

Asymptotic Properties of Nonlinear Least Squares Estimates in Stochastic Regression Models Over a Finite Design Space. Application to Self-Tuning Optimisation ^{*}

Luc Pronzato ^{*}

^{*} *Laboratoire I3S, CNRS/Université de Nice-Sophia Antipolis, Bât Euclide, Les Algorithmes, 2000 route des lucioles, BP 121, 06903 Sophia Antipolis cedex, France (e-mail: pronzato@i3s.unice.fr)*

Abstract: We present new conditions for the strong consistency and asymptotic normality of the least squares estimator in nonlinear stochastic models when the design variables vary in a finite set. The application to self-tuning optimisation is considered, with a simple adaptive strategy that guarantees simultaneously the convergence to the optimum and the strong consistency of the estimates of the model parameters. An illustrative example is presented.

Keywords: Optimal design of experiments; self-tuning optimisation; extremum seeking; penalized optimal design; sequential design; consistency; asymptotic normality.

1. INTRODUCTION

Consider a stochastic regression model with observations

$$Y_k = \eta(\mathbf{x}_k, \bar{\theta}) + \varepsilon_k, \quad k = 1, 2, \dots \quad (1)$$

where $\{\varepsilon_k\}$ is a sequence of i.i.d. random variables with $\mathbb{E}(\varepsilon_1) = 0$ and $\mathbb{E}(\varepsilon_1^2) = \sigma^2 < \infty$, $\{\mathbf{x}_k\}$ is a sequence of design points in $\mathcal{X} \subset \mathbb{R}^d$ and $\eta(\mathbf{x}, \theta)$ is a known function of \mathbf{x} and parameter vector $\theta \in \Theta$, a compact subset of \mathbb{R}^p , with $\bar{\theta}$, the true unknown value of θ , such that $\bar{\theta} \in \text{int}(\Theta)$. We denote \mathcal{F}_k the σ -field generated by $\{Y_1, \dots, Y_k\}$ and assume that \mathbf{x}_k is \mathcal{F}_{k-1} measurable. This setup includes for instance the case of NARX models (nonlinear autoregressive models with exogeneous inputs) where $Y_k = \eta(Y_{k-1}, \dots, Y_{k-a}, u_{k-q}, \dots, u_{k-q-b}, \bar{\theta}) + \varepsilon_k$ where $a, b \in \mathbb{N}$, q is the delay and u_i is the input at stage i .

The unknown $\bar{\theta}$ will be estimated by Least Squares (LS) and we denote

$$S_n(\theta) = \sum_{k=1}^n [Y_k - \eta(\mathbf{x}_k, \theta)]^2$$

and $\hat{\theta}_{LS}^n = \arg \min_{\theta \in \Theta} S_n(\theta)$. We shall suppose that $\eta(\mathbf{x}, \theta)$ is continuously differentiable with respect to $\theta \in \text{int}(\Theta)$ for all $\mathbf{x} \in \mathcal{X}$ and denote $\mathbf{f}_\theta(\mathbf{x}) = \partial \eta(\mathbf{x}, \theta) / \partial \theta$ and

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

the information matrix for parameters θ and design measure ξ (a probability measure on \mathcal{X}). When ξ is the empirical measure ξ_k for $\mathbf{x}_1, \dots, \mathbf{x}_k$ we get the information matrix (normalized, per observation)

$$\mathbf{M}(\xi_k, \theta) = \frac{1}{k} \sum_{i=1}^k \mathbf{f}_\theta(\mathbf{x}_i) \mathbf{f}_\theta^\top(\mathbf{x}_i).$$

In the case of a linear regression model where

$$\eta(\mathbf{x}, \theta) = \mathbf{f}^\top(\mathbf{x})\theta, \quad \forall \mathbf{x} \in \mathcal{X}, \theta \in \Theta, \quad (2)$$

(so that $\mathbf{M}(\xi, \theta)$ does not depend on θ), Lai and Wei [1982] show that the conditions

$$\lambda_{\min}[n\mathbf{M}(\xi_n)] \xrightarrow{\text{a.s.}} \infty, \quad n \rightarrow \infty \quad (3)$$

$$\{\log \lambda_{\max}[n\mathbf{M}(\xi_n)]\}^\rho / \lambda_{\min}[n\mathbf{M}(\xi_n)] \xrightarrow{\text{a.s.}} 0, \quad n \rightarrow \infty \quad (4)$$

for some $\rho > 1$ are sufficient for the strong consistency of the LS estimator $\hat{\theta}_{LS}^n$ when $\{\varepsilon_k\}$ in (1) is a martingale difference sequence and $\sup_n \mathbb{E}(\varepsilon_n^2 | \mathcal{F}_{n-1}) < \infty$ a.s. The case of nonlinear stochastic regression models is considered in [Lai, 1994], where sufficient conditions for strong consistency are given, which reduce to (3) and the Christopheit and Helmes [1980] condition,

$$\lambda_{\max}[n\mathbf{M}(\xi_n)] = \mathcal{O}\{\lambda_{\min}^\rho[n\mathbf{M}(\xi_n)]\} \text{ for some } \rho \in (1, 2), \quad (5)$$

in the case of a linear model.

This paper gives new sufficient conditions for the strong consistency of $\hat{\theta}_{LS}^n$ in nonlinear stochastic models. These conditions, obtained under the assumption that $\{\mathbf{x}_k\}$ lives in a finite set, are much weaker than (3-4). The paper also gives conditions under which $\sqrt{n} \mathbf{M}^{1/2}(\xi_n, \hat{\theta}_{LS}^n)(\hat{\theta}_{LS}^n - \bar{\theta})$ converges in distribution to a normal random variable $\mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{I})$, with $\mathbf{0}$ and \mathbf{I} respectively the p -dimensional null vector and identity matrix. This means that $\mathbf{M}(\xi_n, \hat{\theta}_{LS}^n)$ can be used to characterize the asymptotic precision of the estimation of θ although the sequence of design points is stochastic. Conditions for strong consistency with a finite design space are given in Section 2 and conditions for asymptotic normality in Section 3. The application of

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

these results to self-tuning optimisation is considered in Section 4, where a comparison is made with the results in [Pronzato, 2000, Pronzato and Thierry, 2003]. A simple illustrative example is presented.

2. STRONG CONSISTENCY OF LS ESTIMATES WITH A FINITE DESIGN SPACE

Define

$$D_n(\theta, \bar{\theta}) = \sum_{k=1}^n [\eta(\mathbf{x}_k, \theta) - \eta(\mathbf{x}_k, \bar{\theta})]^2. \quad (6)$$

Next theorem shows that the strong consistency of the LS estimator is a consequence of $D_n(\theta, \bar{\theta})$ tending to infinity fast enough for $\theta \neq \bar{\theta}$. The fact that the design space \mathcal{X} is finite makes the required rate of increase for $D_n(\theta, \bar{\theta})$ quite slow.

Theorem 1. Let $\{\mathbf{x}_i\}$ be a design sequence on a finite set \mathcal{X} . If $D_n(\theta, \bar{\theta})$ given by (6) satisfies

$$\text{for all } \delta > 0, \left[\inf_{\|\theta - \bar{\theta}\| \geq \delta} D_n(\theta, \bar{\theta}) \right] / (\log \log n) \xrightarrow{\text{a.s.}} \infty, \quad (7)$$

then $\hat{\theta}_{LS}^n \xrightarrow{\text{a.s.}} \bar{\theta}$ as $n \rightarrow \infty$. If $D_n(\theta, \bar{\theta})$ simply satisfies

$$\text{for all } \delta > 0, \inf_{\|\theta - \bar{\theta}\| \geq \delta} D_n(\theta, \bar{\theta}) \xrightarrow{P} \infty, \quad (8)$$

then $\hat{\theta}_{LS}^n \xrightarrow{P} \bar{\theta}$ as $n \rightarrow \infty$.

The proof is given in [Pronzato, 2009] and is based on the following lemma from Wu [1981].

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

$$\liminf_{n \rightarrow \infty} \inf_{\|\theta - \bar{\theta}\| \geq \delta} [S_n(\theta) - S_n(\bar{\theta})] > 0 \text{ almost surely,} \quad (9)$$

then $\hat{\theta}_{LS}^n \xrightarrow{\text{a.s.}} \bar{\theta}$ as $n \rightarrow \infty$. If for any $\delta > 0$

$$\text{Prob} \left\{ \inf_{\|\theta - \bar{\theta}\| \geq \delta} [S_n(\theta) - S_n(\bar{\theta})] > 0 \right\} \rightarrow 1, \quad n \rightarrow \infty, \quad (10)$$

then $\hat{\theta}_{LS}^n \xrightarrow{P} \bar{\theta}$ as $n \rightarrow \infty$.

The condition [for all $\theta \neq \bar{\theta}$, $D_n(\theta, \bar{\theta}) \rightarrow \infty$ as $n \rightarrow \infty$] is sufficient for the strong consistency of $\hat{\theta}_{LS}^n$ when the parameter set Θ is finite, see Wu [1981]. From Theorem 1, when \mathcal{X} is finite this condition is also sufficient for the weak consistency of $\hat{\theta}_{LS}^n$ without restriction on Θ . It is proved in [Wu, 1981] to be necessary for the existence of a weakly consistent estimator of $\bar{\theta}$ in a regression model when the errors ε_i are independent with a distribution having a density $\varphi(\cdot)$ positive almost everywhere and absolutely continuous with respect to the Lebesgue measure and with finite Fisher information for location. Notice that a classical condition for strong consistency of LS estimates in nonlinear regression with non-random design is $D_n(\theta, \bar{\theta}) = \mathcal{O}(n)$ for $\theta \neq \bar{\theta}$, see e.g. Jennrich [1969], which is much stronger than (7). Also note that (7) is much less restrictive than the conditions (3-4) for stochastic designs in linear models and than the Christopheit and Helmes [1980] condition (5).

3. ASYMPTOTIC NORMALITY OF LS ESTIMATES WITH A FINITE DESIGN SPACE

Under a fixed design, the information matrix can be considered as a large sample approximation for the variance-covariance matrix of the estimator, thus allowing straightforward statistical inference from the trial. The situation is more complicated for adaptive designs and has been intensively discussed in the literature. The property below gives a simple sufficient condition in the situation where the \mathbf{x}_k 's in (1) belong to a finite set. We use the following regularity assumption for the model.

H_f: For all \mathbf{x} in \mathcal{X} , the components of $\mathbf{f}_\theta(\mathbf{x})$ are continuously differentiable with respect to θ in some open neighborhood of $\bar{\theta}$.

Theorem 3. Assume that $\hat{\theta}_{LS}^n$ is strongly consistent, that **H_f** is satisfied, that the design points belong to a finite set \mathcal{X} and that exists a sequence $\{\mathbf{C}_n\}$ of $p \times p$ deterministic matrices such that $\mathbf{C}_n^{-1} \mathbf{M}^{1/2}(\xi_n, \bar{\theta}) \xrightarrow{P} \mathbf{I}$, with $l_n = \lambda_{\min}(\mathbf{C}_n)$ satisfying $n^{1/4} l_n \rightarrow \infty$ and $\|\hat{\theta}_{LS}^n - \bar{\theta}\| / l_n^2 \xrightarrow{P} 0$ as $n \rightarrow \infty$. Then, $\hat{\theta}_{LS}^n$ satisfies

$$\sqrt{n} \mathbf{M}^{1/2}(\xi_n, \hat{\theta}_{LS}^n) (\hat{\theta}_{LS}^n - \bar{\theta}) \xrightarrow{d} \omega \sim \mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{I}) \quad (11)$$

as $n \rightarrow \infty$.

One may notice that, compared to Wu [1981], we do not require that $(n/\tau_n) \mathbf{M}(\xi_n, \bar{\theta})$ tends to a positive definite matrix for some $\tau_n \rightarrow \infty$ and, compared to Lai and Wei [1982], Lai [1994], we do not require the existence of high-order derivatives of $\eta(\mathbf{x}, \theta)$ w.r.t. θ . On the other hand, we need that $\lambda_{\min}(\mathbf{C}_n)$ decreases more slowly than $n^{-1/4}$.

4. APPLICATION TO SELF-TUNING OPTIMISATION

4.1 Problem statement

We consider a self-tuning optimisation problem where one wishes to minimize some function $\phi(\mathbf{x}, \theta)$ with respect to \mathbf{x} , the unknown parameters $\bar{\theta}$ being estimated from the observations in the model (1). One may have $\phi(\cdot, \cdot) = \eta(\cdot, \cdot)$ but this is not mandatory. In particular, less regularity is required for $\phi(\mathbf{x}, \cdot)$ than for $\eta(\mathbf{x}, \cdot)$ and we shall only use the following assumptions on ϕ .

H_φ-(i): $\phi(\mathbf{x}, \theta)$ is bounded for below and above for any $\mathbf{x} \in \mathcal{X}$ and $\theta \in \Theta$.

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

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

Note that compared to methods based on local gradient approximations, see, e.g. [Manzie and Krstić, 2007], the method is not restricted to a neighborhood of a local minimum of $\phi(\mathbf{x}, \theta)$. On the other hand, we shall assume that \mathbf{x} belongs to a finite space and $\phi(\mathbf{x}, \theta)$ must have a known parametric form.

When the problem is to estimate \mathbf{x}^* , one can resort to optimal design theory and choose the sequence $\{\mathbf{x}_k\}$ in

order to optimize a criterion that measures the precision of the estimation of θ in (1). For instance, one may use a nominal value θ^0 for θ , construct a D -optimal design measure $\xi_D^*(\theta^0)$ on \mathcal{X} maximizing $\log \det \mathbf{M}(\xi, \theta^0)$ and choose \mathbf{x}_k 's such that their empirical measure approaches $\xi_D^*(\theta^0)$. Alternatively, one may relate the design criterion to the estimation of $\mathbf{x}^*(\theta)$, see, e.g., [Chaloner, 1989, Pronzato and Walter, 1993].

In self-tuning optimisation, the design points form a sequence of control variables for the objective of minimizing $\sum_i \phi(\mathbf{x}_i, \theta)$. The trivial optimum solution $\mathbf{x}_i = \mathbf{x}^*(\theta)$ for all i is not feasible since θ is unknown, hence the dual aspect of the control: minimize the objective, help estimate θ . A naive approach (Forced-Certainty-Equivalence control) consists in replacing the unknown θ by its current estimated value at stage k , that is, $\mathbf{x}_{k+1} = \mathbf{x}^*(\hat{\theta}_{LS}^k)$. However, this does not provide enough excitation to estimate θ consistently, see Bozin and Zarrop [1991] for a detailed analysis of the special case $\phi(x, \theta) = \eta(x, \theta) = \theta_1 x + \theta_2 x^2$.

Here we shall consider the same approach as in [Pronzato, 2000, Pronzato and Thierry, 2003] and use

$$\mathbf{x}_{n+1} = \arg \min_{\mathbf{x} \in \mathcal{X}} \left\{ \phi(\mathbf{x}, \hat{\theta}_{LS}^n) - \alpha_n \mathbf{f}_{\hat{\theta}_{LS}^n}^\top(\mathbf{x}) \mathbf{M}^{-1}(\xi_n, \hat{\theta}_{LS}^n) \mathbf{f}_{\hat{\theta}_{LS}^n}(\mathbf{x}) \right\}, \quad (12)$$

with $\alpha_n > 0$; that is, to the current objective $\phi(\mathbf{x}, \hat{\theta}_{LS}^n)$, to be minimized with respect to \mathbf{x} , we add a penalty for poor estimation, $-\alpha_n \mathbf{f}_{\hat{\theta}_{LS}^n}^\top(\mathbf{x}) \mathbf{M}^{-1}(\xi_n, \hat{\theta}_{LS}^n) \mathbf{f}_{\hat{\theta}_{LS}^n}(\mathbf{x})$ (note that \mathbf{x}_{n+1} is \mathcal{F}_n -measurable). Iterations of the similar type can be used to generate an optimal design under a cost-constraint, the cost of an observation at the design point \mathbf{x} for parameters θ being measured by $\phi(\mathbf{x}, \theta)$, see [Pronzato, 2008]. See also Åström and Wittenmark [1989]. The results in Section 4.2 indicate that when \mathcal{X} is finite and the sequence of penalty coefficients α_n decreases slowly enough, the LS estimator $\hat{\theta}_{LS}^n$ is strongly consistent.

When $\hat{\theta}_{LS}^n$ is frozen to a fixed value θ and $\alpha_n \equiv \alpha$ constant, the iteration (12) corresponds to one step of a steepest descent vertex-direction algorithm for the minimisation of $\int_{\mathcal{X}} \phi(\mathbf{x}, \theta) \xi(d\mathbf{x}) - \alpha \log \det \mathbf{M}(\xi, \theta)$, with step-length $1/n$ at stage n . Convergence of the empirical measure ξ_n to an optimal design measure is proved in [Pronzato, 2000] using an argument developed in [Wu and Wynn, 1978]. It is also shown in the same reference that $\int_{\mathcal{X}} \phi(\mathbf{x}, \theta) \xi_n(d\mathbf{x}) \rightarrow \min_{\mathbf{x} \in \mathcal{X}} \phi(\mathbf{x}, \theta)$ as $n \rightarrow \infty$ when α_n decreases to zero and the sequence $\{n\alpha_n\}$ increases to infinity. If, moreover, H_ϕ - (iii) is satisfied then $\xi_n \xrightarrow{w} \delta_{\mathbf{x}^*(\theta)}$ (weak convergence of probability measures), with $\delta_{\mathbf{x}}$ the delta measure at \mathbf{x} . Those results do not require \mathcal{X} to be finite.

The fact that the parameters are estimated in (12) makes the proof of convergence a much more complicated issue. In the case of the linear regression model (2) it is shown in [Pronzato, 2000] that if $\{\alpha_n\}$ is such that $\alpha_n \log n$ decreases to zero and $n\alpha_n/(\log n)^{1+\delta}$ increases to infinity for some $\delta > 0$, then $\int_{\mathcal{X}} \phi(\mathbf{x}, \theta) \xi_n(d\mathbf{x}) \xrightarrow{a.s.} \min_{\mathbf{x} \in \mathcal{X}} \phi(\mathbf{x}, \theta)$ (and $\xi_n \xrightarrow{w} \delta_{\mathbf{x}^*(\theta)}$ a.s. if H_ϕ - (iii) is satisfied). Using Bayesian imbedding, the same properties are shown to hold in [Pronzato and Thierry, 2003] under the weaker conditions

$\alpha_n \rightarrow 0$ and $n\alpha_n \rightarrow \infty$ (however, the almost sure convergence then concerns the product measure $\mu \times Q$ with μ the prior measure for θ and Q the probability measure induced by $\{\varepsilon_k\}$). To the best of our knowledge, no similar result exists for nonlinear regression models.

When $\phi(\mathbf{x}, \theta) = 0$ for all \mathbf{x} the iteration (12) becomes

$$\mathbf{x}_{n+1} = \arg \max_{\mathbf{x} \in \mathcal{X}} \mathbf{f}_{\hat{\theta}_{LS}^n}^\top(\mathbf{x}) \mathbf{M}^{-1}(\xi_n, \hat{\theta}_{LS}^n) \mathbf{f}_{\hat{\theta}_{LS}^n}(\mathbf{x}), \quad (13)$$

which corresponds to one step in the sequential construction of a D -optimal design. Even for this particular situation, and although this method is widely used, very few asymptotic results are available: the developments in [Ford and Silvey, 1980, Wu, 1985, Müller and Pötscher, 1992] only concern a particular example; [Hu, 1998] is specific of Bayesian estimation by posterior mean and does not use a fully sequential design of the form (13); Lai [1994] and Chaudhuri and Mykland [1995] require the introduction of a subsequence of non-adaptive design points to ensure consistency of the estimator and Chaudhuri and Mykland [1993] require that the size of the initial experiment (non-adaptive) grows with the increase in size of the total experiment. Intuitively, the almost sure convergence of $\hat{\theta}_{LS}^n$ to some $\hat{\theta}^\infty$ would be enough to imply the convergence of ξ_n to a D -optimal design measure for $\hat{\theta}^\infty$ and, conversely, convergence of ξ_n to a design ξ_∞ such that $\mathbf{M}(\xi_\infty, \theta)$ is non-singular for any θ would be enough in general to make the estimator consistent. It is thus the interplay between estimation and design iterations (which implies that each design point depends on previous observations) that creates difficulties. As the results below will show, those difficulties disappear when \mathcal{X} is a finite set. Notice that the assumption that \mathcal{X} is finite is seldom limitative since practical considerations often impose such a restriction on possible choices for the design points; this can be contrasted with the much less natural assumption that would consist in considering the feasible parameter set as finite, see, e.g., Caines [1975].

The results below rely on simple arguments based on three ideas. First, we consider iterations of the form

$$\mathbf{x}_{n+1} = \arg \min_{\mathbf{x} \in \mathcal{X}} \left\{ \phi(\mathbf{x}, \hat{\theta}^n) - \alpha_n \mathbf{f}_{\hat{\theta}^n}^\top(\mathbf{x}) \mathbf{M}^{-1}(\xi_n, \hat{\theta}^n) \mathbf{f}_{\hat{\theta}^n}(\mathbf{x}) \right\}, \quad (14)$$

where $\{\hat{\theta}^n\}$ is taken as *any sequence* of vectors in Θ . The asymptotic design properties obtained within this framework thus also apply when $\hat{\theta}^n$ corresponds to $\hat{\theta}_{LS}^n$. Second, when \mathcal{X} is finite we obtain a lower bound on the sampling rate of a subset of points of \mathcal{X} associated with a nonsingular information matrix. Third, we can show that this bound guarantees the strong consistency of $\hat{\theta}_{LS}^n$. With a few additional technicalities, this yields almost sure convergence results for the adaptive designs constructed via (12).

4.2 Asymptotic properties of LS estimates and designs

We shall use the following assumptions on \mathcal{X} .

H_X- (i): \mathcal{X} is finite, $\mathcal{X} = \{\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \dots, \mathbf{x}^{(K)}\}$.

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

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

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

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

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

The case of sequential D -optimal design, corresponding to the iterations (13), is considered in [Pronzato, 2009]. When $\alpha_n \rightarrow \alpha > 0$ ($n \rightarrow \infty$) in (12), the results are similar to those in [Pronzato, 2009] and $\hat{\theta}_{LS}^n \xrightarrow{\text{a.s.}} \bar{\theta}$, $\mathbf{M}(\xi_n, \hat{\theta}_{LS}^n) \xrightarrow{\text{a.s.}} \mathbf{M}(\xi^*, \bar{\theta})$ with ξ^* minimizing $\int_{\mathcal{X}} \phi(\mathbf{x}, \bar{\theta}) \xi(d\mathbf{x}) - \alpha \log \det \mathbf{M}(\xi, \bar{\theta})$. One can take $\mathbf{C}_n = \mathbf{M}^{1/2}(\xi^*, \bar{\theta})$ for all n in Theorem 3 and $\hat{\theta}_{LS}^n$ is asymptotically normal.

The situation is more complicated when the sequence $\{\alpha_n\}$ in (12,14) satisfies the following:

$\mathbf{H}_{\alpha}\text{-}(i)$: $\{\alpha_n\}$ is a non-increasing positive sequence tending to zero as $n \rightarrow \infty$,

the situation considered in the rest of the paper. We then obtain the following lower bound on the sampling rate of nonsingular designs.

Lemma 4. Let $\{\hat{\theta}^n\}$ be an arbitrary sequence in Θ used to generate design points according to (14) in a design space satisfying $\mathbf{H}_{\mathcal{X}}\text{-}(i)$, $\mathbf{H}_{\mathcal{X}}\text{-}(ii)$, with an initialisation such that $\mathbf{M}(\xi_n, \theta)$ is non-singular for all θ in Θ and all $n \geq p$. Let $r_{n,i} = r_n(\mathbf{x}^{(i)})$ denote the number of times $\mathbf{x}^{(i)}$ appears in the sequence $\mathbf{x}_1, \dots, \mathbf{x}_n$, $i = 1, \dots, K$, and consider the associated order statistics $r_{n,1:K} \geq r_{n,2:K} \geq \dots \geq r_{n,K:K}$. Define

$$q^* = \max\{j : \exists \beta > 0 \mid \liminf_{n \rightarrow \infty} r_{n,j:K} / (n\alpha_n) > \beta\}.$$

Then, $\mathbf{H}_{\phi}\text{-}(i)$ and $\mathbf{H}_{\alpha}\text{-}(i)$ imply $q^* \geq p$ with probability one.

For any sequence $\{\hat{\theta}^n\}$ used in (14), the conditions of Lemma 4 ensure the existence of N_1 and $\beta > 0$ such that $r_{n,j:K} > \beta n \alpha_n$ for all $n > N_1$ and all $j = 1, \dots, p$. Under the additional assumption $\mathbf{H}_{\mathcal{X}}\text{-}(iii)$ we thus obtain that $D_n(\theta, \bar{\theta})$ given by (6) satisfies

$$\frac{1}{\log \log n} \inf_{\|\theta - \bar{\theta}\| \geq \delta} D_n(\theta, \bar{\theta}) > \frac{\beta n \alpha_n \epsilon(\delta)}{\log \log n}, \quad n > N_1.$$

Therefore, if $n\alpha_n / \log \log n \rightarrow \infty$ as $n \rightarrow \infty$, $\hat{\theta}_{LS}^n \xrightarrow{\text{a.s.}} \bar{\theta}$ from Theorem 1. Since this holds for any sequence $\{\hat{\theta}^n\}$ in Θ , it is true in particular when $\hat{\theta}_{LS}^n$ is substituted for $\hat{\theta}^n$ in (14). It thus holds for (12).

Using the following assumption

$\mathbf{H}_{\alpha}\text{-}(ii)$: the sequence $\{\alpha_n\}$ is such that $n\alpha_n$ is non-decreasing with $n\alpha_n / \log \log n \rightarrow \infty$ as $n \rightarrow \infty$;

in complement of $\mathbf{H}_{\alpha}\text{-}(i)$, one can show that the adaptive design algorithm (12) is such that $\{\mathbf{x}_n\}$ tends to accumulate at the point of minimum cost for θ .

Theorem 5. Suppose that in the regression model (1) the design points for $n > p$ are generated sequentially according to (12), where α_n satisfies $\mathbf{H}_{\alpha}\text{-}(i)$ and $\mathbf{H}_{\alpha}\text{-}(ii)$. Suppose, moreover, that the first p design points are such that the information matrix is nonsingular for any $\theta \in \Theta$. Then, under $\mathbf{H}_{\mathcal{X}}\text{-}(i-iv)$, $\mathbf{H}_{\phi}\text{-}(i)$ and $\mathbf{H}_{\phi}\text{-}(ii)$ we have $\hat{\theta}_{LS}^n \xrightarrow{\text{a.s.}} \bar{\theta}$ and

$$\int_{\mathcal{X}} \phi(\mathbf{x}, \bar{\theta}) \xi_n(d\mathbf{x}) \xrightarrow{\text{a.s.}} \min_{\mathbf{x} \in \mathcal{X}} \phi(\mathbf{x}, \bar{\theta}), \quad n \rightarrow \infty. \quad (15)$$

If, moreover, $\mathbf{H}_{\phi}\text{-}(iii)$ is satisfied, then

$$\xi_n \xrightarrow{w} \delta_{\mathbf{x}^*(\bar{\theta})} \text{ almost surely, } n \rightarrow \infty. \quad (16)$$

Remark 6. Notice that the condition on the rate of decrease of $\{\alpha_n\}$ in Theorem 5 is weaker than in [Pronzato, 2000] although the model is nonlinear. Also note that using a penalty for poor estimation of the form

$$-C \frac{\det \left[\mathbf{f}_{\hat{\theta}_{LS}^n}(\mathbf{x}) \mathbf{f}_{\hat{\theta}_{LS}^n}^\top(\mathbf{x}) + \sum_{k=1}^n \mathbf{f}_{\hat{\theta}_{LS}^n}(\mathbf{x}_k) \mathbf{f}_{\hat{\theta}_{LS}^n}^\top(\mathbf{x}_k) \right]}{\det \left[\sum_{k=1}^n \mathbf{f}_{\hat{\theta}_{LS}^n}(\mathbf{x}_k) \mathbf{f}_{\hat{\theta}_{LS}^n}^\top(\mathbf{x}_k) \right]}, \quad C > 0,$$

as suggested in [Åström and Wittenmark, 1989] is equivalent to taking $\alpha_n = C/n$ in (12), which does not satisfy $\mathbf{H}_{\alpha}\text{-}(ii)$ and therefore does not guarantee the strong consistency of the LS estimates.

Remark 7. The property (16) does not imply that the \mathbf{x}_k 's generated by (12) converge to $\mathbf{x}^*(\bar{\theta})$. However, the following property is proved in [Pronzato, 2008]. Suppose that \mathcal{X}' is obtained by the discretization of a compact set \mathcal{X} and define $\underline{\mathbf{x}} = \arg \min_{\mathbf{x} \in \mathcal{X}'} \phi(\mathbf{x}, \bar{\theta})$ and, for any design measure ξ on \mathcal{X}' , $\Delta_{\bar{\theta}}(\xi) = \int_{\mathcal{X}'} \phi(\mathbf{x}, \bar{\theta}) \xi(d\mathbf{x}) - \phi(\underline{\mathbf{x}}, \bar{\theta})$. Suppose that there exist designs measures ξ_α on \mathcal{X}' such that $\Delta_{\bar{\theta}}(\xi_\alpha) \geq p\alpha$ and for all $\epsilon > 0$

$$\limsup_{\alpha \rightarrow 0^+} \sup_{\mathbf{x} \in \mathcal{X}', \|\mathbf{x} - \underline{\mathbf{x}}\| > \epsilon} \frac{2\Delta_{\bar{\theta}}(\xi_\alpha) [\mathbf{f}_{\bar{\theta}}^\top(\mathbf{x}) \mathbf{M}^{-1}(\xi_\alpha, \bar{\theta}) \mathbf{f}_{\bar{\theta}}(\mathbf{x})]}{\phi(\mathbf{x}, \bar{\theta}) - \phi(\underline{\mathbf{x}}, \bar{\theta})} < 1.$$

Then the supporting points of an optimal design measure minimizing $\int_{\mathcal{X}'} \phi(\mathbf{x}, \bar{\theta}) \xi(d\mathbf{x}) - \alpha \log \det \mathbf{M}(\xi, \bar{\theta})$ converge to $\underline{\mathbf{x}}$ as $\alpha \rightarrow 0^+$, and, under the conditions of Theorem 5, the design sequence $\{\mathbf{x}_n\}$ on \mathcal{X} will concentrate around $\mathbf{x}^*(\bar{\theta})$ as $n \rightarrow \infty$.

Remark 8. Under the conditions of Theorem 5, there exist N_0 and $\beta > 0$ such that, for all $n > N_0$, $\lambda_{\min}[\mathbf{M}(\xi_n, \bar{\theta})] > \beta \bar{\gamma} \alpha_n$, with $\bar{\gamma}$ as in $\mathbf{H}_{\mathcal{X}}\text{-}(iv)$. The asymptotic normality of $\hat{\theta}_{LS}^n$ is ensured if one can exhibit a sequence $\{\mathbf{C}_n\}$ satisfying the conditions of Theorem 3. For $\lambda_{\min}(\mathbf{C}_n) \sim \alpha_n^{1/2}$ we then obtain that imposing the condition $n^{1/3} \alpha_n \rightarrow \infty$ on the decrease rate of α_n would be enough. A possible construction is based on the matrix $\mathbf{M}^{1/2}(\nu_n, \bar{\theta})$ obtained when $\bar{\theta}$ is substituted for $\hat{\theta}_{LS}^n$ in the iterations (12). However, it remains to be proved that $\mathbf{M}^{-1}(\nu_n, \bar{\theta}) \mathbf{M}(\xi_n, \bar{\theta}) \xrightarrow{p} \mathbf{I}$, which is not obvious (notice that the sequence $\{\hat{\theta}_{LS}^n\}$ becomes highly correlated as n increases).

Remark 9. When the function to be minimized is the model response itself and, moreover, is linear with respect to θ , one can construct analytical approximate solutions for the self-tuning optimizer over a finite horizon N when using a Bayesian approach. In particular, one of the constructions proposed in [Pronzato and Thierry, 2003] is shown to be within $\mathcal{O}(\sigma^4)$ of the optimal solution of

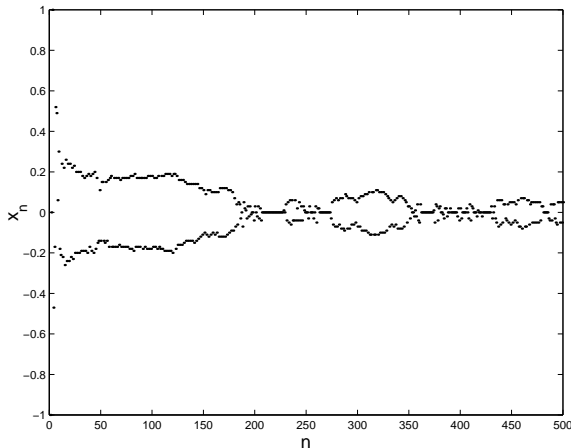


Fig. 1. A typical sequence $\{x_n\}$, $\alpha_n = (\log n)^{-4}$

the self-tuning optimizer problem (the latter being not computable, it corresponds to the solution of a stochastic dynamic programming problem).

4.3 Example

We take $\eta(x, \theta) = \theta_1 \theta_2 \theta_3 + \theta_2 x + \theta_3 (1 + \theta_1^2) x^2 + \theta_1^2 x^3$, $\phi(x, \theta) = \theta_2 + \theta_1 x^2 + \theta_2 \theta_3 x^4$, with $x \in [-1, 1]$ and $\theta = (\theta_1, \theta_2, \theta_3)^\top \in \mathbb{R}$. One can easily check that the model is structurally globally identifiable so that, if the sequence $\{x_k\}$ is rich enough, the LS estimator is unique and θ can be estimated consistently. For the true value $\bar{\theta}$ of the parameters we take $\bar{\theta} = (0, 1, 1)^\top$ which gives $\mathbf{f}_{\bar{\theta}}(x) = (1, x, x^2)^\top$ and $\phi(x, \bar{\theta}) = 1 + x^4$. The optimal design $\xi^*(\alpha)$ on $\mathcal{X}' = [-1, 1]$ that minimizes $\int_{\mathcal{X}'} \phi(x, \bar{\theta}) \xi(dx) - \alpha \log \det \mathbf{M}(\xi, \bar{\theta})$ can be constructed analytically for any α , see [Pronzato, 2008]. For $\alpha < 2/9$, the support points are $-\sqrt{3}(\alpha/2)^{1/4}, 0, \sqrt{3}(\alpha/2)^{1/4}$, with respective weights $1/6, 2/3, 1/6$, showing that $\xi^*(\alpha)$ concentrates around $0 = \arg \min_{x \in \mathcal{X}'} \phi(x, \bar{\theta})$ as α tends to zero.

Figure 1 shows a typical sequence $\{x_n\}$ generated by (12) for $n \geq 3$ when $\sigma = 1$, $x_1 = -1$, $x_2 = 0$, $x_3 = 1$, $\alpha_n = (\log n)^{-4}$ and \mathcal{X} consists of 201 points regularly spaced in $[-1, 1]$. The design points tend to concentrate around $x^*(\bar{\theta}) = 0$ as n increases. This is due to the fact that $\phi(x, \bar{\theta})$ is sufficiently flat around $x^*(\bar{\theta})$. The situation can be much different for other functions ϕ , see for instance the examples in [Pronzato, 2000, Pronzato and Thierry, 2003] where (16) is satisfied but design points are continuously generated far from $\mathbf{x}^*(\bar{\theta})$ (although less and less often as n increases).

Figure 2 presents the corresponding sequence $\{\phi(x_n, \bar{\theta})\}$, showing that convergence to the minimum value 1 is fast despite the model is nonlinear and the observations are very noisy, see Figure 3 for a plot of the sequence $\{Y_n\}$. The evolution of the parameter estimates $\hat{\theta}_{LS}^n$ is presented in Figure 4.

4.4 A concluding remark on the difficulties raised by dynamical systems

Compared to [Choi et al., 2002] where a periodic disturbance of magnitude α plays the role of a persistently

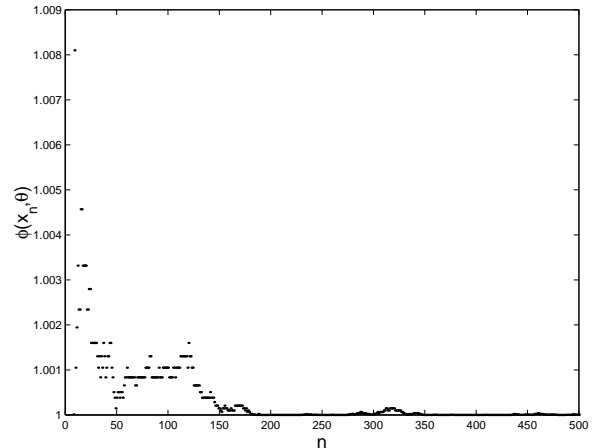


Fig. 2. Corresponding sequence $\{\phi(x_n, \bar{\theta})\}$, $n \geq 8$

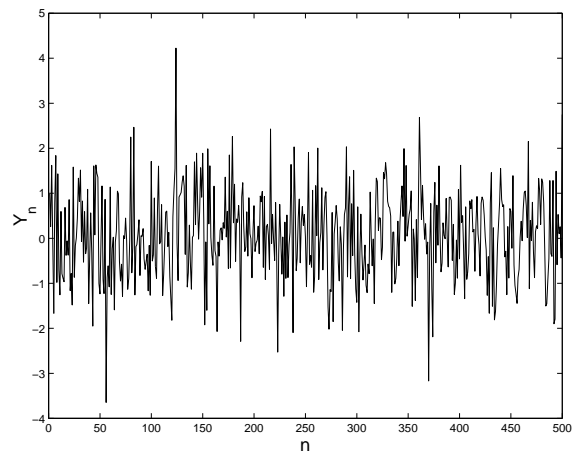


Fig. 3. Sequence of observations $\{Y_n\}$

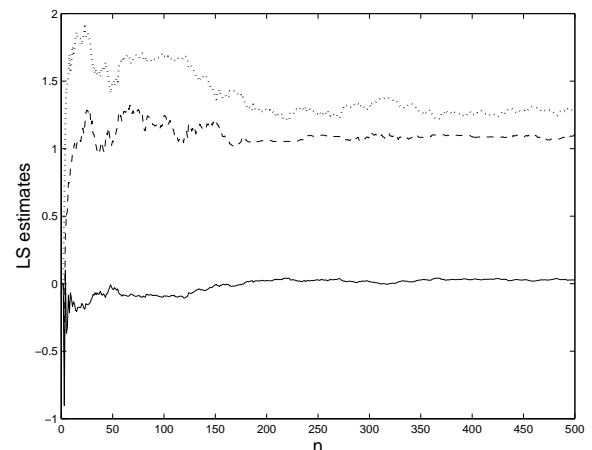


Fig. 4. LS estimates $\hat{\theta}_{LS}^n$ (solid line for θ_1 , dashed-line for θ_2 and dots for θ_3)

exciting input signal and the output converges to a neighborhood $\mathcal{O}(\alpha^2)$ of the optimum, iterations of the form (12) guarantee exact asymptotic convergence to the optimum when the excitation provided by the penalty for poor estimation vanishes slowly enough, see Theorem 5. However, (12) assumes that \mathbf{x}_{n+1} can be chosen freely in a given \mathcal{X} that does not depend on \mathbf{x}_n ; that is, it implicitly assumes that the system to be optimized is static (it is only the

fact that \mathbf{x}_{n+1} depends on $\hat{\theta}_{LS}^n$ that introduces a dynamic feedback), whereas Choi et al. [2002] consider self-tuning optimisation of a dynamic discrete-time system. (See also Krstić [2000], Krstić and Wang [2000] for continuous-time dynamic systems). Within the setup of Choi et al. [2002], it means that we need to observe the input of the static nonlinearity. This is a rather severe limitation, but one that seems difficult to overcome.

To illustrate the problem, consider the classical algorithm for the sequential construction of a D -optimal design with iterations of the form (13), see [Wynn, 1970]. Take the linear model $\eta(x, \theta) = \theta_1 x + \theta_2 x^2$, so that $\mathbf{f}_\theta(x) = \mathbf{f}(x) = (x, x^2)^\top$, and suppose that x_n may only vary within the interval $[-1, 1]$ by increments of $\pm\delta$ in one iteration (so that $x_{n+1} = x_n + u_n$) with $u_n \in \{-\delta, 0, \delta\}$, with $1/\delta$ integer. Also suppose that $\mathbf{M}(\xi_{n_0})$ is non singular for some n_0 and that $x_{n_0} = m\delta \in [0, 1]$ with m a strictly positive integer. One can easily check that the iterations

$$u_n = \arg \max_{u \in \{-\delta, 0, \delta\}} \mathbf{f}^\top(x_n + u_n) \mathbf{M}^{-1}(\xi_n) \mathbf{f}(x_n + u_n)$$

for $n \geq n_0$ do not yield convergence to the D -optimal design measure ξ_D^* (which allocates weights 1/2 at the extreme points ± 1). Indeed, negative values for x_n can only be reached if $x_k = 0$ is selected for some $k \geq n_0$, which is impossible since $\mathbf{f}(0) = \mathbf{0}$. Other types of iterations, perhaps less myopic than (12) and (13) which only look one step-ahead, should thus be considered for general dynamic systems.

5. CONCLUSIONS

Self-tuning optimisation has been considered through an approach based on sequential penalized optimal design. Simple conditions have been given that guarantee the strong consistency of the LS estimator in this context of sequentially determined control variables, under the assumption that they belong to a finite set and that the penalty for poor estimation does not decrease too fast.

REFERENCES

- K.J. Åström and B. Wittenmark. *Adaptive Control*. Addison Wesley, 1989.
- A.S. Bozin and M.B. Zarrop. Self tuning optimizer — convergence and robustness properties. In *Proc. 1st European Control Conf.*, pages 672–677, Grenoble, July 1991.
- P.E. Caines. A note on the consistency of maximum likelihood estimates for finite families of stochastic processes. *Annals of Statistics*, 3(2):539–546, 1975.
- K. Chaloner. Bayesian design for estimating the turning point of a quadratic regression. *Commun. Statist.-Theory Meth.*, 18(4):1385–1400, 1989.
- P. Chaudhuri and P.A. Mykland. Nonlinear experiments: optimal design and inference based likelihood. *Journal of the American Statistical Association*, 88(422):538–546, 1993.
- P. Chaudhuri and P.A. Mykland. On efficiently designing of nonlinear experiments. *Statistica Sinica*, 5:421–440, 1995.
- J.-Y. Choi, M. Krstić, K.B. Ariyur, and J.S. Lee. Extremum seeking control for discrete-time systems. *IEEE Transactions on Automatic Control*, 47(2):318–323, 2002.
- N. Christopeit and K. Helmes. Strong consistency of least squares estimators in linear regression models. *Annals of Statistics*, 8:778–788, 1980.
- I. Ford and S.D. Silvey. A sequentially constructed design for estimating a nonlinear parametric function. *Biometrika*, 67(2):381–388, 1980.
- I. Hu. On sequential designs in nonlinear problems. *Biometrika*, 85(2):496–503, 1998.
- R.I. Jennrich. Asymptotic properties of nonlinear least squares estimation. *Annals of Math. Stat.*, 40:633–643, 1969.
- M. Krstić. Performance improvement and limitations in extremum seeking control. *System & Control Letters*, 39:313–326, 2000.
- M. Krstić and H.-H. Wang. Stability of extremum seeking feedback for general nonlinear dynamic systems. *Automatica*, 36:595–601, 2000.
- T.L. Lai. Asymptotic properties of nonlinear least squares estimates in stochastic regression models. *Annals of Statistics*, 22(4):1917–1930, 1994.
- T.L. Lai and C.Z. Wei. Least squares estimates in stochastic regression models with applications to identification and control of dynamic systems. *Annals of Statistics*, 10(1):154–166, 1982.
- C. Manzie and M. Krstić. Discrete time extremum seeking using stochastic perturbations. In *Proceedings of the 46th Conf. on Decision and Control*, pages 3096–3101, New Orleans, Dec. 2007.
- W.G. Müller and B.M. Pötscher. Batch sequential design for a nonlinear estimation problem. In V.V. Fedorov, W.G. Müller, and I.N. Vuchkov, editors, *Model-Oriented Data Analysis II, Proceedings 2nd IIASA Workshop, St Kyrik (Bulgaria), May 1990*, pages 77–87. Physica Verlag, Heidelberg, 1992.
- L. Pronzato. One-step ahead adaptive D -optimal design on a finite design space is asymptotically optimal. *Metrika*, 2009. to appear.
- L. Pronzato. Penalized optimal designs for dose-finding. Technical Report I3S/RR-2008-18-FR, Laboratoire I3S, CNRS–Université de Nice-Sophia Antipolis, 06903 Sophia Antipolis, France, 2008. <http://www.i3s.unice.fr/mh/RR/rapports.html>.
- L. Pronzato. Adaptive optimisation and D -optimum experimental design. *Annals of Statistics*, 28(6):1743–1761, 2000.
- L. Pronzato and E. Thierry. Sequential experimental design and response optimisation. *Statistical Methods and Applications*, 11(3):277–292, 2003.
- L. Pronzato and E. Walter. Experimental design for estimating the optimum point in a response surface. *Acta Applicandae Mathematicae*, 33:45–68, 1993.
- C.F.J. Wu. Asymptotic theory of nonlinear least squares estimation. *Annals of Statistics*, 9(3):501–513, 1981.
- C.F.J. Wu. Asymptotic inference from sequential design in a nonlinear situation. *Biometrika*, 72(3):553–558, 1985.
- C.F.J. Wu and H.P. Wynn. The convergence of general step-length algorithms for regular optimum design criteria. *Annals of Statistics*, 6(6):1273–1285, 1978.
- H.P. Wynn. The sequential generation of D -optimum experimental designs. *Annals of Math. Stat.*, 41:1655–1664, 1970.