}(\xi^*_\alpha), \mathscr{M}(x)]\, \mu(\mathrm{d}x) = 0$; see Wynn (1982); Fedorov (1989) and Fedorov and Hackl (1997, Chap. 4). (The condition mentioned in those references is that μ has no atoms, but the example in Section 4.4.2 will show that this is not sufficient; extension to arbitrary measures is considered in (Sahm and Schwabe, 2001).)

For $\alpha \in (0,1)$, denote

$$\mathbb{M}(\alpha) = \left\{ \mathbf{M}(\xi_\alpha) = \int_{\mathscr{X}} \mathscr{M}(x)\, \xi_\alpha(\mathrm{d}x) : \xi_\alpha \in \mathscr{P}^+(\mathscr{X}),\, \xi_\alpha \leq \frac{\mu}{\alpha},\, \int_{\mathscr{X}} \xi_\alpha(\mathrm{d}x) = 1 \right\}.$$

In (Pronzato, 2006), it is shown that, for any $\mathbf{M} \in \mathbb{M}^>$,

$$\mathbf{M}^+(\mathbf{M}, \alpha) = \arg \max_{\mathbf{M}' \in \mathbb{M}(\alpha)} F_\Phi(\mathbf{M}, \mathbf{M}') = \frac{1}{\alpha} \int_{\mathscr{X}} \mathbb{I}_{\{F_\Phi[\mathbf{M}, \mathscr{M}(x)] \geq C_{1-\alpha}\}}\, \mathscr{M}(x)\, \mu(\mathrm{d}x)\,, \tag{2.1}$$

where, for any proposition $\mathcal{A}$, $\mathbb{I}_{\{\mathcal{A}\}} = 1$ if $\mathcal{A}$ is true and is zero otherwise, and $C_{1-\alpha} = C_{1-\alpha}(\mathbf{M})$ is an $(1-\alpha)$-quantile of $F_\Phi[\mathbf{M}, \mathscr{M}(X)]$ when $X \sim \mu$ and satisfies

$$\int_{\mathscr{X}} \mathbb{I}_{\{F_\Phi[\mathbf{M}, \mathscr{M}(x)] \geq C_{1-\alpha}(\mathbf{M})\}}\, \mu(\mathrm{d}x) = \alpha\,. \tag{2.2}$$

Therefore, $\mathbf{M}^*_\alpha = \mathbf{M}(\xi^*_\alpha)$ is the optimum information matrix in $\mathbb{M}(\alpha)$ (unique since Φ is strictly concave) if and only if it satisfies $\max_{\mathbf{M}' \in \mathbb{M}(\alpha)} F_\Phi(\mathbf{M}^*_\alpha, \mathbf{M}') = 0$, or equivalently $\mathbf{M}^*_\alpha = \mathbf{M}^+(\mathbf{M}^*_\alpha, \alpha)$, and the constant $C^*_{1-\alpha}$ equals $C_{1-\alpha}(\mathbf{M}^*_\alpha)$; see (Pronzato, 2006, Th. 5); see also Pronzato (2004). Note that $C^*_{1-\alpha} \leq 0$ since $\int_{\mathscr{X}} F_\Phi[\mathbf{M}(\xi^*_\alpha), \mathscr{M}(x)]\, \xi^*_\alpha(\mathrm{d}x) = 0$ and $F_\Phi[\mathbf{M}(\xi^*_\alpha), \mathscr{M}(x)] \geq C^*_{1-\alpha}$ on the support of ξ^*_α.

3 Sequential construction of an optimal bounded design measure

3.1 A stochastic approximation problem

Suppose that the X_i are i.i.d. with μ. The solution of $\mathbf{M} = \mathbf{M}^+(\mathbf{M}, \alpha)$, $\alpha \in (0,1)$, with respect to $\mathbf{M}$ by stochastic approximation yields the iterations

$$\begin{aligned} n_{k+1} &= n_k + \mathbb{I}_{\{F_\Phi[\mathbf{M}_{n_k}, \mathscr{M}(X_{k+1})] \geq C_{1-\alpha}(\mathbf{M}_{n_k})\}}\,, \\ \mathbf{M}_{n_{k+1}} &= \mathbf{M}_{n_k} + \tfrac{1}{n_k+1}\, \mathbb{I}_{\{F_\Phi[\mathbf{M}_{n_k}, \mathscr{M}(X_{k+1})] \geq C_{1-\alpha}(\mathbf{M}_{n_k})\}}\, [\mathscr{M}(X_{k+1}) - \mathbf{M}_{n_k}]\,. \end{aligned} \tag{3.1}$$

Note that $\mathsf{E}\left\{\mathbb{I}_{\{F_\Phi[\mathbf{M},\mathscr{M}(X)]\geq C_{1-\alpha}(\mathbf{M})\}}\,[\mathscr{M}(X)-\mathbf{M}]\right\}=\alpha\,[\mathbf{M}^+(\mathbf{M},\alpha)-\mathbf{M}]$. The almost sure (a.s.) convergence of $\mathbf{M}_{n_k}$ in (3.1) to $\mathbf{M}(\xi^*_\alpha)$ that maximizes $\Phi(\mathbf{M})$ with respect $\mathbf{M}\in\mathbb{M}(\alpha)$ is proved in (Pronzato, 2006) under rather weak assumptions on Φ, $\mathscr{M}$ and μ.

The construction (3.1) requires the calculation of the $(1-\alpha)$-quantile $C_{1-\alpha}(\mathbf{M}_{n_k})$ for all n_k, see (2.2), which is not feasible when μ is unknown and has a prohibitive computational cost when we know μ. For that reason, it is proposed in (Pronzato, 2006) to replace $C_{1-\alpha}(\mathbf{M}_{n_k})$ by the empirical quantile $\widetilde{C}_{\alpha,k}(\mathbf{M}_{n_k})$ that uses the empirical measure $\mu_k=(1/k)\sum_{i=1}^k\delta_{X_i}$ of the X_i that have been observed up to stage k. This construction preserves the a.s. convergence of $\mathbf{M}_{n_k}$ to $\mathbf{M}(\xi^*_\alpha)$ in (3.1), but its computational cost and storage requirement increase with k, which makes it unadapted to situations with very large N. The next section considers the recursive estimation of $C_{1-\alpha}(\mathbf{M}_{n_k})$ and contains the main result of the paper.

3.2 Recursive quantile estimation

The idea is to plug a recursive estimator of the $(1-\alpha)$-quantile $C_{1-\alpha}(\mathbf{M}_{n_k})$ in (3.1). Under mild assumptions, for random variables Z_i that are i.i.d. with distribution function F such that the solution of the equation $F(z)=1-\alpha$ is unique, the recursion

$$\widehat{C}_{k+1}=\widehat{C}_k+\frac{\beta}{k+1}\left(\mathbb{I}_{\{Z_{k+1}\geq\widehat{C}_k\}}-\alpha\right) \tag{3.2}$$

with $\beta>0$ converges a.s. to the quantile $C_{1-\alpha}$ such that $F(C_{1-\alpha})=1-\alpha$. Here, we shall use a construction based on (Tierney, 1983). In that paper, a clever dynamical choice of $\beta=\beta_k$ is shown to provide the optimal asymptotic rate of convergence of $\widehat{C}_k$ towards $C_{1-\alpha}$, with $\sqrt{k}(\widehat{C}_k-C_{1-\alpha})\xrightarrow{\mathrm{d}}\mathscr{N}(0,\alpha(1-\alpha)/f^2(C_{1-\alpha}))$ as $k\to\infty$, where $f(z)=\mathrm{d}F(z)/\mathrm{d}z$ is the p.d.f. of the Z_i — note that it coincides with the asymptotic behavior of the sample (empirical) quantile. The only conditions on F are that $f(z)$ exists for all z and is uniformly bounded, and that f is continuous and positive at the unique root $C_{1-\alpha}$ of $F(z)=1-\alpha$.

There is a noticeable difference, however, with the estimation of $C_{1-\alpha}(\mathbf{M}_{n_k})$: in our case we need to estimate a quantile of $Z_k(X)=F_\Phi[\mathbf{M}_{n_k},\mathscr{M}(X)]$ for $X\sim\mu$, with the distribution of $Z_k(X)$ evolving with k. For that reason, we shall impose a faster dynamic to the evolution of $\widehat{C}_k$, and replace (3.2) by

$$\widehat{C}_{k+1}=\widehat{C}_k+\frac{\beta_k}{(k+1)^q}\left(\mathbb{I}_{\{Z_k(X_{k+1})\geq\widehat{C}_k\}}-\alpha\right) \tag{3.3}$$

for some $q\in(0,1)$. The combination of (3.3) with (3.1) yields a particular nonlinear two-time-scale stochastic approximation scheme. There exist advanced results on the convergence of linear two-time-scale stochastic approximation, see Konda and Tsitsiklis (2004); Dalal et al. (2018). To the best of our knowledge, however, there are few results on convergence for nonlinear schemes. Convergence is shown in (Borkar, 1997) under the assumption of boundedness of the iterates using the ODE method of Ljung (1977); sufficient conditions for stability are provided in (Lakshminarayanan and Bhatnagar, 2017), also using the ODE approach. In the proof of Theorem 3.1 we provide justifications for our construction, based on the analyses and results in the references mentioned above.

The construction is summarized in Algorithm 1 below. The presence of the small number ϵ_1 is only due to technical reasons: setting $z_{k+1}=+\infty$ when $n_k/k<\epsilon_1$ in (3.4) has the effect of always selecting X_{k+1} when less than $\epsilon_1\,k$ points have been selected previously; it ensures that $n_{k+1}/k>\epsilon_1$

for all k and thus that $\mathbf{M}_{n_k}$ always belongs to $\mathbb{M}^{\geq}_{\ell,L}$ for some $\ell > 0$ and $L < \infty$; see Lemma B.1 in Appendix.

Algorithm 1: sequential selection (α given).

0) Choose $k_0 \geq p$, $q \in (1/2, 1)$, $\gamma \in (0, q - 1/2)$, and $0 < \epsilon_1 \ll \alpha$.

1) Initialization: select $X_1, \ldots, X_{k_0}$, compute $\mathbf{M}_{n_{k_0}} = (1/k_0) \sum_{i=1}^{k_0} \mathscr{M}(X_i)$. If $\mathbf{M}_{n_{k_0}}$ is singular, increase k_0 and select the next points until $\mathbf{M}_{n_{k_0}}$ has full rank. Set $k = n_k = k_0$, the number of points selected.

   Compute $\zeta_i = Z_{k_0}(X_i)$, for $i = 1, \ldots, k_0$ and order the ζ_i as $\zeta_{1:k_0} \leq \zeta_{2:k_0} \leq \cdots \leq \zeta_{k_0:k_0}$; denote $k_0^+ = \lceil (1 - \alpha/2)\, k_0 \rceil$ and $k_0^- = \max\{\lfloor (1 - 3\,\alpha/2)\, k_0 \rfloor, 1\}$.

   Initialize $\widehat{C}_{k_0}$ at $\zeta_{\lceil (1-\alpha)\, k_0 \rceil : k_0}$; set $\beta_0 = k_0/(k_0^+ - k_0^-)$, $h = (\zeta_{k_0^+:k_0} - \zeta_{k_0^-:k_0})$, $h_{k_0} = h/k_0^\gamma$ and $\widehat{f}_{k_0} = \left[\sum_{i=1}^{k_0} \mathbb{I}_{\{|\zeta_i - \widehat{C}_{k_0}| \leq h_{k_0}\}}\right] / (2\, k_0\, h_{k_0})$.

2) Iteration $k+1$: collect X_{k+1} and compute $Z_k(X_{k+1}) = F_\Phi[\mathbf{M}_{n_k}, \mathscr{M}(X_{k+1})]$.

$$\begin{array}{ll} \text{If } n_k/k < \epsilon_1 & \text{set } z_{k+1} = +\infty\,; \\ \text{otherwise} & \text{set } z_{k+1} = Z_k(X_{k+1})\,. \end{array} \tag{3.4}$$

   If $z_{k+1} \geq \widehat{C}_k$, update n_k into $n_{k+1} = n_k + 1$ and $\mathbf{M}_{n_k}$ into

$$\mathbf{M}_{n_{k+1}} = \mathbf{M}_{n_k} + \frac{1}{n_k + 1}\, [\mathscr{M}(X_{k+1}) - \mathbf{M}_{n_k}]\,; \tag{3.5}$$

   otherwise, set $n_{k+1} = n_k$.

3) Compute $\beta_k = \min\{1/\widehat{f}_k, \beta_0\, k^\gamma\}$; update $\widehat{C}_k$ using (3.3).

   Set $h_{k+1} = h/(k+1)^\gamma$ and update $\widehat{f}_k$ into

$$\widehat{f}_{k+1} = \widehat{f}_k + \frac{1}{(k+1)^q} \left[\frac{1}{2\, h_{k+1}}\, \mathbb{I}_{\{|Z_k(X_{k+1}) - \widehat{C}_k| \leq h_{k+1}\}} - \widehat{f}_k\right].$$

4) $k \leftarrow k + 1$, return to Step 2.

Note that $\widehat{C}_k$ is updated whatever the value of $Z_k(X_{k+1})$. Recursive quantile estimation by (3.3) follows (Tierney, 1983). To ensure a faster dynamic for the evolution of $\widehat{C}_k$ than for $\mathbf{M}_{n_k}$, we take $q < 1$ instead of $q = 1$ in (Tierney, 1983), and the construction of $\widehat{f}_k$ and the choices of β_k and h_k are modified accordingly. Following the same arguments as in the proof of Proposition 1 of (Tierney, 1983), the a.s. convergence of $\widehat{C}_k$ to $C_{1-\alpha}$ in the modified version of (3.2) is proved in Theorem C.1 (Appendix C).

The next theorem establishes the convergence of the combined stochastic approximation schemes with two time-scales.

Theorem 3.1. *Under H_Φ, H_μ, $H_{\mathscr{M}}$ and $H_{\mu,\mathscr{M}}$, the normalized information matrix $\mathbf{M}_{n_k}$ corresponding to the n_k candidates selected after k iterations of Algorithm 1 converges a.s. to the optimal matrix $\mathbf{M}^*_\alpha$ in $\mathbb{M}(\alpha)$ as $k \to \infty$.*

Proof. Our analysis is based on (Borkar, 1997). We denote by $\mathscr{F}_n$ the increasing sequence of σ-fields generated by the X_i. According to (3.3), we can write $\widehat{C}_{k+1} = \widehat{C}_k + [\beta_k/(k+1)^q]\, V_{k+1}$ with $V_{k+1} = \mathbb{I}_{\{Z_k(X_{k+1}) \geq \widehat{C}_k\}} - \alpha$. Therefore, $\mathsf{E}\{V_{k+1}|\mathscr{F}_k\} = \int_{\mathscr{X}} [\mathbb{I}_{\{Z_k(x) \geq \widehat{C}_k\}} - \alpha]\, \mu(\mathrm{d}x)$ and $\mathsf{var}\{V_{k+1}|\mathscr{F}_k\} = F_k(\widehat{C}_k)[1 - F_k(\widehat{C}_k)]$, with F_k the distribution function of $Z_k(X)$. From Lemma B.1 (Appendix B) and $\mathrm{H}_{\mu,\mathscr{M}}$, F_k has a well defined density f_k for all k, with $f_k(t) > 0$ for all t and f_k bounded. The first part of the proof of Theorem C.1 applies (see Appendix C): $\widehat{f}_k$ is a.s. bounded and β_k is bounded away from zero a.s. Therefore, $\sum_k \beta_k/(k+1)^q \to \infty$ a.s. and $(k+1)^q/[\beta_k\,(k+1)] \to 0$ a.s.; also, $\sum_k [\beta_k/(k+1)^q]^2 < \infty$ since $q - \gamma > 1/2$.

The o.d.e. associated with (3.3), for a fixed matrix $\mathbf{M}$ and thus a fixed $Z(\cdot)$, such that $Z(X) = F_\Phi[\mathbf{M}, \mathscr{M}(X)]$ has the distribution function F and density f, is

$$\frac{\mathrm{d}C(t)}{\mathrm{d}t} = 1 - F[C(t)] - \alpha = F(C_{1-\alpha}) - F[C(t)]\,,$$

where $C_{1-\alpha} = C_{1-\alpha}(\mathbf{M})$ satisfies $F(C_{1-\alpha}) = 1 - \alpha$. Consider the Lyapunov function $L(C) = [F(C) - F(C_{1-\alpha})]^2$. It satisfies $\mathrm{d}L[C(t)]/\mathrm{d}t = -2\, f[C(t)]\, L[C(t)] \leq 0$, with $\mathrm{d}L[C(t)]/\mathrm{d}t = 0$ if and only if $C = C_{1-\alpha}$. Moreover, $C_{1-\alpha}$ is Lipschitz continuous in $\mathbf{M}$; see Lemma D.1 in Appendix D. The conditions for Theorem 1.1 in (Borkar, 1997) are thus satisfied concerning the iterations for $\widehat{C}_k$.

Denote $\widehat{\mathbf{M}}_k = \mathbf{M}_{n_k}$ and $\rho_k = n_k/k$, so that (3.4) implies $k\,\rho_k \geq \epsilon_1\,(k-1)$ for all k; see Lemma B.1 in Appendix. They satisfy

$$\rho_{k+1} = \rho_k + \frac{R_{k+1}}{k+1} \text{ and } \widehat{\mathbf{M}}_{k+1} = \widehat{\mathbf{M}}_k + \frac{\mathbf{\Omega}_{k+1}}{k+1}\,, \tag{3.6}$$

where $R_{k+1} = \mathbb{I}_{\{Z_k(X_{k+1}) \geq \widehat{C}_k\}} - \rho_k$, and $\mathbf{\Omega}_{k+1} = (1/\rho_{k+1})\, \mathbb{I}_{\{Z_k(X_{k+1}) \geq \widehat{C}_k\}} \left[\mathscr{M}(X_{k+1}) - \widehat{\mathbf{M}}_k\right]$. We have $\mathsf{E}\{R_{k+1}|\mathscr{F}_k\} = \int_{\mathscr{X}} \mathbb{I}_{\{Z_k(x) \geq \widehat{C}_k\}}\, \mu(\mathrm{d}x) - \rho_k$ and $\mathsf{var}\{R_{k+1}|\mathscr{F}_k\} = F_k(\widehat{C}_k)[1 - F_k(\widehat{C}_k)]$, with F_k the distribution function of $Z_k(X)$, which, from $\mathrm{H}_{\mu,\mathscr{M}}$, has a well defined density f_k for all k. Also,

$$\mathsf{E}\{\mathbf{\Omega}_{k+1}|\mathscr{F}_k\} \;=\; \frac{\mathcal{I}_k}{\rho_k + \frac{1-\rho_k}{k+1}} = \frac{1 + 1/k}{\rho_k + 1/k}\, \mathcal{I}_k\,,$$

where $\mathcal{I}_k = \int_{\mathscr{X}} \mathbb{I}_{\{Z_k(x) \geq \widehat{C}_k\}} \left[\mathscr{M}(x) - \widehat{\mathbf{M}}_k\right]\, \mu(\mathrm{d}x)$. Denote $\Delta_{k+1} = \mathbf{\Omega}_{k+1} - \mathcal{I}_k/\rho_k$, so that

$$\widehat{\mathbf{M}}_{k+1} = \widehat{\mathbf{M}}_k + \frac{1}{k+1}\,\frac{\mathcal{I}_k}{\rho_k} + \frac{\Delta_{k+1}}{k+1}\,. \tag{3.7}$$

We get $\mathsf{E}\{\Delta_{k+1}|\mathscr{F}_k\} = (\rho_k - 1)/[\rho_k\,(k\,\rho_k + 1)]\,\mathcal{I}_k$ and

$$\begin{aligned}
\mathsf{var}\{\{\Delta_{k+1}\}_{i,j}|\mathscr{F}_k\} \;&=\; \mathsf{var}\{\{\mathbf{\Omega}_{k+1}\}_{i,j}|\mathscr{F}_k\} \\
&=\; \frac{(k+1)^2}{(k\,\rho_k + 1)^2} \left[\int_{\mathscr{X}} \mathbb{I}_{\{Z_k(x) \geq \widehat{C}_k\}}\, \{\mathscr{M}(x) - \widehat{\mathbf{M}}_k\}^2_{i,j}\, \mu(\mathrm{d}x) - \{\mathcal{I}_k\}^2_{i,j}\right]\,,
\end{aligned}$$

where (3.4) implies that $\rho_k > \epsilon_1/2$, and therefore $(k+1)\,(1-\rho_k)/[\rho_k\,(k\,\rho_k + 1)] < 4/\epsilon_1^2$, and $\mathsf{var}\{\{\Delta_{k+1}\}_{i,j}|\mathscr{F}_k\}$ is a.s. bounded from $\mathrm{H}_{\mathscr{M}}$-(i). This implies that $\sum_k \Delta_{k+1}/(k+1) < \infty$ a.s. The limiting o.d.e. associated with (3.6) and (3.7) are

$$\begin{aligned}
\frac{\mathrm{d}\rho(t)}{\mathrm{d}t} \;&=\; \int_{\mathscr{X}} \mathbb{I}_{\{F_\Phi[\widehat{\mathbf{M}}(t), \mathscr{M}(x)] \geq C_{1-\alpha}[\widehat{\mathbf{M}}(t)]\}}\, \mu(\mathrm{d}x) - \rho(t) = \alpha - \rho(t)\,, \\
\frac{\mathrm{d}\widehat{\mathbf{M}}(t)}{\mathrm{d}t} \;&=\; \frac{1}{\rho(t)} \int_{\mathscr{X}} \mathbb{I}_{\{F_\Phi[\widehat{\mathbf{M}}(t), \mathscr{M}(x)] \geq C_{1-\alpha}[\widehat{\mathbf{M}}(t)]\}} \left[\mathscr{M}(x) - \widehat{\mathbf{M}}(t)\right]\, \mu(\mathrm{d}x) \\
&=\; \frac{\alpha}{\rho(t)} \left\{\mathbf{M}^+[\widehat{\mathbf{M}}(t), \alpha] - \widehat{\mathbf{M}}(t)\right\}\,,
\end{aligned}$$

where $\mathbf{M}^+[\widehat{\mathbf{M}}(t), \alpha]$ is defined by (2.1). The first equation implies that $\rho(t)$ converges exponentially fast to α, with $\rho(t) = \alpha + [\rho(0) - \alpha] \exp(-t)$; the second equation gives

$$\frac{\mathrm{d}\Phi[\widehat{\mathbf{M}}(t)]}{\mathrm{d}t} = \mathrm{trace}\left[\nabla_\Phi[\widehat{\mathbf{M}}(t)]\,\frac{\mathrm{d}\widehat{\mathbf{M}}(t)}{\mathrm{d}t}\right] = \frac{\alpha}{\rho(t)} \max_{\mathbf{M}' \in \mathbb{M}(\alpha)} F_\Phi[\widehat{\mathbf{M}}(t), \mathbf{M}'] \geq 0\,,$$

with a strict inequality if $\widehat{\mathbf{M}}(t) \neq \mathbf{M}^*_\alpha$, the optimal matrix in $\mathbb{M}(\alpha)$. The conditions of Theorem 1.1 in (Borkar, 1997) are thus satisfied, and $\widehat{\mathbf{M}}_k$ converges to $\mathbf{M}^*_\alpha$ a.s. ■

Remark 3.1.

(i) Algorithm 1 does not require the knowledge of μ and has minimum storage requirements: apart for the current matrix $\mathbf{M}_{n_k}$, we only need to update the scalar variables $\widehat{C}_k$ and f_k. Its complexity is $\mathcal{O}(d^3\,N)$ in general, considering that the complexity of the calculation of $F_\Phi[\mathbf{M}, \mathscr{M}(X)]$ is $\mathcal{O}(d^3)$. It can be reduced to $\mathcal{O}(d^2\,N)$ when $\mathscr{M}(X)$ has rank one and $\mathbf{M}^{-1}_{n_k}$ is updated instead of $\mathbf{M}_{n_k}$ (see remark *(iii)* below), for D-optimality and Φ_q-optimality with q integer; see Section 2.1. Very long sequences (X_i) can thus be processed.

(ii) Numerical simulations indicate that we do not need to take $q < 1$ in Algorithm 1: (3.3) with $q = 1$ yields satisfactory performance, provided the step-size obeys Kersten's rule and does not decrease at each iteration.

(iii) The substitution of $\mathrm{trace}[\nabla_\Phi(\mathbf{M})\mathscr{M}(X)]$ for $F_\Phi[\mathbf{M}, \mathscr{M}(X)] = \mathrm{trace}\{\nabla_\Phi(\mathbf{M})[\mathscr{M}(X) - \mathbf{M}]\}$ everywhere does not change the behavior of the algorithm. When $\nabla_\Phi(\mathbf{M})$ only depends on $\mathbf{M}^{-1}$ (which is often the case for classical design criteria, see the discussion following the presentation of H_Φ), and if $\mathscr{M}(X)$ is a low rank matrix, it may be preferable to update $\mathbf{M}^{-1}_{n_k}$ instead of $\mathbf{M}_{n_k}$, thereby avoiding matrix inversions. For example, if $\mathscr{M}(X_{k+1}) = I_\varepsilon\,\mathbf{f}(X_{k+1})\mathbf{f}^\top(X_{k+1})$, then, instead of updating (3.5), it is preferable to update the following

$$\mathbf{M}^{-1}_{n_{k+1}} = \left(1 + \frac{1}{n_k}\right)\left[\mathbf{M}^{-1}_{n_k} - \frac{I_\varepsilon\,\mathbf{M}^{-1}_{n_k}\mathbf{f}(X_{k+1})\mathbf{f}^\top(X_{k+1})\mathbf{M}^{-1}_{n_k}}{n_k + I_\varepsilon\,\mathbf{f}^\top(X_{k+1})\mathbf{M}^{-1}_{n_k}\mathbf{f}(X_{k+1})}\right].$$

Low-rank updates of the Cholesky decomposition of the matrix can be considered too.

(iv) Algorithm 1 can be adapted to the case where the number of iterations is fixed (equal to the size N of the candidate set $\mathscr{X}_N$) and the number of candidates n to be selected is imposed. A straightforward modification is to introduce truncation and forced selection: we run the algorithm with $\alpha = n/N$ and, at Step 2, we set $z_{k+1} = -\infty$ (reject X_{k+1}) if $n_k \geq n$ and set $z_{k+1} = +\infty$ (select X_{k+1}) if $n - n_k \geq N - k$. However, this may induce the selection of points X_i carrying little information when k approaches N in case n_k is excessively small. For that reason, adaptation of α to n_k, obtained by substituting $\alpha_k = (n - n_k)/(N - k)$ for the constant α everywhere, seems preferable. This is illustrated by an example in Section 4.2.

(v) The case when μ has discrete components (atoms), or more precisely when there exist subsets of $\mathscr{X}$ of positive measure where $Z_\mathbf{M}(x)$ is constant (see Section 4.4.2), requires additional technical developments which we do not detail here.

A first difficulty is that $\mathrm{H}_\mathscr{M}$-*(ii)* may not be satisfied when the matrices $\mathscr{M}(x)$ do not have full rank, unless we only consider large enough ϵ. Unless ϵ_1 in (3.4) is large enough, Lemma B.1 is not valid, and other arguments are required in the proof of Theorem 3.1. Possible remedies

may consist (a) in adding a regularization matrix $\gamma \mathbf{I}_p$ with a small γ to all matrices $\mathscr{M}(x)$ (which amounts at considering optimal design for Bayesian estimation with a vague prior; see, e.g., Pilz (1983)), or (b) in replacing the condition in (3.4) by [If $\lambda_{\min}(\mathbf{M}_{n_k}) < \epsilon_1$, set $z_{k+1} = +\infty$].

A second difficulty is that $C_{1-\alpha}(\mathbf{M}^*_\alpha)$ may correspond to a point of discontinuity of the distribution function of $F_\Phi[\mathbf{M}^*_\alpha, \mathscr{M}(X)]$. The estimated value f_k of the density of $F_\Phi[\mathbf{M}_{n_k}, \mathscr{M}(X)]$ at $\widehat{C}_k$ (Step 3 of Algorithm 1) may then increase to infinity and β_k tend to zero in (3.3). This can be avoided by setting $\beta_k = \max\{\epsilon_2, \min(1/\widehat{f}_k, \beta_0\, k^\gamma)\}$ for some $\epsilon_2 > 0$.

In (Pronzato, 2006), where empirical quantiles are used, measures needed to be taken to avoid the acceptance of too many points, for instance based on the adaptation of α through $\alpha_k = (n - n_k)/(N - k)$, see remark (iv) above, or via the addition of the extra condition [if $n_k/k > \alpha$, set $z_{k+1} = -\infty$] to (3.4) in case n is not specified. Such measures do not appear to be necessary when quantiles are estimated by (3.3); see the examples in Section 4.4. ◁

4 Examples

We always take $k_0 = 5\,p$, $q = 5/8$, $\gamma = 1/10$ in Algorithm 1 (our simulations indicate that these choices are not critical); we also set $\epsilon_1 = 0$.

4.1 Example 1: quadratic regression with normal independent variables

Take $\mathscr{M}(x) = \mathbf{f}(x)\mathbf{f}^\top(x)$, with $\mathbf{f}(x) = (1,\ x,\ x^2)^\top$ and $\Phi(\mathbf{M}) = \log\det(\mathbf{M})$, and let the X_i be i.i.d. standard normal variables $\mathscr{N}(0,1)$. The D-optimal design for x in an interval $[t, t']$ corresponds to $\xi^* = (1/3)\ \big(\delta_t + \delta_{(t+t')/2} + \delta_{t'}\big)$. In the data thinning problem, the optimal solution corresponds to the selection of X_i in the union of three intervals; that is, with the notation of Section 2, $\mathscr{X}^*_\alpha = (-\infty, -a] \cup [-b, b] \cup [a, \infty)$. The values of a and b are obtained by solving the pair of equations $\int_0^b \varphi(x)\,\mathrm{d}x + \int_a^\infty \varphi(x)\,\mathrm{d}x = \alpha/2$ and $\mathrm{trace}[\mathbf{M}^{-1}(\xi)\mathscr{M}(a)] = \mathrm{trace}[\mathbf{M}^{-1}(\xi)\mathscr{M}(b)]$, with φ the standard normal density and $\mathbf{M}(\xi) = [\int_{-\infty}^{-a} \mathscr{M}(x)\,\varphi(x)\,\mathrm{d}x + \int_{-b}^{b} \mathscr{M}(x)\,\varphi(x)\,\mathrm{d}x + \int_a^\infty \mathscr{M}(x)\,\varphi(x)\,\mathrm{d}x]/\alpha$.

We set the horizon N at 100 000 and consider the two cases $\alpha = 1/2$ and $\alpha = 1/10$. In each case we keep α constant but apply the rule of Remark 3.1-(iv) (truncation/forced selection) to select exactly $n = 50\,000$ and $n = 10\,000$ design points, respectively. For $\alpha = 1/2$, we have $a \simeq 1.0280$, $b \simeq 0.2482$, and $\Phi^*_\alpha = \Phi(\mathbf{M}^*_\alpha) = \max_{\mathbf{M}\in\mathbb{M}(\alpha)} \Phi(\mathbf{M}) \simeq 1.6354$, $C_{1-\alpha}(\mathbf{M}^*_\alpha) \simeq -1.2470$; when $\alpha = 1/10$, we have $a \simeq 1.8842$, $b \simeq 0.0507$, and $\Phi^*_\alpha \simeq 3.2963$, $C_{1-\alpha}(\mathbf{M}^*_\alpha) \simeq -0.8513$. The figures below present results obtained for one simulation (i.e., one random set $\mathscr{X}_N$), but they are rather typical in the sense that different $\mathscr{X}_N$ yield similar behaviors.

Figure 1 shows a smoothed histogram (Epanechnikov kernel, bandwidth equal to 1/1000 of the range of the X_i in $\mathscr{X}_N$) of the design points selected by Algorithm 1, for $\alpha = 1/2$ (left) and $\alpha = 1/10$ (right). There is good adequation with the theoretical optimal density, which corresponds to a truncation of the normal density at values indicated by the vertical dotted lines.

Figure 2 presents the evolution of $\Phi(\mathbf{M}_{n_k})$ as a function of k, together with the optimal value Φ^*_α (horizontal line), for the two choices of α considered (the figures show some similarity on the two panels since the same set $\mathscr{X}_N$ is used for both). Convergence of $\Phi(\mathbf{M}_{n_k})$ to Φ^*_α is fast in both cases; the presence of steps on the evolution of $\Phi(\mathbf{M}_{n_k})$, more visible on the right panel, is due to long subsequences of samples consecutively rejected.

Figure 3 shows the behavior of the final directional derivative $F_\Phi[\mathbf{M}_{n_N}, \mathscr{M}(x)]$, after observation of all X_i in $\mathscr{X}_N$, together with the value of its estimated quantile $\widehat{C}_N$ (horizontal solid

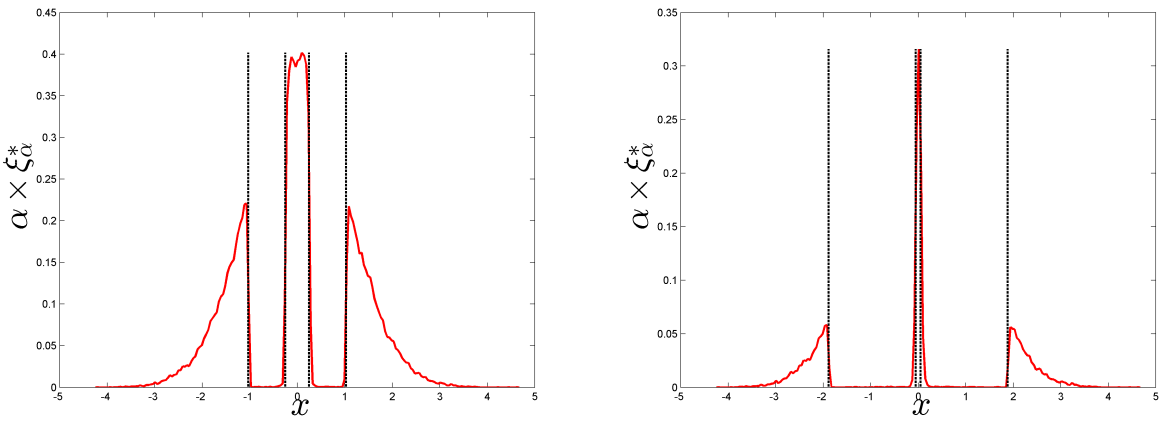

Figure 1: Smoothed histogram of the X_i selected by Algorithm 1; the vertical dotted lines indicate the positions of $-a, -b, b, a$ that define the set $\mathscr{X}^*_\alpha = (-\infty, -a] \cup [-b, b] \cup [a, \infty)$ where $\xi^*_\alpha = \mu/\alpha$; $N = 100\,000$; Left: $\alpha = 1/2$ ($n = 50\,000$); Right: $\alpha = 1/10$ ($n = 10\,000$).

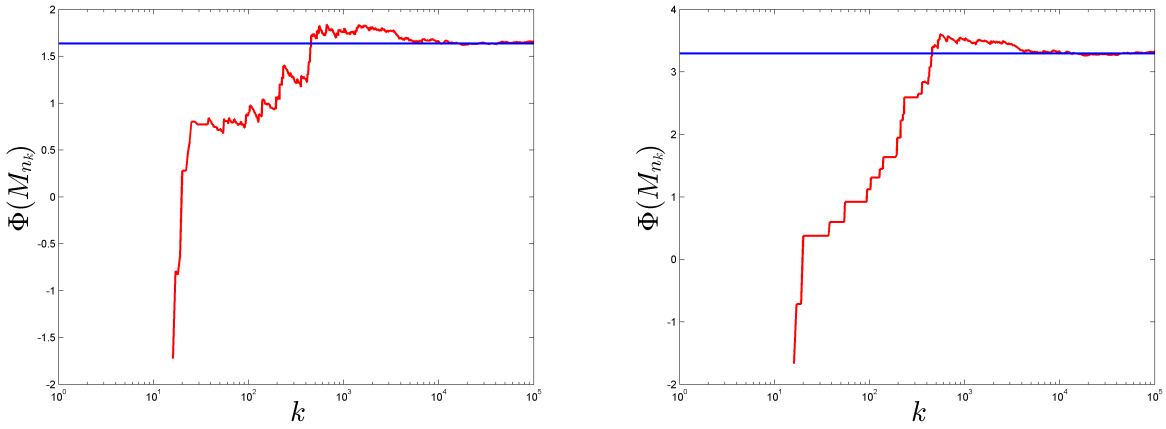

Figure 2: Evolution of $\Phi(\mathbf{M}_{n_k})$ obtained with Algorithm 1 as a function of k (log scale); the horizontal line indicates the optimal value Φ^*_α; $N = 100\,000$; Left: $\alpha = 1/2$ ($n = 50\,000$); Right: $\alpha = 1/10$ ($n = 10\,000$).

line). The theoretical values $C_{1-\alpha}(\mathbf{M}^*_\alpha)$ (horizontal dashed line) and the values $-a, -b, b, a$ where $F_\Phi[\mathbf{M}^*_\alpha, \mathscr{M}(x)] = C_{1-\alpha}(\mathbf{M}^*_\alpha)$ (vertical dashed lines) are also shown ($\widehat{C}_N$ and $C_{1-\alpha}(\mathbf{M}^*_\alpha)$ are indistinguishable on the right panel). Although the figure indicates that $F_\Phi[\mathbf{M}_{n_N}, \mathscr{M}(x)]$ differs significantly from $F_\Phi[\mathbf{M}^*_\alpha, \mathscr{M}(x)]$, they are close enough to allow selection of the most informative X_i, as illustrated by Figures 1 and 2.

Figure 4 shows $\|\mathbf{M}_{n_k} - \mathbf{M}^*_\alpha\|$ (Frobenius norm) as a function of k (log scale), averaged over 1 000 independent repetitions with random samples $\mathscr{X}_N$ of size $N = 10\,000$, for $\alpha = 1/2$. It suggests that $\|\mathbf{M}_{n_k} - \mathbf{M}^*_\alpha\| = \mathcal{O}(1/\sqrt{k})$ for large k, although the conditions in (Konda and Tsitsiklis, 2004) are not satisfied since the scheme we consider is nonlinear. This convergence rate is significantly faster than what is suggested by Dalal et al. (2018). These investigations require further developments and will be pursued elsewhere.

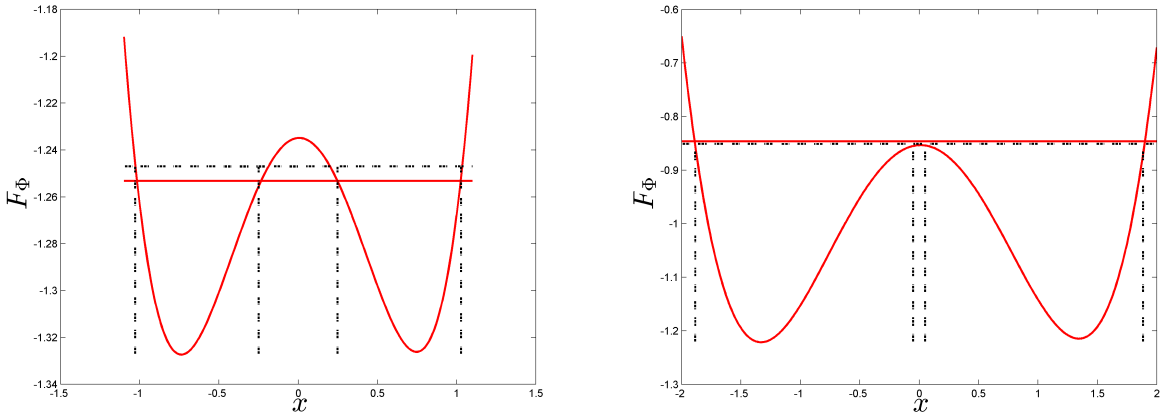

Figure 3: $F_\Phi[\mathbf{M}_{n_N}, \mathscr{M}(x)] = \mathrm{trace}[\mathbf{M}_{n_N}^{-1}\,\mathscr{M}(x)] - 3$ as a function of x (solid line); the horizontal solid (respectively, dashed) line indicates the value of $\widehat{C}_N$ (respectively, $C_{1-\alpha}(\mathbf{M}_\alpha^*)$), the vertical lines indicate the positions of $-a, -b, b, a$ where $F_\Phi[\mathbf{M}_\alpha^*, \mathscr{M}(x)] = C_{1-\alpha}(\mathbf{M}_\alpha^*)$; $N = 100\,000$; Left: $\alpha = 1/2$ ($n = 50\,000$); Right: $\alpha = 1/10$ ($n = 10\,000$).

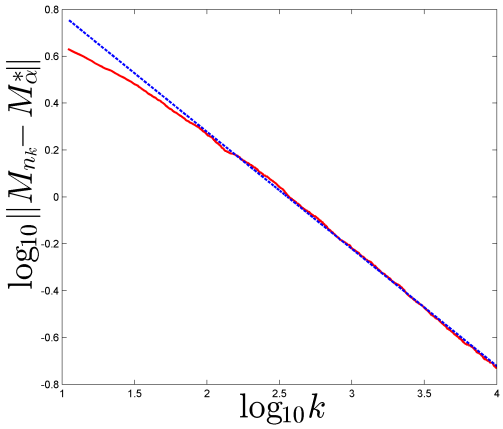

Figure 4: Evolution of $\log_{10}\,\|\mathbf{M}_{n_k} - \mathbf{M}_\alpha^*\|$ as a function of $\log_{10}\,k$ (values averaged over 1 000 random samples $\mathscr{X}_N$); the dashed line has slope $-1/2$ ($\alpha = 1/2$: $n = 5\,000$, $N = 10\,000$).

4.2 Example 2: multilinear regression with normal independent variables

Take $\mathscr{M}(X) = XX^\top$, with $X = (x_1,\ x_2, \ldots, x_d)^\top$, $d > 1$, and $\Phi(\mathbf{M}) = \log\det(\mathbf{M})$, the vectors X_i being i.i.d. $\mathscr{N}(\mathbf{0}, \mathbf{I}_d)$ (so that $p = d$). Denote by $\varphi(\mathbf{x}) = (2\,\pi)^{-d/2}\,\exp(-\|\mathbf{x}\|^2/2)$ the probability density of X. For symmetry reasons, for any $\alpha \in (0,1)$ the optimal (normalized) information matrix is $\mathbf{M}_\alpha^* = \rho_\alpha\,\mathbf{I}_d$, with $\Phi(\mathbf{M}_\alpha^*) = d\,\log\rho_\alpha$, where

$$\begin{aligned}\rho_\alpha = \frac{1}{\alpha}\int_{\|\mathbf{x}\|\geq R_\alpha} x_1^2\,\varphi(\mathbf{x})\,\mathrm{d}\mathbf{x} &= \frac{1}{d\,\alpha}\int_{\|\mathbf{x}\|\geq R_\alpha}\|\mathbf{x}\|^2\,\varphi(\mathbf{x})\,\mathrm{d}\mathbf{x}\\ &= \frac{1}{d\,\alpha}\int_{r\geq R_\alpha} r^2\,\frac{1}{(2\,\pi)^{d/2}}\,\exp(-r^2/2)\,S_d(1)\,r^{d-1}\,\mathrm{d}r\,,\end{aligned}$$

with $S_d(1) = 2\,\pi^{d/2}/\Gamma(d/2)$, the surface area of the d-dimensional unit sphere, and R_α the solution of

$$\alpha = \int_{\|\mathbf{x}\| \geq R_\alpha} \varphi(\mathbf{x})\,\mathrm{d}\mathbf{x} = \int_{r \geq R_\alpha} \frac{1}{(2\,\pi)^{d/2}}\,\exp(-r^2/2)\,S_d(1)\,r^{d-1}\,\mathrm{d}r\,.$$

Since $F_\Phi[\mathbf{M}, \mathscr{M}(X)] = \mathrm{trace}[\mathbf{M}^{-1}\mathscr{M}(X)] - d$, we get $F_\Phi[\mathbf{M}^*_\alpha, \mathscr{M}(X)] = \|\mathbf{x}\|^2/\rho_\alpha - d$, $C_{1-\alpha}(\mathbf{M}^*_\alpha) = R^2_\alpha/\rho_\alpha - d \leq 0$, and $\Phi^*_\alpha = \Phi(\mathbf{M}^*_\alpha)$ is differentiable with respect to α, with $\mathrm{d}\Phi^*_\alpha/\mathrm{d}\alpha = C_{1-\alpha}(\mathbf{M}^*_\alpha)/\alpha$; see Pronzato (2004, Th. 4). Closed-form expressions are available for $d = 2$, with $R_\alpha = \sqrt{-2\,\log\alpha}$ and $\rho_\alpha = 1 - \log\alpha$; R_α and ρ_α can easily be computed numerically for any $d > 2$ and $\alpha \in (0, 1)$. One may notice that, from a result by Harman (2004), the design matrix $\mathbf{M}^*_\alpha$ is optimal for any other orthogonally invariant criterion Φ.

For the linear model with intercept, such that $\mathscr{M}'(X) = \mathbf{f}(X)\mathbf{f}^\top(X)$ with $\mathbf{f}(X) = [1,\ X^\top]^\top$, the optimal matrix is

$$\mathbf{M'}^*_\alpha = \begin{pmatrix} 1 & \mathbf{0}^\top \\ \mathbf{0} & \mathbf{M}^*_\alpha \end{pmatrix}$$

with $\mathbf{M}^*_\alpha = \rho_\alpha\,\mathbf{I}_d$ the optimal matrix for the model without intercept. The same design is thus optimal for both models. Also, when the X_i are i.i.d. $\mathscr{N}(0, \mathbf{\Sigma})$, the optimal matrix $\mathbf{M}^*_{\Sigma,\alpha}$ for $\Phi(\cdot) = \log\det(\cdot)$ simply equals $\mathbf{\Sigma}^{1/2}\,\mathbf{M}^*_\alpha\,\mathbf{\Sigma}^{1/2}$.

Again, we present results obtained for one random set $\mathscr{X}_N$. Figure 5 shows the evolution of $\Phi(\mathbf{M}_{n_k})$ as a function of k for $d = 3$ with $\alpha = 1/1\,000$ and $N = 100\,000$ when we want we select exactly 100 points: the blue dashed-line is when we combine truncation and forced selection; the red solid line is when we adapt α according to $\alpha_k = (n - n_k)/(N - k)$; see Remark 3.1-$(iv)$ — the final values, for $k = N$, are indicated by a triangle and a star, respectively; we only show the evolution of $\Phi(\mathbf{M}_{n_k})$ for k between 10 000 and 100 000 since the curves are confounded for smaller k (they are based on the same $\mathscr{X}_N$). In the first case, the late forced selection of unimportant X_i yields a significant decrease of $\Phi(\mathbf{M}_{n_k})$, whereas adaptation of α anticipates the need of being less selective to reach the target number n of selected points.

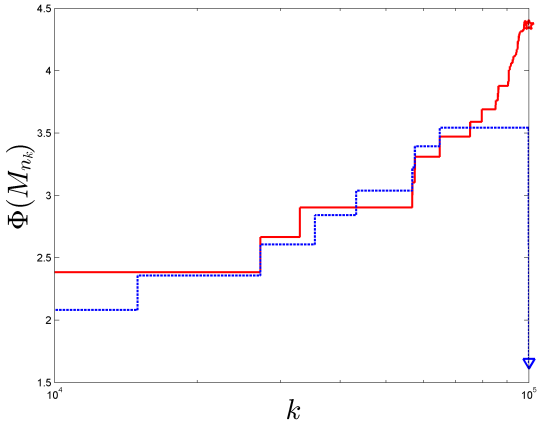

Figure 5: Evolution of $\Phi(\mathbf{M}_{n_k})$ obtained with Algorithm 1 as a function of k (log scale): $d = 3$, $N = 100\,000$, exactly $n = 100$ points are collected using truncation/forced selection (blue dashed line and $\triangledown$) or adaptation of α (red solid line and $\bigstar$); see Remark 3.1-(iv).

Figure 2 has illustrated the convergence of $\Phi(\mathbf{M}_{n_k})$ to Φ^*_α for a fixed α as $k \to \infty$, but in fact what really matters is that n_k tends to infinity: indeed, $\Phi(\mathbf{M}_{n_k})$ does not converge to Φ^*_α if we fix $n_k = n$ and let k tend to infinity, so that $\alpha = n/k$ tends to zero (see also Section 5.1). This is illustrated on the left panel of Figure 6, where $d = 25$ and, from left to right, α equals 0.5 (magenta dotted line), 0.1, 0.05 and 0.01 (red solid line). Since the optimal value Φ^*_α depends on α, here we present the evolution with k of the D-efficiency $[\det(\mathbf{M}_{n_k})/\det(\mathbf{M}^*_\alpha)]^{1/p} = \exp[(\Phi(\mathbf{M}_{n_k}) - \Phi^*_\alpha)/d]$. The right panel is for fixed $\alpha = 0.1$ and varying d, with, from left to right, $d = 5$ (red solid line), 10, 20, 30 and 50 (cyan solid line). As one may expect, performance (slightly) deteriorates as d increases due to the increasing variability of $Z_{\mathbf{M}}(X)$, with $\mathsf{var}[Z_{\mathbf{M}}(X)] = \mathsf{var}[X^\top \mathbf{M} X] = 2\,\mathrm{trace}(\mathbf{M}^2)$.

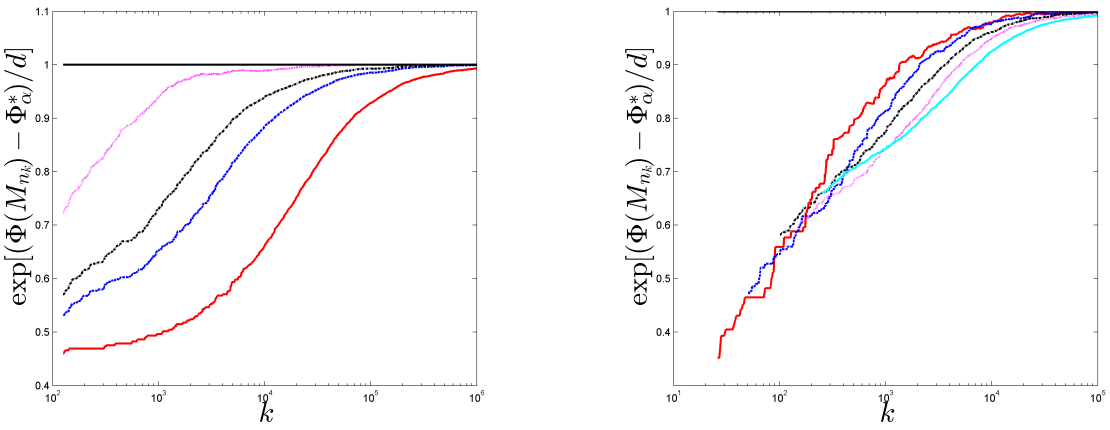

Figure 6: Evolution of D-efficiency of $\mathbf{M}_{n_k}$ obtained with Algorithm 1 as a function of k (log scale); the horizontal line indicates the optimal value 1. Left: $d = 25$ and $\alpha = 0.5$ (magenta dotted line), 0.1 (black), 0.05 (blue) and 0.01 (red solid line). Right: $\alpha = 0.1$ and $d = 5$ (red solid line), 10 (blue), 20 (black), 30 (magenta) and 50 (cyan solid line).

4.3 Example 3: processing a non i.i.d. sequence

When the design points X_i are not from an i.i.d. sequence, Algorithm 1 cannot be used directly and some preprocessing is required. When storage of the whole sequence $\mathscr{X}_N$ is possible, a random permutation can be applied to $\mathscr{X}_N$ before using Algorithm 1. When N is too large for that, for instance in the context of data streaming, and the sequence possesses a structured time-dependence, one may try to identify the dependence model through time series analysis and use forecasting to decide which design points should be selected. The data thinning mechanism is then totally dependent on the model of the sequence, and the investigation of the techniques to be used is beyond the scope of this paper. Examples of the application of a simple scrambling method to the sequence prior to selection by Algorithm 1 are presented below. The method corresponds to Algorithm 2 below; its output sequence $\widetilde{X}_k$ is used as input for Algorithm 1. We do not study the properties of the method in conjunction with the convergence properties of Algorithm 1, which

would require further developments.

Algorithm 2: random scrambling in a buffer.

1) Initialization: choose the buffer size B, set $k = 1$ and $\mathscr{X}^{(1)} = \{X_1, \ldots, X_B\}$.

2) Draw $\widetilde{X}_k$ by uniform sampling within $\mathscr{X}^{(k)}$.

3) Set $\mathscr{X}^{(k+1)} = \mathscr{X}^{(k)} \setminus \{\widetilde{X}_k\} \cup \{X_{B+k}\}$, $k \leftarrow k+1$, return to Step 2.

Direct calculation shows that the probability that $\widetilde{X}_k$ equals X_i is

$$\mathsf{Prob}\{\widetilde{X}_k = X_i\} = \begin{cases} \frac{1}{B}\left(1-\frac{1}{B}\right)^{k-1} & \text{for } 1 \le i \le B \\ \frac{1}{B}\left(1-\frac{1}{B}\right)^{k-1+B-i} & \text{for } B+1 \le i \le B+k-1 \\ 0 & \text{for } B+k-1 < i \end{cases}$$

showing the limits of randomization via Algorithm 2 (in particular, the first points of the sequence X_i will tend to appear first among the $\widetilde{X}_k$). However, the method will give satisfactory results if the size B of the buffer is large enough, as its performance improves as B increases.

As an illustration, we consider the same quadratic regression model as in Example 1, with $\Phi(\mathbf{M}) = \log\det(\mathbf{M})$, in the extreme case where $X_i = x(t_i)$ with $x(t)$ a simple function of t.

First, we consider the extremely unfavorable situation where $x(t) = t$. When t is uniformly distributed on $\mathscr{T} = [0, T]$, the optimal design ξ^*_α selects all points associated with t in $[0, t_a] \cup [t_b, T - t_b] \cup [T - t_a, T]$, for some $t_a < t_b$ in $\mathscr{T}$ satisfying $2\,t_a + T - 2\,t_b = \alpha\,T$. For $\alpha = 1/10$, we get $t_a/T \simeq 0.03227$ and $t_b/T \simeq 0.48227$. The horizontal black line in Figure 7 indicates the optimal value Φ^*_α when $T = 1$. The blue dotted line shows $\Phi(\mathbf{M}_{n_k})$ when Algorithm 1 is directly applied to the points $X_i = i/N$, $i = 1, \ldots, N = 100\,000$. The red line is when randomization via Algorithm 2, with buffer size $B = \alpha N = 10\,000$, is applied first; the dotted curve in magenta is for $B = 3\,\alpha N$. The positive effects of randomization through Algorithm 2 and the influence of the buffer size are visible on the figure. Here, the monotonicity of the X_i, which inhibits early exploration of their range of variation, prevents convergence to the optimum.

We now consider the more favorable case where $X_i = \sin(2\pi\nu i/N)$, $i = 1, \ldots, N = 100\,000$, with $\nu = 5$. The left panel of Figure 8 shows the same information as Figure 7, when Algorithm 1 is applied directly to the X_i (blue dotted line) and after preprocessing with Algorithm 2 with $B = \alpha N$ (red line) and $B = \alpha N/10$ (magenta dotted line). The early exploration of the range of variability of the X_i, possible here thanks to the periodicity of $x(t)$, makes the randomization through Algorithm 2 efficient enough to allow Algorithm 1 to behave correctly when $B = \alpha N$ (red line). The situation improves when B is increased, but naturally deteriorates if B is too small (magenta dotted line). The right panel shows the points $\widetilde{X}_k$ produced by Algorithm 2 (with $B = \alpha N = 10\,000$) which are selected by Algorithm 1. The effect of randomization is visible. For $k < 5\,000$ all points in the buffer are in the interval $[\sin(2\pi\nu(k+B)/N), 1] \subset [-1, 1]$ and the points selected by Algorithm 1 are near the end points or the center of this moving interval. For larger k, randomization is strong enough to maintain the presence of suitable candidates in the buffer for selection by Algorithm 1.

4.4 Examples with $Z_{\mathbf{M}}(x)$ constant on subsets of positive measure

Here we consider situations where $\mathrm{H}_{\mu,\mathscr{M}}$ is violated due to the existence of subsets of $\mathscr{X}$ of positive measure on which $Z_{\mathbf{M}}(x)$ is constant. The model is the same as in Section 4.2, with $X = (x_1,\ x_2, \ldots, x_d)^\top$, $\mathscr{M}(X) = XX^\top$ and $\Phi(\mathbf{M}) = \log\det(\mathbf{M})$.

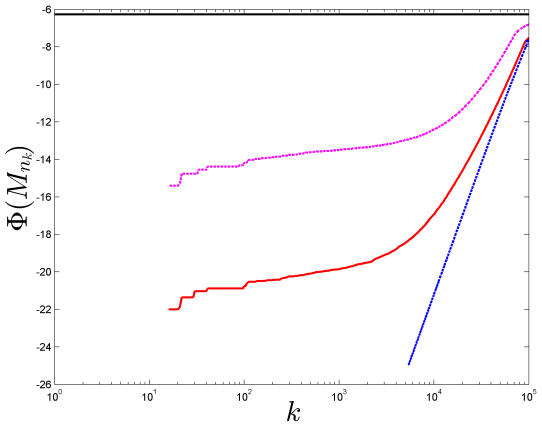

Figure 7: Evolution of $\Phi(\mathbf{M}_{n_k})$ obtained with Algorithm 1 as a function of k (log scale), the horizontal line indicates the optimal value Φ^*_α; $N = 100\,000$, $\alpha = 1/10$; direct processing of the points $X_i = i/N$ (dotted blue line), preprocessing with Algorithm 2 with $B = \alpha N$ (red solid line) and $B = 3\,\alpha N$ (magenta dotted line).

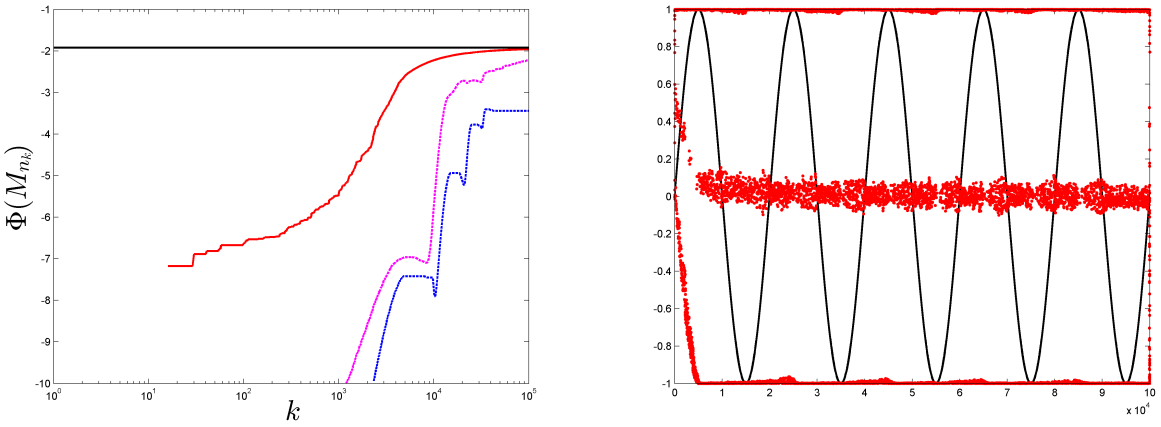

Figure 8: Left: Evolution of $\Phi(\mathbf{M}_{n_k})$ obtained with Algorithm 1 as a function of k (log scale), the horizontal line indicates the optimal value Φ^*_α; $N = 100\,000$, $\alpha = 1/10$; direct processing of the points $X_i = \sin(2\pi\nu i/N)$ (dotted blue line), preprocessing with Algorithm 2 with $B = \alpha N$ (red solid line) and $B = \alpha N/10$ (magenta dotted line). Right: X_i, $i = 1, \ldots, N = 100\,000$ (black) and points $\widetilde{X}_k$ produced by Algorithm 2 with $B = \alpha N$ which are selected by Algorithm 1 (red).

4.4.1 Example 4: μ has discrete components

This is Example 11 in (Pronzato, 2006), where $d = 2$, $\mu = (1/2)\,\mu_{\mathscr{N}} + (1/2)\,\mu_d$, with $\mu_{\mathscr{N}}$ corresponding to the normal distribution $\mathscr{N}(0, 1)$ and μ_d the discrete measure that puts weight $1/4$ at each one of the points $(\pm 1, \pm 1)$. Denote by $\mathscr{B}(r)$ the closed ball centered at the origin $\mathbf{0}$ with radius r, by $\mu[r]$ the measure equal to μ on its complement $\overline{\mathscr{B}(r)}$, and let $\mathsf{e} = \exp(1)$. The optimal matrix

is $\mathbf{M}_\alpha^* = \mathbf{M}(\xi_\alpha^*)$, with ξ_α^* the probability measure defined by:

$$\xi_\alpha^* = \begin{cases} \frac{1}{\alpha}\,\mu[\sqrt{-2\,\log(2\,\alpha)}] & \text{if } 0 < \alpha \leq \frac{1}{2\mathsf{e}}\,, \\ \frac{1}{\alpha}\,\mu[\sqrt{2}] + \frac{1}{\alpha}\,[\alpha - 1/(2\,\mathsf{e})]\mu_d & \text{if } \frac{1}{2\mathsf{e}} < \alpha \leq \frac{1}{2\mathsf{e}} + \frac{1}{2}\,, \\ \frac{1}{\alpha}\,\mu[\sqrt{-2\,\log(2\,\alpha - 1)}] & \text{if } \frac{1}{2\mathsf{e}} + \frac{1}{2} < \alpha < 1\,, \end{cases}$$

with associated Φ-values

$$\Phi(\mathbf{M}_\alpha^*) = \begin{cases} 2\,\log[1 - \log(2\,\alpha)] & \text{if } 0 < \alpha \leq \frac{1}{2\mathsf{e}}\,, \\ 2\,\log\left(1 + \frac{1}{2\mathsf{e}\,\alpha}\right) & \text{if } \frac{1}{2\mathsf{e}} < \alpha \leq \frac{1}{2\mathsf{e}} + \frac{1}{2}\,, \\ 2\,\log\left(1 - \frac{(2\,\alpha - 1)\,\log(2\,\alpha - 1)}{2\,\alpha}\right) & \text{if } \frac{1}{2\mathsf{e}} + \frac{1}{2} < \alpha \leq 1\,. \end{cases}$$

Figure 9 shows the evolution of $\Phi(\mathbf{M}_{n_k})$ as a function of k for $\alpha = 0.5$ (left) and $\alpha = 0.02$ (right). Note that $\alpha < 1/(2\mathsf{e})$ in the second case, but $1/(2\mathsf{e}) < \alpha \leq 1/(2\mathsf{e}) + 1/2$ in the first one, so that ξ_α^* is neither zero nor μ/α on the four points $(\pm 1, \pm 1)$. Figure 9 shows that Algorithm 1 nevertheless behaves satisfactorily in both cases.

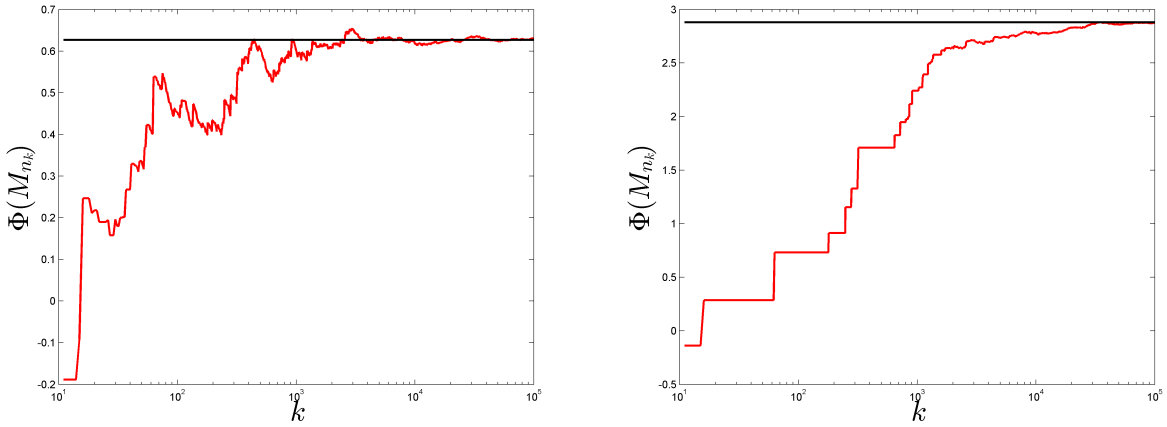

Figure 9: Evolution of $\Phi(\mathbf{M}_{n_k})$ obtained with Algorithm 1 as a function of k (log scale) when $d = 2$ and $\mu = (1/2)\,\mu_{\mathscr{N}} + (1/2)\,\mu_d$ with $\alpha = 0.5$ (left) and $\alpha = 0.02$ (right); the horizontal line indicates the optimal value Φ_α^*.

4.4.2 Example 5: the distribution of $Z_{\mathbf{M}_\alpha^*}(X)$ has discrete components

Let $U[\mathscr{S}_d(\mathbf{0}, r)]$ denote the uniform probability measure on the d-dimensional sphere $\mathscr{S}_d(\mathbf{0}, r)$ with center $\mathbf{0}$ and radius r. The probability measure of the X_i is $\mu = (1/3)\sum_{i=1}^3 U[\mathscr{S}_d(\mathbf{0}, r_i)]$, the mixture of distributions on three nested spheres with radii $r_1 > r_2 > r_3 > 0$. The optimal bounded measure is

$$\xi_\alpha^* = \begin{cases} U[\mathscr{S}_d(\mathbf{0}, r_1)] & \text{if } 0 < \alpha \leq \frac{1}{3}\,, \\ \frac{1}{3\alpha}\,U[\mathscr{S}_d(\mathbf{0}, r_1)] + \frac{\alpha - 1/3}{\alpha} U[\mathscr{S}_d(\mathbf{0}, r_2)] & \text{if } \frac{1}{3} < \alpha \leq \frac{2}{3}\,, \\ \frac{1}{3\alpha}\,\{U[\mathscr{S}_d(\mathbf{0}, r_1)] + U[\mathscr{S}_d(\mathbf{0}, r_2)]\} + \frac{\alpha - 2/3}{\alpha} U[\mathscr{S}_d(\mathbf{0}, r_3)] & \text{if } \frac{2}{3} < \alpha < 1\,, \end{cases}$$

with associated Φ-values

$$\Phi(\mathbf{M}_\alpha^*) = \begin{cases} d\,\log(r_1^2/d) & \text{if } 0 < \alpha \leq \frac{1}{3}\,, \\ d\,\log\left(\frac{r_1^2/3 + (\alpha - 1/3) r_2^2}{\alpha d}\right) & \text{if } \frac{1}{3} < \alpha \leq \frac{2}{3}\,, \\ d\,\log\left(\frac{(r_1^2 + r_2^2)/3 + (\alpha - 2/3) r_3^2}{\alpha d}\right) & \text{if } \frac{2}{3} < \alpha < 1\,. \end{cases}$$

Notice that for $\alpha \in (0, 1/3)$ (respectively, $\alpha \in (1/3, 2/3)$) $\xi^*_\alpha \neq 0$ and $\xi^*_\alpha \neq \mu/\alpha$ on $\mathscr{S}_d(\mathbf{0}, r_1)$ (respectively, on $\mathscr{S}_d(\mathbf{0}, r_2)$) although μ is atomless.

The left panel of Figure 10 gives the evolution with k of the D-efficiency $[\det(\mathbf{M}_{n_k})/\det(\mathbf{M}^*_\alpha)]^{1/p} = \exp[(\Phi(\mathbf{M}_{n_k}) - \Phi^*_\alpha)/d]$, for $\alpha = 0.5$ (red solid line) and $\alpha = 0.2$ (blue dashed line) when $d = 5$. The right panel shows the evolution of the ratio n_k/k for those two situations, with the limiting value α indicated by a horizontal line. Although assumption $\mathrm{H}_{\mu,\mathscr{M}}$ is violated, Algorithm 1 continues to perform satisfactorily.

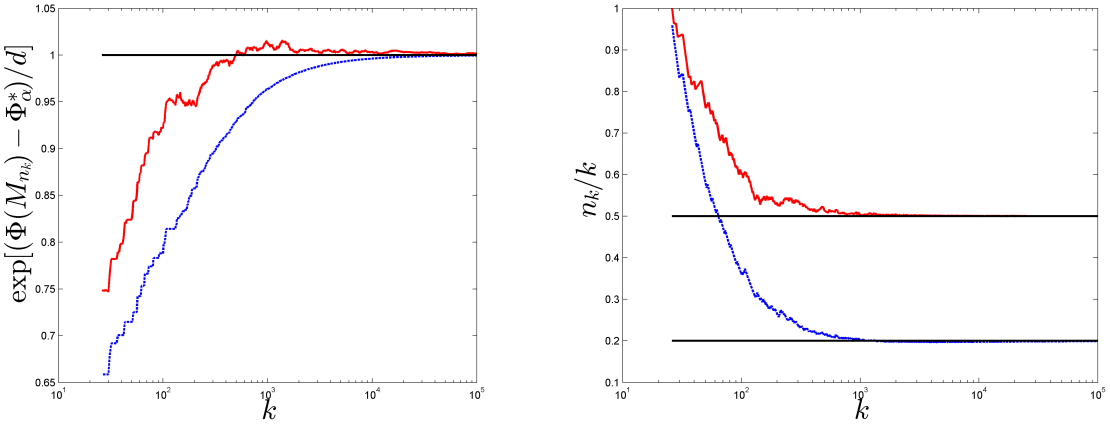

Figure 10: Left: Evolution of D-efficiency of $\mathbf{M}_{n_k}$ obtained with Algorithm 1 as a function of k (log scale) for $\alpha = 0.5$ (red solid line) and $\alpha = 0.2$ (blue dashed line); the horizontal line indicates the optimal value 1; $d = 5$. Right: evolution of the ratio n_k/k in the same simulations.

5 Comparison with other methods

5.1 Case n fixed with large N: comparison with an exchange method

The convergence of $\Phi(\mathbf{M}_{n_k})$ to Φ^*_α in Algorithm 1 relies on the fact that n_k grows like $\mathcal{O}(\alpha\, k)$ for some $\alpha > 0$; see Theorem 3.1. If the number n of points to be selected is fixed, Algorithm 1 does not provide any performance guarantee when applied to a sequence of length $N \to \infty$ (the situation is different when $p = 1$ where an asymptotically optimal construction is available; see Pronzato (2001)). In that case, a method of the exchange type may look more promising, although large values of N entail serious difficulties. Typically, the algorithm is initialized by a n point design chosen within $\mathscr{X}_N$, and at each iteration a temporarily selected X_i is replaced by a better point in $\mathscr{X}_N$. Fedorov's (1972) algorithm considers all $n \times (N - n)$ possible replacements at each iteration ($(N - n)$ instead of N since we do not allow repetitions in the present context); its computational cost is prohibitive for large N. The variants suggested by Cook and Nachtsheim (1980), or the DETMAX algorithm of Mitchell (1974), still require the maximization of a function $g(X_j)$ with respect to $X_j \in \mathscr{X}_N$ at each iteration, which remains unfeasible for very large N. Below, we consider a simplified version where all N points are examined successively, and replacement is accepted when it improves the

current criterion value.

Algorithm 3: sequential exchange (n fixed).

1) Initialization: select $X_1, \ldots, X_n$, set $k = n$ and $\mathscr{X}_k^* = \{X_1, \ldots, X_k\}$, compute $\mathbf{M}_{n,k} = (1/k) \sum_{i=1}^k \mathscr{M}(X_i)$ and $\Phi(\mathbf{M}_{n,k})$.

2) Iteration $k+1$: collect X_{k+1}. If $X_{k+1} \in \mathscr{X}_k^*$, set $\Delta^{(k)}(X_{i^*}, X_{k+1}) = 0$; otherwise compute

$$\Delta^{(k)}(X_{i^*}, X_{k+1}) = \max_{X_i \in \mathscr{X}_k^*} [\Phi\{\mathbf{M}_{n,k} + (1/n)[\mathscr{M}(X_{k+1}) - \mathscr{M}(X_i)]\} - \Phi(\mathbf{M}_{n,k})] \; .$$

If $\Delta^{(k)}(X_{i^*}, X_{k+1}) > 0$, set $\mathscr{X}_{k+1}^* = \mathscr{X}_k^* \setminus \{X_{i^*}\} \cup X_{k+1}$, update $\mathbf{M}_{n,k}$ into $\mathbf{M}_{n,k+1} = \mathbf{M}_{n,k} + (1/n)[\mathscr{M}(X_{k+1}) - \mathscr{M}(X_{i^*})]$, compute $\Phi(\mathbf{M}_{n,k+1})$;

otherwise, set $\mathscr{X}_{k+1}^* = \mathscr{X}_k^*$, $\mathbf{M}_{n,k+1} = \mathbf{M}_{n,k}$.

3) If $k+1 = N$ stop; otherwise, $k \leftarrow k+1$, return to Step 2.

Remark 5.1. When $\mathscr{M}(x)$ has rank one, with $\mathscr{M}(x) = \mathbf{f}(x)\mathbf{f}^\top(x)$ and $\Phi(\mathbf{M}) = \log\det(\mathbf{M})$ or $\Phi(\mathbf{M}) = \det^{1/p}(\mathbf{M})$ (D-optimal design), $\Delta^{(k)}(X_{i^*}, X_{k+1}) > 0$ is equivalent to

$$\mathbf{f}^\top(X_{k+1})\mathbf{M}_{n,k}^{-1}\mathbf{f}^\top(X_{k+1}) - \mathbf{f}^\top(X_{i^*})\mathbf{M}_{n,k}^{-1}\mathbf{f}^\top(X_{i^*}) + \delta^{(k)}(X_{i^*}, X_{k+1}) > 0 \, , \tag{5.1}$$

where

$$\delta^{(k)}(X, X_{k+1}) = \frac{[\mathbf{f}^\top(X_{k+1})\mathbf{M}_{n,k}^{-1}\mathbf{f}^\top(X)]^2 - [\mathbf{f}^\top(X_{k+1})\mathbf{M}_{n,k}^{-1}\mathbf{f}^\top(X_{k+1})][\mathbf{f}^\top(X)\mathbf{M}_{n,k}^{-1}\mathbf{f}^\top(X)]}{n} \, , \tag{5.2}$$

see Fedorov (1972, p. 164). As for Algorithm 1 (see Remark 3.1-(*iii*)), we may update $\mathbf{M}_{n,k}^{-1}$ instead of $\mathbf{M}_{n,k}$ to avoid matrix inversions. For large enough n, the term (5.2) is negligible and the condition is almost $\mathbf{f}^\top(X_{k+1})\mathbf{M}_{n,k}^{-1}\mathbf{f}^\top(X_{k+1}) > \mathbf{f}^\top(X_{i^*})\mathbf{M}_{n,k}^{-1}\mathbf{f}^\top(X_{i^*})$; that is, $F_\Phi[\mathbf{M}_{n,k}, \mathscr{M}(X_{k+1})] > F_\Phi[\mathbf{M}_{n,k}, \mathscr{M}(X_{i^*})]$. This is the condition we use in the example below. It does not guarantee in general that $\Phi(\mathbf{M}_{n,k+1}) > \Phi(\mathbf{M}_{n,k})$ (since $\delta^{(k)}(X_i, X_{k+1}) \leq 0$ from Cauchy-Schwartz inequality), but no significant difference was observed compared with the use of the exact condition (5.1). Algorithm 3 has complexity $\mathcal{O}(nd^3\,N)$ in general (the additional factor n compared with Algorithm 1 is due to the calculation of the maximum over all X_i in $\mathscr{X}_k^*$ at Step 2). ◁

Neither Algorithm 1 with $\alpha = n/N$ and $\mathbf{M}_{n,k} = \mathbf{M}_{n_k}$ nor Algorithm 3 ensures that $\Phi(\mathbf{M}_{n,N})$ tends to $\Phi_\alpha^* = \Phi(\mathbf{M}_\alpha^*)$ as $N \to \infty$. Also, we can expect to have $\Phi(\mathbf{M}_{n,k}) \lesssim \Phi_{n/k}^*$ for all k with Algorithm 3, since under H_Φ the matrix $\mathbf{M}_{n,k}^*$ corresponding to the optimal selection of n distinct points among $\mathscr{X}_k$ satisfies $\mathsf{E}\{\Phi(\mathbf{M}_{n,k}^*)\} \leq \Phi_{n/k}^*$ for all $k \geq n > 0$; see Pronzato (2006, Lemma 3).

Example 6: n fixed and N large We consider the same situation as in Example 2 (Section 4.2), with $X = (x_1,\ x_2, \ldots, x_d)^\top$, $\mathscr{M}(X) = XX^\top$, $\Phi(\mathbf{M}) = \log\det(\mathbf{M})$; the X_i are i.i.d. $\mathscr{N}(\mathbf{0}, \mathbf{I}_d)$, with $p = d = 3$. We still take $k_0 = 5\,p$, $q = 5/8$, $\gamma = 1/10$ in Algorithm 1. We have $\mathsf{E}\{\mathbf{M}_{n_{k_0}}\} = \mathbf{M}(\mu)$ in Algorithm 1, and, when n is large enough, $\mathbf{M}_{n,n} \simeq \mathbf{M}(\mu)$ at Step 1 of Algorithm 3, with $\mathbf{M}(\mu) = \mathbf{I}_d$ and therefore $\Phi(\mathbf{M}_{n,n}) \simeq 0$.

We consider two values of n, $n = 100$ and $n = 1\,000$, with $\alpha = 10^{-3}$ (that is, with $N = 100\,000$ and $N = 1\,000\,000$, respectively). Figure 11 shows the evolutions of $\Phi(\mathbf{M}_{n_k})$ ($k \geq k_0$, Algorithm 1, red solid line) and $\Phi(\mathbf{M}_{n,k})$ ($k \geq n$, Algorithm 3, blue dashed line) as functions of k in those two cases ($n = 100$, left; $n = 1\,000$, right). In order to select n points exactly, adaptation of α is

used in Algorithm 1, see Remark 3.1-(iv). The value of n is too small for $\Phi(\mathbf{M}_{n_k})$ to approach Φ^*_α (indicated by the horizontal black line) in the first case, whereas $n = 1\,000$ is large enough on the right panel; Algorithm 3 performs similarly in both cases and is superior to Algorithm 1 for $n = 100$; the magenta curve with triangles shows $\Phi^*_{n/k}$, $k \geq n$, with $\Phi^*_{n/k} \gtrsim \Phi(\mathbf{M}_{n,k})$ for all k, as expected. ◁

In case it is possible to store the N points X_i, we can replay both algorithms on the same data set in order to increase the final value of Φ for the sample selected. For Algorithm 3, we can simply run the algorithm again on a set $\mathscr{X}_N^{(2)}$ — starting with $k = 1$ at Step 1 since n points have already been selected — with $\mathscr{X}_N^{(2)} = \mathscr{X}_N$ or corresponding to a random permutation of it. Series of runs on sets $\mathscr{X}_N^{(2)}, \mathscr{X}_N^{(3)}, \ldots$ can be concatenated: the fact that Φ can only increase implies convergence for an infinite sequence of runs, but generally to a local maximum only; see the discussion in (Cook and Nachtsheim, 1980, Sect. 2.4). When applied to Example 6, this method was not able to improve the design obtained in the first run of Algorithm 3, with a similar behavior with or without permutations in the construction of the $\mathscr{X}_N^{(i)}$.

Algorithm 1 requires a more subtle modification since points are selected without replacement. First, we run Algorithm 1 with α fixed at n/N on a set $\mathscr{X}_{mN} = \mathscr{X}_N \cup \mathscr{X}_N^{(2)} \cup \cdots \cup \mathscr{X}_N^{(m)}$, where the replications $\mathscr{X}_N^{(i)}$ are all identical to $\mathscr{X}_N$ or correspond to random permutations of it. The values of $\mathbf{M}_{n_{mN}}$ and $\widehat{C}_{mN}$ are then used in a second stage, where the N points $X_1, \ldots, X_N$ in $\mathscr{X}_N$ are inspected sequentially: starting at $k = 0$ and $n_k = 0$, a new point X_{k+1} is selected if $n_k < n$ and $F_\Phi[\mathbf{M}_{n_{mN}}, \mathscr{M}(X_{k+1})] > \widehat{C}_{mN}$ (or if $n - n_k \geq N - k + 1$, see Remark 3.1-(iv)). The set $\mathscr{X}_N$ is thus used $m + 1$ times in total. The idea is that for m large enough, we can expect $\mathbf{M}_{n_{mN}}$ to be close to $\mathbf{M}^*_\alpha$ and $\widehat{C}_{mN}$ to be close to the true quantile $C_{1-\alpha}(\mathbf{M}^*_\alpha)$, whereas the optimal rule for selection is $F_\Phi[\mathbf{M}^*_\alpha, \mathscr{M}(X_{k+1})] > C_{1-\alpha}(\mathbf{M}^*_\alpha)$. Note that the quantile of the directional derivative is not estimated in this second phase, and updating of $\mathbf{M}_{n_k}$ is only used to follow the evolution of $\Phi(\mathbf{M}_{n_k})$ on plots.

Example 6 (continued) The black-dotted line in Figure 11 shows the evolution of $\Phi(\mathbf{M}_{n_k})$ as a function of k in the second phase (for k large enough to have $\Phi(\mathbf{M}_{n_k}) > 0$): we have taken $m = 9$ for $n = 100$ (left), so that $(m + 1)N = 1\,000\,000$ points are used in total (but 10 times the same), and $m = 1$ for $n = 1\,000$ (right), with $2\,000\,000$ points used (twice the same). Figure 12 shows the evolution of $\widehat{C}_k$ for $k = 1, \ldots, mN$, for $n = 100$, $N = 100\,000$, $m = 9$ (left), and $n = 1\,000$, $N = 1\,000\,000$, $m = 1$ (right); the horizontal black line indicates the value of $C_{1-\alpha}(\mathbf{M}^*_\alpha)$. The left panel indicates that $n = 100$ is too small to estimate $C_{1-\alpha}(\mathbf{M}^*_\alpha)$ correctly with Algorithm 1 (note that $m = 4$ would have been enough), which is consistent with the behavior of $\Phi(\mathbf{M}_{n_k})$ observed in Figure 11-left (red solid line). The right panel of Figure 12 shows that $\widehat{C}_k$ has converged before inspection of the $1\,000\,000$ points in $\mathscr{X}_N$, which explains the satisfactory behavior of Algorithm 1 in Figure 11-right. Notice the similarity between the left and right panels of Figure 12 due to the fact that the same value $\alpha = 10^{-3}$ is used in both. Here the $\mathscr{X}_N^{(i)}$ are constructed by random permutations of the points in $\mathscr{X}_N$, but the behavior is similar without.

5.2 Comparison with IBOSS

IBOSS (Information-Based Optimal Subdata Selection, Wang et al. (2019)) is a selection procedure motivated by D-optimality developed in the context of multilinear regression with intercept, where $\mathscr{M}(X) = \mathbf{f}(X)\mathbf{f}^\top(X)$ with $\mathbf{f}(X) = [1,\ X^\top]^\top$. All points X_i in $\mathscr{X}_N$ are processed simultaneously: the d coordinates of the X_i are examined successively; for each $k = 1, \ldots, d$, the r points with largest

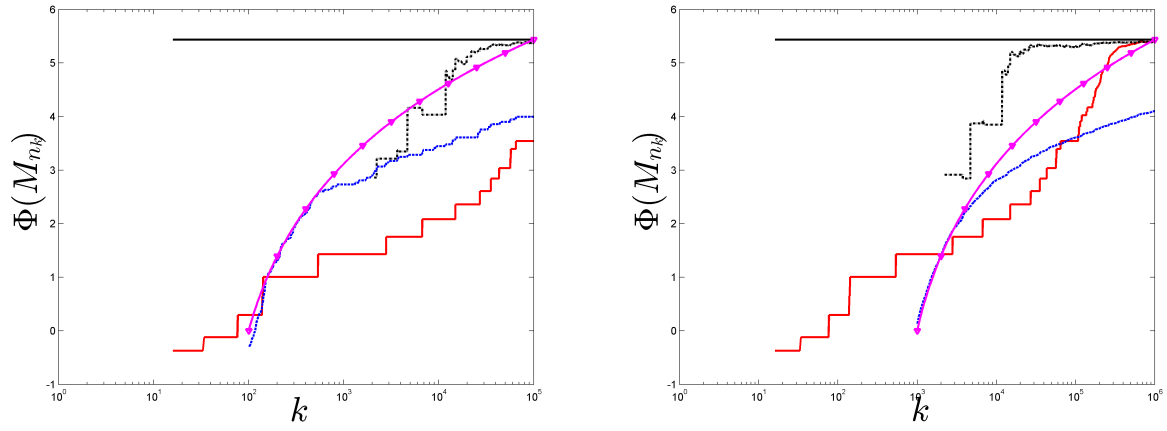

Figure 11: Evolutions of $\Phi(\mathbf{M}_{n_k})$ ($k \geq k_0$, Algorithm 1, red solid line) $\Phi(\mathbf{M}_{n,k})$ ($k \geq n$, Algorithm 3, blue dashed line), and $\Phi^*_{n/k}$ (magenta curve with triangles) as functions of k; the horizontal black line corresponds to Φ^*_α; the black dotted curve shows the evolution of $\Phi(\mathbf{M}_{n_k})$ as a function of k when the selection is based on $F_\Phi[\mathbf{M}_{n_{mN}}, \mathscr{M}(X_{k+1})] > \widehat{C}_{mN}$, with $\mathbf{M}_{n_{mN}}$ and $\widehat{C}_{mN}$ obtained with Algorithm 1 applied to $\mathscr{X}_{mN} = \mathscr{X}_N \cup \mathscr{X}_N^{(2)} \cup \cdots \cup \mathscr{X}_N^{(m)}$. Left: $n = 100$, $N = 100\,000$, $m = 9$; Right: $n = 1\,000$, $N = 1\,000\,000$, $m = 1$.

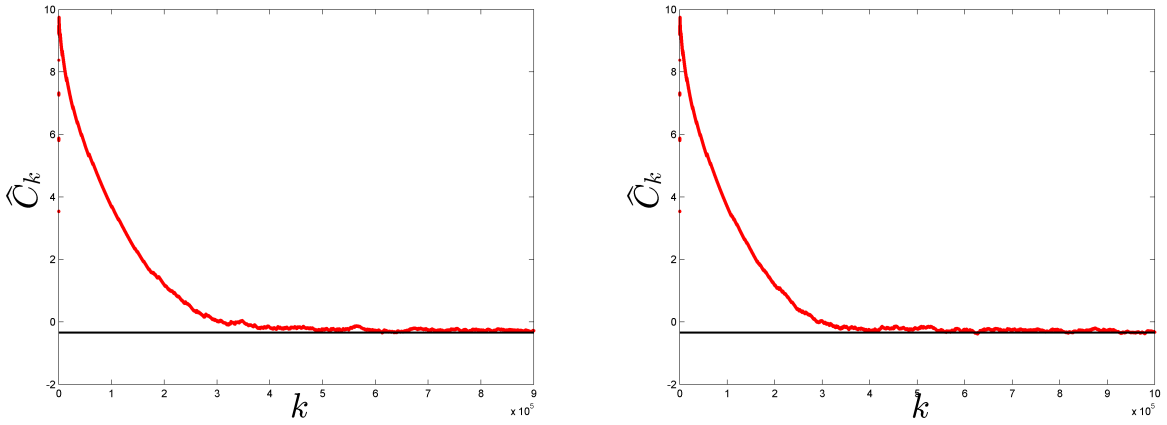

Figure 12: Evolution of $\widehat{C}_k$ in Algorithm 1 when applied to $\mathscr{X}_{mN} = \mathscr{X}_N \cup \mathscr{X}_N^{(2)} \cup \cdots \cup \mathscr{X}_N^{(m)}$, the horizontal black line corresponds to $C_{1-\alpha}(\mathbf{M}^*_\alpha)$. Left: $n = 100$, $N = 100\,000$, $m = 9$; Right: $n = 1\,000$, $N = 1\,000\,000$, $m = 1$.

k-th coordinate and the r points having smallest k-th coordinate are selected (and removed from $\mathscr{X}_N$), where $r = n/(2d)$, possibly with suitable rounding, when exactly n points have to be selected. The design selected is sensitive to the order in which coordinates are inspected. The necessity to find the largest or smallest coordinate values yields a complexity of $\mathcal{O}(d\,N)$; parallelization with simultaneous sorting of each coordinate is possible. Like for any design selection algorithm, the matrix $\mathbf{M}_{n,N}$ obtained with IBOSS satisfies $\mathsf{E}\{\Phi(\mathbf{M}_{n,N})\} \leq \Phi^*_{n/N}$ for all $N \geq n > 0$ (Pronzato, 2006, Lemma 3). The asymptotic performance of IBOSS (the behavior of $\mathbf{M}_{n,N}$ and $\mathbf{M}^{-1}_{n,N}$) for n fixed and N tending to infinity is investigated in (Wang et al., 2019) for X following a multivariate

normal or lognormal distribution. Next property concerns the situation where n is a fraction of N, with $N \to \infty$ and the components of X are independent.

Theorem 5.1. *Suppose that the X_i are i.i.d. with μ satisfying H_μ and, moreover, that their components $\{X_i\}_k$ are independent, with φ_k the p.d.f. of $\{X_1\}_k$ for $k = 1, \ldots, d$. Suppose, without any loss of generality, that coordinates are inspected in the order $1, \ldots, d$. Then, for any $\alpha \in (0, 1]$, the matrix $\mathbf{V}_{n,N} = (1/n) \sum_{j=1}^{n} X_{i_j} X_{i_j}^\top$ corresponding to the n points X_{i_j} selected by IBOSS satisfies $\mathbf{V}_{n,N} \to \mathbf{V}_\alpha^{\mathsf{IBOSS}}$ a.s. when $n = \lfloor \alpha N \rfloor$ and $N \to \infty$, with*

$$\{\mathbf{V}_\alpha^{\mathsf{IBOSS}}\}_{k,k} = \frac{1}{\alpha} \left[\mathsf{E}[\{X\}_k^2] - \pi_k \, s_k(\alpha) \right] , \; k = 1 \ldots, d \,, \tag{5.3}$$

$$\{\mathbf{V}_\alpha^{\mathsf{IBOSS}}\}_{k,k'} = \frac{1}{\alpha} \left[\mathsf{E}[\{X\}_k] \, \mathsf{E}[\{X\}_k'] - \frac{\pi_k \, \pi_{k'}}{1-\alpha} \, m_k(\alpha) \, m_{k'}(\alpha) \right] , \; k \neq k' \,, \tag{5.4}$$

where $\mathsf{E}[\{X\}_k] = \int_{-\infty}^{\infty} x \, \varphi_k(x) \, \mathrm{d}x$, $\mathsf{E}[\{X\}_k^2] = \int_{-\infty}^{\infty} x^2 \, \varphi_k(x) \, \mathrm{d}x$, $\pi_k = (1-\alpha)[d-(k-1)\alpha]/(d-k\alpha)$,

$$s_k(\alpha) = \int_{q_k\left(\frac{\alpha}{2[d-(k-1)\alpha]}\right)}^{q_k\left(1-\frac{\alpha}{2[d-(k-1)\alpha]}\right)} x^2 \, \varphi_k(x) \, \mathrm{d}x \quad \textit{and} \quad m_k(\alpha) = \int_{q_k\left(\frac{\alpha}{2[d-(k-1)\alpha]}\right)}^{q_k\left(1-\frac{\alpha}{2[d-(k-1)\alpha]}\right)} x \, \varphi_k(x) \, \mathrm{d}x \,,$$

with $q_k(\cdot)$ the quantile function for φ_k, satisfying $\int_{-\infty}^{q_k(t)} \varphi_k(u) \, \mathrm{d}u = t$ for any $t \in (0, 1]$.

Proof. By construction, IBOSS asymptotically first selects all points such that $\{X\}_1$ does not belong to $\mathcal{I}_1 = (q_1[\alpha/(2d)], q_1[1-\alpha/(2d)])$, then, among remaining points, all those such that $\{X\}_2 \notin \mathcal{I}_2 = (q_2[\alpha/(2d(1-\alpha/d))], q_1[1-\alpha/(2d(1-\alpha/d))])$. By induction, all points such that $\{X\}_k \notin \mathcal{I}_k = (q_k[\alpha/(2[d-(k-1)\alpha])], q_k[1-\alpha/(2[d-(k-1)\alpha])])$ are selected at stage $k \in \{3, \ldots, d\}$. Denote $\mathbf{x} = (x_1, \ldots, x_d)^\top$. We have

$$\begin{aligned} \{\mathbf{V}_\alpha^{\mathsf{IBOSS}}\}_{k,k} &= \frac{1}{\alpha} \int_{\mathscr{X} \setminus \prod_{\ell=1}^d \mathcal{I}_\ell} x_k^2 \, \varphi(\mathbf{x}) \, \mathrm{d}\mathbf{x} = \frac{1}{\alpha} \left[\int_{\mathscr{X}} x_k^2 \, \varphi(\mathbf{x}) \, \mathrm{d}\mathbf{x} - \int_{\prod_{\ell=1}^d \mathcal{I}_\ell} x_k^2 \, \varphi(\mathbf{x}) \, \mathrm{d}\mathbf{x} \right] \\ &= \frac{1}{\alpha} \left[\mathsf{E}[\{X\}_k^2] - \left(\prod_{\ell \neq k} \Pr\{\{X\}_\ell \in \mathcal{I}_\ell\} \right) \int_{\mathcal{I}_k} x^2 \, \varphi_k(x) \, \mathrm{d}x \right] . \end{aligned}$$

Direct calculation gives $\Pr\{X \in \prod_{k=1}^d \mathcal{I}_k\} = 1 - \alpha$ and

$$\prod_{\ell \neq k} \Pr\{\{X\}_\ell \in \mathcal{I}_\ell\} = \frac{1-\alpha}{\Pr\{\{X\}_k \in \mathcal{I}_k\}} = \frac{1-\alpha}{1-\frac{\alpha}{d-(k-1)\alpha}} = \pi_k \,,$$

which proves (5.3). Similarly,

$$\begin{aligned} \{\mathbf{V}_\alpha^{\mathsf{IBOSS}}\}_{k,k'} &= \frac{1}{\alpha} \int_{\mathscr{X} \setminus \prod_{\ell=1}^d \mathcal{I}_\ell} x_k \, x_{k'} \, \varphi(\mathbf{x}) \, \mathrm{d}\mathbf{x} = \frac{1}{\alpha} \left[\int_{\mathscr{X}} x_k \, x_{k'} \, \varphi(\mathbf{x}) \, \mathrm{d}\mathbf{x} - \int_{\prod_{\ell=1}^d \mathcal{I}_\ell} x_k \, x_{k'} \, \varphi(\mathbf{x}) \, \mathrm{d}\mathbf{x} \right] \\ &= \frac{1}{\alpha} \left[\mathsf{E}[\{X\}_k] \, \mathsf{E}[\{X\}_k'] - \left(\prod_{\ell \neq k, \, \ell \neq k'} \Pr\{\{X\}_\ell \in \mathcal{I}_\ell\} \right) m_k(\alpha) \, m_{k'}(\alpha) \right] , \end{aligned}$$

with

$$\prod_{\ell \neq k, \, \ell \neq k'} \Pr\{\{X\}_\ell \in \mathcal{I}_\ell\} = \frac{1-\alpha}{\Pr\{\{X\}_k \in \mathcal{I}_k\} \, \Pr\{\{X\}_{k'} \in \mathcal{I}_{k'}\}} = \frac{\pi_k \, \pi_{k'}}{1-\alpha} \,,$$

which proves (5.4) and concludes the proof. ■

A key difference between IBOSS and Algorithm 1 is that IBOSS is nonsequential and therefore cannot be used in the streaming setting. Also, IBOSS is motivated by D-optimal design and may not perform well for other criteria, whereas Algorithm 1 converges to the optimal solution when $n = \lfloor \alpha N \rfloor$ and $N \to \infty$ for any criterion satisfying H_Φ. Moreover, IBOSS strongly relies on the assumption that $\mathscr{M}(X) = \mathbf{f}(X)\mathbf{f}^\top(X)$ with $\mathbf{f}(X) = [1,\ X^\top]^\top$ and, as the next example illustrates, it can perform poorly in other situations, in particular when the X_i are functionally dependent.

Example 7: quadratic regression on $[0,1]$ Take $\mathbf{f}(X) = [X,\ X^2]^\top$, with X uniformly distributed in $[0,1]$ and $\Phi(\mathbf{M}) = \log\det(\mathbf{M})$. For $\alpha \le \alpha_* \simeq 0.754160$, the optimal measure ξ^*_α equals μ/α on $[1/2-a, 1/2+b] \cup [1-(\alpha-a-b), 1]$ for some $a > b$ (which are determined by the two equations $F_\Phi[\mathbf{M}^*_\alpha, \mathscr{M}(1/2-a)] = F_\Phi[\mathbf{M}^*_\alpha, \mathscr{M}(1/2+b)] = F_\Phi[\mathbf{M}^*_\alpha, \mathscr{M}(1-(\alpha-a-b))]$). For $\alpha \ge \alpha_*$, $\xi^*_\alpha = \mu/\alpha$ on $[1-\alpha, 1]$. When $n = \lfloor \alpha N \rfloor$ when $N \to \infty$, the matrix $\mathbf{M}^{\mathsf{IBOSS}}_{n,N}$ obtained with IBOSS applied to the points $\mathbf{f}(X_i)$ converges to $\mathbf{M}^{\mathsf{IBOSS}}_\alpha = \mathbf{M}(\xi^{\mathsf{IBOSS}}_\alpha)$, with $\xi^{\mathsf{IBOSS}}_\alpha = \mu/\alpha$ on $[0, \alpha/2] \cup [1-\alpha/2, 1]$. The left panel of Figure 13 shows $\det(\mathbf{M}^*_\alpha)$ (red solid line) and $\det(\mathbf{M}^{\mathsf{IBOSS}}_\alpha)$ (blue dotted line) as functions of $\alpha \in [0,1]$. We have $\det(\mathbf{M}^{\mathsf{IBOSS}}_\alpha) = (1/960)\alpha^2(\alpha^4 + 25 - 40\alpha + 26\alpha^2 - 8\alpha^3)$, which tends to 0 as $\alpha \to 0$. ◁

Next examples show that IBOSS performs more comparably to Algorithm 1 for multilinear regression with intercept, where $\mathscr{M}(X) = \mathbf{f}(X)\mathbf{f}^\top(X)$ with $\mathbf{f}(X) = [1,\ X^\top]^\top$. Its performance may nevertheless be significantly poorer than that of Algorithm 1.

Example 8: multilinear regression with intercept, $\Phi(\mathbf{M}) = \log\det(\mathbf{M})$

X **is uniformly distributed in** $[-1,1]^2$**.** Direct calculation shows that, for any $\alpha \in [0,1]$, the optimal measure ξ^*_α equals μ/α on $[-1,1]^2 \setminus \mathscr{B}_2(\mathbf{0}, R_\alpha)$, with $\mathscr{B}_2(\mathbf{0}, r)$ the open ball centered at the origin with radius r. Here, $R_\alpha = 2\sqrt{(1-\alpha)/\pi}$ when $\alpha \ge 1 - \pi/4$, and $R_\alpha > 1$ is solution of $1 + \pi R^2/4 - \sqrt{R^2-1} - R^2 \arcsin(1/R) = \alpha$ when $\alpha \in (1-\pi/4, 1]$. The associated optimal matrix is diagonal, $\mathbf{M}^*_\alpha = \mathrm{diag}\{1, \rho_\alpha, \rho_\alpha\}$, with

$$\rho_\alpha = \begin{cases} \frac{1}{2\alpha}\left[2/3 - 2(1-\alpha)^2/\pi\right] & \text{if } 0 \le \alpha \le 1 - \pi/4\,, \\ \frac{1}{2\alpha}\left[2/3 + \pi R_\alpha^4/8 - (R_\alpha^4/2)\arcsin(1/R_\alpha) - \sqrt{R_\alpha^2 - 1}\,(R_\alpha^2 + 2)/6\right] & \text{if } 1 - \pi/4 < \alpha \le 1\,. \end{cases}$$

Extension to $d > 2$ is possible but involves complicated calculations.

When $n = \lfloor \alpha N \rfloor$ and $N \to \infty$, the matrix $\mathbf{M}^{\mathsf{IBOSS}}_{n,N}$ obtained with IBOSS converges to $\mathbf{M}^{\mathsf{IBOSS}}_\alpha = \mathbf{M}(\xi^{\mathsf{IBOSS}}_\alpha)$ when $n = \lfloor \alpha N \rfloor$ and $N \to \infty$, with $\xi^{\mathsf{IBOSS}}_\alpha = \mu/\alpha$ on $[-1,1]^2 \setminus ([-1+a, 1-a] \times [-1+b, 1-b])$, with $a = \alpha/2$ and $b = \alpha/(1-\alpha) \ge a$. The matrix $\mathbf{M}^{\mathsf{IBOSS}}_\alpha$ is diagonal, $\mathbf{M}^{\mathsf{IBOSS}}_\alpha = \mathrm{diag}\{1, D_{\alpha,1}, D_{\alpha,2}\}$, where $\mathbf{V}^{\mathsf{IBOSS}}_\alpha = \mathrm{diag}\{D_{\alpha,1}, D_{\alpha,2}\}$ is the matrix in Theorem 5.1 with $D_{\alpha,1} = (8 - 5\alpha + \alpha^2)/12$ and $D_{\alpha,2} = (8 - 11\alpha + 4\alpha^2)/[3(2-\alpha)^2]$. The right panel of Figure 13 shows $\det(\mathbf{M}^*_\alpha)$ (red solid line) and $\det(\mathbf{M}^{\mathsf{IBOSS}}_\alpha)$ (blue dashed line) as functions of $\alpha \in [0,1]$. Note that $\det(\mathbf{M}^*_\alpha) \to 1$ whereas $\det(\mathbf{M}^{\mathsf{IBOSS}}_\alpha) \to 4/9$ when $\alpha \to 0$. The problem is due to selection by IBOSS of points having one coordinate in the central part of the interval. ◁

X **is normally distributed** $\mathscr{N}(\mathbf{0}, \mathbf{I}_d)$**.** The expression of the optimal matrix $\mathbf{M}^*_\alpha$ has been derived in Section 4.2; the asymptotic value for $N \to \infty$ of the matrix $\mathbf{M}_{\lfloor \alpha N \rfloor, N}$ is

$$\mathbf{M}^{\mathsf{IBOSS}}_\alpha = \begin{pmatrix} 1 & \mathbf{0}^\top \\ \mathbf{0} & \mathbf{V}^{\mathsf{IBOSS}}_\alpha \end{pmatrix},$$

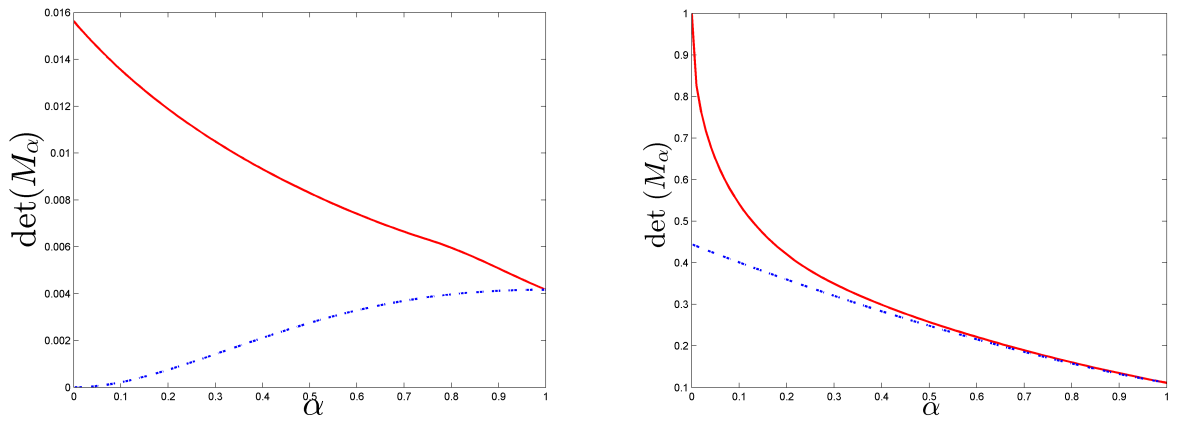

Figure 13: $\det(\mathbf{M}_\alpha^*)$ (red solid line) and $\det(\mathbf{M}_\alpha^{\mathsf{IBOSS}})$ (blue dotted line) as functions of $\alpha \in [0,1]$. Left: quadratic regression on $[0,1]$; Right: multilinear regression with intercept on $[-1,1]^2$.

where the expression of $\mathbf{V}_\alpha^{\mathsf{IBOSS}}$ (here a diagonal matrix) is given in Theorem 5.1. Figure 14 shows the D-efficiency $\det^{1/(d+1)}(\mathbf{M}_\alpha^{\mathsf{IBOSS}})/\det^{1/(d+1)}(\mathbf{M}_\alpha^*)$ as a function of $\alpha \in (0,1]$ for $d = 3$ (left) and $d = 25$ (right), showing that the performance of IBOSS deteriorates as d increases. We also performed series of simulations for $d = 25$, with 100 independent repetitions of selections of $n = \lfloor \alpha N \rfloor$ points within $\mathscr{X}_N$ ($N = 10\,000$) based on IBOSS and Algorithm 1. Due to the small value of N, we apply Algorithm 1 to replications $\mathscr{X}_{mN} = \mathscr{X}_N \cup \mathscr{X}_N^{(2)} \cup \cdots \cup \mathscr{X}_N^{(m)}$ of $\mathscr{X}_N$, see Section 5.1, with $m = 99$ for $\alpha < 0.1$, $m = 9$ for $0.1 \leq \alpha < 0.5$ and $m = 4$ for $\alpha \geq 0.5$. The colored areas on Figure 14 show the variability range for efficiency, corresponding to the empirical mean $\pm$ 2 standard deviations obtained for the 100 repetitions, for IBOSS (green, bottom) and Algorithm 1 (magenta, top); note that variability decreases as $n = \lfloor \alpha N \rfloor$ increases. The approximation of $\mathbf{M}_{n,N}$ obtained with IBOSS by the asymptotic matrix $\mathbf{M}_\alpha^{\mathsf{IBOSS}}$ is quite accurate although N is rather small; Algorithm 1 (incorporating m repetitions of $\mathscr{X}_N$) performs significantly better than IBOSS although the setting is particularly favorable to IBOSS — it is significantly slower than IBOSS, however, when m is large.

6 Conclusions and further developments

We have proposed a sequential subsampling method for experimental design (Algorithm 1) that converges to the optimal solution when the length of the sequence tends to infinity and a fixed proportion of design points is selected. Since the method only needs to keep the memory of the current information matrix associated with the design problem (or its inverse), and to update a pair of scalar variables (an estimated quantile, and an estimate of the p.d.f. value at the quantile), it can be applied to sequences of arbitrary length and is suitable for data streaming.

We have not tried to optimize the choice of initialization and tuning parameters in Algorithm 1. Although it does not seem critical (the same tuning has been used in all the examples presented), there is certainly an opportunity to improve, in particular concerning β_0 and $\widehat{C}_{k_0}$ (for instance, using the information that $C_{1-\alpha}^* \leq 0$ whereas $\widehat{C}_{k_0} > 0$ for small α with the initialization we use).

We have only considered the case of linear models, where the information matrix does not depend on unknown parameters (equivalent to local optimum design in case of a nonlinear model),

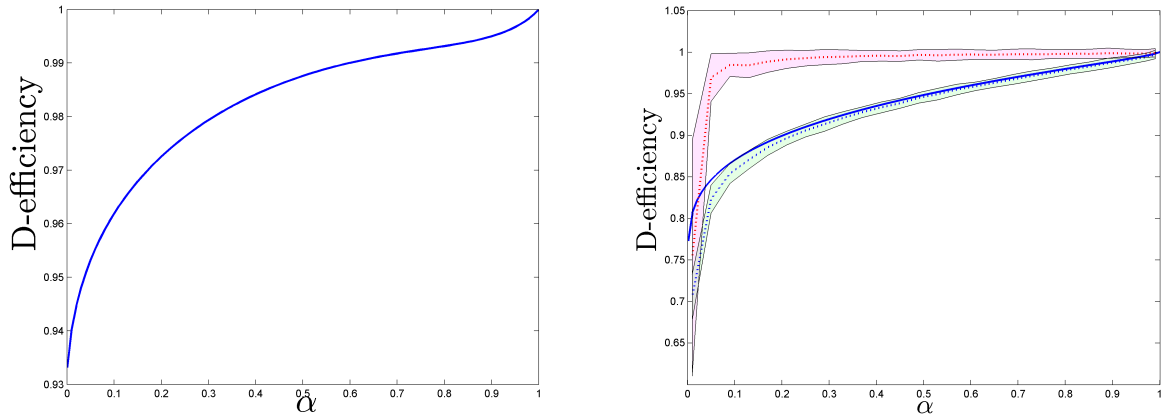

Figure 14: D-efficiency of IBOSS (blue solid line) as a function of $\alpha \in (0, 1]$ for $d = 3$ (left) and $d = 25$ (right). The enveloppes on the right panel show the empirical mean efficiency $\pm$ 2 standard deviations obtained for 100 independent repetitions with $n = \lfloor \alpha N \rfloor$ and $N = 10\,000$ for IBOSS (green, bottom) and Algorithm 1 (magenta, top).

but extension to online parameter estimation in a nonlinear model with $\mathscr{M}(x) = \mathscr{M}(x, \boldsymbol{\theta})$ would not require important modifications. Denote by $\hat{\boldsymbol{\theta}}^n$ the estimated value of the parameters after observation at the n design points selected, $X_{i_1}, \ldots, X_{i_n}$, say. Then, we can use $\mathbf{M}_{n_{k_0}} = (1/k_0) \sum_{i=1}^{k_0} \mathscr{M}(X_i, \hat{\boldsymbol{\theta}}^{k_0})$ at Step 1 of Algorithm 1, and $\mathbf{M}_{n_k+1}$ given by (3.5) can be replaced by $\mathbf{M}_{n_k+1} = [1/(n_k + 1)] \, [\sum_{j=1}^{n_k} \mathscr{M}(X_{i_j}, \hat{\boldsymbol{\theta}}^{n_k}) + \mathscr{M}(X_{k+1}, \hat{\boldsymbol{\theta}}^{n_k})]$ at Step 2. Recursive estimation can be used for $k > k_0$ to reduce computational cost. For instance for maximum likelihood estimation, with the notation of Section 1, we can update $\hat{\boldsymbol{\theta}}^{n_k}$ as

$$\hat{\boldsymbol{\theta}}^{n_k+1} = \hat{\boldsymbol{\theta}}^{n_k} + \frac{1}{n_k + 1} \mathbf{M}_{n_k+1}^{-1} \frac{\partial \log \varphi_{X_{k+1},\theta}(Y_{k+1})}{\partial \theta}\bigg|_{\theta=\hat{\boldsymbol{\theta}}^{n_k}}$$

when X_{k+1} is selected; see Ljung and Söderström (1983); Tsypkin (1983). A further simplification would be to update $\mathbf{M}_{n_k}$ as $\mathbf{M}_{n_{k+1}} = \mathbf{M}_{n_k} + [1/(n_k + 1)] \, [\mathscr{M}(X_{k+1}, \hat{\boldsymbol{\theta}}^{n_k}) - \mathbf{M}_{n_k}]$. When the X_i are i.i.d. with μ satisfying H_μ, the strong consistency of $\hat{\boldsymbol{\theta}}^{n_k}$ holds with such recursive schemes under rather general conditions when all X_i are selected. Showing that this remains true when only a proportion α is selected by Algorithm 1 requires technical developments outside the scope of this paper, but we anticipate that $\mathbf{M}_{n_k} \to \mathbf{M}^*_{\alpha,\overline{\boldsymbol{\theta}}}$ a.s., with $\mathbf{M}^*_{\alpha,\overline{\boldsymbol{\theta}}}$ the optimal matrix for the true value $\overline{\boldsymbol{\theta}}$ of the model parameters.

Algorithm 1 can be viewed as an adaptive version of the treatment allocation method presented in (Metelkina and Pronzato, 2017): consider the selection or rejection of X_i as the allocation of individual i to treatment 1 (selection) or 2 (rejection), with respective contributions $\mathscr{M}_1(X_i) = \mathscr{M}(X_i)$ or $\mathscr{M}_2(X_i) = 0$ to the collection of information; count a cost of one for allocation to treatment 1 and zero for rejection. Then, the doubly-adaptive sequential allocation (4.6) of Metelkina and Pronzato (2017) that optimizes a compromise between information and cost exactly coincides with Algorithm 1 where $\widehat{C}_k$ is frozen to a fixed C, i.e., without Step 3. In that sense, the two-time-scale stochastic approximation procedure of Algorithm 1 opens the way to the development of adaptive treatment allocation procedures where the proportion of individuals allocated to the poorest treatment could be adjusted online to a given target.

Finally, the designs obtained with the proposed thinning procedure are model-based: when the model is wrong, ξ_α^* is no longer optimal for the true model. Model-robustness issues are not considered in the paper and would require specific developments, following for instance the approach in (Wiens, 2005; Nie et al., 2018).

A Maximum of $\Phi(\mathbf{M}_{n_k})$

The property below is stated without proof in (Pronzato, 2006). We provide here a formal proof based on results on conditional value-at-risk by Rockafellar and Uryasev (2000) and Pflug (2000).

Lemma A.1. *Suppose that $n_k/k \to \alpha$ as $k \to \infty$. Then, under H_Φ and $H_{\mathscr{M}}$, for any choice of n_k points X_i among k points i.i.d. with μ, we have $\limsup_{k\to\infty} \Phi(\mathbf{M}_{n_k,k}) \le \Phi(\mathbf{M}_\alpha^*)$ a.s., where $\mathbf{M}_\alpha^*$ maximizes $\Phi(\mathbf{M})$ with respect to $\mathbf{M} \in \mathbb{M}(\alpha)$.*

Proof. Denote by $\mathbf{M}_{n_k,k}^*$ the matrix that corresponds to choosing n_k distinct candidates that maximize $\Phi(\mathbf{M}_{n_k,k})$. The concavity of Φ implies

$$\Phi(\mathbf{M}_{n_k,k}^*) \le \Phi(\mathbf{M}_\alpha^*) + \mathrm{trace}[\nabla_\Phi(\mathbf{M}_\alpha^*)(\mathbf{M}_{n_k,k}^* - \mathbf{M}_\alpha^*)]\,. \tag{A.1}$$

The rest of the proof consists in deriving an upper bound on the second term on the right-hand side of (A.1).

Denote $z_i = \mathrm{trace}[\nabla_\Phi(\mathbf{M}_\alpha^*)\mathscr{M}(X_i)]$ for all $i = 1, \ldots, k$ and let the $z_{i:k}$ denote the version sorted by decreasing values. Since Φ is increasing for Loewner ordering, $\Phi(\mathbf{M}) \le \Phi(\mathbf{M} + \mathbf{z}\mathbf{z}^\top)$ for any $\mathbf{M} \in \mathbb{M}^{\ge}$ and any $\mathbf{z} \in \mathbb{R}^p$, and concavity implies $\Phi(\mathbf{M} + \mathbf{z}\mathbf{z}^\top) \le \Phi(\mathbf{M}) + \mathbf{z}^\top \nabla_\Phi(\mathbf{M})\mathbf{z}$, showing that $\nabla_\Phi(\mathbf{M}) \in \mathbb{M}^{\ge}$. Therefore, $z_{i:k} \ge 0$ for all i.

First, we may notice that $\mathrm{trace}[\nabla_\Phi(\mathbf{M}_\alpha^*)\mathbf{M}_{n_k,k}^*] \le (1/n_k)\sum_{i=1}^{n_k} z_{i:k}$ and that

$$\mathrm{trace}[\nabla_\Phi(\mathbf{M}_\alpha^*)\mathbf{M}_\alpha^*] = \frac{1}{\alpha}\int_{\mathscr{X}} \mathbb{I}_{\{\mathrm{trace}[\nabla_\Phi(\mathbf{M}_\alpha^*)\mathscr{M}(x)] \ge c_{1-\alpha}\}}\, \mathrm{trace}[\nabla_\Phi(\mathbf{M}_\alpha^*)\mathscr{M}(x)]\, \mu(\mathrm{d}x)$$

with $c_{1-\alpha} \ge 0$ and such that $\int_{\mathscr{X}} \mathbb{I}_{\{\mathrm{trace}[\nabla_\Phi(\mathbf{M}_\alpha^*)\mathscr{M}(x)] \ge c_{1-\alpha}\}}\, \mu(\mathrm{d}x) = \alpha$; see (2.2).

Following Rockafellar and Uryasev (2000); Pflug (2000), we then define the functions $g(x;\beta,a) = a + (1/\beta)\,[\mathrm{trace}[\nabla_\Phi(\mathbf{M}_\beta^*)\mathscr{M}(x) - a]^+$, $x \in \mathscr{X}$, $\beta \in (0,1)$, $a \in \mathbb{R}$. We can then write, for any $\beta \in (0,1)$,

$$\mathrm{trace}[\nabla_\Phi(\mathbf{M}_\beta^*)\mathbf{M}_\beta^*] = \mathsf{E}\{g(X;\beta,c_{1-\beta})\} = \inf_a \mathsf{E}\{g(X;\beta,a)\} \ge c_{1-\beta}\,, \tag{A.2}$$

and

$$\frac{1}{n_k}\sum_{i=1}^{n_k} z_{i:k} = \mathsf{E}_{\mu_k}\{g(X;\alpha_k,z_{n_k:k})\} = \inf_a \mathsf{E}_{\mu_k}\{g(X;\alpha_k,a)\}\,,$$

where $\alpha_k = n_k/k \in (\alpha/2, 1]$ for all k larger than some k_1 and where $\mathsf{E}_{\mu_k}\{\cdot\}$ denotes expectation for the empirical measure $\mu_k = (1/k)\sum_{i=1}^k \delta_{X_i}$.

Next, we construct an upper bound on $z_{n_k:k}$. For $k > k_1$, the matrix $\mathbf{M}_k = (1/k)\sum_{i=1}^k \mathscr{M}(X_i)$ satisfies

$$\mathrm{trace}[\nabla_\Phi(\mathbf{M}_\alpha^*)\mathbf{M}_k] = (1/k)\sum_{i=1}^{k} z_{i:k} \ge (n_k/k)\, z_{n_k:k} > (\alpha/2)\, z_{n_k:k}\,. \tag{A.3}$$

Now, $\mathbf{M}_\alpha^* = \mathbf{M}(\xi_\alpha^*)$ with $\xi_\alpha^* = \mu/\alpha$ on a set $\mathscr{X}_\alpha^* \subset \mathscr{X}$ and $\xi_\alpha^* = 0$ elsewhere, and $\mu(\mathscr{X}_\alpha^*) = \alpha\,\xi_\alpha^*(\mathscr{X}_\alpha^*) = \alpha\,\xi_\alpha^*(\mathscr{X}) = \alpha$. $\mathrm{H}_{\mathscr{M}}$-(ii) then implies that $\lambda_{\min}(\mathbf{M}_\alpha^*) = (1/\alpha)\,\lambda_{\min}[\int_{\mathscr{X}_\alpha^*} \mathscr{M}(x)\,\mu(\mathrm{d}x)] >$

ℓ_α/α, and H_Φ implies that $\|\nabla_\Phi(\mathbf{M}^*_\alpha)\| < A(\ell_\alpha/\alpha) < \infty$. Therefore $\mathrm{trace}[\nabla_\Phi(\mathbf{M}^*_\alpha)\mathbf{M}(\mu)] < A_\alpha = A(\ell_\alpha/\alpha)\sqrt{pB}$ from $\mathrm{H}_{\mathscr{M}}$-(i). Since $\mathrm{trace}[\nabla_\Phi(\mathbf{M}^*_\alpha)\mathbf{M}_k]$ tends to $\mathrm{trace}[\nabla_\Phi(\mathbf{M}^*_\alpha)\mathbf{M}(\mu)]$ a.s. as $k \to \infty$, (A.3) implies that there exists a.s. k_2 such that, for all $k > k_2$, $z_{n_k:k} < A_\alpha/(4\alpha)$.

To summarize, (A.1) implies

$$\begin{aligned}\Phi(\mathbf{M}^*_{n_k,k}) &\leq \Phi(\mathbf{M}^*_\alpha) + \mathsf{E}_{\mu_k}\{g(X;\alpha_k,z_{n_k:k})\} - \mathsf{E}\{g(X;\alpha,c_{1-\alpha})\} \\ &\leq \Phi(\mathbf{M}^*_\alpha) + |\mathsf{E}_{\mu_k}\{g(X;\alpha_k,z_{n_k:k})\} - \mathsf{E}\{g(X;\alpha_k,c_{1-\alpha_k})\}| \\ &\quad + \left|\mathrm{trace}[\nabla_\Phi(\mathbf{M}^*_{\alpha_k})\mathbf{M}^*_{\alpha_k}] - \mathrm{trace}[\nabla_\Phi(\mathbf{M}^*_\alpha)\mathbf{M}^*_\alpha]\right| .\end{aligned}$$

The last term tends to zero as k tends to infinity, due to (A.2) and the continuity of conditional value-at-risk; see (Rockafellar and Uryasev, 2002, Prop. 13). Since $c_{1-\alpha_k} \leq \mathrm{trace}[\nabla_\Phi(\mathbf{M}^*_{\alpha_k})\mathbf{M}^*_{\alpha_k}]$, see (A.2), and $\alpha_k \to \alpha$, for all k large enough we have $c_{1-\alpha_k} \leq 2\,\mathrm{trace}[\nabla_\Phi(\mathbf{M}^*_\alpha)\mathbf{M}^*_\alpha]$. Denote $\bar{a} = \max\{A_\alpha/(4\alpha), 2\,\mathrm{trace}[\nabla_\Phi(\mathbf{M}^*_\alpha)\mathbf{M}^*_\alpha]\}$. The second term can then be rewritten as

$$\begin{aligned}|\mathsf{E}_{\mu_k}\{g(X;\alpha_k,z_{n_k:k})\} - \mathsf{E}\{g(X;\alpha_k,c_{1-\alpha_k})\}| &= \left|\inf_{a\in[0,\bar{a}]} \mathsf{E}_{\mu_k}\{g(X;\alpha_k,a)\} - \inf_{a\in[0,\bar{a}]} \mathsf{E}\{g(X;\alpha_k,a)\}\right| \\ &\leq \max_{a\in[0,\bar{a}]} |\mathsf{E}_{\mu_k}\{g(X;\alpha_k,a)\} - \mathsf{E}\{g(X;\alpha_k,a)\}| .\end{aligned}$$

The functions $g(\cdot;t,a)$ with $t \in (\alpha/2, 1]$, $a \in [0,\bar{a}]$, form a Glivenko-Cantelli class; see (van der Vaart, 1998, p. 271). It implies that $\max_{a\in[0,\bar{a}]} |\mathsf{E}_{\mu_k}\{g(X;\alpha_k,a)\} - \mathsf{E}\{g(X;\alpha_k,a)\}| \to 0$ a.s., which concludes the proof. ∎

The class of functions $g(\cdot;t,a)$ is in fact Donsker (van der Vaart, 1998, p. 271). The strict concavity of $\Phi(\cdot)$ implies that optimal matrices are unique, and in complement of Lemma A.1 we get $\|\mathbf{M}^*_{\lfloor\alpha k\rfloor,k} - \mathbf{M}^*_\alpha\| = \mathcal{O}_p(1/\sqrt{k})$. Note that when an optimal bounded design measure ξ^*_α is known, a selection procedure such that $n_k/k \to \alpha$ and $\Phi(\mathbf{M}_{n_k,k}) \to \Phi(\mathbf{M}^*_\alpha)$ a.s. is straightforwardly available: simply select the points that belong to the set $\mathscr{X}^*_\alpha$ on which $\xi^*_\alpha = \mu/\alpha$.

B Non degeneracy of $\mathbf{M}_{n_k}$

To invoke $\mathrm{H}_{\mu,\mathscr{M}}$ in order to ensure the existence of a density $\varphi_{\mathbf{M}_{n_k}}$ having the required properties for all k (which is essential for the convergence of Algorithm 1, see Theorem 3.1), we need to guarantee that $\mathbf{M}_{n_k} \in \mathbb{M}^{\geq}_{\ell,L}$ for all k, for some ℓ and L. This is the object of the following lemma.

Lemma B.1. *Under $\mathrm{H}_{\mathscr{M}}$, when $\epsilon_1 > 0$ in Algorithm 1, $n_{k+1}/k > \epsilon_1$ for all k and there exists a.s. $\ell > 0$ and $L < \infty$ such that $\mathbf{M}_{n_k} \in \mathbb{M}^{\geq}_{\ell,L}$ for all $k > k_0$.*

Proof. Since the first k_0 points are selected, we have $n_k/k = 1 > \epsilon_1$ for $k \leq k_0$. Let k_* be the first k for which $n_k/k < \epsilon_1$. It implies that $n_{k_*} = n_{k_*-1} > (k_*-1)\,\epsilon_1$, and (3.4) implies that $n_{k_*+1} = n_{k_*} + 1$. Therefore, $n_{k_*+1}/k_* > \epsilon_1 + (1-\epsilon_1)/k_* > \epsilon_1$, and $n_k/(k-1) > \epsilon_1$ for all $k > 1$.

If the n_k points were chosen randomly, $n_k > (k-1)\,\epsilon_1$ would be enough to obtain that, from $\mathrm{H}_{\mathscr{M}}$, $\lambda_{\min}(\mathbf{M}_{n_k}) > \ell_{\epsilon_1}/2$ and $\lambda_{\max}(\mathbf{M}_{n_k}) < \sqrt{B}/2$ for all k larger than some k_1. However, here the situation is more complicated since points are accepted or rejected according to a sequential decision rule, and a more sophisticated argumentation is required. An expedite solution is to consider the worst possible choices of n_k points, that yield the smallest value of $\lambda_{\min}(\mathbf{M}_{n_k})$ and largest value of $\lambda_{\max}(\mathbf{M}_{n_k})$. This approach is used in Lemma B.2 presented below, which permits to conclude the proof. ∎

Lemma B.2. *Under $H_{\mathscr{M}}$, any matrix $\mathbf{M}_{n_k}$ obtained by choosing n_k points out of k independently distributed with μ and such that $n_k/k > \epsilon > 0$ satisfies $\liminf_{k\to\infty} \lambda_{\min}(\mathbf{M}_{n_k}) > \ell$ and $\limsup_{k\to\infty} \lambda_{\max}(\mathbf{M}_{n_k}) < L$ a.s. for some $\ell > 0$ and $L < \infty$.*

Proof. We first construct a lower bound on $\liminf_{k\to\infty} \lambda_{\min}(\mathbf{M}_{n_k})$. Consider the criterion $\Phi_\infty^+(\mathbf{M}) = \lambda_{\min}(\mathbf{M})$, and denote by $\mathbf{M}^*_{n_k,k}$ the n_k-point design matrix that minimizes Φ_∞^+ over the design space formed by k points X_i i.i.d. with μ. We can write $\mathbf{M}^*_{n_k,k} = (1/n_k)\sum_{i=1}^{n_k} \mathscr{M}(X_{k_i})$, where the k_i correspond to the indices of positive u_i in the minimization of $f(\mathbf{u}) = \Phi_\infty^+[\sum_{i=1}^k u_i \mathscr{M}(X_i)]$ with respect to $\mathbf{u} = (u_1, \ldots, u_k)$ under the constraints $u_i \in \{0,1\}$ for all i and $\sum_i u_i = n_k$. Obviously, any matrix $\mathbf{M}_{n_k}$ obtained by choosing n_k distinct points X_i among $X_1, \ldots, X_k$ satisfies $\lambda_{\min}(\mathbf{M}_{n_k}) \geq \lambda_{\min}(\mathbf{M}^*_{n_k,k})$.

For any $\mathbf{M} \in \mathbb{M}^{\geq}$, denote $\mathscr{U}(\mathbf{M}) = \{\mathbf{u} \in \mathbb{R}^p : \|\mathbf{u}\| = 1,\ \mathbf{M}\mathbf{u} = \lambda_{\min}(\mathbf{M})\mathbf{u}\}$. Then, for any $\mathbf{u} \in \mathscr{U}(\mathbf{M}^*_{n_k,k})$, $\mathbf{u}^\top \mathbf{M}^*_{n_k,k}\mathbf{u} = \lambda_{\min}(\mathbf{M}^*_{n_k,k}) = \min_{\mathbf{v}\in\mathbb{R}^p:\,\|\mathbf{v}\|=1} \mathbf{v}^\top \mathbf{M}^*_{n_k,k}\mathbf{v} = (1/n_k)\sum_{i=1}^{n_k} z_{i:k}(\mathbf{u})$, where the $z_{i:k}(\mathbf{u})$ correspond to the values of $\mathbf{u}^\top \mathscr{M}(X_i)\mathbf{u}$ sorted by increasing order for $i = 1, \ldots, k$. For any $m \in \{1, \ldots, n_k - 1\}$, we thus have

$$\lambda_{\min}(\mathbf{M}^*_{n_k,k}) \geq \frac{1}{m}\sum_{i=1}^{m} z_{i:k}(\mathbf{u}) \geq \lambda_{\min}(\mathbf{M}^*_{m,k})\,,$$

showing that the worst situation corresponds to the smallest admissible n_k; that is, we only have to consider the case when $n_k/k \to \epsilon$ as $k \to \infty$.

Since Φ_∞^+ is concave, for any $\mathbf{M}' \in \mathbb{M}^{\geq}$ we have

$$\lambda_{\min}(\mathbf{M}') \leq \lambda_{\min}(\mathbf{M}^*_{n_k,k}) + F_{\Phi_\infty^+}(\mathbf{M}^*_{n_k,k}, \mathbf{M}')\,, \tag{B.1}$$

where $F_{\Phi_\infty^+}(\mathbf{M}, \mathbf{M}') = \min_{\mathbf{u}\in\mathscr{U}(\mathbf{M})} \mathbf{u}^\top(\mathbf{M}' - \mathbf{M})\mathbf{u}$ is the directional derivative of Φ_∞^+ at $\mathbf{M}$ in the direction $\mathbf{M}'$.

For any $\alpha \in (0,1)$ and any $\xi_\alpha \leq \mu/\alpha$, there exists a set $\mathscr{X}_\alpha \subset \mathscr{X}$ such that $\xi_\alpha \geq (1-\alpha)\mu$ on $\mathscr{X}_\alpha$ and $\mu(\mathscr{X}_\alpha) \geq \alpha^2$. Indeed, any set $\mathscr{Z}$ on which $\xi_\alpha < (1-\alpha)\mu$ is such that $\xi_\alpha(\mathscr{Z}) < (1-\alpha)\,\mu(\mathscr{Z}) \leq (1-\alpha)$; therefore, taking $\mathscr{X}_\alpha = \mathscr{X} \setminus \mathscr{Z}$, we get $\mu(\mathscr{X}_\alpha) \geq \alpha\,\xi_\alpha(\mathscr{X}_\alpha) \geq \alpha^2$. Denote $\alpha_k = n_k/k$, with $\alpha_k > \epsilon$ and $\alpha_k \to \epsilon$ as $k \to \infty$, and take any $\mathbf{M}' = \mathbf{M}(\xi_{\alpha_k}) \in \mathbb{M}(\alpha_k)$. Applying $\mathrm{H}_{\mathscr{M}}$-(ii) to the set $\mathscr{X}_{\alpha_k}$ defined above, we get

$$\begin{aligned}
\lambda_{\min}(\mathbf{M}') = \lambda_{\min}\left(\int_{\mathscr{X}} \mathscr{M}(x)\,\xi_{\alpha_k}(\mathrm{d}x)\right) &\geq \lambda_{\min}\left(\int_{\mathscr{X}_{\alpha_k}} \mathscr{M}(x)\,\xi_{\alpha_k}(\mathrm{d}x)\right) \\
&\geq (1-\alpha_k)\,\lambda_{\min}\left(\int_{\mathscr{X}_{\alpha_k}} \mathscr{M}(x)\,\mu(\mathrm{d}x)\right) > (1-\alpha_k)\,\ell_{\alpha_k^2}\,.
\end{aligned}$$

For k larger than some k_1 we have $\alpha_k \in (\epsilon, 2\epsilon)$, and therefore $\lambda_{\min}(\mathbf{M}') > c_\epsilon = (1-2\epsilon)\,\ell_{\epsilon^2} > 0$. The inequality (B.1) thus gives, for $k > k_1$,

$$c_\epsilon < \lambda_{\min}(\mathbf{M}^*_{n_k,k}) + \min_{\mathbf{u}\in\mathscr{U}(\mathbf{M}^*_{n_k,k})}\ \min_{\mathbf{M}'\in\mathbb{M}(\alpha_k)} \mathbf{u}^\top(\mathbf{M}' - \mathbf{M}^*_{n_k,k})\mathbf{u}\,. \tag{B.2}$$

The rest of the proof follows from results on conditional value-at-risk by Rockafellar and Uryasev (2000) and Pflug (2000). For a fixed $\mathbf{u} \in \mathbb{R}^p$, $\mathbf{u} \neq \mathbf{0}$, and $\alpha \in (0,1)$, we have

$$\min_{\mathbf{M}'\in\mathbb{M}(\alpha)} \mathbf{u}^\top\mathbf{M}'\mathbf{u} = \frac{1}{\alpha}\int_{\mathscr{X}} \mathbb{I}_{\{\mathbf{u}^\top\mathscr{M}(x)\mathbf{u} \leq a_\alpha(\mathbf{u})\}}\,[\mathbf{u}^\top\mathscr{M}(x)\mathbf{u}]\,\mu(\mathrm{d}x)\,,$$

where the α-quantile $a_\alpha(\mathbf{u})$ satisfies $\int_{\mathscr{X}} \mathbb{I}_{\{\mathbf{u}^\top \mathscr{M}(x)\mathbf{u} \le a_\alpha(\mathbf{u})\}}\, \mu(\mathrm{d}x) = \alpha$. For any $a \in \mathbb{R}$ and $\mathbf{u} \in \mathbb{R}^p$, denote

$$h(x;\alpha,a,\mathbf{u}) = a - \frac{1}{\alpha}\,[a - \mathbf{u}^\top \mathscr{M}(x)\mathbf{u}]^+\,,\ x \in \mathscr{X}\,.$$

We can write $\min_{\mathbf{M}' \in \mathbb{M}(\alpha)} \mathbf{u}^\top \mathbf{M}'\mathbf{u} = \mathsf{E}\{h(X;\alpha,a_\alpha(\mathbf{u}),\mathbf{u})\} = \sup_{a\in\mathbb{R}} \mathsf{E}\{h(X;\alpha,a,\mathbf{u})\}$, where the expectation is with respect to X distributed with μ (Rockafellar and Uryasev, 2000). Also, from Pflug (2000), for any $\mathbf{u} \in \mathscr{U}(\mathbf{M}^*_{n_k,k})$ we can write $\mathbf{u}^\top \mathbf{M}^*_{n_k,k}\mathbf{u} = \mathsf{E}_{\mu_k}\{h(X;\alpha_k,z_{n_k:k}(\mathbf{u}),\mathbf{u})\} = \sup_{a\in\mathbb{R}} \mathsf{E}_{\mu_k}\{h(X;\alpha_k,a,\mathbf{u})\}$, where $\mathsf{E}_{\mu_k}\{\cdot\}$ denotes expectation for the empirical measure $\mu_k = (1/k)\sum_{i=1}^k \delta_{X_i}$.

Now, from H$_{\mathscr{M}}$-(i), for any $\mathbf{u} \in \mathbb{R}^p$ with $\|\mathbf{u}\| = 1$,

$$(1-\alpha)\,a_\alpha(\mathbf{u}) \le \int_{\mathscr{X}} \mathbb{I}_{\{\mathbf{u}^\top \mathscr{M}(x)\mathbf{u} > a_\alpha(\mathbf{u})\}}\,[\mathbf{u}^\top \mathscr{M}(x)\mathbf{u}]\,\mu(\mathrm{d}x) < \sqrt{B}\,. \tag{B.3}$$

We also have $(k-n_k)\,z_{n_k:k}(\mathbf{u}) \le \sum_{i=n_k+1}^k z_{i:k}(\mathbf{u}) \le \sum_{i=1}^k z_{i:k}(\mathbf{u}) = k\,(\mathbf{u}^\top \mathbf{M}_k \mathbf{u}) \le k\,\lambda_{\max}(\mathbf{M}_k)$, with $\mathbf{M}_k \to \mathbf{M}(\mu)$ a.s. as $k \to \infty$. Denote $\overline{z_\epsilon} = 2\sqrt{B}/(1-2\epsilon)$; since $\alpha_k \to \epsilon$, from H$_{\mathscr{M}}$2-(i) there exists a.s. k_2 such that, for all $k > k_2$, $z_{n_k:k}(\mathbf{u}) < \overline{z_\epsilon}$ and, from (B.3), $a_{\alpha_k}(\mathbf{u}) < \overline{z_\epsilon}$.

Therefore, for large enough k, for any $\mathbf{u} \in \mathscr{U}(\mathbf{M}^*_{n_k,k})$,

$$\begin{aligned}\min_{\mathbf{M}' \in \mathbb{M}(\alpha_k)} \mathbf{u}^\top(\mathbf{M}' - \mathbf{M}^*_{n_k,k})\mathbf{u} &= \mathsf{E}\{h(X;\alpha_k,a_{\alpha_k}(\mathbf{u}),\mathbf{u})\} - \mathsf{E}_{\mu_k}\{h(X;\alpha_k,z_{n_k:k}(\mathbf{u}),\mathbf{u})\} \\ &\le \sup_{a\in[0,\overline{z_\epsilon}]} |\mathsf{E}\{h(X;\alpha_k,a,\mathbf{u})\} - \mathsf{E}_{\mu_k}\{h(X;\alpha_k,a,\mathbf{u})\}|\,.\end{aligned}$$

The functions $h(\cdot;\alpha,a,\mathbf{u})$ with $\alpha \in (\epsilon,2\epsilon)$, $a \in [0,\overline{z_\epsilon}]$ and $\mathbf{u} \in \mathbb{R}^p$, $\|\mathbf{u}\| = 1$, form a Glivenko-Cantelli class; see (van der Vaart, 1998, p. 271). This implies that there exists a.s. k_3 such that

$$\max_{\mathbf{u}\in\mathbb{R}^p:\|\mathbf{u}\|=1} \sup_{a\in[0,\overline{z_\epsilon}]} |\mathsf{E}\{h(X;\alpha_k,a,\mathbf{u})\} - \mathsf{E}_{\mu_k}\{h(X;\alpha_k,a,\mathbf{u})\}| < c_\epsilon/2\,,\quad \forall k > k_3\,.$$

Therefore, from (B.2), $\lambda_{\min}(\mathbf{M}^*_{n_k,k}) > c_\epsilon/2$ for all $k > k_3$, which concludes the first part of the proof.

We construct now an upper bound on $\limsup_{k\to\infty} \lambda_{\max}(\mathbf{M}_{n_k})$ following steps similar to the above developments but exploiting now the convexity of the criterion $\mathbf{M} \to 1/\Phi^+_{-\infty}(\mathbf{M}) = \lambda_{\max}(\mathbf{M})$. Its directional derivative is $F_{1/\Phi^+_{-\infty}}(\mathbf{M},\mathbf{M}') = \max_{\mathbf{u}\in\mathscr{U}(\mathbf{M})} \mathbf{u}^\top(\mathbf{M}'-\mathbf{M})\mathbf{u}$, with $\mathscr{U}(\mathbf{M}) = \{\mathbf{u} \in \mathbb{R}^p : \|\mathbf{u}\| = 1\,,\ \mathbf{M}\mathbf{u} = \lambda_{\max}(\mathbf{M})\mathbf{u}\}$. Denote by $\mathbf{M}^*_{n_k,k}$ the n_k-point design matrix that maximizes $1/\Phi^+_{-\infty}$ over the design space formed by k points X_i i.i.d. with μ. We can write $\mathbf{M}^*_{n_k,k} = (1/n_k)\sum_{i=1}^{n_k} \mathscr{M}(X_{k_i})$, where the k_i correspond to the indices of positive u_i in the maximization of $f(\mathbf{u}) = \lambda_{\max}[\sum_{i=1}^k u_i\,\mathscr{M}(X_i)]$ with respect to $\mathbf{u} = (u_1,\ldots,u_k)$ under the constraints $u_i \in \{0,1\}$ for all i and $\sum_i u_i = n_k$. Any matrix $\mathbf{M}_{n_k}$ obtained by selecting n_k distinct points X_i among $X_1,\ldots,X_k$ satisfies $\lambda_{\max}(\mathbf{M}_{n_k}) \le \lambda_{\max}(\mathbf{M}^*_{n_k,k})$.

For any $\mathbf{u} \in \mathscr{U}(\mathbf{M}^*_{n_k,k})$ we can write $\mathbf{u}^\top \mathbf{M}^*_{n_k,k}\mathbf{u} = \lambda_{\max}(\mathbf{M}^*_{n_k,k}) = \max_{\mathbf{v}\in\mathbb{R}^p:\,\|\mathbf{v}\|=1} \mathbf{v}^\top \mathbf{M}^*_{n_k,k}\mathbf{v} = (1/n_k)\sum_{i=1}^{n_k} z_{i:k}(\mathbf{u})$, where the $z_{i:k}(\mathbf{u})$ correspond to the values of $\mathbf{u}^\top \mathscr{M}(X_i)\mathbf{u}$ sorted by decreasing order for $i = 1,\ldots,k$. For any $m \in \{1,\ldots,n_k-1\}$, we thus have

$$\lambda_{\max}(\mathbf{M}^*_{n_k,k}) \le \frac{1}{m}\sum_{i=1}^m z_{i:k}(\mathbf{u}) \le \lambda_{\max}(\mathbf{M}^*_{m,k})\,,$$

showing that the worst case corresponds to the smallest admissible n_k, and we can restrict our attention to the case when $\alpha_k = n_k/k \to \epsilon$ as $k \to \infty$.

The convexity of $1/\Phi^+_{-\infty}$ implies that, for any $\mathbf{M}' \in \mathbb{M}^{\geq}$,

$$\lambda_{\max}(\mathbf{M}') \geq \lambda_{\max}(\mathbf{M}^*_{n_k,k}) + F_{1/\Phi^+_{-\infty}}(\mathbf{M}^*_{n_k,k}, \mathbf{M}')\,. \tag{B.4}$$

Take $\mathbf{M}' \in \mathbb{M}(\alpha_k)$, corresponding to some ξ_{α_k}. From H$_{\mathscr{M}}$-(i),

$$\lambda_{\max}(\mathbf{M}') = \lambda_{\max}\left[\int_{\mathscr{X}} \mathscr{M}(x)\,\xi_{\alpha_k}(\mathrm{d}x)\right] \leq \frac{1}{\alpha_k}\,\lambda_{\max}[\mathbf{M}(\mu)] < \frac{\sqrt{B}}{\alpha_k}\,.$$

Therefore, there exists some k_1 such that, for all $k > k_1$, $\lambda_{\max}(\mathbf{M}') < 2\sqrt{B}/\epsilon$, and (B.4) gives

$$\frac{2\sqrt{B}}{\epsilon} \geq \lambda_{\max}(\mathbf{M}^*_{n_k,k}) + \max_{\mathbf{u}\in\mathscr{U}(\mathbf{M}^*_{n_k,k})} \max_{\mathbf{M}'\in\mathbb{M}(\alpha_k)} \mathbf{u}^\top(\mathbf{M}' - \mathbf{M}^*_{n_k,k})\mathbf{u}\,.$$

For $a \in \mathbb{R}$, $\alpha \in (0,1)$ and $\mathbf{u} \in \mathbb{R}^p$, denote $h(x;\alpha,a,\mathbf{u}) = a + (1/\alpha)[\mathbf{u}^\top\mathscr{M}(x)\mathbf{u} - a]^+$, $x \in \mathscr{X}$. We have $\lambda_{\max}(\mathbf{M}^*_{n_k,k}) = (1/n_k)\sum_{i=1}^{n_k} z_{i:k}(\mathbf{u}) = \mathsf{E}_{\mu_k}\{h(X;\alpha_k, z_{n_k:k}(\mathbf{u}),\mathbf{u})\} = \inf_a \mathsf{E}_{\mu_k}\{h(X;\alpha_k,a,\mathbf{u})\}$, $\mathbf{u} \in \mathscr{U}(\mathbf{M}^*_{n_k,k})$, with $z_{n_k:k}(\mathbf{u})$ satisfying $0 \leq n_k\, z_{n_k:k}(\mathbf{u}) \leq \sum_{i=1}^{n_k} z_{i:k}(\mathbf{u}) < \sum_{i=1}^{k} z_{i:k}(\mathbf{u}) = k\,\lambda_{max}(\mathbf{M}_k)$. Also, for any $\alpha \in (0,1)$ and $\mathbf{u} \in \mathbb{R}^p$, $\mathbf{u} \neq \mathbf{0}$, $\max_{\mathbf{M}'\in\mathbb{M}(\alpha)} \mathbf{u}^\top\mathbf{M}'\mathbf{u} = \mathsf{E}\{h(X;\alpha,a_\alpha(\mathbf{u}),\mathbf{u})\} = \inf_a \mathsf{E}\{h(X;\alpha,a,\mathbf{u})\}$, where $a_\alpha(\mathbf{u})$ satisfies $\int_{\mathscr{X}} \mathbb{I}_{\{\mathbf{u}^\top\mathscr{M}(x)\mathbf{u}\geq a_\alpha(\mathbf{u})\}}\,\mu(\mathrm{d}x) = \alpha$, and H$_{\mathscr{M}}$-$(i)$ implies that $0 \leq a_\alpha(\mathbf{u}) \leq (1/\alpha)\int_{\mathscr{X}} \mathbb{I}_{\{\mathbf{u}^\top\mathscr{M}(x)\mathbf{u}\geq a_\alpha(\mathbf{u})\}}\,\mathbf{u}^\top\mathscr{M}(x)\mathbf{u}\,\mu(\mathrm{d}x) < \sqrt{B}/\alpha$. Since $\alpha_k = n_k/k \to \epsilon$ and $\mathbf{M}_k \to \mathbf{M}(\mu)$ a.s., there exists a.s. k_2 such that, for all $k > k_2$, $0 \leq a_{\alpha_k}(\mathbf{u}) < 2\sqrt{B}/\epsilon$ and $0 \leq z_{n_k:k}(\mathbf{u}) < 2\sqrt{B}/\epsilon$. This implies that, for $\mathbf{u} \in \mathscr{U}(\mathbf{M}^*_{n_k,k})$ and $k > k_2$,

$$\begin{aligned}\max_{\mathbf{M}'\in\mathbb{M}(\alpha_k)} \mathbf{u}^\top(\mathbf{M}' - \mathbf{M}^*_{n_k,k})\mathbf{u} &= \mathsf{E}\{h(X;\alpha_k,a_{\alpha_k}(\mathbf{u}),\mathbf{u})\} - \mathsf{E}_{\mu_k}\{h(X;\alpha_k,z_{n_k:k}(\mathbf{u}),\mathbf{u})\} \\ &\leq \sup_{a\in[0,2\sqrt{B}/\epsilon]} \left|\mathsf{E}\{h(X;\alpha_k,a,\mathbf{u})\} - \mathsf{E}_{\mu_k}\{h(X;\alpha_k,a,\mathbf{u})\}\right|\,.\end{aligned}$$

The rest of the proof is similar to the case above for $\lambda_{\min}$, using the fact that the functions $h(\cdot;\alpha,a,\mathbf{u})$ with $\alpha \in (\epsilon, 2\epsilon)$, $a \in [0, 2\sqrt{B}/\epsilon]$ and $\mathbf{u} \in \mathbb{R}^p$, $\|\mathbf{u}\| = 1$, form a Glivenko-Cantelli class. ■

C Convergence of $\widehat{C}_k$

We consider the convergence properties of (3.3) when the matrix $\mathbf{M}_k$ is fixed, that is,

$$\widehat{C}_{k+1} = \widehat{C}_k + \frac{\beta_k}{(k+1)^q}\,\left(\mathbb{I}_{\{Z_{k+1}\geq\widehat{C}_k\}} - \alpha\right)\,, \tag{C.1}$$

where the Z_k have a fixed distribution with uniformly bounded density f such that $f(C_{1-\alpha}) > 0$. We follow the arguments of Tierney (1983). The construction of β_k is like in Algorithm 1, with $\beta_k = \max\{\min(1/\widehat{f}_k, \beta_0\,k^\gamma)\}$ and $\widehat{f}_k$ following the recursion

$$\widehat{f}_{k+1} = \widehat{f}_k + \frac{1}{(k+1)^q}\,\left[\frac{1}{2\,h_{k+1}}\,\mathbb{I}_{\{|Z_{k+1}-\widehat{C}_k|\leq h_{k+1}\}} - \widehat{f}_k\right] \tag{C.2}$$

with $h_k = h/k^\gamma$.

Theorem C.1. *Let $\alpha \in (0,1)$, $\beta_0 > 0$, $h > 0$, $1/2 < q \leq 1$, $0 < \gamma < q - 1/2$. Let F be a distribution function such that $f(t) = \mathrm{d}F(t)/\mathrm{d}t$ exists for all t, is uniformly bounded, and is strictly positive in a neighborhood of $C_{1-\alpha}$, the unique value of C such that $F(C) = 1-\alpha$. Let (X_i) be an i.i.d. sequence distributed with F and define $\widehat{C}_k$ and $\widehat{f}_k$ by (C.1) and (C.2) respectively, with $\beta_k = \min\{1/\widehat{f}_k, \beta_0\, k^\gamma\}$ and $h_k = h/k^\gamma$. Then, $\widehat{C}_k \to C_{1-\alpha}$ a.s. when $k \to \infty$.*

Proof. We first show that $\widehat{f}_k$ is a.s. bounded. From the mean-value theorem, there exists a t_k in $[\widehat{C}_k - h_{k+1}, \widehat{C}_k + h_{k+1}]$ such that $\Pr\{|Z_{k+1} - \widehat{C}_k| \leq h_{k+1}\} = F(\widehat{C}_k + h_{k+1}) - F(\widehat{C}_k - h_{k+1}) = 2\, h_{k+1}\, f(t_k)$. Denote $\omega_{k+1} = \mathbb{I}_{\{|Z_{k+1} - \widehat{C}_k| \leq h_{k+1}\}} - 2\, h_{k+1}\, f(t_k)$. We can write

$$\widehat{f}_{k+1} = (1 - B_k)\, \widehat{f}_k + A_k + A'_k$$

where $B_k = 1/[(k+1)^q]$, $A_k = \omega_{k+1}/[2\, h_{k+1}\, (k+1)^q]$ and $A'_k = B_k\, f(t_k)$. Therefore,

$$\widehat{f}_{k+1} = \widehat{f}_1 \prod_{i=1}^{k}(1 - B_i) + \sum_{j=1}^{k}(A_j + A'_j) \prod_{i=j+1}^{k} (1 - B_i)\,.$$

We have $\prod_{i=1}^k (1 - B_i) < \exp(-\sum_{i=1}^k B_i) \to 0$ as $k \to \infty$ since $q \leq 1$. Next, for $h_k = h/k^\gamma$ and $0 < \gamma < q - 1/2$, $\sum_k 1/[h_k\, k^q]^2 < \infty$, $\sum_{j=1}^k A_j$ forms an $\mathscr{L}^2$-bounded martingale and therefore converges a.s. to some limit. Lemma 2 of Albert and Gardner (1967, p. 190) then implies that $\sum_{j=1}^k A_j \prod_{i=j+1}^k (1 - B_i) \to 0$ a.s. as $k \to \infty$. Consider now the term $T_k = \sum_{j=1}^k A'_j \prod_{i=j+1}^k (1 - B_i)$. Since f is bounded, $A'_j < \bar{f}\, B_j$ for some $\bar{f} < \infty$ and

$$T_k < \bar{f} \sum_{j=1}^{k} B_j \prod_{i=j+1}^{k} (1 - B_i) = \bar{f} \left[1 - \prod_{i=1}^{k}(1 - B_i)\right] < \bar{f}\,,$$

where the equality follows from Albert and Gardner (1967, Lemma 1, p. 189). This shows that $\widehat{f}_k$ is a.s. bounded. Therefore, $\beta_k = \min\{1/\widehat{f}_k, \beta_0\, k^\gamma\}$ is a.s. bounded away from zero.

We consider now the convergence of (C.1). Following Tierney (1983), define

$$D_k = \frac{\beta_k}{(k+1)^q} \left\{\mathbb{I}_{\{Z_{k+1} \geq \widehat{C}_k\}} - [1 - F(\widehat{C}_k)]\right\} \text{ and } E_k = \frac{\beta_k}{(k+1)^q} \frac{F(\widehat{C}_k) - (1-\alpha)}{\widehat{C}_k - C_{1-\alpha}}\,.$$

Denote by $\mathscr{F}_k$ the increasing sequence of σ-fields generated by the X_i; we have $\mathsf{E}\{D_k|\mathscr{F}_k\} = 0$ and $\mathsf{E}\{D_k^2|\mathscr{F}_k\} = \beta_k^2\, F(\widehat{C}_k)\,[1 - F(\widehat{C}_k)]/(k+1)^{2q}$. We can rewrite (C.1) as $\widehat{C}_{k+1} - C_{1-\alpha} = (\widehat{C}_k - C_{1-\alpha})\,(1 - E_k) + D_k$, which gives

$$\mathsf{E}\{(\widehat{C}_{k+1} - C_{1-\alpha})^2|\mathscr{F}_k\} = (\widehat{C}_k - C_{1-\alpha})^2\,(1 - E_k)^2 + \frac{\beta_k^2}{(k+1)^{2q}}\, F(\widehat{C}_k)\,[1 - F(\widehat{C}_k)]\,.$$

$E_k \geq 0$ for all k, $[F(\widehat{C}_k) - (1-\alpha)]/(\widehat{C}_k - C_{1-\alpha})$ is bounded since f is bounded, and therefore $E_k \to 0$. Since $\beta_k \leq \beta_0\, k^\gamma$ and $0 < \gamma < q - 1/2$, $\sum_k \beta_k^2/(k+1)^{2q} < \infty$. Robbins-Siegmund Theorem (1971) then implies that $\widehat{C}_k$ converges a.s. to some limit and that $\sum_k (\widehat{C}_k - C_{1-\alpha})^2\,[1 - (1 - E_k)^2] < \infty$ a.s.; since $E_k \to 0$, we obtain $\sum_k (\widehat{C}_k - C_{1-\alpha})^2\, E_k < \infty$ a.s. Since $q \leq 1$, β_k is a.s. bounded away from zero, and f is strictly positive in a neighborhood of $C_{1-\alpha}$, we obtain that $\sum_k E_k = \infty$, implying that $\widehat{C}_k \to C_{1-\alpha}$ a.s., which concludes the proof. ∎

Tierney (1983) uses $q = 1$; the continuity of f at $C_{1-\alpha}$ then implies that $f_k \to f(C_{1-\alpha})$ a.s., and his construction also achieves the optimal rate of convergence of $\widehat{C}_k$ to $C_{1-\alpha}$, with $\sqrt{k}(\widehat{C}_k - C_{1-\alpha}) \xrightarrow{\mathrm{d}} \mathscr{N}(0, \alpha(1-\alpha)/f^2(C_{1-\alpha}))$ as $k \to \infty$.

D Lipschitz continuity of $C_{1-\alpha}(\mathbf{M})$

Lemma D.1. *Under H_Φ and $H_{\mu,\mathscr{M}}$, the $(1-\alpha)$-quantile $C_{1-\alpha}(\mathbf{M})$ of the distribution $F_{\mathbf{M}}$ of $Z_{\mathbf{M}}(X) = F_\Phi[\mathbf{M}, \mathscr{M}(X)]$ is a Lipschitz continuous function of $\mathbf{M} \in \mathbb{M}^{\geq}_{\ell,L}$.*

Proof. For any $\mathbf{A} \in \mathbb{M}^{>}$, define the random variable $T_{\mathbf{A}}(X) = \mathrm{trace}[\mathbf{A}.\mathscr{M}(X)]$ and denote $G_{\mathbf{A}}$ its distribution function and $Q_{1-\alpha}(\mathbf{A})$ the associated $(1-\alpha)$-quantile. We have $Z_{\mathbf{M}}(X) = T_{\nabla_\Phi(\mathbf{M})}(X) - \mathrm{trace}[\nabla_\Phi(\mathbf{M})\mathbf{M}]$, and therefore

$$C_{1-\alpha}(\mathbf{M}) = Q_{1-\alpha}[\nabla_\Phi(\mathbf{M})] - \mathrm{trace}[\nabla_\Phi(\mathbf{M})\mathbf{M}]\,. \tag{D.1}$$

We fist show that $\mathrm{trace}[\nabla_\Phi(\mathbf{M})\mathbf{M}]$ is Lipschitz continuous in $\mathbf{M}$. Indeed, for any $\mathbf{M}$, $\mathbf{M}'$ in $\mathbb{M}^{\geq}_{\ell,L}$, we have

$$\begin{aligned}\big|\mathrm{trace}[\nabla_\Phi(\mathbf{M}')\mathbf{M}'] - \mathrm{trace}[\nabla_\Phi(\mathbf{M})\mathbf{M}]\big| &\leq \|\mathbf{M}'\|\,\|\nabla_\Phi(\mathbf{M}') - \nabla_\Phi(\mathbf{M})\| + \|\nabla_\Phi(\mathbf{M})\|\,\|\mathbf{M}' - \mathbf{M}\| \\ &< [L\sqrt{p}\,K_\ell + A(\ell)]\,\|\mathbf{M}' - \mathbf{M}\|\,, \end{aligned}\tag{D.2}$$

where we used H_Φ and the fact that $\mathbf{M}, \mathbf{M}' \in \mathbb{M}^{\geq}_{\ell,L}$.

Consider now $G_{\mathbf{A}}$ and $G_{\mathbf{A}'}$ for two matrices $\mathbf{A}$ and $\mathbf{A}'$ in $\mathbb{M}^{>}$. We have

$$G_{\mathbf{A}'}(t) - G_{\mathbf{A}}(t) = \int_{\mathscr{X}} \left(\mathbb{I}_{\{\mathrm{trace}[\mathbf{A}'.\mathscr{M}(x)] \leq t\}} - \mathbb{I}_{\{\mathrm{trace}[\mathbf{A}.\mathscr{M}(x)] \leq t\}}\right) \mu(\mathrm{d}x)\,,$$

and therefore

$$\begin{aligned}|G_{\mathbf{A}'}(t) - G_{\mathbf{A}}(t)| &\leq \mathsf{Prob}\left\{\min\{\mathrm{trace}[\mathbf{A}'\mathscr{M}(X)], \mathrm{trace}[\mathbf{A}.\mathscr{M}(X)]\} \leq t \right. \\ &\qquad \left. \leq \max\{\mathrm{trace}[\mathbf{A}'\mathscr{M}(X)], \mathrm{trace}[\mathbf{A}.\mathscr{M}(X)]\}\right\} \\ &\leq \mathsf{Prob}\left\{\mathrm{trace}[(\mathbf{A} - \|\mathbf{A} - \mathbf{A}'\|\mathbf{I}_p)\mathscr{M}(X)] \leq t \leq \mathrm{trace}[(\mathbf{A} + \|\mathbf{A} - \mathbf{A}'\|\mathbf{I}_p)\mathscr{M}(X)]\right\}\,,\end{aligned}$$

with $\mathbf{I}_p$ the $p \times p$ identity matrix. Since $\mathbf{A} - \lambda_{\min}(\mathbf{A})\,\mathbf{I}_b \in \mathbb{M}^{\geq}$, denoting $b_1 = 1 - \|\mathbf{A} - \mathbf{A}'\|/\lambda_{\min}(\mathbf{A})$ and $b_2 = 1 + \|\mathbf{A} - \mathbf{A}'\|/\lambda_{\min}(\mathbf{A})$, we obtain

$$\begin{aligned}|G_{\mathbf{A}'}(t) - G_{\mathbf{A}}(t)| &\leq \mathsf{Prob}\left\{b_1\ \mathrm{trace}[\mathbf{A}.\mathscr{M}(X)] \leq t \leq b_2\ \mathrm{trace}[\mathbf{A}.\mathscr{M}(X)]\right\} \\ &= \mathsf{Prob}\left\{\mathrm{trace}[\mathbf{A}.\mathscr{M}(X)] \leq \frac{t}{b_1} \bigwedge \mathrm{trace}[\mathbf{A}.\mathscr{M}(X)] \geq \frac{t}{b_2}\right\} \\ &= G_{\mathbf{A}}(t/b_1) - G_{\mathbf{A}}(t/b_2)\,.\end{aligned}\tag{D.3}$$

In the rest of the proof we show that $Q_{1-\alpha}[\nabla_\Phi(\mathbf{M})]$ is Lipschitz continuous in $\mathbf{M}$. Take two matrices $\mathbf{M}, \mathbf{M}' \in \mathbb{M}^{\geq}_{\ell,L}$, and consider the associated $(1-\alpha)$-quantiles $Q_{1-\alpha}[\nabla_\Phi(\mathbf{M})]$ and $Q_{1-\alpha}[\nabla_\Phi(\mathbf{M}')]$, which we shall respectively denote $Q_{1-\alpha}$ and $Q'_{1-\alpha}$ to simplify notation. From $\mathrm{H}_{\mu,\mathscr{M}}$, the p.d.f. $\psi_{\mathbf{M}}$ associated with $G_{\nabla_\Phi(\mathbf{M})}$ is continuous at $Q_{1-\alpha}$ and satisfies $\psi_{\mathbf{M}}(Q_{1-\alpha}) > \epsilon_{\ell,L}$. From the identities

$$\int_{-\infty}^{Q_{1-\alpha}} \psi_{\mathbf{M}}(z)\,\mathrm{d}z = \int_{-\infty}^{Q'_{1-\alpha}} \psi_{\mathbf{M}'}(z)\,\mathrm{d}z = 1 - \alpha\,,$$

we deduce

$$\left|\int_{Q_{1-\alpha}}^{Q'_{1-\alpha}} \psi_{\mathbf{M}}(z)\,\mathrm{d}z\right| = \left|\int_{-\infty}^{Q'_{1-\alpha}} [\psi_{\mathbf{M}'}(z) - \psi_{\mathbf{M}}(z)]\,\mathrm{d}z\right| = \left|G_{\nabla_\Phi(\mathbf{M}')}(Q'_{1-\alpha}) - G_{\nabla_\Phi(\mathbf{M})}(Q'_{1-\alpha})\right|\,. \tag{D.4}$$

From H_Φ, when substituting $\nabla_\Phi(\mathbf{M})$ for $\mathbf{A}$ and $\nabla_\Phi(\mathbf{M}')$ for $\mathbf{A}'$ in b_1 and b_2, we get $b_1 > B_1 = 1 - K_\ell\|\mathbf{M}'-\mathbf{M}\|/a(L)$ and $b_2 < B_2 = 1 + K_\ell\|\mathbf{M}'-\mathbf{M}\|/a(L)$, showing that $Q'_{1-\alpha} \to Q_{1-\alpha}$ as $\|\mathbf{M}'-\mathbf{M}\| \to 0$. Therefore, there exists some β_1 such that, for $\|\mathbf{M}'-\mathbf{M}\| < \beta_1$ we have

$$\left| \int_{Q_{1-\alpha}}^{Q'_{1-\alpha}} \psi_{\mathbf{M}}(z)\,\mathrm{d}z \right| > \frac{1}{2}\,\left|Q'_{1-\alpha} - Q_{1-\alpha}\right|\,\epsilon_{\ell,L}\,. \tag{D.5}$$

Using (D.3), we also obtain for $\|\mathbf{M}'-\mathbf{M}\|$ smaller than some β_2

$$\begin{aligned}
\left|G_{\nabla_\Phi(\mathbf{M}')}(Q'_{1-\alpha}) - G_{\nabla_\Phi(\mathbf{M})}(Q'_{1-\alpha})\right| &\leq G_{\nabla_\Phi(\mathbf{M})}(Q'_{1-\alpha}/B_1) - G_{\nabla_\Phi(\mathbf{M})}(Q'_{1-\alpha}/B_2) \\
&< 2\psi_{\mathbf{M}}(Q'_{1-\alpha})\left(\frac{Q'_{1-\alpha}}{B_1} - \frac{Q'_{1-\alpha}}{B_2}\right) \\
&< 4\|\mathbf{M}'-\mathbf{M}\|\,\psi_{\mathbf{M}}(Q'_{1-\alpha})\,Q'_{1-\alpha}\,\frac{a(L)}{K_\ell\left(a^2(L)/K_\ell^2 - \|\mathbf{M}'-\mathbf{M}\|^2\right)}\,.
\end{aligned}$$

Therefore, when $\|\mathbf{M}'-\mathbf{M}\| < a(L)/(K_\ell\sqrt{2})$,

$$\left|G_{\nabla_\Phi(\mathbf{M}')}(Q'_{1-\alpha}) - G_{\nabla_\Phi(\mathbf{M})}(Q'_{1-\alpha})\right| < \kappa\,\|\mathbf{M}-\mathbf{M}'\|$$

with $\kappa = 8\bar{\varphi}_{\mathbf{M}}\,Q'_{1-\alpha}\,K_\ell/a(L)$, where $\bar{\varphi}_{\mathbf{M}}$ is the upper bound on $\varphi_{\mathbf{M}}$ in $\mathrm{H}_{\mu,\mathscr{M}}$. Using (D.4) and (D.5) we thus obtain, for $\|\mathbf{M}'-\mathbf{M}\|$ small enough,

$$\left|Q_{1-\alpha}[\nabla_\Phi(\mathbf{M}')] - Q_{1-\alpha}[\nabla_\Phi(\mathbf{M})]\right| < 2\kappa/\epsilon_{\ell,L}\,\|\mathbf{M}-\mathbf{M}'\|\,,$$

which, combined with (D.2) and (D.1), completes the proof. ■

Acknowledgements

The work of the first author was partly supported by project INDEX (INcremental Design of EXperiments) ANR-18-CE91-0007 of the French National Research Agency (ANR).

References

Albert, A. and Gardner, L. (1967). *Stochastic Approximation and Nonlinear Regression*. MIT Press, Cambridge, MA.

Albright, S. and Derman, C. (1972). Asymptotically optimal policies for the stochastic sequential assignment problem. *Management Science*, 19(1):46–51.

Borkar, V. (1997). Stochastic approximation with two time scales. *Systems & Control Letters*, 29:291–294.

Cook, R. and Nachtsheim, C. (1980). A comparison of algorithms for constructing exact d-optimal designs. *Technometrics*, 22(3):315–324.

Dalal, G., Szörényi, B., Thoppe, G., and Mannor, S. (2018). Finite sample analysis of two-timescale stochastic approximation with applications to reinforcement learning. *Proc. of Machine Learning Research*, 75:1–35.

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

Fedorov, V. (1989). Optimal design with bounded density: optimization algorithms of the exchange type. *Journal Statist. Planning and Inference*, 22:1–13.

Fedorov, V. and Hackl, P. (1997). *Model-Oriented Design of Experiments*. Springer, Berlin.

Harman, R. (2004). Lower bounds on efficiency ratios based on ϕ_p-optimal designs. In Di Bucchianico, A., Läuter, H., and Wynn, H., editors, *mODa'7 – Advances in Model–Oriented Design and Analysis, Proceedings of the 7th Int. Workshop, Heeze (Netherlands)*, pages 89–96, Heidelberg. Physica Verlag.

Kesten, H. (1958). Accelerated stochastic approximation. *The Annals of Mathematical Statistics*, 29(1):41–59.

Konda, V. and Tsitsiklis, J. (2004). Convergence rate of linear two-time-scale stochastic approximation. *The Annals of Applied Probability*, 14(2):796–819.

Lakshminarayanan, C. and Bhatnagar, S. (2017). A stability criterion for two timescale stochatic approximation. *Automatica*, 79:108–114.

Ljung, L. (1977). Analysis of recursive stochastic algorithms. *IEEE Transactions on Automatic Control*, 22(4):551–575.

Ljung, L. and Söderström, T. (1983). *Theory and Practice of Recursive Identification*. MIT Press, Cambridge, MA.

Metelkina, A. and Pronzato, L. (2017). Information-regret compromise in covariate-adaptive treatment allocation. *The Annals of Statistics*, 45(5):2046–2073.

Mitchell, T. (1974). An algorithm for the construction of "D-optimal" experimental designs. *Technometrics*, 16:203–210.

Nie, R., Wiens, D., and Zhai, Z. (2018). Minimax robust active learning for approximately specified regression models. *Canadian Journal of Statistics*, 46(1):104–122.

Pflug, G. (2000). Some remarks on the value-at-risk and the conditional value-at-risk. In Uryasev, S., editor, *Probabilistic Constrained Optimization*, pages 272–281. Springer.

Pilz, J. (1983). *Bayesian Estimation and Experimental Design in Linear Regression Models*, volume 55. Teubner-Texte zur Mathematik, Leipzig. (also Wiley, New York, 1991).

Pronzato, L. (2001). Optimal and asymptotically optimal decision rules for sequential screening and resource allocation. *IEEE Transactions on Automatic Control*, 46(5):687–697.

Pronzato, L. (2004). A minimax equivalence theorem for optimum bounded design measures. *Statistics & Probability Letters*, 68:325–331.

Pronzato, L. (2006). On the sequential construction of optimum bounded designs. *Journal of Statistical Planning and Inference*, 136:2783–2804.

Pronzato, L. and Pázman, A. (2013). *Design of Experiments in Nonlinear Models. Asymptotic Normality, Optimality Criteria and Small-Sample Properties*. Springer, LNS 212, New York.

Pukelsheim, F. (1993). *Optimal Experimental Design*. Wiley, New York.

Robbins, H. and Siegmund, D. (1971). A convergence theorem for non negative almost supermartingales and some applications. In Rustagi, J., editor, *Optimization Methods in Statistics*, pages 233–257. Academic Press, New York.

Rockafellar, R. and Uryasev, S. (2000). Optimization of conditional value-at-risk. *Journal of Risk*, 2:21–42.

Rockafellar, R. and Uryasev, S. (2002). Conditional value-at-risk for general loss distributions. *Journal of Banking & Finance*, 26:1443–1471.

Sahm, M. and Schwabe, R. (2001). A note on optimal bounded designs. In Atkinson, A., Bogacka, B., and Zhigljavsky, A., editors, *Optimum Design 2000*, chapter 13, pages 131–140. Kluwer, Dordrecht.

Tierney, L. (1983). A space-efficient recursive procedure for estimating a quantile of an unknown distribution. *SIAM Journal on Scientific and Statistical Computing*, 4(4):706–711.

Tsypkin, Y. (1983). Optimality in identification of linear plants. *International Journal of Systems Science*, 14(1):59–74.

van der Vaart, A. (1998). *Asymptotic Statistics*. Cambridge University Press, Cambridge.

Wang, H., Yang, M., and Stufken, J. (2019). Information-based optimal subdata selection for big data linear regression. *Journal of the American Statistical Association*, 114(525):393–405.

Wiens, D. (2005). Robustness in spatial studies II: minimax design. *Environmetrics*, 16(2):205–217.

Wynn, H. (1982). Optimum submeasures with applications to finite population sampling. In Gupta, S. and Berger, J., editors, *Statistical Decision Theory and Related Topics III. Proc. 3rd Purdue Symp., vol. 2*, pages 485–495. Academic Press, New York.