The Annals of Statistics
2012, Vol. 40, No. 6, 2850–2876
DOI: 10.1214/12-AOS1053

OPTIMAL TWO-STAGE PROCEDURES FOR ESTIMATING LOCATION AND SIZE OF THE MAXIMUM OF A MULTIVARIATE REGRESSION FUNCTION

BY EDUARD BELITSER, SUBHASHIS GHOSAL[1] AND HARRY VAN ZANTEN

Eindhoven University of Technology, North Carolina State University and University of Amsterdam

We propose a two-stage procedure for estimating the location $\boldsymbol{\mu}$ and size M of the maximum of a smooth d-variate regression function f. In the first stage, a preliminary estimator of $\boldsymbol{\mu}$ obtained from a standard nonparametric smoothing method is used. At the second stage, we "zoom-in" near the vicinity of the preliminary estimator and make further observations at some design points in that vicinity. We fit an appropriate polynomial regression model to estimate the location and size of the maximum. We establish that, under suitable smoothness conditions and appropriate choice of the zooming, the second stage estimators have better convergence rates than the corresponding first stage estimators of $\boldsymbol{\mu}$ and M. More specifically, for α-smooth regression functions, the optimal nonparametric rates $n^{-(\alpha-1)/(2\alpha+d)}$ and $n^{-\alpha/(2\alpha+d)}$ at the first stage can be improved to $n^{-(\alpha-1)/(2\alpha)}$ and $n^{-1/2}$, respectively, for $\alpha > 1 + \sqrt{1 + d/2}$. These rates are optimal in the class of all possible sequential estimators. Interestingly, the two-stage procedure resolves "the curse of the dimensionality" problem to some extent, as the dimension d does not control the second stage convergence rates, provided that the function class is sufficiently smooth. We consider a multi-stage generalization of our procedure that attains the optimal rate for any smoothness level $\alpha > 2$ starting with a preliminary estimator with any power-law rate at the first stage.

1. Introduction. In many applications, it is of interest to estimate the location and size of the extremum of a univariate or multivariate regression function. For instance, an oil company may be interested in determining the best location for drilling a well in a confined region. Based on information obtained from drilling at a few preliminary locations in the region, the goal is to obtain an estimate of the best location and the amount of the reserve based on these noisy measurements.

Suppose we observe noisy measurements of an unknown regression function $f:\mathbb{R}^d \to \mathbb{R}$, sampled at points from some compact, convex set $D \subset \mathbb{R}^d$,

$$Y_k = f(\mathbf{x}_k) + \xi_k, \qquad \mathbf{x}_k \in D \subset \mathbb{R}^d, \qquad k = 1, \dots, n, \tag{1}$$

Received September 2011; revised September 2012.

[1]Supported by a grant from NWO of the Netherlands. Part of this work was done when S. Ghosal was visiting EURANDOM, Eindhoven.

MSC2010 subject classifications. Primary 62L12; secondary 62G05, 62H12, 62L05.

Key words and phrases. Two-stage procedure, optimal rate, sequential design, multi-stage procedure, adaptive estimation.

where the ξ_k's are independent zero mean errors with $\mathrm{Var}(\xi_k) = \sigma^2$. Clearly, for estimating any feature of f, the estimation error increases with σ^2. Thus among all error distributions satisfying $\mathrm{Var}(\xi_k) \le \sigma^2$ for every k, the homoscedasticity condition $\mathrm{Var}(\xi_k) = \sigma^2$ is the least favorable. This shows that the latter condition can be relaxed to the former without increasing the bound on error of estimation, and the obtained rates under the homoscedasticity condition remains minimax optimal under the larger heteroscedastic model.

Assume that f has a unique maximum at $\boldsymbol{\mu}$ in the interior of D, that is,

$$\max_{\mathbf{x}\in D} f(\mathbf{x}) = f(\boldsymbol{\mu}) = M, \qquad f(\mathbf{x}) < f(\boldsymbol{\mu}) \qquad \text{for all } \mathbf{x} \neq \boldsymbol{\mu}. \tag{2}$$

If the function f is sufficiently smooth, then the gradient $\nabla f(\boldsymbol{\mu}) = 0$ and the Hessian matrix of f at $\boldsymbol{\mu}$ is nonpositive definite. The goal is to estimate the maximum of the regression function $M = f(\boldsymbol{\mu})$ and its location $\boldsymbol{\mu}$.

Clearly, the choice of the design points $\{\mathbf{x}_k, k = 1, \dots, n\}$ significantly influences the estimation accuracy. There are two basic design settings: fixed in advance (or randomly sampled from a chosen distribution) and sequential, where one is allowed to use the information obtained from an earlier sample to determine subsequent design points. If the design is fixed and nothing is known about the location of the maximum, the design points should be "almost uniformly" spread out all over the set of interest D. The problem of estimating the location and size of extrema of nonparametric regression functions for the fixed design situation has been studied by many authors. The one-dimensional case is thoroughly investigated, whereas the study in the multivariate situation has been limited; see Müller (1985, 1989), Shoung and Zhang (2001), Facer and Müller (2003) and the references therein. The minimax rate for estimating the maximum value of the function ranging over an α-smooth nonparametric class (e.g., isotropic Hölder class defined below) is $n^{-\alpha/(2\alpha+d)}$. As to the estimation of the location of the maximum, it is a folklore that the minimax rate is the same as the minimax rate for estimating the first derivative of the regression function, which is given by $n^{-(\alpha-1)/(2\alpha+d)}$. In the setting of estimating the mode μ of a univariate twice differentiable density f, Hasminskii (1979) showed that under the assumption that $f''(\mu) < 0$, the lower bound for the minimax risk rate is of the order $n^{-1/5}$ consistent with the rate $n^{-(\alpha-1)/(2\alpha+d)}$. Klemelä (2005) considered the problem of adaptive estimation of the mode of a multivariate density with a bounded support that satisfies, in a neighborhood of the mode, a smoothness condition of a level higher than 2.

If we can choose a design point before making each observation using the data obtained so far, then we are in the classical sequential design setting. Kiefer and Wolfowits (1952) introduced a Robbins–Monro type of algorithm to estimate the mode μ of f in the univariate framework. Blum (1954) proposed a multivariate version of their algorithm which allows to estimate the location $\boldsymbol{\mu}$ of the maximum of a multivariate regression function f. Since then, this Kiefer–Wolfowits–Blum recursive algorithm has been extended in many directions by many authors.

The main fact is that the algorithm converges to $\boldsymbol{\mu}$ with the rate $n^{-1/3}$ under the assumption that the regression function f is three times differentiable. More generally, Chen (1988) and Polyak and Tsybakov (1990) established that, in the sequential design setting, the minimax rate for estimating the location of the maximum of α-smooth regression functions is $n^{-(\alpha-1)/(2\alpha)}$. Dippon (2003) proposed a general class of randomized gradient recursive algorithms which attain the optimal convergence rate. Mokkadem and Pelletier (2007) considered the problem of simultaneously estimating, in the sequential design setting, the location and the size of the maximum of a regression function that is three times continuously differentiable. They proposed a companion recursive procedure to the Kiefer–Wolfowits–Blum algorithm so that, by applying both the companion and the Kiefer–Wolfowits–Blum algorithms, one can simultaneously estimate the location and size of the maximum of regression functions in an on-line regime. Interestingly, in a sequential design setting, the convergence rate for estimating the maximum itself $M = f(\boldsymbol{\mu})$ can, in principle, attain the parametric rate $n^{-1/2}$. The companion procedure of Mokkadem and Pelletier (2007) for estimating the maximum can also achieve the parametric rate $n^{-1/2}$, but this companion procedure must use different design points than those used in the Kiefer–Wolfowits–Blum procedure.

In this paper, we propose a two-stage strategy to tackle the problem of simultaneously estimating the location $\boldsymbol{\mu}$ and size M of the maximum of the regression function f according to the observation scheme (1). This is an approach in between the two above described frameworks—global fixed design and a fully sequential design. Often, from an operational point of view, fully sequential sampling can be expensive, whereas a two-stage procedure is much simpler to implement. Our findings establish that the two-stage procedure can be properly designed to match the strength of a fully sequential procedure. Moreover, the same design scheme can be used to obtain the optimal rates for estimating both $\boldsymbol{\mu}$ and M.

Now we describe the two-stage procedure. We construct a preliminary estimator $\tilde{\boldsymbol{\mu}}$ of $\boldsymbol{\mu}$ by spending a portion of our sampling budget to make observations over a relatively uniform grid of points in the area of interest and applying some standard nonparametric smoothing method for the fixed design setting based on this initial set of data. Additional prior information, if available, may also be used to reduce the span of the design points or to more efficiently choose design points leading to increased accuracy of the preliminary estimator. At the second stage, we "zoom-in" on a neighborhood of $\tilde{\boldsymbol{\mu}}$ of an appropriate size δ_n, to be called the *localization parameter*. The idea is that if this vicinity is "small enough," that is, the preliminary estimator $\tilde{\boldsymbol{\mu}}$ converges to $\boldsymbol{\mu}$, the regression function f can be accurately approximated by a Taylor polynomial within the vicinity of $\tilde{\boldsymbol{\mu}}$. We then spend the remaining portion of the sampling budget to gather further observations at appropriately chosen design points in the vicinity of $\tilde{\boldsymbol{\mu}}$. Finally, we fit a polynomial regression model on the new set of data and show that the remainder of the expansion is appropriately small, provided that the preliminary estimator $\tilde{\boldsymbol{\mu}}$ has sufficient accuracy. This procedure leads to improved estimators of $\boldsymbol{\mu}$ and M and does not use knowledge of the noise variance σ^2. The last step in our approach is reminiscent of the nonparametric methodology of local polynomial regression in case of fixed design setting; see Fan and Gijbels (1996). Our two-stage procedure is motivated by the recent work of Lan, Banerjee and Michailidis (2009) and Tang, Banerjee and Michailidis (2011), who, respectively, considered such procedures for estimating change points in a regression function and the level point of a univariate monotone regression function. Motivating grounds for a two-stage approach were nicely described by them. The principal differences between their and our techniques are that we consider smooth rather than step or monotone functions, and we use polynomial regression of an appropriate degree in the second stage rather than regression based on step or linear functions, respectively, used by them.

The results for estimating $\boldsymbol{\mu}$ and M under the fully sequential setting, which we are aware of, all follow the Robbins–Monro procedure, where the next design point depends only on the previous observation and does not incorporate all available information up to the current moment. In this setting, one makes observations only along a certain path of design points, eventually leading to the location of the maximum. In our two-stage approach, one also gets the global estimate of the regression function from the first stage all over the area of interest, which may be useful in some practical situations. We also get an accompanying estimator for the size of the maximum M (in fact, for all the relevant derivatives at the location of the maximum) in a natural way, while in a Robbins–Monro type sequential design, one needs to adjust the design points to estimate M. This can place serious constraints on the available budget since typically both $\boldsymbol{\mu}$ and M need to be estimated.

Our main result gives a decomposition of the convergence rate of the second stage estimator as the sum of an approximation term and a stochastic term, similar to the classical bias-variance trade-off. An implication of the main result is as follows. Suppose we take a preliminary nonparametric estimator $\tilde{\boldsymbol{\mu}}$ with the optimal single-stage convergence rate $n^{-(\alpha-1)/(2\alpha+d)}$. Then by applying our two-stage procedure with an appropriate choice of the localization parameter δ_n, we obtain optimal (for the sequential design setting) convergence rates, $n^{-(\alpha-1)/(2\alpha)}$ and $n^{-1/2}$, respectively, under the condition on the smoothness parameter $\alpha > 1 + \sqrt{1 + d/2}$. Note that $n^{-1/2}$ is also the "oracle rate" for estimating M corresponding to taking n samples at the "perfect location" $\boldsymbol{\mu}$. Thus, for α-smooth regression functions, the second stage improves the rates in estimating $\boldsymbol{\mu}$ and M from the nonparametric rates $n^{-(\alpha-1)/(2\alpha+d)}$ and $n^{-\alpha/(2\alpha+d)}$ to the optimal sequential rates $n^{-(\alpha-1)/(2\alpha)}$ and $n^{-1/2}$, respectively. Curiously, the dimension d disappears from powers in the second stage convergence rates. However, the curse of dimensionality is still present in a milder form through the constraint $\alpha > 1 + \sqrt{1 + d/2}$. For instance, if $\alpha > 3$, then the second stage rates are optimal for $d = 1, \dots, 6$. We can resolve the curse of dimensionality completely by considering a multi-stage generalization of the two-stage procedure, obtained by iterating the second stage operation on the estimator obtained in the second stage, and continuing the iteration sufficiently many times. We shall show that after an appropriate number of stages, the optimal convergence rates are attained for any $\alpha > 2$. In fact, even if we start with a not necessarily optimal preliminary estimator at the first stage (as long as it has a convergence rate of a power-law type), this multi-stage approach will lead to the optimal resulting stage after a finite number of stages. The number of stages depends on the smoothness of the regression function and the quality (convergence rate) of the preliminary estimator from the first stage. The method still uses knowledge of the smoothness level α in its formulation, and hence is not adaptive for estimating $\boldsymbol{\mu}$. Nevertheless, the multi-stage procedure achieves the optimal rate $n^{-1/2}$ for estimating M without using the knowledge of α.

The paper is organized as follows. In Section 2, we introduce the notation and assumptions. Section 3 describes the two-stage procedure and states the main result. The multi-stage generalization is discussed in Section 4, and some simulation results are given in Section 5. Proofs are presented in Section 6. Some auxiliary results are given in the Appendix.

2. Notation, preliminaries and assumptions. We describe the notation and conventions to be used in this paper. All asymptotic relations and symbols [like $O(\delta_n)$, $o(\delta_n)$, $O_p(\delta_n)$, $o_p(\delta_n)$ etc.] will refer to the asymptotic regime $n \to \infty$; here $c_n = O(\delta_n)$ [resp., $c_n = o(\delta_n)$] means that that c_n/δ_n is bounded (resp., $c_n/\delta_n \to 0$) and for a stochastic sequence X_n, $X_n = O_p(\delta_n)$ [resp., $X_n = o_p(\delta_n)$] means that that $\mathrm{P}\{|X_n| \le K c_n\} \to 1$ for some constant K (resp., $\mathrm{P}\{|X_n| < \varepsilon\delta_n\} \to 1$ for all $\varepsilon > 0$). For numerical sequences β_n and β_n', by $\beta_n \ll \beta_n'$ (or $\beta_n' \gg \beta_n$) we mean that $\beta_n = o(\beta_n')$, while by $\beta_n' \gtrsim \beta_n$, we mean that $\beta_n = O(\beta_n')$. By $\beta_n \asymp \beta_n'$ we mean that $\beta_n = O(\beta_n')$ and $\beta_n' = O(\beta_n)$. Let $\mathbb{N}$ stand for $\{0, 1, 2, \ldots\}$. For a set S, denote by $|S|$ the number of elements in S. Vectors are represented by bold symbols and can be upper or lowercase English or Greek letters. All vectors are in the column format with the corresponding nonbold letters with subscripts denoting the components, that is, for $\mathbf{x}, \mathbf{x}_k \in \mathbb{R}^d$, $\mathbf{x} = (x_1, \ldots, x_d)$ and $\mathbf{x}_k = (x_{k,1}, \ldots, x_{k,d})$. By $\|\mathbf{x}\|$ for a vector $\mathbf{x}$, we mean the usual Euclidean norm of $\mathbf{x} \in \mathbb{R}^d$. Matrices are also written in bold and only uppercase English letters are used to denote them. If $\mathbf{A}$ is a matrix, $\|\mathbf{A}\|$ will stand for a norm on the space of matrices such as the operator norm defined by $\|\mathbf{A}\| = \sup_{\|\mathbf{x}\| \le 1} \|\mathbf{A}\mathbf{x}\|$. Let $B(\mathbf{c}, R) = \{\mathbf{z} \in \mathbb{R}^d : \|\mathbf{z} - \mathbf{c}\| \le R\}$ denote a ball in $\mathbb{R}^d$ with center $\mathbf{c} \in \mathbb{R}^d$ and radius $R > 0$. Define a cube around a point $\mathbf{a} = (a_1, \ldots, a_d) \in \mathbb{R}^d$ with an edge length 2δ by

$$C(\mathbf{a}, \delta) = \{\mathbf{x} \in \mathbb{R}^d : x_k \in [a_k - \delta, a_k + \delta], k = 1, \ldots, d\} \subset \mathbb{R}^d. \tag{3}$$

If $\mathbf{a} = \mathbf{0}$, then we write $C(\delta)$ for $C(\mathbf{0}, \delta)$.

We shall use the multi-index notation $\mathbf{i} = (i_1, \ldots, i_d) \in \mathbb{N}^d$. For a multi-index $\mathbf{i}$, a vector $\mathbf{x} \in \mathbb{R}^d$ and a sufficiently smooth function f of d variables, define

$$|\mathbf{i}| = \sum_{k=1}^d i_k, \qquad \mathbf{i}! = \prod_{k=1}^d i_k!, \qquad \mathbf{x}^{\mathbf{i}} = \prod_{k=1}^d x_k^{i_k}, \qquad \mathbf{D}^{\mathbf{i}} f(\mathbf{x}_0) = \frac{\partial^{|\mathbf{i}|} f(\mathbf{x})}{\partial x_1^{i_1} \cdots \partial x_d^{i_d}}\Big|_{\mathbf{x}=\mathbf{x}_0}.$$

For $k, d, r \in \mathbb{N}$, define

$$\mathbb{I}_k(d) = \{\mathbf{i} \in \mathbb{N}^d : i_1 + \cdots + i_d = k\}, \qquad \mathbb{I}(r, d) = \bigcup_{k=0}^{r} \mathbb{I}_k(d),$$

with $\mathbb{I}_0(d) = \{(0, \ldots, 0)\}$. For convenience in writing, $\mathbb{I}(r, d)$ will be enumerated by stacking elements of $\mathbb{I}_0(d), \mathbb{I}_1(d), \ldots, \mathbb{I}_r(d)$, in that order. Within each $\mathbb{I}_k(d)$, the elements are arranged following the lexicographic (or dictionary) ordering. Observe that $\mathbb{I}_k(d)$ and $\mathbb{I}_l(d)$ introduced above are disjoint if $k \neq l$. The cardinality $|\mathbb{I}_k(d)|$ is the number of d-tuples $(k_1, \ldots, k_d) \in \mathbb{N}^d$ such that $k_1 + \cdots + k_d = k$, or equivalently, the number of ways to put k balls in d boxes. Thus $|\mathbb{I}_k(d)| = \binom{d+k-1}{k} = \binom{d+k-1}{d-1}$, and hence

$$|\mathbb{I}(r, d)| = \sum_{k=0}^{r} |\mathbb{I}_k(d)| = \sum_{k=0}^{r} \binom{d+k-1}{d-1}.$$

In particular, $|\mathbb{I}(r, 1)| = r + 1$.

For an $\alpha \in \mathbb{R}$, let $\lceil \alpha \rceil$ be the smallest integer bigger than or equal to α. Then $r_\alpha = \lceil \alpha - 1 \rceil$ stands for the largest integer which is strictly less than α. Clearly, if $\alpha \in \mathbb{N}$, then $r_\alpha = \alpha - 1$.

For $\alpha, L > 0$ and a compact, convex set $D \subseteq \mathbb{R}^d$, introduce an isotropic Hölder functional class $\mathcal{H}_d(\alpha, L, D)$, consisting of r_α-times differentiable functions $f : D \to \mathbb{R}$ such that

$$|f(\mathbf{x}) - P_{f,\mathbf{x}_0}(\mathbf{x})| \leq L\|\mathbf{x} - \mathbf{x}_0\|^\alpha, \qquad \mathbf{x}, \mathbf{x}_0 \in D, \tag{4}$$

where

$$P_{f,\mathbf{x}_0}(\mathbf{x}) = \sum_{\mathbf{i} \in \mathbb{I}(r_\alpha, d)} \frac{1}{\mathbf{i}!} \mathbf{D}^{\mathbf{i}} f(\mathbf{x}_0)(\mathbf{x} - \mathbf{x}_0)^{\mathbf{i}} \tag{5}$$

is the Taylor polynomial of order r_α obtained by expansion of f about the point $\mathbf{x}_0$.

Put $q(\alpha, d) = |\mathbb{I}(r_\alpha, d)| - 1$. Observe that the total number of terms in the d-variate Taylor polynomial $P_{f,\mathbf{x}_0}(\mathbf{x})$ of order r_α defined in (5) is $q(\alpha, d) + 1$.

For a function $g : \mathbb{R}^d \to \mathbb{R}$ such that all second-order partial derivatives of g exist at a point $\mathbf{x}_0 \in \mathbb{R}^d$, denote by $Hg(\mathbf{x}_0)$ the Hessian matrix of the function g at the point $\mathbf{x}_0$, whose (i, j)th entry is given by $\frac{\partial^2 g(\mathbf{x}_0)}{\partial x_j \partial x_i}$, $i, j = 1, \ldots, d$. Notice that if g has continuous second order partial derivatives at $\mathbf{x}_0$, then the Hessian matrix $Hg(\mathbf{x}_0)$ is symmetric, and hence its eigenvalues must be real. For a symmetric matrix $\mathbf{M}$, denote by $\lambda_{\min}(\mathbf{M})$ and $\lambda_{\max}(\mathbf{M})$ the smallest and the largest eigenvalues of $\mathbf{M}$, respectively.

Consider the model (1) with $f : D \to \mathbb{R}$. We now describe the assumptions on f to be used throughout the paper.

(A1) The function $f(\mathbf{x})$, $\mathbf{x} \in D \subseteq \mathbb{R}^d$, allows extension on a slightly bigger set $D^\varepsilon = \bigcup_{\mathbf{x}\in D} B(\mathbf{x}, \varepsilon)$ for some $\varepsilon > 0$ (in order to avoid boundary effects) and belongs to an isotropic Hölder functional class $\mathcal{H}_d(\alpha, L, D^\varepsilon)$ defined by (4), with $L > 0$ and $\alpha > 2$.

(A2) There is a unique point $\boldsymbol{\mu}$ in the interior $\mathring{D}$ of D that maximizes the function f on D, that is, $M = \sup_{\mathbf{x}\in D} f(\mathbf{x}) = \max_{\mathbf{x}\in \mathring{D}} f(\mathbf{x}) = f(\boldsymbol{\mu})$ and $\lambda_{\max}(Hf(\boldsymbol{\mu})) < 0$.

Note that conditions (A1) and (A2) imply that $\nabla f(\boldsymbol{\mu}) = \mathbf{0}$, and the Hessian $Hf(\boldsymbol{\mu})$ is a symmetric and negative definite matrix. Besides, as $\alpha > 2$, the Hessian matrix $Hf(\mathbf{x})$ is continuous and therefore for some $\kappa, \lambda_0 > 0$,

$$\sup_{\mathbf{x}\in B(\boldsymbol{\mu},\kappa)} \lambda_{\max}\big(Hf(\mathbf{x})\big) \le -\lambda_0. \tag{6}$$

Notice that constants κ, λ_0 depend on f. If we do not pursue any uniformity over f in our results, then the condition (6) follows from (A1), (A2) and can therefore be used in the proofs. However, when uniformizing the results over a functional class, this condition becomes autonomous and must be added to the description of the functional class; see Remark 1 below.

3. The two-stage procedure. For a column vector $\boldsymbol{\vartheta} = (\vartheta_{\mathbf{i}} : \mathbf{i} \in \mathbb{I}(r_\alpha, d))^T$, introduce the multivariate polynomial function

$$f_{\boldsymbol{\vartheta}}(\mathbf{x}) = \sum_{\mathbf{i}\in\mathbb{I}(r_\alpha,d)} \vartheta_{\mathbf{i}} \mathbf{x}^{\mathbf{i}} = \sum_{k=0}^{r_\alpha} \sum_{\mathbf{i}\in\mathbb{I}_k(d)} \vartheta_{\mathbf{i}} \mathbf{x}^{\mathbf{i}}. \tag{7}$$

We now describe the two-stage procedure for estimating the parameters $(\boldsymbol{\mu}, M)$. The first stage consists of the first two steps and the steps 3–5 comprise the second stage.

(1) The first stage starts as follows. For $\upsilon \in (0, 1)$, choose first stage design budget, that is, $n_1 \in \mathbb{N}$ such that $0 < n_1 < n$, $n_1/n \to \upsilon$. Find n_1 design points $\{\tilde{\mathbf{x}}_i, i = 1, \dots, n_1\}$ approximately uniformly over the set D in the sense that, for some $c_1, c_2 > 0$, the family of balls $\{B(\tilde{\mathbf{x}}_i, c_1 n^{-1/d}), i = 1, \dots, n_1\}$ covers D and $\|\tilde{\mathbf{x}}_i - \tilde{\mathbf{x}}_j\| \ge c_2 n^{-1/d}$ for $i \ne j$.

Observe the data $\mathcal{D}_1^* = \{(\tilde{\mathbf{x}}_k, \tilde{Y}_k), k = 1, \dots, n_1\}$, $\tilde{Y}_k = f(\tilde{\mathbf{x}}_k) + \tilde{\xi}_k$, $k = 1, \dots, n_1$, according to the model (1).

(2) Using $\mathcal{D}_1^*$, construct a preliminary consistent estimator $\tilde{\boldsymbol{\mu}}$ of $\boldsymbol{\mu}$. For $d = 1$, one may use the kernel estimator of Müller (1989) and for $d \ge 2$, its multivariate generalization given by Facer and Müller (2003).

(3) Let $n_2 = n - n_1$ be the remaining portion of the design budget, and let l be the smallest integer that satisfies $2l \ge r_\alpha$. Assume that $n_2 = n_3(2l + 1)^d$ for some $n_3 \in \mathbb{N}$, which is always possible to arrange. Note that $n_3 \ge cn$ for some constant

$c > 0$. Introduce a *localization parameter* $\delta_n > 0$, $\delta_n \to 0$, and define the set

$$\prod_{i=1}^{d}\{\tilde{\mu}_k + j_i\delta_n,\ j_i = 0, \pm 1, \pm 2, \dots, \pm l\} = \{\tilde{\mathbf{d}}_1, \dots, \tilde{\mathbf{d}}_{(2l+1)^d}\},$$

which consists of $(2l + 1)^d$ different points from the d-dimensional cube $C(\tilde{\boldsymbol{\mu}}, l\delta_n)$.

Now introduce the second stage design points $\{\mathbf{x}_k, k = 1, \dots, n_2\}$ in such a way that $|I_j| = n_3$ for all $j = 1, \dots, (2l + 1)^d$, where $I_j = \{1 \le k \le n_2 : \mathbf{x}_k = \tilde{\mathbf{d}}_j\}$. In other words, each point among the $(2l + 1)^d$ different points from the set $\{\tilde{\mathbf{d}}_1, \dots, \tilde{\mathbf{d}}_{(2l+1)^d}\}$ is repeated $n_3 = n_2/(2l + 1)^d$ times in the second stage design $\{\mathbf{x}_k, k = 1, \dots, n_2\}$. Observe the data $\mathcal{D}_2^* = \{(\mathbf{x}_k, Y_k), k = 1, \dots, n_2\}$, $Y_k = f(\mathbf{x}_k) + \xi_k, k = 1, \dots, n_2$, according to the model (1).

(4) Introduce the column vectors $\mathbf{Y} = (Y_1, \dots, Y_{n_2})^T$, $\bar{\mathbf{x}}_k = (\mathbf{x}_k^{\mathbf{i}} : \mathbf{i} \in \mathbb{I}(r_\alpha, d))^T$, $k = 1, \dots, n_2$, and form the data-matrix $\mathbf{X} = (\bar{\mathbf{x}}_1, \dots, \bar{\mathbf{x}}_{n_2})^T$ of dimension $n_2 \times (q(r_\alpha, d) + 1)$. Now using $\mathcal{D}_2^*$, fit a polynomial regression model of order r_α by

$$\tilde{\boldsymbol{\vartheta}} = \arg\min_{\boldsymbol{\vartheta}} \sum_{k=1}^{n_2}\bigl(Y_k - f_{\boldsymbol{\vartheta}}(\mathbf{x}_k)\bigr)^2 = \arg\min_{\boldsymbol{\vartheta}} \|\mathbf{Y} - \mathbf{X}\boldsymbol{\vartheta}\|^2,$$

where the polynomial $f_{\boldsymbol{\vartheta}}$ is introduced by (7). The unique least squares solution is given by $\tilde{\boldsymbol{\vartheta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}$, since $\mathbf{X}$ is full-rank by Lemma 1 below. Intuitively, this is expected since the number of observations $n_2 \ge (2l + 1)^d \ge (r_\alpha + 1)^d \ge |\mathbb{I}(r_\alpha, d)| = q(\alpha, d) + 1$.

(5) Finally, define the two-stage estimator $(\hat{\boldsymbol{\mu}}, \hat{M})$ of $(\boldsymbol{\mu}, M)$ by

$$\hat{\boldsymbol{\mu}} = \arg\max_{\mathbf{x}\in C(\tilde{\boldsymbol{\mu}}, l\delta_n)} f_{\tilde{\boldsymbol{\vartheta}}}(\mathbf{x}), \qquad \hat{M} = f_{\tilde{\boldsymbol{\vartheta}}}(\hat{\boldsymbol{\mu}}). \tag{8}$$

Note that $(\hat{\boldsymbol{\mu}}, \hat{M})$ depends on the first-stage estimator and the localization parameter δ_n introduced in step 3.

Clearly the construction of the two-stage procedure does not assume the knowledge of the error variance σ^2 provided that the preliminary estimator $\tilde{\boldsymbol{\mu}}$ also does not use this knowledge. Furthermore, the two-stage approach simultaneously estimates $\boldsymbol{\mu}$ and M, since the same design points for both estimators are used in the procedure. The two-stage procedure also provides improved estimators for all the relevant derivatives of f at $\boldsymbol{\mu}$; see Remark 7 below.

The following theorem gives the rate of convergence of the two-stage procedure for any smoothness level $\alpha > 2$.

THEOREM 1. *Suppose that the localization parameter δ_n satisfies $\sqrt{n}\delta_n^2 \to \infty$ and $\|\tilde{\boldsymbol{\mu}} - \boldsymbol{\mu}\| = o_p(\delta_n)$. Then under conditions* (A1) *and* (A2),

$$\|\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}\| = O_p\bigl(n^{-1/2}\delta_n^{-1}\bigr) + O_p\bigl(\delta_n^{\alpha-1}\bigr) \tag{9}$$

and

$$\hat{M} - M = O_p(n^{-1/2}) + O_p(\delta_n^\alpha). \tag{10}$$

Condition $\|\tilde{\boldsymbol{\mu}} - \boldsymbol{\mu}\| = o_p(\delta_n)$ has a clear heuristic interpretation: at the second stage, one should not localize more than what the accuracy of the estimation procedure allows at the first stage. Actually, it is sufficient to assume that $\mathrm{P}(\|\tilde{\boldsymbol{\mu}} - \boldsymbol{\mu}\| \le K\delta_n) \to 1$ for some K, but the dependence of K on unknown quantities will complicate the analysis.

We first observe that there is always a rate improvement from the first stage to the second if δ_n is chosen properly. To see this, let ε_n be the rate of convergence of $\tilde{\boldsymbol{\mu}}$. Since ε_n cannot be better than the optimal rate of convergence of all possible sequential procedures, which is $n^{-(\alpha-1)/2\alpha}$, we have $\varepsilon_n \gtrsim n^{-(\alpha-1)/(2\alpha)}$. Choose $\delta_n = \max(m_n\varepsilon_n, n^{-1/(2\alpha)})$, where m_n is a positive sequence going to infinity sufficiently slowly. Then $\|\tilde{\boldsymbol{\mu}} - \boldsymbol{\mu}\| = o_p(\delta_n)$ and $\sqrt{n}\delta_n^2 \to \infty$ (as $\alpha > 2$) are satisfied, and hence it remains to show that $\delta_n^{\alpha-1} = O(\varepsilon_n)$ and $n^{-1/2}\delta_n^{-1} = O(\varepsilon_n)$. If $\varepsilon_n \ll n^{-1/(2\alpha)}$, then $\delta_n = n^{-1/(2\alpha)}$ and the second stage rate is $n^{-1/2}\delta_n^{-1} = \delta_n^{\alpha-1} = n^{-(\alpha-1)/(2\alpha)} = O(\varepsilon_n)$, and the order improves strictly unless $\varepsilon_n \asymp n^{-(\alpha-1)/(2\alpha)}$. Clearly, in this case, the choice of δ_n is optimal as it balances the "order of variability" $n^{-1/2}\delta_n^{-1}$ and the "order of bias" $\delta_n^{\alpha-1}$. On the other hand, if $\varepsilon_n \gtrsim n^{-1/(2\alpha)}$, $\delta_n = m_n\varepsilon_n$, so $n^{-1/2}\delta_n^{-1} = o(\delta_n^{\alpha-1})$ and the second stage rate is $\delta_n^{\alpha-1} = m_n^{\alpha-1}\varepsilon_n^{\alpha-1} \ll \varepsilon_n$, since $\alpha > 2$ and m_n grows sufficiently slowly. Note that the "optimal choice" $\delta_n = n^{-1/(2\alpha)}$ is prohibited in this case since we need $\varepsilon_n = o(\delta_n)$. For estimating M, the rate of convergence of the two-stage procedure clearly is $\max(n^{-1/2}, m_n^\alpha\varepsilon_n^\alpha)$, which matches the optimal rate $n^{-1/2}$ if $\varepsilon_n \ll n^{-1/(2\alpha)}$. Of course, if the choice of δ_n is too big, then the rates for estimating $\boldsymbol{\mu}$ or M may deteriorate in the second stage.

Clearly, it is natural to use a preliminary estimator $\tilde{\boldsymbol{\mu}}$ with the fastest possible convergence rate $\varepsilon_n = n^{-(\alpha-1)/(2\alpha+d)}$ for any nonsequential procedure. Then the two-stage estimator will lead to the best possible convergence rates $n^{-(\alpha-1)/(2\alpha)}$ and $n^{-1/2}$ for estimating $\boldsymbol{\mu}$ and M, respectively, among all sequential procedures, provided that $\varepsilon_n = o(n^{-1/(2\alpha)})$. This condition holds when $\frac{\alpha-1}{2\alpha+d} > \frac{1}{2\alpha}$, or equivalently, $\alpha > 1 + \sqrt{1 + d/2}$. Indeed, under this condition, the two-stage procedure achieves the optimal rates $n^{-(\alpha-1)/(2\alpha)}$ and $n^{-1/2}$ for estimating $\boldsymbol{\mu}$ and M, respectively, even when a rate-optimal estimator is not used, as long as the convergence rate ε_n of the preliminary estimator is faster than $n^{-1/(2\alpha)}$. If the condition $\varepsilon_n = o(n^{-1/(2\alpha)})$ fails, the two-stage procedure does not give the optimal rate. In Section 4, we discuss a multi-stage generalization that can achieve optimal rate starting with almost any first-stage estimator.

The following corollary summarizes our conclusions.

COROLLARY 1. *Suppose that $\alpha > 1 + \sqrt{1 + d/2}$ and conditions* (A1), (A2) *hold. If the convergence rate of the preliminary estimator is faster than $n^{-1/(2\alpha)}$*

and the localization parameter is $\delta_n = n^{-1/(2\alpha)}$, *then*

$$\|\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}\| = O_p\big(n^{-(\alpha-1)/2\alpha}\big), \qquad \hat{M} - M = O_p\big(n^{-1/2}\big).$$

Interestingly, dimension d, which affects the first-stage optimal convergence rates $n^{-(\alpha-1)/(2\alpha+d)}$ and $n^{-\alpha/(2\alpha+d)}$ for estimating $\boldsymbol{\mu}$ and M, respectively, does not affect the corresponding two-stage and fully sequential optimal convergence rates $n^{-(\alpha-1)/(2\alpha)}$ and $n^{-1/2}$. Thus the curse of dimensionality is nearly avoided by the two-stage procedure, provided that the regression function is sufficiently smooth to ensure $\alpha > 1 + \sqrt{1 + d/2}$. The lower bound in this inequality increases with the dimension d. Notice that if $\alpha > 3$, the corollary yields the optimal rates for estimating $\boldsymbol{\mu}$ and M for all the dimensions for which $3 \geq 1 + \sqrt{1 + d/2}$, that is, up to dimension $d = 6$, including the most important dimensions $d = 1, 2, 3$.

REMARK 1. We can formulate a uniform version of Theorem 1. By inspecting the proofs, we see that all the bounds for the two-stage procedure can be made uniform over the Hölder class $\mathcal{H}_d(\alpha, L, D)$ if we additionally require relation (6) for some $\kappa, \lambda_0 > 0$, the uniform boundedness of all the partial derivatives involved in the definition of $\mathcal{H}_d$ and the uniformity of the first stage estimator.

To be more specific, for some positive α, L, L_1, κ, λ_0, κ_1, δ and ε, such that $\alpha > 2$ and $\kappa_1 \leq \kappa$, and a compact convex $D \subseteq \mathbb{R}^d$, introduce the following conditions:

($\tilde{\mathrm{A}}1$) $f \in \mathcal{H}_d(\alpha, L, D^\varepsilon)$ and $\sup_{\mathbf{x} \in D^\varepsilon} |\mathbf{D}^{\mathbf{i}} f(\mathbf{x})| \leq L_1$ for all $\mathbf{i} \in \mathbb{I}(r_\alpha, d)$.

($\tilde{\mathrm{A}}2$) There is a unique point $\boldsymbol{\mu} \in \mathring{D}$ that maximizes the function f on D, $\sup_{\mathbf{x} \in D} f(\mathbf{x}) = \max_{\mathbf{x} \in \mathring{D}} f(\mathbf{x}) = f(\boldsymbol{\mu})$. Moreover, $f(\boldsymbol{\mu}) \geq f(\mathbf{x}) + \delta$ for all $\mathbf{x} \notin B(\boldsymbol{\mu}, \kappa_1)$ and $\sup_{\mathbf{x} \in B(\boldsymbol{\mu}, \kappa)} \lambda_{\max}(Hf(\mathbf{x})) \leq -\lambda_0$.

Let $\tilde{\mathcal{H}}_d$ be the class of functions which satisfy ($\tilde{\mathrm{A}}1$) and ($\tilde{\mathrm{A}}2$). Then Theorem 1 holds uniformly in $f \in \tilde{\mathcal{H}}$, provided $\|\tilde{\boldsymbol{\mu}} - \boldsymbol{\mu}\| = o_p(\delta_n)$ holds uniformly over $\tilde{\mathcal{H}}$. Condition ($\tilde{\mathrm{A}}1$) is a strengthened version of (A1), namely (A1) is complemented by the requirement of uniform boundedness of all the relevant partial derivatives. Condition ($\tilde{\mathrm{A}}2$) is in turn a stronger version of (A2): relation (6) is included in ($\tilde{\mathrm{A}}2$) with common κ and λ for the whole class, and the existence of a unique location $\boldsymbol{\mu}$ of maximum is strenthened by the requirement of the uniform separation of the maximum function value $f(\boldsymbol{\mu})$ from the function values outside $B(\boldsymbol{\mu}, \kappa_1)$. Inside this vicinity, as $\kappa_1 \leq \kappa$, the separation of the maximum can be characterized by the Taylor expansion and (6); see the arguments in (32) below. This uniform separation condition is essential to make the first stage rate for $\tilde{\boldsymbol{\mu}}$ uniform over the functional class. Note that the separation condition for any particular function holds by the compactness of D and the uniqueness of the location of the maximum.

On the other hand, the two-stage procedure can achieve improved rates only under a local Hölder condition satisfied in a neighborhood of $\boldsymbol{\mu}$ provided that a first stage estimator with sufficiently good rate is available as a preliminary estimator.

This is due to the fact that the second stage design points are chosen close to the preliminary estimate, and hence close to the true maximum location $\boldsymbol{\mu}$.

REMARK 2. Almost sure convergence of $\hat{\boldsymbol{\mu}}$ and $\hat{M}$ can be obtained assuming that the preliminary estimator convergence rate is given in the almost sure sense. This will follow from the estimates given in Lemmas 2 and 3. Under additional moment conditions on the error distribution, almost sure convergence rate of a kernel-type estimator can be found.

4. Multi-stage procedures and resolving the curse of dimensionality. Theorem 1 shows that for estimating the maxima $\boldsymbol{\mu}$ and maximum value M of a Hölder α-smooth function f, $\alpha > 2$, starting with an estimator $\tilde{\boldsymbol{\mu}}$ having convergence rate ε_n and localization parameter δ_n, a two-stage estimator has an improved rate of convergence, unless ε_n is already equal to the optimal rate $n^{-(\alpha-1)/(2\alpha)}$. More precisely, assuming that $\varepsilon_n = O(n^{-\gamma})$ for some $\gamma > 0$ and choosing $\delta_n = \max(m_n\varepsilon_n, n^{-1/(2\alpha)})$, where $m_n \to \infty$ is a slowly varying sequence, the convergence rate for estimating $\boldsymbol{\mu}$ improves to the optimal rate $n^{-(\alpha-1)/(2\alpha)}$ if $\varepsilon_n = o(n^{-1/(2\alpha)})$ and to $\varepsilon_n^{\alpha-1}$ up to a slowly varying factor, if $\varepsilon_n \gtrsim n^{-1/(2\alpha)}$. Although the latter rate is not optimal, further improvement in rate can be achieved by applying the two-stage technique again, using the estimator obtained in the second stage as the new preliminary estimator, and repeating the procedure until the optimal rate is obtained. After k iterations of the two-stage procedure, the convergence rate thus becomes $\varepsilon_n^{(\alpha-1)^k}$ up to a slowly varying factor, provided that $\varepsilon_n^{(\alpha-1)^k} \gg n^{-(\alpha-1)/(2\alpha)}$. Let k_1 be the largest integer such that the last relation holds for $k = k_1$. Then iterating the two-stage procedure $k_1 + 1$ times, the resulting estimator will have the optimal convergence rate $n^{-(\alpha-1)/(2\alpha)}$. Thus the final multi-stage procedure has convergence rate completely free of the dimension d and applies to any smoothness level $\alpha > 2$. In order to apply the procedure in $k_1 + 1$ stages, one will need to split the observation budget in $k_1 + 1$ parts following the description given in step 3 of the procedure.

If we are interested only in estimating the maximum M, we may be able to stop earlier when applying the multi-stage procedure. In this case, the target optimal rate is $n^{-1/2}$. The two-stage estimator has convergence rate given by $\max\{\varepsilon_n^{\alpha}, n^{-1/2}\}$. Hence the optimal rate will be obtained at stage $k_2 + 1$, where k_2 is the largest integer integer such that $\varepsilon_n^{\alpha^k} \gg n^{-1/2}$.

REMARK 3. The smoothness level α needs to be strictly greater than 2 to control the error in the second-order Taylor approximation of the underlying multivariate regression function. As α gets closer to 2, the required number of stages in the multi-stage procedure increases without bound.

Consider now the adaptive version of our original estimation problem, where the problem is to estimate $\boldsymbol{\mu}$ at the optimal rate $n^{-(\alpha-1)/(2\alpha)}$ and M at rate $n^{-1/2}$

without knowing the smoothness level α. Since the choice of the localization parameter δ_n depends on the knowledge of α it is not possible to apply the two-stage procedure, and hence a multi-stage adaptive estimator for $\boldsymbol{\mu}$ is not possible. However, for estimating M, it is possible to construct a multi-stage procedure with convergence rate $n^{-1/2}$ without knowing α, as long as $\alpha > 2$. Start with an estimator for $\boldsymbol{\mu}$ which converges at rate $n^{-\beta}$ for all Hölder 2-smooth functions. For instance, the rate $n^{-1/(4+d)}$ is possible in dimension d by the results of Müller (1989) and Facer and Müller (2003). Then by applying Theorem 1 with $\delta_n = m_n n^{-\beta}$, where $m_n \to \infty$ is a slowly varying sequence, the rate of convergence for estimating M improves to $\max\{m_n^2 n^{-2\beta}, n^{-1/2}\}$ in stage two. Repeating the two-stage procedure k times, thus the rate will improve to $n^{-1/2}$, whenever $2^k\beta > \frac{1}{2}$, or $k > (\log(1/\beta)/\log 2) - 1$. In particular, starting with the one-stage optimal estimator having convergence rate $n^{-1/(4+d)}$, the required number of stages to achieve $n^{-1/2}$ rate at all Hölder 2-smooth functions is the smallest integer greater than $(\log(4 + d)/\log 2) - 1$, since repeating the two-stage procedure beyond k given above does not hurt the rate.

5. Simulations. In this section, we compare the performance of the two-stage procedure with an equivalent single-stage procedure. We consider the bivariate case $d = 2$ and take a regression function $f : [0, 1]^2 \to \mathbb{R}$ defined by

$$f(x, y) = 5x(x - 1)y(y - 1)\sin(11x)\sin(11y)$$

(the smooth surface in the top left panel of Figure 1). In the first stage the function is observed with Gaussian noise with standard deviation $\sigma = 0.1$ on a regular 25 by 25 grid (gray points in the same panel). Using standard local linear regression, a surface is fit through these points (the surface in the top right panel of Figure 1) and the point where this fitted function is maximal serves as the stage one estimator $\tilde{\boldsymbol{\mu}}$ (the red point in the same panel). Next we take $\delta = 0.1$ and generate 70 new observations at each of the nine points $(\tilde{\mu}_1 + j_1\delta, \tilde{\mu}_2 + j_2\delta)$, $j_1, j_2 = 0, \pm 1$ (gray points in the top right panel of Figure 1). Finally a quadratic surface is fitted through these new data points (the surface in the lower panel of Figure 1) and the location of the maximum is the final second stage estimator $\hat{\boldsymbol{\mu}}$ (the green point in the figure). The implementation of the procedure is rather straightforward. In the statistical language R, we used the standard function `loess` in the first stage to fit the surface using the first stage observations, and we used the function `lm` to fit the quadratic surface using the second stage observations.

Note that in total we have used $25 \times 25 + 9 \times 70 = 1255$ observations. It is illustrative to compare our procedure to a single stage estimator that uses about the same amount of regularly spaced observations. The closest is a regular 36 by 36 grid, which contains 1296 points. We make noisy observations of the function f at these grid points, again corrupted by centered Gaussian noise with standard deviation 0.1. We consider the estimator for the location of the maximum of f that

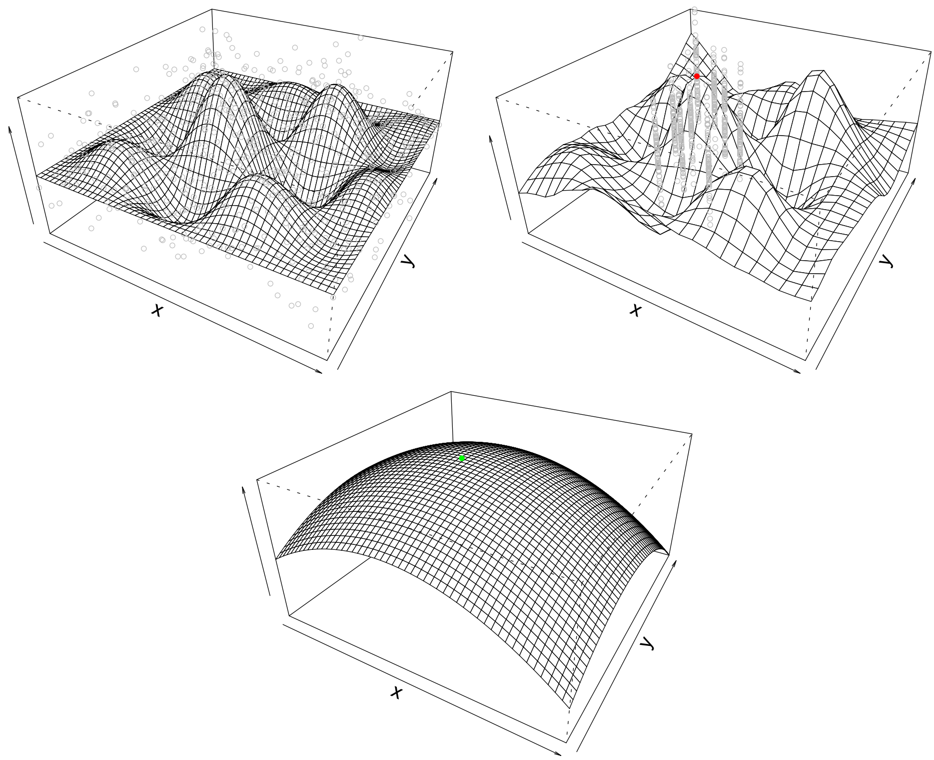

FIG. 1. *Top left: true regression function (surface) and stage one observations (gray points). Top right: surface fitted through stage one observations, the initial estimator (red point) and the stage two observations (gray points). Bottom: quadratic surface fitted through the stage two observations and final estimator (green point).*

is obtained by fitting a locally linear surface through these data points and computing the location where this is maximal. Obviously, the quality of this estimator depends on the bandwidth that is used (or span parameter, as it is called in the R function `loess`). To obtain a fair comparison with our two-stage estimator, we should make an optimal choice. We achieve this by repeating the experiment a large number of times with different bandwidths and computing numerical mean squared errors (MSEs). The result is shown in the left panel of Figure 2. The numerical MSE is minimal for the bandwidth choice $h = 0.085$.

To compare the mean-squared error of the single-stage estimator based on this regular grid, we replicated the experiment 10,000 times and computed the Monte-Carlo average of the squared difference between the estimate and the true maximum. The results are shown in the left boxplot in the right panel of Figure 2. Similarly we carried out the two-stage procedure 10,000 times (with bandwidth $h = 0.085$ in stage one and $\delta = 0.1$) and computed the errors as well. These are shown in the right boxplot in the right panel of Figure 2. It is clear that the two-

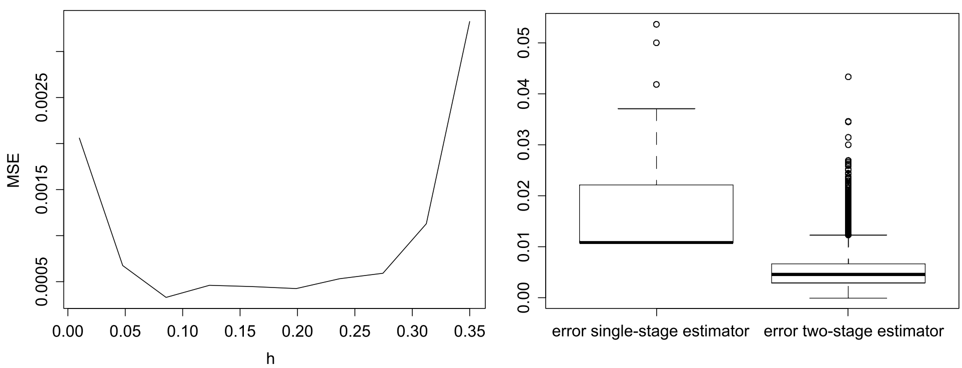

FIG. 2. *Left*: *MSE of the single-stage estimator against the bandwidth used*. *Right*: *boxplot of the errors of the single-stage estimator* (*with optimal bandwidth choice* $h = 0.085$) *and the errors of our two-stage procedure* (*with the same bandwidth choice and* $\delta = 0.1$).

stage estimator performs better in this situation, in terms of the mean-squared error. This is in spite of the fact that the two-stage estimator has used less observations, namely 1255 in total compared to 1296 used by the single-stage estimator.

In practice the quality of our procedure clearly depends on the quality of the estimator that is used in the first stage and also on the choice of the localization parameter δ. In this simulation example, where we use local linear regression in the first stage, the quality of the estimator therefore depends on the bandwidth used in stage one. To investigate the dependence of the performance on this parameter we carried out the simulation study described above for a range of bandwidths. The results are shown in the left panel of Figure 3. The solid line gives the MSE of our estimator as a function of the bandwidth used in stage one. The dashed line is the MSE of the optimal single stage estimator described above. The plot shows

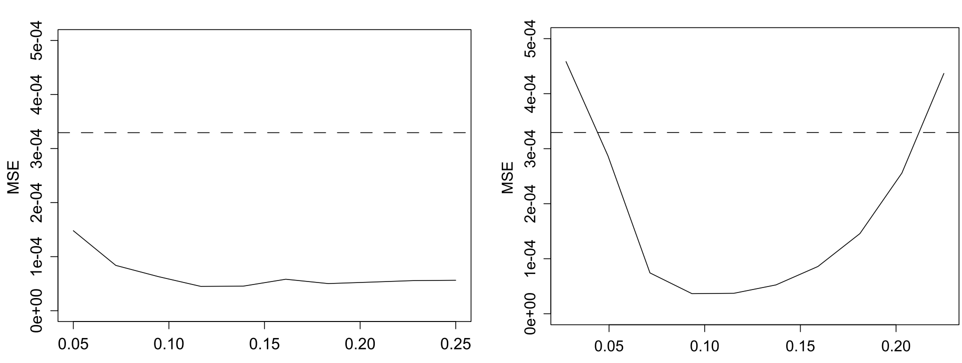

FIG. 3. *MSE of the two-stage estimator against the bandwidth* h *used in the first stage* (*left*), *and against the localization parameter* δ (*right*). *The dashed line is the MSE of the single-stage estimator with optimal bandwidth choice*.

that in fact for a range of bandwidths the two stage procedure performs better than the single stage procedure. Similarly, the right panel of Figure 3 describes the performance of the two-stage procedure as a function of the localization parameter δ. Again there is a range of possible δ's for which we obtain an improved performance, but choosing δ too small or too large deteriorates the quality. In practice one might, for instance, use cross-validation type methods to set the tuning parameters h and δ. Further research is needed to find theoretically sound methods.

6. Proofs. Throughout this section, α and d are are kept fixed. To simplify notation, we abbreviate $\mathbb{I}(r_\alpha, d)$ by $\mathbb{I}$, $\mathbb{I}_k(d)$ by $\mathbb{I}_k$ and $q(\alpha, d)$ by q.

First we introduce several quantities we are going to use in the sequel. Define $\mathbf{z}_k = \mathbf{x}_k - \tilde{\boldsymbol{\mu}}$, $k = 1, \dots, n_2$, and reformulate definition (8) by representing the involved quantities in terms of the newly defined shifted design points $\mathbf{z}_k$, $k = 1, \dots, n_2$. Let $\mathbf{d}_j = \tilde{\mathbf{d}}_j - \tilde{\boldsymbol{\mu}}$, $j = 1, \dots, (2l+1)^d$. Then for all $k = 1, \dots, n_2$,

$$\mathbf{z}_k \in \{0, \pm\delta_n, \dots, \pm l\delta_n\}^d = \{\mathbf{d}_1, \dots, \mathbf{d}_{(2l+1)^d}\} \subset C(l\delta_n), \tag{11}$$

so that each of the distinct $(2l+1)^d$ points are repeated $n_3 = n_2/(2l+1)^d$ times in the new design set $\{\mathbf{z}_k, k = 1, \dots, n_2\}$. Using definition (7), define an estimator $\hat{\boldsymbol{\theta}}$ by equating the two polynomials

$$f_{\hat{\boldsymbol{\theta}}}(\mathbf{x} - \tilde{\boldsymbol{\mu}}) = f_{\tilde{\boldsymbol{\vartheta}}}(\mathbf{x}) \qquad \text{equivalently } \hat{\boldsymbol{\theta}} = (\mathbf{Z}^T\mathbf{Z})^{-1}\mathbf{Z}^T\mathbf{Y}, \tag{12}$$

where

$$\mathbf{Z} = (\bar{\mathbf{z}}_1, \dots, \bar{\mathbf{z}}_{n_2})^T, \qquad \bar{\mathbf{z}}_k = (\mathbf{z}_k^{\mathbf{i}} : \mathbf{i} \in \mathbb{I})^T \tag{13}$$

and $\mathbf{z}_k = \mathbf{x}_k - \tilde{\boldsymbol{\mu}}$, $k = 1, \dots, n_2$. The matrix $\mathbf{Z}^T\mathbf{Z}$ is invertible by Lemma 1 below. We thus obtain an equivalent description of the estimator $(\hat{\boldsymbol{\mu}}, \hat{M})$ given by (8) in terms of the polynomial $f_{\hat{\boldsymbol{\theta}}}(\mathbf{z})$ defined by (7), with $\hat{\boldsymbol{\theta}}$ defined by (12),

$$\hat{\boldsymbol{\mu}} = \tilde{\boldsymbol{\mu}} + \mathring{\boldsymbol{\mu}}, \qquad \hat{M} = f_{\hat{\boldsymbol{\theta}}}(\mathring{\boldsymbol{\mu}}) \qquad \text{where } \mathring{\boldsymbol{\mu}} = \arg\max_{\mathbf{z} \in C(l\delta_n)} f_{\hat{\boldsymbol{\theta}}}(\mathbf{z}), \tag{14}$$

$C(l\delta_n) = [-l\delta_n, l\delta_n]^d \subset \mathbb{R}^d$. In doing this shifting trick, we make the computations easier because the matrix $\mathbf{Z}^T\mathbf{Z}$ will have a lot of zero entries as the design points $\mathbf{z}_k$'s are symmetrically centered around zero in each dimension rather than being centered around $\tilde{\boldsymbol{\mu}}$.

Next, let the vector $\boldsymbol{\theta} = (\theta_{\mathbf{i}} : \mathbf{i} \in \mathbb{I})^T$ be defined by the equality of the two polynomials $f_{\boldsymbol{\theta}}(\mathbf{x} - \tilde{\boldsymbol{\mu}}) = P_{f,\boldsymbol{\mu}}(\mathbf{x})$, where $P_{f,\boldsymbol{\mu}}$ is the Taylor expansion of f of order r_α around $\boldsymbol{\mu}$ defined by (5),

$$\sum_{\mathbf{i} \in \mathbb{I}} \theta_{\mathbf{i}}(\mathbf{x} - \tilde{\boldsymbol{\mu}})^{\mathbf{i}} = f(\boldsymbol{\mu}) + \sum_{\mathbf{i} \in \mathbb{I}, |\mathbf{i}| \geq 2} \frac{\mathbf{D}^{\mathbf{i}} f(\boldsymbol{\mu})}{\mathbf{i}!}(\mathbf{x} - \boldsymbol{\mu})^{\mathbf{i}}, \tag{15}$$

here we have used the condition $\nabla f(\boldsymbol{\mu}) = \mathbf{0}$, due to (A1) and (A2). Thus, $\boldsymbol{\theta}$ is a random vector depending on f, $\boldsymbol{\mu}$ and $\tilde{\boldsymbol{\mu}}$. From (15) it follows that

$$\mathbf{i}!\theta_{\mathbf{i}} = \mathbf{D}^{\mathbf{i}} P_{f,\boldsymbol{\mu}}(\tilde{\boldsymbol{\mu}}), \qquad \mathbf{D}^{\mathbf{i}} f(\boldsymbol{\mu}) = \mathbf{D}^{\mathbf{i}} f_{\boldsymbol{\theta}}(\boldsymbol{\mu} - \tilde{\boldsymbol{\mu}}), \qquad \mathbf{i} \in \mathbb{I}. \tag{16}$$

The next lemma ensures that the estimator (12) is well defined; that is, $\mathbf{Z}^T\mathbf{Z}$ is invertible.

LEMMA 1. *The columns of matrix* $\mathbf{Z}$ (*and* $\mathbf{X}$) *defined by* (13) *are linearly independent.*

PROOF. Consider the matrix $\mathbf{Z}$; the same proof applies to $\mathbf{X}$.

For multi-indices $\mathbf{i} \in \mathbb{N}^p$ and $\mathbf{j} \in \mathbb{N}^s$, define the concatenation operation $(\mathbf{i}, \mathbf{j}) = (i_1, \dots, i_p, j_1, \dots, j_s) \in \mathbb{N}^{p+s}$. In particular, for $k \in \mathbb{N}$, $\mathbf{i} \in \mathbb{N}^p$, $(k, \mathbf{i}) = (k, i_1, \dots, i_p)$. Introduce the following notation: for $l = 0, 1, \dots, d-1$ and $\mathbf{z} = (z_1, \dots, z_d) \in \mathbb{R}^d$, define $\mathbf{i}_{-l} = (i_{l+1}, \dots, i_d) \in \mathbb{N}^{d-l}$, $\mathbf{z}_{-l} = (z_{l+1}, \dots, z_d) \in \mathbb{R}^{d-l}$ and the set $\mathbb{I}_{-l}(i_1, \dots, i_l) = \{\mathbf{i} \in \mathbb{N}^{d-l} : (i_1, \dots, i_l, \mathbf{i}) \in \mathbb{I}\}$.

Let $\mathbf{C_i}$, $\mathbf{i} \in \mathbb{I}$, be the columns of the matrix $\mathbf{Z}$. We need to show that $\sum_{\mathbf{i} \in \mathbb{I}} \lambda_{\mathbf{i}} \mathbf{C_i} = \mathbf{0}$ implies that $\lambda_{\mathbf{i}} = 0$ for all $\mathbf{i} \in \mathbb{I}$. The equality $\sum_{\mathbf{i} \in \mathbb{I}} \lambda_{\mathbf{i}} \mathbf{C_i} = \mathbf{0}$ is equivalent to

$$0 = \sum_{\mathbf{i} \in \mathbb{I}} \lambda_{\mathbf{i}} \mathbf{z}_k^{\mathbf{i}}, \qquad k = 1, 2, \dots, n_2.$$

Among $\{\mathbf{z}_1, \dots, \mathbf{z}_{n_2}\}$, only $(2l+1)^d$ are distinct—$\{\mathbf{d}_1, \dots, \mathbf{d}_{(2l+1)^d}\}$ given by (11). Thus, for all $\mathbf{z} \in \{\mathbf{d}_1, \dots, \mathbf{d}_{(2l+1)^d}\}$,

$$0 = \sum_{\mathbf{i} \in \mathbb{I}} \lambda_{\mathbf{i}} \mathbf{z}^{\mathbf{i}} = \sum_{i_1=0}^{r_\alpha} z_1^{i_1} \sum_{\mathbf{i}_{-1} \in \mathbb{I}_{-1}(i_1)} \lambda_{i_1 \mathbf{i}_{-1}} \mathbf{z}_{-1}^{\mathbf{i}_{-1}}.$$

For a fixed $\mathbf{z}_{-1} = (z_2, \dots, z_d)$, the right-hand side of the last relation is a polynomial of order r_α in variable z_1. But we have $2l + 1 > r_\alpha$ different design values $\{j\delta_n : j = 0, \pm 1, \pm 2, \dots, \pm l\}$ of the variable z_1 for which this polynomial must take the zero value. This forces all the coefficients of this polynomial to be zero. Thus we have that

$$0 = \sum_{\mathbf{i}_{-1} \in \mathbb{I}_{-1}(i_1)} \lambda_{i_1 \mathbf{i}_{-1}} \mathbf{z}_{-1}^{\mathbf{i}_{-1}}, \qquad i_1 = 0, 1, \dots, r_\alpha$$

for all possible design values of $\mathbf{z}_{-1} = (z_2, \dots, z_d)$. Iterating the above reasoning up to the variable z_d leads to, for all $i_1, \dots, i_{d-1} = 0, 1, \dots, r_\alpha$, $z_d \in \{0, \pm\delta_n, \pm 2\delta_n, \dots, \pm l\delta_n\}$,

$$0 = \sum_{i_d \in \mathbb{I}_{-(d-1)}(i_1, i_2, \dots, i_{d-1})} \lambda_{i_1 i_2, \dots, i_{d-1} i_d} z_d^{i_d},$$

from which we derive that $\lambda_{\mathbf{i}} = 0$ for all $\mathbf{i} \in \mathbb{I}$. $\square$

REMARK 4. In the case $d = 1$, $\mathbf{X}$ and $\mathbf{Z}$ are Vandermonde matrices.

The next lemma shows that the second stage data $\mathcal{D}_2^*$ can be regarded as coming approximately from a certain polynomial regression model.

LEMMA 2. *Assume* (A1) *and let the data* $\{(\mathbf{x}_k, Y_k), k = 1, \dots, n_2\}$ *and* $\boldsymbol{\xi} = (\xi_1, \dots, \xi_{n_2})^T$ *be given by the second stage observation scheme,* $\mathbf{Y} = (Y_1, \dots, Y_{n_2})^T$, $\boldsymbol{\eta} = (\eta_1, \dots, \eta_{n_2})^T$ *with* $\eta_k = f(\mathbf{x}_k) - P_{f,\boldsymbol{\mu}}(\mathbf{x}_k)$. *Then* $\mathbf{Y} = \mathbf{Z}\boldsymbol{\theta} + \boldsymbol{\eta} + \boldsymbol{\xi}$, *where* $\mathbf{Z}$ *and* $\boldsymbol{\theta}$ *are defined by* (13) *and* (15), *respectively, and* $\boldsymbol{\eta}$ *is independent of* $\boldsymbol{\xi}$. *Moreover, for some constants* C_1, C_2 *and uniformly in* $k \in \{1, 2, \dots, n_2\}$,

$$|\eta_k| \le C_1 \delta_n^\alpha + C_2 \|\tilde{\boldsymbol{\mu}} - \boldsymbol{\mu}\|^\alpha. \tag{17}$$

PROOF. Since $\eta_k = f(\mathbf{x}_k) - P_{f,\boldsymbol{\mu}}(\mathbf{x}_k)$, by (13) and (15),

$$\begin{aligned} Y_k &= f(\mathbf{x}_k) + \xi_k = P_{f,\boldsymbol{\mu}}(\mathbf{x}_k) + \eta_k + \xi_k \\ &= f_{\boldsymbol{\theta}}(\mathbf{x}_k - \tilde{\boldsymbol{\mu}}) + \eta_k + \xi_k = \bar{\mathbf{z}}_k^T \boldsymbol{\theta} + \eta_k + \xi_k. \end{aligned}$$

Clearly, $\boldsymbol{\eta}$ is independent of $\boldsymbol{\xi}$ by definition. It remains to show (17). Apply the c_r-inequality, $|a + b|^r \le \max(1, 2^{r-1})(|a|^r + |b|^r)$, $r > 0$, (4) and the fact that $\|\mathbf{x}_k - \tilde{\boldsymbol{\mu}}\| \le \sqrt{dl}\delta_n$, $k = 1, \dots, n_2$, to obtain (17)

$$\begin{aligned} |\eta_k| &= |f(\mathbf{x}_k) - P_{f,\boldsymbol{\mu}}(\mathbf{x}_k)| \le L\|\mathbf{x}_k - \boldsymbol{\mu}\|^\alpha \\ &\le L c_\alpha (\|\mathbf{x}_k - \tilde{\boldsymbol{\mu}}\|^\alpha + \|\tilde{\boldsymbol{\mu}} - \boldsymbol{\mu}\|^\alpha) \le C_1 \delta_n^\alpha + C_2 \|\tilde{\boldsymbol{\mu}} - \boldsymbol{\mu}\|^\alpha. \end{aligned} \qquad \square$$

Lemma 1 ensures that the matrix $\mathbf{Z}^T\mathbf{Z}$ is nonsingular. The following lemma describes the asymptotic behavior of the elements of its inverse. For notational convenience, below we enumerate the rows and columns of matrices starting from 0.

Enumerate $\mathbb{I}$ by arranging their elements in the order described in Section 2, which we denote by $\mathbf{i}_0, \dots, \mathbf{i}_q$, respectively.

LEMMA 3. *The* (i, j)*th element* h_{ij} *of* $(\mathbf{Z}^T\mathbf{Z})^{-1}$ *satisfies*

$$h_{ij} = O\big(n^{-1}\delta_n^{-(|\mathbf{i}_i| + |\mathbf{i}_j|)}\big), \qquad i, j = 0, 1, \dots, q.$$

PROOF. Since $\mathbf{z}_k^{\mathbf{0}} = 1$ for all $k = 1, \dots, n_2$, we have

$$\mathbf{Z}^T\mathbf{Z} = \begin{pmatrix} n_2 & \sum_{k=1}^{n_2} \mathbf{z}_k^{\mathbf{i}_1} & \cdots & \sum_{k=1}^{n_2} \mathbf{z}_k^{\mathbf{i}_q} \\ \sum_{k=1}^{n_2} \mathbf{z}_k^{\mathbf{i}_1} & \sum_{k=1}^{n_2} \mathbf{z}_k^{\mathbf{i}_1}\mathbf{z}_k^{\mathbf{i}_1} & \cdots & \sum_{i=1}^{n_2} \mathbf{z}_k^{\mathbf{i}_1}\mathbf{z}_k^{\mathbf{i}_q} \\ & & \cdots & \\ \sum_{k=1}^{n_2} \mathbf{z}_k^{\mathbf{i}_q} & \sum_{i=1}^{n_2} \mathbf{z}_k^{\mathbf{i}_1}\mathbf{z}_k^{\mathbf{i}_q} & \cdots & \sum_{i=1}^{n_2} \mathbf{z}_k^{\mathbf{i}_q}\mathbf{z}_k^{\mathbf{i}_q} \end{pmatrix}.$$

Then, for some constants a_{ij}, $i, j = 0, 1, \dots, q$, we rewrite the symmetric matrix $\mathbf{Z}^T\mathbf{Z}$ as follows:

$$\mathbf{Z}^T\mathbf{Z} = n_3 \begin{pmatrix} a_{00} & a_{01}\delta_n^{|\mathbf{i}_1|} & \cdots & a_{0q}\delta_n^{|\mathbf{i}_q|} \\ a_{10}\delta_n^{|\mathbf{i}_1|} & a_{11}\delta_n^{|\mathbf{i}_1| + |\mathbf{i}_1|} & \cdots & a_{1q}\delta_n^{|\mathbf{i}_1| + |\mathbf{i}_q|} \\ & & \cdots & \\ a_{q0}\delta_n^{|\mathbf{i}_q|} & a_{q1}\delta_n^{|\mathbf{i}_q| + |\mathbf{i}_1|} & \cdots & a_{qq}\delta_n^{|\mathbf{i}_q| + |\mathbf{i}_q|} \end{pmatrix}.$$

Some entries are easy to compute. For example, $a_{00} = (2l + 1)^d$ since $n_2 = (2l+1)^d n_3$. Moreover, there are many zeros due to the symmetry of the design. In particular, $a_{ij} = 0$ for all $i, j \in \{0, 1, \dots, q\}$ such that $|\mathbf{i}_i| + |\mathbf{i}_j|$ is an odd number. However we are not concerned about the exact values a_{ij} but only about nonsingularity of the matrix $\mathbf{A}$ with (i, j)th entry equal to a_{ij}, $i, j = 0, \dots, q$.

Let $\mathbf{\Delta}$ be the diagonal matrix with elements $\delta_n^{|\mathbf{i}_j|}$, $j = 0, 1, \dots, q$, in that order. Now notice that $\mathbf{Z}^T\mathbf{Z} = n_3\mathbf{\Delta}\mathbf{A}\mathbf{\Delta}$. Since $\mathbf{Z}^T\mathbf{Z}$ is nonsingular by Lemma 1, it follows that $\mathbf{A}$ is also invertible. Therefore $(\mathbf{Z}^T\mathbf{Z})^{-1} = n_3^{-1}\mathbf{\Delta}^{-1}\mathbf{A}^{-1}\mathbf{\Delta}^{-1}$. Denote by a^{ij} the (i, j)th entry of the constant matrix $\mathbf{A}^{-1}$ and recall that $n_3 \geq cn$. Then for $i, j = 0, 1, \dots, q$,

$$h_{ij} = n_3^{-1} a^{ij} \delta_n^{-(|\mathbf{i}_i|+|\mathbf{i}_j|)} = O\big(n^{-1}\delta_n^{-(|\mathbf{i}_i|+|\mathbf{i}_j|)}\big). \qquad \square$$

REMARK 5. For $d = 1$ and even r_α, we have $2l + 1 = r_\alpha + 1$. Put $b_0 = r_\alpha + 1$, $b_m = 0$ for all odd $m \in \{1, \dots, 2r_\alpha\}$, and for each even $m \in \{1, \dots, 2r_\alpha\}$

$$b_m = 2\big(1 + 2^m + 3^m + \cdots + l^m\big) = 2\big\{1 + 2^m + \cdots + (r_\alpha/2)^m\big\}.$$

Then the entries of $\mathbf{A}$ can be computed as follows: Since $n_2 = (2l + 1)n_3$, $\sum_{k=1}^{n_2} z_k^m = 0$ for each odd $m \in \{1, \dots, 2r_\alpha\}$ and

$$\sum_{k=1}^{n_2} z_k^m = 2n_3\big\{l^m\delta_n^m + (l-1)^m\delta_n^m + \cdots + \delta_n^m\big\} = n_3\delta_n^m b_m$$

for each even $m \in \{1, \dots, 2r_\alpha\}$, we obtain that $a_{ij} = b_{i+j}$.

The case of odd r_α can be treated similarly leading to slightly different constants.

LEMMA 4. *Assume* (A1), *and let* $\hat{\boldsymbol{\theta}}$ *and* $\boldsymbol{\theta}$ *be defined by* (12) *and* (15), *respectively. Then*

$$\hat{\theta}_{\mathbf{i}} = \theta_{\mathbf{i}} + O_p\big(n^{-1/2}\delta_n^{-|\mathbf{i}|}\big) + O\big(\delta_n^{\alpha-|\mathbf{i}|}\big) + O\big(\|\tilde{\boldsymbol{\mu}} - \boldsymbol{\mu}\|^\alpha \delta_n^{-|\mathbf{i}|}\big), \qquad \mathbf{i} \in \mathbb{I}.$$

PROOF. Using (12) and Lemma 2, write

$$\hat{\boldsymbol{\theta}} - \boldsymbol{\theta} = \big(\mathbf{Z}^T\mathbf{Z}\big)^{-1}\mathbf{Z}^T\mathbf{Y} - \boldsymbol{\theta} = \big(\mathbf{Z}^T\mathbf{Z}\big)^{-1}\mathbf{Z}^T(\boldsymbol{\eta} + \boldsymbol{\xi}). \tag{18}$$

Since $\mathrm{E}(\boldsymbol{\xi}) = \mathbf{0}$ and $\mathrm{Cov}((\mathbf{Z}^T\mathbf{Z})^{-1}\mathbf{Z}^T\boldsymbol{\xi}) = \sigma^2(\mathbf{Z}^T\mathbf{Z})^{-1}$, the order of the term $(\mathbf{Z}^T\mathbf{Z})^{-1}\mathbf{Z}^T\boldsymbol{\xi}$ is determined by the diagonal entries of the matrix $(\mathbf{Z}^T\mathbf{Z})^{-1}$. Hence, by Lemma 3, we have

$$\big(\mathbf{Z}^T\mathbf{Z}\big)^{-1}\mathbf{Z}^T\boldsymbol{\xi} = \big(O_p\big(n^{-1/2}\big), O_p\big(n^{-1/2}\delta_n^{-|\mathbf{i}_1|}\big), \dots, O_p\big(n^{-1/2}\delta_n^{-|\mathbf{i}_q|}\big)\big)^T. \tag{19}$$

In view of (11), $\mathbf{z}_k \in C(l\delta_n)$, so that $|\mathbf{z}_k^{\mathbf{i}}| \le c\delta_n^{|\mathbf{i}|}$, $k = 1, \dots, n_2$, $\mathbf{i} \in \mathbb{I}$. Using this, (17), $n_2 \le n$ and Lemma 3, we obtain that

$$
\begin{aligned}
(\mathbf{Z}^T\mathbf{Z})^{-1}\mathbf{Z}^T\boldsymbol{\eta} &= \begin{pmatrix} h_{00}\sum_{k=1}^{n_2}\mathbf{z}_k^{\mathbf{i}_0}\eta_k + \cdots + h_{0q}\sum_{k=1}^{n_2}\mathbf{z}_k^{\mathbf{i}_q}\eta_k \\ h_{10}\sum_{k=1}^{n_2}\mathbf{z}_k^{\mathbf{i}_0}\eta_k + \cdots + h_{1q}\sum_{i=k}^{n_2}\mathbf{z}_k^{\mathbf{i}_q}\eta_k \\ \vdots \\ h_{q0}\sum_{k=1}^{n_2}\mathbf{z}_k^{\mathbf{i}_0}\eta_k + \cdots + h_{qq}\sum_{k=1}^{n_2}\mathbf{z}_k^{\mathbf{i}_q}\eta_k \end{pmatrix} \\
&= \begin{pmatrix} O(\delta_n^{\alpha}) + O(\|\tilde{\boldsymbol{\mu}} - \boldsymbol{\mu}\|^{\alpha}) \\ O(\delta_n^{\alpha-|\mathbf{i}_1|}) + O(\|\tilde{\boldsymbol{\mu}} - \boldsymbol{\mu}\|^{\alpha}\delta_n^{-|\mathbf{i}_1|}) \\ \vdots \\ O(\delta_n^{\alpha-|\mathbf{i}_q|}) + O(\|\tilde{\boldsymbol{\mu}} - \boldsymbol{\mu}\|^{\alpha}\delta_n^{-|\mathbf{i}_q|}) \end{pmatrix}.
\end{aligned} \tag{20}
$$

Combining relations (18), (19) and (20) completes the proof of the lemma. $\square$

Let $\mathbf{1}_j$, $j = 1, \dots, d$, be the d standard unit vectors of $\mathbb{R}^d$, that is, $\mathbf{1}_j$ has 1 at the jth coordinate and zeros at other $d-1$ coordinates. Notice that $\mathbb{I}_1 = \{\mathbf{1}_j : j = 1, \dots, d\}$.

LEMMA 5. *Assume* (A1) *and let* $f_{\boldsymbol{\vartheta}}$, $\hat{\boldsymbol{\theta}}$ *and* $\boldsymbol{\theta}$ *be defined by* (7), (12) *and* (15), *respectively. If* $\|\boldsymbol{\mu} - \tilde{\boldsymbol{\mu}}\| = o_p(\delta_n)$, *then*

$$\nabla f_{\hat{\boldsymbol{\theta}}}(\boldsymbol{\mu} - \tilde{\boldsymbol{\mu}}) - \nabla f_{\boldsymbol{\theta}}(\boldsymbol{\mu} - \tilde{\boldsymbol{\mu}}) = O_p(n^{-1/2}\delta_n^{-1}) + O_p(\delta_n^{\alpha-1}).$$

PROOF. The jth coordinate of the vector $\nabla f_{\boldsymbol{\theta}}(\mathbf{x})$ is

$$
\begin{aligned}
\frac{\partial f_{\boldsymbol{\theta}}(\mathbf{x})}{\partial x_j} &= \sum_{\mathbf{i}\in\mathbb{I}} \boldsymbol{\theta}_{\mathbf{i}} \frac{\partial \mathbf{x}^{\mathbf{i}}}{\partial x_j} = \sum_{\mathbf{i}\in\mathbb{I}:\, i_j \ge 1} \boldsymbol{\theta}_{\mathbf{i}} \frac{\partial \mathbf{x}^{\mathbf{i}}}{\partial x_j} = \sum_{\mathbf{i}\in\mathbb{I}:\, i_j\ge 1} i_j \theta_{\mathbf{i}} \mathbf{x}^{\mathbf{i}-\mathbf{1}_j} \\
&= \theta_{\mathbf{1}_j} + \sum_{\mathbf{i}\in\mathbb{I}:\, i_j\ge 1, |\mathbf{i}|\ge 2} i_j \theta_{\mathbf{i}} \mathbf{x}^{\mathbf{i}-\mathbf{1}_j},
\end{aligned}
$$

where $\mathbf{1}_j \in \mathbb{I}_1$. Then, for each $j = 1, \dots, d$,

$$
\begin{aligned}
&\frac{\partial f_{\hat{\boldsymbol{\theta}}}(\boldsymbol{\mu} - \tilde{\boldsymbol{\mu}})}{\partial x_j} - \frac{\partial f_{\boldsymbol{\theta}}(\boldsymbol{\mu} - \tilde{\boldsymbol{\mu}})}{\partial x_j} \\
&\quad = \hat{\theta}_{\mathbf{1}_j} - \theta_{\mathbf{1}_j} + \sum_{\mathbf{i}\in\mathbb{I}:\, i_j\ge 1, |\mathbf{i}|\ge 2} i_j (\hat{\theta}_{\mathbf{i}} - \theta_{\mathbf{i}})(\boldsymbol{\mu} - \tilde{\boldsymbol{\mu}})^{\mathbf{i}-\mathbf{1}_j}.
\end{aligned} \tag{21}
$$

Now we bound the right-hand side of (21). Since $\mathbf{1}_j \in \mathbb{I}_1$, that is, $|\mathbf{1}_j| = 1$ for $j = 1, \dots, d$, and $\|\boldsymbol{\mu} - \tilde{\boldsymbol{\mu}}\| = o_p(\delta_n)$, we obtain by Lemma 4 that

$$\begin{aligned}\hat{\theta}_{\mathbf{1}_j} - \theta_{\mathbf{1}_j} &= O_p(n^{-1/2}\delta_n^{-1}) + O(\delta_n^{\alpha-1}) + O(\|\tilde{\boldsymbol{\mu}} - \boldsymbol{\mu}\|^{\alpha}\delta_n^{-1}) \\ &= O_p(n^{-1/2}\delta_n^{-1}) + O_p(\delta_n^{\alpha-1}), \qquad j = 1, \dots, d.\end{aligned} \tag{22}$$

The same argument applies to each term of the sum in the right-hand side of (21): for all $\mathbf{i} \in \mathbb{I}$ such that $i_j \ge 1$ and $|\mathbf{i}| \ge 2$

$$\begin{aligned}&|(\hat{\theta}_{\mathbf{i}} - \theta_{\mathbf{i}})(\boldsymbol{\mu} - \tilde{\boldsymbol{\mu}})^{\mathbf{i}-\mathbf{1}_j}| \\ &\quad \le |\hat{\theta}_{\mathbf{i}} - \theta_{\mathbf{i}}|\|\boldsymbol{\mu} - \tilde{\boldsymbol{\mu}}\|^{|\mathbf{i}-\mathbf{1}_j|} \\ &\quad = [O_p(n^{-1/2}\delta_n^{-|\mathbf{i}|}) + O(\delta_n^{\alpha-|\mathbf{i}|}) + O(\|\tilde{\boldsymbol{\mu}} - \boldsymbol{\mu}\|^{\alpha}\delta_n^{-|\mathbf{i}|})]\|\boldsymbol{\mu} - \tilde{\boldsymbol{\mu}}\|^{|\mathbf{i}|-1} \\ &\quad = o_p(n^{-1/2}\delta_n^{-1}) + o_p(\delta_n^{\alpha-1}) + o_p(\|\tilde{\boldsymbol{\mu}} - \boldsymbol{\mu}\|^{\alpha}\delta_n^{-1}) \\ &\quad = o_p(n^{-1/2}\delta_n^{-1}) + o_p(\delta_n^{\alpha-1}).\end{aligned} \tag{23}$$

There are fixed number of terms in the sum from (21) and the constant i_j is at most r_α. Combining this with (22) and (23), we see that the main term in (21) is $\hat{\theta}_{\mathbf{1}_j} - \theta_{\mathbf{1}_j}$ and therefore

$$\|\nabla f_{\hat{\boldsymbol{\theta}}}(\boldsymbol{\mu} - \tilde{\boldsymbol{\mu}}) - \nabla f_{\boldsymbol{\theta}}(\boldsymbol{\mu} - \tilde{\boldsymbol{\mu}})\| = O_p(n^{-1/2}\delta_n^{-1}) + O_p(\delta_n^{\alpha-1}). \qquad \square$$

For an $(s \times p)$-matrix $\mathbf{A}$, let $\|\mathbf{A}\| = \sup_{\mathbf{x}\in\mathbb{R}^p : \|\mathbf{x}\|\le 1} \|\mathbf{A}\mathbf{x}\|$ be the operator norm for the rest of this section and define the maximum norm $\|\mathbf{A}\|_{\max} = \max_{i,j} |a_{ij}|$, where a_{ij} are the entries of the matrix $\mathbf{A}$. These norms are related by

$$\|\mathbf{A}\|_{\max} \le \|\mathbf{A}\| \le \sqrt{sp}\|\mathbf{A}\|_{\max}. \tag{24}$$

LEMMA 6. *Assume* (A1), (A2), $\|\boldsymbol{\mu} - \tilde{\boldsymbol{\mu}}\| = o_p(\delta_n)$ *and* $\sqrt{n}\delta_n^2 \to \infty$. *For* $\boldsymbol{\mu}^* \in \mathbb{R}^d$ *such that* $\|\boldsymbol{\mu}^*\| = o_p(1)$ *and for any fixed* $\varepsilon \in (0, 1)$, *let*

$$B_n = \{\|Hf(\boldsymbol{\mu}) - Hf_{\hat{\boldsymbol{\theta}}}(\boldsymbol{\mu}^*)\| \le (1-\varepsilon)\|(Hf(\boldsymbol{\mu}))^{-1}\|^{-1}\}. \tag{25}$$

Then $\mathrm{P}(B_n) \to 1$ *as* $n \to \infty$, *on the event* B_n, $(Hf_{\hat{\boldsymbol{\theta}}}(\boldsymbol{\mu}^*))^{-1}$ *exists and*

$$\|(Hf(\boldsymbol{\mu}))^{-1} - (Hf_{\hat{\boldsymbol{\theta}}}(\boldsymbol{\mu}^*))^{-1}\| = o_p(1).$$

PROOF. Clearly, by the smoothness of a polynomial,

$$Hf_{\boldsymbol{\theta}}(\mathbf{z}) = Hf_{\boldsymbol{\theta}}(\mathbf{0}) + O(\|\mathbf{z}\|) \qquad \text{as } \|\mathbf{z}\| \to 0. \tag{26}$$

We note that the elements of the matrix $Hf_{\boldsymbol{\theta}}(\mathbf{0})$ [resp., $Hf_{\hat{\boldsymbol{\theta}}}(\mathbf{0})$] are linear combinations of $\theta_{\mathbf{i}}$ (resp., $\hat{\theta}_{\mathbf{i}}$), $\mathbf{i} \in \mathbb{I}_2$. From Lemma 4 and the conditions $\alpha > 2$, $\|\boldsymbol{\mu} - \tilde{\boldsymbol{\mu}}\| = o_p(\delta_n)$ and $\sqrt{n}\delta_n^2 \to \infty$, we obtain that

$$\hat{\theta}_{\mathbf{i}} - \theta_{\mathbf{i}} = O_p(n^{-1/2}\delta_n^{-2}) + O(\delta_n^{\alpha-2}) + O(\|\tilde{\boldsymbol{\mu}} - \boldsymbol{\mu}\|^{\alpha}\delta_n^{-2}) = o_p(1), \qquad \mathbf{i} \in \mathbb{I}_2,$$

where vector $\boldsymbol{\theta}$ is defined by (15). Therefore, entry-wise

$$Hf_{\hat{\boldsymbol{\theta}}}(\mathbf{0}) = Hf_{\boldsymbol{\theta}}(\mathbf{0}) + o_p(1). \tag{27}$$

By (A1), (A2) and the definition (15) of $\boldsymbol{\theta}$, $Hf(\boldsymbol{\mu}) = Hf_{\boldsymbol{\theta}}(\boldsymbol{\mu} - \tilde{\boldsymbol{\mu}})$. This and (26) imply that entry-wise

$$Hf(\boldsymbol{\mu}) = Hf_{\boldsymbol{\theta}}(\boldsymbol{\mu} - \tilde{\boldsymbol{\mu}}) = Hf_{\boldsymbol{\theta}}(\mathbf{0}) + O(\|\boldsymbol{\mu} - \tilde{\boldsymbol{\mu}}\|). \tag{28}$$

Combining (26), (27) and (28) leads to the following entry-wise relation:

$$\begin{aligned}
Hf_{\hat{\boldsymbol{\theta}}}(\boldsymbol{\mu}^*) &= Hf_{\hat{\boldsymbol{\theta}}}(\mathbf{0}) + O(\|\boldsymbol{\mu}^*\|) \\
&= Hf_{\boldsymbol{\theta}}(\mathbf{0}) + o_p(1) + O(\|\boldsymbol{\mu}^*\|) \\
&= Hf(\boldsymbol{\mu}) + O(\|\boldsymbol{\mu} - \tilde{\boldsymbol{\mu}}\|) + o_p(1) + O(\|\boldsymbol{\mu}^*\|) \\
&= Hf(\boldsymbol{\mu}) + o_p(1).
\end{aligned}$$

Then $\|Hf_{\hat{\boldsymbol{\theta}}}(\boldsymbol{\mu}^*) - Hf(\boldsymbol{\mu})\|_{\max} = o_p(1)$ and hence, by (24),

$$\|Hf(\boldsymbol{\mu}) - Hf_{\hat{\boldsymbol{\theta}}}(\boldsymbol{\mu}^*)\| = o_p(1). \tag{29}$$

Next, since (A1) and (A2) imply (6), $\lambda_{\min}(Hf(\boldsymbol{\mu})) \le \cdots \le \lambda_{\max}(Hf(\boldsymbol{\mu})) \le -\lambda_0 < 0$. Hence, $\|(Hf(\boldsymbol{\mu}))^{-1}\| = -(\lambda_{\max}(Hf(\boldsymbol{\mu})))^{-1} \le \lambda_0^{-1}$, or

$$\lambda_0 \le \|(Hf(\boldsymbol{\mu}))^{-1}\|^{-1}. \tag{30}$$

Define the event $C_n = \{\|Hf(\boldsymbol{\mu}) - Hf_{\hat{\boldsymbol{\theta}}}(\boldsymbol{\mu}^*)\| \le (1-\varepsilon)\lambda_0\}$. Using (30) and Lemma 11, we obtain that

$$C_n \subset B_n \subset \{(Hf_{\hat{\boldsymbol{\theta}}}(\boldsymbol{\mu}^*))^{-1} \text{ exists}\}.$$

In view of (29), $\mathrm{P}(C_n) \to 1$ and hence $\mathrm{P}(B_n) \to 1$. Finally, by applying (29), (30) and Lemma 11 again, we get that on the event B_n

$$\begin{aligned}
\|(Hf(\boldsymbol{\mu}))^{-1} - (Hf_{\hat{\boldsymbol{\theta}}}(\boldsymbol{\mu}^*))^{-1}\| &\le \varepsilon^{-1}\|(Hf(\boldsymbol{\mu}))^{-1}\|^2 \|Hf(\boldsymbol{\mu}) - Hf_{\hat{\boldsymbol{\theta}}}(\boldsymbol{\mu}^*)\| \\
&\le \varepsilon^{-1}\lambda_0^{-2}\|Hf(\boldsymbol{\mu}) - Hf_{\hat{\boldsymbol{\theta}}}(\boldsymbol{\mu}^*)\| = o_p(1). \qquad \square
\end{aligned}$$

REMARK 6. Lemma 6 would still hold if we only assumed that $\|\boldsymbol{\mu} - \tilde{\boldsymbol{\mu}}\| = O_p(\delta_n)$ instead of $\|\boldsymbol{\mu} - \tilde{\boldsymbol{\mu}}\| = o_p(\delta_n)$.

LEMMA 7. *Assume* (A1), (A2), $\|\boldsymbol{\mu} - \tilde{\boldsymbol{\mu}}\| = o_p(\delta_n)$, $\sqrt{n}\delta_n^2 \to \infty$ *and let* $A_n = \{\mathring{\boldsymbol{\mu}} \in C(2l\delta_n/3)\}$, *where the estimator* $\mathring{\boldsymbol{\mu}}$ *is defined by* (14). *Then* $\mathrm{P}(A_n) \to 1$ *as* $n \to \infty$.

PROOF. Bound $\mathrm{P}(A_n^c)$ by

$$\mathrm{P}(\mathring{\boldsymbol{\mu}} \notin C(2l\delta_n/3), \boldsymbol{\mu} - \tilde{\boldsymbol{\mu}} \in C(l\delta_n/3)) + \mathrm{P}(\boldsymbol{\mu} - \tilde{\boldsymbol{\mu}} \notin C(l\delta_n/3)). \tag{31}$$

The second term converges to zero by the condition $\|\boldsymbol{\mu} - \tilde{\boldsymbol{\mu}}\| = o_p(\delta_n)$.

For a symmetric matrix $\mathbf{M}$ and any $\mathbf{x} \in \mathbb{R}^d$, $\lambda_{\min}(\mathbf{M})\|\mathbf{x}\|^2 \le \mathbf{x}^T\mathbf{M}\mathbf{x} \le \lambda_{\max}(\mathbf{M})\|\mathbf{x}\|^2$. Recall that $\nabla f(\boldsymbol{\mu}) = \mathbf{0}$ and (6) follow from (A1) and (A2). Then, for $\boldsymbol{\mu} \in C(\tilde{\boldsymbol{\mu}}, l\delta_n/3)$ and $\mathbf{x} \in C(\tilde{\boldsymbol{\mu}}, l\delta_n) \setminus C(\tilde{\boldsymbol{\mu}}, 2l\delta_n/3)$, by using Taylor's expansion, $\nabla f(\boldsymbol{\mu}) = \mathbf{0}$ and (6), we have

$$\begin{aligned} f(\mathbf{x}) &= f(\boldsymbol{\mu}) + \frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^T Hf(\boldsymbol{\mu}^*)(\mathbf{x} - \boldsymbol{\mu}) \\ &\le f(\boldsymbol{\mu}) - \frac{\lambda_0}{2}\|\mathbf{x} - \boldsymbol{\mu}\|^2 \le f(\boldsymbol{\mu}) - c\delta_n^2 \end{aligned} \tag{32}$$

for some positive constant $c = c(\lambda_0, l)$ and sufficiently large n such that $\|\boldsymbol{\mu}^* - \boldsymbol{\mu}\| \le \kappa$, with $\kappa > 0$ from (6).

Next, by using (4), (15) and the c_r-inequality,

$$f_{\boldsymbol{\theta}}(\mathbf{z}) = P_{f,\boldsymbol{\mu}}(\mathbf{z} + \tilde{\boldsymbol{\mu}}) = f(\mathbf{z} + \tilde{\boldsymbol{\mu}}) + O(\|\mathbf{z}\|^\alpha) + O(\|\tilde{\boldsymbol{\mu}} - \boldsymbol{\mu}\|^\alpha).$$

Now we combine this with Lemma 4 and the conditions $\alpha > 2$, $\|\boldsymbol{\mu} - \tilde{\boldsymbol{\mu}}\| = o_p(\delta_n)$ and $\sqrt{n}\delta_n^2 \to \infty$ to obtain that, uniformly in $\mathbf{z} \in C(l\delta_n)$,

$$\begin{aligned} f_{\hat{\boldsymbol{\theta}}}(\mathbf{z}) &= f_{\boldsymbol{\theta}}(\mathbf{z}) + O_p(n^{-1/2}) + O(\delta_n^\alpha) \\ &= f(\mathbf{z} + \tilde{\boldsymbol{\mu}}) + O(\|\mathbf{z}\|^\alpha) + O(\|\tilde{\boldsymbol{\mu}} - \boldsymbol{\mu}\|^\alpha) + O_p(n^{-1/2}) + O(\delta_n^\alpha) \\ &= f(\mathbf{z} + \tilde{\boldsymbol{\mu}}) + o_p(\delta_n^2). \end{aligned} \tag{33}$$

Recall that $\mathring{\boldsymbol{\mu}} \in C(l\delta_n)$ by the definition (14). By (32) and (33), we see that the event

$$\begin{aligned} &\{\mathring{\boldsymbol{\mu}} \notin C(2l\delta_n/3), \boldsymbol{\mu} - \tilde{\boldsymbol{\mu}} \in C(l\delta_n/3)\} \\ &\quad = \{\mathring{\boldsymbol{\mu}} + \tilde{\boldsymbol{\mu}} \in C(\tilde{\boldsymbol{\mu}}, l\delta_n) \setminus C(\tilde{\boldsymbol{\mu}}, 2l\delta_n/3), \boldsymbol{\mu} \in C(\tilde{\boldsymbol{\mu}}, l\delta_n/3)\} \end{aligned}$$

implies the event

$$\begin{aligned} f(\boldsymbol{\mu}) - c\delta_n^2 &\ge f(\mathring{\boldsymbol{\mu}} + \tilde{\boldsymbol{\mu}}) = f_{\hat{\boldsymbol{\theta}}}(\mathring{\boldsymbol{\mu}}) + o_p(\delta_n^2) \\ &\ge f_{\hat{\boldsymbol{\theta}}}(\boldsymbol{\mu} - \tilde{\boldsymbol{\mu}}) + o_p(\delta_n^2) = f(\boldsymbol{\mu}) + o_p(\delta_n^2), \end{aligned}$$

leading to

$$\mathrm{P}(\mathring{\boldsymbol{\mu}} \notin C(2l\delta_n/3), \boldsymbol{\mu} - \tilde{\boldsymbol{\mu}} \in C(l\delta_n/3)) \le \mathrm{P}(c\delta_n^2 \le o_p(\delta_n^2)) \to 0$$

as $n \to \infty$. Combined with (31), this completes the proof of the lemma. $\square$

PROOF OF THEOREM 1. By (A1) and (A2), $\nabla f(\boldsymbol{\mu}) = \mathbf{0}$. According to the definition (15) of the polynomial $f_{\boldsymbol{\theta}}$,

$$\mathbf{0} = \nabla f(\boldsymbol{\mu}) = \nabla P_{f,\boldsymbol{\mu}}(\boldsymbol{\mu}) = \nabla f_{\boldsymbol{\theta}}(\boldsymbol{\mu} - \tilde{\boldsymbol{\mu}}). \tag{34}$$

By (14), $\max_{\mathbf{z}\in C(l\delta_n)} f_{\hat{\boldsymbol\theta}}(\mathbf{z}) = f_{\hat{\boldsymbol\theta}}(\mathring{\boldsymbol\mu})$. If this maximum is not attained on the boundary of $C(l\delta_n)$, then $\nabla f_{\hat{\boldsymbol\theta}}(\mathring{\boldsymbol\mu})$ must be zero. Hence we have that on the event $A_n = \{\mathring{\boldsymbol\mu} \in C(2l\delta_n/3)\}$

$$\mathbf{0} = \nabla f_{\hat{\boldsymbol\theta}}(\mathring{\boldsymbol\mu}) = \nabla f_{\hat{\boldsymbol\theta}}(\boldsymbol\mu - \tilde{\boldsymbol\mu}) + Hf_{\hat{\boldsymbol\theta}}(\boldsymbol\mu^*)\big(\mathring{\boldsymbol\mu} - (\boldsymbol\mu - \tilde{\boldsymbol\mu})\big), \tag{35}$$

where $\boldsymbol\mu^* = (\mu_1^*, \dots, \mu_d^*) = \lambda\mathring{\boldsymbol\mu} + (1-\lambda)(\boldsymbol\mu - \tilde{\boldsymbol\mu})$ for some $\lambda \in [0,1]$. Thus $\|\boldsymbol\mu^*\| = O(\|\mathring{\boldsymbol\mu}\|) + O(\|\boldsymbol\mu - \tilde{\boldsymbol\mu}\|) = O_p(\delta_n) = o_p(1)$.

By Lemma 6, $(Hf_{\hat{\boldsymbol\theta}}(\boldsymbol\mu^*))^{-1}$ exists on the event B_n defined by (25). Relations (34) and (35) imply that on the event $A_n \cap B_n$

$$\begin{aligned}
\hat{\boldsymbol\mu} - \boldsymbol\mu &= -\big(Hf_{\hat{\boldsymbol\theta}}(\boldsymbol\mu^*)\big)^{-1}\nabla f_{\hat{\boldsymbol\theta}}(\boldsymbol\mu - \tilde{\boldsymbol\mu})\\
&= -\big(Hf_{\hat{\boldsymbol\theta}}(\boldsymbol\mu^*)\big)^{-1}\big(\nabla f_{\hat{\boldsymbol\theta}}(\boldsymbol\mu - \tilde{\boldsymbol\mu}) - \nabla f_{\boldsymbol\theta}(\boldsymbol\mu - \tilde{\boldsymbol\mu})\big)\\
&= -\big(Hf(\boldsymbol\mu)\big)^{-1}\big(\nabla f_{\hat{\boldsymbol\theta}}(\boldsymbol\mu - \tilde{\boldsymbol\mu}) - \nabla f_{\boldsymbol\theta}(\boldsymbol\mu - \tilde{\boldsymbol\mu})\big) + r_n,
\end{aligned} \tag{36}$$

where $r_n = [(Hf(\boldsymbol\mu))^{-1} - (Hf_{\hat{\boldsymbol\theta}}(\boldsymbol\mu^*))^{-1}](\nabla f_{\hat{\boldsymbol\theta}}(\boldsymbol\mu - \tilde{\boldsymbol\mu}) - \nabla f_{\boldsymbol\theta}(\boldsymbol\mu - \tilde{\boldsymbol\mu}))$ is the remainder term.

By Lemma 5 and (30), we bound the norm of the first term on the right-hand side of (36) as

$$\begin{aligned}
&\big\|\big(Hf(\boldsymbol\mu)\big)^{-1}\big(\nabla f_{\hat{\boldsymbol\theta}}(\boldsymbol\mu - \tilde{\boldsymbol\mu}) - \nabla f_{\boldsymbol\theta}(\boldsymbol\mu - \tilde{\boldsymbol\mu})\big)\big\|\\
&\qquad \le \lambda_0^{-1}\big\|\nabla f_{\hat{\boldsymbol\theta}}(\boldsymbol\mu - \tilde{\boldsymbol\mu}) - \nabla f_{\boldsymbol\theta}(\boldsymbol\mu - \tilde{\boldsymbol\mu})\big\| = O_p(\gamma_n),
\end{aligned}$$

where $\gamma_n = n^{-1/2}\delta_n^{-1} + \delta_n^{\alpha-1}$. Therefore $\|r_n\| = o_p(1)O_p(\gamma_n) = o_p(\gamma_n)$ on the event B_n by Lemmas 5 and 6. Consequently on the event $A_n \cap B_n$, we have

$$\|\hat{\boldsymbol\mu} - \boldsymbol\mu\| = O_p\big(n^{-1/2}\delta_n^{-1} + \delta_n^{\alpha-1}\big) = O_p(\gamma_n). \tag{37}$$

For any constant $\rho > 0$,

$$\mathrm{P}\big(\|\hat{\boldsymbol\mu} - \boldsymbol\mu\| > \rho\gamma_n\big) \le \mathrm{P}\big(\{\|\hat{\boldsymbol\mu} - \boldsymbol\mu\| > \rho\gamma_n\} \cap A_n \cap B_n\big) + \mathrm{P}\big(A_n^c\big) + \mathrm{P}\big(B_n^c\big).$$

The first term on the right-hand side can be made arbitrarily small by choosing ρ sufficiently large in view of (37), uniformly in n, while the other two terms converge to zero by Lemmas 6 and 7. This proves (9).

It remains to prove (10). From (15) it follows that

$$M = f(\boldsymbol\mu) = f_{\boldsymbol\theta}(\boldsymbol\mu - \tilde{\boldsymbol\mu}) = \sum_{\mathbf{i}\in\mathbb{I}} \theta_{\mathbf{i}}(\boldsymbol\mu - \tilde{\boldsymbol\mu})^{\mathbf{i}},$$

so that, according to (14), $\hat M - M$ can be written as

$$\begin{aligned}
&f_{\hat{\boldsymbol\theta}}(\mathring{\boldsymbol\mu}) - f_{\boldsymbol\theta}(\boldsymbol\mu - \tilde{\boldsymbol\mu})\\
&\quad = \sum_{\mathbf{i}\in\mathbb{I}}\big[\hat\theta_{\mathbf{i}}\mathring{\boldsymbol\mu}^{\mathbf{i}} - \theta_{\mathbf{i}}(\boldsymbol\mu - \tilde{\boldsymbol\mu})^{\mathbf{i}}\big]\\
&\quad = \hat\theta_{\mathbf{i}_0} - \theta_{\mathbf{i}_0} + \sum_{\mathbf{i}\in\mathbb{I}, |\mathbf{i}|\ge 1}(\hat\theta_{\mathbf{i}} - \theta_{\mathbf{i}})\mathring{\boldsymbol\mu}^{\mathbf{i}} + \sum_{\mathbf{i}\in\mathbb{I}, |\mathbf{i}|\ge 1}\theta_{\mathbf{i}}\big[\mathring{\boldsymbol\mu}^{\mathbf{i}} - (\boldsymbol\mu - \tilde{\boldsymbol\mu})^{\mathbf{i}}\big].
\end{aligned} \tag{38}$$

By Lemma 4, the first term in (38) is

$$\hat{\theta}_{\mathbf{i}_0} - \theta_{\mathbf{i}_0} = O_p(n^{-1/2}) + O_p(\delta_n^\alpha). \tag{39}$$

From (14), (9) and the conditions $\sqrt{n}\delta_n^2 \to \infty$, $\alpha > 2$, $\|\boldsymbol{\mu} - \tilde{\boldsymbol{\mu}}\| = o_p(\delta_n)$, it follows that

$$\begin{aligned} \|\mathring{\boldsymbol{\mu}}\| &= \|\hat{\boldsymbol{\mu}} - \tilde{\boldsymbol{\mu}}\| \le \|\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}\| + \|\boldsymbol{\mu} - \tilde{\boldsymbol{\mu}}\| \\ &= O_p(n^{-1/2}\delta_n^{-1}) + O_p(\delta_n^{\alpha-1}) + O(\|\boldsymbol{\mu} - \tilde{\boldsymbol{\mu}}\|) = o_p(\delta_n). \end{aligned} \tag{40}$$

Using (40) and Lemma 4, each term in the second sum of (38)

$$|(\hat{\theta}_{\mathbf{i}} - \theta_{\mathbf{i}})\mathring{\boldsymbol{\mu}}^{\mathbf{i}}| \le |\hat{\theta}_{\mathbf{i}} - \theta_{\mathbf{i}}| \|\mathring{\boldsymbol{\mu}}\|^{|\mathbf{i}|} = o_p(n^{-1/2}) + o_p(\delta_n^\alpha),$$

so that, as there are a fixed number of terms in the sum,

$$\sum_{\mathbf{i}\in\mathbb{I}, |\mathbf{i}|\ge 1} (\hat{\theta}_{\mathbf{i}} - \theta_{\mathbf{i}})\mathring{\boldsymbol{\mu}}^{\mathbf{i}} = o_p(n^{-1/2}) + o_p(\delta_n^\alpha). \tag{41}$$

Now consider the third sum in (38). Combining Lemma 8 with (9), (40) and the condition $\|\boldsymbol{\mu} - \tilde{\boldsymbol{\mu}}\| = o_p(\delta_n)$, we obtain that for any $\mathbf{i} \in \mathbb{I}$, $|\mathbf{i}| \ge 1$,

$$\begin{aligned} |\mathring{\boldsymbol{\mu}}^{\mathbf{i}} - (\boldsymbol{\mu} - \tilde{\boldsymbol{\mu}})^{\mathbf{i}}| &\le \|\mathring{\boldsymbol{\mu}} - (\boldsymbol{\mu} - \tilde{\boldsymbol{\mu}})\| \sum_{k=1}^{|\mathbf{i}|} \|\mathring{\boldsymbol{\mu}}\|^{|\mathbf{i}|-k} \|\boldsymbol{\mu} - \tilde{\boldsymbol{\mu}}\|^{k-1} \\ &= \|\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}\| \sum_{k=1}^{|\mathbf{i}|} \|\mathring{\boldsymbol{\mu}}\|^{|\mathbf{i}|-k} \|\boldsymbol{\mu} - \tilde{\boldsymbol{\mu}}\|^{k-1} \\ &= o_p(n^{-1/2}\delta_n^{|\mathbf{i}|-2}) + o_p(\delta_n^{\alpha+|\mathbf{i}|-2}). \end{aligned} \tag{42}$$

Since $\mathbf{D}^{\mathbf{i}} P_{f,\boldsymbol{\mu}}(\mathbf{x})$, $\mathbf{i} \in \mathbb{I}$, are continuous, they are bounded over the compact set D, so that $\theta_{\mathbf{i}} = O_p(1)$, $\mathbf{i} \in \mathbb{I}$, in view of (16). Because of this and (42),

$$\sum_{\mathbf{i}\in\mathbb{I}, |\mathbf{i}|\ge 2} \theta_{\mathbf{i}}[\mathring{\boldsymbol{\mu}}^{\mathbf{i}} - (\boldsymbol{\mu} - \tilde{\boldsymbol{\mu}})^{\mathbf{i}}] = o_p(n^{-1/2}) + o_p(\delta_n^\alpha). \tag{43}$$

It remains to handle separately the terms in the third sum of (38) over $\mathbf{i} \in \mathbb{I}_1$, that is, $\mathbf{i} \in \mathbb{I}$ such that $|\mathbf{i}| = 1$. Due to (16) and the condition $\|\boldsymbol{\mu} - \tilde{\boldsymbol{\mu}}\| = o_p(\delta_n)$,

$$\theta_{\mathbf{i}} = \mathbf{D}^{\mathbf{i}} P_{f,\boldsymbol{\mu}}(\tilde{\boldsymbol{\mu}}) = O(\|\boldsymbol{\mu} - \tilde{\boldsymbol{\mu}}\|) = o_p(\delta_n), \qquad \mathbf{i} \in \mathbb{I}, |\mathbf{i}| = 1. \tag{44}$$

Then (42) and (44) imply that

$$\sum_{\mathbf{i}\in\mathbb{I}, |\mathbf{i}|=1} \theta_{\mathbf{i}}[\mathring{\boldsymbol{\mu}}^{\mathbf{i}} - (\boldsymbol{\mu} - \tilde{\boldsymbol{\mu}})^{\mathbf{i}}] = o_p(n^{-1/2}) + o_p(\delta_n^\alpha).$$

Finally, combining the last display with (38), (39), (41) and (43) completes the proof of (10). $\square$

REMARK 7. The above argument for estimating the parameter $M = f(\boldsymbol{\mu})$ can be refined for the problem of estimating any mixed derivative $\mathbf{D}^{\mathbf{i}} f(\boldsymbol{\mu})$, for $\mathbf{i} \in \mathbb{I}$, $|\mathbf{i}| \geq 2$. One can take the estimator $\mathbf{D}^{\mathbf{i}} f_{\hat{\boldsymbol{\theta}}}(\mathring{\boldsymbol{\mu}})$ and establish in a similar way that

$$\mathbf{D}^{\mathbf{i}} f_{\hat{\boldsymbol{\theta}}}(\mathring{\boldsymbol{\mu}}) - \mathbf{D}^{\mathbf{i}} f(\boldsymbol{\mu}) = O_p(n^{-1/2}\delta_n^{-|\mathbf{i}|}) + O_p(\delta_n^{\alpha-|\mathbf{i}|}), \qquad \mathbf{i} \in \mathbb{I}, |\mathbf{i}| \geq 2.$$

APPENDIX

LEMMA 8. *For any* $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$ *and any* $\mathbf{i} \in \mathbb{N}^d$ *such that* $|\mathbf{i}| \geq 1$,

$$|\mathbf{x}^{\mathbf{i}} - \mathbf{y}^{\mathbf{i}}| \leq \|\mathbf{x} - \mathbf{y}\| \sum_{k=1}^{|\mathbf{i}|} \|\mathbf{x}\|^{|\mathbf{i}|-k} \|\mathbf{y}\|^{k-1}.$$

PROOF. We prove the lemma by induction in dimension. For $d = 1$,

$$x^i - y^i = (x - y) \sum_{k=1}^{i} x^{i-k} y^{k-1}$$

and the statement follows.

Now we handle the inductive step. Suppose the statement is true for all dimensions $k = 1, \ldots, d-1$. We want to show that it also holds for the dimension d. Without loss of generality assume that $i_1 > 0$. Recall the notation $\mathbf{x}_{-1} = (x_2, \ldots, x_d)$, $\mathbf{i}_{-1} = (i_2, \ldots, i_d)$ that we used in Lemma 1. We have

$$\begin{aligned}\mathbf{x}^{\mathbf{i}} - \mathbf{y}^{\mathbf{i}} &= \mathbf{x}^{\mathbf{i}} - x_1^{i_1} \mathbf{y}_{-1}^{\mathbf{i}_{-1}} + x_1^{i_1} \mathbf{y}_{-1}^{\mathbf{i}_{-1}} - \mathbf{y}^{\mathbf{i}} \\ &= x_1^{i_1} (\mathbf{x}_{-1}^{\mathbf{i}_{-1}} - \mathbf{y}_{-1}^{\mathbf{i}_{-1}}) + (x_1^{i_1} - y_1^{i_1}) \mathbf{y}_{-1}^{\mathbf{i}_{-1}}.\end{aligned}$$

Obviously, $|x_1| \leq \|\mathbf{x}\|$, $\|\mathbf{x}_{-1}\| \leq \|\mathbf{x}\|$ and $|\mathbf{i}| = |\mathbf{i}_{-1}| + i_1$. Using these relations and the assumption of the inductive step, we obtain that

$$\begin{aligned}|x_1^{i_1}(\mathbf{x}_{-1}^{\mathbf{i}_{-1}} - \mathbf{y}_{-1}^{\mathbf{i}_{-1}})| &\leq |x_1|^{i_1} \|\mathbf{x}_{-1} - \mathbf{y}_{-1}\| \sum_{k=1}^{|\mathbf{i}_{-1}|} \|\mathbf{x}_{-1}\|^{|\mathbf{i}_{-1}|-k} \|\mathbf{y}_{-1}\|^{k-1} \\ &\leq \|\mathbf{x} - \mathbf{y}\| \sum_{k=1}^{|\mathbf{i}|-i_1} \|\mathbf{x}\|^{|\mathbf{i}|-k} \|\mathbf{y}\|^{k-1}\end{aligned}$$

and

$$\begin{aligned}|(x_1^{i_1} - y_1^{i_1}) \mathbf{y}_{-1}^{\mathbf{i}_{-1}}| &\leq \|\mathbf{y}_{-1}\|^{\mathbf{i}_{-1}} |x_1 - y_1| \sum_{k=1}^{i_1} |x_1|^{i_1-k} |y_1|^{k-1} \\ &\leq \|\mathbf{x} - \mathbf{y}\| \sum_{k=|\mathbf{i}|-i_1+1}^{|\mathbf{i}|} \|\mathbf{x}\|^{|\mathbf{i}|-k} \|\mathbf{y}\|^{k-1}.\end{aligned}$$

Combining the last three relations, we obtain the desired result. $\square$

Below, we consider $s \times s$ matrices and let $\mathbf{I}$ denote the identity matrix of order s. Let $\|\mathbf{A}\|$ be some norm on the space of $s \times s$ matrices satisfying the multiplicative property $\|\mathbf{AB}\| \le \|\mathbf{A}\|\|\mathbf{B}\|$. For example, the operator norm satisfies this property.

LEMMA 9 (Banach's lemma). *Let* $\mathbf{M}$ *be a matrix with* $\|\mathbf{M}\| < 1$. *Then* $\mathbf{I} - \mathbf{M}$ *is invertible*, $(\mathbf{I} - \mathbf{M})^{-1} = \mathbf{I} + \mathbf{M} + \mathbf{M}^2 + \cdots$ *and* $\|(\mathbf{I} - \mathbf{M})^{-1}\| \le (1 - \|\mathbf{M}\|)^{-1}$.

The proof of Banach's lemma can be found in many textbooks on functional analysis. The next two lemmas are essentially adopted from Facer and Müller (2003) with some modifications.

LEMMA 10. *Let* $\mathbf{V}$ *be invertible and* $\mathbf{W}$ *be such that* $\|\mathbf{W}\| < \|\mathbf{V}^{-1}\|^{-1}$. *Then* $\mathbf{V} + \mathbf{W}$ *is invertible and*

$$(\|\mathbf{V}\| + \|\mathbf{W}\|)^{-1} \le \|(\mathbf{V} + \mathbf{W})^{-1}\| \le \frac{\|\mathbf{V}^{-1}\|}{1 - \|\mathbf{V}^{-1}\mathbf{W}\|}.$$

PROOF. Since $\|\mathbf{V}^{-1}\mathbf{W}\| < 1$ due to the condition $\|\mathbf{W}\| < \|\mathbf{V}^{-1}\|^{-1}$, the matrix $(\mathbf{I} + \mathbf{V}^{-1}\mathbf{W})$ is invertible and $\|(\mathbf{I} + \mathbf{V}^{-1}\mathbf{W})^{-1}\| \le (1 - \|\mathbf{V}^{-1}\mathbf{W}\|)^{-1}$ by Banach's lemma. Therefore, $\mathbf{V} + \mathbf{W} = \mathbf{V}(\mathbf{I} + \mathbf{V}^{-1}\mathbf{W})$ is also invertible and

$$\begin{aligned}\|(\mathbf{V} + \mathbf{W})^{-1}\| &= \|(\mathbf{I} + \mathbf{V}^{-1}\mathbf{W})^{-1}\mathbf{V}^{-1}\| \\ &\le \|\mathbf{V}^{-1}\|\|(\mathbf{I} + \mathbf{V}^{-1}\mathbf{W})^{-1}\| \le \frac{\|\mathbf{V}^{-1}\|}{1 - \|\mathbf{V}^{-1}\mathbf{W}\|}.\end{aligned}$$

Now, using $\|\mathbf{V} + \mathbf{W}\| \le \|\mathbf{V}\| + \|\mathbf{W}\|$ and the invertibility of $\mathbf{V} + \mathbf{W}$, we obtain $\|(\mathbf{V} + \mathbf{W})^{-1}\| \ge \|\mathbf{V} + \mathbf{W}\|^{-1} \ge (\|\mathbf{V}\| + \|\mathbf{W}\|)^{-1}$. $\square$

LEMMA 11. *Let* $\mathbf{A}$ *be invertible and* $\mathbf{B}$ *be such that* $\|\mathbf{A} - \mathbf{B}\| \le (1 - \varepsilon)\|\mathbf{A}^{-1}\|^{-1}$ *for some* $\varepsilon \in (0, 1]$. *Then* $\mathbf{B}$ *is invertible and*

$$\|\mathbf{B}^{-1} - \mathbf{A}^{-1}\| \le \varepsilon^{-1}\|\mathbf{A}^{-1}\|^2\|\mathbf{A} - \mathbf{B}\|.$$

PROOF. Write $\mathbf{B} = \mathbf{A} + (\mathbf{B} - \mathbf{A})$ and apply Lemma 10 with $\mathbf{V} = \mathbf{A}$ and $\mathbf{W} = \mathbf{B} - \mathbf{A}$ to conclude that $\mathbf{B}$ is invertible and, as $\|\mathbf{A}^{-1}(\mathbf{B} - \mathbf{A})\| \le 1 - \varepsilon$ by the condition of the lemma,

$$\|\mathbf{B}^{-1}\| \le \frac{\|\mathbf{A}^{-1}\|}{1 - \|\mathbf{A}^{-1}(\mathbf{B} - \mathbf{A})\|} \le \frac{\|\mathbf{A}^{-1}\|}{1 - (1 - \varepsilon)} = \varepsilon^{-1}\|\mathbf{A}^{-1}\|.$$

By using the last relation, we complete the proof,

$$\begin{aligned}\|\mathbf{B}^{-1} - \mathbf{A}^{-1}\| &\le \|\mathbf{A}^{-1}\|\|\mathbf{A}\mathbf{B}^{-1} - \mathbf{I}\| \\ &\le \|\mathbf{A}^{-1}\|\|\mathbf{A} - \mathbf{B}\|\|\mathbf{B}^{-1}\| \le \varepsilon^{-1}\|\mathbf{A}^{-1}\|^2\|\mathbf{A} - \mathbf{B}\|.\end{aligned}$$ $\square$

Acknowledgments. The comments of the referees and the Associate Editor helped us to better organize the paper.

REFERENCES

BLUM, J. R. (1954). Multidimensional stochastic approximation methods. *Ann. Math. Statistics* **25** 737–744. MR0065092

CHEN, H. (1988). Lower rate of convergence for locating a maximum of a function. *Ann. Statist.* **16** 1330–1334. MR0959206

DIPPON, J. (2003). Accelerated randomized stochastic optimization. *Ann. Statist.* **31** 1260–1281. MR2001650

FACER, M. R. and MÜLLER, H.-G. (2003). Nonparametric estimation of the location of a maximum in a response surface. *J. Multivariate Anal.* **87** 191–217. MR2007268

FAN, J. and GIJBELS, I. (1996). *Local Polynomial Modelling and Its Applications*. *Monographs on Statistics and Applied Probability* **66**. Chapman & Hall, London. MR1383587

HASMINSKIĬ, R. Z. (1979). Lower bound for the risks of nonparametric estimates of the mode. In *Contributions to Statistics* (J. Jureckova, ed.) 91–97. Reidel, Dordrecht. MR0561262

KIEFER, J. and WOLFOWITZ, J. (1952). Stochastic estimation of the maximum of a regression function. *Ann. Math. Statistics* **23** 462–466. MR0050243

KLEMELÄ, J. (2005). Adaptive estimation of the mode of a multivariate density. *J. Nonparametr. Stat.* **17** 83–105. MR2112688

LAN, Y., BANERJEE, M. and MICHAILIDIS, G. (2009). Change-point estimation under adaptive sampling. *Ann. Statist.* **37** 1752–1791. MR2533471

MOKKADEM, A. and PELLETIER, M. (2007). A companion for the Kiefer–Wolfowitz–Blum stochastic approximation algorithm. *Ann. Statist.* **35** 1749–1772. MR2351104

MÜLLER, H.-G. (1985). Kernel estimators of zeros and of location and size of extrema of regression functions. *Scand. J. Stat.* **12** 221–232. MR0817940

MÜLLER, H.-G. (1989). Adaptive nonparametric peak estimation. *Ann. Statist.* **17** 1053–1069. MR1015137

POLYAK, B. T. and TSYBAKOV, A. B. (1990). Optimal orders of accuracy for search algorithms of stochastic optimization. *Problems Inform. Transmission* **26** 126–133.

SHOUNG, J.-M. and ZHANG, C.-H. (2001). Least squares estimators of the mode of a unimodal regression function. *Ann. Statist.* **29** 648–665. MR1865335

TANG, R., BANERJEE, M. and MICHAILIDIS, G. (2011). A two-stage hybrid procedure for estimating an inverse regression function. *Ann. Statist.* **39** 956–989. MR2816344

E. BELITSER
DEPARTMENT OF MATHEMATICS
EINDHOVEN UNIVERSITY OF TECHNOLOGY
P.O. BOX 513
5600 MB EINDHOVEN
THE NETHERLANDS
E-MAIL: e.n.belitser@tue.nl

S. GHOSAL
DEPARTMENT OF STATISTICS
NORTH CAROLINA STATE UNIVERSITY
4276 SAS HALL, 2311 STINSON DRIVE
RALEIGH, NORTH CAROLINA 27695-8203
USA
E-MAIL: sghosal@stat.ncsu.edu

H. VAN ZANTEN
KORTEWEG-DE VRIES INSTITUTE FOR MATHEMATICS
UNIVERSITY OF AMSTERDAM
P.O. BOX 94248
1090 GE AMSTERDAM
THE NETHERLANDS
E-MAIL: j.h.vanzanten@uva.nl