

OPTIMAL TWO-STAGE PROCEDURES FOR ESTIMATING LOCATION AND SIZE OF MAXIMUM OF MULTIVARIATE REGRESSION FUNCTIONS

BY EDUARD BELITSER[†], SUBHASHIS GHOSAL^{‡,*}
AND HARRY VAN ZANTEN[†]

Eindhoven University of Technology[†] and North Carolina State University[‡]

Abstract We propose a two-stage procedure for estimating the location $\boldsymbol{\mu}$ and size M of maximum of smooth d -variate regression functions $f(\mathbf{x})$. 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 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 with $\alpha > 1 + \sqrt{1 + d}/2$, 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, which are the optimal rates in the class of all possible sequential estimators. Interestingly, the two-stage procedure resolves the curse of 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 obtains the optimal rate for any smoothness level $\alpha > 2$ starting with a preliminary estimator with any power-law rate at the first stage. Adaptive properties of multi-stage approach are also discussed.

1. Introduction. In many applications, it is of interest to estimate the location and size of the extremum of a regression function in an additive nonparametric regression model, possibly with more than one regressors. For instance, an oil company may be interested in determining the best location for drilling a well in a confined region in terms of output. Based on information obtained from drilling at a few preliminary locations in the

*Part of this work was done when the second author was visiting EURANDOM, Eindhoven, supported by a grant from NWO of the Netherlands.

AMS 2000 subject classifications: Primary 62L12; secondary 62G05, 62H12, 62L05

Keywords and phrases: two-stage procedure, optimal rate in sequential design setting, multi-stage adaptive

region, the goal is obtain an estimate of the best location and the amount of the reserve based on these noisy measurements.

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

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

where ξ_k 's are independent zero mean errors with $\text{Var}(\xi_k) = \sigma^2$. It is easy to verify that this homoscedasticity condition can be relaxed to $\text{Var}(\xi_k) \leq \sigma^2$ for all our convergence rate bounds below except that optimality of the rates cannot be concluded. Assume that f has a unique maximum at $\boldsymbol{\mu}$ on D at $\boldsymbol{\mu}$ in the interior of D , i.e.,

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

If the function f is sufficiently smooth, the gradient $\nabla f(\boldsymbol{\mu}) = 0$ and the Hessian matrix of f at $\boldsymbol{\mu}$ is negative 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 $\{\mathbf{x}_k, k = 1, \dots, n\}$ is essential for this problem. 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 extremum, 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; Müller (1985), (1989), Shoung and Zhang (2001), and 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 (for example, isotropic Hölder class defined below) is $n^{-\alpha/(2\alpha+d)}$ up to a logarithmic factor. As to the estimation of the location of the extremum, the minimax rate coincides with 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 $n^{-1/5}$. 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 level higher than 2.

If we can choose the design point for each observation of the regression function f , then we are in the classical sequential design setting. Kiefer and Wolfowitz (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 μ of the maximum of multivariate regression functions. Since then, this Kiefer-Wolfowitz-Blum recursive algorithm has been extended in many directions by many authors. The main fact is that the algorithm converges to μ 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 algorithm which attains the optimal convergence rate. Mokkadem and Pelletier (2007) considered the problem of simultaneous estimation, 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-Wolfowitz-Blum algorithm so that, by applying both the companion and the Kiefer-Wolfowitz-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 in estimating the maximum itself $M = f(\mu)$ 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-Wolfowitz-Blum procedure.

In this paper, we propose a two-stage strategy to tackle the problem of simultaneously estimating the location μ and size M of the maximum of the regression function f according to the observation scheme (1). In a way, 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 plan. Our findings establish that the two-stage procedure can be properly designed to match the strength of a fully sequential procedure.

Now we describe the two-stage procedure. We construct a preliminary estimator $\tilde{\mu}$ of μ by spending a portion of our sampling budget to make

observations over a relatively uniformly 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” 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 on our approach is reminiscent to 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 Tang, Banerjee and Michailidis (2011), who considered such a procedure for estimating the level point of a univariate monotone regression function. Motivating grounds for two-stage approach were nicely described by them. The principal differences between their and our techniques are that we consider smooth rather than monotone functions and need to use polynomial regression of an appropriate degree in the second stage rather than linear regression 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 sequential design Robbins-Monro type settings one needs to adjust the design points to estimate M . This can place serious constrain on the available budget if both $\boldsymbol{\mu}$ and M need to be estimated using the sequential procedure.

Our main result gives a decomposition of the convergence rate of the second stage estimator as the sum of an approximation term (analog of the bias term) and the stochastic term (analog of the variance 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 (in the minimax sense, with respect to a class of α -smooth regression functions) convergence rate $n^{-(\alpha-1)/(2\alpha+d)}$ in the fixed design setting. 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 $\alpha > 1 + \sqrt{1+d/2}$. Thus, for α -smooth regression functions, the second stage improves the rates in estimating $\boldsymbol{\mu}$ and M from 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 $\alpha > 3$, 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 procedures, 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 a finite number of stages the optimal convergence rates are attained. 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. In its basic form, the method still uses knowledge of the smoothness level α in its formulation, and hence is not adaptive. However, if we know that $\alpha \geq \alpha_1$ for some $\alpha_1 > 2$, we can apply the multi-stage approach with $\alpha = \alpha_1$ to derive the rate $n^{-(\alpha_1-1)/(2\alpha_1)}$ in estimating the location of the maximum. Remarkably, this approach does provides a completely adaptive estimator with the optimal rate $n^{-1/2}$ for the size of the maximum of the regression function.

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

2. Notations, preliminaries and assumptions. We describe the notations and conventions to be used in this paper. For numerical sequences β_n and β'_n , by $\beta_n \ll \beta'_n$ (or $\beta'_n \gg \beta_n$) we mean that $\beta_n = o_p(\beta'_n)$ and we say that there is an improvement from β'_n to β_n . By $\beta_n \asymp \beta'_n$ we mean that $\beta_n = O_p(\beta'_n)$ and $\beta'_n = O(\beta_n)$. All asymptotic relations and symbols (like $O(\delta_n)$, $O_p(n^{-1/2})$ etc.) will refer to the asymptotic regime $n \rightarrow \infty$ unless otherwise is specified. Let \mathbb{N} stand for $\{0, 1, 2, \dots\}$. For a set S , denote by $|S|$ the number of elements in S . Introduce vector notations $\mathbf{x}, \mathbf{x}_k \in \mathbb{R}^d$: $\mathbf{x} = (x_1, \dots, x_d)$ and $\mathbf{x}_k = (x_{k,1}, \dots, x_{k,d})$. By $\|\mathbf{x}\|$ for a vector \mathbf{x} , we mean the usual Euclidean norm of $\mathbf{x} \in \mathbb{R}^d$. If \mathbf{A} is a matrix, $\|\mathbf{A}\|$ will stand for a norm on the space of matrices such as the operator norm $\|\mathbf{A}\| = \sup_{\|\mathbf{x}\| \leq 1} \|\mathbf{A}\mathbf{x}\|$. Let $B(\mathbf{c}, R) = \{\mathbf{z} \in \mathbb{R}^d : \|\mathbf{z} - \mathbf{c}\| \leq R\}$ 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 $\mathbf{a} = (a_1, \dots, a_d) \in \mathbb{R}^d$ with an edge length 2δ by

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

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

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

$$|\mathbf{i}| = \sum_{k=1}^d i_k, \quad \mathbf{i}! = \prod_{k=1}^d i_k!, \quad \mathbf{x}^{\mathbf{i}} = \prod_{k=1}^d x_k^{i_k}, \quad \partial^{\mathbf{i}} f(\mathbf{x}_0) = \left. \frac{\partial^k f(\mathbf{x})}{\partial x_1^{i_1} \dots \partial x_d^{i_d}} \right|_{\mathbf{x}=\mathbf{x}_0}.$$

Also introduce the special multi-indexes $\mathbf{i}_0 = (0, \dots, 0)$ and

$$\mathbf{i}_k = (0, \dots, 0, 1, 0, \dots, 0), \quad k = 1, \dots, d,$$

with 1 on the k th place.

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

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

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

$\binom{d+k-1}{k} = \binom{d+k-1}{d-1}$, and hence

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

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

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 $\mathcal{H}_d(\alpha, L) = \mathcal{H}_d(\alpha, L, D)$, consisting of r_α -times differentiable functions $f : D \rightarrow \mathbb{R}$ such that

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

where

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

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

For a function $g : \mathbb{R}^d \rightarrow \mathbb{R}$ such that all second-order partial derivatives of g exist at point $\mathbf{x}_0 \in \mathbb{R}^d$, denote by $Hg(\mathbf{x}_0)$ the Hessian matrix of the function g at point \mathbf{x}_0 :

$$(6) \quad Hg(\mathbf{x}_0) = \left(\left(\frac{\partial^2 g(\mathbf{x}_0)}{\partial x_j \partial x_i} \right) \right)_{i=1, j=1}^{d, d}.$$

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 biggest eigenvalues of \mathbf{M} , respectively.

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

- (A1) The function $f(\mathbf{x})$, $\mathbf{x} \in D \subseteq \mathbb{R}^d$, belongs to an isotropic Hölder functional class $\mathcal{H}_d(\alpha, L, D)$ defined by (4), with $L > 0$ and $\alpha > 2$.
- (A2) There is a unique point $\boldsymbol{\mu}$ in the interior $\overset{\circ}{D}$ of D that maximizes the function f on D , i.e. $M = \max_{\mathbf{x} \in D} f(\mathbf{x}) = \sup_{\mathbf{x} \in \overset{\circ}{D}} f(\mathbf{x}) = f(\boldsymbol{\mu})$, $\boldsymbol{\mu} \in \overset{\circ}{D}$.
- (A3) There exist $\kappa, \lambda_0 > 0$ such that $\sup_{\mathbf{x} \in B(\boldsymbol{\mu}, \kappa)} \lambda_{\max}(Hf(\mathbf{x})) \leq -\lambda_0 < 0$.

Put $q = q(\alpha, d) = |\mathbf{I}(r_\alpha, d)| - 1$ and $\mathbf{I} = \mathbf{I}(r_\alpha, d)$ from now on. Thus \mathbf{I} consists of $q + 1$ elements. Observe that the total number of terms in the d -variate Taylor polynomial $P_{f, \mathbf{x}_0}(\mathbf{x})$ of order $r_\alpha = \lceil \alpha - 1 \rceil$ defined in (5) has also cardinality $q + 1$.

REMARK 1. Without loss of generality, assume further that the regression function $f(\mathbf{x})$ originally defined on the set D , allows extension on an ϵ -neighborhood $D_\epsilon = \bigcup_{\mathbf{x} \in D} B(\mathbf{x}, \epsilon)$ of D for some $\epsilon > 0$, preserving the Hölder $\mathcal{H}_d(\alpha, L)$ -smoothness on D_ϵ , in order to avoid boundary effects.

REMARK 2. Conditions (A1)–(A3) are not completely independent. For example, conditions (A1)–(A2) imply in particular that $\nabla f(\boldsymbol{\mu}) = \mathbf{0}$ and the Hessian $Hf(\boldsymbol{\mu})$ is symmetric and negative definite matrix. Besides, as $\alpha > 2$, the Hessian matrix $Hf(\mathbf{x})$ is continuous and therefore for some κ, λ_0 depending on the underlying function f ,

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

So, if we do not pursue any uniformity over f , then condition (A3) follows from (A1) and (A2) and is in principle redundant. However, condition (A3) imposes a uniformity requirement on the constants κ and λ_0 , which is necessary when generalizing the result uniformly over a functional class.

3. The two-stage procedure. For a column vector $\boldsymbol{\vartheta} = (\vartheta_{\mathbf{i}}, \mathbf{i} \in \mathbf{I}(r_\alpha, d)) = (\vartheta_{\mathbf{i}_0}, \vartheta_{\mathbf{i}_1}, \dots, \vartheta_{\mathbf{i}_q})^T$, introduce the multivariate polynomial function

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

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

1. The first stage starts is as follows. Take a fraction $v \in (0, 1)$ of the design budget, i.e. $n_1 = n_1(n) \in \mathbb{N}$ such that $0 < n_1 < n$, $n_1/n \rightarrow v$. Next, allocate 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 $\|\tilde{\mathbf{x}}_i - \tilde{\mathbf{x}}_j\| \geq c_2 n^{-1/d}$ for $i \neq j$.

Observe the data according to the model (1): $D_1 = \{(\tilde{\mathbf{x}}_i, \tilde{Y}_i), i = 1, \dots, n_1\}$, $\tilde{Y}_i = f(\tilde{\mathbf{x}}_i) + \tilde{\xi}_i, i = 1, \dots, n_1$.

2. Using the data D_1 , construct a preliminary consistent estimator $\tilde{\boldsymbol{\mu}} = \tilde{\boldsymbol{\mu}}(D_1) = (\tilde{\mu}_1, \dots, \tilde{\mu}_d)$ of $\boldsymbol{\mu} = (\mu_1, \dots, \mu_d)$:

$$\|\tilde{\boldsymbol{\mu}} - \boldsymbol{\mu}\| = o_p(1) \quad \text{as } n \rightarrow \infty.$$

One can use for example a kernel estimator from Müller (1989) for $d = 1$ and 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 $2l \geq r_\alpha$. Assume without loss of generality that $n_2 = n_3(2l+1)^d$ for some $n_3 \in \mathbb{N}$. Notice that $n_3 \geq cn$ for some constant $c > 0$. Introduce a *localization parameter* $\delta_n > 0$, $\delta_n \rightarrow 0$, and define the set

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

which consists of $(2l+1)^d$ different points from a 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 $I_j = \{1 \leq k \leq n_2 : \mathbf{x}_k = \tilde{\mathbf{d}}_j\}$. In words, each point among $(2l+1)^d$ 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 $D_2 = \{(\mathbf{x}_k, Y_k), k = 1, \dots, n_2\}$ according to the model (1):

$$(8) \quad Y_k = f(\mathbf{x}_k) + \xi_k, \quad k = 1, \dots, n_2.$$

4. Introduce the vectors $\mathbf{Y} = (Y_1, \dots, Y_{n_2})^T$, $\mathbf{X}_k = (\mathbf{x}_k^{i_0}, \mathbf{x}_k^{i_1}, \dots, \mathbf{x}_k^{i_q})$, $k = 1, \dots, n_2$, and the matrix $\mathbf{X} = (\mathbf{X}_1^T, \dots, \mathbf{X}_{n_2}^T)^T$ of dimension $n_2 \times (q+1)$. Now fit the data D_2 given by (8) in the polynomial regression model of order r_α :

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

where the polynomial $f_{\boldsymbol{\theta}}$ is introduced by (7). The unique least squares solution is given by $\tilde{\boldsymbol{\theta}} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{Y}$ as long as \mathbf{X} has linearly independent columns. This is assured by our choice of the second stage design points $\{\mathbf{x}_k, k = 1, \dots, n_2\}$ as we have at least $r_\alpha + 1$ (in fact, $2l+1 \geq r_\alpha + 1$) distinct design points in each dimension and certainly $n_2 \geq (2l+1)^d \geq (r_\alpha + 1)^d \geq |\mathbf{I}(r_\alpha, d)| = q + 1$. A formal proof of the columns independence of the matrix \mathbf{X} is given below in Lemma 1.

5. Finally, derive the estimator $(\hat{\boldsymbol{\mu}}, \hat{M}) = (\hat{\boldsymbol{\mu}}(\delta_n, \tilde{\boldsymbol{\mu}}), \hat{M}(\delta_n, \tilde{\boldsymbol{\mu}}))$ of $(\boldsymbol{\mu}, M)$ as follows:

$$(9) \quad \hat{\boldsymbol{\mu}} = \arg \max_{\mathbf{x} \in C(\tilde{\boldsymbol{\mu}}, l, \delta_n)} f_{\hat{\boldsymbol{\theta}}}(\mathbf{x}), \quad \hat{M} = f_{\hat{\boldsymbol{\theta}}}(\hat{\boldsymbol{\mu}}),$$

where cube C is defined by (3), $\tilde{\boldsymbol{\mu}}$ is the preliminary estimator for $\boldsymbol{\mu}$ from the first stage of the estimation procedure and δ_n is the localization parameter introduced at step 3.

Notice that the estimators $\hat{\boldsymbol{\mu}} = \hat{\boldsymbol{\mu}}(\delta_n, \tilde{\boldsymbol{\mu}})$ and $\hat{M} = \hat{M}(\delta_n, \tilde{\boldsymbol{\mu}})$ depend on the localization parameter δ_n and the estimator $\tilde{\boldsymbol{\mu}}$ from the first stage of the estimation procedure. We however do not emphasize this dependence most of the time in order to avoid overloaded notations.

Now we are ready to formulate the main result.

THEOREM 1. *Assume (A1)–(A3) and let $(\hat{\boldsymbol{\mu}}, \hat{M}) = (\hat{\boldsymbol{\mu}}(\delta_n, \tilde{\boldsymbol{\mu}}), \hat{M}(\delta_n, \tilde{\boldsymbol{\mu}}))$ be the two-stage estimator defined by (9) (or equivalently, by (21)), where the localization parameter δ_n and the preliminary estimator $\tilde{\boldsymbol{\mu}}$ are such that $\|\tilde{\boldsymbol{\mu}} - \boldsymbol{\mu}\| = o_p(\delta_n)$ and $n^{-1/2} = o(\delta_n^2)$. Then, as $n \rightarrow \infty$,*

$$(10) \quad \|\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}\| = O_p(n^{-1/2}\delta_n^{-1}) + O_p(\delta_n^{\alpha-1})$$

and

$$(11) \quad \hat{M} - M = O_p(n^{-1/2}) + O_p(\delta_n^\alpha).$$

REMARK 3. The two-stage approach provides simultaneously the estimators for the location and the size of the maximum of α -smooth regression functions, since the same design points for the both estimators are used in the procedure.

REMARK 4. The two-stage approach is adaptive with respect to the noise variance σ^2 provided the preliminary estimator $\tilde{\boldsymbol{\mu}}$ does not use knowledge of σ^2 .

REMARK 5. The condition $\|\tilde{\boldsymbol{\mu}} - \boldsymbol{\mu}\| = o_p(\delta_n)$ has a clear heuristic interpretation: one should not localize at the second stage more than what the accuracy of the estimation procedure allows at the first stage. This condition can be relaxed to $\delta_n \geq K\|\tilde{\boldsymbol{\mu}} - \boldsymbol{\mu}\|$ for sufficiently large K and all sufficiently large n .

REMARK 6. A generalization of the main theorem is possible for an anisotropic Hölder class, i.e. with possibly different smoothness properties in different directions. Presumably, the approach will go through with the adjustments for the localization parameter $\boldsymbol{\delta}_n = (\delta_{1,n}, \dots, \delta_{d,n})$, which will become a vector with localization parameters pertinent to the different smoothness properties in different directions as coordinates.

REMARK 7. One can formulate a version of the above theorem, in which the main statements hold uniformly over regression functions from certain nonparametric functional class. This class is essentially a Hölder smoothness class $\mathcal{H}_d(\alpha, L, D)$ (with $\alpha > 2$), restricted by certain conditions. By checking the proofs, one can see that all the second stage results in the paper can be made uniform over the Hölder class $\mathcal{H}_d(\alpha, L, D)$ if we assume additionally the uniform boundedness of all the partial derivatives involved in the definition of $\mathcal{H}_d(\alpha, L, D)$ and uniformity holds in the first stage results.

The following class can serve as an example. Let $\tilde{\mathcal{H}}(\alpha, L, D, L_1, \kappa_1, \delta, \kappa, \lambda_0)$ consists of all functions $f : D \rightarrow \mathbb{R}$, such that the following conditions are fulfilled:

- ($\tilde{\text{A1}}$) For $\alpha > 2$ and $L > 0$, $f \in \mathcal{H}_d(\alpha, L, D)$, and $\sup_{\mathbf{x} \in D} |\partial^{\mathbf{i}} f(\mathbf{x})| \leq L_1$ for all $\mathbf{i} \in \mathbf{I}(r_\alpha, d)$.
- ($\tilde{\text{A2}}$) There is a unique point $\boldsymbol{\mu} \in \mathring{D}$ that maximizes the function f on D : $\max_{\mathbf{x} \in D} f(\mathbf{x}) = \sup_{\mathbf{x} \in \mathring{D}} f(\mathbf{x}) = f(\boldsymbol{\mu})$. Moreover, there exist $\delta, \kappa_1 > 0$ such that $\kappa_1 \leq \kappa$ and $f(\boldsymbol{\mu}) \geq f(\mathbf{x}) + \delta$ for all $\mathbf{x} \notin B(\boldsymbol{\mu}, \kappa_1)$, with κ from condition ($\tilde{\text{A3}}$).
- ($\tilde{\text{A3}}$) There exist $\kappa, \lambda_0 > 0$ such that $\sup_{\mathbf{x} \in B(\boldsymbol{\mu}, \kappa)} \lambda_{\max}(Hf(\mathbf{x})) \leq -\lambda_0 < 0$.

As we can see, condition ($\tilde{\text{A3}}$) coincides with (A3), condition ($\tilde{\text{A1}}$) is a strengthened version of (A1), namely (A1) is complemented by the requirement of uniform boundedness of all the relevant partial derivatives. Condition ($\tilde{\text{A2}}$) is also a stronger version of (A2) and mainly intended to provide the uniformity of the rate for the first stage estimator $\tilde{\boldsymbol{\mu}}$. The existence of the maximum is complemented by the uniform separation of the maximum from the set outside of the vicinity $B(\boldsymbol{\mu}, \kappa_1)$. Inside this vicinity, the separateness of the maximum can be characterized by the Taylor expansion up to the second order term(s) by virtue of condition (A3), much in the same way as it is done in relation (41) in the proof of Lemma 7.

REMARK 8. One can also relax the condition (A1) by imposing the Hölder condition only for a small ϵ -neighborhood $B(\boldsymbol{\mu}, \epsilon)$ instead of D , that is, assume $f \in \mathcal{H}_d(\alpha, L, B(\boldsymbol{\mu}, \epsilon))$.

REMARK 9. Since the condition of the theorem $\|\tilde{\boldsymbol{\mu}} - \boldsymbol{\mu}\| = o_p(\delta_n)$ needs to be satisfied, it is only natural to use such a preliminary estimator $\tilde{\boldsymbol{\mu}}$ from the first stage that has the fastest convergence rate. The first stage corresponds to the fixed design situation and one folklore fact is that the optimal convergence rate in estimating the mode of an α -smooth function is the same as in estimating its first derivative, with the optimal (minimax) convergence rate $n^{-(\alpha-1)/(2\alpha+d)}$. Thus the optimal estimator $\tilde{\boldsymbol{\mu}}$ satisfies

$$(12) \quad \|\tilde{\boldsymbol{\mu}} - \boldsymbol{\mu}\| = O_p(n^{-(\alpha-1)/(2\alpha+d)}).$$

The optimal choice for the localization parameter δ_n in estimating $\boldsymbol{\mu}$ is formally determined by balancing the terms $\delta_n^{\alpha-1}$ and $n^{-1/2}\delta_n^{-1}$ and is clearly $\delta_n = cn^{-1/(2\alpha)}$. This would lead to the convergence rates $n^{-(\alpha-1)/(2\alpha)}$ and $n^{-1/2}$ in estimating the location $\boldsymbol{\mu}$ and the maximum $M = f(\boldsymbol{\mu})$ respectively. However, the condition of the theorem $\|\tilde{\boldsymbol{\mu}} - \boldsymbol{\mu}\| = o_p(\delta_n)$ must be satisfied. It turns out that the condition $\|\tilde{\boldsymbol{\mu}} - \boldsymbol{\mu}\| = o_p(\delta_n)$ holds with optimal $\tilde{\boldsymbol{\mu}}$ satisfying (12) and the optimal choice of the localization parameter $\delta_n = cn^{-1/(2\alpha)}$ if $\alpha > 1 + \sqrt{1 + d/2}$. This is summarized in the following corollary.

COROLLARY 1. *Let the relation (12) hold, the localization parameter $\delta_n = cn^{-1/(2\alpha)}$ and let the conditions (A1)–(A3) be fulfilled with $\alpha > 1 + \sqrt{1 + d/2}$. Then*

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

Indeed, in order for the condition $\|\tilde{\boldsymbol{\mu}} - \boldsymbol{\mu}\| = o_p(\delta_n)$ to hold,

$$O_p(n^{-(\alpha-1)/(2\alpha+d)}) = \|\tilde{\boldsymbol{\mu}} - \boldsymbol{\mu}\| = o_p(\delta_n) = o_p(n^{-1/(2\alpha)}),$$

which is valid if $\frac{\alpha-1}{2\alpha+d} > \frac{1}{2\alpha}$, or equivalently $\alpha > 1 + \sqrt{1 + d/2}$. Other two conditions of Theorem 1, $\alpha > 2$ and $n^{-1/2} = o(\delta_n^2)$, are also fulfilled. Indeed, $\alpha > 1 + \sqrt{1 + d/2} > 2$ and hence the condition $n^{-1/2} = o(\delta_n^2) = o(n^{-1/\alpha})$ holds true as well.

Notice that if $\alpha > 3$, the corollary yields the optimal rates in estimating $(\boldsymbol{\mu}, M)$ for all the dimensions for which $3 \geq 1 + \sqrt{1 + d/2}$, i.e. up to the dimension $d = 6$, including the most important dimensions $d = 1, 2, 3$.

The above corollary reveals interesting features of the two-stage approach, which we describe in the following remarks.

REMARK 10. The above second-stage rates $n^{-(\alpha-1)/(2\alpha)}$ and $n^{-1/2}$ for estimating $\boldsymbol{\mu}$ and $M = f(\boldsymbol{\mu})$ improve upon the corresponding optimal rates for the fixed design case (first stage estimators), $n^{-(\alpha-1)/(2\alpha+d)}$ and $n^{-\alpha/(2\alpha+d)}$

respectively. Notice that the dimension effect d disappears in the second stage convergence rate unlike the rate of the optimal first-stage estimator. It seems as if the two-stage approach fixed the curse of dimensionality. Actually, it still persists although in a much milder form: the regression function must be sufficiently smooth to ensure $\alpha > 1 + \sqrt{1 + d/2}$. The lower bound in this inequality increases with the dimension d .

REMARK 11. Note that the rates $n^{-(\alpha-1)/2\alpha}$ and $n^{-1/2}$ are the lower bounds for the minimax rate in estimating respectively the location and the size of the maximum of α -smooth regression functions in case of sequential design; see Chen (1988), Polyak and Tsybakov (1990) and Dippon (2003). Then the rates $n^{-(\alpha-1)/2\alpha}$ and $n^{-1/2}$ also lower bound the two-stage rates in estimating respectively the location and the size of the maximum of α -smooth regression functions. Thus, the obtained second-stage rates $n^{-(\alpha-1)/(2\alpha)}$ and $n^{-1/2}$ of simultaneous estimation procedures for $\boldsymbol{\mu}$ and $M = f(\boldsymbol{\mu})$ are optimal and cannot be improved. This illustrates another appealing feature of the two-stage approach in that the rates are as good as those of the completely flexible sequential approach under the smoothness assumption $\alpha > 1 + \sqrt{1 + d/2}$. In the next section, we show that we can get rid of this restriction completely by considering multi-stage procedures.

4. Multi-stage procedures and resolving the curse of dimensionality. Let a numerical sequence $\delta_n^{(1)}$ be such that $\|\tilde{\boldsymbol{\mu}} - \boldsymbol{\mu}\| = O_p(\delta_n^{(1)})$. Recall that δ_n is the localization parameter and denote further $\delta_n^{(2)} = \delta_n^{(2)}(\delta_n) = n^{-1/2}\delta_n^{-1} + \delta_n^{\alpha-1}$, $\Delta_n^{(1)} = 1$ and $\Delta_n^{(2)} = \Delta_n^{(2)}(\delta_n) = n^{-1/2} + \delta_n^\alpha$.

Let us look at our two-stage procedure from a somewhat more general point of view. We can think of $\tilde{\boldsymbol{\mu}}$ and $\delta_n^{(1)}$ as quantification of some prior knowledge about $\boldsymbol{\mu}$.

Since $\delta_n = o(1)$ and $\alpha > 0$, there is always an improvement from $\Delta_n^{(1)}$ to $\Delta_n^{(2)}$. As we have already established in the previous section, the best rate for $\delta_n^{(2)}$ we can get is $n^{-(\alpha-1)/(2\alpha)}$. Indeed, an improvement from $\delta_n^{(1)}$ to $\delta_n^{(2)}$ occurs if $\delta_n^{(2)} \ll \delta_n^{(1)}$, which yields the range of δ_n that lead to an improvement:

$$n^{-1/2}(\delta_n^{(1)})^{-1} \ll \delta_n \ll (\delta_n^{(1)})^{1/(\alpha-1)}.$$

From this relations it follows that it is possible to improve (i.e., there is a non-void choice for δ_n) if $n^{-1/2}(\delta_n^{(1)})^{-1} \ll (\delta_n^{(1)})^{1/(\alpha-1)}$, or equivalently

$$(13) \quad \delta_n^{(1)} \gg n^{-(\alpha-1)/(2\alpha)}.$$

The conclusion is that there is no improvement if $\delta_n^{(1)} = O(n^{-(\alpha-1)/(2\alpha)})$. In other words, it does not make sense to apply the second stage of our procedure if our prior knowledge is already good enough. Assume therefore that (13) is fulfilled.

The best choice $\delta_n \asymp n^{-1/(2\alpha)}$ leads to the optimal $\delta_n^{(2)} = O(n^{-(\alpha-1)/(2\alpha)})$, and other choices of δ_n give the following:

$$(14) \quad \delta_n^{(2)} \asymp \delta_n^{\alpha-1} \quad \text{if} \quad n^{-1/(2\alpha)} = O(\delta_n),$$

$$(15) \quad \delta_n^{(2)} \asymp n^{-1/2} \delta_n^{-1} \quad \text{if} \quad \delta_n = O(n^{-1/(2\alpha)}).$$

Clearly, if possible, the best strategy is to use the optimal $\delta_n \asymp n^{-1/(2\alpha)}$ in order to obtain the optimal (unimprovable) $\delta_n^{(2)} \asymp n^{-(\alpha-1)/(2\alpha)}$. Now we want to elucidate when it is possible to use this optimal $\delta_n \asymp n^{-1/(2\alpha)}$ and when not, and which δ_n should be chosen in the latter case.

In order for Theorem 1 to hold, the conditions to be validated are $n^{-1/4} \ll \delta_n$ and $\delta_n^{(1)} \ll \delta_n$, which gives two lower bounds for the choice of the localization parameter δ_n . Since $\alpha > 2$, we have that $n^{-(\alpha-1)/(2\alpha)} \ll n^{-1/4} \ll n^{-1/(2\alpha)}$, so that the optimal $\delta_n \asymp n^{-1/(2\alpha)}$ automatically satisfies the first lower bound. It is how the second lower bound $\delta_n^{(1)} \ll \delta_n$ is related to $n^{-1/(2\alpha)}$ that determines whether we can choose the localization parameter optimally or not. We consider two cases.

Case I: $\delta_n^{(1)} \ll n^{-1/(2\alpha)}$.

In this case, we simply take $\delta_n = n^{-1/(2\alpha)}$ and remark that the conditions $\alpha > 2$, $\|\tilde{\boldsymbol{\mu}} - \boldsymbol{\mu}\| = O_p(\delta_n^{(1)}) = o_p(\delta_n)$ and $n^{-1/2} = o(\delta_n^2)$ are fulfilled. Finally, apply Theorem 1 to derive the optimal rates:

$$\begin{aligned} \|\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}\| &= O_p(\delta_n^{(2)}) = O_p(n^{-(\alpha-1)/(2\alpha)}), \\ \|\hat{M} - M\| &= O_p(\Delta_n^{(2)}) = O_p(n^{-1/2}). \end{aligned}$$

As we already discussed in the previous section, the optimal nonparametric estimator at the first stage corresponds to the choice $\delta_n^{(1)} = n^{-(\alpha-1)/(2\alpha+d)}$ and the condition $\|\tilde{\boldsymbol{\mu}} - \boldsymbol{\mu}\| = O_p(\delta_n^{(1)}) = o_p(\delta_n)$ boils down to $\alpha > 1 + \sqrt{1 + d/2}$ which is described by Corollary 1.

Case II: $n^{-1/(2\alpha)} = O(\delta_n^{(1)})$.

In this case we cannot use the optimal choice $\delta_n = n^{-1/(2\alpha)}$ because the condition of Theorem 1 $\|\tilde{\boldsymbol{\mu}} - \boldsymbol{\mu}\| = o_p(\delta_n)$ may not be fulfilled. Suppose that we have a preliminary estimator $\tilde{\boldsymbol{\mu}}$ such that $\|\tilde{\boldsymbol{\mu}} - \boldsymbol{\mu}\| = O_p(\delta_n^{(1)})$, with $\delta_n^{(1)} = n^{-\beta}$ for some $\beta > 0$. Since $\beta > 1/(2\alpha)$ is covered by the previous case,

assume $0 < \beta \leq 1/(2\alpha)$. Clearly, the closer our choice of δ_n to the optimal $n^{-1/(2\alpha)}$, the bigger improvement in the rate is achieved. We take $\delta_n = M_n \delta_n^{(1)} = M_n n^{-\beta}$, with $M_n \rightarrow \infty$ slowly as compared to a power law (say, $M_n = \log \log n$). Then the conditions $\alpha > 2$, $\|\tilde{\boldsymbol{\mu}} - \boldsymbol{\mu}\| = O_p(\delta_n^{(1)}) = o_p(\delta_n)$ and $n^{-1/2} = o(\delta_n^2)$ (as $2\beta \leq 1/\alpha < 1/2$) are fulfilled. We apply Theorem 1 and use (14) to derive

$$\begin{aligned}\delta_n^{(2)} &= n^{-1/2} \delta_n^{-1} + \delta_n^{\alpha-1} \asymp \delta_n^{\alpha-1} = M_n^{\alpha-1} n^{-\beta(\alpha-1)}, \\ \Delta_n^{(2)} &= n^{-1/2} + \delta_n^\alpha \asymp \delta_n^\alpha = M_n^\alpha n^{-\beta\alpha}.\end{aligned}$$

Thus we have improved from $\delta_n^{(1)}$ to $\delta_n^{(2)}$ and from $\Delta_n^{(1)}$ to $\Delta_n^{(2)}$, but did not attain the optimal rates yet.

We can use the following strategy in this case. We split the observation budget in m portions and apply an m -stage procedure, using one portion at each stage. The third stage is the same as the second, with $\delta_n^{(2)}$ instead of $\delta_n^{(1)}$ and $\Delta_n^{(2)}$ instead of $\Delta_n^{(1)}$, the resulting rates of the third stage are called $\delta_n^{(3)}$ and $\Delta_n^{(3)}$, and so on. We apply a new stage as long as we are in Case II so that at the l -th stage we improve from $\delta_n^{(l-1)}$ to

$$\delta_n^{(l)} \asymp M_n^{g_{l-1}(\alpha)} n^{-\beta(\alpha-1)^{l-1}},$$

and from $\Delta_n^{(l-1)}$ to

$$\Delta_n^{(l)} \asymp M_n^{h_{l-1}(\alpha)} n^{-\beta\alpha^{l-1}},$$

with some polynomials g_{l-1} and h_{l-1} of order $l-1$. As soon as we fall into the situation of Case I, i.e. the smallest $k \in \mathbb{N}$ such that $n^{-\beta(\alpha-1)^{k-1}} \ll n^{-1/(2\alpha)}$, we apply the very last $(k+1)$ -th stage with the optimal $\delta_n \asymp n^{-1/(2\alpha)}$ to obtain the optimal rates $\delta_n^{(k+1)} \asymp n^{-(\alpha-1)/(2\alpha)}$ and $\Delta_n^{(k+1)} \asymp n^{-1/2}$. The necessary number of stages is given by $k_1 + 1$, where k_1 is the smallest natural number such that

$$(16) \quad (\alpha - 1)^{k_1 - 1} \beta \geq \frac{1}{2\alpha} \quad \text{or} \quad k_1 \geq \frac{\log(1/(2\alpha\beta))}{\log(\alpha - 1)} + 1.$$

REMARK 12. If we are interested only in estimating the maximum M , we can stop earlier when applying multi-stage procedure in Case II. Namely, at stage $k_2 + 1$ if $k_2 < k_1$, where k_2 is the smallest integer such that $\Delta_n^{(k_2)} = O(n^{-1/2})$, or

$$(17) \quad \alpha^{k_2 - 1} \beta > \frac{1}{2}, \quad \text{or} \quad k_2 > \frac{\log(1/(2\beta))}{\log \alpha} + 1.$$

REMARK 13. The value of the smoothness parameter α needs to be strictly greater than 2 to control the error in the second order Taylor approximation of the underlying multivariate regression function. The closer α gets to 2, the bigger the number of stages needed in the multi-stage procedures gets.

REMARK 14. When we start in a situation of Case II at the first stage, we need to apply k_1 (with k_1 given by (16)) more stages according to the above described multi-stage approach in order to get a situation of Case I for the first time and then we can use the optimal $\delta_n \asymp n^{-1/(2\alpha)}$. In fact, this would be the best strategy, because if we apply the next stage in the same way as we did before (i.e. taking $\delta_n = \delta_n^{(k_1+1)} M_n \ll n^{-1/(2\alpha)}$), then, according to (15), the resulting rate of this stage will be determined by the first term $n^{-1/2}\delta_n$ and will not be optimal. It may be better than the previous one, but it can also be worse. Thus, if we do not take the optimal choice of δ_n at stages when we are allowed to do so (Case I), we will keep on jumping around the optimal rate.

REMARK 15. Consider now the adaptive version of our original estimation problem: suppose now that we do not know the “true” smoothness parameter α , we only know that $\alpha \geq 2 + \epsilon$ for some $\epsilon > 0$. In this case, our regression function is certainly α_1 -smooth, with $\alpha_1 = 2 + \epsilon$, and we can apply our multi-stage approach with $\alpha = \alpha_1$. The number of stages k_1 is determined by (16) with $\alpha_1 = 2 + \epsilon$ and is the smallest natural number such that $1 + \lceil \log(1/(2(2 + \epsilon)\beta)) \rceil / \log(1 + \epsilon)$. At the last stage we take the “optimal” $\delta_n \asymp n^{-1/(2\alpha_1)}$ to derive the resulting rate $n^{-(\alpha_1-1)/(2\alpha_1)}$ in estimating the location of the maximum of the regression function. This is an adaptive procedure and will work for all $\alpha > 2 + \epsilon$. For certain (not too high) values of α , this procedure gives an improvement as related to the optimal nonparametric adaptive estimators of the location at the first stage. We can speak of some kind of partial adaptiveness.

Remarkably, the multi-stage approach does provides a completely adaptive estimator for the size of the maximum of the regression function. Indeed, we simply apply $k_2 + 1$ stages, with k_2 defined by (17), where $\alpha = \alpha_1 = 2 + \epsilon$. Then the resulting rate is $\Delta_n^{(k_2)} = O(n^{-1/2})$, i.e. we attained the best possible rate for the sequential setting without using the knowledge of the smoothness parameter α . This is due to the key property that the first term $O(n^{-1/2})$ in the risk of the estimator for the size of the maximum does not depend on the localization parameter δ_n .

5. Proofs. First we introduce several quantities we are going to use in the sequel. Now we define $\mathbf{z}_k = \mathbf{x}_k - \tilde{\boldsymbol{\mu}}$, $k = 1, \dots, n_2$ and reformulate the definition (9) 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 $k = 1, \dots, n_2$

$$(18) \quad \mathbf{z}_k \in \{0, \pm\delta_n, \dots, \pm l\delta_n\}^d = \{\mathbf{d}_1, \dots, \mathbf{d}_{(2l+1)^d}\} \subset C(l\delta_n),$$

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

$$f_{\hat{\boldsymbol{\theta}}}(\mathbf{x} - \tilde{\boldsymbol{\mu}}) = f_{\hat{\boldsymbol{\theta}}}(\mathbf{x}).$$

Equivalently,

$$(19) \quad \hat{\boldsymbol{\theta}} = (\mathbf{Z}^T \mathbf{Z})^{-1} \mathbf{Z}^T \mathbf{Y},$$

where

$$(20) \quad \mathbf{Z} = (\mathbf{Z}_1^T, \dots, \mathbf{Z}_{n_2}^T)^T, \quad \mathbf{Z}_k = (\mathbf{z}_k^{\mathbf{i}_0}, \mathbf{z}_k^{\mathbf{i}_1}, \dots, \mathbf{z}_k^{\mathbf{i}_q})$$

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 the same arguments as for $\mathbf{X}^T \mathbf{X}$, namely by Lemma 1. We thus obtain an equivalent description of the estimator $(\hat{\boldsymbol{\mu}}, \hat{M})$ given by (9) in terms of the polynomial $f_{\hat{\boldsymbol{\theta}}}(\mathbf{z})$ defined by (7) and (19):

$$(21) \quad \hat{\boldsymbol{\mu}} = \tilde{\boldsymbol{\mu}} + \hat{\boldsymbol{\mu}}, \quad \hat{M} = f_{\hat{\boldsymbol{\theta}}}(\hat{\boldsymbol{\mu}}), \quad \text{where } \hat{\boldsymbol{\mu}} = \arg \max_{\mathbf{z} \in C(l\delta_n)} f_{\hat{\boldsymbol{\theta}}}(\mathbf{z}),$$

$C(l\delta_n) = [-l\delta_n, l\delta_n]^d \subset \mathbb{R}^d$. 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 around $\tilde{\boldsymbol{\mu}}$.

Next, let the vector $\boldsymbol{\theta} = (\theta_{\mathbf{i}_0}, \theta_{\mathbf{i}_1}, \dots, \theta_{\mathbf{i}_q})^T$ be defined by the equality of the two polynomials $f_{\boldsymbol{\theta}}(\mathbf{x} - \tilde{\boldsymbol{\mu}}) = P_{f, \boldsymbol{\mu}}(\mathbf{x})$:

$$(22) \quad \sum_{\mathbf{i} \in \mathbf{I}} \theta_{\mathbf{i}} (\mathbf{x} - \tilde{\boldsymbol{\mu}})^{\mathbf{i}} = \sum_{\mathbf{i} \in \mathbf{I}} \frac{\partial^{\mathbf{i}} f(\boldsymbol{\mu})}{\mathbf{i}!} (\mathbf{x} - \boldsymbol{\mu})^{\mathbf{i}} = f(\boldsymbol{\mu}) + \sum_{\mathbf{i} \in \mathbf{I}, |\mathbf{i}| \geq 2} \frac{\partial^{\mathbf{i}} f(\boldsymbol{\mu})}{\mathbf{i}!} (\mathbf{x} - \boldsymbol{\mu})^{\mathbf{i}},$$

where we used the condition $\nabla f(\boldsymbol{\mu}) = \mathbf{0}$, due to (A1)–(A2). Thus, $\boldsymbol{\theta}$ is a random vector depending on f , $\boldsymbol{\mu}$ and $\tilde{\boldsymbol{\mu}}$. From (22) it follows that

$$(23) \quad \mathbf{i}! \theta_{\mathbf{i}} = \partial^{\mathbf{i}} P_{f, \boldsymbol{\mu}}(\tilde{\boldsymbol{\mu}}), \quad \partial^{\mathbf{i}} f(\boldsymbol{\mu}) = \partial^{\mathbf{i}} f_{\boldsymbol{\theta}}(\boldsymbol{\mu} - \tilde{\boldsymbol{\mu}}), \quad \mathbf{i} \in \mathbf{I} = \{\mathbf{i}_0, \dots, \mathbf{i}_q\}.$$

The next lemma ensures that the estimator (19) is well defined, i.e. $\mathbf{Z}^T \mathbf{Z}$ is invertible.

LEMMA 1. *The columns of matrix \mathbf{Z} (and \mathbf{X}) defined by (20) are linearly independent.*

PROOF. Consider the matrix \mathbf{Z} , the same proof holds for matrix \mathbf{X} .

Introduce the concatenation operation for multi-indices: for $\mathbf{i} = (i_1, \dots, i_p) \in \mathbb{N}^p$ and $\mathbf{j} = (j_1, \dots, j_s) \in \mathbb{N}^s$, let $\mathbf{k} = \mathbf{ij} = (i_1, \dots, i_p, j_1, \dots, j_s) \in \mathbb{N}^{p+s}$. 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 notations: for $l = 0, 1, \dots, d-1$ and $\mathbf{z} = (z_1, \dots, z_d) \in \mathbb{R}^d$,

$$\begin{aligned} \mathbf{I}_{-l}(i_1, \dots, i_l) &= \{\mathbf{i} \in \mathbb{N}^{d-l} : i_1 \dots i_l \mathbf{i} \in \mathbf{I}\}, \\ \mathbf{i}_{-l} &= (i_{l+1}, \dots, i_d) \in \mathbb{N}^{d-l}, \\ \mathbf{z}_{-l} &= (z_{l+1}, \dots, z_d) \in \mathbb{R}^{d-l}. \end{aligned}$$

Let \mathbf{C}_i , $\mathbf{i} \in \mathbf{I}$, be the columns of the matrix \mathbf{Z} . We are to show that $\sum_{\mathbf{i} \in \mathbf{I}} \lambda_i \mathbf{C}_i = 0$ implies that $\lambda_i = 0$ for all $\mathbf{i} \in \mathbf{I}$. The equality $\sum_{\mathbf{i} \in \mathbf{I}} \lambda_i \mathbf{C}_i = 0$ is equivalent to

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

Among $\{\mathbf{z}_1, \dots, \mathbf{z}_{n_2}\}$, there are $(2l+1)^d$ different design points $\{\mathbf{d}_1, \dots, \mathbf{d}_{(2l+1)^d}\}$ given by (18). Thus, for all $\mathbf{z} \in \{\mathbf{d}_1, \dots, \mathbf{d}_{(2l+1)^d}\}$,

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

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 in variable z_1 . But we have $2l+1 > r$ 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. Therefore, the coefficients of this polynomial must be zero because otherwise this degree r polynomial will have more than r zeros. Thus we have that

$$0 = \sum_{\mathbf{i}_{-1} \in \mathbf{I}_{-1}(i_1)} \lambda_{i_1 \mathbf{i}_{-1}} \mathbf{z}_{-1}^{\mathbf{i}_{-1}}, \quad i_1 = 0, 1, \dots, r,$$

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 $i_1, \dots, i_{d-1} = 0, 1, \dots, r$, $z_d \in \{0, \pm\delta_n, \pm 2\delta_n, \dots, \pm l\delta_n\}$,

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

from which we derive that $\lambda_i = 0$ for all $\mathbf{i} \in \mathbf{I}$. □

REMARK 16. In case $d = 1$, \mathbf{X} and \mathbf{Z} are Vandermonde matrices.

The next lemma claims basically that the second stage data D_2 given by (8) can be regarded as coming approximately from a certain parametric polynomial model.

LEMMA 2. Assume (A1) and let the data $\{(\mathbf{x}_i, Y_i), i = 1, \dots, n_2\}$ and $\boldsymbol{\xi} = (\xi_1, \dots, \xi_{n_2})^T$ be given by the second stage observation scheme (8), $\mathbf{Y} = (Y_1, \dots, Y_{n_2})^T$, $\boldsymbol{\eta} = (\eta_1, \dots, \eta_{n_2})^T$ 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 (20) and (22) respectively, and $\boldsymbol{\eta}$ is independent of $\boldsymbol{\xi}$. Moreover, for some constants C_1, C_2 and uniformly in $k \in \{1, 2, \dots, n_2\}$,

$$(24) \quad |\eta_k| \leq C_1 \delta_n^\alpha + C_2 \|\tilde{\boldsymbol{\mu}} - \boldsymbol{\mu}\|^\alpha = O(\delta_n^\alpha) + O_p(\|\tilde{\boldsymbol{\mu}} - \boldsymbol{\mu}\|^\alpha).$$

PROOF. Since $\eta_k = f(\mathbf{x}_k) - P_{f, \boldsymbol{\mu}}(\mathbf{x}_k)$, then, by (8), (20) and (22),

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

Clearly, $\boldsymbol{\eta}$ is independent of $\boldsymbol{\xi}$ by definition. It remains to show (24). Recall the c_r -inequality: $|a + b|^r \leq \max(1, 2^{r-1})(|a|^r + |b|^r)$ for any $r > 0$. Apply this inequality, (4) and the fact that $\|\mathbf{x}_k - \tilde{\boldsymbol{\mu}}\| \leq d^{1/2} l \delta_n$, $k = 1, \dots, n_2$, to obtain (24):

$$\begin{aligned} |\eta_k| &= |f(\mathbf{x}_k) - P_{f, \boldsymbol{\mu}}(\mathbf{x}_k)| \leq L \|\mathbf{x}_k - \boldsymbol{\mu}\|^\alpha \\ &\leq L c_\alpha (\|\mathbf{x}_k - \tilde{\boldsymbol{\mu}}\|^\alpha + \|\tilde{\boldsymbol{\mu}} - \boldsymbol{\mu}\|^\alpha) = O(\delta_n^\alpha) + O_p(\|\tilde{\boldsymbol{\mu}} - \boldsymbol{\mu}\|^\alpha). \end{aligned}$$

□

Lemma 1 ensures that the matrix $\mathbf{Z}^T \mathbf{Z}$ is non-singular. 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: for an $(s + 1) \times (p + 1)$ matrix \mathbf{A} , $\mathbf{A} = ((a_{ij})) = ((a_{ij}))_{i=0, j=0}^{s, p}$.

LEMMA 3. Let \mathbf{Z} be defined by (20) and $(\mathbf{Z}^T \mathbf{Z})^{-1} = ((d_{ij}))_{i=0, j=0}^{q, q}$. Then

$$d_{ij} = O\left(n^{-1} \delta_n^{-(|i| + |j|)}\right), \quad i, j = 0, 1, \dots, q.$$

PROOF. Since $\mathbf{z}_k^{\mathbf{i}_0} = \mathbf{z}_k^{\mathbf{0}} = 1$ for all $k = 1, \dots, n_2$, we have

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

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

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

Some entries are easy to compute. For example, $a_{00} = (2l+1)^d$ since $n_2 = (2l+1)^d n_3$ and 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 with the exact values a_{ij} ; the only important fact is that the matrix $\mathbf{A} = ((a_{ij}))$ is not degenerate as is shown below.

It is not difficult to express the determinant of $\mathbf{Z}^T \mathbf{Z}$ as follows:

$$\det(\mathbf{Z}^T \mathbf{Z}) = \det(\mathbf{A}) n_3^{q+1} \delta_n^m,$$

where $m = 2 \sum_{k=0}^q |\mathbf{i}_k| = 2 \sum_{\mathbf{i} \in \mathbf{I}} |\mathbf{i}| = 2 \sum_{k=1}^r k |\mathbf{I}_k|$. In particular, for $d = 1$, $m = 2 \sum_{k=1}^r k = r(r+1)$. Indeed, when computing the determinant of the matrix $\mathbf{Z}^T \mathbf{Z}$, we factor out n_3 from each row and factor out $\delta_n^{|\mathbf{i}_i|}$ times from the i -th row and $|\mathbf{i}_j|$ times from the j -th column to reduce the calculations to the determinant of matrix \mathbf{A} . Since the matrix $\mathbf{Z}^T \mathbf{Z}$ is not degenerate by Lemma 1, the last relation implies that $\det(\mathbf{A}) \neq 0$. Let $C_{ij}(\mathbf{Z}^T \mathbf{Z})$ and $C_{ij}(\mathbf{A})$ be the cofactors of the (i, j) -th entries of the matrices $\mathbf{Z}^T \mathbf{Z}$ and \mathbf{A} respectively. Similarly, we derive

$$C_{ij}(\mathbf{Z}^T \mathbf{Z}) = C_{ij}(\mathbf{A}) n_3^q \delta_n^{m - (|\mathbf{i}_i| + |\mathbf{i}_j|)}, \quad i, j = 0, 1, \dots, q.$$

Denote by a^{ij} the (i, j) -th entry of the constant matrix \mathbf{A}^{-1} and recall that $n_3 \geq cn$. The lemma follows from the last two relations: for $i, j = 0, 1, \dots, q$,

$$\begin{aligned} d_{ji} &= \frac{C_{ij}(\mathbf{Z}^T \mathbf{Z})}{\det(\mathbf{Z}^T \mathbf{Z})} = \frac{C_{ij}(\mathbf{A}) n_3^q \delta_n^{m - (|\mathbf{i}_i| + |\mathbf{i}_j|)}}{\det(\mathbf{A}) n_3^{q+1} \delta_n^m} \\ &= a^{ij} n_3^{-1} \delta_n^{-(|\mathbf{i}_i| + |\mathbf{i}_j|)} = O\left(n^{-1} \delta_n^{-(|\mathbf{i}_i| + |\mathbf{i}_j|)}\right). \end{aligned}$$

□

REMARK 17. In case $d = 1$ and even r , we have $2l + 1 = r + 1$. Then the entries of the matrix $\mathbf{A} = ((a_{ij}))_{i=0, j=0}^{r, r}$ can be computed as follows. Since $n_2 = (2l + 1)n_3$, $\sum_{i=1}^{n_2} z_i^k = 0$ for each odd $k \in \{1, \dots, 2r\}$ and

$$\sum_{i=1}^{n_2} z_i^k = 2n_3 \{l^k \delta_n^k + (l-1)^k \delta_n^k + \dots + \delta_n^k\} = n_3 \delta_n^k a_k,$$

for each even $k \in \{1, \dots, 2r\}$, we obtain that $a_{ij} = a_{i+j}$, where $a_0 = r + 1$, $a_k = 0$ for all odd $k \in \{1, \dots, 2r\}$, and for each even $k \in \{1, \dots, 2r\}$

$$a_k = 2(1 + 2^k + 3^k + \dots + l^k) = 2\{1 + 2^k + \dots + (r/2)^k\}.$$

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

LEMMA 4. Assume (A1) and let $\hat{\boldsymbol{\theta}}$ and $\boldsymbol{\theta}$ be defined by respectively (19) and (22). Then

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

PROOF. By invoking (19) and Lemma 2, write

$$(25) \quad \hat{\boldsymbol{\theta}} - \boldsymbol{\theta} = (\mathbf{Z}^T \mathbf{Z})^{-1} \mathbf{Z}^T \mathbf{Y} - \boldsymbol{\theta} = (\mathbf{Z}^T \mathbf{Z})^{-1} \mathbf{Z}^T (\boldsymbol{\eta} + \boldsymbol{\xi}).$$

Since $E(\boldsymbol{\xi}) = \mathbf{0}$ and $\text{Cov}((\mathbf{Z}^T \mathbf{Z})^{-1} \mathbf{Z}^T \boldsymbol{\xi}) = \sigma^2 (\mathbf{Z}^T \mathbf{Z})^{-1}$, 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

$$(26) \quad (\mathbf{Z}^T \mathbf{Z})^{-1} \mathbf{Z}^T \boldsymbol{\xi} = \begin{pmatrix} O_p(n^{-1/2}) \\ O_p(n^{-1/2} \delta_n^{-|\mathbf{i}_1|}) \\ \vdots \\ O_p(n^{-1/2} \delta_n^{-|\mathbf{i}_q|}) \end{pmatrix}.$$

In view of (18), $\mathbf{z}_k \in C(l\delta_n)$, so that $|\mathbf{z}_k^{\mathbf{i}}| \leq c\delta_n^{|\mathbf{i}|}$, $k = 1, \dots, n_2$, $\mathbf{i} \in \mathbf{I}$. Using this, (24) from Lemma 2, the fact that $n_2 \leq c_1 n$ and again Lemma 3,

we obtain that

$$(27) \quad (\mathbf{Z}^T \mathbf{Z})^{-1} \mathbf{Z}^T \boldsymbol{\eta} = \begin{pmatrix} d_{00} \sum_{k=1}^{n_2} \mathbf{z}_k^{\mathbf{i}_0} \eta_k + \dots + d_{0q} \sum_{k=1}^{n_2} \mathbf{z}_k^{\mathbf{i}_q} \eta_k \\ d_{10} \sum_{k=1}^{n_2} \mathbf{z}_k^{\mathbf{i}_0} \eta_k + \dots + d_{1q} \sum_{i=k}^{n_2} \mathbf{z}_k^{\mathbf{i}_q} \eta_k \\ \vdots \\ d_{q0} \sum_{k=1}^{n_2} \mathbf{z}_k^{\mathbf{i}_0} \eta_k + \dots + d_{qq} \sum_{k=1}^{n_2} \mathbf{z}_k^{\mathbf{i}_q} \eta_k \end{pmatrix} = \begin{pmatrix} O(\delta_n^\alpha) + O_p(\|\tilde{\boldsymbol{\mu}} - \boldsymbol{\mu}\|^\alpha) \\ O(\delta_n^{\alpha-|\mathbf{i}_1|}) + O_p(\|\tilde{\boldsymbol{\mu}} - \boldsymbol{\mu}\|^\alpha \delta_n^{-|\mathbf{i}_1|}) \\ \vdots \\ O(\delta_n^{\alpha-|\mathbf{i}_q|}) + O_p(\|\tilde{\boldsymbol{\mu}} - \boldsymbol{\mu}\|^\alpha \delta_n^{-|\mathbf{i}_q|}) \end{pmatrix}.$$

Combining the relations (25), (26) and (27) completes the proof of the lemma. \square

Fix a $k \in \{1, \dots, d\}$. For all $\mathbf{i} = (i_1, \dots, i_d) \in \mathbf{I}$ such that $i_k \geq 1$, define an operator

$$(28) \quad \mathcal{D}_k(\mathbf{i}) = (i_1, \dots, i_{k-1}, i_k - 1, i_{k+1}, \dots, i_d).$$

Its iterates $\mathcal{D}_j \mathcal{D}_k(\mathbf{i})$ are defined for $\mathbf{i} = (i_1, \dots, i_d) \in \mathbf{I}$ such that $i_k, i_j \geq 1$ for $j \neq k$ and $\mathcal{D}_k^2(\mathbf{i})$ for \mathbf{i} such that $i_k \geq 2$. Notice that $|\mathcal{D}_k(\mathbf{i})| = |\mathbf{i}| - 1$.

LEMMA 5. *Assume (A1) and let $f_{\boldsymbol{\theta}}$, $\hat{\boldsymbol{\theta}}$ and $\boldsymbol{\theta}$ be defined by (7), (19) and (22) 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. Recall that $\mathbf{I}_0 = \{\mathbf{i}_0\}$, $\mathbf{I}_1 = \{\mathbf{i} \in \mathbb{N}^d : |\mathbf{i}| = 1\} = \{\mathbf{i}_1, \dots, \mathbf{i}_d\}$, where $\mathbf{i}_k = (0, \dots, 0, 1, 0, \dots, 0)$, where 1 is at the k -th place and other $d-1$ coordinates are zeros. The k -th coordinate of the vector $\nabla f_{\boldsymbol{\theta}}(\mathbf{x})$ is

$$\begin{aligned} \frac{\partial f_{\boldsymbol{\theta}}(\mathbf{x})}{\partial x_k} &= \sum_{\mathbf{i} \in \mathbf{I}} \theta_{\mathbf{i}} \frac{\partial \mathbf{x}^{\mathbf{i}}}{\partial x_k} = \sum_{\mathbf{i} \in \mathbf{I}: i_k \geq 1} \theta_{\mathbf{i}} \frac{\partial \mathbf{x}^{\mathbf{i}}}{\partial x_k} = \sum_{\mathbf{i} \in \mathbf{I}: i_k \geq 1} i_k \theta_{\mathbf{i}} \mathbf{x}^{\mathcal{D}_k(\mathbf{i})} \\ &= \theta_{\mathbf{i}_k} + \sum_{\mathbf{i} \in \mathbf{I}: i_k \geq 1, |\mathbf{i}| \geq 2} i_k \theta_{\mathbf{i}} \mathbf{x}^{\mathcal{D}_k(\mathbf{i})}, \end{aligned}$$

where operator \mathcal{D}_k is defined by (28). Then, for each $k = 1, \dots, d$,

$$(29) \quad \frac{\partial f_{\hat{\boldsymbol{\theta}}}(\boldsymbol{\mu} - \tilde{\boldsymbol{\mu}})}{\partial x_k} - \frac{\partial f_{\boldsymbol{\theta}}(\boldsymbol{\mu} - \tilde{\boldsymbol{\mu}})}{\partial x_k} = \hat{\theta}_{\mathbf{i}_k} - \theta_{\mathbf{i}_k} + \sum_{\mathbf{i} \in \mathbf{I}: i_k \geq 1, |\mathbf{i}| \geq 2} i_k (\hat{\theta}_{\mathbf{i}} - \theta_{\mathbf{i}}) (\boldsymbol{\mu} - \tilde{\boldsymbol{\mu}})^{\mathcal{D}_k(\mathbf{i})}.$$

Now we bound the right hand side of (29). Since $|\mathbf{i}_k| = 1$ for $k = 1, \dots, d$ and $\|\boldsymbol{\mu} - \tilde{\boldsymbol{\mu}}\| = o_p(\delta_n)$, we obtain by Lemma 4 that

$$\begin{aligned} \hat{\theta}_{\mathbf{i}_k} - \theta_{\mathbf{i}_k} &= O_p(n^{-1/2}\delta_n^{-1}) + O(\delta_n^{\alpha-1}) + O_p(\|\tilde{\boldsymbol{\mu}} - \boldsymbol{\mu}\|^\alpha \delta_n^{-1}) \\ (30) \quad &= O_p(n^{-1/2}\delta_n^{-1}) + O(\delta_n^{\alpha-1}), \quad k = 1, \dots, d. \end{aligned}$$

The same argument applies to each term of the sum in the right hand side of (29):

$$\begin{aligned} &|(\hat{\theta}_{\mathbf{i}} - \theta_{\mathbf{i}})(\boldsymbol{\mu} - \tilde{\boldsymbol{\mu}})^{\mathcal{D}_k(\mathbf{i})}| \\ &\leq |\hat{\theta}_{\mathbf{i}} - \theta_{\mathbf{i}}| \|\boldsymbol{\mu} - \tilde{\boldsymbol{\mu}}\|^{|\mathcal{D}_k(\mathbf{i})|} \\ &= \left[O_p(n^{-1/2}\delta_n^{-|\mathbf{i}|}) + O(\delta_n^{\alpha-|\mathbf{i}|}) + O_p(\|\tilde{\boldsymbol{\mu}} - \boldsymbol{\mu}\|^\alpha \delta_n^{-|\mathbf{i}|}) \right] \|\boldsymbol{\mu} - \tilde{\boldsymbol{\mu}}\|^{|\mathbf{i}|-1} \\ &= o_p(n^{-1/2}\delta_n^{-1}) + o_p(\delta_n^{\alpha-1}) + o_p(\|\tilde{\boldsymbol{\mu}} - \boldsymbol{\mu}\|^\alpha \delta_n^{-1}) \\ (31) \quad &= o_p(n^{-1/2}\delta_n^{-1}) + o_p(\delta_n^{\alpha-1}). \end{aligned}$$

There are finitely many terms in the sum from (29) and the constant i_k is at most r . Combining estimates in (30) and (31), we obtain

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

□

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

$$(32) \quad \|\mathbf{A}\|_{\max} \leq \|\mathbf{A}\| \leq \sqrt{sp} \|\mathbf{A}\|_{\max}.$$

LEMMA 6. Assume (A1), (A3), $\|\boldsymbol{\mu} - \tilde{\boldsymbol{\mu}}\| = O_p(\delta_n)$ and $n^{-1/2} = o(\delta_n^2)$. For a $\boldsymbol{\mu}^* \in \mathbb{R}^d$ such that $\|\boldsymbol{\mu}^*\| = o_p(1)$ and for any fixed $\epsilon \in (0, 1)$, introduce the event

$$(33) \quad B_n = \left\{ \|Hf(\boldsymbol{\mu}) - Hf_{\hat{\boldsymbol{\theta}}}(\boldsymbol{\mu}^*)\| \leq (1 - \epsilon) \|(Hf(\boldsymbol{\mu}))^{-1}\|^{-1} \right\},$$

where $f_{\boldsymbol{\theta}}$ and $\hat{\boldsymbol{\theta}}$ are defined by (7) and (19) respectively. Then $P(B_n) \rightarrow 1$ as $n \rightarrow \infty$. Moreover, 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. We first establish some properties of the Hessian matrix $Hf_{\boldsymbol{\vartheta}}(\mathbf{z})$. In view of the definition (7) of the polynomial $f_{\boldsymbol{\vartheta}}$, the entries of the matrix $Hf_{\boldsymbol{\vartheta}}(\mathbf{z})$ are continuous with respect to $\boldsymbol{\vartheta}$ and \mathbf{z} . Namely, we have entry-wise that

$$(34) \quad Hf_{\boldsymbol{\vartheta}}(\mathbf{z}) = Hf_{\boldsymbol{\vartheta}}(\mathbf{0}) + O(\|\mathbf{z}\|) \quad \text{as} \quad \|\mathbf{z}\| \rightarrow 0.$$

Next, note that the elements of the matrix $Hf_{\boldsymbol{\vartheta}}(\mathbf{0})$ are the multiples of $\vartheta_{\mathbf{i}}$, $\mathbf{i} \in \mathbf{I}_2$. In fact, if $\mathbf{i} = \mathbf{i}_i + \mathbf{i}_j$, $i, j \in \{1, \dots, d\}$, then the (i, j) -th entry of the matrix $Hf_{\boldsymbol{\vartheta}}(\mathbf{0})$ is a multiple of $\vartheta_{\mathbf{i}}$, that is, entry-wise

$$(35) \quad Hf_{\boldsymbol{\vartheta}+\boldsymbol{\Delta}}(\mathbf{0}) = Hf_{\boldsymbol{\vartheta}}(\mathbf{0}) + O(\|\boldsymbol{\Delta}_{\mathbf{I}_2}\|), \quad \text{as} \quad \|\boldsymbol{\Delta}_{\mathbf{I}_2}\| \rightarrow 0,$$

where $\boldsymbol{\Delta}_{\mathbf{I}_2} = (\Delta_{\mathbf{i}}, \mathbf{i} \in \mathbf{I}_2)$ is the subvector of $\boldsymbol{\Delta} = (\Delta_{\mathbf{i}}, \mathbf{i} \in \mathbf{I})$ with the coordinates indexed by $\mathbf{I}_2 \subset \mathbf{I}$.

Now, by using Lemma 4 and the conditions $\alpha > 2$, $\|\boldsymbol{\mu} - \tilde{\boldsymbol{\mu}}\| = O_p(\delta_n)$ and $n^{-1/2} = o(\delta_n^2)$, we obtain that

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

where vector $\boldsymbol{\theta}$ is defined by (22). As $Hf_{\hat{\boldsymbol{\theta}}}(\mathbf{0})$ depends only on $\hat{\theta}_{\mathbf{i}}$, $\mathbf{i} \in \mathbf{I}_2$, the last relation and (35) yield that entry-wise

$$(36) \quad Hf_{\hat{\boldsymbol{\theta}}}(\mathbf{0}) = Hf_{\boldsymbol{\theta}}(\mathbf{0}) + o_p(1).$$

By the definition (22) of $\boldsymbol{\theta}$, $Hf(\boldsymbol{\mu}) = Hf_{\boldsymbol{\theta}}(\boldsymbol{\mu} - \tilde{\boldsymbol{\mu}})$. This and (34) imply that entry-wise

$$(37) \quad Hf(\boldsymbol{\mu}) = Hf_{\boldsymbol{\theta}}(\boldsymbol{\mu} - \tilde{\boldsymbol{\mu}}) = Hf_{\boldsymbol{\theta}}(\mathbf{0}) + O_p(\|\boldsymbol{\mu} - \tilde{\boldsymbol{\mu}}\|).$$

Combining (34), (36) and (37) leads to the following entry-wise relation:

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

Then $\|Hf_{\hat{\boldsymbol{\theta}}}(\boldsymbol{\mu}^*) - Hf(\boldsymbol{\mu})\|_{\max} = o_p(1)$ and hence, by (32),

$$(38) \quad \|Hf(\boldsymbol{\mu}) - Hf_{\hat{\boldsymbol{\theta}}}(\boldsymbol{\mu}^*)\| = o_p(1).$$

Next, we have that $\lambda_{\min}(Hf(\boldsymbol{\mu})) \leq \dots \leq \lambda_{\max}(Hf(\boldsymbol{\mu})) \leq -\lambda_0 < 0$, so that

$$\|(Hf(\boldsymbol{\mu}))^{-1}\| = -(\lambda_{\max}(Hf(\boldsymbol{\mu})))^{-1} \leq \lambda_0^{-1},$$

or

$$(39) \quad \lambda_0 \leq \|(Hf(\boldsymbol{\mu}))^{-1}\|^{-1}.$$

Define an event $C_n = \{\|Hf(\boldsymbol{\mu}) - Hf_{\hat{\boldsymbol{\theta}}}(\boldsymbol{\mu}^*)\| \leq (1 - \epsilon)\lambda_0\}$.

Using (39) and Lemma 11, we obtain that

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

In view of (37), $P(C_n) \rightarrow 1$ and hence $P(B_n) \rightarrow 1$. Finally, by applying (38), (39) and Lemma 11 again, we get that on the event B_n

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

□

LEMMA 7. Assume (A1)–(A3), $\|\boldsymbol{\mu} - \tilde{\boldsymbol{\mu}}\| = o_p(\delta_n)$, $n^{-1/2} = o(\delta_n^2)$ and let $A_n = \{\dot{\boldsymbol{\mu}} \in C(2l\delta_n/3)\}$, where the estimator $\dot{\boldsymbol{\mu}}$ is defined by (21). Then $P(A_n) \rightarrow 1$ as $n \rightarrow \infty$.

PROOF. Write

$$(40) \quad \begin{aligned} P(A_n^c) &\leq P(\dot{\boldsymbol{\mu}} \notin C(2l\delta_n/3), \boldsymbol{\mu} - \tilde{\boldsymbol{\mu}} \in C(l\delta_n/3)) \\ &\quad + P(\boldsymbol{\mu} - \tilde{\boldsymbol{\mu}} \notin C(l\delta_n/3)). \end{aligned}$$

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 $\mathbf{x} \in \mathbb{R}^d$, $\lambda_{\min}(\mathbf{M})\|\mathbf{x}\|^2 \leq \mathbf{x}^T \mathbf{M} \mathbf{x} \leq \lambda_{\max}(\mathbf{M})\|\mathbf{x}\|^2$. Therefore, by Conditions (A1)–(A3), for $\boldsymbol{\mu} \in C(\tilde{\boldsymbol{\mu}}, l\delta_n/3)$ and $\mathbf{x} \in C(\tilde{\boldsymbol{\mu}}, l\delta_n) \setminus C(\tilde{\boldsymbol{\mu}}, 2l\delta_n/3)$, we have

$$(41) \quad \begin{aligned} f(\mathbf{x}) &= f(\boldsymbol{\mu}) + \frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^T Hf(\boldsymbol{\mu}^*)(\mathbf{x} - \boldsymbol{\mu}) \\ &\leq f(\boldsymbol{\mu}) - \frac{\lambda_0}{2}\|\mathbf{x} - \boldsymbol{\mu}\|^2 \leq f(\boldsymbol{\mu}) - c\delta_n^2 \end{aligned}$$

for some positive constant $c = c(\lambda_0, l)$ and sufficiently large n such that $\|\boldsymbol{\mu}^* - \boldsymbol{\mu}\| \leq \kappa$, with $\kappa > 0$ from the condition (A2).

Next, by using (4), (22) and the c_r -inequality

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

Now we combine this with Lemma 4 and the conditions $\alpha > 2$, $\|\boldsymbol{\mu} - \tilde{\boldsymbol{\mu}}\| = o_p(\delta_n)$ and $n^{-1/2} = o(\delta_n^2)$ to obtain that, uniformly in $\mathbf{z} \in C(l\delta_n)$,

$$\begin{aligned} f_{\hat{\boldsymbol{\theta}}}(\mathbf{z}) &= f_{\boldsymbol{\theta}}(\mathbf{z}) + O_p(n^{-1/2}) + O(\delta_n^\alpha) \\ &= f(\mathbf{z} + \tilde{\boldsymbol{\mu}}) + O(\|\mathbf{z}\|^\alpha) + O_p(\|\tilde{\boldsymbol{\mu}} - \boldsymbol{\mu}\|^\alpha) + O_p(n^{-1/2}) + O(\delta_n^\alpha) \\ (42) \quad &= f(\mathbf{z} + \tilde{\boldsymbol{\mu}}) + o_p(\delta_n^2). \end{aligned}$$

Recall that $\hat{\boldsymbol{\mu}} \in C(l\delta_n)$ by the definition (21). By (41) and (42), we see that the event

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

implies the event

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

leading to

$$P(\hat{\boldsymbol{\mu}} \notin C(2l\delta_n/3), \boldsymbol{\mu} - \tilde{\boldsymbol{\mu}} \in C(l\delta_n/3)) \leq P(c\delta_n^2 \leq o_p(\delta_n^2)) \rightarrow 0$$

as $n \rightarrow \infty$. Combined with (40), this completes the proof of the lemma. \square

PROOF OF THEOREM 1. In view of conditions (A1)–(A3), $\nabla f(\boldsymbol{\mu}) = \mathbf{0}$ and the Hessian matrix $Hf(\boldsymbol{\mu})$ is negative definite. According to the definition (22) of the polynomial $f_{\boldsymbol{\theta}}$,

$$(43) \quad \mathbf{0} = \nabla f(\boldsymbol{\mu}) = \nabla P_{f, \boldsymbol{\mu}}(\boldsymbol{\mu}) = \nabla f_{\boldsymbol{\theta}}(\boldsymbol{\mu} - \tilde{\boldsymbol{\mu}}).$$

By (21), $\max_{\mathbf{z} \in C(l\delta_n)} f_{\hat{\boldsymbol{\theta}}}(\mathbf{z}) = f_{\hat{\boldsymbol{\theta}}}(\hat{\boldsymbol{\mu}})$. If this maximum is not attained on the boundary of $C(l\delta_n)$, then $\nabla f_{\hat{\boldsymbol{\theta}}}(\hat{\boldsymbol{\mu}})$ must be zero. Hence we have that on the event $A_n = \{\hat{\boldsymbol{\mu}} \in C(2l\delta_n/3)\}$

$$(44) \quad \mathbf{0} = \nabla f_{\hat{\boldsymbol{\theta}}}(\hat{\boldsymbol{\mu}}) = \nabla f_{\hat{\boldsymbol{\theta}}}(\boldsymbol{\mu} - \tilde{\boldsymbol{\mu}}) + Hf_{\hat{\boldsymbol{\theta}}}(\boldsymbol{\mu}^*)(\hat{\boldsymbol{\mu}} - (\boldsymbol{\mu} - \tilde{\boldsymbol{\mu}})),$$

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

By Lemma 6, $(Hf_{\hat{\boldsymbol{\theta}}}(\boldsymbol{\mu}^*))^{-1}$ exists on the event B_n defined by (33). The relations (43) and (44) imply that on the event $A_n \cap B_n$

$$\begin{aligned} \hat{\boldsymbol{\mu}} - \boldsymbol{\mu} &= -(Hf_{\hat{\boldsymbol{\theta}}}(\boldsymbol{\mu}^*))^{-1} \nabla f_{\hat{\boldsymbol{\theta}}}(\boldsymbol{\mu} - \tilde{\boldsymbol{\mu}}) \\ &= -(Hf_{\hat{\boldsymbol{\theta}}}(\boldsymbol{\mu}^*))^{-1} (\nabla f_{\hat{\boldsymbol{\theta}}}(\boldsymbol{\mu} - \tilde{\boldsymbol{\mu}}) - \nabla f_{\boldsymbol{\theta}}(\boldsymbol{\mu} - \tilde{\boldsymbol{\mu}})) \\ (45) \quad &= -(Hf(\boldsymbol{\mu}))^{-1} (\nabla f_{\hat{\boldsymbol{\theta}}}(\boldsymbol{\mu} - \tilde{\boldsymbol{\mu}}) - \nabla f_{\boldsymbol{\theta}}(\boldsymbol{\mu} - \tilde{\boldsymbol{\mu}})) + r_n, \end{aligned}$$

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 (39), we bound the norm of the first term in the right hand side of (45) as

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

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

$$(46) \quad \|\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}\| = O_p(n^{-1/2}\delta_n^{-1} + \delta_n^{\alpha-1}) = O_p(\gamma_n).$$

For any constant $\rho > 0$,

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

The first term on the right hand side can be made arbitrarily small by choosing ρ sufficiently large in view of (46), uniformly in n , while the other two terms converge to zero by Lemmas 6 and 7. This proves (10).

It remains to prove (11). From (22) it follows that

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

so that, according to (21), $\hat{M} - M$ can be written as

$$\begin{aligned} & f_{\hat{\boldsymbol{\theta}}}(\hat{\boldsymbol{\mu}}) - f_{\boldsymbol{\theta}}(\boldsymbol{\mu} - \tilde{\boldsymbol{\mu}}) = \sum_{\mathbf{i} \in \mathbf{I}} [\hat{\theta}_{\mathbf{i}} \hat{\boldsymbol{\mu}}^{\mathbf{i}} - \theta_{\mathbf{i}}(\boldsymbol{\mu} - \tilde{\boldsymbol{\mu}})^{\mathbf{i}}] \\ (47) \quad & = \hat{\theta}_{\mathbf{i}_0} - \theta_{\mathbf{i}_0} + \sum_{\mathbf{i} \in \mathbf{I}, |\mathbf{i}| \geq 1} (\hat{\theta}_{\mathbf{i}} - \theta_{\mathbf{i}}) \hat{\boldsymbol{\mu}}^{\mathbf{i}} + \sum_{\mathbf{i} \in \mathbf{I}, |\mathbf{i}| \geq 1} \theta_{\mathbf{i}} [\hat{\boldsymbol{\mu}}^{\mathbf{i}} - (\boldsymbol{\mu} - \tilde{\boldsymbol{\mu}})^{\mathbf{i}}]. \end{aligned}$$

By Lemma 4, the first term in (47) is

$$(48) \quad \hat{\theta}_{\mathbf{i}_0} - \theta_{\mathbf{i}_0} = O_p(n^{-1/2}) + O_p(\delta_n^{\alpha}).$$

From (21), (10) and the conditions $n^{-1/2} = o(\delta_n^2)$, $\alpha > 2$, $\|\boldsymbol{\mu} - \tilde{\boldsymbol{\mu}}\| = o_p(\delta_n)$, it follows that

$$\begin{aligned} & \|\hat{\boldsymbol{\mu}}\| = \|\hat{\boldsymbol{\mu}} - \tilde{\boldsymbol{\mu}}\| \leq \|\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}\| + \|\boldsymbol{\mu} - \tilde{\boldsymbol{\mu}}\| \\ (49) \quad & = O_p(n^{-1/2}\delta_n^{-1}) + O_p(\delta_n^{\alpha-1}) + O_p(\|\boldsymbol{\mu} - \tilde{\boldsymbol{\mu}}\|) = o_p(\delta_n). \end{aligned}$$

Using (49) and Lemma 4, we evaluate each term in the second sum of (47)

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

so that, as there are finitely many terms in the sum,

$$(50) \quad \sum_{\mathbf{i} \in \mathbf{I}, |\mathbf{i}| \geq 1} (\hat{\theta}_{\mathbf{i}} - \theta_{\mathbf{i}})\dot{\boldsymbol{\mu}}^{\mathbf{i}} = o_p(n^{-1/2}) + o_p(\delta_n^\alpha).$$

Now consider the third sum in (47). Combining Lemma 8 with (10), (49) and the condition $\|\boldsymbol{\mu} - \tilde{\boldsymbol{\mu}}\| = o_p(\delta_n)$, we obtain that for any $\mathbf{i} \in \mathbf{I}$ such that $|\mathbf{i}| \geq 1$

$$\begin{aligned} |\dot{\boldsymbol{\mu}}^{\mathbf{i}} - (\boldsymbol{\mu} - \tilde{\boldsymbol{\mu}})^{\mathbf{i}}| &\leq \|\dot{\boldsymbol{\mu}} - (\boldsymbol{\mu} - \tilde{\boldsymbol{\mu}})\| \sum_{k=1}^{|\mathbf{i}|} \|\dot{\boldsymbol{\mu}}\|^{|\mathbf{i}|-k} \|\boldsymbol{\mu} - \tilde{\boldsymbol{\mu}}\|^{k-1} \\ &= \|\dot{\boldsymbol{\mu}} - \boldsymbol{\mu}\| \sum_{k=1}^{|\mathbf{i}|} \|\dot{\boldsymbol{\mu}}\|^{|\mathbf{i}|-k} \|\boldsymbol{\mu} - \tilde{\boldsymbol{\mu}}\|^{k-1} \\ (51) \quad &= o_p(n^{-1/2} \delta_n^{|\mathbf{i}|-2}) + o_p(\delta_n^{\alpha+|\mathbf{i}|-2}). \end{aligned}$$

Since $\partial^{\mathbf{i}} P_{f, \boldsymbol{\mu}}(\mathbf{x})$, $\mathbf{i} \in \mathbf{I}$, are continuous, they are bounded over the compact set D , so that $\theta_{\mathbf{i}} = O_p(1)$, $\mathbf{i} \in \mathbf{I}$, in view of (23). Because of this and (51),

$$(52) \quad \sum_{\mathbf{i} \in \mathbf{I}, |\mathbf{i}| \geq 2} \theta_{\mathbf{i}} [\dot{\boldsymbol{\mu}}^{\mathbf{i}} - (\boldsymbol{\mu} - \tilde{\boldsymbol{\mu}})^{\mathbf{i}}] = o_p(n^{-1/2}) + o_p(\delta_n^\alpha).$$

It remains to handle separately the terms in the third sum of (47) over $\mathbf{i} \in \mathbf{I}_1$, i.e. such $\mathbf{i} \in \mathbf{I}$ that $|\mathbf{i}| = 1$. Due to (22) and the condition $\|\boldsymbol{\mu} - \tilde{\boldsymbol{\mu}}\| = o_p(\delta_n)$,

$$(53) \quad \theta_{\mathbf{i}} = \partial^{\mathbf{i}} P_{f, \boldsymbol{\mu}}(\tilde{\boldsymbol{\mu}}) = O_p(\|\boldsymbol{\mu} - \tilde{\boldsymbol{\mu}}\|) = o_p(\delta_n), \quad \mathbf{i} \in \mathbf{I}, |\mathbf{i}| = 1.$$

Then (51) and (53) imply that

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

Finally, combining the last display with (47), (48), (50) and (52) completes the proof of (11). \square

REMARK 18. The above argument for the estimating the parameter $M = f(\boldsymbol{\mu})$ can be refined for the problem of estimating any mixed derivative $\partial^{\mathbf{i}} f(\boldsymbol{\mu})$, for $\mathbf{i} \in \mathbf{I}$, $|\mathbf{i}| \geq 2$. One can take the estimator $\partial^{\mathbf{i}} f_{\hat{\boldsymbol{\theta}}}(\hat{\boldsymbol{\mu}})$ and establish in a similar way that

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

6. Appendix.

LEMMA 8. For any $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$ and any $\mathbf{i} \in \mathbf{I}$ 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. The basis is $d = 1$. If $d = 1$, then

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

and the statement follows.

Now we handle the inductive step. Suppose the statement is true for all dimensions $k = 1, \dots, d-1$. Let us prove that it also holds for the dimension d . Without loss of generality assume that $i_1 > 0$. Recall the notations $\mathbf{x}_{-1} = (x_2, \dots, x_d)$, $\mathbf{i}_{-1} = (i_2, \dots, i_d)$ 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 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

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

Combining the last three relations completes the proof of the lemma:

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

□

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

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

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*

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

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 $\|(\mathbf{I} + \mathbf{V}^{-1}\mathbf{W})^{-1}\| \leq (1 - \|\mathbf{V}^{-1}\mathbf{W}\|)^{-1}$ by Banach's lemma. Therefore, $\mathbf{V} + \mathbf{W} = \mathbf{V}(\mathbf{I} + \mathbf{V}^{-1}\mathbf{W})$ is also invertible and

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

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

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

$$\|\mathbf{B}^{-1} - \mathbf{A}^{-1}\| \leq \epsilon^{-1}\|\mathbf{A}^{-1}\|^2\|\mathbf{A} - \mathbf{B}\|.$$

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 $\|\mathbf{A}^{-1}(\mathbf{B} - \mathbf{A})\| \leq 1 - \epsilon$ by the condition of the lemma,

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

By using the last relation, we complete the proof:

$$\begin{aligned}\|\mathbf{B}^{-1} - \mathbf{A}^{-1}\| &\leq \|\mathbf{A}^{-1}\| \|\mathbf{A}\mathbf{B}^{-1} - \mathbf{I}\| \\ &\leq \|\mathbf{A}^{-1}\| \|\mathbf{A} - \mathbf{B}\| \|\mathbf{B}^{-1}\| \leq \epsilon^{-1} \|\mathbf{A}^{-1}\|^2 \|\mathbf{A} - \mathbf{B}\|.\end{aligned}$$

□

REFERENCES

- [1] BLUM, J. R. (1954). Multidimensional stochastic approximation methods. *Ann. Math. Statistics.* **25** 737–744.
- [2] CHEN, H. (1988). Lower rate of convergence for locating the maximum of a function. *Ann. Statist.* **16** 1330–1334.
- [3] DIPPON, J. (2003). Accelerated randomized stochastic optimization. *Ann. Statist.* **31** 1260–1281.
- [4] 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.
- [5] FAN, J. and GIJBELS, I. (1996). *Local Polynomial modeling and its applications – Theory and Methodologies*. New York: Chapman & Hall.
- [6] HASMINSKII, R. Z. (1979). Lower bound for the risks of the nonparametric estimates of the mode. In *Contribution to Statistics: Hajek Memorial Volume* (J. Jureckova, ed.) 91–97. Academia, Prague.
- [7] KIEFER, J. and WOLFOWITZ, J. (1952). Stochastic estimation of the maximum of a regression function. *Ann. Math. Statistics.* **23** 462–466.
- [8] KLEMELÄ, J. (2005). Adaptive estimation of the mode of a multivariate density. *J. Nonparametr. Stat.* **17** 83–105.
- [9] MOKKADEM, A. and PELLETIER, M. (2007). A companion for the Kiefer-Wolfowitz-Blum stochastic approximation algorithm. *Ann. Statist.* **35** 1749–1772.
- [10] MÜLLER, H.-G. (1985). Kernel estimation of zeros and of location and size of extrema of regression functions. *Scand. J. Statist.* **12** 221–232.
- [11] MÜLLER, H.-G. (1989). Adaptive nonparametric peak estimation. *Ann. Statist.* **17** 1053–1069.
- [12] POLYAK, B. T. and TSYBAKOV, A. B. (1990). Optimal order of accuracy for search algorithms of stochastic optimization. *Problems Inform. Transmission* **26** 126–133.
- [13] TANG, R., BANERJE, M. and MICHAILIDIS, G. (2011). A two-stage hybrid procedure for estimating an inverse regression function. *Ann. Statist.* To appear.
- [14] SHOUNG., J.-M. and ZHANG, C.-H. (2001). Least squares estimation of the mode of a unimodal regression function. *Ann. Statist.* **29** 648–665.

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

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