Universal Bias Reduction in Estimation of Smooth Additive Function in High Dimensions

Fan Zhou ZHOUFAN066@GMAIL.COM
Cognitive Computing Lab, Baidu Research, 10900 NE 8th St. Bellevue, WA 98004, USA

Ping Li LIPING98@GMAIL.COM
Cognitive Computing Lab, Baidu Research, 10900 NE 8th St. Bellevue, WA 98004, USA

Cun-Hui Zhang CZHANG@STAT.RUTGERS.EDU
Department of Statistics, Rutgers University. The work of Cun-Hui Zhang was conducted as a consulting researcher at Baidu Research – 10900 NE 8th St. Bellevue, WA 98004, USA.

Editors: Shipra Agrawal and Pranjal Awasthi

Abstract

Suppose we observe $\mathbf{x}_j = \boldsymbol{\theta} + \boldsymbol{\varepsilon}_j$, $j = 1, ..., n$ where $\boldsymbol{\theta} \in \mathbb{R}^d$ is an unknown parameter and $\boldsymbol{\varepsilon}_j$ are i.i.d. random noise vectors satisfying some general distribution. We study the estimation of $f(\boldsymbol{\theta}) := \sum_{i=1}^d f_i(\theta_i)$ when $f : \mathbb{R}^d \to \mathbb{R}$ is a given smooth additive function and d is large. Inspired by a recent work on studying the estimation of $f(\boldsymbol{\theta})$ under Gaussian shift model via a Fourier analytical approach, we propose a new estimator that can be implemented easily and computed fast. We show that the new estimator achieves effective bias reduction universally under minimum moment constraint. Further, we establish its asymptotic normality which implies the new estimator is asymptotically efficient. When f_i is sufficiently smooth and d is large, such properties make the new estimator rate optimal. Efficient computation of the new estimator and the minimum requirement of noise make this work more applicable to real world applications.

Keywords: bias reduction, universality, high dimensional estimation, additive model, asymptotic normality

1. Introduction

We observe

$$\mathbf{x}_j = \boldsymbol{\theta} + \boldsymbol{\varepsilon}_j, \tag{1.1}$$

with $\mathbf{x}_j$, $j = 1, ..., n$ being noisy observations of an unknown parameter $\boldsymbol{\theta} \in \mathbb{R}^d$, and $\boldsymbol{\varepsilon}_j \in \mathbb{R}^d$ being i.i.d. copies of a random vector $\boldsymbol{\varepsilon}$ that satisfies some general distribution. We study the estimation of the function value $f(\boldsymbol{\theta})$ when $f : \mathbb{R}^d \to \mathbb{R}$ is a given smooth function with an additive structure:

$$f(\boldsymbol{\theta}) := \sum_{i=1}^d f_i(\theta_i). \tag{1.2}$$

The major motivation of this work is the following: when the dimension parameter d is a fixed constant or relatively small compared with the sample size n, one can use the naive plug-in estimator $f(\bar{\mathbf{x}})$ with the sample mean $\bar{\mathbf{x}}$ to serve the purpose. Because the overall bias of $f(\bar{\mathbf{x}})$ can be bounded by the sum of bias of estimation for each $f_i(\theta_i)$, which is roughly of the same order due to d being treated as a constant. However, if d is large and comparable to n, say $d = n^\alpha$ for some $\alpha \in (0, 1)$, $f(\bar{\mathbf{x}})$ is no longer a good choice since its large bias makes it sub-optimal (Koltchinskii and Zhilova,

2021a; Zhou and Li, 2021; Zhou et al., 2021) even for smooth f. As a consequence, to develop an effective estimator with a high dimensional parameter $\boldsymbol{\theta}$ posed a challenge. Recent works are mostly focused on specific distributions of noise. Though some are claimed to able to apply to other noise scenarios, generalization of their theoretical guarantee which usually heavily relies on the specific distribution's properties is equally hard. So a universal method with theoretical guarantee is still missing. In this article, we make an attempt to fill this gap.

Model (1.1) is the so-called measurement error model (1.1) (Carroll et al., 2006) and it models the scenario when the underlying parameter $\boldsymbol{\theta}$ always comes with noise either by unintentional machine measurement error or by intentional add-on noise due to privacy concern. As it's ubiquitous in real world applications, early studies on this topic can be dated back to Levit (1976, 1978); Ibragimov et al. (1986); Bickel and Ritov (1988); Nemirovskii (1991); Birgé and Massart (1995); Laurent (1996); Lepski et al. (1999); Nemirovskii (2000). Other works related to this model focused on the estimation of f itself when f is not given, we refer to Fan and Truong (1993); Carroll et al. (2006); Han and Park (2018) and the references therein. On the other hand, functionals with an additive structure as in (1.2) are perhaps the most important and widely used ones in statistical learning such as boosting methods (Friedman, 2001) or generalized additive models (Hastie, 2017). Unlike linear regression, in those applications the regression function, i.e. the conditional mean is expressed as nonlinear summation of $\boldsymbol{\theta}$'s variables:

$$\mathbb{E}[y|\boldsymbol{\theta}] = \sum_{i=1}^{d} f_i(\theta_i). \tag{1.3}$$

Other important examples of additive functions include: the loss function of general machine learning problems; $\|\boldsymbol{\theta}\|_p^p$, the ℓ_p-norm of $\boldsymbol{\theta}$; or the entropy of a discrete probability distribution: $f(\boldsymbol{\theta}) = \sum_{i=1}^{d} -\theta_i \log \theta_i$. Historically, two types of functional estimation with an additive structure are extensively studied: one is the linear functional, i.e. $f(\boldsymbol{\theta}) = \sum_{i=1}^{d} \theta_i$, see Donoho and Liu (1987, 1991); Klemelä and Tsybakov (2001); Cai and Low (2005a) and the references therein. The other is the quadratic functional, i.e. $f(\boldsymbol{\theta}) = \sum_{i=1}^{d} \theta_i^2$, see Donoho and Nussbaum (1990); Cai and Low (2005b); Klemelä (2006); Laurent and Massart (2000) and the references therein. Recently we see a resurgence of interests in studying minimax theory of those topics under sparsity class in Gaussian shift model, see Collier et al. (2017); Collier and Comminges (2019). Another line of exciting works focus on minimax estimation of non-smooth additive models, see Cai and Low (2011); Jiao et al. (2015); Wu and Yang (2016, 2019); Carpentier and Verzelen (2019); Collier et al. (2020)

Related works: Several methods have been developed recently to study the problem in high dimensions. One is based on an iterative bootstrap technique with a more statistical flavor see Jiao and Han (2020); Koltchinskii (2020); Koltchinskii and Zhilova (2021a,b). Jiao and Han (2020) used this method to study the problem in one dimensional case under binomial model. While Koltchinskii (2020); Koltchinskii and Zhilova (2021a,b) developed this method into its general form and studied general smooth functional estimation in Gaussian shift model. Zhou and Li (2021) applied this iterative bootstrap method to study the Gaussian shift model when f is an additive function with Hölder smoothness. The key idea is to use bootstrap chain and several rounds of re-sampling to reduce bias in an iterative way.

Another type of methods is rooted in approximation theory which seeks to replace f by its approximation to achieve de-biasing. For instance, Jiao and Han (2020) used Taylor expansion of $f(\bar{\mathbf{x}})$

at $\boldsymbol{\theta}$ to approximate $f(\boldsymbol{\theta})$. Collier and Comminges (2019) used Hermite polynomial approximations of f and studied the estimation of general additive functional under Gaussian shift model. Zhou and Li (2019) used approximations from Fourier analysis and Littlewood-Paley theory, which we will explain in detail in Section 3. In other interesting related works such as Acharya et al. (2017); Hao and Orlitsky (2019), the authors tried to plug in different MLE of $\boldsymbol{\theta}$ such as profile maximum likelihood estimator other than the sample mean $\bar{\mathbf{x}}$. In Hao and Li (2020a,b), the authors adopted Bessel smoothing to improve the estimations of a collection of summary statistics, in particular for estimating the number of unseen species. In Zhou et al. (2021), the authors developed a Taylor expansion type method and studied the problem when f is a general functional that can be smooth or non-smooth. In Koltchinskii (2022), the author developed a jackknife type bias reduction method and studied a similar problem when f is Hölder smooth and $\boldsymbol{\theta}$ is the covariance operator of Gaussian random variables in Banach space.

Major contributions: The works mentioned above are either based on model (1.1) under a specific distribution and/or requires to know the distribution of the noise in advance which is quite unrealistic, or are computationally intensive when it comes to implementation as re-sampling is needed. In this article, we solve this by proposing a new estimator and relaxing the distribution requirement of noise to its most general form only assuming finite second moment. We use a Fourier analytical approach originated from a brilliant idea in the seminal work of Kolmogorov (1950) on unbiased estimation and a recent work (Zhou and Li, 2019) which studied the estimation of $f(\boldsymbol{\theta})$ with a general smooth function f under Gaussian shift model. Nevertheless, both estimators in Zhou and Li (2019, 2021) rely heavily on the Gaussian assumption in terms of estimator construction and theoretical analysis, and can hardly be generalized to a distribution free setting. As a major contribution, we introduce a new estimator and close the theory gap by developing some new analysis tools.

Roadmap: In Section 3, we explain the ideas and intuitions of the construction of our estimator, which has a simple expression and is faster to compute compared with other approaches such as iterative bootstrap. In Section 4, we establish an upper bound on bias of the new estimator. Our analysis is distribution free and only requires finite second moment of the noise. The bound indicates the new estimator achieves an effective bias reduction and is quite different from that of general smooth function f without an additive structure. In Section 5, we show the estimator scaled by the inverse of the Fisher information for estimation of $f(\boldsymbol{\theta})$ is normally distributed around the ground truth $f(\boldsymbol{\theta})$, indicating the estimator has optimal asymptotic variance implied by Cramér-Rao bound and is asymptotically efficient. Thanks to f's additive structure, the asymptotic normality can be achieved under much looser constraint on smoothness compared with the general smooth function case. In Section 6, simulations are presented to validate our theory. The new estimator is not only effective in bias reduction but also maintains near optimal variance, which eventually leads to its optimal performance on MSE for large d. To summarize, the simple explicit expression of the new estimator and the minimum requirement of noise setting implies universality of the proposed estimator.

2. Preliminaries

Notation: Boldface uppercase letter $\mathbf{X} \in \mathbb{R}^{d\times n}$ denotes the data matrix, and boldface lowercase letter $\mathbf{x}$ denotes a vector. We use $\|\cdot\|$ to denote the ℓ_2-norm of a vector. We use $\mathcal{F}$ and $\mathcal{F}^{-1}$ to denote the Fourier transform (FT) and inverse Fourier transform (IFT) respectively. The conventional notation "$\Rightarrow$" denotes weak convergence, i.e. convergence in distribution and use "$\xrightarrow{p}$" to denote

convergence in probability. We use $\mathcal{S}'$ to denote the set of all complex-valued tempered distributions on $\mathbb{R}$ and $L^p(\mathbb{R})$ to denote the L^p spaces. Given nonnegative numbers a and b, $a \lesssim b$ means $a \leq Cb$ for a numerical constant C, and $a \asymp b$ means $a \lesssim b$ and $b \lesssim a$. $a \wedge b = \min\{a, b\}$ and $a \vee b = \max\{a, b\}$.

2.1. Minimum assumptions on noise

Assumption 1 *Let $\boldsymbol{\varepsilon}$ be a random vector in $\mathbb{R}^d$, assume that $\mathbb{E}[\boldsymbol{\varepsilon}] = 0$.*

Assumption 2 *Let $\boldsymbol{\varepsilon} = (\epsilon_1, \ldots, \epsilon_d)^T$ be a random vector in $\mathbb{R}^d$, assume that the variance of each component ϵ_i is uniformly bounded by some constant σ^2 independent of d and n, i.e. $\mathbb{E}[\epsilon_i^2] \leq \sigma^2$, $\forall$ $i = 1, ...d$.*

2.2. A Besov-type norm and the function class

Before we get into the function class of our interest, we introduce a Besov-type norm to characterize smoothness of f. Given $\psi : \mathbb{R} \to \mathbb{R}$, we define

$$\|\psi\|_{s,\infty,1} := \frac{1}{\sqrt{2\pi}} \int_{\mathbb{R}} \big|\mathcal{F}\psi(\zeta)\big|\big(1 \vee |\zeta|^s\big) d\zeta, \tag{2.1}$$

and we are interested in functions that residing in the following function class

$$\mathcal{F}^s(M) := \big\{\psi \in \mathcal{S}' : \|\psi\|_{s,\infty,1} < M\big\}. \tag{2.2}$$

The norm defined in (2.1) is similar as the definition of Sobolev norm or Bessel-Potential norm, see Chap. 2.2 Triebel (1983). Clearly, the parameter s characterizes the smoothness of ψ as it controls how fast $\mathcal{F}\psi$ would decay when ζ approaches infinity. Note that the smoothness index s can be related to classical Sobolev smoothness or Hölder smoothness through embedding theorems Sec. 2.5.7 Triebel (1983). Except for these two classical classes of functions, analytical and entire functions also satisfy (2.2). Another example in (2.2) is the mixture model, one can check that with some distribution function $G(x)$ and absolute constant C:

$$\psi := \int e^{-(\theta - x)^2/2} dG(x), \quad \|\psi\|_{s,\infty,1} \leq C. \tag{2.3}$$

We assume that f is equipped with a homogeneous additive structure, i.e.

$$f = \sum_{i=1}^{d} f_i, \quad f_i \in \mathcal{F}^s(M). \tag{2.4}$$

Especially, we use $\mathcal{F}_d^s(M) := \mathcal{F}^s(M) \bigoplus \cdots \bigoplus \mathcal{F}^s(M)$ to denote the function class. Then an example of (1.2) of mixture model with coefficients π_i and candidate distributions $G_i(x)$ follows as

$$f(\boldsymbol{\theta}) = \sum_{i=1}^{d} \pi_i \int e^{-(\theta_i - x)^2/2} dG_i(x). \tag{2.5}$$

3. Construction of the estimator

As we have mentioned, the plug-in estimator's large bias can make it sub-optimal for both smooth (Koltchinskii and Zhilova, 2021a). The main purpose of our new estimator is to effectively reduce the bias and make it controllable. Our idea is based on a Fourier analytical approach developed by Kolmogorov (1950); Zhou and Li (2019) to study the Gaussian case, and exploit the smoothness property of f to reduce the bias of estimation of each component $f_i(\theta_i)$ to the order $O(n^{-(s\wedge 2)/2})$.

In a recent work (Zhou and Li, 2019), the authors proposed an estimator under Gaussian noise as follows

$$g(\bar{\mathbf{x}}) := \frac{1}{(2\pi)^{d/2}} \int_{\Omega} \mathcal{F}f(\boldsymbol{\zeta}) e^{\langle \boldsymbol{\Sigma}\boldsymbol{\zeta},\boldsymbol{\zeta}\rangle/2n} e^{i\boldsymbol{\zeta}^T \bar{\mathbf{x}}} d\boldsymbol{\zeta}. \tag{3.1}$$

where $\Omega := \{\boldsymbol{\zeta} : \|\boldsymbol{\zeta}\| \le R\}$

denotes the truncated regime of the support of $\mathcal{F}f$. The key point of this estimator is that the factor $e^{\langle \boldsymbol{\Sigma}\boldsymbol{\zeta},\boldsymbol{\zeta}\rangle/2n}$ in (3.1) exactly cancels the characteristic function $\mathbb{E}\big[e^{i\boldsymbol{\zeta}^T\bar{\boldsymbol{\varepsilon}}}\big]$ of the noise which makes $g(\bar{\mathbf{x}})$ an unbiased estimator of the analytical part of f, i.e.

$$\mathbb{E}_{\boldsymbol{\theta}}\big[g(\bar{\mathbf{x}})\big] = f^N(\boldsymbol{\theta}); \quad f^N(\boldsymbol{\theta}) := \frac{1}{(2\pi)^{d/2}} \int_{\Omega} \mathcal{F}f(\boldsymbol{\zeta}) \cdot e^{i\boldsymbol{\zeta}^T\boldsymbol{\theta}} d\boldsymbol{\zeta}. \tag{3.2}$$

On the other hand, they showed the remainder $\tilde{f}^N$ of the decomposition $f := f^N + \tilde{f}^N$

$$\tilde{f}^N(\boldsymbol{\theta}) := \frac{1}{(2\pi)^{d/2}} \int_{\mathbb{R}^d\setminus\Omega} \mathcal{F}f(\boldsymbol{\zeta}) \cdot e^{i\boldsymbol{\zeta}^T\boldsymbol{\theta}} d\boldsymbol{\zeta} \tag{3.3}$$

is uniformly small when f is sufficiently smooth. So the overall bias of $g(\bar{\mathbf{x}})$ is small. However, when $\boldsymbol{\varepsilon}$ is non-Gaussian, (3.2) no longer holds and its theoretical analysis can hardly be generalized. Inspired by this approach, we construct a new estimator with modifications to realize the idea above. Firstly, we introduce some critical ingredients. We consider the one-dimensional case where $x_j = \theta + \epsilon_j$, $j = 1, \ldots, n$, and ϵ_j's are i.i.d. copies of a random variable ϵ. Given $\zeta \in \mathbb{R}$, we denote by

$$g_{\zeta,n}(x_j, \bar{x}) := \frac{(\zeta x_j - \zeta\bar{x})^2}{2n^2}. \tag{3.4}$$

Then we define the following operator on a given function $f_1 : \mathbb{R} \to \mathbb{R}$:

$$\mathcal{T}(f_1) := \frac{1}{\sqrt{2\pi}} \int_{\Omega} \mathcal{F}f_1(\zeta) \Big(1 + \sum_{j=1}^{n} g_{\zeta,n}(x_j, \bar{x})\Big) e^{i\zeta\bar{x}} d\zeta, \tag{3.5}$$

where $\Omega := \{\zeta \in \mathbb{R} : |\zeta| \le R\}$ is a truncated region of the support of $\mathcal{F}f_1$. As a result, we apply $\mathcal{T}$ to each component function f_i with noisy observations of each coordinate θ_i and introduce our new estimator as

$$g^*(\mathbf{X}) := \sum_{i=1}^{d} \mathcal{T}(f_i). \tag{3.6}$$

Here $\mathbf{X} \in \mathbb{R}^{d\times n}$ denotes the observation data matrix under model (1.1) with $\mathbf{x}_j$ being its j-th column. $\mathcal{T}$ is applied to f_i using the i-th row of $\mathbf{X}$.

To understand the effectiveness of $g^*(\mathbf{X})$ on bias reduction on a single component f_i, we compare it with its plug-in counterpart. Recall that the plug-in estimate of f_1 can be written as

$$f_1(\bar{x}) = \frac{1}{\sqrt{2\pi}} \int \mathcal{F}f_1(\zeta) \cdot e^{i\zeta\bar{x}} d\zeta = \frac{1}{\sqrt{2\pi}} \int \mathcal{F}f_1(\zeta) \cdot e^{i\zeta(\bar{\epsilon}+\theta_1)} d\zeta. \tag{3.7}$$

Take a closer look at its bias through the expansion of $e^{i\zeta\bar{\epsilon}}$, one can check that the bias $\mathbb{E}[f_1(\bar{x})] - f_1(\theta_1)$ can be written as

$$\frac{1}{\sqrt{2\pi}} \int_\Omega \mathcal{F}f_1(\zeta) \cdot e^{i\zeta\theta_1} \cdot \mathbb{E}\Big[-\frac{(\zeta\bar{\epsilon})^2}{2} \Big] d\zeta + \mathcal{R}_1. \tag{3.8}$$

The remainder $\mathcal{R}_1$ shall be of a smaller order when f_1 is sufficiently smooth. On the other hand, as for the bias of $\mathcal{T}(f_1)$, one can check that $\mathbb{E}\big[\mathcal{T}(f_1)\big] - f_1(\theta_1)$ can be written as

$$\frac{1}{\sqrt{2\pi}} \int_\Omega \mathcal{F}f_1(\zeta) \cdot e^{i\zeta\theta_1} \cdot \mathbb{E}\Big[-\frac{(\zeta\bar{\epsilon})^2}{2n} \Big] d\zeta + \mathcal{R}_2, \tag{3.9}$$

where again $\mathcal{R}_2$ is the remainder and shall be of a smaller order. By comparing (3.8) and (3.9), roughly speaking, $\mathcal{T}(f_1)$ reduces the bias of leading term of $f_1^N(\bar{x})$ by a scaling factor n^{-1} if everything else is well controlled.

Remark 1 *Another interpretation of our estimator is that we are using a data-driven approach to eliminate the effect caused by the first term appeared in* (3.8)*. To be more specific, the expectation appeared in the first term of* (3.8) *is exactly the variance of $\bar{\epsilon}$ multiplied by a scaling factor. What we do in* (3.5)*, i.e. $\sum_{j=1}^n g_{\zeta,n}(x_j, \bar{x})$, is to use the data to estimate the expectation term in* (3.8) *and cancel it by adding it to the integral. The estimator* (3.6) *corrects the bias only up to the second order of the expansion of $e^{i\zeta\bar{\epsilon}}$ due to Assumption 2. Ideally, higher orders of bias correction can be achieved if higher moments of $\boldsymbol{\varepsilon}$ are assumed to be finite. Such process can be done by carefully enumerating and matching the terms in the expansion of $e^{i\zeta\bar{\epsilon}}$. Similar process are described in detail by Jiao and Han (2020) when the authors used the Taylor series to approximate f to achieve bias-reduction. Note that the recent work by Zhou et al. (2021) developed this Taylor expansion method into its general form, they also reduced the distribution constraint to minimum momentum requirement.*

In next section, we will give a theoretical justification of this observation and prove an upper bound on its bias. As shown in Section 6, simulation results show that our estimator (3.6) has a clear advantage over the plug-in estimator on bias reduction. Meanwhile, the simple expression of (3.5) makes it easy to implement and the use of FT makes it fast to compute which is an advantage over the computationally-intensive bootstrap approach (Koltchinskii and Zhilova, 2021a; Zhou and Li, 2021) which needs to re-sample the noise.

4. Bias reduction

We now provide a theoretical justification of our intuition explained in Section 3 by proving an upper bound on the bias of the proposed estimator (3.6). As we shall see, when we choose the cut-off radius $R \asymp \sqrt{n}$ as defined in the operator $\mathcal{T}$ in (3.5), the upper bound is of the order $O(d \cdot n^{(s\wedge 2)/2})$.

Unlike the previous works (Koltchinskii and Zhilova, 2021a; Zhou and Li, 2019, 2021; Collier et al., 2017) in which the analysis relies heavily on the Gaussian assumption, Theorem 2 is established by only requiring finite second moment of each coordinate of $\boldsymbol{\varepsilon}_j$. This not only introduces new ideas in the analysis but also opens the door for the estimator to a much wider range of applications since Gaussian assumption is often inaccurate and limited in practice.

Theorem 2 *Under model (1.1), suppose that the noise vector $\boldsymbol{\varepsilon}_j$, $j = 1, ..., n$ are i.i.d. copies of a random vector $\boldsymbol{\varepsilon}$ satisfying Assumption 1-2. For any given $f \in \mathcal{F}_d^s(M)$, take $R \asymp \sqrt{n}$ in (3.5). Then for some numerical constant C_1^*, the estimator in (3.6) satisfies*

$$\left|\mathbb{E}_{\boldsymbol{\theta}}\left[g^*(\mathbf{X})\right] - f(\boldsymbol{\theta})\right| \leq C_1^* M\sigma^2\Big(d \cdot n^{-1} \vee d \cdot n^{-s/2}\Big). \tag{4.1}$$

Remark 3 *Bound (4.1) indicates the bias of $g^*(\mathbf{X})$ is of the order $O(d \cdot n^{-(s\wedge 2)/2})$ given σ, M are fixed constants. Firstly, it differs from the bound on the bias $O((d/n)^{s/2})$ for general functional estimation without an additive structure (Koltchinskii and Zhilova, 2021a; Zhou and Li, 2019). The additive structure does help to improve the rate on bias by decoupling the dependence between s and d. Meanwhile, under Gaussian noise the bound on bias achieved by the previous methods (Zhou and Li, 2021) in additive models is $O(d \cdot n^{-s/2})$ for all $s > 1$ uniformly. For the case $s > 2$, the new estimator pays the price for universality by sacrificing some reduction on bias. In fact, one shall achieve higher order bias-reduction if we assume higher finite moments of $\boldsymbol{\varepsilon}$, say the bound would be $O(n^{-(3\wedge s)/2})$ if third moment is finite. This can be done by replacing (3.4) with higher order approximations of $e^{i\zeta\bar{\epsilon}}$ while the analysis is more complicated.*

As shown in Section 6, compared with $f(\bar{\mathbf{x}})$, the new estimator is effective in bias reduction which achieves optimal performance on MSE for large d while $f(\bar{\mathbf{x}})$ fails.

Proof [Proof of Theorem 2] Due to page limit, we outline the key steps of this proof. Detailed proofs of ancillary lemmas are deferred to supplementary material. Given the additive structure of f, we have

$$\left|\mathbb{E}_{\boldsymbol{\theta}}\left[g^*(\mathbf{X})\right] - f(\boldsymbol{\theta})\right| = \Big|\sum_{i=1}^{d}\Big(\mathbb{E}_{\boldsymbol{\theta}}\left[\mathcal{T}(f_i)\right] - f_i(\theta_i)\Big)\Big| \leq \sum_{i=1}^{d}\left|\mathbb{E}_{\theta_i}\left[\mathcal{T}(f_i)\right] - f_i(\theta_i)\right|. \tag{4.2}$$

In the following, we will focus on bounding $\left|\mathbb{E}_{\theta_i}\left[\mathcal{T}(f_i)\right] - f_i(\theta_i)\right|$. When considering a single component f_i, we abuse the notation a little bit. We use $\epsilon_j^{(i)}$ or $\epsilon_j\ (x_j)$, $j = 1, ..., n$ to denote the i.i.d. noise of θ_i and it shall not cause any ambiguity. Given the decomposition $f_i(\theta) = f_i^N(\theta_i) + \tilde{f}_i^N(\theta_i)$, with $f_i^N(\theta_i) := (2\pi)^{-1/2}\int_\Omega \mathcal{F}f_i^N(\zeta)e^{i\zeta\theta_i}d\zeta$, we have

$$\left|\mathbb{E}_{\theta_i}\left[\mathcal{T}(f_i)\right] - f_i(\theta_i)\right| \leq \left|\mathbb{E}_{\theta_i}\left[\mathcal{T}(f_i)\right] - f_i^N(\theta_i)\right| + \left|\tilde{f}_i^N(\theta_i)\right|. \tag{4.3}$$

Based on definition (3.5), we replace f_i^N by $\mathcal{F}^{-1}[\mathcal{F}f_i^N(\zeta)]$ in the first term on the right hand side in (4.3) and get

$$\left|\mathbb{E}_{\theta_i}\left[\mathcal{T}(f_i)\right] - f_i^N(\theta_i)\right| = \Big|\frac{1}{\sqrt{2\pi}}\int_\Omega \mathcal{F}f_i^N(\zeta)\mathbb{E}\Big[\Big(1 + \sum_{j=1}^{n} g_{\zeta,n}(x_j, \bar{x})\Big)e^{i\zeta\bar{\epsilon}} - 1\Big]e^{i\zeta\theta_i}d\zeta\Big| \leq \mathrm{I} + \mathrm{II} + \mathrm{III}$$

with $g_j := g_{\zeta,n}(x_j, \bar{x})$, $g_j^o := g_{\zeta,n}(\epsilon_j, 0)$ and

$$\mathrm{I} := \frac{1}{\sqrt{2\pi}} \int_\Omega \Big|\mathcal{F} f_i^N(\zeta)\Big| \cdot \mathbb{E}\Big|\Big(1 + \sum_{j=1}^n g_j\Big) - \prod_{j=1}^n \Big(1 + g_j^o\Big)\Big| d\zeta$$

$$\mathrm{II} := \frac{1}{\sqrt{2\pi}} \int_\Omega \Big|\mathcal{F} f_i^N(\zeta)\Big| \cdot \mathbb{E}\Big|\prod_{j=1}^n \Big(1 + g_j^o\Big) e^{i\zeta\bar{\epsilon}} - \prod_{j=1}^n \Big(1 + \frac{i\zeta\epsilon_j}{n}\Big)\Big| d\zeta$$

$$\mathrm{III} := \Big|\frac{1}{\sqrt{2\pi}} \int_\Omega \mathcal{F} f_i^N(\zeta) \cdot \mathbb{E}\Big[\prod_{j=1}^n \Big(1 + \frac{i\zeta\epsilon_j}{n}\Big) - 1\Big] e^{i\zeta\theta_i} d\zeta\Big|.$$

Then, we will bound I, II and III respectively. Firstly, we bound the first term I through a comparison between $g_{\zeta,n}(x_j, \bar{x})$ and $g_{\zeta,n}(\epsilon_j, 0)$. Observe that $g_{\zeta,n}(x_j, \bar{x})$ can be written as $g_{\zeta,n}(\epsilon_j, \bar{\epsilon})$:

$$g_{\zeta,n}(x_j, \bar{x}) = \frac{(\zeta x_j - \zeta\bar{x})^2}{2n^2} = \frac{(\zeta\epsilon_j - \zeta\bar{\epsilon})^2}{2n^2} = g_{\zeta,n}(\epsilon_j, \bar{\epsilon}).$$

We use the following lemma which plays a key role in bounding the first term I.

Lemma 4 *Given ϵ_j, $j = 1, ..., n$ be i.i.d. copies of a random variable ϵ satisfying $\mathbb{E}[\epsilon] = 0$ and $\mathbb{E}[\epsilon^2] = \sigma^2$. Then the following bound holds*

$$\mathbb{E}\Big|\Big(1 + \sum_{j=1}^n g_{\zeta,n}(x_j, \bar{x})\Big) - \prod_{j=1}^n \Big(1 + g_{\zeta,n}(\epsilon_j, 0)\Big)\Big| \le \frac{2\zeta^2\sigma^2}{n}\Big(1 \vee \frac{\zeta^2\sigma^2}{n}\Big). \tag{4.4}$$

For the second term II, by a standard swapping argument we have

$$\mathbb{E}\Big|\prod_{j=1}^n \Big(1 + g_j^o\Big) e^{i\zeta\epsilon_j/n} - \prod_{j=1}^n \Big(1 + \frac{i\zeta\epsilon_j}{n}\Big)\Big| \le \sum_{j=1}^n \Big(\mathbb{E}|1 + g_j^o|\Big)^{(j-1)_+} \mathbb{E}\Big[\Big|g_j^o e^{\frac{i\zeta\epsilon_j}{n}} - \frac{i\zeta\epsilon_j}{n}\Big|\Big] \Big(\mathbb{E}\Big|1 + \frac{i\zeta\epsilon_j}{n}\Big|\Big)^{(n-j)_+}.$$

where $x_+ = \max(0, x)$. We use the following lemma to handle the term $\mathbb{E}\big[\big|g_j^o e^{i\zeta\epsilon_j/n} - i\zeta\epsilon_j/n\big|\big]$.

Lemma 5 *For any $x \in \mathbb{R}$, the following bound holds*

$$\big|(1 + x^2/2)e^{ix} - (1 + ix)\big| \le \min\{|x|^2, |x|^3\}. \tag{4.5}$$

Note that by Lemma 5, we replace x by $\zeta\epsilon_j/n$, it gives

$$\mathbb{E}\Big[\Big|g_j^o e^{i\zeta\epsilon_j/n} - \frac{i\zeta\epsilon_j}{n}\Big|\Big] \le \mathbb{E}|\zeta\epsilon_j/n|^2 \wedge \mathbb{E}|\zeta\epsilon_j/n|^3 \le \zeta^2\sigma^2/n^2. \tag{4.6}$$

Similarly, by Lyapunov's inequality

$$\mathbb{E}\Big|1 + i\zeta\epsilon_j/n\Big| \le \Big(\mathbb{E}\big[\big|1 + i\zeta\epsilon_j/n\big|^2\big]\Big)^{1/2} \le \sqrt{1 + 1/n} \tag{4.7}$$

and

$$\mathbb{E}\big|1 + g_j^o\big| \le 1 + \zeta^2\sigma^2/2n^2 \le 1 + 1/2n. \tag{4.8}$$

Then (4.6), (4.7) and (4.8) implies that

$$\mathbb{E}\Big|\prod_{j=1}^n \Big(1 + g_j^o\Big) e^{i\zeta\epsilon_j/n} - \prod_{j=1}^n \Big(1 + \frac{i\zeta\epsilon_j}{n}\Big)\Big| \le n \cdot \frac{e\zeta^2\sigma^2}{n^2} \le \frac{e\zeta^2\sigma^2}{n}. \tag{4.9}$$

For the third term III, due to independence and $\mathbb{E}[\epsilon_j] = 0$, it naturally holds

$$\mathbb{E}\Big[\prod_{j=1}^{n}\Big(1+\frac{i\zeta\epsilon_j}{n}\Big)-1\Big] = \Big[\prod_{j=1}^{n}\mathbb{E}\Big(1+\frac{i\zeta\epsilon_j}{n}\Big)-1\Big] = 0. \tag{4.10}$$

Combining (4.4), (4.9) and (4.10), we get

$$\begin{aligned}\big|\mathbb{E}_{\theta_i}\big[\mathcal{T}(f_i)\big] - f_i^N(\theta_i)\big| &= \Big|\frac{1}{\sqrt{2\pi}}\int_\Omega \mathcal{F}f_i^N(\zeta)\mathbb{E}\Big[\Big(1+\sum_{j=1}^{n}g_j\Big)e^{i\zeta\bar{\epsilon}}-1\Big]e^{i\zeta\theta_i}d\zeta\Big| \\ &\lesssim \frac{1}{\sqrt{2\pi}}\int_\Omega \Big|\mathcal{F}f_i^N(\zeta)\Big|\cdot\frac{\sigma^2\zeta^2}{n}d\zeta\end{aligned} \tag{4.11}$$

Here we consider two cases: 1. when $1 < s < 2$, in this case, one can check that given $|\zeta| < \sqrt{n}$ (4.11) implies for some numerical constant C_2

$$\big|\mathbb{E}_{\theta_i}\big[\mathcal{T}(f_i)\big] - f_i^N(\theta_i)\big| \le C_2\frac{\|f_i\|_{s,\infty,1}}{n}\cdot\Big(\frac{|\zeta|^{2-s}}{n^{(2-s)/2}}\Big) \le C_2\frac{\|f_i\|_{s,\infty,1}}{n};$$

2. when $s \ge 2$, in this case, one can check that similarly for some numerical constant C_3

$$\big|\mathbb{E}_{\theta_i}\big[\mathcal{T}(f_i)\big] - f_i^N(\theta_i)\big| \le C_3\|f_i\|_{s,\infty,1}/n. \tag{4.12}$$

On the other hand, as for the term $\tilde{f}^N(\theta_i)$, this can be bounded by

$$|\tilde{f}^N(\theta_i)| \le \frac{1}{\sqrt{2\pi}}\int_{\mathbb{R}\setminus\Omega}\Big|\mathcal{F}f_i(\theta_i)\Big|\cdot\frac{|\zeta|^s}{R^s}d\zeta \le \|f_i\|_{s,\infty,1}\cdot R^{-s}$$

As a consequence, combining the above analysis and (4.2), we get

$$\big|\mathbb{E}_{\boldsymbol{\theta}}\big[g^*(\mathbf{X})\big] - f(\boldsymbol{\theta})\big| \le C^*(d/n \vee dR^{-s}),$$

which completes the proof of Theorem 2. ■

5. Asymptotic normality

In this section, we establish the asymptotic normality of the proposed estimator $g^*(\mathbf{X})$. Especially, we show that $g^*(\mathbf{X})$ is normally distributed around the true parameter $f(\boldsymbol{\theta})$ as sample size n goes to infinity. Such results are very meaningful in practice and provide theoretical guarantee when one intends to build confidence intervals of the true parameter using the estimator. In the last section, we showed that the bias of $g^*(\mathbf{X})$ is well controlled. Aside from that, we still need to show that the variance of $g^*(\mathbf{X})$ is still well controlled. As there is the bias-variance trade-off phenomenon in statistical learning theory, we want to make sure that the new estimator doesn't simply sacrifice its variance to reduce its bias. According to the well-known Cramér-Rao bound, $\sigma^2\|\nabla f(\boldsymbol{\theta})\|^2/n$ is the best possible variance for any unbiased estimator of $f(\boldsymbol{\theta})$. In the following Theorem 6, we show that the asymptotic variance of $g^*(\mathbf{X})$ is indeed $\sigma^2\|\nabla f(\boldsymbol{\theta})\|^2/n$. In other words, the new estimator does not only achieve effective bias-reduction, it also achieves the best variance in asymptotic sense. This further implies that the new estimator is asymptotically efficient.

Theorem 6 *Under model (1.1), suppose that the noise vector $\boldsymbol{\varepsilon}_j$, $j = 1, ..., n$ are i.i.d. copies of a random vector $\boldsymbol{\varepsilon}$ satisfying Assumption 1-2. Let $g^*(\mathbf{X})$ be the estimator defined in (3.6) by taking $R \asymp \sqrt{n}$. Assume that for a given $f \in \mathcal{F}_d^s(M)$ with $s > 1$, $\|\nabla f(\boldsymbol{\theta})\| \asymp \sqrt{d}$ and $d = n^\alpha$, $\alpha \in (0, 1)$. Then when $s \geq 2$*

$$\frac{\sqrt{n}\big(g^*(\mathbf{X}) - f(\boldsymbol{\theta})\big)}{\sigma\big\|\nabla f(\boldsymbol{\theta})\big\|} \Rightarrow \mathcal{N}(0, 1), \quad \text{as} \quad n \to \infty. \tag{5.1}$$

where $\mathcal{N}(0, 1)$ is the standard normal distribution. Moreover, when $1 < s < 2$, if $s > 1 + \alpha$

$$\frac{\sqrt{n}\big(g^*(\mathbf{X}) - f(\boldsymbol{\theta})\big)}{\sigma\big\|\nabla f(\boldsymbol{\theta})\big\|} \Rightarrow \mathcal{N}(0, 1), \quad \text{as} \quad n \to \infty. \tag{5.2}$$

Remark 7 *Theorem 6 shows that aside from the benefit of bias reduction, the proposed estimator's variance is still reasonable. In fact, it shows that the asymptotic variance of the proposed estimator is optimal. Meanwhile, Theorem 6 indicates that our estimator $g^*(\mathbf{X})$ is asymptotically normally distributed around the true parameter $f(\boldsymbol{\theta})$ with an asymptotic variance $\sigma^2\|\nabla f(\boldsymbol{\theta})\|^2/n$. This indicates the new estimator is asymptotically efficient and validates its effectiveness in bias-reduction. On the other hand, when $s \geq 2$, asymptotic normality holds for all dimensions without any smoothness constraint. This is due to f's special additive structure and is different from the result for general smooth function without any specific structure, see Koltchinskii and Zhilova (2021a); Zhou and Li (2019, 2021). As Koltchinskii and Zhilova (2021a) recently discovered that the smoothness has to be above certain dimension related threshold in order to achieve asymptotic normality. As for the case $1 < s < 2$, it still requires $s > 1 + \alpha$ which is also needed to achieve asymptotic normality for additive models via other methods such as iterative bootstrap under Gaussian shift model. Currently, we don't know whether this requirement is essential or just a technical barrier.*

Remark 8 *Together with (4.1), Theorem 6 shows that when $s \geq 2$*

$$\mathbb{E}\big[g^*(\mathbf{X}) - f(\boldsymbol{\theta})\big]^2 \lesssim \sigma^2 d/n. \tag{5.3}$$

This is the optimal rate on estimation of $f(\boldsymbol{\theta})$ with additive structure for sufficiently smooth f and non-sparse $\boldsymbol{\theta}$. As we shall see in Section 6, thanks to its effective bias reduction, our estimator's performance aligns well with this rate while $f(\bar{\mathbf{x}})$'s derails when d becomes large.

Proof [Proof of Theorem 6.] Due to page limit, we outline the key steps of this proof. Detailed proofs of the key lemmas are deferred to supplementary material. By the definition of (3.6), we have the following decomposition

$$g^*(\mathbf{X}) - f(\boldsymbol{\theta}) = \sum_{i=1}^{d} \Big(\mathcal{T}(f_i) - f_i^N(\theta_i) - \tilde{f}_i^N(\theta_i)\Big).$$

Recall that from the proof of Theorem 2 we showed

$$g^*(\mathbf{X}) - f(\boldsymbol{\theta}) = \sum_{i=1}^{d} \Big(\frac{1}{\sqrt{2\pi}} \int_\Omega \mathcal{F} f_i^N(\zeta) \cdot \Big(\prod_{j=1}^{n} \Big(1 + \frac{i\zeta\epsilon_j^{(i)}}{n}\Big) - 1\Big) e^{i\zeta\theta_i} d\zeta + \tilde{\mathcal{R}} \tag{5.4}$$

where $\tilde{\mathcal{R}}$ denotes the remainder and $\epsilon_j^{(i)}$ denotes the i-th coordinate of $\boldsymbol{\varepsilon}_j$.

The first step is to show that under the condition of Theorem 6, $\sqrt{n}\tilde{\mathcal{R}}/\sigma\|\nabla f(\boldsymbol{\theta})\|$ converges to 0 in probability. It follows directly from Theorem 2 and we state it in the following lemma.

Lemma 9 *Under the condition of Theorem 6, if 1.* $s \geq 2$ *or 2.* $1 < s < 2$ *and* $s > 1 + \alpha$*, then*

$$\frac{\sqrt{n} \cdot \tilde{\mathcal{R}}}{\sigma \left\| \nabla f(\boldsymbol{\theta}) \right\|} \xrightarrow{p} 0. \tag{5.5}$$

The rest will be dealing with the first term in (5.4). We expand the product inside the integral as

$$\prod_{j=1}^{n} \left(1 + i\zeta \epsilon_j / n\right) = 1 + S_1(\zeta) + \cdots + S_n(\zeta) \tag{5.6}$$

with $S_u(\zeta) = \sum_{\{j_1, \ldots, j_u\} \subset [n]} \prod_{1 \leq j_1 \neq \cdots \neq j_u \leq n} (i\zeta \epsilon_{j_k} / n)$. Therefore,

$$\begin{aligned} &\sum_{i=1}^{d} \Big(\frac{1}{\sqrt{2\pi}} \int_{\Omega} \mathcal{F} f_i^N(\zeta) \cdot \Big(\prod_{j=1}^{n} \Big(1 + \frac{i\zeta \epsilon_j^{(i)}}{n}\Big) - 1 \Big) e^{i\zeta\theta_i} d\zeta \\ &= \sum_{i=1}^{d} \frac{1}{\sqrt{2\pi}} \int_{\Omega} \mathcal{F} f_i^N(\zeta) \Big(S_1(\zeta) + \cdots + S_n(\zeta) \Big) e^{i\zeta\theta_i} d\zeta. \end{aligned} \tag{5.7}$$

The next major step is to show that the term associated with $S_1(\zeta)$ in (5.7) scaled by $\sigma \left\| \nabla f(\boldsymbol{\theta}) \right\| / \sqrt{n}$ converges in distribution to the standard normal $\mathcal{N}(0, 1)$. We claim it as the following lemma.

Lemma 10 *Under the condition of Theorem 6, we have when* $n \to \infty$

$$\frac{\sqrt{n} \cdot \left(\sum_{i=1}^{d} \frac{1}{\sqrt{2\pi}} \int_{\Omega} \mathcal{F} f_i^N(\zeta) \cdot S_1(\zeta) \cdot e^{i\zeta\theta_i} d\zeta \right)}{\sigma \left\| \nabla f(\boldsymbol{\theta}) \right\|} \Rightarrow \mathcal{N}(0, 1). \tag{5.8}$$

The final step is to show that the sum of remaining terms (5.7) except for $S_1(\zeta)$ scaled by $\sigma \left\| \nabla f(\boldsymbol{\theta}) \right\| / \sqrt{n}$ converges in probability to 0 under the condition of Theorem 6. This is done by the key observation that S_k $k \geq 2$ are completely degenerate U-statistic of order k. We claim this result in the following lemma.

Lemma 11 *Under the condition of Theorem 6, suppose that 1. if* $s \geq 2$ *or 2. if* $1 < s < 2$ *and* $s > 1 + \alpha$*. Then as* $n \to \infty$

$$\frac{\sqrt{n} \cdot \left(\sum_{i=1}^{d} \frac{1}{\sqrt{2\pi}} \int_{\Omega} \mathcal{F} f_i^N(\zeta) \sum_{k=2}^{n} S_k(\zeta) e^{i\zeta\theta_i} d\zeta \right)}{\sigma \left\| \nabla f(\boldsymbol{\theta}) \right\|} \xrightarrow{p} 0. \tag{5.9}$$

Combine the results of (5.5), (5.8) and (5.9), by Slutsky's Theorem we get

$$\frac{\sqrt{n} \cdot \left(g^*(\mathbf{X}) - f(\boldsymbol{\theta}) \right)}{\sigma \left\| \nabla f(\boldsymbol{\theta}) \right\|} \Rightarrow \mathcal{N}(0, 1), \quad \text{as} \quad n \to \infty, \tag{5.10}$$

which completes the proof of Theorem 6. ■

6. Numerical simulation

We present numerical simulation results of our estimator (3.6) (denoted as Adaptive) and compare it with plug-in estimator $f(\bar{\mathbf{x}})$ (denoted as Plug-in) and the estimator (3.1) introduced for Gaussian case (denoted as EXP). The difference between Adaptive and EXP is that the latter replaces $g_{\zeta,n}(x_j, \bar{x})$ by the exponential term $e^{\zeta^2\sigma^2}$. Note that to implement EXP, one has to know the underlying variance σ^2 of noise, while Adaptive doesn't need that. This automatically makes our estimator adaptive and data-driven. The unknown parameters $\boldsymbol{\theta} \in \mathbb{R}^d$ are randomly generated that θ_i is uniformly distributed over $[0.4, 0.6]$. We set $\sigma = 1$ and $n = 10^3$. For the dimension factor, we set $d = n^\alpha$ and α ranges from 0.5 to 0.95 with an incremental size 0.05. The noise $\boldsymbol{\varepsilon}$ we use is an isotropic random vector satisfying multivariate Student's t-distribution with degree of freedom $\nu = 3, 4$. Note that when $\nu = 4$, the forth moment does not exist respectively. The additive function we use has a homogeneous Hölder smoothness structure: $f(\boldsymbol{\theta}) := \sum_{i=1}^d h(\theta_i) = \sum_{i=1}^d \theta_i^s$. More experiments are in the supplement.

Firstly, we compare the performance of Adaptive, EXP and Plug-in on bias-reduction, variance and MSE with $h(\theta)$ of two different smoothness. We simulated the bias, variance and MSE from 1000 independent runs.The simulation results are presented in the first and the second row of Figure 1.

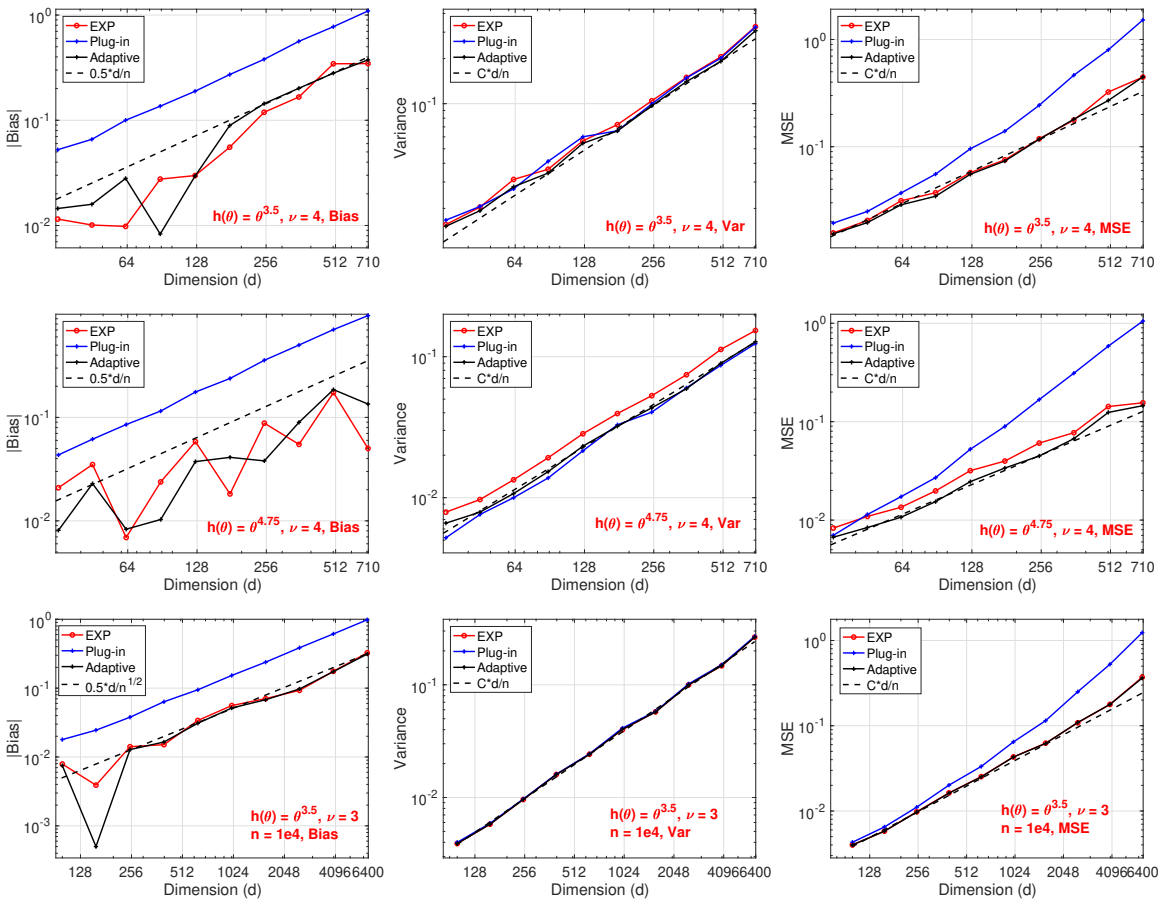

Figure 1: Comparison on bias, variance, and MSE.

As we can see, both EXP and Adaptive have clear bias-reduction compared with Plug-in. The black dash line is the upper bound we show in Theorem 2. The simulation results validate Theorem 2. Meanwhile, the similar performance of EXP and Adaptive on bias reduction also verifies that our idea of using $g_{\zeta,n}$ as an approximation of the exponential term is very effective. As for the variance comparison, we plot $\sigma^2\|f(\boldsymbol{\theta})\|^2/n$ as black dash lines. It is supposed to be the optimal variance. Plug-in and Adaptive almost have the same variance that matches the black dash lines while EXP's is bigger in the case $h(\theta) = \theta^{4.75}$. This validates our Theorem 6 which shows that our estimator has the optimal asymptotic variance. Given this and its advantages over Plug-in on bias reduction, it shows Adaptive is superior. This is further reflected in the comparisons of MSE which show that when the dimension d is large, Adaptive has very obvious reduction in MSE thanks to its small bias. The black dash lines in MSE comparisons are of the order $O(d/n)$ which is the optimal rate. As we can see, Plug-in is far from optimal while our estimator Adaptive matches it well. In terms of comparison with EXP, the advantage of Adaptive is evident: 1. the first two figures in the middle column shows that EXP's variance can be unstable when s is larger for heavy-tailed noise, while Adaptive's performance is consistent; 2. To implement Adaptive, we don't need to know the variance of the noise. Moreover, there is no theoretical guarantee for EXP when it applies to heavy-tailed noise, while we provide it with Adaptive, which is a major contribution of this work.

7. Conclusion

We proposed a new estimator to study the estimation of $f(\boldsymbol{\theta}) = \sum_{i=1}^d f_i(\theta_i)$ based on noisy observations of $\boldsymbol{\theta}$ when f is a given smooth additive function and d is large. Our major contribution is that we not only introduced a new estimator that can be efficiently computed and has good performance in practice, but also introduced new analysis to provide theoretical justification of its performance in a distribution free setting under minimum moment constraint of noise whereas previous methods are either under a specific noise distribution or computationally intensive.

As we have mentioned in Remark 3.1 and 4.2, the current approach is based on the second order Taylor expansion of $e^{i\zeta\bar{\epsilon}}$. This is due to only the second moment of noise are assumed. When higher moments of noise exist, one can seek to achieve better bias reduction by approximation of higher order terms in expansion of $e^{i\zeta\bar{\epsilon}}$ as what we did using $\sum_{j=1}^n g_{\zeta,n}(x_j, \bar{x})$. In short, our method can be easily generalized to a moment adaptive approach to achieve sharp bias correction for such problems. However, theoretical analysis of such generalization can be more complicated and beyond the scope.

References

Jayadev Acharya, Hirakendu Das, Alon Orlitsky, and Ananda Theertha Suresh. A unified maximum likelihood approach for estimating symmetric properties of discrete distributions. In *Proceedings of the 34th International Conference on Machine Learning (ICML)*, pages 11–21, Sydney, Australia, 2017.

Peter J Bickel and Yaacov Ritov. Estimating integrated squared density derivatives: sharp best order of convergence estimates. *Sankhyā: The Indian Journal of Statistics, Series A*, pages 381–393, 1988.

Lucien Birgé and Pascal Massart. Estimation of integral functionals of a density. *The Annals of Statistics*, 23(1):11–29, 1995.

T Tony Cai and Mark G Low. On adaptive estimation of linear functionals. *The Annals of Statistics*, 33(5):2311–2343, 2005a.

T Tony Cai and Mark G Low. Nonquadratic estimators of a quadratic functional. *The Annals of Statistics*, 33(6):2930–2956, 2005b.

T Tony Cai and Mark G Low. Testing composite hypotheses, hermite polynomials and optimal estimation of a nonsmooth functional. *The Annals of Statistics*, 39(2):1012–1041, 04 2011.

Alexandra Carpentier and Nicolas Verzelen. Adaptive estimation of the sparsity in the gaussian vector model. *The Annals of Statistics*, 47(1):93–126, 02 2019.

Raymond J Carroll, David Ruppert, Leonard A Stefanski, and Ciprian M Crainiceanu. *Measurement error in nonlinear models: a modern perspective*. CRC press, 2006.

Olivier Collier and Laëtitia Comminges. Minimax optimal estimators for general additive functional estimation. *arXiv preprint arXiv:1908.11070*, 2019.

Olivier Collier, Laëtitia Comminges, and Alexandre B Tsybakov. Minimax estimation of linear and quadratic functionals on sparsity classes. *The Annals of Statistics*, 45(3):923–958, 2017.

Olivier Collier, Laëtitia Comminges, and Alexandre B Tsybakov. On estimation of nonsmooth functionals of sparse normal means. *Bernoulli*, 26(3):1989–2020, 08 2020.

David L Donoho and Richard Chieng Liu. *On minimax estimation of linear functionals*. University of California (Berkeley). Department of Statistics, 1987.

David L Donoho and Richard Chieng Liu. Geometrizing rates of convergence, iii. *The Annals of Statistics*, 19(2):668–701, 1991.

David L Donoho and Michael Nussbaum. Minimax quadratic estimation of a quadratic functional. *Journal of Complexity*, 6(3):290–323, 1990.

Rick Durrett. *Probability: Theory and Examples*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 5 edition, 2019.

Jianqing Fan and Young K Truong. Nonparametric regression with errors in variables. *The Annals of Statistics*, pages 1900–1925, 1993.

Jerome H. Friedman. Greedy function approximation: A gradient boosting machine. *The Annals of Statistics*, 29(5):1189–1232, 2001.

Kyunghee Han and Byeong U Park. Smooth backfitting for errors-in-variables additive models. *Annals of Statistics*, 46(5):2216–2250, 2018.

Yi Hao and Ping Li. Bessel smoothing and multi-distribution property estimation. In *Proceedings of the Conference on Learning Theory (COLT)*, pages 1817–1876, Virtual Event [Graz, Austria], 2020a.

Yi Hao and Ping Li. Optimal prediction of the number of unseen species with multiplicity. In *Advances in Neural Information Processing Systems (NeurIPS)*, virtual, 2020b.

Yi Hao and Alon Orlitsky. The broad optimality of profile maximum likelihood. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 10989–11001, Vancouver, Canada, 2019.

Trevor J Hastie. Generalized additive models. In *Statistical models in S*, pages 249–307. Routledge, 2017.

Il'dar Abdullovich Ibragimov, Arkadi Nemirovskii, and Rafail Zalmanovich Khas' minskii. Some problems of nonparametric estimation in the gaussian white noise. *Teoriya Veroyatnostei i ee Primeneniya*, 31(3):451–466, 1986.

Jiantao Jiao and Yanjun Han. Bias correction with Jackknife, Bootstrap, and Taylor series. *IEEE Trans. Inf. Theory*, 66(7):4392–4418, 2020.

Jiantao Jiao, Kartik Venkat, Yanjun Han, and Tsachy Weissman. Minimax estimation of functionals of discrete distributions. *IEEE Trans. Inf. Theory*, 61(5):2835–2885, 2015.

Jussi Klemelä. Sharp adaptive estimation of quadratic functionals. *Probability theory and related fields*, 134(4):539–564, 2006.

Jussi Klemelä and Alexandre B Tsybakov. Sharp adaptive estimation of linear functionals. *The Annals of Statistics*, 29(6):1567–1600, 2001.

Andrei Nikolaevich Kolmogorov. Unbiased estimates. *Izvestiya Rossiiskoi Akademii Nauk. Seriya Matematicheskaya*, 14(4):303–326, 1950.

Vladimir Koltchinskii. Asymptotically efficient estimation of smooth functionals of covariance operators. *Journal of the European Mathematical Society*, 23(3):765–843, 2020.

Vladimir Koltchinskii. Estimation of smooth functionals of covariance operators: jackknife bias reduction and bounds in terms of effective rank. *arXiv preprint arXiv:2205.10280*, 2022.

Vladimir Koltchinskii and Mayya Zhilova. Efficient estimation of smooth functionals in Gaussian shift models. *Annales de l'Institut Henri Poincaré, Probabilités et Statistiques*, 57(1):351 – 386, 2021a.

Vladimir Koltchinskii and Mayya Zhilova. Estimation of smooth functionals in normal models: bias reduction and asymptotic efficiency. *The Annals of Statistics*, 49(5):2577–2610, 2021b.

Béatrice Laurent. Efficient estimation of integral functionals of a density. *The Annals of Statistics*, 24(2):659–681, 1996.

Béatrice Laurent and Pascal Massart. Adaptive estimation of a quadratic functional by model selection. *The Annals of Statistics*, pages 1302–1338, 2000.

Oleg Lepski, Arkadi Nemirovskii, and Vladimir Spokoiny. On estimation of the Lr norm of a regression function. *Probability Theory and Related Fields*, 113(2):221–253, 1999.

Boris Ya Levit. On the efficiency of a class of non-parametric estimates. *Theory of Probability & Its Applications*, 20(4):723–740, 1976.

Boris Ya Levit. Asymptotically efficient estimation of nonlinear functionals. *Problemy Peredachi Informatsii*, 14(3):65–72, 1978.

Arkadi Nemirovskii. On necessary conditions for efficient estimation of functionals of a nonparametric signal in white noise. *Theory of Probability & Its Applications*, 35(1):94–103, 1991.

Arkadi Nemirovskii. Topics in non-parametric. *Ecole d'Eté de Probabilités de Saint-Flour*, 28:85, 2000.

Hans Triebel. *Theory of Function Spaces*. Birkhäuser Basel, 1st edition, 1983.

Yihong Wu and Pengkun Yang. Minimax rates of entropy estimation on large alphabets via best polynomial approximation. *IEEE Trans. Inf. Theory*, 62(6):3702–3720, 2016.

Yihong Wu and Pengkun Yang. Chebyshev polynomials, moment matching, and optimal estimation of the unseen. *The Annals of Statistics*, 47(2):857–883, 04 2019.

Fan Zhou and Ping Li. A Fourier analytical approach to estimation of smooth functions in gaussian shift model. *arXiv preprint arXiv:1911.02010*, 2019.

Fan Zhou and Ping Li. Optimal estimation of high dimensional smooth additive function based on noisy observations. In *Proceedings of the 38th International Conference on Machine Learning (ICML)*, pages 12813–12823, Virtual Event, 2021.

Fan Zhou, Ping Li, and Cun-Hui Zhang. High-order statistical functional expansion and its application to some nonsmooth problems. *arXiv preprint arXiv:2112.15591*, 2021.

8. Proofs of Key Lemmas in the Main Paper

8.1. Proof of Lemma 4

Proof [Proof of Lemma 4.] According to our definition in (3.4), we have

$$\prod_{j=1}^{n}\Big(1+g_{\zeta,n}(\epsilon_j,0)\Big)=1+\sum_{j=1}^{n}g_{\zeta,n}(\epsilon_j,0)+\tilde{\mathcal{R}}_1, \tag{8.1}$$

where we use $\tilde{\mathcal{R}}_1$ to denote the remainder. Then we can write the left hand side of (4.4) as

$$\begin{aligned}\mathbb{E}\Big|\Big(1+\sum_{j=1}^{n}g_{\zeta,n}(x_j,\bar{x})\Big)-\prod_{j=1}^{n}\Big(1+g_{\zeta,n}(\epsilon_j,0)\Big)\Big| \\ \leq \mathbb{E}\Big|\sum_{j=1}^{n}(g_{\zeta,n}(\epsilon_j,\bar{\epsilon})-g_{\zeta,n}(\epsilon_j,0)\Big|+\mathbb{E}\tilde{\mathcal{R}}_1.\end{aligned} \tag{8.2}$$

To bound the first term on the right hand side of (8.2), we have

$$\begin{aligned}\mathbb{E}\Big|\sum_{j=1}^{n}(g_{\zeta,n}(\epsilon_j,\bar{\epsilon})-g_{\zeta,n}(\epsilon_j,0)\Big| &= \mathbb{E}\Big|\sum_{j=1}^{n}\Big(\frac{(\zeta\epsilon_j)^2}{2n^2}-\frac{(\zeta\epsilon_j-\zeta\bar{\epsilon})^2}{2n^2}\Big)\Big| \\ &= \mathbb{E}\Big[\frac{(\zeta\bar{\epsilon})^2}{n}\Big]=\frac{\zeta^2\sigma^2}{n^2}.\end{aligned} \tag{8.3}$$

Next, we bound the second term on the right hand side of (8.2). Observe that

$$\tilde{\mathcal{R}}_1:=\sum_{j_1\neq j_2}g_{\zeta,n}(\epsilon_{j_1},0)g_{\zeta,n}(\epsilon_{j_2},0)+\cdots+g_{\zeta,n}(\epsilon_1,0)g_{\zeta,n}(\epsilon_2,0)\cdots g_{\zeta,n}(\epsilon_n,0).$$

Due to independence and under the condition $\sigma^2R^2<n$, we have

$$\mathbb{E}\tilde{\mathcal{R}}_1=\sum_{k=2}^{n}\binom{n}{k}\Big(\mathbb{E}\big[g_{\zeta,n}(\epsilon_k,0)\big]\Big)^k\leq\sum_{k=2}^{n}\binom{n}{k}\Big(\frac{\zeta^2\sigma^2}{2n^2}\Big)^k\leq\sum_{k=2}^{n}\frac{1}{k!}\Big(\frac{\zeta^2\sigma^2}{2n}\Big)^k\leq\frac{\zeta^4\sigma^4}{4n^2} \tag{8.4}$$

■

8.2. Proof of Lemma 5

Proof [Proof of Lemma 5] Integrating by parts gives

$$\int_0^x(x-s)^ne^{is}ds=\frac{x^{n+1}}{n+1}+\frac{i}{n+1}\int_0^x(x-s)^{n+1}e^{is}ds. \tag{8.5}$$

Set $n=0$, we have

$$\int_0^x e^{is}ds=x+i\int_0^x(x-s)^ne^{is}ds. \tag{8.6}$$

By rearranging the terms, we get

$$e^{ix} = 1 + ix + i^2 \int_0^x (x-s)e^{is}ds. \tag{8.7}$$

Now, one can check that

$$\begin{aligned}(1+x^2/2)e^{ix} - (1+ix) &= x^2e^{ix}/2 + i^2\int_0^x (x-s)e^{is}ds \\ &= \int_0^x i\Big(x(x-u) - \frac{(x-u)^2}{2}\Big)e^{i(x-u)}du.\end{aligned} \tag{8.8}$$

When $|x| \le 1$, (8.8) implies that

$$(1+x^2/2)e^{ix} - (1+ix) \le |x|^3. \tag{8.9}$$

On the other hand, when $|x| > 1$, integration by parts gives

$$\begin{aligned}\big|(1+x^2/2)e^{ix} - (1+ix)\big| &= \Big|\int_0^x i\Big(x(x-u) - \frac{(x-u)^2}{2}\Big)e^{i(x-u)}du\Big| \\ &= \Big|i\Big(x^2 - \frac{x^2}{2}\Big)e^{ix} + \int_0^x i\Big(x-(x-u)\Big)e^{i(x-u)}du\Big| \\ &\le 2\cdot|x|^2/2 \le |x|^2.\end{aligned} \tag{8.10}$$

Combine (8.9) and (8.10), this completes the proof of Lemma 5. ■

8.3. Proof of Lemma 9

Proof [Proof of Lemma 9.] By Theorem 2, we have

$$\mathbb{E}\Big|\frac{\sqrt{n}\cdot\tilde{\mathcal{R}}}{\sigma\|\nabla f(\boldsymbol{\theta})\|}\Big| \lesssim \frac{1}{\|\nabla f(\boldsymbol{\theta})\|}\cdot\Big(\frac{d}{\sqrt{n}} \bigvee \frac{d\sqrt{n}}{R^s}\Big). \tag{8.11}$$

Apparently, under condition $d = n^\alpha$, $\alpha \in (0,1)$ and $R \asymp \sqrt{n}$ and $\|\nabla f(\boldsymbol{\theta})\| \asymp \sqrt{d}$, we have case 1: if $s \ge 2$:

$$\mathbb{E}\Big|\frac{\sqrt{n}\cdot\tilde{\mathcal{R}}}{\sigma\|\nabla f(\boldsymbol{\theta})\|}\Big| \lesssim \sqrt{\frac{d}{n}} \to 0, \quad as \;\; n\to\infty; \tag{8.12}$$

or case 2: if $1 < s < 2$ and $s > 1+\alpha$:

$$\mathbb{E}\Big|\frac{\sqrt{n}\cdot\tilde{\mathcal{R}}}{\sigma\|\nabla f(\boldsymbol{\theta})\|}\Big| \lesssim \frac{\sqrt{d}}{n^{(s-1)/2}} \to 0, \quad as \;\; n\to\infty. \tag{8.13}$$

(8.12) and (8.13) imply that under the condition of Theorem 6 and by Lemma 2.2.2 in Durrett (2019),

$$\frac{\sqrt{n}\cdot\tilde{\mathcal{R}}}{\sigma\|\nabla f(\boldsymbol{\theta})\|} \xrightarrow{p} 0. \tag{8.14}$$

■

8.4. Proof of Lemma 10

Proof [Proof of Lemma 10.] Note that for the term associated with $S_1(\zeta)$, we have

$$\begin{aligned}\sum_{i=1}^{d}\frac{1}{\sqrt{2\pi}}\int_{\Omega}\mathcal{F}f_i^N(\zeta)\cdot S_1(\zeta)\cdot e^{i\zeta\theta_i}d\zeta &= \sum_{i=1}^{d}\frac{1}{\sqrt{2\pi}}\int_{\Omega}\mathcal{F}f_i^N(\zeta)\Big(\sum_{j=1}^{n}i\zeta\epsilon_j^{(i)}/n\Big)e^{i\zeta\theta_i}d\zeta \\ &= \sum_{j=1}^{n}\sum_{i=1}^{d}\frac{1}{\sqrt{2\pi}}\int_{\Omega}\mathcal{F}f_i^N(\zeta)\Big(i\zeta\epsilon_j^{(i)}/n\Big)e^{i\zeta\theta_i}d\zeta\end{aligned} \tag{8.15}$$

and for each fixed i and j, a single term on the right hand side is a zero mean random variable with the following variance

$$\begin{aligned}&\frac{1}{2\pi}\int_{\Omega}\int_{\Omega}\mathcal{F}f_i^N(\zeta_1)\mathcal{F}f_i^N(\zeta_2)e^{i\zeta_1\theta_i}e^{-i\zeta_2\theta_i}\cdot\frac{\sigma^2\zeta_1\zeta_2}{n^2}d\zeta_1d\zeta_2 \\ &=\frac{1}{2\pi}\int_{\Omega}\int_{\Omega}\mathcal{F}f_i^N(\zeta_1)\mathcal{F}f_i^N(\zeta_2)e^{i\zeta_1\theta_i}e^{-i\zeta_2\theta_i}\cdot\sigma^2n^{-2}(i\zeta_1)(-i\zeta_2)d\zeta_1d\zeta_2 \\ &=\sigma^2n^{-2}\big((f_i^N)'(\theta_i)\big)^2.\end{aligned} \tag{8.16}$$

As a consequence, (8.15) is a sum of n i.i.d. random variables with variance $\sigma^2n^{-1}\big\|\nabla f^N(\boldsymbol{\theta})\big\|^2$. By standard Central Limit Theorem,

$$\frac{\sqrt{n}\cdot\big(\sum_{i=1}^{d}\frac{1}{\sqrt{2\pi}}\int_{\Omega}\mathcal{F}f_i^N(\zeta)\cdot S_1(\zeta)\cdot e^{i\zeta\theta_i}d\zeta\big)}{\sigma\big\|\nabla f^N(\boldsymbol{\theta})\big\|}\Rightarrow\mathcal{N}(0,1). \tag{8.17}$$

Next, we still need to replace the variance $\sigma^2n^{-1}\big\|\nabla f^N(\boldsymbol{\theta})\big\|^2$ by $\sigma^2n^{-1}\big\|\nabla f(\boldsymbol{\theta})\big\|^2$. We use the following lemma.

Lemma 12 *For a given $f\in\mathcal{F}_d^s(M)$, we denote by $f^N(\boldsymbol{\theta}):=\sum_{i=1}^{d}f_i^N(\theta_i)$. Then we have*

$$\big|\sigma n^{-1/2}\big\|\nabla f^N(\boldsymbol{\theta})\big\|-\sigma n^{-1/2}\big\|\nabla f(\boldsymbol{\theta})\big\|\big|\le M\sigma\sqrt{d/n}\cdot R^{1-s}. \tag{8.18}$$

Observe that

$$\begin{aligned}&\frac{\sqrt{n}\cdot\big(\sum_{i=1}^{d}\frac{1}{\sqrt{2\pi}}\int_{\Omega}\mathcal{F}f_i^N(\zeta)\cdot S_1(\zeta)\cdot e^{i\zeta\theta_i}d\zeta\big)}{\sigma\big\|\nabla f(\boldsymbol{\theta})\big\|} \\ &=\frac{\sqrt{n}\cdot\big(\sum_{i=1}^{d}\frac{1}{\sqrt{2\pi}}\int_{\Omega}\mathcal{F}f_i^N(\zeta)\cdot S_1(\zeta)\cdot e^{i\zeta\theta_i}d\zeta\big)}{\sigma\big\|\nabla f^N(\boldsymbol{\theta})\big\|}\cdot\frac{\big\|\nabla f^N(\boldsymbol{\theta})\big\|}{\big\|\nabla f(\boldsymbol{\theta})\big\|}.\end{aligned} \tag{8.19}$$

According to Lemma 12, under the condition of Theorem 6 we have $\big\|\nabla f^N(\boldsymbol{\theta})\big\|/\big\|\nabla f(\boldsymbol{\theta})\big\|\to0$ as $n\to\infty$. Together with (8.17), we have

$$\frac{\sqrt{n}\cdot\big(\sum_{i=1}^{d}\frac{1}{\sqrt{2\pi}}\int_{\Omega}\mathcal{F}f_i^N(\zeta)\cdot S_1(\zeta)\cdot e^{i\zeta\theta_i}d\zeta\big)}{\sigma\big\|\nabla f(\boldsymbol{\theta})\big\|}\Rightarrow\mathcal{N}(0,1),\ as\ n\to\infty. \tag{8.20}$$

■

8.5. Proof of Lemma 12

Proof [Proof of Lemma 12.] By triangle inequality,

$$\left|\sigma n^{-1/2}\left\|\nabla f^N(\boldsymbol{\theta})\right\| - \sigma n^{-1/2}\left\|\nabla f(\boldsymbol{\theta})\right\|\right| \leq \sigma n^{-1/2}\left\|\nabla f^N(\boldsymbol{\theta}) - \nabla f(\boldsymbol{\theta})\right\|. \tag{8.21}$$

By the analysis in (8.16),

$$\sigma n^{-1/2}\left\|\nabla f^N(\boldsymbol{\theta}) - \nabla f(\boldsymbol{\theta})\right\| \leq \sigma n^{-1/2}\sqrt{\sum_{i=1}^{d}\left\|f_i\right\|_{s,\infty,1}^2}\cdot R^{1-s} \leq \sigma\sqrt{d/n}\cdot R^{1-s} \tag{8.22}$$

■

8.6. Proof of Lemma 11

Proof [Proof of Lemma 11.] As for the case $u \geq 2$ for a fixed i,

$$\begin{aligned}
&\mathbb{E}\Big|\frac{1}{\sqrt{2\pi}}\int_\Omega \mathcal{F}f_i^N(\zeta)\sum_{u=2}^{n} S_u(\zeta)e^{i\zeta\theta_i}d\zeta\Big| \leq \frac{1}{\sqrt{2\pi}}\int_\Omega \Big|\mathcal{F}f_i^N(\zeta)\Big|\mathbb{E}\Big|\sum_{u=2}^{n} S_u(\zeta)\Big|d\zeta \\
&\leq \frac{1}{\sqrt{2\pi}}\int_\Omega \Big|\mathcal{F}f_i^N(\zeta)\Big|\sqrt{\mathbb{E}\Big|\sum_{u=2}^{n} S_u(\zeta)\Big|^2}d\zeta = \frac{1}{\sqrt{2\pi}}\int_\Omega \Big|\mathcal{F}f_i^N(\zeta)\Big|\sqrt{\sum_{u=2}^{n}\mathbb{E}\big|S_u(\zeta)\big|^2}d\zeta \\
&\lesssim \frac{1}{\sqrt{2\pi}}\int_\Omega \Big|\mathcal{F}f_i^N(\zeta)\Big|\cdot\sqrt{\sum_{u=2}^{n}\frac{(\sigma^2\zeta^2)^u}{n^u u!}}d\zeta \leq \frac{1}{\sqrt{2\pi}}\int_\Omega \Big|\mathcal{F}f_i^N(\zeta)\Big|\cdot\frac{\sigma^2\zeta^2}{n}d\zeta \\
&\lesssim 2\sigma^2\left\|f_i\right\|_{s,\infty,1}\cdot n^{-(2\wedge s)/2}.
\end{aligned} \tag{8.23}$$

The second line is due to Jensen's inequality; the third inequality is due to the fact that S_u are completely degenerate U-statistics of order u, orthogonal to each other, and with variance

$$\mathbb{E}\Big[|S_u(\zeta)|^2\Big] = \binom{n}{u}\left(\mathbb{E}\big[|\zeta\epsilon_1/n|^2\big]\right)^u \leq \left(\mathbb{E}\big[|\zeta\epsilon_1/n^{1/2}|^2\big]\right)^u/u! = (\sigma^2\zeta^2/n)^u/u!, \tag{8.24}$$

and $\zeta^2/n \leq 1$. As a result, we get

$$\frac{\mathbb{E}\Big|\sum_{i=1}^{d}\frac{1}{\sqrt{2\pi}}\int_\Omega \mathcal{F}f_i^N(\zeta)\sum_{k=2}^{n} S_k(\zeta)e^{i\zeta\theta_i}d\zeta\Big|}{\sigma n^{-1/2}\left\|\nabla f(\boldsymbol{\theta})\right\|} \leq \frac{2M\sigma d\cdot n^{-(2\wedge s)/2}}{n^{-1/2}\left\|\nabla f(\boldsymbol{\theta})\right\|}. \tag{8.25}$$

One can check that for $s \geq 2$, and $\|\nabla f(\boldsymbol{\theta})\| \asymp \sqrt{d}$, $d = n^\alpha$ with $\alpha \in (0,1)$ when $n \to \infty$,

$$\frac{\mathbb{E}\Big|\sum_{i=1}^{d}\frac{1}{\sqrt{2\pi}}\int_\Omega \mathcal{F}f_i(\zeta)\sum_{k=2}^{n} S_k(\zeta)e^{i\zeta\theta_i}d\zeta\Big|}{\sigma n^{-1/2}\left\|\nabla f(\boldsymbol{\theta})\right\|} \leq \frac{2\sigma d\cdot n^{-(2\wedge s)/2}}{n^{-1/2}\left\|\nabla f(\boldsymbol{\theta})\right\|} \lesssim \frac{\sqrt{d}}{\sqrt{n}} \to 0. \tag{8.26}$$

When $1 < s < 2$, and $\|\nabla f(\boldsymbol{\theta})\| \asymp \sqrt{d}$, $d = n^\alpha$ with $\alpha \in (0,1)$, and $s > 1 + \alpha$

$$\frac{\mathbb{E}\Big|\sum_{i=1}^{d}\frac{1}{\sqrt{2\pi}}\int_\Omega \mathcal{F}f_i(\zeta)\sum_{k=2}^{n} S_k(\zeta)e^{i\zeta\theta_i}d\zeta\Big|}{\sigma n^{-1/2}\left\|\nabla f(\boldsymbol{\theta})\right\|} \leq \frac{2\sigma d\cdot n^{-s/2}}{n^{-1/2}\left\|\nabla f(\boldsymbol{\theta})\right\|} \lesssim \frac{\sqrt{d}}{n^{(s-1)/2}} \to 0. \tag{8.27}$$

Both (8.26) and (8.27) imply that under the condition of Theorem 6,

$$\frac{\sum_{i=1}^{d}\frac{1}{\sqrt{2\pi}}\int_{\Omega}\mathcal{F}f_i^N(\zeta)\sum_{k=2}^{n}S_k(\zeta)e^{i\zeta\theta_i}d\zeta}{\sigma n^{-1/2}\left\|\nabla f(\boldsymbol{\theta})\right\|}\xrightarrow{p}0,\ \ as\ \ n\to\infty \tag{8.28}$$

■

Appendix A. Additional Experiments and More Discussion

A.1. Analytical approximation of $f(\theta)$

In this section, we compare how well $f:\mathbb{R}\to\mathbb{R}$ can be approximated by the Fourier analytical approximations developed in Zhou and Li (2019) and in this paper through MATLAB FFT implementations. By doing this, on one hand, we would like to give the readers a concrete understanding of the approximation analysis we did in Section 3 and Section 4; and on the other, we show how this can be related to the choice of the tuning parameter R appeared in our estimator. We denote by $f(\theta)$ as a single component function in the additive model (1.2) and

$$g(\theta):=\frac{1}{\sqrt{2\pi}}\int_{\Omega}\mathcal{F}f(\zeta)e^{\sigma^2\zeta^2/2n}e^{i\theta\zeta}d\zeta;\ \ h(\theta):=\frac{1}{\sqrt{2\pi}}\int_{\Omega}\mathcal{F}f(\zeta)\Big(1+\sum_{j=1}^{n}g_{\zeta,n}(x_j,\bar{x})\Big)e^{i\zeta\theta}d\zeta,$$

Here $\Omega:=\{\zeta\in\mathbb{R}:|\zeta|\le R\}$. $g(\theta)$ is exactly the estimator used in Zhou and Li (2019) for one-dimensional case while $h(\theta)$ is the implementation of $\mathcal{T}(f_i)$ as in (3.5). In other words, $g(\theta)$ and $f(\theta)$ are the analytic approximations of $f(\theta)$.

We compare $g(\theta)$ and $h(\theta)$ with $f(\theta)$ in Figure 2 using different R and different range of θ. The true function $f(\theta)$ is plotted as the red curve and $g(\theta)$ as black and $h(\theta)$ as blue. As we can see, both approximations work well in the middle of the range of θ. In the first row of Figure 2, we used the

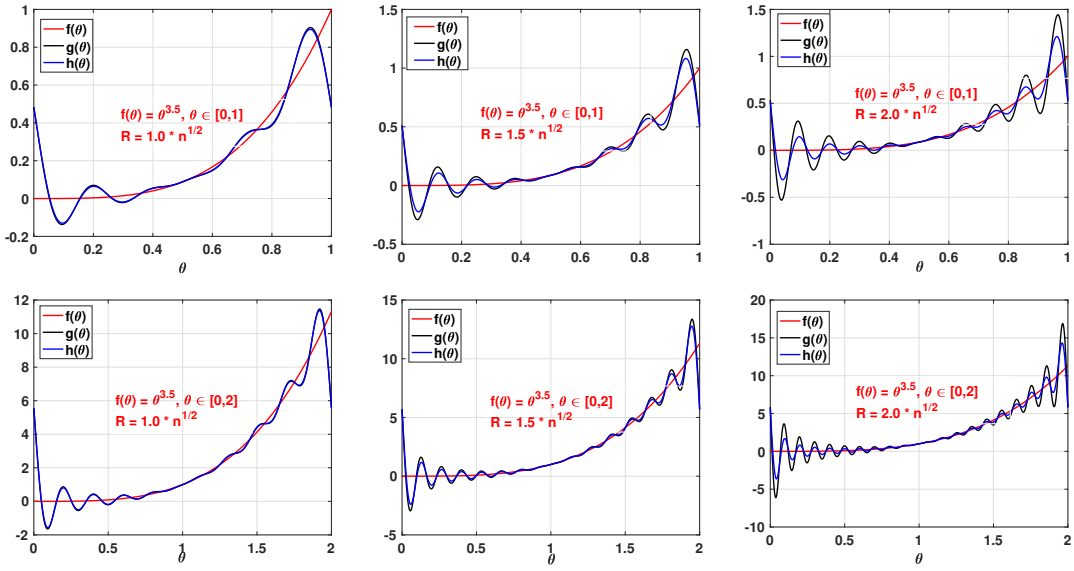

Figure 2: Analytical approximation of $f(\theta)$ with different R

range $\theta \in [0, 1]$, and both approximations work the best around 0.5. In the second row, we used the range of $\theta \in [0, 2]$, and they work the best around 1. This gives us an important practical instruction when implementing such estimators, to get the best performance, one needs to figure out a proper range of θ to implement such analytical approximations first. For instance, if θ is close to 1, clearly using range $[0, 1]$ will give poor results as these approximations itself are far from accurate. Although such phenomena are due to numerical implementations and not reflected in the analysis.

Another key observation of Figure 2 is that when R is relatively small, the approximation using the exponential function as in $g(\theta)$ is very similar to that of using the quadratic terms as in $h(\theta)$ (our estimator). One can hardly notice the difference between them in the first column of Figure 2. As R increases, one can see the clear difference between $g(\theta)$ and $h(\theta)$ near the boundary of the range. However, as the part of the approximation we use is in the middle, the difference reflected in the estimators' performance is minor as we have already seen in Figure 1. This validates the effectiveness of our approximation by replacing the exponential term using $g_{\zeta,n}$.

A.2. Choice of the constant matters

As we briefly mentioned in the main paper that our theory suggests the truncation radius $R = c^*\sqrt{n}$ and the choice of the constant c^* matters. In this section, we show numerical evidence of this. In Figure 3, we present the performance of our estimator with different choices of c^* under various

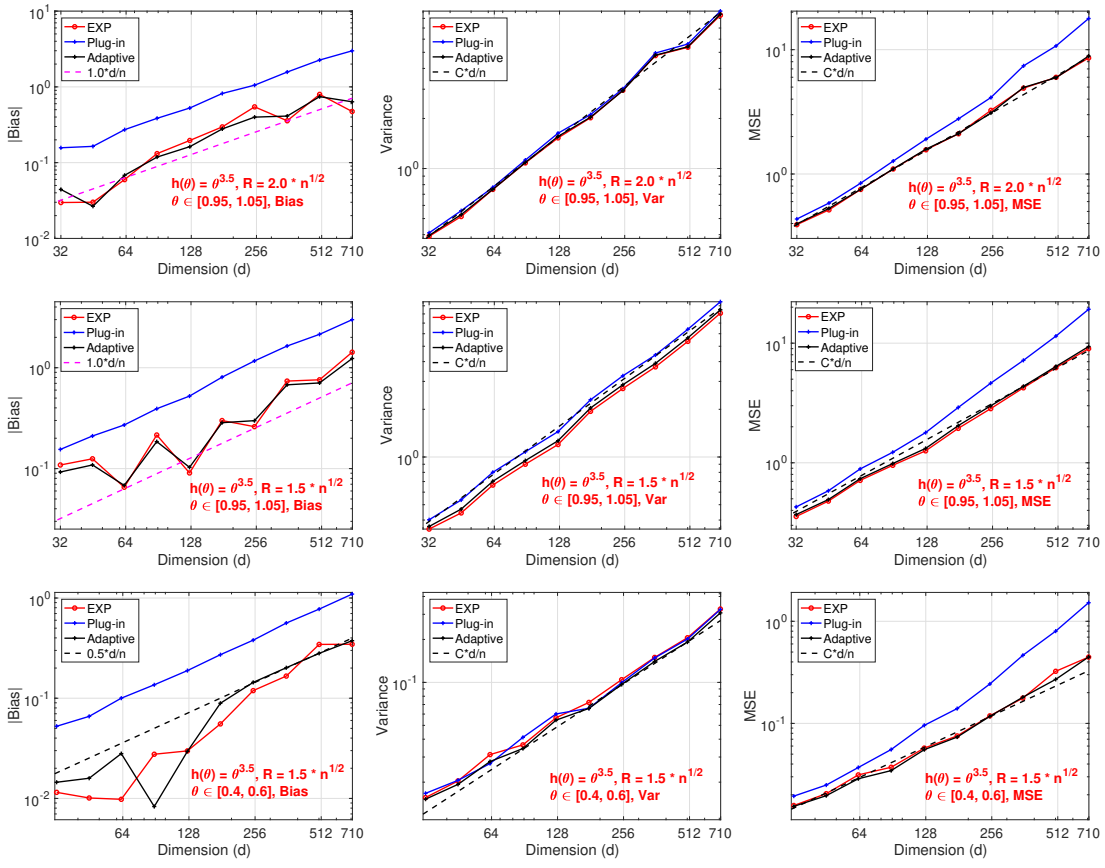

Figure 3: Different choices of R lead to different performance

settings. The difference between the first row and the second row in Figure 3 is that we use $c^* = 2.0$ in the first and $c^* = 1.5$ in the second, and we keep everything else the same. As we can see, $c^* = 1.5$ in the second row leads to smaller variance but larger bias compared with the first row with $c^* = 2.0$. As we have explained in Section 6, the black dash line in variance comparison shall be the optimal asymptotic variance and it matches that of Plug-in. Although both give the same and optimal performance on MSE, clearly $R = 2\sqrt{n}$ shall be a better choice. However, when we change the range of θ to $[0.4, 0.6]$ as shown in the third row, one can see $R = 1.5\sqrt{n}$ leads to quite different performance compared with that of when $\theta \in [0.95, 1.05]$ with $R = 1.5\sqrt{n}$. This time the variance of both EXP and Adaptive matches that of Plug-in. It behaves similarly to the case when $\theta \in [0.95, 1.05]$ with $R = 2\sqrt{n}$. Clearly, the choice of c^* matters with different θ. Currently, we don't know whether there is an optimal choice of c^* for each setting or not. If yes, a data driven procedure to determine this constant can be meaningful.

A.3. Experiments with heavy-tailed distribution

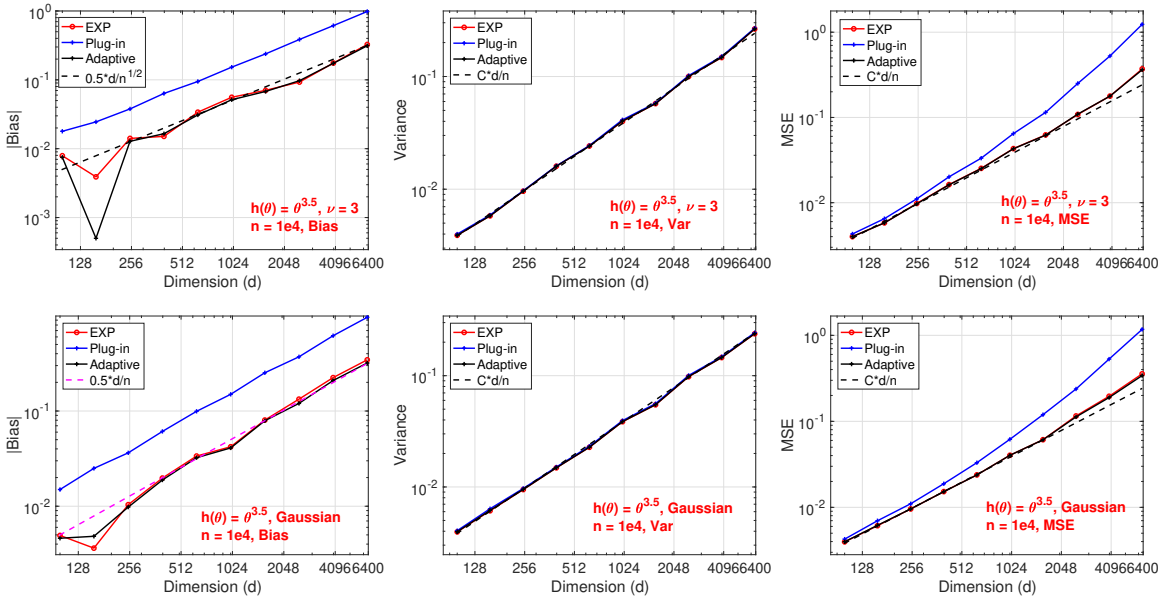

Figure 4: Experiments with different distributions of noise

As we have already shown in Section 6, our estimator works well with Student's t-distribution with 4-degrees of freedom. It is well-known that such distributions only have finite third moment. In this section, we explore more different noise distributions to test our estimator. In Figure 4, we test our estimator with Student's t-distribution with 3-degrees of freedom. In this case, the noise has only finite second moment. As we can see, in the first row of Figure 4, our estimator still works well and its performance aligns with our theory. In fact, in the second row of Figure 4, we change to Gaussian noise and keep everything else the same. The performance of our estimator under Gaussian noise is similar to that in the first row. This validates the distribution-free feature of our estimator. One should notice that under Gaussian noise, EXP typically has better performance compared with Adaptive. In Figure 4, the performances are similar due to the fact that we used the same truncation parameter R

for both and the best choice of R for EXP is not necessarily the best choice for Adaptive. In general, one can tune this R to make EXP achieve its best performance. But according to our experience, the difference is not that big.

A.4. Bias reduction leads to better confidence intervals

Table 1: Classification accuracies for naive Bayes and flexible Bayes on various data sets.

α	d	$f(\boldsymbol{\theta})$	$f(\bar{\mathbf{x}})$	$g^*(\mathbf{X})$
0.50	100	9.4025	[9.4179, 9.4261]	$[\mathbf{9.3999}, \mathbf{9.4086}]$
0.55	158	14.3505	[14.611, 14.622]	$[\mathbf{14.345}, \mathbf{14.355}]$
0.60	251	46.363	[46.588, 46.616]	$[\mathbf{46.348}, \mathbf{46.376}]$
0.65	398	36.2775	[36.326, 36.341]	$[\mathbf{36.271}, \mathbf{36.287}]$
0.70	631	57.2348	[57.319, 57.340]	$[\mathbf{57.235}, \mathbf{57.257}]$
0.75	1000	89.5941	[89.735, 89.760]	$[\mathbf{89.593}, \mathbf{89.619}]$
0.80	1585	141.3395	[141.54, 141.57]	$[\mathbf{141.33}, \mathbf{141.42}]$

In the main paper, we proved two major theorems: Theorem 2 says our estimator is universally effective in bias reduction when, and Theorem 6 says our estimator as a random variable is normally distributed around the true parameter when n is large. A direct application of such results is that we can use our estimators to build confidence intervals of the true parameter. In this section, we compare the 95%-confidence interval built by our estimator Adaptive and that of Plug-in. The data is shown in Table 1. The noise we use is generated by Student's t-distribution with 4-degrees of freedom. The confidence interval is based on 10^4 i.i.d. copies of $g^*(\mathbf{X})$ by fitting a normal distribution using MATLAB. As we can see, confidence intervals based on our estimator are accurate and always better than the ones built based on Plug-in at all levels of dimension. As one can see, the true parameters always fall outside the ones built based on Plug-in estimators which makes the proposed estimator meaningful.