The Annals of Statistics
2013, Vol. 41, No. 3, 1542–1568
DOI: 10.1214/13-AOS1126

NONPARAMETRIC REGRESSION WITH THE SCALE DEPENDING ON AUXILIARY VARIABLE

BY SAM EFROMOVICH[1]

University of Texas at Dallas

The paper is devoted to the problem of estimation of a univariate component in a heteroscedastic nonparametric multiple regression under the mean integrated squared error (MISE) criteria. The aim is to understand how the scale function should be used for estimation of the univariate component. It is known that the scale function does not affect the rate of the MISE convergence, and as a result sharp constants are explored. The paper begins with developing a sharp-minimax theory for a pivotal model $Y = f(X) + \sigma(X, \mathbf{Z})\varepsilon$, where ε is standard normal and independent of the predictor X and the auxiliary vector-covariate $\mathbf{Z}$. It is shown that if the scale $\sigma(x, \mathbf{z})$ depends on the auxiliary variable, then a special estimator, which uses the scale (or its estimate), is asymptotically sharp minimax and adaptive to unknown smoothness of $f(x)$. This is an interesting conclusion because if the scale does not depend on the auxiliary covariate $\mathbf{Z}$, then ignoring the heteroscedasticity can yield a sharp minimax estimation. The pivotal model serves as a natural benchmark for a general additive model $Y = f(X) + g(\mathbf{Z}) + \sigma(X, \mathbf{Z})\varepsilon$, where ε may depend on $(X, \mathbf{Z})$ and have only a finite fourth moment. It is shown that for this model a data-driven estimator can perform as well as for the benchmark. Furthermore, the estimator, suggested for continuous responses, can be also used for the case of discrete responses. Bernoulli and Poisson regressions, that are inherently heteroscedastic, are particular considered examples for which sharp minimax lower bounds are obtained as well. A numerical study shows that the asymptotic theory sheds light on small samples.

1. Introduction.

We begin the Introduction with a simple model which will allow us to explain the setting and the problem, then formulate studied extensions and finish with terminology used in the paper.

1.1. *Pivotal regression model.*

In order to set the stage for a variety of considered problems, it is convenient to begin with a simple nonparametric regression model

$$Y = f(X) + \sigma(X, \mathbf{Z})\varepsilon, \tag{1.1}$$

which will serve as a pivot for all other models. In (1.1) Y is the response, X is the univariate random predictor of interest and $\mathbf{Z} := (Z_1, \dots, Z_D)$ is the vector of

Received September 2012; revised April 2013.
[1]Supported in part by NSF Grant DMS-09-0679 and NSA Grant H982301310212.
MSC2010 subject classifications. Primary 62G07, 62C05; secondary 62E20.
Key words and phrases. Adaptation, lower bound, MISE, sharp minimax.

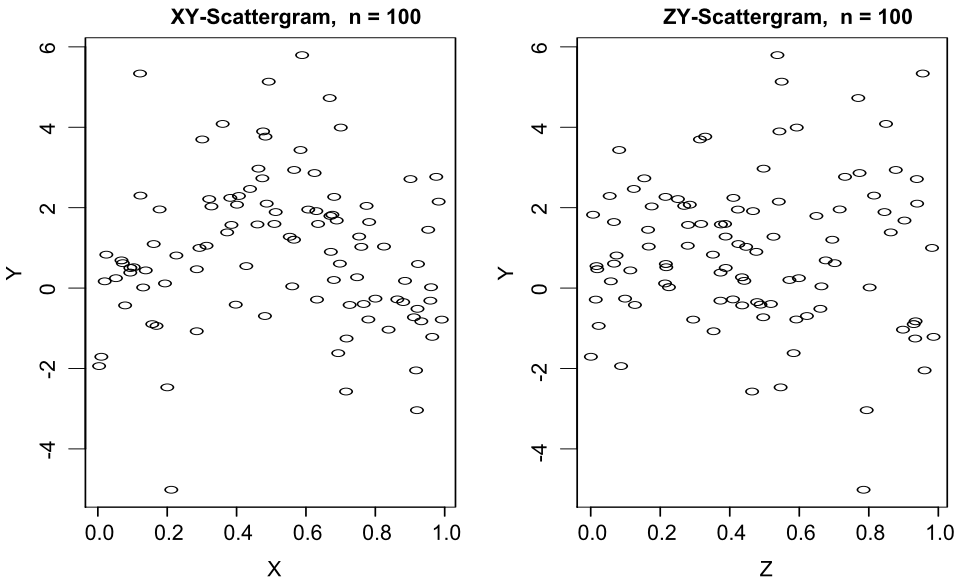

FIG. 1. *Scattergrams for a data simulated according to model* (1.1) *with* $D = 1$.

random auxiliary covariates, $\sigma(x, \mathbf{z})$ is the scale function [$\sigma^2(x, \mathbf{z})$ is also called the variance or volatility] and ε is a standard normal error independent of $(X, \mathbf{Z})$. It is assumed that $(X, \mathbf{Z})$ has a joint density $p(x, \mathbf{z})$ supported on $[0, 1]^{1+D}$, and in what follows $p(x)$ denotes the (marginal) density of X. The problem is to estimate the nonparametric regression function $f(x)$ based on a sample of size n from $(X, \mathbf{Z}, Y)$.

Figure 1 illustrates model (1.1) for a particular case $D = 1$ and $n = 100$ (more details will be revealed shortly). The data is volatile (compare with "typical" data studied in [7, 13, 20, 35]), and it is difficult to visualize an underlying regression. The XY-scattergram suggests a number of possible outliers, but here we do know that these are not outliers, and they are due to heteroscedasticity that can be observed in the ZY-scattergram. Typically, for such a data with two covariates one would definitely attempt to use a multiple or additive regression to explain or reduce the volatility in XY-scattergram and to improve visualization of the underlying regression. However, here we do know that there is no additive component in z. The only hope to help a nonparametric estimator is to use a known (or estimated) scale function. But is this worthwhile to do, and if the answer is "yes," then how one should proceed? Before presenting the answer, let us return to describing the studied setting and known results.

1.2. *Pivotal problem.* To be specific about smoothness of $f(x)$ and because we are going to study minimax constants, let us assume that $f(x)$ belongs to a Sobolev class $\mathcal{S}(\alpha, Q) := \{f(x) : f(x) = \sum_{j=0}^{\infty} \theta_j \varphi_j(x), \varphi_0(x) := 1, \varphi_j(x) :=$

$2^{1/2}\cos(\pi jx)$, $j \ge 1$, $\sum_{j=0}^{\infty}[1+(\pi j)^{2\alpha}]\theta_j^2 \le Q < \infty, x \in [0,1], \alpha \ge 1\}$. Furthermore, the risk of an estimate $\check{f}(x)$ is defined by the mean integrated squared error (MISE) $E\{\int_0^1 (\check{f}(x) - f(x))^2\,dx\}$.

The above-presented discussion of a simulation exhibited in Figure 1 raises the following question. Suppose that, apart of $f(x)$, the statistician knows everything about regression (1.1). Should one use the scale function (and correspondingly the auxiliary variable) in a regression estimator? To warm up the reader, let us consider several arguments against and for using the scale. Against: (a1) A majority of nonparametric research is devoted to rates of the MISE convergence. For the considered setting the rate is $n^{-2\alpha/(2\alpha+1)}$, and then practically all known estimators can attain this rate without using the scale; see [13–15, 22]. (a2) There is a widely held opinion that regression estimation is "...relatively insensitive to heteroscedasticity..." as discussed in [35]. (a3) This is probably the strongest argument against using/estimating the scale. Let us consider a particular case $\sigma(x, \mathbf{z}) = \sigma(x)$ and assume that $p(x)$ and $\sigma(x)$ are positive and have bounded derivatives on $[0,1]$. Then in [12] the following sharp minimax lower bound is established:

$$\inf_{\check{f}^*} \sup_{f \in \mathcal{S}(\alpha,Q)} E\left\{\int_0^1 \bigl(\check{f}^*(x; p, \sigma, \alpha, Q) - f(x)\bigr)^2\,dx\right\} \ge P(\alpha, Q)\bigl[d_1(p,\sigma)n^{-1}\bigr]^{2\alpha/(2\alpha+1)}\bigl(1+o_n(1)\bigr), \tag{1.2}$$

where the infimum is taken over all possible $\check{f}^*$ based on a sample $\{(X_1, Y_1), \dots, (X_n, Y_n)\}$, the design density $p(x)$, the scale function $\sigma(x)$ and parameters (α, Q) that define the underlying Sobolev class. In (1.2)

$$P(\alpha, Q) := \bigl[\alpha/\pi(\alpha+1)\bigr]^{2\alpha/(2\alpha+1)}\bigl[Q(2\alpha+1)\bigr]^{1/(2\alpha+1)} \tag{1.3}$$

is the Pinsker constant [31], and

$$d_1 := d_1(p,\sigma) := \int_0^1 \frac{\sigma^2(x)}{p(x)}\,dx \tag{1.4}$$

is the coefficient of difficulty which is equal to one in the classical case of the unit scale and uniform design, and here and in what follows $o_n(1)$'s denote generic sequences which vanish as $n \to \infty$. Furthermore, in [12] (see also [7]) it is shown that there exists an estimator based solely on data (in what follows referred to as E-estimator) that does not estimate the scale $\sigma(x)$, "ignores" the heteroscedasticity and nonetheless attains the lower bound (1.2). In other words, the "ignore-heteroscedasticity" methodology may yield a sharp-minimax estimation. Furthermore, according to [7, 12] the E-estimator performs well for small samples.

Typical arguments in favor of using/estimating the scale are as follows: (f1) Scale affects the constant of the MISE convergence, and constants may be more important for small samples than rates [7, 28, 29]; (f2) Weighted regression (with weights depending on the scale) is a familiar remedy for heteroscedasticity [13,

15, 16, 19, 22, 32, 35]; (f3) It is reasonable to believe that using the scale may improve an estimator.

Because there are many rate-optimal estimators, to answer the raised pivotal question it is natural to explore a sharp-minimax estimation, that is, estimation with best constant and rate of the MISE convergence. It will be shown shortly that for the model (1.1) the lower bound (1.2) [with the infimum taken over all possible $\check{f}^*$ based on a sample of size n from $(X, \mathbf{Z}, Y)$, all nuisance functions defining the model (1.6) and parameters (α, Q)] still holds with d_1 being replaced by

$$d := d(p,\sigma) := \int_0^1 \frac{dx}{p(x)E\{\sigma^{-2}(X,\mathbf{Z})|X=x\}}. \tag{1.5}$$

The E-estimator, if it is naïvely used for model (1.1), is consistent and even rate minimax, and supremum (over the Sobolev class) of its MISE is equal to the right-hand side of (1.2) with d_1 being replaced by $d_2 := E\{\sigma^2(X,\mathbf{Z})p^{-2}(X)\}$. The latter, according to the Cauchy–Schwarz inequality, is larger than d whenever the scale depends on the auxiliary variable.

We conclude that for the scale depending on an auxiliary variable, the E-estimator, which ignores heteroscedasticity, is no longer sharp minimax. As a result, it is reasonable to explore a regression estimator that uses the scale to attain the sharp minimaxity. The underlying idea of the proposed estimator, based on the developed asymptotic theory, is to use weighted responses $w_l Y_l$ with weights

$$w_l(p,\sigma) := p^{-1}(X_l)\frac{\sigma^{-2}(X_l,\mathbf{Z}_l)}{E\{\sigma^{-2}(X_l,\mathbf{Z}_l)|X_l\}}.$$

Note that: $p^{-1}(X_l)$ is a well-known weight in a univariate sharp-minimax regression [7]; If $\sigma(x,\mathbf{z}) = \sigma(x)$, then the weight does not depend on the scale; Given $X_l = x_l$, conditional expectation $E\{\sigma^{-2}(x_l,\mathbf{Z}_l)|X_l = x_l\}$ is the best estimate (predictor) of $\sigma^{-2}(x_l,\mathbf{Z}_l)$ under the MSE criteria, and the better the estimation is, the closer the weight will be to $p^{-1}(X_l)$; In the light of the foregoing, the proposed weight may be of a special benefit to the case of independent X and $\mathbf{Z}$; The weights should help in dealing with "outliers" created by heteroscedasticity in auxiliary covariates. To shed additional light on the made comments, let us return to Figure 1. The underlying model is defined in Section 4 where it is revealed that the used scale is $\sigma(x,z) = \sigma(z)$ and X and Z are independent. [The interested reader can also look at the identical left diagram in Figure 2 where the solid line shows the underlying regression $f(x)$.] We can now realize that "outliers" in the XY-scattergram are created by the heteroscedasticity in z and the independence of Z from X which creates a chaotic placement of "outliers" in the scattergram.

1.3. *Extensions*. The following extensions of the model (1.1) will be considered:

(i) Model (1.1) is a natural benchmark for a general additive model

$$Y = f(X) + g(\mathbf{Z}) + \sigma(X,\mathbf{Z})\varepsilon, \tag{1.6}$$

where $g(\mathbf{z})$ is a nuisance D-dimensional additive component integrated to zero on $[0,1]^D$. There is a vast literature devoted to univariate additive models [15, 16, 18, 19, 21, 23–25, 34, 36], with the most advanced sharp-minimax result due to Horowitz, Klemela and Mammen [21] where, for the case of a known $\sigma(x,\mathbf{z})=\sigma$, $g(\mathbf{z})=g_1(z_1)+\cdots+g_D(z_D)$ with differentiable univariate additive components, and known parameters α, Q and σ, a shrinkage estimator $\check{f}(x,\alpha,Q,\sigma)$ is proposed such that for any $C>0$,

$$\sup_{f\in\mathcal{S}(\alpha,Q)}\Pr\Bigg((n/d_1)^{2\alpha/(2\alpha+1)}P^{-1}(\alpha,Q) \times E\Bigg\{\int_0^1\big(\check{f}(x,\alpha,Q,\sigma)-f(x)\big)^2dx|(X_1,\mathbf{Z}_1),\dots,(X_n,\mathbf{Z}_n)\Bigg\} > 1+C\Bigg) = o_n(1).$$

We will show shortly that without any assumption on the structure of unknown $g(\mathbf{z})$ there exists a data-driven sharp-minimax estimator. In other words, the presence of a nuisance additive component $g(\mathbf{z})$ affects neither minimax rate, nor the sharp minimax constant, nor the ability of adaptive estimation.

(ii) It is of interest to relax the assumption about independence between the regression error and covariates as well as the assumption about normal distribution of the error. It will be shown shortly that the MISE of the proposed regression estimator still attains the minimax lower bound (1.2), with d_1 being replaced by d, whenever the regression error satisfies

$$E\{\varepsilon|X,\mathbf{Z}\}=0,\qquad E\{\varepsilon^2|X,\mathbf{Z}\}=1,\qquad E\{\varepsilon^4|X,\mathbf{Z}\}<C<\infty\qquad\text{a.s.}\tag{1.7}$$

To compare with a known assumption for a univariate regression, in [12] for model (1.1) with $\sigma(x,\mathbf{z})=\sigma(x)$ the proposed adaptive estimation assumes independence of the predictor and regression error ε plus a finite eighth moment of the regression error.

(iii) Extension (ii) is a natural bridge to other classical heteroscedastic models as well as to discrete responses. In this paper Bernoulli and Poisson regressions, that are inherently heteroscedastic, are considered. Note that these regressions create a new issue of satisfying bona fide properties of the regression function, and the following extension is instrumental in solving the issue.

(iv) As we shall see shortly, it is worthwhile to replace a single Sobolev class $\mathcal{S}(\alpha,Q)$ by a family $\mathcal{F}$ of function classes that includes Sobolev, local Sobolev (introduced in Golubev [17]) and shrinking (toward a pivotal regression function)

Sobolev classes as particular cases. Namely, set

$$\mathcal{F} := \mathcal{F}(f_0, \rho_n, M_n, \alpha, Q)$$

$$:= \Bigg\{ f(x) : f(x) = \sum_{j=0}^{M_n-1} \int_0^1 f_0(u)\varphi_j(u)\,du\varphi_j(x) I(M_n > 0) + \sum_{j \ge M_n} \theta_j \varphi_j(x),$$

$$x \in [0,1], \tag{1.8}$$

$$\sup_{x\in[0,1]} |f_0(x)| < \infty, \int_0^1 f_0^2(x)\,dx < \infty, \theta_j := \int_0^1 f(u)\varphi_j(u)\,du,$$

$$\sum_{j \ge M_n} [1 + (\pi j)^{2\alpha}]\theta_j^2 \le Q < \infty, \sup_{x\in[0,1]} \Bigg| \sum_{j \ge M_n} \theta_j \varphi(x) \Bigg| < \rho_n,$$

$$\alpha \ge 1, 0 \le M_n < n^{1/(2\alpha+1)} / \ln^2(n), \rho_n > n^{-1/(2\alpha+1)} \ln(n) \Bigg\}.$$

Here $f_0(x)$ is a bona fide (e.g., positive for Poisson regression) regression function which will be referred to as a *pivot*, $I(\cdot)$ is the indicator and the last line in (1.8) specifies restrictions on α and numerical sequences ρ_n and M_n.

1.4. *Comments on the family $\mathcal{F}$ and minimax approach.* (a) With respect to a classical Sobolev class $\mathcal{S}(\alpha, Q)$, we have $\mathcal{S}(\alpha, Q) = \mathcal{F}(0, \infty, 0, \alpha, Q)$, and if the pivot is constant $f_0(x) = C$, $C < Q^{1/2}$, then $\mathcal{F}(C, \rho_n, 1, \alpha, Q - C^2) \subset \mathcal{S}(\alpha, Q)$. As a result, the classical Sobolev class is a particular (not changing with n) member of the family. A function f from the family is not farther than ρ_n in L_∞-norm from the pivot. Furthermore, if $M_n > 0$, then on M_n low frequencies the regression function f is equal to the pivot, and on higher frequencies it is not farther than ρ_n in L_∞-norm and not farther than $([1 + (\pi M_n)^{2\alpha}]^{-1} Q)^{1/2}$ in L_2-norm. As a result, if either ρ_n or M_n^{-1} vanishes as $n \to \infty$, the set of considered regression functions shrinks toward the pivot. This allows us to conclude that the family $\mathcal{F}$ includes local Sobolev classes shrinking in L_2-norm, or L_∞-norm, or in both norms to the pivot. Two other shrinking properties are $\mathcal{F}(f_0, \rho, M, \alpha, Q) \subset \mathcal{F}(f_0, \rho + \gamma, M, \alpha, Q)$ and $\mathcal{F}(0, \rho, M + \gamma, \alpha, Q) \subset \mathcal{F}(0, \rho, M, \alpha, Q)$, $\gamma > 0$. Let us also note that a local Sobolev class, proposed in Golubev [17], can be written as $f_0 + \mathcal{F}(0, \rho, 0, \alpha, Q)$ where $f_0 \in \mathcal{S}(\alpha', Q')$, $\alpha' > \alpha$ and $\rho > 0$. Furthermore, let us note that $n^{1/(2\alpha+1)}$ is the classical number of Fourier coefficients that should be estimated by a rate-minimax estimator; this sheds light on the upper bound in the last line of (1.8) for considered M_n. The lower bound for considered ρ_n is due to a specific least favorable prior distribution of parameters which is used in establishing the minimax lower bound.

(b) It may be convenient to think about both the function family (1.8) and the minimax approach in terms of the game theory. There are three players in a minimax game: the dealer, nature and the statistician. The game is defined by: (i) an

underlying *model* [here a regression model (1.6)]; (ii) assumptions about *nuisance functions* [here the additive component $g(\mathbf{z})$, scale $\sigma(x,\mathbf{z})$, distribution of the error ε and the design density $p(x,\mathbf{z})$]; (iii) *parameters* of a family $\mathcal{F}$ which defines a class of estimated functions $f(x)$ [here $\mathcal{F}$ is defined in (1.8) and the parameters are the pivotal regression $f_0(x)$, sequences M_n and ρ_n and Sobolev's α and Q]. The game begins with the dealer dealing nuisance functions and parameters of $\mathcal{F}$ to nature. This deal must satisfy assumptions of the game. Then for each n nature chooses a regression function f from the dealt $\mathcal{F}$ and generates a sample of size n using f and the dealt model. The dealer and the statistician, using the sample, estimate f. The dealer knows everything apart of estimated f, the statistician knows the sample, all assumptions of the game plus may know some nuisance functions (like the design density in controlled regressions or the distribution of error in special regression models like Poisson). Nature tries to select most difficult regression function f for estimation, and the dealer and the statistician try to estimate it with the smallest MISE. The dealer has an advantage of knowing the dealt $\mathcal{F}$ and nuisance functions, and therefore the dealer's MISE may serve as a lower bound (benchmark) for the statistician.

(c) Using family (1.8) of function classes in place of a single Sobolev class allows us to answer (at least partially) to a familiar criticism of a minimax approach that the statistician cares only about the worst case scenario regression from $\mathcal{S}(\alpha, Q)$ which can be far from an underlying regression function. This is where introducing a pivot whose smoothness is not restricted, together with the possibility to consider shrinking function classes, shines.

(d) The smaller a function class is, the smaller the minimax MISE (for the dealer and the statistician) may be. This is where the imposed restriction [see the last line in (1.8)] on the dealer's choice of deals comes into the play. As we shall see shortly, none of the legitimate deals (which may imply local and/or shrinking function classes) changes a sharp lower bound known for a classical Sobolev class $\mathcal{S}(\alpha, Q)$. On the other hand, not all estimates, that are sharp minimax for Sobolev classes, are even rate minimax for particular deals. For instance, classical estimates based on the Pinsker smoothing, used for a univariate regression model in Efromovich [7] and an additive regression model in Horowitz, Klemela and Mammen [21], are sharp minimax for a Sobolev class, but not even rate minimax for $\mathcal{F}$ whenever pivot f_0 and sequence M_n are such that $\sum_{j=1}^{M_n} j^{2\alpha}[\int_0^1 f_0(x)\varphi_j(x)\,dx]^2 \to \infty$ as $n \to \infty$. In other words, if the pivot is not a Sobolev function of order α, then the famous Pinsker smoothing is no longer even rate minimax. We will prove this assertion in the Appendix (see [11]).

1.5. *Terminology.* The aforementioned approach [Section 1.4(b)] allows us to introduce the following terminology. *Estimator* is a statistic based on a sample, made assumptions and, if known, on nuisance functions defining model (1.6). In what follows we will explicitly state what nuisance functions, if any, are known.

Dealer-estimator knows everything about model (1.6) apart of the regression function f chosen by nature and also knows the dealt class (1.8). As an example, we may say that (1.2) is the lower bound for the minimax MISE where the supremum is taken over all regression functions from $\mathcal{S}(\alpha, Q)$, and the infimum is taken over all possible dealer-estimators. *Oracle-estimator* knows everything that a dealer-estimator does plus a regression function f chosen by nature. As we shall see shortly, they may be useful in suggesting a good estimator.

The context of the paper is as follows. Section 2 presents main theoretical results. Section 3 presents the methodology, estimators and a discussion of assumptions and results, for a ladder of regression models where each model is of interest on its own. Section 4 is devoted to a numerical study. Proofs, notes and more discussion can be found in the online Appendix (see [11]).

2. Main results. We begin with lower bounds and then show that they are sharp (attainable) by estimators.

2.1. *Lower bounds for dealer-estimators.* Using terminology of the Introduction, our aim is to propose a lower minimax bound for all possible dealer-estimators that know: (i) A sample of size n; (ii) Model (1.6) where nuisance functions $g(\mathbf{z})$, $\sigma(x, \mathbf{z})$ and joint design density $p(x, \mathbf{z})$ are given and ε is an independent standard normal random variable; (iii) Pivot f_0, constants α and Q and sequences ρ_n and M_n used to define a family (1.8). In other words, a dealer-estimator $\tilde{f}^*$ knows everything apart of a regression function f and

$$\tilde{f}^*(x) := \tilde{f}^*\big(x, (X, \mathbf{Z}, Y)^n, f_0(x), g(\mathbf{z}), p(x, \mathbf{z}), \sigma(x, \mathbf{z}), \rho_n, M_n, \alpha, Q\big). \tag{2.1}$$

Here $(X, \mathbf{Z}, Y)^n := ((X_1, \mathbf{Z}_1, Y_1), \dots, (X_n, \mathbf{Z}_n, Y_n))$ denotes a sample.

Please note that, for the dealer who knows the additive component $g(\mathbf{z})$, model (1.6) is equivalent to the pivotal model (1.1).

ASSUMPTION 2.1. In models (1.1) and (1.6) the regression error ε is standard normal and independent of $(X, \mathbf{Z})$.

ASSUMPTION 2.2. The joint design density $p(x, \mathbf{z})$ of $(X, \mathbf{Z})$ is supported on $[0, 1]^{D+1}$, and $\max(|\ln(p(x, \mathbf{z}))|, |\ln(\sigma(x, \mathbf{z}))|)$ is bounded on $[0, 1]^{D+1}$. Function $\mathcal{I}(x) := \int_{[0,1]^D} p(x, \mathbf{z})\sigma^{-2}(x, \mathbf{z})\, d\mathbf{z}$ is Riemann integrable on $[0, 1]$.

THEOREM 2.1. *Let Assumptions 2.1 and 2.2 hold. Then for models (1.1) and (1.6) the following lower minimax bound for dealer-estimators (2.1) holds*:

$$\begin{aligned}\inf_{\tilde{f}^*} \sup_{f \in \mathcal{F}(f_0, \rho_n, M_n, \alpha, Q)} & E\left\{\int_0^1 \big[\tilde{f}^*(x) - f(x)\big]^2 dx\right\} \\ & \geq P(\alpha, Q)\left[\int_0^1 \frac{dx}{\int_{[0,1]^D} p(x, \mathbf{z})\sigma^{-2}(x, \mathbf{z})\, d\mathbf{z}} n^{-1}\right]^{2\alpha/(2\alpha+1)} (1 + o_n(1)),\end{aligned} \tag{2.2}$$

where the infimum is taken over all possible dealer-estimators $\tilde{f}^*$, *and* $P(\alpha, Q)$ *is defined in* (1.3).

Remember that $\mathcal{S}(\alpha, Q) = \mathcal{F}(0, \infty, 0, \alpha, Q)$, and this implies that the lower bound also holds for classical Sobolev classes. Let us also note that for the case $\sigma(x, \mathbf{z}) = \sigma(x)$, with positive and having bounded derivatives on $[0, 1]$ functions $p(x)$ and $\sigma(x)$ and Sobolev regression functions, the lower bound (2.2) is known from [12] where it is established via the equivalence (between regression and filtering in white noise) principle. In this paper a different technique of finding a lower bound is employed which allows us to relax the assumptions.

The lower bound (2.2) is challenging for an estimator to match because the dealer knows everything apart from an underlying regression function. Nonetheless, as we shall see shortly, it is possible to propose an estimator that matches performance of the best dealer-estimator.

Now let us consider two classical discrete nonparametric regression models, *Bernoulli and Poisson* [7, 15]. They may be defined as (1.6) where now the distribution of ε depends on $(X, \mathbf{Z})$ and $Y \in \{0, 1\}$ in the Bernoulli case and $Y \in \{0, 1, \ldots\}$ in the Poisson case. Another way to describe these models is as follows: (i) For Bernoulli regression we observe a sample from $(X, \mathbf{Z}, Y)$ where Y is Bernoulli and $\Pr(Y = 1|X, \mathbf{Z}) = f(X) + g(\mathbf{Z})$; (ii) For Poisson regression we observe a sample from $(X, \mathbf{Z}, Y)$ where Y is Poisson and $E\{Y|X, \mathbf{Z}\} = f(X) + g(\mathbf{Z})$. Furthermore, there is an extra bona fide restriction on estimated regression functions. For Bernoulli case a regression function takes on values between zero and one, and for Poisson case a regression function is positive. This is the place where using a pivot and local/shrinking classes becomes handy.

These two regressions are inherently heteroscedastic because for the Bernoulli regression

$$\sigma^2(x, \mathbf{z}) := \sigma^2_{fg}(x, \mathbf{z}) := \big[f(x) + g(\mathbf{z})\big]\big[1 - f(x) - g(\mathbf{z})\big] \tag{2.3}$$

and for the Poisson regression

$$\sigma^2(x, \mathbf{z}) := \sigma^2_{fg}(x, \mathbf{z}) := f(x) + g(\mathbf{z}). \tag{2.4}$$

This is another specific of these regressions because the scale function contains extra information about the estimand (the regression function). Can this information help and improve the minimax MISE convergence? As the following result shows, the answer is "no."

THEOREM 2.2. *Consider the above-described Bernoulli and Poisson regressions. Suppose that Assumption* 2.2 *holds with correspondingly defined scale functions* (2.3) *or* (2.4), *and in* (1.8) $M_n \to \infty$ *as* $n \to \infty$. *For all* $(x, \mathbf{z}) \in [0, 1]^{D+1}$ *it is assumed that the pivot* $f_0(x)$, *used in* (1.8), *satisfies* $0 < C_* \le f_0(x) + g(\mathbf{z})$ *and*

additionally for the Bernoulli regression $f_0(x) + g(\mathbf{z}) \le C^* < 1$. *Then for both regressions*,

$$\inf_{\tilde{f}^*} \sup_{f \in \mathcal{F}(f_0,\rho_n,M_n,\alpha,Q)\cap\mathcal{F}^*(g)} E\left\{\int_0^1 [\tilde{f}^*(x) - f(x)]^2\,dx\right\} \tag{2.5}$$
$$\ge P(\alpha, Q)\left[\int_0^1 \frac{dx}{\int_{[0,1]^D} p(x,\mathbf{z})\sigma^{-2}_{f_0 g}(x,\mathbf{z})\,d\mathbf{z}} n^{-1}\right]^{2\alpha/(2\alpha+1)} (1 + o_n(1)),$$

where the infimum is taken over all possible dealer-estimators $\tilde{f}^*$, $\mathcal{F}^*(g)$ *is a class of all bona fide* f *and* $P(\alpha, Q)$ *is defined in* (1.3).

As we see, the lower oracle's bounds are the same for the normal regression with continuous responses and Bernoulli and Poisson regressions with discrete responses; this can be explained by the fact that conditional distributions of responses, given covariates, belong to exponential families [6, 27].

The following result, whose proof and a specific dealer-estimator can be found in the Appendix (see [11]), shows that the lower bounds are sharp.

THEOREM 2.3. *Lower bounds* (2.2) *and* (2.5) *are attainable by a dealer-estimator* $\check{f}^*(x)$, *that is*, $\sup_{f\in\mathcal{S}(f_0,\rho_n,M_n,\alpha,Q)} E\{\int_0^1[\check{f}^*(x) - f(x)]^2\,dx\}$ *is not greater than the right-hand sides of* (2.2) *and* (2.5) *for the Normal and Bernoulli/Poisson regressions considered in Theorems* 2.1 *and* 2.2, *respectively.*

2.2. *Sharpness of lower bounds for estimators*. Our aim is to show that an estimator can match performance of a dealer-estimator, that is, an estimator can be adaptive (to underlying function class and nuisance functions in a regression model) and sharp minimax.

Introduce: a tensor-product cosine basis $\{\psi_{\mathbf{s}}(\mathbf{v}) := \prod_{t=1}^D \varphi_{s_t}(v_t)$, $\mathbf{s} := (s_1, \dots, s_D) \in \{0, 1, \dots\}^D$, $\mathbf{v} := (v_1, \dots, v_D) \in [0,1]^D\}$, l_∞-norm $\|\mathbf{s}\|_\infty := \max(s_1, \dots, s_D)$, analytic function class $\mathcal{A} := \mathcal{A}(\beta_0, \dots, \beta_D, Q_1) := \{q(x,\mathbf{z}) : q(x,\mathbf{z}) := \sum_{i,\mathbf{s}} \pi_{i\mathbf{s}}\varphi_i(x)\psi_{\mathbf{s}}(\mathbf{z}), |\pi_{i\mathbf{s}}| \le Q_1[e^{\beta_0 i} + \sum_{k=1}^D e^{\beta_k s_k}]^{-1}, \min(\beta_0, \beta_1, \dots, \beta_D) > 0, Q_1 < \infty\}$ and a k-variate Sobolev class $\mathcal{S}_k := \{q(x_1, \dots, x_k) : q(x_1, \dots, x_k) = \sum_{i_1,\dots,i_k=0}^\infty q_{i_1,\dots,i_k} \prod_{s=1}^k \varphi_{i_s}(x_{i_s}), \sum_{i_1,\dots,i_k=0}^\infty [1 + \sum_{s=1}^k (2\pi i_s)^{2k}] \times q^2_{i_1,\dots,i_k} \le Q_2 < \infty\}$; see [7, 30, 35]. Parameters of the classes are unknown to the statistician. In what follows ν's are generic nonnegative constants that are used as powers, and C's are generic positive constants used as factors.

For convenience of future references, let us introduce an array of assumptions.

ASSUMPTION 2.3. The following assumptions may be used in different propositions:

(a) Assumption 2.2 holds and regression error ε satisfies (1.7).

(b) Nuisance additive component $g(\mathbf{z})$ is bounded and integrable on $[0,1]^D$ to zero.

(c) The design density satisfies for some $\nu > 0$,

$$\sum_{(j,\mathbf{s})\in\{0,1,\dots\}^{D+1}} \left| \int_{[0,1]^{D+1}} p(x,\mathbf{z})\varphi_j(x)\psi_{\mathbf{s}}(\mathbf{z})\,dx\,d\mathbf{z} \right| \le C\ln^{\nu}(n) \tag{2.6}$$

and for some positive constant ν_0 and any $t > n^{\nu_0}$,

$$\sum_{j=0}^{\infty}\sum_{\|\mathbf{s}\|>t} \left[\int_{[0,1]^{D+1}} p(x,\mathbf{z})\varphi_j(x)\psi_{\mathbf{s}}(\mathbf{z})\,dx\,d\mathbf{z} \right]^2 \le C\ln^{\nu}(n)t^{-D}. \tag{2.7}$$

(d) The L_2-approximation of additive component $g(\mathbf{z})$ satisfies for any $t > 0$ and some $\nu > 0$

$$\sum_{\|\mathbf{s}\|>t} \left[\int_{[0,1]^D} g(\mathbf{z})\psi_{\mathbf{s}}(\mathbf{z})\,d\mathbf{z} \right]^2 \le Ct^{-\nu}. \tag{2.8}$$

(e) Two constants, c_* and c^*, are given such that $0 < c_* \le \sigma^2(x,\mathbf{z}) \le c^* < \infty$.

(f) Design density $p(x,\mathbf{z})$ belongs to an analytic class $\mathcal{A}$.

(g) Design density $p(x,\mathbf{z})$ belongs to a $(D+1)$-variate Sobolev class $\mathcal{S}_{D+1}$ and nuisance component $g(\mathbf{z})$ belongs to a D-variate Sobolev class $\mathcal{S}_D$.

Let us note that: in part (c) a larger class of densities is allowed for larger n; if in part (g) we additionally assume that $g(\mathbf{z}) = \sum_{r=1}^{D} g_r(z_r)$, then the familiar assumption $g_r \in \mathcal{S}_1$, $r = 1,\dots,D$, is sufficient and the corresponding proof can be found in the Appendix (see [11]).

The following proposition presents a ladder of settings, each of interest on its own, for which sharp-minimax and adaptive estimation is possible. A discussion of the settings and proposed estimators will be presented in Section 3.

THEOREM 2.4. *Consider a general additive regression model* (1.6) *with the regression error that may depend on covariates* $(X,\mathbf{Z})$ *and satisfying* (1.7). *Then for each of the following five sets of assumptions there exists an estimator that is sharp minimax and matches performance of the dealer-estimator outlined in Theorem* 2.3:

(1) *Additive component* $g(\mathbf{z})$, *design density and scale are known and Assumption* 2.3(a) *holds*.

(2) *Design density and scale are known and Assumption* 2.3(a)–(d) *holds*.

(3) *Design density is known and Assumption* 2.3(a)–(e) *holds*.

(4) *Assumption* 2.3(a), (b), (d), (e), (f) *holds*.

(5) *Assumption* 2.3(a), (b), (e), (g) *holds*.

This result implies the following proposition.

COROLLARY 2.1. *Consider Bernoulli and Poisson regression models discussed in Theorem 2.2. Then the assertion of Theorem 2.4 holds, and the same estimators attain the minimax lower bound of Theorem 2.2.*

3. Estimation. We begin with an explanation of the *methodology* of sharp-minimax estimation. Two technical results are presented for a general regression model. The former is about a blockwise-shrinkage oracle-estimator which is adaptive and sharp-minimax. The latter is about sufficient conditions for an estimator to mimic the oracle. These two results shed light on the underlying methodology of constructing sharp-minimax estimators and are of interest on their own. Then we are presenting specific estimators for each setting considered in Theorem 2.4.

To propose a blockwise-shrinkage oracle-estimator, let $\{B_k, k = 1, 2, \ldots\}$ be a partition of nonnegative integers [frequencies of the cosine basis $\{\varphi_j(x), j = 0, 1, \ldots\}$] into nonoverlapping blocks of cardinality (length) L_k such that $\max(j : j \in B_k) < \min(j : j \in B_{k+1})$. The blockwise-shrinkage oracle-estimator is defined as

$$\hat{f}^*(x) := \sum_{k=1}^{K_n} \mu_k \sum_{j \in B_k} \hat{\theta}_j \varphi_j(x), \tag{3.1}$$

where K_n is some positive, nondecreasing and integer-valued sequence,

$$\mu_k := \frac{\Theta_k}{\Theta_k + dn^{-1}} \tag{3.2}$$

is the oracle's shrinkage coefficient for frequencies from the block B_k, $d := d(p, \sigma)$ is the coefficient of difficulty (1.5) that appears in the lower bounds (2.2) and (2.5),

$$\Theta_k := L_k^{-1} \sum_{j \in B_k} \theta_j^2 \tag{3.3}$$

is the Sobolev functional which defines the average energy of $f(x)$ on frequencies from the block B_k. A statistic $\hat{\theta}_j$, used in (3.1), is an appropriate estimator of the Fourier coefficient $\theta_j = \int_0^1 f(x)\varphi_j(x)\,dx$. For the purposes of this paper, the oracle should be able to suggest a statistic whose mean squared error (MSE) satisfies

$$E\{(\hat{\theta}_j - \theta_j)^2\} \le dn^{-1}\bigl(1 + o_n(1) + o_j(1) + a_j^2 \ln^{\nu}(n)\bigr), \tag{3.4}$$

where d is defined in (1.5), and its squared bias satisfies

$$\bigl[E\{\hat{\theta}_j\} - \theta_j\bigr]^2 \le n^{-1}\bigl[o_n(1) + o_j(1) + a_j^2 \ln^{\nu}(n)\bigr]. \tag{3.5}$$

Here and in what follows $\{a_j^2\}$'s are generic summable sequences ($\sum_{j=0}^{\infty} a_j^2 < \infty$) and ν's are generic nonnegative constants that are used in powers.

The following result explains why it is worthwhile to consider the oracle-estimator (3.1).

LEMMA 3.1. *Suppose that in* (3.1) *the sequence* K_n *is large enough to satisfy the inequality* $\sum_{k=1}^{K_n} L_k > n^{1/(2\alpha+1)} \ln(\ln(n+20))$, *and* (3.4)–(3.5) *hold. Then*

$$\sup_{f\in\mathcal{F}(f_0,\infty,M_n,\alpha,Q)} E\left\{\int_0^1 \bigl(\hat{f}^*(x)-f(x)\bigr)^2\,dx\right\} \le P(\alpha,Q)(d/n)^{2\alpha/(2\alpha+1)}\bigl(1+o_n(1)\bigr). \tag{3.6}$$

Let us make several comments about this result: (i) Lemma 3.1 does not refer to or is based on a specific regression model; (ii) It was explained in the Introduction that $\mathcal{F}(0,\infty,0,\alpha,Q)=\mathcal{S}(\alpha,Q)$ and thus the presented upper bound holds for classical Sobolev classes; (iii) Using lower bounds of Section 2 and relation $\mathcal{F}(f_0,\rho_n,M_n,\alpha,Q)\subset\mathcal{F}(f_0,\infty,M_n,\alpha,Q)$, we conclude that the oracle-estimator is adaptive and sharp-minimax.

Now we are in a position to describe the proposed methodology of developing a data-driven estimator that mimics the oracle-estimator and is sharp-minimax.

Let us introduce several new sequences and specific blocks used from now on. Set: $b_n := \lfloor \ln(n+20)\rfloor$ where $\lfloor x\rfloor$ denotes the largest integer which is at most x; $c_n := \lfloor \ln(b_n)\rfloor$; $m := \lfloor n/(7c_n)\rfloor$ and it is assumed that n is large enough so $m>3$; $L_k := 1$ for $k=1,2,\dots,b_n$ and $L_k := \lfloor (1+b_n^{-1})^k\rfloor$ for $k>b_n$; K_n is the smallest integer such that $\sum_{k=1}^{K_n} L_k > n^{1/3}c_n$; $B_k := \{k-1\}$ for $k=1,2,\dots,b_n$ and $B_k := \{\sum_{s=1}^{k-1} L_s, \sum_{s=1}^{k-1} L_s+1,\dots,\sum_{s=1}^{k} L_s-1\}$ for $b_n<k\le K_n$.

Let us comment on the specific choice of blocks. The first b_n blocks have unit lengths, and this choice is motivated by good performance for small samples. Then the length of blocks increases geometrically but in such a way that $L_{k+1}/L_k = 1+o_n(1)$. This choice is motivated by the asymptotic analysis together with a good performance for small samples. Let us note that the number of considered blocks, K_n, is of order $\ln^2(n)$. The largest length of the blocks, L_{K_n}, is of order $n^{1/3}[\ln(\ln(n))]/\ln(n)$. The total number of estimated low frequency Fourier coefficients is of order $n^{1/3}\ln(\ln(n))$. This choice is explained by the fact that the sum of not estimated squared Fourier coefficients is of order $o_n(1)n^{-2\alpha/(2\alpha+1)}$ whenever $\alpha\ge 1$. Another way to look at this choice is as follows. It is known [4, 6, 8–10, 13–15] that for Sobolev's functions of order α at most $n^{1/(2\alpha+1)}c_n$ first Fourier coefficients should be estimated, and this defines the choice of K_n. Furthermore, if it is additionally known that $\alpha\ge\alpha_0$, then the total number can be changed to $n^{1/(2\alpha_0+1)}c_n$.

The following proposition explains how to develop an estimator that matches performance of the oracle.

LEMMA 3.2. *Suppose that there exist two arrays of statistics* $\{\hat{\Theta}_k, k=1,\dots,K_n\}$ *and* $\{\hat{\theta}_j, j=0,\dots,\sum_{k=1}^{K_n} L_k\}$, *and a statistic* $\hat{d}$ *such that the two arrays and* $\hat{d}$ *are mutually independent, the array* $\{\hat{\theta}_j\}$ *satisfies* (3.4)–(3.5), *the array*

$\{\hat\Theta_k\}$ satisfies for some positive constants C_1 and ν_1

$$E\{(\hat\Theta_k-\Theta_k)^4\}\le C_1L_k^{-1}b_n^{\nu_1}n^{-2}(\Theta_k+n^{-1})^2, \tag{3.7}$$

and the statistic $\hat d$ satisfies for some constant $C_2\ge 1$

$$E\left\{\frac{(\hat d-d)^2}{\hat d}\right\}=o_n(1),\qquad \hat d\in\left[(C_2b_n)^{-1/4},(C_2b_n)^{1/4}\right]\qquad a.s. \tag{3.8}$$

Then the blockwise-shrinkage estimator

$$\hat f(x):=\sum_{k=1}^{K_n}\frac{\hat\Theta_k}{\hat\Theta_k+\hat dn^{-1}}I\bigl(\hat\Theta_k>(b_nn)^{-1}\bigr)\sum_{j\in B_k}\hat\theta_j\varphi_j(x), \tag{3.9}$$

which mimics the oracle-estimator (3.1), inherits the sharp-minimax property of the oracle-estimator described in Lemma 3.1, namely

$$\begin{aligned}&\sup_{f\in\mathcal F(f_0,\infty,M_n,\alpha,Q)}E\left\{\int_0^1\bigl(\hat f(x)-f(x)\bigr)^2\,dx\right\}\\&\qquad\le P(\alpha,Q)(d/n)^{2\alpha/(2\alpha+1)}\bigl(1+o_n(1)\bigr).\end{aligned} \tag{3.10}$$

Now we are in a position to consider settings (1)–(5) of Theorem 2.4 in turn, and propose corresponding statistics $\{\hat\theta_j,\hat\Theta_k,\hat d\}$ used in the estimator (3.9).

3.1. *Known additive component, design and scale.* This is the case where model (1.6) transforms into the pivotal model (1.1). Because nuisance additive component $g(\mathbf z)$ is known, without loss of generality we could assume that $g(\mathbf z)=0$ or replace Y by $Y-g(\mathbf Z)$. However, we do not do this because we would like to indicate what may be done for the case of unknown g. Our idea is to mimic oracle (3.1) via application of Lemma 3.2. To do this, we need to suggest estimators for Sobolev functionals Θ_k and Fourier coefficients θ_j; note that the coefficient of difficulty d, defined in (1.5), is known. Set

$$\hat\theta_j:=\frac{1}{n-2m}\sum_{l=2m+1}^{n}\frac{[Y_l-\tilde f_{-j}(X_l)-g(\mathbf Z_l)]\sigma^{-2}(X_l,\mathbf Z_l)\varphi_j(X_l)}{\mathcal I(X_l)}, \tag{3.11}$$

where

$$\mathcal I(x):=\int_{[0,1]^D}p(x,\mathbf z)\sigma^{-2}(x,\mathbf z)\,d\mathbf z, \tag{3.12}$$

$$\tilde f_{-j}(x)=m^{-1}\sum_{l=1}^{m}\sum_{i\in\mathcal N_{-j}}\frac{(Y_l-g(\mathbf Z_l)\varphi_j(X_l)}{p(X_l)}\varphi_i(x) \tag{3.13}$$

and $\mathcal{N}_{-j} := \{0, 1, \dots, b_n\} \setminus \{j\}$. Note that $\tilde{f}_{-j}(x)$ estimates $f_{-j}(x) := f(x) - \theta_j \varphi_j(x)$. Further,

$$\hat{\Theta}_k := \frac{2}{m(m-1)} \sum_{m+1 \le l_1 < l_2 \le 2m} L_k^{-1} \sum_{j \in B_k} \frac{Y_{l_1} Y_{l_2} \varphi_j(X_{l_1}) \varphi_j(X_{l_2})}{p(X_{l_1}, \mathbf{Z}_{l_1}) p(X_{l_2}, \mathbf{Z}_{l_2})}, \tag{3.14}$$

and note that this is U-statistic and unbiased estimate of Θ_k. This special form of the estimator $\hat{\Theta}_k$ (it is different from those used in [4, 7, 8, 12]) implies existence of the fourth moment of $\hat{\Theta}_k$ given existence of the fourth moment of the regression error. Another remark is that we may use the marginal density of X in place of the joint design probability density if $Y_l - g(\mathbf{Z}_l)$ is used in the numerator of (3.14) in place of Y_l.

Let us comment on the estimator (3.11) of Fourier coefficients θ_j. First, the statistic $\tilde{f}_{-j}$ is subtracted from the response to decrease the MSE. If the subtraction is skipped then in (3.4) we would have a larger factor $(d + \int_0^1 f^2(x)\,dx)$ in place of the wished d. Second, the estimator uses weights (remember the discussion in the Introduction)

$$w_l = \frac{\sigma^{-2}(X_l, \mathbf{Z}_l)}{\mathcal{I}(X_l)} = \frac{\sigma^{-2}(X_l, \mathbf{Z}_l)}{p(X_l) E\{\sigma^{-2}(X, \mathbf{Z}) | X = X_l\}}. \tag{3.15}$$

This choice of weights yields the wished properties (3.4)–(3.5). Note that if $\sigma(x, \mathbf{z}) = \sigma(x)$, then weights (3.15) do not depend on the scale.

PROPOSITION 3.1. *Consider setting* (1) *of Theorem* 2.4. *Then the blockwise-shrinkage regression estimator* (3.9) *where* $\hat{\Theta}_k$ *is defined in* (3.14) *and* $\hat{\theta}_j$ *in* (3.11), *is adaptive to* $(f_0(x), \rho_n, M_n, \alpha, Q)$ *and sharp minimax, that is, its MISE satisfies* (3.10).

An interesting outcome of the proposition is that no smoothness of the pivotal regression function is required for adaptive sharp-minimax estimation, and that regression error may depend on covariates and have only the fourth moment.

REMARK 3.1. In Section 4, where estimators are tested on small samples, we will study D-estimator which is the above-defined estimator without splitting data. Similarly, all other proposed estimators, when used for small samples, do not split data.

3.2. *Known design and scale.* Here the main complication is an unknown additive nuisance component $g(\mathbf{z})$. To mimic the oracle we need to "remove" the nuisance component from the response, and this is a familiar approach in the additive models literature. As it is shown in the Appendix (see [11]), this straightforward approach requires an extra assumption about smoothness of the scale. Because the main topic of the paper is heteroscedasticity, it is of interest to assume as little

as possible about the scale function. Furthermore, let us remind the reader that estimation of the scale function is a complicated statistical problem on its own because quality of estimation depends on smoothness of the regression function and the scale function [3]. As a result, even if for now the scale function is known, it is desirable to assume as little as possible about its properties and then later use a simple estimator of the scale.

The recommended approach is to replace the known $\sigma^{-2}(x,\mathbf{z})$ by its Fejér approximation of order b_n,

$$\begin{aligned}&\sigma_{b_n}^{-2}(x,\mathbf{z})\\ &:= b_n^{-1}\sum_{t=0}^{b_n-1}\sum_{\|(i,\mathbf{s})\|_\infty\le t}\left[\int_{[0,1]^{D+1}}\sigma^{-2}(u,\mathbf{v})\varphi_i(u)\psi_{\mathbf{s}}(\mathbf{v})\,du\,d\mathbf{v}\varphi_i(x)\psi_{\mathbf{s}}(\mathbf{z})\right]\\ &=:\sum_{\|(i,\mathbf{s})\|_\infty<b_n}\eta_{i\mathbf{s}}\varphi_i(x)\psi_{\mathbf{s}}(\mathbf{z}).\end{aligned}\tag{3.16}$$

Here $\eta_{i\mathbf{s}}$ are Fejér coefficients (note that they depend on the order b_n). The Fejér approximation has a unique property of preserving the range of approximated $\sigma^{-2}(x,\mathbf{z})$; see more about this nice trigonometric approximation in [2, 7, 33, 37]. Note that while using Fejér's approximation is important, the choice of its order (here b_n) is flexible. We also replace known $\mathcal{I}(x)$ by the corresponding approximation

$$\mathcal{I}_{b_n}(x) := \int_{[0,1]^D} p(x,\mathbf{z})\sigma_{b_n}^{-2}(x,\mathbf{z})\,d\mathbf{z}\tag{3.17}$$

$$= \sum_{t=0}^{\infty}\sum_{\|(i,\mathbf{s})\|_\infty<b_n}\pi_{t\mathbf{s}}\eta_{i\mathbf{s}}\varphi_t(x)\varphi_i(x),\tag{3.18}$$

where $\pi_{t\mathbf{s}} := \int_{[0,1]^{D+1}} p(x,\mathbf{z})\varphi_t(x)\psi_{\mathbf{s}}(\mathbf{z})\,dx\,d\mathbf{z}$ are Fourier coefficients of the known design density.

Introduce estimates for $f_{-j}(x)$, Θ_k, $g(\mathbf{z})$, and θ_j in turn. Write

$$\tilde{f}_{-j}(x) := m^{-1}\sum_{l=1}^{m}\sum_{i\in\mathcal{N}_{-j}}\frac{Y_l\varphi_i(X_l)}{p(X_l,\mathbf{Z}_l)}\varphi_i(x),\tag{3.19}$$

where $\mathcal{N}_{-j}$ is the same as in (3.13),

$$\tilde{g}(\mathbf{z}) := m^{-1}\sum_{l=2m+1}^{3m}\sum_{\mathbf{r}\in\mathcal{N}_g}\frac{Y_l\psi_{\mathbf{r}}(\mathbf{Z}_l)}{p(X_l,\mathbf{Z}_l)}\psi_{\mathbf{r}}(\mathbf{z})\tag{3.20}$$

is the projection series estimator of $g(\mathbf{z})$ with $\mathcal{N}_g := \{0,1,\dots,N_g\}^D\setminus\{0\}^D$ and $N_g := \lfloor n^{1/D}/b_n^{2/D}\rfloor$, and

$$\hat{\theta}_j := (n-3m)^{-1}\sum_{l=3m+1}^{n}\frac{[Y_l-\tilde{f}_{-j}(X_l)-\tilde{g}(\mathbf{Z}_l)]\sigma_{b_n}^{-2}(X_l,\mathbf{Z}_l)\varphi_j(X_l)}{\mathcal{I}_{b_n}(X_l)}.\tag{3.21}$$

PROPOSITION 3.2. *Consider setting* (2) *of Theorem* 2.4. *Then the estimator* (3.9), *with* $\hat{\Theta}_k$ *defined in* (3.14), $\hat{\theta}_j$ *in* (3.21) *and* $\hat{d} = d$ *defined in* (1.5), *is adaptive and sharp minimax, that is, its MISE satisfies* (3.10).

Note that no regularity/smoothness of the scale is assumed (it can be even discontinuous), but we added a very mild assumption (2.8) on how well the nuisance additive component can be approximated by the trigonometric basis. For instance, (2.8) holds if in each variable the function $g(z_1, \dots, z_D)$ is piecewise Lipschitz of some positive order (note that Lipschitz functions of order $\beta < 1$ are often referred to as Hölder functions) [7]. The reason why the proposed Fejér approximation of $\sigma^{-2}(x, \mathbf{z})$ helps is due to the fact that it is just a weighted sum of first b_n Fourier terms of $\sigma^{-2}(x, \mathbf{z})$, that is, the approximation is an extremely smooth function. At the same time, the approximation is sufficient for mimicking the scale and satisfying (3.4)–(3.5). While this result is of interest on its own, it plays a key role in the case of an unknown scale because it indicates that a rough estimator of the scale may be sufficient for a sharp-minimax and adaptive estimation.

3.3. *Known design.* This is a familiar regression problem which includes, as a particular case, controlled design regressions [7, 13, 15, 35]. The main issue now is an appropriate estimation of the scale. In the assumption for setting (3) of Theorem 2.4 we still do not impose any restriction on smoothness of an underlying scale $\sigma(x, \mathbf{z})$ and have not added a new assumption about the additive nuisance component $g(\mathbf{z})$. On the other hand, we added Assumption 2.3(e) which requires knowledge of the range of the scale function. If the latter is unknown, then some information, on how well the scale can be approximated by the trigonometric basis, is required; see Remark A.3 in the Appendix (see [11]).

Following Lemma 3.2 we need to propose an estimate of the coefficient of difficulty d defined in (1.5), and, following Section 3.2, we need to propose an estimate of $\sigma_{b_n}^{-2}(x, \mathbf{z})$. We begin with the explanation of how to construct an estimate of d. Remember that, according to Lemma 3.2, an estimator should be independent of all other statistics. To estimate the scale function we begin with a truncated projection estimate of $q(x, \mathbf{z}) := f(x) + g(\mathbf{z})$,

$$\tilde{q}_1(x, \mathbf{z}) := \max\Biggl(-b_n, \min\Biggl(b_n, m^{-1} \sum_{l=3m+1}^{4m} \sum_{\|(i,\mathbf{r})\|_\infty < b_n} \frac{Y_l \varphi_i(X_l) \psi_{\mathbf{r}}(\mathbf{Z}_l)}{p(X_l, \mathbf{Z}_l)} \times \varphi_i(x) \psi_{\mathbf{r}}(\mathbf{z})\Biggr)\Biggr), \tag{3.22}$$

which is used in the following bona fide projection estimator of $\sigma^2(x, \mathbf{z})$:

$$\tilde{\sigma}_1^2(x, \mathbf{z}) := \max\Biggl(c_*, \min\Biggl(c^*, \sum_{\|(i,\mathbf{r})\|_\infty < b_n} \tilde{\sigma}_{1i\mathbf{r}} \varphi_i(x) \psi_{\mathbf{r}}(\mathbf{z})\Biggr)\Biggr). \tag{3.23}$$

Here $\tilde{\sigma}_{1i\mathbf{r}}$ is the estimate of Fourier coefficients $\sigma_{i\mathbf{r}}$ of $\sigma^2(x,\mathbf{z})$,

$$\sigma_{i\mathbf{r}} := \int_{[0,1]^{D+1}} \sigma^2(x,\mathbf{z})\varphi_i(x)\psi_{\mathbf{r}}(\mathbf{z})\,dx\,d\mathbf{z},$$

and the proposed estimate, motivated by the method of moments, is

$$\tilde{\sigma}_{1i\mathbf{r}} := m^{-1}\sum_{l=4m+1}^{5m} \frac{(Y_l - \tilde{q}_1(X_l,\mathbf{Z}_l))^2}{p(X_l,\mathbf{Z}_l)}\varphi_i(X_l)\psi_{\mathbf{r}}(\mathbf{Z}_l). \tag{3.24}$$

With the bona fide estimate (3.23) of $\sigma^2(x,\mathbf{z})$ at hand, we plug it in (1.5) and get

$$\tilde{d} := \int_0^1 \frac{dx}{\int_{[0,1]^D} p(x,\mathbf{z})\tilde{\sigma}_1^{-2}(x,\mathbf{z})\,d\mathbf{z}}. \tag{3.25}$$

Now we are utilizing the same approach to estimate $\sigma_{b_n}^{-2}(x,\mathbf{z})$ used by the estimator $\hat{\theta}_j$. Remember that, to follow the recipe of Lemma 3.2, this estimate should be independent of $\tilde{d}$. We define it similarly to (3.22)–(3.24),

$$\tilde{q}(x,\mathbf{z}) := \max\biggl(-b_n, \min\biggl(b_n, m^{-1}\sum_{l=5m+1}^{6m}\sum_{\|(i,\mathbf{r})\|_\infty<b_n} \frac{Y_l\varphi_i(X_l)\psi_{\mathbf{r}}(\mathbf{Z}_l)}{p(X_l,\mathbf{Z}_l)} \times \varphi_i(x)\psi_{\mathbf{r}}(\mathbf{z})\biggr)\biggr) \tag{3.26}$$

and

$$\tilde{\sigma}^2(x,\mathbf{z}) := \max\biggl(c_*, \min\biggl(c^*, \sum_{\|(i,\mathbf{r})\|_\infty<b_n} \tilde{\sigma}_{i\mathbf{r}}\varphi_i(x)\psi_{\mathbf{r}}(\mathbf{z})\biggr)\biggr), \tag{3.27}$$

where

$$\tilde{\sigma}_{i\mathbf{r}} := m^{-1}\sum_{l=6m+1}^{7m} \frac{(Y_l - \tilde{q}(X_l,\mathbf{Z}_l))^2}{p(X_l,\mathbf{Z}_l)}\varphi_i(X_l)\psi_{\mathbf{r}}(\mathbf{Z}_l). \tag{3.28}$$

Note that now the estimate $\tilde{\sigma}^2(x,\mathbf{z})$ plays the role of $\sigma^2(x,\mathbf{z})$, and then we apply the Fejér approximation (3.16) to the estimate (3.27) and get the estimate of $\sigma_{b_n}^{-2}(x,\mathbf{z})$,

$$\begin{aligned}&\tilde{\sigma}_{b_n}^{-2}(x,\mathbf{z})\\ &:= b_n^{-1}\sum_{t=0}^{b_n-1}\sum_{\|(i,\mathbf{s})\|_\infty\le t}\biggl[\int_{[0,1]^{D+1}} \tilde{\sigma}^{-2}(u,\mathbf{v})\varphi_i(u)\psi_{\mathbf{s}}(\mathbf{v})\,du\,d\mathbf{v}\varphi_i(x)\psi_{\mathbf{s}}(\mathbf{z})\biggr]\\ &=: \sum_{\|(i,\mathbf{s})\|_\infty<b_n} \tilde{\eta}_{i\mathbf{s}}\varphi_i(x)\psi_{\mathbf{s}}(\mathbf{z}).\end{aligned} \tag{3.29}$$

Further, following (3.17) and (3.18), we define the plug-in estimate of $\mathcal{I}_{b_n}(x)$,

$$\begin{aligned}\tilde{\mathcal{I}}_{b_n}(x) &:= \int_{[0,1]^D} p(x,\mathbf{z})\tilde{\sigma}_{b_n}^{-2}(x,\mathbf{z})\,d\mathbf{z} \\ &= \sum_{t=0}^{\infty}\sum_{\|(i,\mathbf{s})\|_\infty<b_n} \pi_{t\mathbf{s}}\tilde{\eta}_{i\mathbf{s}}\varphi_t(x)\varphi_i(x).\end{aligned} \tag{3.30}$$

Finally, mimicking (3.21), we introduce a new estimator of Fourier coefficients θ_j,

$$\hat{\theta}_j := (n-7m)^{-1}\sum_{l=7m+1}^{n}\frac{[Y_l - \tilde{f}_{-j}(X_l) - \tilde{g}(\mathbf{Z}_l)]\tilde{\sigma}_{b_n}^{-2}(X_l,\mathbf{Z}_l)\varphi_j(X_l)}{\tilde{\mathcal{I}}_{b_n}(X_l)}. \tag{3.31}$$

Here $\tilde{f}_{-j}$ and $\tilde{g}$ are estimates (3.19) and (3.20).

PROPOSITION 3.3. *Consider setting* (3) *of Theorem* 2.4. *Then the estimator* (3.9), *with* $\hat{\Theta}_k$ *defined in* (3.14), $\hat{\theta}_j$ *defined in* (3.31) *and* $\hat{d} = \tilde{d}$ *defined in* (3.25), *is adaptive and sharp minimax*, *that is*, *its MISE satisfies* (3.10).

Note that a rough estimate of the scale is sufficient, and no assumption about smoothness of an underlying scale function is made.

3.4. *Unknown nuisance functions*. Here we relax the last assumption that the design density p is known. We are considering setting (4) of Theorem 2.4 (with analytic $p \in \mathcal{A}$) and setting (5) (with Sobolev $p \in \mathcal{S}_{D+1}$) simultaneously to highlight similarities and differences in proposed estimators. We will use the indicator $I(p \notin \mathcal{A}) = 1$ for the case of setting (5). Remember that Sobolev classes were discussed in the Introduction, a nice discussion of analytic functions can be found in [1, 26, 30, 37] and in [7] they are recommended for modeling and approximation of a wide variety of densities for the case of small data sets.

Because now the design is unknown, all previously defined estimates become dealer-estimates, and we will use a standard plug-in technique of using a density estimate in place of an unknown design density. To follow the recipe of Lemma 3.2, we need to plug-in independent design density estimates in different oracle-estimates, and this forces us to rewrite one more time all statistics. This is a good review of what we have done so far. Remember our notation $b_n := \lfloor \ln(n+20) \rfloor$, $c_n := \lfloor \ln(b_n) \rfloor$, and set $\mathcal{N}_p := \{0,1,\dots,N_p\}^{D+1}$, $N_p := \lfloor b_n c_n \rfloor I(p \in \mathcal{A}) + \lfloor n^{1/3(D+1)} \rfloor I(p \notin \mathcal{A})$. Note that N_p is a traditional minimax cutoff for the studied densities. Set $m := \lfloor n/[(21)c_n] \rfloor$, $\mathcal{M}_s := \{(s-1)m+1, (s-1)m+2, \dots, sm\}$, and introduce nine identical (but based on different subsamples) truncated minimax projection density estimates [5, 7]

$$\tilde{p}_s(x,\mathbf{z}) := \max\Bigg(c_n^{-1}, m^{-1}\sum_{l\in\mathcal{M}_s}\sum_{(i,\mathbf{r})\in\mathcal{N}_p}\varphi_i(X_l)\psi_{\mathbf{r}}(\mathbf{Z}_l)\varphi_i(x)\psi_{\mathbf{r}}(\mathbf{z})\Bigg), \tag{3.32}$$

where $s = 1, \dots, 9$. We have truncated the projection density estimate from below by c_n^{-1} because its reciprocal will be used.

Now we can define statistics used by the proposed estimator. The first one is the estimator mimicking dealer-estimator (3.25) of the coefficient of difficulty d. We begin with mimicking dealer-estimates (3.22) and (3.23) used in (3.25). Write

$$\tilde{q}_1(x, \mathbf{z}) := \max\Bigg(-b_n, \min\Bigg(b_n, m^{-1} \sum_{l \in \mathcal{M}_{10}} \sum_{\|(i,\mathbf{r})\|_\infty < b_n} \frac{Y_l \varphi_i(X_l) \psi_{\mathbf{r}}(\mathbf{Z}_l)}{\tilde{p}_1(X_l, \mathbf{Z}_l)} \times \varphi_i(x) \psi_{\mathbf{r}}(\mathbf{z})\Bigg)\Bigg), \tag{3.33}$$

$$\tilde{\sigma}_{1i\mathbf{r}} := m^{-1} \sum_{l \in \mathcal{M}_{11}} \frac{(Y_l - \tilde{q}_1(X_l, \mathbf{Z}_l))^2}{\tilde{p}_2(X_l, \mathbf{Z}_l)} \varphi_i(X_l) \psi_{\mathbf{r}}(\mathbf{Z}_l), \tag{3.34}$$

$$\tilde{\sigma}_1^2(x, \mathbf{z}) := \max\Bigg(c_*, \max\Bigg(c^*, \sum_{\|(i,\mathbf{r})\|_\infty < b_n} \tilde{\sigma}_{1i\mathbf{r}} \varphi_i(x) \psi_{\mathbf{r}}(\mathbf{z})\Bigg)\Bigg). \tag{3.35}$$

These statistics allow us to define the estimate of d [compare with (3.25)],

$$\tilde{d} := \int_0^1 \frac{dx}{\int_{[0,1]^D} \tilde{p}_3(x, \mathbf{z}) \tilde{\sigma}_1^{-2}(x, \mathbf{z})\, d\mathbf{z}}. \tag{3.36}$$

Now we consider a number of statistics used to calculate $\hat{\theta}_j$ and $\hat{\Theta}_k$. Following (3.19), set $\mathcal{N}_{-j} := \{\{0, 1, \dots, b_n\} \setminus \{j\}\} I(p \in \mathcal{A}) + \{\{0, 1, \dots, \lfloor n^{1/3} \rfloor\} \setminus \{j\}\} I(p \notin \mathcal{A})$ and define the estimate of $f_{-j}(x) := f(x) - \theta_j \varphi_j(x)$ as

$$\tilde{f}_{-j}(x) := m^{-1} \sum_{l \in \mathcal{M}_{12}} \sum_{i \in \mathcal{N}_{-j}} \frac{Y_l \varphi_i(X_l)}{\tilde{p}_4(X_l, \mathbf{Z}_l)} \varphi_i(x). \tag{3.37}$$

Following (3.20), we define the estimate of the additive nuisance component $g(\mathbf{z})$ as

$$\tilde{g}(\mathbf{z}) := m^{-1} \sum_{l \in \mathcal{M}_{13}} \sum_{\mathbf{r} \in \mathcal{N}_g} \frac{Y_l \psi_{\mathbf{r}}(\mathbf{Z}_l)}{\tilde{p}_5(X_l, \mathbf{Z}_l)} \psi_{\mathbf{r}}(\mathbf{z}), \tag{3.38}$$

where

$$\mathcal{N}_g := \{\{0, 1, \dots, \lfloor n^{1/D}/b_n^{2D} \rfloor\}^D \setminus \{0\}^D\} I(p \in \mathcal{A}) + \{\{0, 1, \dots, \lfloor n^{1/(3D)} \rfloor\}^D \setminus \{0\}^D\} I(p \notin \mathcal{A}).$$

Now we are following (3.26)–(3.30) and estimates $\sigma_{b_n}^{-2}(x, \mathbf{z})$ and $\mathcal{I}_{b_n}(x)$. Write

$$\tilde{q}(x, \mathbf{z}) := \max\Bigg(-b_n, \min\Bigg(b_n, m^{-1} \sum_{l \in \mathcal{M}_{14}} \sum_{\|(i,\mathbf{r})\|_\infty < b_n} \frac{Y_l \varphi_i(X_l) \psi_{\mathbf{r}}(\mathbf{Z}_l)}{\tilde{p}_6(X_l, \mathbf{Z}_l)} \times \varphi_i(x) \psi_{\mathbf{r}}(\mathbf{z})\Bigg)\Bigg) \tag{3.39}$$

for the estimate of $q(x,\mathbf{z}) := f(x) + g(\mathbf{z})$. This allows us to estimate Fourier coefficients $\sigma_{i\mathbf{r}}$ of the squared scale function by

$$\tilde{\sigma}_{i\mathbf{r}} := m^{-1} \sum_{l \in \mathcal{M}_{15}} \frac{(Y_l - \tilde{q}(X_l, \mathbf{Z}_l))^2}{\tilde{p}_7(X_l, \mathbf{Z}_l)} \varphi_i(X_l) \psi_{\mathbf{r}}(\mathbf{Z}_l). \tag{3.40}$$

Then, following (3.27), we can define a truncated projection estimate of the squared scale function

$$\tilde{\sigma}^2(x,\mathbf{z}) := \max\Bigg(c_*, \max\Bigg(c^*, \sum_{\|(i,\mathbf{r})\|_\infty < b_n} \tilde{\sigma}_{i\mathbf{r}} \varphi_i(x) \psi_{\mathbf{r}}(\mathbf{z})\Bigg)\Bigg). \tag{3.41}$$

In addition to density estimates (3.32), let us introduce three identical (but based on different subsamples) density estimates. Set $N_p^* := N_p I(p \in \mathcal{A}) + \lfloor n^{1/(2(D+2))} \rfloor I(p \notin \mathcal{A})$, $\mathcal{N}_p^* := \{0, 1, \dots, N_p^*\}^{D+1}$, and for $s = 1, 2, 3$, define

$$\check{p}_s(x,\mathbf{z}) := \max\Bigg(c_n^{-1}, m^{-1} \sum_{l \in \mathcal{M}_{15+s}} \sum_{(i,\mathbf{r}) \in \mathcal{N}_p^*} \varphi_i(X_l) \psi_{\mathbf{r}}(\mathbf{Z}_l) \varphi_i(x) \psi_{\mathbf{r}}(\mathbf{z})\Bigg). \tag{3.42}$$

Note that, with respect to (3.32), the estimate (3.42) is changed only for the case of Sobolev design densities where a larger cutoff (implying a smaller bias) is used; a discussion of why the change is needed and what are the other options can be found in the Appendix (see [11]).

Now we can introduce estimates for $\sigma_{b_n}^{-2}(x,\mathbf{z})$ and $\mathcal{I}_{b_n}(x)$. Following the methodology of (3.29) and (3.30) we set

$$\begin{aligned} &\tilde{\sigma}_{b_n}^{-2}(x,\mathbf{z}) \\ &:= b_n^{-1} \sum_{t=0}^{b_n - 1} \sum_{\|(i,\mathbf{s})\|_\infty \le t} \Bigg[\int_{[0,1]^{D+1}} \tilde{\sigma}^{-2}(u,\mathbf{v}) \varphi_i(u) \psi_{\mathbf{s}}(\mathbf{v})\, du\, d\mathbf{v} \varphi_i(x) \psi_{\mathbf{s}}(\mathbf{z})\Bigg] \\ &=: \sum_{\|(i,\mathbf{s})\|_\infty < b_n} \tilde{\eta}_{i\mathbf{s}} \varphi_i(x) \psi_{\mathbf{s}}(\mathbf{z}) \end{aligned} \tag{3.43}$$

and [note that the estimate (3.42) is used]

$$\begin{aligned} \tilde{\mathcal{I}}_{b_n}(x) &:= \int_{[0,1]^D} \check{p}_1(x,\mathbf{z}) \tilde{\sigma}_{b_n}^{-2}(x,\mathbf{z})\, d\mathbf{z} \\ &= \sum_{t=0}^{N_p^*} \sum_{\|(i,\mathbf{s})\|_\infty < b_n} \check{\pi}_{t\mathbf{s}} \tilde{\eta}_{i\mathbf{s}} \varphi_t(x) \varphi_i(x), \end{aligned} \tag{3.44}$$

where $\check{\pi}_{t\mathbf{s}} := \int_{[0,1]^{D+1}} \check{p}_1(x,\mathbf{z}) \varphi_t(x) \psi_{\mathbf{s}}(\mathbf{z})\, dx\, d\mathbf{z}$ are Fourier coefficients of the density estimate.

Only for the case of a Sobolev design density do we need to calculate statistics

$$\hat{q}_{-j,s}(x) := m^{-1} \sum_{l \in \mathcal{M}_{18+s}} \Bigg[\sum_{i \in \mathcal{N}_{-j}} \frac{Y_l \varphi_i(X_l)}{\tilde{p}_{7+s}(X_l, \mathbf{Z}_l)} \varphi_i(x) + \sum_{\mathbf{r} \in \mathcal{N}_g} \frac{Y_l \psi_{\mathbf{r}}(\mathbf{Z}_l)}{\tilde{p}_{7+s}(X_l, \mathbf{Z}_l)} \psi_{\mathbf{r}}(\mathbf{z}) \Bigg], \qquad s = 1, 2. \tag{3.45}$$

Here $\mathcal{N}_{-j}$ and $\mathcal{N}_g$ are defined above line (3.37) and below line (3.38), respectively.

This finishes all preliminary calculations. Now we can define a new estimator for Sobolev functionals,

$$\hat{\Theta}_k := \frac{2}{m(m-1)} \times \sum_{l_1, l_2 \in \mathcal{M}_{21}, l_1 < l_2} L_k^{-1} \sum_{j \in B_k} \frac{[Y_{l_1} - I(p \notin \mathcal{A}) \hat{q}_{-j,1}(X_{l_1}, \mathbf{Z}_{l_1})] \varphi_j(X_{l_1})}{\check{p}_2(X_{l_1}, \mathbf{Z}_{l_1})} \times \frac{[Y_{l_2} - I(p \notin \mathcal{A}) \hat{q}_{-j,2}(X_{l_2}, \mathbf{Z}_{l_2})] \varphi_j(X_{l_2})}{\check{p}_3(X_{l_2}, \mathbf{Z}_{l_2})} \tag{3.46}$$

and, mimicking dealer-estimate (3.31) of Fourier coefficients θ_j, define

$$\hat{\theta}_j := (n - 21m)^{-1} \sum_{l=21m+1}^{n} \frac{[Y_l - \tilde{f}_{-j}(X_l) - \tilde{g}(\mathbf{Z}_l)] \tilde{\sigma}_{b_n}^{-2}(X_l, \mathbf{Z}_l) \varphi_j(X_l)}{\tilde{\mathcal{I}}_{b_n}(X_l)}. \tag{3.47}$$

Here $\tilde{f}_{-j}$, $\tilde{g}$, $\tilde{\sigma}_{b_n}^{-2}$ and $\tilde{\mathcal{I}}_{b_n}$ are defined in (3.37), (3.38), (3.43) and (3.44), respectively.

PROPOSITION 3.4. *Consider settings* (4) *and* (5) *of Theorem* 2.4. *Assume that* $I(p \in \mathcal{A}) = 1$ *and* $I(p \notin \mathcal{A}) = 1$ *indicate that settings* (4) *and* (5) *are considered, respectively. Then estimator* (3.9), *with* $\hat{\Theta}_k$ *defined in* (3.46), $\hat{\theta}_j$ *defined in* (3.47) *and* $\hat{d} = \tilde{d}$ *defined in* (3.36), *is adaptive and sharp minimax, that is, its MISE satisfies* (3.10).

REMARK 3.2. In what follows the proposed data-driven estimator, calculated without splitting data and with $I(p \in \mathcal{A}) = 1$, is referred to as S-estimator.

Propositions 3.1 and 3.4 imply that the pivotal model (1.1) is a fair benchmark for the general additive model (1.6), and this proves the conjecture made in the Introduction. More discussion, notes and remarks can be found in the Appendix (see [11]).

4. Numerical study. We begin with the following Monte Carlo study. The underlying model is (1.6) where $D = 1$, $g(z) = 0$, the joint design density $p(x, z) =$

$I((x,z)\in[0,1]^2)$, the scale function is $\sigma(x,z)=e^{\lambda z/2}$ and the regression error is standard normal and independent of the covariates (X,Z). We use $\lambda\in\{1,2,3\}$ and four sample sizes $n\in\{50,100,200,400\}$. Figure 1 illustrates a particular simulation with $n=100$ and $\lambda=2$.

We are exploring 4 different estimation procedures with the first two being sharp-minimax for model (1.6) and the last two being sharp-minimax for the model (1.1) with $\sigma(x,z)=\sigma(x)$. The first one is D-estimator defined in Remark 3.1. It knows a sample of size n from (X,Z,Y) and all nuisance functions in the underlying model (1.6). This dealer-estimator serves as a benchmark for the data-driven S-estimator defined in Remark 3.2. The third estimator is the E-estimator of [7, 12] and it was discussed in the Introduction. E-estimator ignores the heteroscedasticity but nonetheless for the considered experiment with $g(z)=0$ it is rate-minimax. In what follows an E-estimator based on a sample of size n from (X,Y) will be referred to as the En-estimator. The last estimator is also an E-estimator which is based on a larger sample of size m. Namely, the larger sample includes the sample of size n from (X,Y), available to the three previous estimators, and then we add extra $m-n$ observations from (X,Y). Here m is the rounded up $nd_2/d=n\int_0^1 e^{\lambda z}\,dz\int_0^1 e^{-\lambda z}\,dz$; remember the discussion below line (1.5). We will refer to this estimator as the Em-estimator to stress that it is based on a larger sample of size m. The underlying idea of exploring Em-estimator is as follows. According to the asymptotic theory, D- and S-estimators, based on a sample of a sufficiently large size n, should have the same MISE as Em-estimator which ignores the heteroscedasticity but can use extra $m-n$ observations. We will test this asymptotic conclusion shortly.

Figure 2 shows us a particular simulation, underlying regression (the solid line) and four estimates (explained in the caption) with their ISE. For the data, shown in the left diagram, all three estimates do a very good job under the difficult circumstances, but their ISEs (denoted as ISED, ISES and ISEEn, resp.) reveal that the D-estimate is better than the S-estimate, and the En-estimate lags behind. All three estimates give us a fair visualization of the bell-type and symmetric about 0.5 underlying regression function. Furthermore, it is practically impossible to see a difference between the D- and S-estimates. This highlights the sensitivity of the ISE criterion. The main issue with the En-estimate is in its tails, but they do reflect the underlying pattern of the shown scattergram (remember that En-estimator knows only the XY-scattergram and has no access to observations of Z). The right diagram shows us a scattergram with 38 observations added from (X,Y). The Em-estimate (remember that the same E-estimator is used in the left and right diagrams) yields a much better fit than the En-estimate, and its ISE (denoted as ISEEm) is close to the ISED and ISES.

For each of 12 particular experiments, defined by the scale function and the sample size, we conduct 1000 simulations and then calculate average ISE (AISE) for the four estimates. Table 1 presents ratios $R_1:=\text{AISES}/\text{AISED}$, $R_2:=\text{AISEEn}/\text{AISES}$ and $R_3:=\text{AISEEm}/\text{AISED}$.

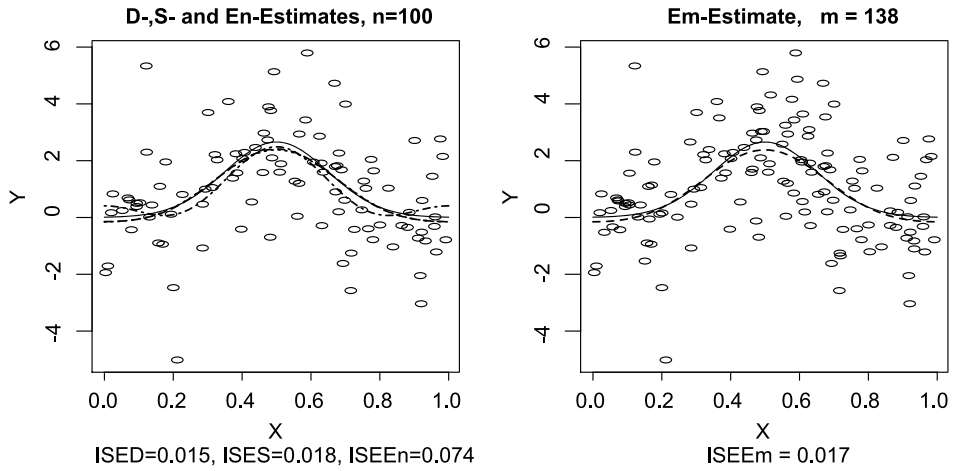

FIG. 2. *Simulated data according to model* (1.6) *with* $f(x)$ *being the Normal* [7], *page* 18, *and shown by the solid line*, $D = 1$, $g(z) = 0$, $\sigma(x, z) = e^z$ *and* $p(x, z) = I((x, z) \in [0, 1]^2)$. *The left scattergram is the same as in the left diagram of Figure* 1, *the right scattergram exhibits the same* 100 *observations plus* 38 *additional ones*, *so the total sample size is* $m = 138$. *All estimators know that the underlying model is* (1.6), *but only the D-estimator knows everything else*, *except for the regression function. The left scattergram is overlaid by the D-estimate*, *S-estimate and En-estimate shown by the dashed*, *dotted and dashed-dotted lines*, *respectively. The dashed line in the right diagram shows the Em-estimate.*

The observed values of ratio $R_1 = \text{AISES}/\text{AISED}$ indicate that, with the exception of the smallest sample size $n = 50$, the proposed data-driven S-estimator does mimic performance of the dealer-estimator. The ratio $R_2 = \text{AISEEn}/\text{AISES}$

TABLE 1
Results of Monte Carlo simulations

		n			
λ		**50**	**100**	**200**	**400**
1	m	54	108	216	432
	R_1, R_2, R_3	1.04, 0.87, 0.90	1.02, 1.08, 1.02	1.04, 1.12, 0.98	1.01, 1.21, 1.13
	R_4, R_5, R_6	1.12, 1.14, 1.15	1.08, 1.09, 1.10	1.07, 1.08, 1.09	1.03, 1.03, 1.04
2	m	69	138	276	552
	R_1, R_2, R_3	1.03, 0.94, 0.79	1.09, 1.21, 1.01	1.06, 1.20, 0.95	1.02, 1.26, 0.96
	R_4, R_5, R_6	1.09, 1, 11, 1, 14	1.14, 1.15, 1.17	1.09, 1.10, 1.12	1.04, 1.05, 1.05
3	m	100	201	403	806
	R_1, R_2, R_3	1.61, 1.03, 0.68	1.11, 1.24, 0.85	1.09, 1.63, 0.96	1.06, 1.51, 0.91
	R_4, R_5, R_6	1.78, 1.85, 1.92	1.18, 1.21, 1.24	1.12, 1.14, 1.15	1.09, 1.10, 1.11

shows that even for the scale function with a moderate heteroscedasticity ($\lambda = 1$) it may be useful to take into account the scale in regression estimation. Furthermore, the observed values of R_2 indicate that a correct usage of the scale in regression estimation becomes paramount for regressions with pronounced heteroscedasticity. Now let us look at the ratio $R_3 = \text{AISEEm}/\text{AISED}$. The asymptotic theory asserts that the Em-estimator, based on m observations, should have the same MISE as the D-estimator based on n observations (remember Figure 2). As we see, results of the numerical study indicate that the asymptotic theory sheds light on performance of the estimators for small samples. Furthermore, please look at the sample sizes m that make the MISE of Em-estimator equal to the dealer's MISE. Even for the case $\lambda = 1$ we need the 8 percent increase, and the required sample size doubles for $\lambda = 3$.

Now let us repeat simulations three more times using nuisance additive components $g_1(z) = z - 1/2$, $g_2(z) = z^2 - 1/3$ and $g_3(z) = z + z^3 - 3/4$ in place of $g(z) = 0$. We are interested in the effect of a nuisance component on estimation of f, which can be evaluated via comparison of performances of the data-driven S-estimator and the D-estimator which knows an underlying nuisance component $g_s(z)$. Results are shown in Table 1 via $R_{3+s} := \text{AISES}_s/\text{AISED}$, $s = 1, 2, 3$, where AISES_s is calculated for the case of sth nuisance component. Note that now R_1 serves as a benchmark for R_{3+s}, and we may conclude that S-estimator does a good job in adapting to the presence of a nuisance component.

Overall, the presented numerical results indicate that: (a) Similarly to [7, 12, 28, 29], the asymptotic theory, which takes into account constants, does shed light on small samples; (b) It is worthwhile to use the scale in regression estimation whenever the scale may depend on auxiliary variables.

Conclusion: It is well known that in a nonparametric heteroscedastic regression the scale function affects the MISE. At the same time, less is known about optimal use of (or even necessity to use) the scale function in regression estimation. The pivotal setting, studied in the paper, is a heteroscedastic regression (1.1) with a univariate regression function, a multivariate scale and a normal regression error which is independent of the covariates. For this setting a sharp-minimax theory of data-driven and adaptive estimation is developed. The outcome is interesting because, depending on the scale function, the scale may or may not be recommended for use by a sharp-minimax regression estimator. Namely, if the scale does not depend on the auxiliary variable, then a sharp-minimax regression estimation does not require knowing, using or estimation of the scale, but otherwise using the scale yields a sharp-minimax MISE. Several extensions of the pivotal model are also considered: (i) The general additive model (1.6) for which model (1.1) can be considered as a benchmark. It is shown that the benchmark is fair meaning that an estimator attains the same minimax MISE for the two models. Special attention is devoted to assumptions on the nuisance functions. In particular, it is shown that no smoothness of the scale is required for the sharp-minimax regression estimation. This is an important conclusion in light of the known minimax result

about the effect of the smoothness of a regression function on the scale estimation. Furthermore, the result holds under a mild assumption on regularity of the multivariate additive component; (ii) The regression error may not necessarily be normal; it suffices that it has only four moments, and it may depend on the covariates; (iii) Response may be discrete with particular examples being classical Bernoulli and Poisson regressions. A numerical study indicates that the developed sharp-minimax asymptotic theory sheds light on performance of estimators for small samples.

Acknowledgments. The author is grateful for the helpful and constructive comments of the Editors, Tony Cai and Runze Li, an Associate Editor and two referees.

SUPPLEMENTARY MATERIAL

Appendix: Notes and proofs (DOI: 10.1214/13-AOS1126SUPP; .pdf). Appendix contains proofs and notes.

REFERENCES

[1] BARY, N. K. (1964). *A Treatise on Trigonometric Series*. Pergament Press, Oxford.
[2] BERENS, H. and XU, Y. (1996). Fejér means for multivariate Fourier series. *Math. Z.* **221** 449–465. MR1381592
[3] CAI, T. T., LEVINE, M. and WANG, L. (2009). Variance function estimation in multivariate nonparametric regression with fixed design. *J. Multivariate Anal.* **100** 126–136. MR2460482
[4] EFROMOVICH, S. (1985). Nonparametric estimation of a density with unknown smoothness. *Theory Probab. Appl.* **30** 557–568.
[5] EFROMOVICH, S. (1989). On sequential nonparametric estimation of a density. *Theory Probab. Appl.* **34** 228–239.
[6] EFROMOVICH, S. (1996). On nonparametric regression for IID observations in a general setting. *Ann. Statist.* **24** 1125–1144. MR1401841
[7] EFROMOVICH, S. (1999). *Nonparametric Curve Estimation*: *Methods*, *Theory*, *and Applications*. Springer, New York. MR1705298
[8] EFROMOVICH, S. (2000). On sharp adaptive estimation of multivariate curves. *Math. Methods Statist.* **9** 117–139. MR1780750
[9] EFROMOVICH, S. (2007). Optimal nonparametric estimation of the density of regression errors with finite support. *Ann. Inst. Statist. Math.* **59** 617–654. MR2397734
[10] EFROMOVICH, S. (2011). Nonparametric estimation of the anisotropic probability density of mixed variables. *J. Multivariate Anal.* **102** 468–481. MR2755009
[11] EFROMOVICH, S. (2013). Supplement to "Nonparametric regression with the scale depending on auxiliary variable." DOI:10.1214/13-AOS1126SUPP.
[12] EFROMOVICH, S. and PINSKER, M. (1996). Sharp-optimal and adaptive estimation for heteroscedastic nonparametric regression. *Statist. Sinica* **6** 925–942. MR1422411
[13] EUBANK, R. L. (1999). *Nonparametric Regression and Spline Smoothing*, 2nd ed. *Statistics*: *Textbooks and Monographs* **157**. Dekker, New York. MR1680784
[14] FAN, J. (1992). Design-adaptive nonparametric regression. *J. Amer. Statist. Assoc.* **87** 998–1004. MR1209561

[15] FAN, J. and GIJBELS, I. (1996). *Local Polynomial Modelling and Its Applications*. *Monographs on Statistics and Applied Probability* **66**. Chapman & Hall, London. MR1383587
[16] FAN, J. and YAO, Q. (2003). *Nonlinear Time Series*: *Nonparametric and Parametric Methods*. Springer, New York. MR1964455
[17] GOLUBEV, G. K. (1991). LAN in problems of estimation of functions and lower bounds for quadratic risks. *Theory Probab. Appl*. **36** 152–157.
[18] GOLUBEV, G. K. (1992). Asymptotically minimax estimation of a regression function in an additive model. *Probl. Inf. Transm*. **28** 101–112.
[19] HASTIE, T. J. and TIBSHIRANI, R. J. (1990). *Generalized Additive Models*. *Monographs on Statistics and Applied Probability* **43**. Chapman & Hall, London. MR1082147
[20] HOFFMANN, M. and LEPSKI, O. (2002). Random rates in anisotropic regression. *Ann. Statist*. **30** 325–396. MR1902892
[21] HOROWITZ, J., KLEMELÄ, J. and MAMMEN, E. (2006). Optimal estimation in additive regression models. *Bernoulli* **12** 271–298. MR2218556
[22] HOROWITZ, J. L. (2009). *Semiparametric and Nonparametric Methods in Econometrics*. Springer, New York. MR2535631
[23] HOROWITZ, J. L. and MAMMEN, E. (2004). Nonparametric estimation of an additive model with a link function. *Ann. Statist*. **32** 2412–2443. MR2153990
[24] HUANG, J., HOROWITZ, J. L. and WEI, F. (2010). Variable selection in nonparametric additive models. *Ann. Statist*. **38** 2282–2313. MR2676890
[25] JIANG, J., FAN, Y. and FAN, J. (2010). Estimation in additive models with highly or nonhighly correlated covariates. *Ann. Statist*. **38** 1403–1432. MR2662347
[26] KAHANE, J.-P. (1985). *Some Random Series of Functions*, 2nd ed. *Cambridge Studies in Advanced Mathematics* **5**. Cambridge Univ. Press, Cambridge. MR0833073
[27] LEMAN, È. (1991). *Theory of Point Estimation*. Wadsworth, New York.
[28] MARRON, J. S., ADAK, S., JOHNSTONE, I., NEUMANN, N. and PATIL, P. (1998). Exact risk analysis of wavelet regression. *J. Comput. Graph. Statist*. **7** 278–309.
[29] MARRON, J. S. and WAND, M. P. (1992). Exact mean integrated squared error. *Ann. Statist*. **20** 712–736. MR1165589
[30] NIKOL'SKIĬ, S. M. (1975). *Approximation of Functions of Several Variables and Imbedding Theorems*. Springer, New York. MR0374877
[31] PINSKER, M. S. (1980). Optimal filtering a square integrable signal in Gaussian white noise. *Probl. Inf. Transm*. **16** 52–68.
[32] SERFLING, R. J. (1980). *Approximation Theorems of Mathematical Statistics*. Wiley, New York. MR0595165
[33] SHAPIRO, V. L. (1964). Fourier series in several variables. *Bull. Amer. Math. Soc*. (*N.S.*) **70** 48–93. MR0158222
[34] STONE, C. J. (1985). Additive regression and other nonparametric models. *Ann. Statist*. **13** 689–705. MR0790566
[35] WASSERMAN, L. (2006). *All of Nonparametric Statistics*. Springer, New York. MR2172729
[36] ZHANG, S. and WONG, M. (2003). Wavelet threshold estimation for additive models. *Ann. Statist*. **31** 152–173.
[37] ZIGMUND, A. (1968). *Trigonometric Series*, 2nd ed. Cambridge Univ. Press, Cambridge.

DEPARTMENT OF MATHEMATICAL SCIENCES
UNIVERSITY OF TEXAS AT DALLAS
RICHARDSON, TEXAS 7580
USA
E-MAIL: efrom@utdallas.edu