Metrika (2021) 84:851–874
https://doi.org/10.1007/s00184-020-00803-0

Convergence of least squares estimators in the adaptive Wynn algorithm for some classes of nonlinear regression models

Fritjof Freise[1] · Norbert Gaffke[2] · Rainer Schwabe[2]

Received: 9 March 2020 / Accepted: 23 December 2020 / Published online: 8 February 2021

Abstract
The paper continues the authors' work (Freise et al. The adaptive Wynn-algorithm in generalized linear models with univariate response. arXiv:1907.02708, 2019) on the adaptive Wynn algorithm in a nonlinear regression model. In the present paper the asymptotics of adaptive least squares estimators under the adaptive Wynn algorithm is studied. Strong consistency and asymptotic normality are derived for two classes of nonlinear models: firstly, for the class of models satisfying a condition of 'saturated identifiability', which was introduced by Pronzato (Metrika 71:219–238, 2010); secondly, a class of generalized linear models. Further essential assumptions are compactness of the experimental region and of the parameter space together with some natural continuity assumptions. For asymptotic normality some further smoothness assumptions and asymptotic homoscedasticity of random errors are needed and the true parameter point is required to be an interior point of the parameter space.

Keywords Approximate design · D-optimality · Adaptive estimation · Strong consistency · Asymptotic normality · Generalized linear model

Mathematics Subject Classification 62L05 · 62F12 · 62J02

✉ Norbert Gaffke
norbert.gaffke@ovgu.de

Fritjof Freise
fritjof.freise@tiho-hannover.de

Rainer Schwabe
rainer.schwabe@ovgu.de

[1] Department of Biometry, Epidemiology and Information Processing, University of Veterinary Medicine Hannover, 30559 Hannover, Germany

[2] Faculty of Mathematics, University of Magdeburg, 39016 Magdeburg, Germany

1 Introduction

The classical algorithm of Wynn (1970) for D-optimal design in linear regression models has motivated a particular scheme for sequential adaptive design in nonlinear regression models, see Freise (2016), Pronzato (2010) and Freise et al. (2019). We refer to this scheme as 'the adaptive Wynn algorithm'. In a previous paper (Freise et al. 2019) of the authors the asymptotics of the sequences of designs and maximum likelihood estimators under the adaptive Wynn algorithm was studied, for the important class of generalized linear models with univariate response. In the present paper the asymptotics of least squares estimators (LSEs) under the adaptive Wynn algorithm is studied, firstly, for the class of nonlinear models satisfying a condition of 'saturated identifiability' and, secondly, for a class of generalized linear models. As a main result, strong consistency of the adaptive LSEs is shown and, as a consequence, the asymptotic D-optimality of the generated design sequence is obtained (almost surely). Here 'D-optimality' means local D-optimality at the true parameter point. Moreover, asymptotic normality of the adaptive LSEs is obtained, where the asymptotic covariance matrix is given by the inverse of the locally D-optimal information matrix at the true parameter point. This shows in particular, that compared to the nonadaptive LSEs under any fixed design the adaptive LSEs from the adaptive Wynn algorithm are asymptotically more efficient in the sense of a smaller determinant of the asymptotic covariance matrix, unless of course, a fixed design is used which is locally D-optimal at the true parameter point. However, the true parameter point is unknown, thus preventing the use of that design. In contrast to the classical concept of a fixed design, the sequential adaptive method provided by the adaptive Wynn algorithm is not affected by the unknown parameter point: asymptotically the true parameter point can be identified and the adaptive designs become D-optimal. Note that other, more practically motivated adaptive procedures for design and estimation restrict to a finite number of adaptation stages, e.g. only two stages as in Dette et al. (2013) and Lane et al. (2014). Those papers addressed problems on asymptotic efficiency of adaptive maximum likelihood estimators for particular two-stage adaptive procedures, when the sample sizes of the stages go to infinity. However, the present paper is exclusively concerned with the adaptive Wynn algorithm which is an infinite-stage adaptive procedure.

Next we give an outline of our framework and the adaptive Wynn algorithm. Suppose a nonlinear regression model with a real valued mean response $\mu(x, \theta)$, $x \in \mathcal{X}$, $\theta \in \Theta$, where $\mathcal{X}$ and Θ are the experimental region and the parameter space, respectively. Suppose that a family of $\mathbb{R}^p$-valued functions f_θ, $\theta \in \Theta$, defined on $\mathcal{X}$ has been identified such that the $p \times p$ matrix $f_\theta(x)\, f_\theta^T(x)$ is the elementary information matrix of $x \in \mathcal{X}$ at $\theta \in \Theta$. Note that a vector $a \in \mathbb{R}^p$ is written as a column vector and a^T denotes its transposed which is thus a p-dimensional row vector. A design ξ is a probability measure on $\mathcal{X}$ with finite support. That is, ξ is described by its support, denoted by $\mathrm{supp}(\xi)$, which is a nonempty finite subset of $\mathcal{X}$, and by its weights $\xi(x)$ for $x \in \mathrm{supp}(\xi)$ which are positive real numbers with $\sum_{x \in \mathrm{supp}(\xi)} \xi(x) = 1$. The information matrix of a design ξ at $\theta \in \Theta$ is defined by

$$M(\xi, \theta) = \sum_{x \in \mathrm{supp}(\xi)} \xi(x)\, f_\theta(x)\, f_\theta^T(x), \tag{1.1}$$

which is a nonnegative definite $p \times p$ matrix.

In applications the family $f_\theta, \theta \in \Theta$, will be related to the mean response $\mu(x, \theta)$, $x \in \mathcal{X}$, $\theta \in \Theta$. Usually, $\Theta \subseteq \mathbb{R}^p$ and $f_\theta(x)$ is given by the gradient of $\mu(x, \theta)$ w.r.t. θ for each fixed x, or by a scalar multiple of that gradient where the scalar factor is a function of θ and x, cf. Atkinson et al. (2014), Lemma 1. For particular classes of models described below, consistency of the adaptive LSEs will be achieved without any relation between the family $f_\theta, \theta \in \Theta$, and the mean response $\mu(x, \theta)$, $x \in \mathcal{X}$, $\theta \in \Theta$, whereas asymptotic normality of the adaptive LSEs will require the gradient relation with scalar factor equal to 1. Throughout we assume the following basic conditions (B1)–(B4).

(B1) The experimental region $\mathcal{X}$ is a compact metric space.

(B2) The parameter space Θ is a compact metric space.

(B3) The real-valued mean response function $(x, \theta) \mapsto \mu(x, \theta)$, defined on the Cartesian product space $\mathcal{X} \times \Theta$, is continuous.

(B4) The family $f_\theta, \theta \in \Theta$, of $\mathbb{R}^p$-valued functions on $\mathcal{X}$ satisfies:

(i) for each $\theta \in \Theta$ the image $f_\theta(\mathcal{X})$ spans $\mathbb{R}^p$;

(ii) the function $(x, \theta) \mapsto f_\theta(x)$ is continuous on $\mathcal{X} \times \Theta$.

By $\mathbb{N}$ and $\mathbb{N}_0$ we denote the set of all positive integers and all nonnegative integers, respectively. By δ_x for any $x \in \mathcal{X}$ we denote the one-point probability distribution on $\mathcal{X}$ concentrated at the point x. The adaptive Wynn algorithm collects iteratively design points $x_i \in \mathcal{X}$, $i \in \mathbb{N}$, while adaptively estimating θ on the basis of the current design points and observed (real valued) responses at those points. In greater detail the algorithm reads as follows.

Adaptive Wynn algorithm

(o) *Initialization*

A positive integer $n_{\mathrm{st}} \in \mathbb{N}$ and design points $x_1, \ldots, x_{n_{\mathrm{st}}} \in \mathcal{X}$ are chosen such that the starting design $\xi_{n_{\mathrm{st}}} = \frac{1}{n_{\mathrm{st}}} \sum_{i=1}^{n_{\mathrm{st}}} \delta_{x_i}$ has positive definite information matrices, i.e., for all $\theta \in \Theta$ the information matrix $M(\xi_{n_{\mathrm{st}}}, \theta)$ is positive definite. Observed responses $y_1, \ldots, y_{n_{\mathrm{st}}} \in \mathbb{R}$ at the design points $x_1, \ldots, x_{n_{\mathrm{st}}}$ are taken, and an initial parameter estimate $\theta_{n_{\mathrm{st}}} \in \Theta$ is calculated,

$$\theta_{n_{\mathrm{st}}} = \widehat{\theta}_{n_{\mathrm{st}}}(x_1, y_1, \ldots, x_{n_{\mathrm{st}}}, y_{n_{\mathrm{st}}}) \in \Theta.$$

(i) *Iteration*

At stage $n \geq n_{\mathrm{st}}$ the current data is given by the points $x_1, \ldots, x_n \in \mathcal{X}$ which form the design $\xi_n = \frac{1}{n} \sum_{i=1}^{n} \delta_{x_i}$, and by the observed responses $y_1, \ldots, y_n \in \mathbb{R}$ at $x_1, \ldots, x_n$, respectively, along with a parameter estimate $\theta_n \in \Theta$,

$$\theta_n = \widehat{\theta}_n(x_1, y_1, \ldots, x_n, y_n) \in \Theta. \tag{1.2}$$

The iteration rule is given by

$$x_{n+1} = \arg \max_{x \in \mathcal{X}} f_{\theta_n}^T(x)\, M^{-1}(\xi_n, \theta_n)\, f_{\theta_n}(x). \tag{1.3}$$

An observation $y_{n+1} \in \mathbb{R}$ of the response at x_{n+1} is taken and a new parameter estimate θ_{n+1} based on the augmented data is computed,

$$\theta_{n+1} = \widehat{\theta}_{n+1}(x_1, y_1, \ldots, x_n, y_n, x_{n+1}, y_{n+1}) \in \Theta.$$

Replace n by $n+1$ and repeat the iteration step (i). □

Of course, Eq. (1.3) requires the information matrix $M(\xi_n, \theta_n)$ to be positive definite at each stage $n \geq n_{\mathrm{st}}$. In fact, this is ensured by the choice of the initial design $\xi_{n_{\mathrm{st}}}$ since, obviously, the sequence of designs ξ_n, $n \geq n_{\mathrm{st}}$ satisfies

$$\xi_{n+1} = \frac{n}{n+1}\,\xi_n + \frac{1}{n+1}\,\delta_{x_{n+1}}, \tag{1.4}$$

$$M(\xi_{n+1}, \theta) = \frac{n}{n+1}\,M(\xi_n, \theta) + \frac{1}{n+1}\,f_\theta(x_{n+1})\,f_\theta^T(x_{n+1}), \quad \theta \in \Theta, \tag{1.5}$$

from which one concludes by induction that $M(\xi_n, \theta)$ is positive definite for all $n \geq n_{\mathrm{st}}$ and all $\theta \in \Theta$. The existence of an initial design $\xi_{n_{\mathrm{st}}}$ as required will be shown in Section 2, Lemma 1. However, we have no general method or algorithm for constructing an initial design according to 'step (o)' of the algorithm. For some important classes of models there exists a saturated initial design (i.e., $n_{\mathrm{st}} = p$) which can easily be constructed, see Remark 1 in Sect. 2.

The algorithm uses, in particular, an observed response y_i at each current design point x_i. So the generated sequence of design points, x_i, $i \in \mathbb{N}$, and the corresponding sequence of designs ξ_n, $n \geq n_{\mathrm{st}}$, are random sequences with a particular dependence structure caused by Eqs. (1.2) and (1.3). An appropriate stochastic model will be stated in Sect. 3 which was used in Freise et al. (2019) and goes back to Lai and Wei (1982), Lai (1994), and Chen et al. (1999). In particular, the generated sequence x_i, $i \in \mathbb{N}$, and the observed responses y_i, are viewed as values of random variables X_i, $i \in \mathbb{N}$, and Y_i, $i \in \mathbb{N}$, respectively, following a stochastic model which we call an 'adaptive regression model'. Our formulation of the adaptive Wynn algorithm is a description of the paths of the stochastic process (X_i, Y_i), $i \in \mathbb{N}$.

The estimators $\widehat{\theta}_n$, $n \geq n_{\mathrm{st}}$, employed by the algorithm to produce the estimates θ_n, $n \geq n_{\mathrm{st}}$, in (1.2), may be any estimators of θ such that their values are in Θ and $\widehat{\theta}_n$ is a function of the data $x_1, y_1, \ldots, x_n, y_n$ available at stage n. Such estimators will be called adaptive estimators. Later, strong consistency of $\widehat{\theta}_n$, $n \geq n_{\mathrm{st}}$, will be required. In Sect. 3 we focus on adaptive LSEs $\widehat{\theta}_n^{(\mathrm{LS})}$, i.e.,

$$\widehat{\theta}_n^{(\mathrm{LS})}(x_1, y_1, \ldots, x_n, y_n) = \arg\min_{\theta \in \Theta} \sum_{i=1}^{n} \big\{ y_i - \mu(x_i, \theta) \big\}^2.$$

Note that when dealing with the adaptive LSEs we will not necessarily assume that the estimators $\widehat{\theta}_n$, $n \geq n_{\mathrm{st}}$, employed by the algorithm are given by the LSEs. Below two alternative conditions on the nonlinear model will be introduced. Either condition will ensure strong consistency of the LSEs, irrespective which adaptive estimators $\widehat{\theta}_n$ are used in the algorithm. One condition is that of 'saturated identifiability' (SI), which

was introduced and employed by Pronzato (2010) in the case of a finite experimental region $\mathcal{X}$.

(SI) If $z_1, \ldots, z_p \in \mathcal{X}$ are pairwise distinct design points then the $\mathbb{R}^p$-valued function on Θ given by $\theta \mapsto \big(\mu(z_1,\theta), \ldots, \mu(z_p,\theta)\big)^T$ is an injection, i.e., if $\theta, \theta' \in \Theta$ and $\mu(z_j,\theta) = \mu(z_j,\theta')$, $1 \le j \le p$, then $\theta = \theta'$.

Recall that p is the dimension of the functions f_θ. As mentioned above, in many applications one has $\Theta \subseteq \mathbb{R}^p$ and $f_\theta(x)$ is given by the gradient of $\mu(x,\theta)$ w.r.t. θ for all $x \in \mathcal{X}$. So, in these cases, p also coincides with the dimension of the parameter vector θ.

The other (alternative) condition states that the model is essentially a generalized linear model. We call this condition (GLM*) where the 'star' is to distinguish it from a similar 'condition (GLM)' employed in Freise et al. (2019) which addressed only to the family f_θ ignoring the mean response function μ.

(GLM*) $\Theta \subseteq \mathbb{R}^p$, $\mu(x,\theta) = G\big(f^T(x)\,\theta\big)$, and $f_\theta(x) = \psi(x,\theta)\, f(x)$ for all $(x,\theta) \in \mathcal{X}\times\Theta$, where $f : \mathcal{X} \longrightarrow \mathbb{R}^p$ and $\psi : \mathcal{X}\times\Theta \longrightarrow (0,\infty)$ are continuous functions, $G : I \longrightarrow \mathbb{R}$ is a differentiable function on an open interval $I \subseteq \mathbb{R}$ with continuous and positive derivative, $G'(u) > 0$ for all $u \in I$, and I covers the values $f^T(x)\,\theta$ for all $(x,\theta) \in \mathcal{X}\times\Theta$.

Note that G is called the inverse link function. If the functions f_θ are assumed to be the gradients (w. r. t. θ) of μ then the form of μ assumed in (GLM*) already implies $f_\theta(x) = \psi(x,\theta)\, f(x)$ with $\psi(x,\theta) = G'\big(f^T(x)\,\theta\big)$ for all $(x,\theta) \in \mathcal{X}\times\Theta$.

Example 1 (cf. Pronzato 2010, Examples 2 and 3; Hu (1998), Examples 2 and 3) Let $p = 2$, $\Theta \subseteq (0,\infty)^2$, $\mathcal{X} \subseteq (0,\infty)$, and consider two regression models,

$$\text{(a)}\ \ \mu(x,\theta) = \frac{\theta_1 x}{\theta_2 + x}, \quad \text{(b)}\ \ \mu(x,\theta) = \theta_1 \exp\big(-\theta_2 x\big),$$

where $\theta = (\theta_1,\theta_2)^T$. Models (a) and (b) are called the 'Michaelis-Menten model' and the 'exponential decay model', respectively. Both models (a) and (b) satisfy condition (SI) which can be seen as follows. Consider model (a). Let $z_1, z_2 \in \mathcal{X}$ with $z_1 < z_2$ and $\theta = (\theta_1,\theta_2)^T, \theta' = (\theta_1',\theta_2')^T \in \Theta$ such that $\theta_1 z_j/(\theta_2 + z_j) = \theta_1' z_j/(\theta_2' + z_j)$ for $j = 1, 2$. Then

$$\frac{\theta_1}{\theta_1'} = \frac{\theta_2 + z_1}{\theta_2' + z_1} = \frac{\theta_2 + z_2}{\theta_2' + z_2}. \tag{1.6}$$

Since $\partial/\partial t \left\{(\theta_2 + t)/(\theta_2' + t)\right\} = (\theta_2' - \theta_2)/(\theta_2' + t)^2$ for $t \in (0,\infty)$, the function $t \mapsto (\theta_2 + t)/(\theta_2' + t)$ on $(0,\infty)$ is either increasing (if $\theta_2 < \theta_2'$) or decreasing (if $\theta_2 > \theta_2'$) or constant (if $\theta_2 = \theta_2'$). So, by $z_1 < z_2$ and (1.6), one gets $\theta_2 = \theta_2'$ and, again by (1.6), $\theta_1 = \theta_1'$, hence $\theta = \theta'$. So (SI) holds for model (a). Consider model (b). Let $z_1, z_2 \in \mathcal{X}$ with $z_1 < z_2$ and $\theta = (\theta_1,\theta_2)^T, \theta' = (\theta_1',\theta_2')^T \in \Theta$ such that $\theta_1 \exp\big(-\theta_2 z_j\big) = \theta_1' \exp\big(-\theta_2' z_j\big)$ for $j = 1, 2$. Then

$$\frac{\theta_1}{\theta_1'} = \exp\big\{(\theta_2 - \theta_2')z_1\big\} = \exp\big\{(\theta_2 - \theta_2')z_2\big\},$$

and by $z_1 < z_2$ this yields $\theta_2' = \theta_2$, and hence $\theta_1' = \theta_1$. So $\theta' = \theta$ and (SI) has been verified for model (b). □

Example 2 generalized linear regression.
Let $\Theta \subseteq \mathbb{R}^p$ and $\mu(x, \theta) = G\big(f^T(x)\,\theta\big)$ for all $(x, \theta) \in \mathcal{X} \times \Theta$, where $f : \mathcal{X} \longrightarrow \mathbb{R}^p$ is a continuous function, and $G : I \longrightarrow \mathbb{R}$ is an increasing continuous function on an interval $I \subseteq \mathbb{R}$ with $\big\{f^T(x)\,\theta : (x, \theta) \in \mathcal{X} \times \Theta\big\} \subseteq I$. Let $z_1, \ldots, z_p \in \mathcal{X}$ be pairwise distinct and $\theta, \theta' \in \Theta$. Clearly, $\mu(z_j, \theta) = \mu(z_j, \theta')$ for all $j = 1, \ldots, p$ is equivalent to $f^T(z_j)\,(\theta - \theta') = 0$ for all $j = 1, \ldots, p$. So, for the present model, condition (SI) is equivalent to the following.
If $\theta, \theta' \in \Theta$ and $z_1, \ldots, z_p \in \mathcal{X}$ pairwise distinct such that $\theta - \theta'$ is orthogonal to the vectors $f(z_1), \ldots, f(z_p)$, then $\theta = \theta'$.
Assume that the parameter space Θ is nondegenerate in the sense that there is no hyperplane of $\mathbb{R}^p$ covering Θ. Then the set of differences $\{\theta - \theta' : \theta, \theta' \in \Theta\}$ spans $\mathbb{R}^p$, and hence condition (SI) is equivalent to the following condition.
(Ch) *If $z_1, \ldots, z_p \in \mathcal{X}$ are pairwise distinct then the vectors $f(z_1), \ldots, f(z_p)$ are linearly independent.*
The notation 'Ch' stands for Chebyshev since, as it can easily be seen, condition (Ch) holds if and only if the (real-valued) component functions $f_1, \ldots, f_p$, say, of f constitute a Chebyshev system, i.e., every linear combination $\sum_{j=1}^{p} c_j f_j(x)$, with coefficients $c_j \in \mathbb{R}$ $(1 \le j \le p)$ not all equal to zero, has at most $p - 1$ distinct zeros on $\mathcal{X}$, cf. Karlin (1968), p. 24. Usually, Chebyshev systems are considered to be functions of one real variable defined on an interval of the real line. The most prominent example is given by the monomials $1, x, \ldots, x^{p-1}$ on some interval. For example, condition (Ch) or (SI), respectively, does not hold for generalized first order regression in two variables on a rectangle $\mathcal{X} = [a_1, b_1] \times [a_2, b_2]$, where $f(x) = (1, x_1, x_2)^T$ for all $x = (x_1, x_2) \in \mathcal{X}$. In this example we have $p = 3$. Let $z^{(1)}, z^{(2)}, z^{(3)} \in \mathcal{X}$ be pairwise distinct. One easily verifies that the vectors $f(z^{(1)})$, $f(z^{(2)})$, $f(z^{(3)})$ are linearly independent if and only if the three points $z^{(1)}$, $z^{(2)}$, $z^{(3)}$ do not lie on a common line. So (Ch) and hence (SI) do not hold. As a consequence from the present Example 2 one may guess that the class of models satisfying condition (SI) is not very large. In particular, generalized linear regression models with more than one regressors will usually not be members of that class. However, such models will be included in our derivations by the class given by condition (GLM*). □

The employed stochastic model for the adaptive Wynn algorithm includes a martingale difference scheme for the error variables, see Sect. 3. Limit theorems for martingales can be applied: a Strong Law of Large Numbers and a Central Limit Theorem to prove strong consistency and asymptotic normality, respectively, of the adaptive LSEs, see Theorems 1 and 2. Some auxiliary results are the content of Sect. 2. The proofs of the results of Sects. 2 and 3 are presented in the Appendix.

2 Auxiliary results

Throughout we assume (B1)–(B4) as introduced in the previous section. Note, however, that (B3) will not play a role in this section. Firstly, we give a proof of the existence of an initial design as required in the algorithm.

Lemma 1 *There exist an* $n_{\mathrm{st}} \in \mathbb{N}$ *and design points* $x_1, \ldots, x_{n_{\mathrm{st}}} \in \mathcal{X}$ *such that for every* $\theta \in \Theta$ *the vectors* $f_\theta(x_1), \ldots, f_\theta(x_{n_{\mathrm{st}}})$ *span* $\mathbb{R}^p$. *Hence, for such* x_i, $1 \le i \le n_{\mathrm{st}}$, *the design* $\xi_{n_{\mathrm{st}}} = \frac{1}{n_{\mathrm{st}}} \sum_{i=1}^{n_{\mathrm{st}}} \delta_{x_i}$ *has the property that its information matrix* $M(\xi_{n_{\mathrm{st}}}, \theta)$ *is positive definite for all* $\theta \in \Theta$.

Remark 1 Some popular nonlinear regression models, preferably those with a scalar regressor variable x, i.e., $\mathcal{X} \subseteq \mathbb{R}$, enjoy a further 'Chebyshev property' (Ch*), which was essentially assumed by Pronzato (2010) as condition $\mathbf{H}_{\mathcal{X}}$– (iv). It states that condition (Ch) from Example 2 holds for each function f_θ:

(Ch*) If $z_1, \ldots, z_p \in \mathcal{X}$ are pairwise distinct and $\theta \in \Theta$ then the vectors $f_\theta(z_1), \ldots, f_\theta(z_p)$ are linearly independent.

If (Ch*) holds then a suitable initial design for the algorithm is provided by any saturated design, i.e., choose $n_{\mathrm{st}} = p$ and pairwise distinct design points $x_1, \ldots, x_p \in \mathcal{X}$. Note also that in the proof of the lemma, assuming (Ch*), one has $U(\theta) = \Theta$ for all $\theta \in \Theta$ and hence $r = 1$ and $n_{\mathrm{st}} = p$. As an example consider the regression models from Example 1, (a) and (b), together with the assumption that f_θ is given by the gradient of $\mu(x, \theta)$ w. r. t. θ, for all $(x, \theta) \in \mathcal{X} \times \Theta$, which yields:

$$\text{(a)}\quad f_\theta(x) = \left(\frac{x}{\theta_2 + x}\,,\; -\frac{\theta_1 x}{(\theta_2 + x)^2} \right)^T, \qquad \text{(b)}\quad f_\theta(x) = \exp(-\theta_2 x) \left(1\,,\; -\theta_1 x \right)^T$$

for all $x \in \mathcal{X} \subseteq (0\,,\,\infty)$ and $\theta = (\theta_1, \theta_2)^T \in \Theta \subseteq (0\,,\,\infty)^2$. In either case (a) or (b), it is straightforward to show that for any given $\theta \in \Theta$ and $x, x' \in \mathcal{X}$, $x < x'$, the two vectors $f_\theta(x)$, $f_\theta(x')$ are linearly independent and hence both models (a) and (b) satisfy (Ch*).

In general, however, Lemma 1 and its proof does not give a practical way of finding a value of r and thus of $n_{\mathrm{st}} = rp$, or even an upper bound on them. Another condition ensuring the existence of an initial design with $n_{\mathrm{st}} = p$ is condition (GLM) considered in Freise et al. (2019) which is the second half of condition (GLM*) from Sect. 1.

(GLM) $f_\theta(x) = \psi(x, \theta)\, f(x)$ for all $(x, \theta) \in \mathcal{X} \times \Theta$, where $\psi : \mathcal{X} \times \Theta \longrightarrow (0\,,\,\infty)$ and $f : \mathcal{X} \longrightarrow \mathbb{R}^p$ are continuous functions.

In fact, (GLM) together with our basic assumption (B4) (i) implies that the image $f(\mathcal{X})$ spans $\mathbb{R}^p$. So one can find p points $x_1, \ldots, x_p \in \mathcal{X}$ such that the vectors $f(x_1), \ldots, f(x_p)$ are linearly independent, and by (GLM) for every $\theta \in \Theta$ the vectors $f_\theta(x_1), \ldots, f_\theta(x_p)$ are linearly independent. So the saturated design $\xi_p = \frac{1}{p} \sum_{i=1}^{p} \delta_{x_i}$ is an appropriate initial design for the algorithm. □

Let any path of the adaptive Wynn algorithm be given as described in the previous section. In particular, x_i, $i \in \mathbb{N}$, is the sequence of design points and $\xi_n = \frac{1}{n} \sum_{i=1}^{n} \delta_{x_i}$, $n \ge n_{\mathrm{st}}$, is the corresponding sequence of designs. For the following two lemmas no assumption on the employed adaptive estimators $\widehat{\theta}_n$, $n \ge n_{\mathrm{st}}$, is needed. In other

words, the sequence θ_n, $n \geq n_{\mathrm{st}}$, of parameter estimates appearing in the path may be arbitrary.

We denote the distance function in the compact metric space $\mathcal{X}$ by $\mathrm{d}_{\mathcal{X}}(x, z), x, z \in \mathcal{X}$. If S_1 and S_2 are nonempty subsets of $\mathcal{X}$ then the distance $\mathrm{d}_{\mathcal{X}}(S_1, S_2)$ of S_1 and S_2 is defined by $\mathrm{d}_{\mathcal{X}}(S_1, S_2) = \inf\{\mathrm{d}_{\mathcal{X}}(x, z) : x \in S_1,\ z \in S_2\}$. In case that $S_1 = \{x\}$ is a singleton we write $\mathrm{d}_{\mathcal{X}}(x, S_2)$ instead of $\mathrm{d}_{\mathcal{X}}(\{x\}, S_2)$. If S is a nonempty subset of $\mathcal{X}$ then the diameter of S is defined by $\mathrm{diam}(S) = \sup\{\mathrm{d}_{\mathcal{X}}(x, z) : x, z \in S\}$.

Lemma 2 *Suppose $p \geq 2$. Let $\varepsilon > 0$ be given. Then there exist $d > 0$ and $n_0 \geq n_{\mathrm{st}}$ such that*

$$\xi_n(S) \leq \frac{1}{p} + \varepsilon \text{ for all } \emptyset \neq S \subseteq \mathcal{X} \text{ with } \mathrm{diam}(S) \leq d \text{ and all } n \geq n_0.$$

Lemma 3 *Suppose $p \geq 2$. There exist $n_0 \geq n_{\mathrm{st}}$, $\pi_0 > 0$, and $d_0 > 0$ such that the following holds.*

For each $n \geq n_0$ there are p subsets $S_{1,n}, S_{2,n}, \ldots, S_{p,n}$ of $\mathcal{X}$ such that
$\xi_n(S_{j,n}) \geq \pi_0$, $1 \leq j \leq p$, $\mathrm{diam}(S_{j,n}) \leq d_0$, $1 \leq j \leq p$, *and*
$\mathrm{d}_{\mathcal{X}}(S_{j,n}, S_{k,n}) \geq d_0$, $1 \leq j < k \leq p$.

Remark 2 In the case that $\mathcal{X}$ is finite it is easily seen that in Lemma 3 the subsets $S_{1,n}, \ldots, S_{p,n}$ can be chosen to be singletons for all $n \geq n_0$. So, in this case, the lemma yields the result of Lemma 2 of Pronzato (2010). □

3 Convergence of least squares estimators

For an analysis of the adaptive Wynn algorithm, the generated sequence x_i, $i \in \mathbb{N}$, of design points and the observed (real valued) responses y_i, are viewed as values of random variables X_i, $i \in \mathbb{N}$, and Y_i, $i \in \mathbb{N}$, respectively, whose dependence structure is described by the following two assumptions (A1) and (A2), see Lai and Wei (1982), Lai (1994) and Chen et al. (1999), see also (Freise et al. 2019), The model thereby stated might be called an 'adaptive regression model'. By $\overline{\theta}$ we denote the true point of the parameter space Θ governing the data. All the random variables appearing in this section are thought to be defined on a common probability space $(\Omega, \mathcal{F}, \mathbb{P}_{\overline{\theta}})$, where Ω is a nonempty set, $\mathcal{F}$ is a sigma-field of subsets of Ω, and $\mathbb{P}_{\overline{\theta}}$ is a probability measure on $\mathcal{F}$ corresponding to the true parameter point $\overline{\theta}$. We assume, as before, the basic conditions (B1)–(B4), and now additionally the following conditions (A1) and (A2) constituting the adaptive regression model.

(A1) There is a given nondecreasing sequence of sub-sigma-fields of $\mathcal{F}$, $\mathcal{F}_0 \subseteq \mathcal{F}_1 \subseteq \ldots \subseteq \mathcal{F}_n \subseteq \ldots$ such that for each $i \in \mathbb{N}$ the random variable X_i is $\mathcal{F}_{i-1}$-measurable and the random variable Y_i is $\mathcal{F}_i$-measurable.
(A2) $Y_i = \mu(X_i, \overline{\theta}) + e_i$ with real-valued square integrable random errors e_i such that $\mathrm{E}\big(e_i \,\big|\, \mathcal{F}_{i-1}\big) = 0$ a.s. for all $i \in \mathbb{N}$, and $\sup_{i \in \mathbb{N}} \mathrm{E}\big(e_i^2 \,\big|\, \mathcal{F}_{i-1}\big) < \infty$ a.s.

As before, $\widehat{\theta}_n$, $n \geq n_{\mathrm{st}}$, are the adaptive estimators employed by the algorithm, now viewed as random variables, $\widehat{\theta}_n = \widehat{\theta}_n(X_1, Y_1, \ldots, X_n, Y_n)$. Of course, a desirable

property of these estimators would be strong consistency, i.e., almost sure convergence to $\overline{\theta}$ (as $n \to \infty$), for short $\widehat{\theta}_n \xrightarrow{\text{a.s.}} \overline{\theta}$.

Remark 3 As shown in our previous paper (Freise et al. 2019), Corollary 3.2, if the estimators $\widehat{\theta}_n$ are strongly consistent, then the sequence ξ_n, $n \geq n_{\text{st}}$, of (random) designs generated by the algorithm is almost surely asymptotically D-optimal, in the sense that $M(\xi_n, \widehat{\theta}_n) \xrightarrow{\text{a.s.}} M(\xi^*_{\overline{\theta}}, \overline{\theta})$, where $\xi^*_{\overline{\theta}}$ is a locally D-optimal design at $\overline{\theta}$. In fact, the conclusion of that corollary is stronger: if the estimators $\widehat{\theta}_n$ are strongly consistent then $M(\xi_n, \widetilde{\theta}_n) \xrightarrow{\text{a.s.}} M(\xi^*_{\overline{\theta}}, \overline{\theta})$ holds for *every* strongly consistent sequence of Θ-valued estimators $\widetilde{\theta}_n$. □

The next result yields strong consistency of the adaptive LSEs $\widehat{\theta}^{(\text{LS})}_n$ for any adaptive estimators $\widehat{\theta}_n$ employed by the algorithm, provided that condition (SI) or condition (GLM*) holds.

Theorem 1 *Assume that condition (SI) or condition* (GLM*) *is satisfied. Then, irrespective of the employed sequence of adaptive estimators $\widehat{\theta}_n$ in the algorithm, the sequence of adaptive LSEs $\widehat{\theta}^{(\text{LS})}_n$ is strongly consistent: $\widehat{\theta}^{(\text{LS})}_n \xrightarrow{\text{a.s.}} \overline{\theta}$.*

Remark 4 A crucial point in the proof of Theorem 1 is the result of 'Step 2' stating that $\inf_{\theta \in C(\overline{\theta}, \varepsilon)} D_n(\theta, \overline{\theta})$ goes to infinity at least as fast as n a. s. when $n \to \infty$. This is due to the adaptive Wynn algorithm, while the results of 'Step 1' and 'Step 3' only use the model assumptions (A1) and (A2). More general adaptive sampling schemes modeled by (A1) and (A2) were addressed to in Pronzato (2009). By Theorem 1 of that paper, in case of a finite design space, strong consistency of adaptive LSEs already holds if the adaptive scheme is such that

$$\inf_{\theta \in C(\overline{\theta}, \varepsilon)} D_n(\theta, \overline{\theta}) \Big/ (\log n)^{\rho} \xrightarrow{\text{a.s.}} \infty$$

for all $\varepsilon > 0$, for some $\rho > 1$. In Theorem 1 of Pronzato (2010), see also Remark 1 in Pronzato (2009), it was claimed that in case of i. i. d. error variables e_i, $i \in \mathbb{N}$, and for a finite design space the condition may be weakened to

$$\inf_{\theta \in C(\overline{\theta}, \varepsilon)} D_n(\theta, \overline{\theta}) \Big/ (\log \log n) \xrightarrow{\text{a.s.}} \infty.$$

However, the proof of the latter in Pronzato (2010), pp. 210–211, is doubtful since the classical Law of Iterated Logarithm is applied to *random* subsequences of the error variables. A proof should rather use a martingale structure and, in particular, a Law of Iterated Logarithm for martigales. Unfortunately, we have not found the appropriate arguments. □

For deriving asymptotic normality of the adaptive least squares estimators further assumptions are needed. Firstly, the 'gradient condition' (B5) on the family of functions f_θ, $\theta \in \Theta$, and the mean response μ is added to conditions (B1)–(B4). Secondly, two additional conditions (L) and (AH) on the error variables in (A1)–(A2) are imposed, where 'L' stands for 'Lindeberg' and 'AH' for 'asymptotic homoscedasticity'.

(B5) $\Theta \subseteq \mathbb{R}^p$ (endowed with the usual Euclidean metric), $\text{int}(\Theta) \neq \emptyset$, where $\text{int}(\Theta)$ denotes the interior of Θ as a subset of $\mathbb{R}^p$, the function $\theta \mapsto \mu(x,\theta)$ is twice differentiable on $\text{int}(\Theta)$ for each fixed $x \in \mathcal{X}$, with gradients and Hessian matrices denoted by $\nabla\mu(x,\theta) = \left(\frac{\partial}{\partial\theta_1}\mu(x,\theta), \ldots, \frac{\partial}{\partial\theta_p}\mu(x,\theta)\right)^T$ and $\nabla^2\mu(x,\theta) = \left(\frac{\partial^2}{\partial\theta_i\partial\theta_j}\mu(x,\theta)\right)_{1\le i,j\le p}$, respectively, for $\theta = (\theta_1,\ldots,\theta_p)^T \in \text{int}(\Theta)$ and $x \in \mathcal{X}$. It is assumed that the functions $(x,\theta) \mapsto \nabla\mu(x,\theta)$ and $(x,\theta) \mapsto \nabla^2\mu(x,\theta)$ are continuous on $\mathcal{X} \times \text{int}(\Theta)$ and

$$f_\theta(x) = \nabla\mu(x,\theta) \quad \text{for all } x \in \mathcal{X} \text{and all } \theta \in \text{int}(\Theta).$$

For a subset $A \subseteq \Omega$ we denote by $\mathbf{1}(A)$ the function on Ω which is constantly equal to 1 on A and is constantly equal to 0 on $\Omega\backslash A$.

(L) $\frac{1}{n}\sum_{i=1}^n \mathrm{E}\left(e_i^2\mathbf{1}\left(|e_i| > \varepsilon\sqrt{n}\right)\middle|\mathcal{F}_{i-1}\right) \xrightarrow{\text{a.s.}} 0$ for all $\varepsilon > 0$.
(AH) $\mathrm{E}\left(e_n^2\middle|\mathcal{F}_{n-1}\right) \xrightarrow{\text{a.s.}} \sigma^2(\overline{\theta})$ for some positive real constant $\sigma(\overline{\theta})$.

The following two conditions (L') and (L") are less technical than the Lindeberg condition (L), and each of them implies (L).
(L') $\sup_{i\in\mathbb{N}} \mathrm{E}\left(|e_i|^\alpha\middle|\mathcal{F}_{i-1}\right) < \infty$ a.s. for some real $\alpha > 2$.
(L") The random variables e_i, $i \in \mathbb{N}$, are identically distributed, and e_i, $\mathcal{F}_{i-1}$ are independent for each $i \in \mathbb{N}$.

In fact, from (L'), observing the trivial inequality $e_i^2\mathbf{1}\left(|e_i| > \varepsilon\sqrt{n}\right) \le |e_i|^\alpha/(\varepsilon\sqrt{n})^{\alpha-2}$, it follows that

$$\frac{1}{n}\sum_{i=1}^n \mathrm{E}\left(e_i^2\mathbf{1}\left(|e_i| > \varepsilon\sqrt{n}\right)\middle|\mathcal{F}_{i-1}\right) \le \frac{1}{(\varepsilon\sqrt{n})^{\alpha-2}} \sup_{i\in\mathbb{N}} \mathrm{E}\left(|e_i|^\alpha\middle|\mathcal{F}_{i-1}\right) \xrightarrow{\text{a.s.}} 0.$$

From (L") it follows for all $i \in \mathbb{N}$

$$\mathrm{E}\left(e_i^2\mathbf{1}\left(|e_i| > \varepsilon\sqrt{n}\right)\middle|\mathcal{F}_{i-1}\right) = \mathrm{E}\left(e_i^2\mathbf{1}\left(|e_i| > \varepsilon\sqrt{n}\right)\right) = \mathrm{E}\left(e_1^2\mathbf{1}\left(|e_1| > \varepsilon\sqrt{n}\right)\right) \text{ a.s.}$$

Hence

$$\frac{1}{n}\sum_{i=1}^n \mathrm{E}\left(e_i^2\mathbf{1}\left(|e_i| > \varepsilon\sqrt{n}\right)\middle|\mathcal{F}_{i-1}\right) = \mathrm{E}\left(e_1^2\mathbf{1}\left(|e_1| > \varepsilon\sqrt{n}\right)\right) \text{ a.s.}$$

and the expectation on the r.h.s. converges to zero as $n \to \infty$. Note also that (L") implies $\mathrm{E}\left(e_i^2\middle|\mathcal{F}_{i-1}\right) = \mathrm{E}\left(e_1^2\right) = \sigma^2(\overline{\theta})$, say. Excluding the trivial case $\sigma^2(\overline{\theta}) = 0$, we see that condition (L") also implies condition (AH).

Remark 5 Condition (L') was employed by Lai and Wei (1982), Theorem 1 of that paper, and by Chen et al. (1999), condition (C4) on p. 1161 of that paper. Condition

(L") meets the assumption of i.i.d. error variables of Pronzato (2010) for a particular choice of the sequence of sub-sigma-fields $\mathcal{F}_i$, $i \in \mathbb{N}_0$. □

The k-dimensional normal distribution with expectation 0 and covariance matrix C is denoted by $\mathrm{N}(0, C)$, where C is a positive definite $k \times k$ matrix. In particular, $\mathrm{N}(0, I_k)$ is the k-dimensional standard normal distribution, where I_k denotes the $k \times k$ identity matrix. For a sequence W_n of $\mathbb{R}^k$-valued random variables, convergence in distribution of W_n (as $n \to \infty$) to a k-dimensional normal distribution $\mathrm{N}(0, C)$ is abbreviated by $W_n \xrightarrow{\mathrm{d}} \mathrm{N}(0, C)$. In the following theorem asymptotic normality of the adaptive least squares estimators $\widehat{\theta}_n^{(\mathrm{LS})}$ is established. To some extent our proof is similar to that of Theorem 2 in Pronzato (2009), though the assumptions are different. Note that, by our Theorem 1, the assumptions of strong consistency of the adaptive estimators $\widehat{\theta}_n$ employed by the algorithm and of strong consistency of the adaptive LSEs $\widehat{\theta}_n^{(\mathrm{LS})}$ are met if $\widehat{\theta}_n = \widehat{\theta}_n^{(\mathrm{LS})}$, $n \geq n_{\mathrm{st}}$, and if one of the conditions (SI) and (GLM*) holds.

Theorem 2 *Assume conditions (B5), (L), and (AH). Moreover, assume that* $\overline{\theta} \in \mathrm{int}(\Theta)$ *and the sequences* $\widehat{\theta}_n$ *and* $\widehat{\theta}_n^{(\mathrm{LS})}$ *of adaptive estimators employed by the algorithm and adaptive LSEs, respectively, are strongly consistent, i.e.,* $\widehat{\theta}_n \xrightarrow{\text{a.s.}} \overline{\theta}$ *and* $\widehat{\theta}_n^{(\mathrm{LS})} \xrightarrow{\text{a.s.}} \overline{\theta}$. *Then, denoting by* $M_*(\overline{\theta}) = M\big(\xi_{\overline{\theta}}^*, \overline{\theta}\big)$ *the information matrix of a locally D-optimal design at* $\overline{\theta}$, *one has*

$$\sqrt{n}\,\big(\widehat{\theta}_n^{(\mathrm{LS})} - \overline{\theta}\big) \xrightarrow{\mathrm{d}} \mathrm{N}\big(0, \sigma^2(\overline{\theta})\, M_*^{-1}(\overline{\theta})\big).$$

For illustration of the achieved convergence results, we present some simulations for the Michaelis-Menten model from Example 1, case (a). For the exponemtial decay model of part (b) of Example 1 we obtained similar simulation results which will not be reported here.

Example 3: Simulation. Assume the Michaelis-Menton model with $p = 2$ parameters,

$$\mu(x, \theta) = \frac{\theta_1 x}{\theta_2 + x},$$
$$f_\theta(x) = \nabla\mu(x, \theta) = \Big(\frac{x}{\theta_2 + x}\,,\, -\frac{\theta_1 x}{(\theta_2 + x)^2}\Big)^T$$

for $x \in \mathcal{X}$ and $\theta = (\theta_1, \theta_2)^T \in \Theta$. Let the experimental region be given by the interval $\mathcal{X} = [\,0.5\,,\,5\,]$ and the parameter space be the square $\Theta = [\,0.1\,,\,10\,]^2$. The true parameter point is chosen to be $\overline{\theta} = (1, 1)^T$. The error variables e_i, $i \in \mathbb{N}$, in (A1) are assumed to be i.i.d normally distributed with expectation zero and variance $\sigma^2(\overline{\theta}) = 0.04$. By simulations, S=10,000 (pieces of) paths $X_i^{(s)}$, $Y_i^{(s)}$, $1 \leq i \leq 500$, where $s = 1, \ldots, S$, of the adaptive Wynn algorithm were generated, where the employed adaptive estimators $\widehat{\theta}_n$ were chosen to be the adaptive LSEs: $\widehat{\theta}_n = \widehat{\theta}_n^{(\mathrm{LS})}$ for all n. The computation of the paths of adaptive LSEs was done by using R Core Team (2020). A fixed initial design $\xi_{n_{\mathrm{st}}}$ was used: The three-point design with equal weights

on the boundary points and the mid-point of the experimental interval, that is, $n_{\mathrm{st}} = 3$ and $x_1 = 0.5$, $x_2 = 2.75$, $x_3 = 5$. Figure 1 illustrates the (almost sure) asymptotic D-optimality of the design sequence generated by the adaptive Wynn algorithm, which is ensured by Theorem 1 and Remark 3. The simulated paths $\xi_n^{(s)}$, $3 \le n \le 500$, where $s = 1, \ldots, S$, yield D-efficiencies

$$\mathrm{eff}(\xi_n^{(s)}) \;=\; \Big\{ \det\big(M(\xi_n^{(s)}, \overline{\theta})\big) \big/ \det\big(M_*(\overline{\theta})\big) \Big\}^{1/2},$$

where $M^*(\overline{\theta}) = M(\xi_{\overline{\theta}}^*, \overline{\theta})$ is the information matrix of the locally D-optimal design at $\overline{\theta}$ which is the two-point design giving equal weights $1/2$ to the points $5/7$ and 5, see Bates and Watts (1988), pp. 125–126, and hence

$$M_*(\overline{\theta}) = \frac{1}{2} f_{\overline{\theta}}(\frac{5}{7}) f_{\overline{\theta}}^T(\frac{5}{7}) + \frac{1}{2} f_{\overline{\theta}}(5) f_{\overline{\theta}}^T(5) = \begin{bmatrix} 0.4340 & -0.1085 \\ -0.1085 & 0.0392 \end{bmatrix}.$$

For each $n \in \{3, \ldots, 500\}$, particular quantiles of the 'data' $\mathrm{eff}(\xi_n^{(s)})$, $1 \le s \le S$, are reported in Fig. 1: minimum, 10%-quantile, 25%-quantile, median, 75%-quantile, 90%-quantile, and maximum. For example, the 10%-quantile curve shows that after less than 50 iterations more than 90% of the simulated paths yield efficiencies at least 0.9. The asymptotic normality of the adaptive LSEs ensured by Theorem 2 suggests that n-times the mean squared error matrix of $\widehat{\theta}_n^{(\mathrm{LS})}$ at $\overline{\theta}$ should converge to the asymptotic covasriance matrix $\sigma^2(\overline{\theta})\, M_*^{-1}(\overline{\theta})$. In fact, this was observed for n-times the simulated mean squared error matrix. Figure 2 shows a plot of the $(2, 2)$-entry of that matrix, that is, n-times the simulated mean squared error of the second component $\widehat{\theta}_{2,n}^{(\mathrm{LS})}$ of $\widehat{\theta}_n^{(\mathrm{LS})}$. The $(2, 2)$-entry of the asymptotic covariance matrix is approximately 3.32, indicated by the horizontal line in the figure. Finally, by Fig. 3 the approximate normal distribution of $\widehat{\theta}_{2,n}^{(\mathrm{LS})}$ is visualized at $n = 250$ via a suitable histogram of the simulated estimates (along with a fitted normal density) and a normal qq-plot. The simulation yielded similar graphics for the first component of $\widehat{\theta}_n^{(\mathrm{LS})}$.

4 Discussion

The adaptive Wynn algorithm for nonlinear regression provides a particular adaptive sampling scheme which has been motivated by the classical iterative procedure established by Wynn (1970) for generating D-optimal designs under a linear model. In the nonlinear situation the strong consistency of the adaptive estimators employed by the algorithm is crucial for ensuring that the generated adaptive design sequence is asymptotically D-optimal in the sense of local D-optimality at the true parameter point (which, of course, is unknown). Note that choosing a locally optimal design would be the best if the true parameter point was known. The focus of the present paper is on adaptive least squares estimators (LSEs), and their strong consistency has been proved for two relevant classes of nonlinear (univariate) regression models: those which have the property of 'saturated identifiability' (SI) and, secondly, those which satisfy a con-

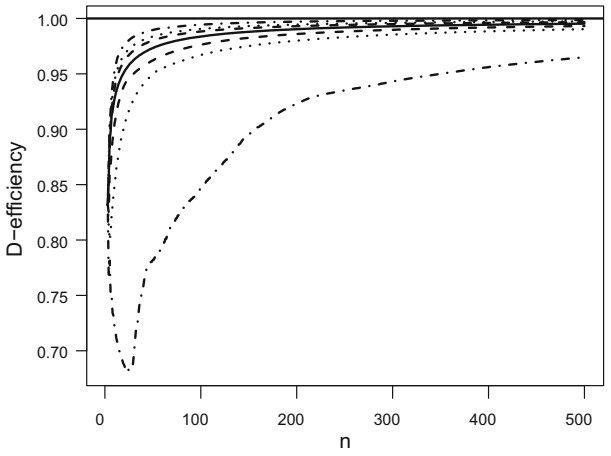

Fig. 1 Quantile curves of the simulated D-efficiencies. From bottom to top: minimum (dots and dashes), 10%-quantile (dotted), 25%-quantile (dashed), median (solid), 75%-quantile (dashed), 90%-quantile (dotted), maximum (dots and dashes)

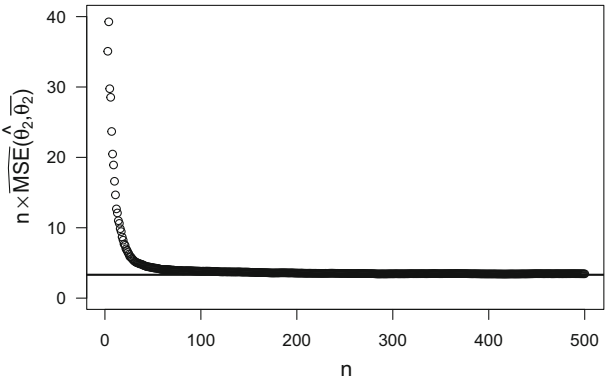

Fig. 2 Plot of n-times the simulated variance of $\widehat{\theta}_{2,n}^{(\mathrm{LS})}$. The horizontal line indicates the asymptotic variance (≈ 3.32)

dition (GLM*) including the class of generalized linear models. Condition (SI) seems to be restricted to models with a real valued regressor variable as the examples in Sect. 1 have shown. Generalized linear models (GLMs) constitute a great and important class of models. However, for a GLM, maximum likelihood or weighted least squares will usually be preferable to ordinary (unweighted) least squares as studied in the present paper. We note that adaptive maximum likelihood estimation in GLMs under the adaptive Wynn algorithm was studied in our previous paper (Freise et al. 2019). A challenging question for future research is to find extensions to other classes of models than just the (SI) class and, with regard to GLMs, to include weighted least squares estimation.

On the basis of strong consistency of adaptive LSEs their asymptotic normality has been established, under further suitable assumptions. The asymptotic covariance

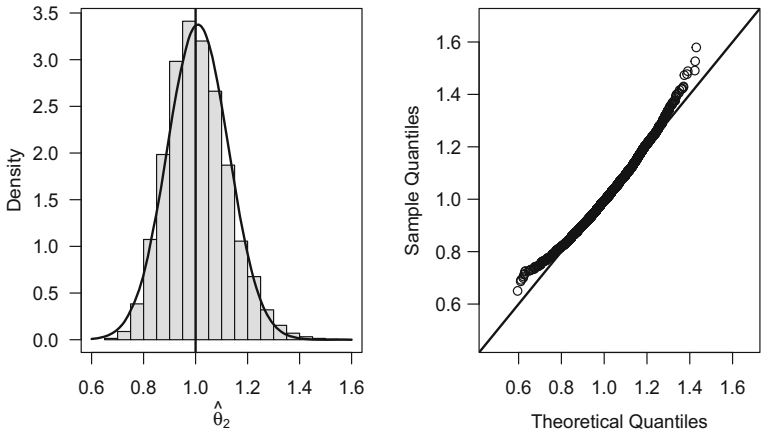

Fig. 3 Step $n = 250$: histogram estimate and fitted normal density (left) and normal QQ-plot (right) of $\widehat{\theta}_{2,n}^{(\mathrm{LS})}$, from the simulation

matrix is given by the inverse information matrix of the locally D-optimal design at the true parameter point. Thus, when measuring the size of a (nonsingular) covariance matrix by its determinant, the result implies the asymptotic efficiency of the adaptive LSEs. Two assumptions may be restrictive: firstly, the information matrices are built by a local first order Taylor expansion of the response (locally at a parameter point), without any scaling adjustment for possible variance heterogeneity. Secondly, a condition of 'asymptotic homoscedasticity' (AH) on the random errors has been imposed, that is, their (conditional) variances are assumed to become asymptotically constant, where the asymptotic variance may depend on the true parameter point. Both assumptions correspond to ordinary (unweighted) least squares employed by the adaptive LSEs under consideration. Again, extensions of the results are desirable which employ weaker assumptions to include models with variance heterogeneity as the majority of GLMs.

While the adaptive Wynn algorithm collects one point at each step and was therefore called 'one-step-ahead algorithm' by Pronzato (2010), an alternative approach is to collect more points at each step. For a special model, a related concept of 'batch sequential design' was employed by Müller and Pötscher (1992). In a forthcoming paper (Freise et al. 2020) we study a sequential adaptive algorithm for D-optimal design which we have called a 'p-step-ahead algorithm', since p design points are collected at each step. An idea of that algorithm was sketched by Ford et al. (1992) in the introduction of their paper, p. 570.

Funding Open Access funding enabled and organized by Projekt DEAL.

Compliance with ethical standards

Conflict of interest On behalf of all authors, the corresponding author states that there is no conflict of interest.

A Proofs

Proof of Lemma 1 By (B4)-(i), for each $\theta \in \Theta$ there exist p design points $z_1(\theta), \ldots, z_p(\theta) \in \mathcal{X}$ such that the vectors $f_\theta\big(z_1(\theta)\big), \ldots, f_\theta\big(z_p(\theta)\big)$ are linearly independent. By (B2) and (B4) (ii), for each $\theta \in \Theta$ the set

$$U(\theta) \,=\, \Big\{\tau \in \Theta \,:\, \det\big[f_\tau\big(z_1(\theta)\big), \ldots, f_\tau\big(z_p(\theta)\big)\big] \neq 0\Big\}$$

is an open set in the (compact) metric space Θ, and $\theta \in U(\theta)$. Hence, trivially, $\Theta = \bigcup_{\theta\in\Theta} U(\theta)$, and by (B2) there is an $r \in \mathbb{N}$ and points $\theta_1, \ldots, \theta_r \in \Theta$ such that $\Theta = \bigcup_{j=1}^r U(\theta_j)$. Denote $x_{ij} = z_i(\theta_j)$, $1 \le i \le p$, $1 \le j \le r$. Then, for every $\tau \in \Theta$ the set of vectors

$$\big\{f_\tau(x_{ij}) \,:\, 1 \le i \le p,\ 1 \le j \le r\big\}$$

spans $\mathbb{R}^p$. In fact, for any given $\tau \in \Theta$ there is some $j_0 \in \{1, \ldots, r\}$ with $\tau \in U(\theta_{j_0})$ hence $\det\big[f_\tau(x_{1j_0}), \ldots, f_\tau(x_{pj_0})\big] \neq 0$, i.e., the vectors $f_\tau(x_{1j_0}), \ldots, f_\tau(x_{pj_0})$ constitute a basis of $\mathbb{R}^p$. So, for $n_{\rm st} = pr$ and $x_1, \ldots, x_{n_{\rm st}}$ being a relabelled family of the points x_{ij}, $1 \le i \le p, 1 \le j \le r$, the vectors $f_\tau(x_1), \ldots, f_\tau(x_{n_{\rm st}})$ span $\mathbb{R}^p$, and hence the information matrix $M(\xi_{n_{\rm st}}, \tau) = \frac{1}{n_{\rm st}} \sum_{i=1}^{n_{\rm st}} f_\tau(x_i)\, f_\tau^T(x_i)$ is positive definite. □

Proof of Lemma 2 Without loss of generality we may assume $\varepsilon < 1 - p^{-1}$. In Freise et al. (2019) we introduced the positive real constants

$$\gamma = \sup_{x\in\mathcal{X},\,\theta\in\Theta} \|f_\theta(x)\| \quad \text{and} \quad \kappa = \inf_{\|v\|=1,\,\theta\in\Theta} \max_{x\in\mathcal{X}} \big\{v^T f_\theta(x)\big\}^2, \tag{A.1}$$

and Lemma 2.3 of that paper stated the following.

> If $0 < \eta < 1 - p^{-1/2}$, $n \ge n_{\rm st}$, and $S \subseteq \mathcal{X}$ are given such that
> $\|f_{\theta_n}(x) - f_{\theta_n}(z)\| \le \eta\kappa/\gamma$ for all $x, z \in S$ and $\xi_n(S) > (1-\eta)^{-2} p^{-1}$,
> then $x_{n+1} \notin S$. (A.2)

Choose $\eta := 1 - \big(1 + p\varepsilon/2\big)^{-1/2}$. Then $0 < \eta < 1 - p^{-1/2}$ and $(1-\eta)^{-2} p^{-1} = p^{-1} + \varepsilon/2$. By (B1), (B2), and (B4) the function $(x, \theta) \mapsto f_\theta(x)$ is uniformly

continuous on its compact domain $\mathcal{X} \times \Theta$. So there exists a $d > 0$ such that

$$\text{if } x, z \in \mathcal{X} \text{ and } \mathrm{d}_{\mathcal{X}}(x, z) \le d \text{ then } \|f_\theta(x) - f_\theta(z)\| \le \eta\kappa/\gamma \ \forall\, \theta \in \Theta. \tag{A.3}$$

We show that d fulfills the requirement of the lemma. Let $\emptyset \neq S \subseteq \mathcal{X}$ with $\mathrm{diam}(S) \le d$. By (A.3) and (A.2) the sequence $\xi_n(S)$, $n \ge n_{\mathrm{st}}$, has the property that for all $n \ge n_{\mathrm{st}}$,

$$\xi_{n+1}(S) = \frac{n}{n+1}\xi_n(S) \quad \text{if } \xi_n(S) > \frac{1}{p} + \frac{\varepsilon}{2},$$
$$\xi_{n+1}(S) \le \xi_n(S) + \frac{1}{n+1} \quad \text{if } \xi_n(S) \le \frac{1}{p} + \frac{\varepsilon}{2}.$$

An application of Lemma 2.1 in Freise et al. (2019) to the sequence $\beta_n := \xi_n(S)$, $n \ge n_{\mathrm{st}}$, and $\beta := \frac{1}{p} + \frac{\varepsilon}{2}$, $\widetilde{\beta} := \frac{1}{p} + \varepsilon$ yields that

$$\xi_n(S) \le \frac{1}{p} + \varepsilon \ \text{ for all } n \ge n_0 := \left\lceil \left(\frac{1}{p} + \frac{\varepsilon}{2}\right)^{-1} \right\rceil \cdot \max\left\{n_{\mathrm{st}}, \lceil 2/\varepsilon \rceil\right\},$$

where $\lceil a \rceil$, for $a \in \mathbb{R}$, denotes the smallest integer greater than or equal to a. Since n_0 does not depend on the particular set S the result follows. □

Proof of Lemma 3 Fix an ε with $0 < \varepsilon < \{p(p-1)\}^{-1}$. Choose $d > 0$ and $n_0 \ge n_{\mathrm{st}}$ according to Lemma 2. By compactness of $\mathcal{X}$ there is a positive integer q and nonempty subsets $R_1, \ldots, R_q$ of $\mathcal{X}$ such that

$$\mathcal{X} = \bigcup_{\ell=1}^{q} R_\ell \quad \text{and} \quad \mathrm{diam}(R_\ell) \le d/3 \ \text{ for all } \ell = 1, \ldots, q.$$

We show that $n_0, \pi_0 := \{p^{-1} - (p-1)\varepsilon\}/q$, and $d_0 := d/3$ satisfy the requirements of the assertion. To this end let $n \ge n_0$ be given. We construct inductively subsets $S_{j,n}$, $1 \le j \le p$, as required.

$\underline{j = 1}$: Clearly, $\sum_{\ell=1}^{q} \xi_n(R_\ell) \ge 1$. Choose $\ell_n \in \{1, \ldots, q\}$ achieving the maximum value of $\xi_n(R_\ell)$, $1 \le \ell \le q$, and set $S_{1,n} := R_{\ell_n}$. Then $\xi_n(S_{1,n}) = \max_{1 \le \ell \le q} \xi_n(R_\ell) \ge 1/q \ge \pi_0$ and $\mathrm{diam}(S_{1,n}) = \mathrm{diam}(R_{\ell_n}) \le d_0$.

Induction step: Let an $r \in \{1, \ldots, p-1\}$ be given along with subsets $S_{1,n}, \ldots, S_{r,n}$ of $\mathcal{X}$ such that $\xi_n(S_{j,n}) \ge \pi_0$ and $\mathrm{diam}(S_{j,n}) \le d_0$, $1 \le j \le r$, and $\mathrm{d}_{\mathcal{X}}(S_{j,n}, S_{k,n}) \ge d_0$, $1 \le j < k \le r$. Let $\overline{S}_{j,n} := \left\{x \in \mathcal{X} : \mathrm{d}_{\mathcal{X}}(x, S_{j,n}) \le d_0\right\}$, $1 \le j \le r$. As it is easily seen, $\mathrm{diam}(\overline{S}_{j,n}) \le 3d_0 = d$ and hence $\xi_n(\overline{S}_{j,n}) \le \frac{1}{p} + \varepsilon$. So for $T_{r,n} := \bigcup_{j=1}^{r} \overline{S}_{j,n}$ one has $\xi_n(T_{r,n}) \le r\left(\frac{1}{p} + \varepsilon\right)$, and hence

$$\xi_n(\mathcal{X} \backslash T_{r,n}) \ge 1 - r\left(\frac{1}{p} + \varepsilon\right) \ge 1 - (p-1)\left(\frac{1}{p} + \varepsilon\right) = p^{-1} - (p-1)\varepsilon.$$

Observing that $\mathcal{X}\backslash T_{r,n} = \bigcup_{\ell=1}^{q}(R_\ell \backslash T_{r,n})$ one gets

$$p^{-1} - (p-1)\varepsilon \le \sum_{\ell=1}^{q} \xi_n(R_\ell \backslash T_{r,n}).$$

Choose $\ell_n \in \{1, \ldots, q\}$ which achieves the maximum value of $\xi_n(R_\ell \backslash T_{r,n})$, $1 \le \ell \le q$, and set $S_{r+1,n} := R_{\ell_n} \backslash T_{r,n}$. Then $\xi_n(S_{r+1,n}) = \max_{1\le\ell\le q} \xi_n\big(R_\ell \backslash T_{r,n}\big) \ge \{p^{-1} - (p-1)\varepsilon\}/q = \pi_0$ and $\operatorname{diam}(S_{r+1,n}) \le \operatorname{diam}(R_{\ell_n}) \le d_0$. Moreover for each $j = 1, \ldots, r$, since $S_{r+1,n} \cap \overline{S}_{j,n} = \emptyset$, one has $\mathrm{d}_{\mathcal{X}}(x, S_{j,n}) > d_0$ for all $x \in S_{r+1,n}$ and hence $\mathrm{d}_{\mathcal{X}}(S_{r+1,n}, S_{j,n}) \ge d_0$. So we have subsets $S_{1,n}, \ldots, S_{r,n}, S_{r+1,n}$ such that

$$\begin{aligned} &\xi_n(S_{j,n}) \ge \pi_0 \ \text{ and } \ \operatorname{diam}(S_{j,n}) \le d_0, \quad 1 \le j \le r+1, \\ &\mathrm{d}_{\mathcal{X}}(S_{j,n}, S_{k,n}) \ge d_0, \quad 1 \le j < k \le r+1. \end{aligned}$$

This completes the inductive construction and the proof of the lemma. □

Proof of Theorem 1 Define for all $n \in \mathbb{N}$ and $\theta \in \Theta$ random variables

$$S_n(\theta) := \sum_{i=1}^{n} \big\{Y_i - \mu(X_i, \theta)\big\}^2 \quad \text{and} \quad D_n(\theta, \overline{\theta}) := \sum_{i=1}^{n} \big\{\mu(X_i, \theta) - \mu(X_i, \overline{\theta})\big\}^2.$$

The proof is divided into three steps. For $\varepsilon > 0$ we denote $C(\overline{\theta}, \varepsilon) := \big\{\theta \in \Theta : \mathrm{d}_\Theta(\theta, \overline{\theta}) \ge \varepsilon\big\}$, where d_Θ denotes the distance function in Θ.

Step 1. Show that for all $\varepsilon > 0$ with $C(\overline{\theta}, \varepsilon) \ne \emptyset$,

$$\left| \frac{1}{n} \left\{ \inf_{\theta \in C(\overline{\theta}, \varepsilon)} S_n(\theta) - S_n(\overline{\theta}) \right\} - \frac{1}{n} \inf_{\theta \in C(\overline{\theta}, \varepsilon)} D_n(\theta, \overline{\theta}) \right| \xrightarrow{\text{a.s.}} 0.$$

Step 2. Show that for all $\varepsilon > 0$ with $C(\overline{\theta}, \varepsilon) \ne \emptyset$,

$$\liminf_{n\to\infty} \left\{ \frac{1}{n} \inf_{\theta \in C(\overline{\theta}, \varepsilon)} D_n(\theta, \overline{\theta}) \right\} > 0 \ \text{ a.s.}$$

Step 3. Conclude from the results of Step 1 and Step 2 that for all $\varepsilon > 0$ with $C(\overline{\theta}, \varepsilon) \ne \emptyset$,

$$\inf_{\theta \in C(\overline{\theta}, \varepsilon)} S_n(\theta) - S_n(\overline{\theta}) \xrightarrow{\text{a.s.}} \infty. \tag{A.4}$$

From (A.4), applying Lemma 1 of Wu (1981), one gets $\widehat{\theta}_n^{(\mathrm{LS})} \xrightarrow{\text{a.s.}} \overline{\theta}$.
Ad Step 1. As in Pronzato (2010), p. 230, one calculates

$$S_n(\theta) - S_n(\overline{\theta}) = D_n(\theta, \overline{\theta}) + 2W_n(\theta, \overline{\theta}), \quad \text{where}$$

$$W_n(\theta,\overline{\theta}) := \sum_{i=1}^{n}\big(\mu(X_i,\overline{\theta})-\mu(X_i,\theta)\big)\,e_i.$$

It follows that

$$\left|\frac{1}{n}\left\{\inf_{\theta\in C(\overline{\theta},\varepsilon)} S_n(\theta) - S_n(\overline{\theta})\right\} - \frac{1}{n}\inf_{\theta\in C(\overline{\theta},\varepsilon)} D_n(\theta,\overline{\theta})\right| \le \frac{2}{n}\sup_{\theta\in\Theta}\big|W_n(\theta,\overline{\theta})\big|.$$

Applying Lemma 3.1, part (c), in Freise et al. (2019) with $h(x,\theta)=\mu(x,\overline{\theta})-\mu(x,\theta)$, $(x,\theta)\in\mathcal{X}\times\Theta$, the result of Step 1 follows.

Ad Step 2 in case that condition (SI) holds.
Consider any path x_i, y_i, $i\in\mathbb{N}$, and θ_n, $n\ge n_{\rm st}$ of the sequences X_i, Y_i, $i\in\mathbb{N}$, and $\widehat{\theta}_n$, $n\ge n_{\rm st}$. Firstly, consider the simple case $p=1$. Then condition (SI) implies that $\mu(x,\theta)\ne\mu(x,\overline{\theta})$ for all $\theta\in C(\overline{\theta},\varepsilon)$, and hence by (B3)

$$c_\varepsilon := \inf_{x\in\mathcal{X}}\big\{\mu(x,\theta)-\mu(x,\overline{\theta})\big\}^2 > 0.$$

It follows that $\frac{1}{n}\inf_{\theta\in C(\overline{\theta},\varepsilon)} D_n(\theta,\overline{\theta}) \ge c_\varepsilon$ for all n and, in particular, its limit inferior is positive. Now let $p\ge 2$. According to Lemma 3, choose $n_0\ge n_{\rm st}$, $\pi_0>0$, $d_0>0$, and subsets $S_{1,n},\ldots,S_{p,n}\subseteq\mathcal{X}$ for all $n\ge n_0$. Define a subset of the p-fold product space $\mathcal{X}^p$ by

$$\Delta := \big\{(z_1,\ldots,z_p)\in\mathcal{X}^p : \mathrm{d}_{\mathcal{X}}(z_j,z_k)\ge d_0,\ \ 1\le j<k\le p\big\}.$$

By (SI), $\sum_{j=1}^{p}\big(\mu(z_j,\theta)-\mu(z_j,\overline{\theta})\big)^2>0$ for all $(z_1,\ldots,z_p)\in\Delta$ and all $\theta\ne\overline{\theta}$. By (B1) the set Δ is compact and by (B2) the set $C(\overline{\theta},\varepsilon)$ is compact. So, together with (B3), one concludes that the following infimum c_ε is positive,

$$c_\varepsilon := \inf\left\{\sum_{j=1}^{p}\big\{\mu(z_j,\theta)-\mu(z_j,\overline{\theta})\big\}^2 : (z_1,\ldots,z_p)\in\Delta,\ \theta\in C(\overline{\theta},\varepsilon)\right\}.$$

For all $n\ge n_0$ and all permutations σ of $\{1,\ldots,p\}$ the Cartesian product $S_n^\sigma := S_{\sigma(1),n}\times S_{\sigma(2),n}\times\ldots\times S_{\sigma(p),n}$ is a subset of Δ, hence $R_n := \bigcup_\sigma S_n^\sigma\subseteq\Delta$. Note that $S_n^\sigma\cap S_n^\tau=\emptyset$ for any two different permutations σ and τ. Consider the p-fold product measure ξ_n^p. Then, for all σ and all $n\ge n_0$ one has $\xi_n^p(S_n^\sigma)=\prod_{j=1}^{p}\xi_n(S_{\sigma(j),n})\ge\pi_0^p$ and hence $\xi_n(R_n)\ge p!\,\pi_0^p$. So

$$\int_{\mathcal{X}^p}\sum_{j=1}^{p}\big\{\mu(z_j,\theta)-\mu(z_j,\overline{\theta})\big\}^2\,\mathrm{d}\xi_n^p(z_1,\ldots,z_p) \ge c_\varepsilon\,p!\,\pi_0^p \ \text{ for all } n\ge n_0 \text{ and } \theta\in C(\overline{\theta},\varepsilon).$$

The integral on the l.h.s. of that inequality is equal to

$$p \int_{\mathcal{X}} \big\{\mu(z,\theta) - \mu(z,\overline{\theta})\big\}^2 \,\mathrm{d}\xi_n(z) \;=\; \frac{p}{n} \sum_{i=1}^{n} \big\{\mu(x_i,\theta) - \mu(x_i,\overline{\theta})\big\}^2.$$

It follows that

$$\begin{aligned} &\frac{1}{n} \inf_{\theta \in C(\overline{\theta},\varepsilon)} \sum_{i=1}^{n} \big\{\mu(x_i,\theta) - \mu(x_i,\overline{\theta})\big\}^2 \\ &\quad \geq\; c_\varepsilon\, (p-1)!\, \pi_0^p \quad \forall\, n \geq n_0, \end{aligned}$$

which implies that the limit inferior of the l.h.s. of that inequality is positive.

Ad Step 2 in case that condition (GLM*) holds.
Again, consider any path x_i, y_i, $i \in \mathbb{N}$, and θ_n, $n \geq n_{\mathrm{st}}$ of the sequences X_i, Y_i, $i \in \mathbb{N}$, and $\widehat{\theta}_n$, $n \geq n_{\mathrm{st}}$. By continuity of f and compactness of $\mathcal{X} \times \Theta$, there is a compact subinterval $J \subseteq I$ such that $\{f^T(x)\,\theta \,:\, (x,\theta) \in \mathcal{X} \times \Theta\} \subseteq J$. Since G is differentiable on I with continuous and positive derivative G', one gets from the mean value theorem that $|G(u) - G(v)| \geq b|u - v|$ for all $u, v \in J$ with $b := \min_{w \in J} G'(w) > 0$. So, for all $i \in \mathbb{N}$ and all $\theta \in C(\overline{\theta}, \varepsilon)$,

$$\big|\mu(x_i,\theta) - \mu(x_i,\overline{\theta})\big| \;=\; \big|G\big(f^T(x_i)\,\theta\big) - G\big(f^T(x_i)\,\overline{\theta}\big)\big| \;\geq\; b\,\big|f^T(x_i)\,(\theta - \overline{\theta})\big|.$$

From this we get for all $\theta \in C(\overline{\theta}, \varepsilon)$, denoting $a_\theta = (\theta - \overline{\theta})/\|\theta - \overline{\theta}\|$,

$$\begin{aligned} D_n(\theta,\overline{\theta}) &= \sum_{i=1}^{n} \big\{\mu(x_i,\theta) - \mu(x_i,\overline{\theta})\big\}^2 \\ &\geq\; b^2\varepsilon^2 \sum_{i=1}^{n} \big\{f^T(x_i)\,a_\theta\big\}^2 \;=\; b^2\varepsilon^2 n \int_{\mathcal{X}} \big\{f^T(x)\,a_\theta\big\}^2 \mathrm{d}\xi_n(x). \end{aligned} \tag{A.5}$$

By Theorem 2.6 and Lemma 2.5 of Freise et al. (2019) there exist $n_0 \geq n_{\mathrm{st}}$, $\rho > 0$, and $\alpha \in (0\,,\,1)$ such that for all $n \geq n_0$ and all normalized coefficient vectors $a \in \mathbb{R}^p$, $\|a\| = 1$, one has $\xi_n\big(\big\{x \in \mathcal{X} \,:\, |f_{\theta_n}^T(x)\,a| \leq \rho\big\}\big) \leq \alpha$. From $f_{\theta_n}(x) = \psi(x,\theta_n)\,f(x)$ one gets

$$\big\{x \in \mathcal{X} \,:\, |f_{\theta_n}^T(x)\,a| > \rho\big\} \;\subseteq\; \big\{x \in \mathcal{X} \,:\, |f^T(x)\,a| > \rho/\psi_{\max}\big\},$$

where $\psi_{\max} := \max_{(x,\theta) \in \mathcal{X} \times \Theta} \psi(x,\theta)$ which is positive and finite. Note that

$$\xi_n\Big(\big\{x \in \mathcal{X} \,:\, |f^T(x)\,a| > \rho/\psi_{\max}\big\}\Big) \;\geq\; 1 - \alpha \;\text{ for all } n \geq n_0.$$

From this and from (A.5) it follows that for all $n \geq n_0$ and all $\theta \in C(\overline{\theta}, \varepsilon)$,

$$D_n(\theta, \overline{\theta}) \;\geq\; b^2\varepsilon^2 n\,(\rho/\psi_{\max})^2(1-\alpha),$$

and hence

$$\frac{1}{n}\inf_{\theta\in C(\overline{\theta},\varepsilon)} D_n(\theta, \overline{\theta}) \;\geq\; b^2\varepsilon^2(\rho/\psi_{\max})^2(1-\alpha),$$

and the result of Step 2 follows.

Ad Step 3. By the results of Step 1 and Step 2,

$$\liminf_{n\to\infty}\frac{1}{n}\left\{\inf_{\theta\in C(\overline{\theta},\varepsilon)} S_n(\theta) - S_n(\overline{\theta})\right\} \;=\; \liminf_{n\to\infty}\frac{1}{n}\inf_{\theta\in C(\overline{\theta},\varepsilon)} D_n(\theta,\overline{\theta}) \;>\; 0 \;\text{ a.s.}$$

Hence $\inf_{\theta\in C(\overline{\theta},\varepsilon)} S_n(\theta) - S_n(\overline{\theta}) \xrightarrow{\text{a.s.}} \infty$. □

Proof of Theorem 2 Choose a compact ball $\overline{B}$ centered at $\overline{\theta}$ and such that $\overline{B} \subseteq \text{int}(\Theta)$. By the strong consistency of the sequence of adaptive LSEs there is a random variable N with values in $\mathbb{N} \cup \{\infty\}$ such that $N < \infty$ a.s. and $\widehat{\theta}_n^{(\text{LS})} \in \overline{B}$ on $\{N \leq n\}$ for all integers $n \geq n_{\text{st}}$. Note that, since N is almost surely finite, $\mathbf{1}(N \leq n) \xrightarrow{\text{a.s.}} 1$ as $n \to \infty$. Recall our notation introduced earlier: $S_n(\theta) = \sum_{i=1}^n \{Y_i - \mu(X_i,\theta)\}^2$, $n \geq n_{\text{st}}$, $\theta \in \Theta$. For the gradients of $S_n(\theta)$ w.r.t. θ one obtains, using (B5),

$$\nabla S_n(\theta) \;=\; -2\sum_{i=1}^n \{Y_i - \mu(X_i,\theta)\}\,\nabla\mu(X_i,\theta), \quad \theta \in \text{int}(\Theta). \tag{A.6}$$

On $\{N \leq n\}$ the gradient at $\widehat{\theta}_n^{(\text{LS})}$ is equal to zero, and hence $\nabla S_n\big(\widehat{\theta}_n^{(\text{LS})}\big) - \nabla S_n(\overline{\theta}) = -\nabla S_n(\overline{\theta})$. That equation yields, inserting from (A.6) and $Y_i = \mu(X_i,\overline{\theta}) + e_i$ from (A2), along with some

$$\begin{aligned}\sum_{i=1}^n e_i\nabla\mu(X_i,\overline{\theta}) \;&=\; \sum_{i=1}^n \{\mu(X_i,\widehat{\theta}_n^{(\text{LS})}) - \mu(X_i,\overline{\theta})\}\,\nabla\mu(X_i,\widehat{\theta}_n^{(\text{LS})}) \\ &\quad - \sum_{i=1}^n e_i\,\{\nabla\mu(X_i,\widehat{\theta}_n^{(\text{LS})}) - \nabla\mu(X_i,\overline{\theta})\} \qquad \text{on } \{N\leq n\}.\end{aligned} \tag{A.7}$$

We firstly show that

$$n^{-1/2}\sigma^{-1}(\overline{\theta})\,M_*^{-1/2}(\overline{\theta})\sum_{i=1}^n e_i\nabla\mu(X_i,\overline{\theta}) \xrightarrow{\text{d}} \text{N}(0, I_p). \tag{A.8}$$

To this end, according to the Cramér-Wold device, let $v \in \mathbb{R}^p$, $v^T v = 1$, be given. Denote $Z_i := \sigma^{-1}(\overline{\theta})\, v^T M_*^{-1/2}(\overline{\theta}) \nabla\mu(X_i, \overline{\theta})$ and $\widetilde{e}_i := e_i Z_i$, $i \in \mathbb{N}$. Abbreviating the random variables on the l.h.s. of (A.8) by W_n, one has $v^T W_n = n^{-1/2} \sum_{i=1}^n \widetilde{e}_i$. The random variable Z_i is $\mathcal{F}_{i-1}$-measurable for all $i \in \mathbb{N}$, and the Z_i, $i \in \mathbb{N}$, are uniformly bounded: $|Z_i| \le c$ for all $i \in \mathbb{N}$ for some positive real constant c. Hence the sequence of partial sums $\sum_{i=1}^n \widetilde{e}_i$, is a martingale w.r.t. $\mathcal{F}_n$, $n \in \mathbb{N}$, and we can apply Corollary 3.1 of Hall and Heyde (1980) which states that the following two conditions (a) and (b) together imply the distributional convergence $n^{-1/2} \sum_{i=1}^n \widetilde{e}_i \xrightarrow{\mathrm{d}} \mathrm{N}(0, 1)$.

(a) $\frac{1}{n} \sum_{i=1}^n \mathrm{E}\big(\widetilde{e}_i^2 \big| \mathcal{F}_{i-1}\big) \xrightarrow{\text{a.s.}} 1$, (b) $\frac{1}{n} \sum_{i=1}^n \mathrm{E}\Big(\widetilde{e}_i^2\, \mathbf{1}\big(|\widetilde{e}_i| > \varepsilon\sqrt{n}\big) \big| \mathcal{F}_{i-1}\Big) \xrightarrow{\text{a.s.}} 0$ for all $\varepsilon > 0$.

Condition (b) follows from condition (L) since

$$\mathrm{E}\Big(\widetilde{e}_i^2\, \mathbf{1}\big(|\widetilde{e}_i| > \varepsilon\sqrt{n}\big) \,\big|\, \mathcal{F}_{i-1}\Big) \le c^2 \mathrm{E}\Big(e_i^2\, \mathbf{1}\big(|e_i| > (\varepsilon/c)\sqrt{n}\big) \,\big|\, \mathcal{F}_{i-1}\Big).$$

To verify (a) we write

$$\begin{aligned} \frac{1}{n} \sum_{i=1}^n \mathrm{E}\big(\widetilde{e}_i^2 \big| \mathcal{F}_{i-1}\big) &= \frac{1}{n} \sum_{i=1}^n \mathrm{E}\big(e_i^2 \big| \mathcal{F}_{i-1}\big)\, Z_i^2 \\ &= \frac{1}{n} \sum_{i=1}^n \big\{ \mathrm{E}\big(e_i^2 \big| \mathcal{F}_{i-1}\big) - \sigma^2(\overline{\theta}) \big\}\, Z_i^2 \;+\; \sigma^2(\overline{\theta}) \frac{1}{n} \sum_{i=1}^n Z_i^2. \end{aligned}$$

By (AH) and $|Z_n| \le c$ for all $n \in \mathbb{N}$ one has $\big[\mathrm{E}\big(e_n^2 \big| \mathcal{F}_{n-1}\big) - \sigma^2(\overline{\theta})\big]\, Z_n^2 \xrightarrow{\text{a.s.}} 0$ and hence $\frac{1}{n} \sum_{i=1}^n \big\{ \mathrm{E}\big(e_i^2 \big| \mathcal{F}_{i-1}\big) - \sigma^2(\overline{\theta}) \big\}\, Z_i^2 \xrightarrow{\text{a.s.}} 0$. By the definition of Z_i, $i \in \mathbb{N}$, and by (B5),

$$\begin{aligned} \sigma^2(\overline{\theta}) \frac{1}{n} \sum_{i=1}^n Z_i^2 &= v^T M_*^{-1/2}(\overline{\theta}) \Big\{ \frac{1}{n} \sum_{i=1}^n \nabla\mu(X_i, \overline{\theta})\, \nabla^T \mu(X_i, \overline{\theta}) \Big\} M_*^{-1/2}(\overline{\theta})\, v \\ &= v^T M_*^{-1/2}(\overline{\theta})\, M(\xi_n, \overline{\theta})\, M_*^{-1/2}(\overline{\theta})\, v \xrightarrow{\text{a.s.}} 1, \end{aligned}$$

where the final convergence is implied by $M(\xi_n, \overline{\theta}) \xrightarrow{\text{a.s.}} M_*(\overline{\theta})$, see Remark 3 above. This proves (a) and hence (A.8). Next we show that

$$\begin{aligned} & n^{-1/2} \sum_{i=1}^n \big\{ \mu(X_i, \widehat{\theta}_n^{(\mathrm{LS})}) - \mu(X_i, \overline{\theta}) \big\}\, \nabla\mu(X_i, \widehat{\theta}_n^{(\mathrm{LS})}) \\ &= \big\{ M(\xi_n, \widehat{\theta}_n^{(\mathrm{LS})}) + A_n \big\} \big\{ n^{1/2} \big(\widehat{\theta}_n^{(\mathrm{LS})} - \overline{\theta} \big) \big\}, \\ &\quad \text{with a sequence} A_n, n \ge n_{\mathrm{st}}, \text{ of random } p \times p \text{ matrices such that } A_n \xrightarrow{\text{a.s.}} 0. \end{aligned} \tag{A.9}$$

By the mean value theorem, for each n there are (random) points $\widetilde{\theta}_{i,n}$, $1 \le i \le n$, on the line segment joining $\widehat{\theta}_n^{(\mathrm{LS})}$ and $\overline{\theta}$ such that

$$\mu(X_i, \widehat{\theta}_n^{(\mathrm{LS})}) - \mu(X_i, \overline{\theta}) \;=\; \nabla^T \mu(X_i, \widetilde{\theta}_{i,n}) \left(\widehat{\theta}_n^{(\mathrm{LS})} - \overline{\theta}\right).$$

So we can write, again using (B5),

$$\begin{aligned}
& n^{-1/2} \sum_{i=1}^{n} \left\{\mu(X_i, \widehat{\theta}_n^{(\mathrm{LS})}) - \mu(X_i, \overline{\theta})\right\} \nabla \mu(X_i, \widehat{\theta}_n^{(\mathrm{LS})}) \\
&= n^{-1} \sum_{i=1}^{n} \nabla \mu(X_i, \widehat{\theta}_n^{(\mathrm{LS})}) \, \nabla^T \mu(X_i, \widetilde{\theta}_{i,n}) \left\{n^{1/2} \left(\widehat{\theta}_n^{(\mathrm{LS})} - \overline{\theta}\right)\right\} \\
&= \left[M(\xi_n, \widehat{\theta}_n^{(\mathrm{LS})}) \;+\; \frac{1}{n} \sum_{i=1}^{n} \nabla \mu(X_i, \widehat{\theta}_n^{(\mathrm{LS})}) \left\{\nabla \mu(X_i, \widetilde{\theta}_{i,n}) - \nabla \mu(X_i, \widehat{\theta}_n^{(\mathrm{LS})})\right\}^T\right] \\
&\quad \left\{n^{1/2} (\widehat{\theta}_n^{(\mathrm{LS})} - \overline{\theta})\right\}.
\end{aligned}$$

For $A_n := \frac{1}{n} \sum_{i=1}^{n} \nabla \mu(X_i, \widehat{\theta}_n^{(\mathrm{LS})}) \left\{\nabla \mu(X_i, \widetilde{\theta}_{i,n}) - \nabla \mu(X_i, \widehat{\theta}_n^{(\mathrm{LS})})\right\}^T$ we get, using the Frobenius norm in the space of $p \times p$ matrices, i.e., $\|A\|_{\mathrm{F}} = \left[\mathrm{trace}(AA^T)\right]^{1/2}$,

$$\begin{aligned}
\|A_n\|_{\mathrm{F}} \;&\le\; \frac{1}{n} \sum_{i=1}^{n} \left\|\nabla \mu(X_i, \widehat{\theta}_n^{(\mathrm{LS})}) \left\{\nabla \mu(X_i, \widetilde{\theta}_{i,n}) - \nabla \mu(X_i, \widehat{\theta}_n^{(\mathrm{LS})})\right\}^T\right\|_{\mathrm{F}} \\
&= \frac{1}{n} \sum_{i=1}^{n} \left\|\nabla \mu(X_i, \widehat{\theta}_n^{(\mathrm{LS})})\right\| \cdot \left\|\nabla \mu(X_i, \widetilde{\theta}_{i,n}) - \nabla \mu(X_i, \widehat{\theta}_n^{(\mathrm{LS})})\right\|,
\end{aligned}$$

where we have used that $\left\|vw^T\right\|_{\mathrm{F}} = \|v\| \cdot \|w\|$ for $v, w \in \mathbb{R}^p$. By compactness of $\mathcal{X} \times \overline{B}$ and uniform continuity of $\nabla \mu(x, \theta)$ on $\mathcal{X} \times \overline{B}$, one has

$$\overline{c} \;:=\; \sup_{\theta \in \overline{B},\, x \in \mathcal{X}} \left\|\nabla \mu(x, \theta)\right\| \;< \infty.$$

From $\max_{1 \le i \le n} \|\widehat{\theta}_n^{(\mathrm{LS})} - \widetilde{\theta}_{i,n}\| \le \|\widehat{\theta}_n^{(\mathrm{LS})} - \overline{\theta}\| \xrightarrow{\text{a.s.}} 0$ (as $n \to 0$) and, again, by the uniform continuity of $\nabla \mu(x, \theta)$ on $\mathcal{X} \times \overline{B}$, one gets

$$\|A_n\|_{\mathrm{F}} \le \overline{c} \max_{1 \le i \le n} \left\|\nabla \mu(X_i, \widetilde{\theta}_{i,n}) - \nabla \mu(X_i, \widehat{\theta}_n^{(\mathrm{LS})})\right\| \;\xrightarrow{\text{a.s.}}\; 0$$

which proves (A.9). Next we show that

$$n^{-1/2} \sum_{i=1}^{n} e_i \left\{\nabla \mu(X_i, \widehat{\theta}_n^{(\mathrm{LS})}) - \nabla \mu(X_i, \overline{\theta})\right\} \;=\; B_n \left\{n^{-1/2} \left(\widehat{\theta}_n^{(\mathrm{LS})} - \overline{\theta}\right)\right\}, \quad \text{(A.10)}$$

with a sequence B_n of $p \times p$ random matrices such that $B_n \xrightarrow{\text{a.s.}} 0$.

Let $v \in \mathbb{R}^p$ be arbitrarily given. We can write, applying the mean value theorem,

$$\begin{aligned} & v^T\Big[n^{-1/2}\sum_{i=1}^{n} e_i\big\{\nabla\mu(X_i,\widehat{\theta}_n^{(\mathrm{LS})}) - \nabla\mu(X_i,\overline{\theta})\big\}\Big] \\ &= n^{-1/2}\sum_{i=1}^{n} e_i\big\{v^T\nabla\mu(X_i,\widehat{\theta}_n^{(\mathrm{LS})}) - v^T\nabla\mu(X_i,\overline{\theta})\big\} \\ &= n^{-1/2}\sum_{i=1}^{n} e_i\, v^T\nabla^2\mu\big(X_i,\widetilde{\theta}_{i,n}(v)\big)\big(\widehat{\theta}_n^{(\mathrm{LS})} - \overline{\theta}\big) \\ &= \frac{1}{n}\sum_{i=1}^{n} e_i\, v^T\nabla^2\mu\big(X_i,\widetilde{\theta}_{i,n}(v)\big)\big\{n^{1/2}\big(\widehat{\theta}_n^{(\mathrm{LS})} - \overline{\theta}\big)\big\}, \end{aligned} \tag{A.11}$$

where $\nabla^2\mu(x,\theta)$ denotes the Hessian matrix (matrix of second partial derivatives) of μ w.r.t. θ for fixed $x \in \mathcal{X}$, and $\widetilde{\theta}_{i,n}(v)$, $1 \le i \le n$, are suitable (random) points on the line segment joining $\widehat{\theta}_n^{(\mathrm{LS})}$ and $\overline{\theta}$. Let $b_n(v) := \frac{1}{n}\sum_{i=1}^{n} e_i\nabla^2\mu\big(X_i,\widetilde{\theta}_{i,n}(v)\big)\, v$ and write $b_n(v) = b_n^{(1)}(v) + b_n^{(2)}(v)$, where

$$\begin{aligned} b_n^{(1)}(v) &:= \frac{1}{n}\sum_{i=1}^{n} e_i\nabla^2\mu(X_i,\overline{\theta})\, v \quad \text{and} \quad b_n^{(2)}(v) \\ b_n^{(2)}(v) &:= \frac{1}{n}\sum_{i=1}^{n} e_i\big\{\nabla^2\mu\big(X_i,\widetilde{\theta}_{i,n}(v)\big)\, v - \nabla^2\mu(X_i,\overline{\theta})\, v\big\}. \end{aligned}$$

Applying Lemma 3.1 (b) in Freise et al. (2019) to each component of $b_n^{(1)}(v)$ one gets $b_n^{(1)}(v) \xrightarrow{\text{a.s.}} 0$. The uniform continuity of $(x,\theta) \mapsto \nabla^2\mu(x,\theta)\, v$ on $\mathcal{X} \times \overline{B}$ and $\max_{1\le i\le n}\|\widetilde{\theta}_{i,n}(v) - \overline{\theta}\| \le \|\widehat{\theta}_n^{(\mathrm{LS})} - \overline{\theta}\| \xrightarrow{\text{a.s.}} 0$ imply that $\max_{1\le i\le n}\big\|\nabla^2\mu\big(X_i,\widetilde{\theta}_{i,n}(v)\big)\, v - \nabla^2\mu(X_i,\overline{\theta})\, v\big\| \xrightarrow{\text{a.s.}} 0$. By Lemma 3.1 (a) in Freise et al. (2019), $\limsup_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n}|e_i| < \infty$ a.s., and hence

$$\|b_n^{(2)}(v)\| \le \max_{1\le i\le n}\big\|\nabla^2\mu\big(X_i,\widetilde{\theta}_{i,n}(v)\big)\, v - \nabla^2\mu(X_i,\overline{\theta})\, v\big\|\,\frac{1}{n}\sum_{i=1}^{n}|e_i| \xrightarrow{\text{a.s.}} 0.$$

Observing (A.11) we have thus obtained that for every $v \in \mathbb{R}^p$

$$v^T\Big[n^{-1/2}\sum_{i=1}^{n} e_i\big\{\nabla\mu(X_i,\widehat{\theta}_n^{(\mathrm{LS})}) - \nabla\mu(X_i,\overline{\theta})\big\}\Big] = b_n^T(v)\big\{n^{1/2}\big(\widehat{\theta}_n^{(\mathrm{LS})} - \overline{\theta}\big)\big\},$$

where $b_n(v) \xrightarrow{\text{a.s.}} 0$. Specializing to the elementary unit vectors $v^{(\ell)}$, $1 \le \ell \le p$, and taking the matrix B_n with rows $b_n^T(v^{(\ell)})$, $1 \le \ell \le p$, one gets (A.10). So, by (A.7),

(A.8), (A.9), and (A.10) one gets

$$\sigma^{-1}(\overline{\theta})\, M_*^{-1/2}(\overline{\theta}) \left\{ M(\xi_n, \widehat{\theta}_n^{(\mathrm{LS})}) + A_n - B_n \right\} \left\{ \sqrt{n}\big(\widehat{\theta}_n^{(\mathrm{LS})} - \overline{\theta}\big) \right\} \xrightarrow{\mathrm{d}} \mathrm{N}(0, I_p),$$

where $A_n \xrightarrow{\text{a.s.}} 0$ and $B_n \xrightarrow{\text{a.s.}} 0$. According to Remark 3 one has $M(\xi_n, \widehat{\theta}_n^{(\mathrm{LS})}) \xrightarrow{\text{a.s.}} M_*(\overline{\theta})$, and using standard properties of convergence in distribution one gets

$$\sqrt{n}\big(\widehat{\theta}_n^{(\mathrm{LS})} - \overline{\theta}\big) \xrightarrow{\mathrm{d}} \mathrm{N}\big(0, \sigma^2(\overline{\theta})\, M_*^{-1}(\overline{\theta})\big).$$

□

References

Atkinson AC, Fedorov VV, Herzberg AM, Zhang R (2014) Elemental information matrices and optimal experimental design for generalized regression models. J Stat Plan Inference 144:81–91

Bates DM, Watts DG (1988) Nonlinear regression analysis and its applications. Wilew, Ney York

Chen K, Hu I, Ying Z (1999) Strong consistency of maximum quasi-likelihood estimators in generalized linear models with fixed and adaptive designs. Ann Stat 27:1155–1163

Dette H, Bornkamp B, Bretz F (2013) On the efficiency of two-stage response-adaptive designs. Stat Med 32:1646–1660

Ford I, Torsney B, Wu CFJ (1992) The use of a canonical form in the construction of locally optimal designs for non-linear problems. J R Stat Soc B 54(2):569–583

Freise F (2016) On convergence of the maximum likelihood estimator in adaptive designs. Dissertation, University of Magdeburg

Freise F, Gaffke N, Schwabe R (2019) The adaptive Wynn-algorithm in generalized linear models with univariate response. arXiv:1907.02708 [math.ST]

Freise F, Gaffke N, Schwabe R (2020) A p-step-ahead sequential adaptive algorithm for D-optimal nonlinear regression design. Technical report, University of Magdeburg

Hall P, Heyde CC (1980) Martingale limit theory and its application. Academic Press, New York

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

Karlin S (1968) Total positivity. Stanford University Press, Stanford

Lai TL (1994) Asymptotic properties of nonlinear least squares estimates in stochastic regression models. Ann Stat 22:1917–1930

Lai TL, Wei CZ (1982) Least squares estimates in stochastic regression models with applications to identification and control of dynamic systems. Ann Stat 10:154–166

Lane A, Yao P, Flournoy N (2014) Information in a two-stage adaptive optimal design. J Stat Plan Inference 144:173–187

Müller WG, Pötscher BM (1992) Batch sequential design for a nonlinear estimation problem. In: Fedorov VV, Vuchkov TN (eds) Model oriented data-analysis. Physika-Verlag, Heidelberg, pp 77–87

Pronzato L (2009) Asymptotic properties of nonlinear estimates in stochastic models with finite design space. Stat Probab Lett 79:2307–2313

Pronzato L (2010) One-step ahead adaptive D-optimal design on a finite design space is asymptotically optimal. Metrika 71:219–238

R Core Team (2020) R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. **http://www.R-project.org/**

Wu C-F (1981) Asymptotic theory of nonlinear least-squares estimation. Ann Stat 9:501–513

Wynn H (1970) The sequential generation of D-optimum experimental designs. Ann Math Stat 5:1655–1664