Optimal Estimation of High Dimensional Smooth Additive Function Based on Noisy Observations

Fan Zhou, Ping Li
Cognitive Computing Lab
Baidu Research
10900 NE 8th St Bellevue WA 98004 USA
{fanzhou, liping11}@baidu.com

Abstract

Given $\mathbf{x}_j = \boldsymbol{\theta} + \boldsymbol{\varepsilon}_j$, $j = 1, ..., n$ where $\boldsymbol{\theta} \in \mathbb{R}^d$ is an unknown parameter and $\boldsymbol{\varepsilon}_j$ are i.i.d. Gaussian noise vectors, we study the estimation of $f(\boldsymbol{\theta})$ for a given smooth function $f : \mathbb{R}^d \to \mathbb{R}$ equipped with an additive structure. We inherit the idea from a recent work which introduced an effective bias reduction technique through iterative bootstrap and derive a bias-reducing estimator. By establishing its normal approximation results, we show that the proposed estimator can achieve asymptotic normality with a looser constraint on smoothness compared with general smooth function due to the additive structure. Such results further imply that the proposed estimator is asymptotically efficient. Both upper and lower bounds on mean squared error are proved which shows the proposed estimator is minimax optimal for the smooth class considered. Numerical simulation results are presented to validate our analysis and show its superior performance of the proposed estimator over the plug-in approach in terms of bias reduction and building confidence intervals.

1. Introduction

We consider the model

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

with $\boldsymbol{\varepsilon}_j$, $j = 1, ..., n$ 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 Gaussian random vectors $\boldsymbol{\varepsilon} \sim \mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{I}_d)$. The goal of this paper is to study the estimation of the function value $f(\boldsymbol{\theta})$ when $f : \mathbb{R}^d \to \mathbb{R}$ is a given smooth function that has an additive structure:

$$f(\boldsymbol{\theta}) := \sum_{j=1}^{d} f_j(\theta_j) \tag{1.2}$$

Especially, we are interested in the high dimensional setting where the dimension can grow with the sample size, namely, $d = n^\alpha$ with $\alpha \in (0, 1)$. Model (1.1) is ubiquitous in real world applications since it models data observations with measurement error. We refer to Carroll et al. (2006) for a thorough study on this particular topic. Meanwhile, estimation of smooth functionals of parameters in high and even infinite dimension space has a long history in the statistics community, Important references include but not limited to Levit (1976; 1978); Ibragimov and KHas'Minskii (2013); Ibragimov et al. (1986); Bickel and Ritov (1988); Nemirovski (1991; 2000); Birgé and Massart (1995); Laurent (1996); Lepski et al. (1999). Among them, two types of functionals with an additive structure are intensively studied. One is the linear functional, see Donoho and Liu (1987; 1991); Cai and Low (2005a); Klemelä and Tsybakov (2001) and the references therein. The other is the quadratic functional, see Donoho and Nussbaum (1990); Cai and Low (2005b); Klemelä (2006); Laurent and Massart (2000) and the references therein.

A natural way to approach the problem is to take the sample mean $\bar{\mathbf{x}}$ of the noisy observations and use the plug-in estimator $f(\bar{\mathbf{x}})$. Since $\bar{\mathbf{x}}$ is the maximum likelihood estimator (MLE) of $\boldsymbol{\theta}$, then one may think that $f(\bar{\mathbf{x}})$ should be good to serve the purpose. Indeed, when d is fixed, it works. However, in high dimensional regime when d is large such estimators can introduce large bias due to the factor d in the convergence rate. Such a concern acts as a major driven force of several recent works which developed new methods to effectively reduce the bias. One way to address this issue is based on an iterative bootstrap technique, see Jiao and Han (2020); Koltchinskii (2020); Koltchinskii and Zhilova (2021). The others turn to use approximation methods to replace f by its approximation to achieve de-biasing. For

Proceedings of the 38^{th} International Conference on Machine Learning, PMLR 139, 2021.

instance, Han et al. (2020) used polynomial approximations of f and studied L_r-norm estimation of a specific functional under Gaussian white noise model, and Zhou and Li (2019) used the ideas in Fourier analysis and Littlewood-Paley theory to build estimators of $f(\boldsymbol{\theta})$ under model (1.1). In other interesting related works such as Acharya et al. (2017); Hao and Orlitsky (2019), the authors studied similar problems using profile maximum likelihood estimator of $\boldsymbol{\theta}$ under discrete distribution setting. The smooth functions studied in Koltchinskii and Zhilova (2021); Zhou and Li (2019) are usually in general forms without any specific structure. Given that functionals with an additive structure are perhaps the most important ones used in machine learning such as boosting methods (Friedman, 2001) or generalized additive models (Hastie and Tibshirani, 1986), we think it is worthwhile to investigate how to use those methods to study estimation of smooth additive models with large d and what are the implications of such specific structure.

As we have already mentioned, specific additive models such as linear and quadratic functionals have been extensively studied during the past few decades. Recently we see a resurgence of interests in studying minimax theory of those topics under sparsity class and Gaussian shift model (1.1), see Collier et al. (2017; 2018). Both linear and quadratic functionals are specific smooth additive models contained in the function class we consider in this article. The results in this paper can reproduce some of their results for quadratic functionals in the so called dense regime since we don't assume any sparsity constraint. Recently Collier and Comminges (2019) studied minimax estimation of a type of general additive functionals based on Hermite polynomials and approximation theory. One common ground between their estimator and ours is that both estimators can be unbiased when f is a polynomial up to certain degree. Another line of exciting results focus on minimax estimation of nonsmooth additive models, see Cai and Low (2011); Carpentier and Verzelen (2019); Jiao et al. (2015); Wu and Yang (2016; 2019); Collier et al. (2020).

In their original work (Koltchinskii and Zhilova, 2021), the authors studied the problem over a quite general smooth function class without exploiting any specific structure of the function itself. In this article, we exploit the additive structure and apply the bias-reduction technique of iterative bootstrap introduced by Koltchinskii and Zhilova (2021) to each component function and construct an estimator for smooth additive functions with each component residing in Hölder class $\Sigma(\beta, L)$ which is a fundamental function class in nonparametric estimation. Our major contribution is on the theory front. We developed new concentration bounds to study the estimation of general smooth additive functions in high dimensional regime, which makes this work the first to study this problem. It turns out that such results can be used to reproduce some classical results in linear and quadratic functional estimation. Those normal approximation results are also new which lay the foundation for building effective confidence intervals of the true parameter.

Contributions and paper organization. The paper is organized as follows: in Section 3, we exploit the additive structure and propose an explicit formula of the estimator accordingly. Then we derive an upper bound on the bias for this estimator. As a byproduct, we show that the proposed estimator is unbiased if each component is a polynomial up to degree $\ell = \lfloor \beta \rfloor$. In Section 4, by using some truncation technique and tools in Gaussian concentration, we establish a concentration inequality which is a major tool to establish asymptotic normality. In Section 5, we prove normal approximation bounds for the estimator. Such results imply that the estimator scaled by the Fisher information for estimation of $f(\boldsymbol{\theta})$ is normally distributed around the ground truth $f(\boldsymbol{\theta})$ for large n. This kind of results can be useful to build confidence intervals in real world applications. Additionally, we provide bounds on mean squared error (MSE). Those results also show that the proposed estimator has optimal asymptotic variance implied by Cramér-Rao bound and it is asymptotically efficient. As we shall see, the convergence rates on bias and normal approximation we obtained are quite different from the existing results of general smooth class. One can benefit from such a specific additive structure of the function itself in terms of achieving bias reduction and normal approximation with much looser smoothness constraint. In Section 6, we prove a minimax lower bound which shows that when non-trivial bias reduction is introduced, i.e., $\beta \geq 2$, the proposed estimator is minimax optimal. In Section 7, numerical simulations are presented to validate our analysis. Especially, we propose solutions to the adaptation issues on the computational aspects. Our simulations show lucrative improvements of the estimator in bias reduction and MSE reduction when d is large. The confidence intervals built upon the proposed estimator is noticeably more accurate than those based on the plug-in approach.

2. Preliminaries

2.1. Notations

We use boldface uppercase letter $\mathbf{X}$ to denote a matrix and boldface lowercase letter $\mathbf{x}$ to denote a vector. We use $\|\cdot\|$ to denote the ℓ_2-norm of a vector, and $\|\cdot\|_{L^p}$ to denote the L^p-norm of a function. We use the conventional notation $\Rightarrow$ to denote weakly convergence or convergence in distribution. Throughout the paper, given nonnegative a and b, $a \lesssim b$ means that $a \leq Cb$ with a numerical constant C and $a \lesssim_L b$ means that $a \leq C(L)b$ with $C(L)$ being a constant involving L; $a \asymp b$ means that $a \lesssim b$ and $b \lesssim a$. $a \wedge b = \min\{a, b\}$ and $a \vee b = \max\{a, b\}$.

2.2. Hölder Class and Additive Structure

We first introduce Hölder class (see Tsybakov (2009)) which is a fundamental function class in nonparametric estimation.

Definition 1. *Let β and L be two positive real numbers. The Hölder class $\Sigma(\beta, L)$ on $T \subset \mathbb{R}$ is defined as the set of $\ell = \lfloor \beta \rfloor$ times differentiable functions $f : T \to \mathbb{R}$ with derivative $f^{(\ell)}$ satisfying*

$$\left| f^{(\ell)}(x) - f^{(\ell)}(x') \right| \leq L|x - x'|^{\beta - \ell}, \ \ \forall x, x' \in T. \quad (2.1)$$

The parameters β and ℓ characterize the smoothness of Hölder class $\Sigma(\beta, L)$.

Next, we introduce the function class $\mathcal{F}_d^\beta$ of our interest. It contains functions with an additive structure where each component belongs to the Hölder space $\Sigma(\beta, L)$ and the derivatives up to the order $\ell = \lfloor \beta \rfloor$ are uniformly bounded by some constant L' in the domain T.

$$\mathcal{F}_d^\beta := \Big\{ f(\boldsymbol{\theta}) = \sum_{j=1}^d f_j(\theta_j) : f_j \in \Sigma(\beta, L), \\ \text{and } \left\| f^{(k)} \right\|_{L^\infty} \leq L', \text{ for } k = 0, 1, 2, ..., \ell, \ \forall\, \boldsymbol{\theta} \in T \Big\}.$$

Note that for a given $f \in \mathcal{F}_d^\beta$ with L, L' and β being fixed constants, the value of f can be as large as the order $O(d)$. This makes the function class $\mathcal{F}_d^\beta$ larger than those considered in Koltchinskii and Zhilova (2021) in nature where both the norm of the function and the norm of its gradient vector are bounded by some constants and are independent of the dimension factor d. Clearly, here for our model, both can depend on d.

3. Bias Reduction

As we have already mentioned, a natural estimator of $f(\boldsymbol{\theta})$ under model (1.1) is the plug-in estimator $f(\bar{\mathbf{x}})$. However, such an estimator can be very inaccurate when the dimension d is large due to the introduction of large bias, see Collier et al. (2017); Zhou and Li (2019); Koltchinskii and Zhilova (2021). Thus, non-trivial bias reduction technique is needed. The idea of construction of the bias-reducing estimator in this article is that we exploit f's additive structure and apply the bias reduction technique of iterative bootstrap introduced by Koltchinskii and Zhilova (2021) to each component function. Intuitively, if the bias of each component f_j can be small enough, then the summation of the bias should also be small, so is the bias of the estimator. As long as we can control its variance well, then this bias reduction should be effective. Indeed, as we shall see in Section 5, the proposed estimator not only have small bias, but also has an optimal variance in asymptotic sense. We briefly summarize the beautiful idea of iterative bootstrap as follows: denote by $\mathcal{T}$ the following linear operator

$$\mathcal{T} g(\boldsymbol{\theta}) := \mathbb{E}_{\boldsymbol{\theta}} g(\boldsymbol{\theta}) = \mathbb{E} g(\boldsymbol{\theta} + \bar{\boldsymbol{\varepsilon}}),$$

where $\bar{\boldsymbol{\varepsilon}} = n^{-1} \sum_{j=1}^n \boldsymbol{\varepsilon}_j$, and denote by $\mathcal{I}$ as the identity operator and further denote by $\mathcal{B} := \mathcal{T} - \mathcal{I}$. To create an estimator $g(\bar{\mathbf{x}})$ of $f(\boldsymbol{\theta})$ with small bias, one wants to solve the integral equation $\mathcal{T} g(\boldsymbol{\theta}) = (\mathcal{I} + \mathcal{B}) g(\boldsymbol{\theta}) = f(\boldsymbol{\theta})$ as accurate as possible. However, solving such an integral equation itself is challenging. Instead, a finite approximation of the Neumann series $(\mathcal{I} + \mathcal{B})^{-1} = (\mathcal{I} - \mathcal{B} + \mathcal{B}^2 - \mathcal{B}^3 + \dots)$ to create the following estimator

$$f_k(\bar{\mathbf{x}}) := \sum_{j=0}^k (-1)^j \mathcal{B}^j f(\bar{\mathbf{x}}). \quad (3.1)$$

In Lemma 1, we derive an explicit formula of $\mathcal{B}^j f(\bar{\mathbf{x}})$ for our model and then use it to construct an estimator. We assume that $\bar{\boldsymbol{\varepsilon}} \sim \mathcal{N}(0, \sigma_\xi^2 \mathbf{I}_d)$ with $\sigma_\xi^2 = \sigma^2 / n$.

Lemma 1. Suppose that $f \in \mathcal{F}_d^\beta$ with $\ell = \lfloor \beta \rfloor$. Then for $k = 1, ..., \ell$

$$\mathcal{B}^k f(\boldsymbol{\theta}) = \mathbb{E}_{\tau, \boldsymbol{\xi}} \Big[\sum_{j=1}^d f_j^{(k)}(\theta_j + \sum_{i=1}^k \tau_i \xi_{ij}) \prod_{i=1}^k \xi_{ij} \Big], \quad (3.2)$$

where τ_i, $i = 1, ..., k$ are i.i.d. random variables uniformly distributed on $[0, 1]$, and $\boldsymbol{\xi}_i = (\xi_{i1}, ..., \xi_{id})^T$, $i = 1, ..., k$ are i.i.d. copies of a Gaussian random vector $\boldsymbol{\xi} \sim \mathcal{N}(\mathbf{0}, \sigma_\xi^2 \mathbf{I}_d)$. All τ_i's are independent of $\boldsymbol{\xi}_i$'s.

The proof of Lemma 1 is deferred to Appendix A.1. Once we have Lemma 1, we introduce the estimator as follows: under model (1.1), given an $f \in \mathcal{F}_d^\beta$, we define

$$f_{\ell - 1}(\bar{\mathbf{x}}) := f(\bar{\mathbf{x}}) + \sum_{k=1}^{\ell - 1} (-1)^k \mathcal{B}^k f(\bar{\mathbf{x}}). \quad (3.3)$$

Here $\mathcal{B}^k f$ is the same defined as in (3.2) with $\sigma_\xi^2 := \sigma^2 / n$. Note that here $\boldsymbol{\xi}$ has the same distribution as $\bar{\boldsymbol{\varepsilon}}$.

Remark 1. Two practical concerns raise when one implements the estimator (3.3): 1) Since $\boldsymbol{\xi}$ has the same distribution as $\bar{\boldsymbol{\varepsilon}}$, one needs to know the distribution of the noise in advance. However, such distribution is unknown in practice since we don't know the variance σ^2 of $\boldsymbol{\xi}$. This can be addressed by using the sample variance computed from the data as a surrogate. 2) There is an expectation $\mathbb{E}_{\tau, \xi} \big[\sum_{j=1}^d f_j^{(k)} \Big(\bar{x}_j + \sum_{i=1}^k \tau_i \xi_{ij} \Big) \prod_{i=1}^k \xi_{ij} \big]$ appearing in (3.3). Oftentimes, it may be difficult to derive an explicit expression of this term. Our solution is to use the empirical mean of sampling N i.i.d. copies of $\sum_{j=1}^d f_j^{(k)} \Big(\bar{x}_j + \sum_{i=1}^k \tau_i \xi_{ij} \Big) \prod_{i=1}^k \xi_{ij}$. We will have a more detailed discussion on these in Section 7.

Remark 2. When $\beta \in (1,2)$, $\ell - 1 = 0$, in this case the estimator (3.3) is simply the trivial plug-in estimator $f(\bar{\mathbf{x}})$. Only when $\beta \geq 2$, the iterative bootstrap procedure comes into the picture and non-trivial bias reduction is introduced.

In Theorem 3.1, we derive an upper bound on the bias of the proposed estimator (3.3) under model (1.1).

Theorem 3.1. *Under model (1.1), suppose that $f \in \mathcal{F}_d^\beta$ with $\beta > 1$ and $\beta \notin \mathbb{N}^*$. Consider the estimator $f_{\ell-1}(\bar{\mathbf{x}})$ with $\ell = \lfloor \beta \rfloor$ defined as in (3.3). Then the following bound on the bias holds:*

$$\left|\mathbb{E} f_{\ell-1}(\bar{\mathbf{x}}) - f(\boldsymbol{\theta})\right| \leq C(L,\beta)\sigma^\beta n^{-\beta/2} d. \tag{3.4}$$

where the constant $C(L,\beta) := \frac{L\Gamma(\beta-\ell+3/2)(\sqrt{2})^\beta}{(\sqrt{\pi})^\ell(\beta-\ell+1)}$ and $\Gamma(\cdot)$ is the gamma function.

The proof of Theorem 3.1 is deferred to Appendix A.2. As we shall see in Section 7.2, the numerical simulation results show this estimator leads substantial improvements in bias reduction compared with $f(\bar{\mathbf{x}})$.

Remark 3. To understand bound (3.4), the rate on the bias is of the order $O(n^{-\beta/2}d)$ given that σ, L and β are fixed constants. It differs from the typical bound on the bias $O((d/n)^{\beta/2})$ for general functional estimation in Gaussian shift model, see Koltchinskii and Zhilova (2021). The d factor in bound (3.4) is due to its additive structure. Basically, if the components f_j's are very similar to each other, then the summation of bias for estimation of each $f_j(\theta_j)$ should be linear in d.

Remark 4. On the other hand, as one can see unlike the bound for general smooth function, the specific additive structure decouples the smoothness index β from the dimension parameter d. This makes bias reduction more obvious for functions with large smoothness index β when d is large. Indeed, as we shall see in Section 7, when α is close to 1, larger β can contribute to obvious reduction on MSE due to better bias correction. In other words, when f is sufficiently smooth, large value of d will not cause as much trouble as those in general functional estimation in high dimensions. This can be seen as a benefit of the additive structure.

The following corollary as a byproduct of Theorem 3.1 shows that if f is a multivariate polynomial of degree at most ℓ, then the estimator defined in (3.3) is an unbiased estimator of $f(\boldsymbol{\theta})$.

Corollary 3.1. Suppose that f has a structure as in (1.2) and each component f_j is a polynomial of order at most ℓ for each j. Then $f_{\ell-1}(\bar{\mathbf{x}})$ defined as in (3.3) is an unbiased estimator of $f(\boldsymbol{\theta})$.

Remark 5. Two classical examples of additive functionals which are extensively studied, namely linear functional, i.e., $f(\boldsymbol{\theta}) := \sum_{j=1}^d \theta_j$ and quadratic functional, i.e., $f(\boldsymbol{\theta}) := \sum_{j=1}^d \theta_j^2$, fall into this category. This indicates that the proposed estimator is in a general form. In fact, if each component $f_j(\theta_j)$ is a higher order polynomial, this estimator suggests that one only need to do a few more steps of bootstrap in order to achieve unbiasedness. As we shall see in later sections, the results we established for this estimator can reproduce those minimax rates on MSE for both linear and quadratic functionals estimation.

4. Concentration Inequalities

As we have mentioned, when we turn to iterative bootstrap to achieve bias reduction for each component function f_j, we still need to show that the variance of the proposed estimator can be well controlled so that we are not sacrificing variance for smaller bias. In this section we resolve this issue by proving a concentration inequality which essentially implies that the estimator is well concentrated around its mean so its variance is still well controlled.

For a given function $g : \mathbb{R}^d \to \mathbb{R}$, we consider the first order Taylor expansion of $g(\bar{\mathbf{x}})$ around $\boldsymbol{\theta}$, and get

$$g(\bar{\mathbf{x}}) = g(\boldsymbol{\theta} + \bar{\boldsymbol{\varepsilon}}) = g(\boldsymbol{\theta}) + \left\langle \nabla g(\boldsymbol{\theta}), \bar{\boldsymbol{\varepsilon}} \right\rangle + S_g(\boldsymbol{\theta}; \bar{\boldsymbol{\varepsilon}}) \tag{4.1}$$

where $S_g(\boldsymbol{\theta}; \bar{\boldsymbol{\varepsilon}})$ denotes the remainder of $g(\bar{\mathbf{x}})$. $g(\boldsymbol{\theta}) + \left\langle \nabla g(\boldsymbol{\theta}), \bar{\boldsymbol{\varepsilon}} \right\rangle$ is the linear approximation of $g(\bar{\mathbf{x}})$ around $\boldsymbol{\theta}$, which is clearly a Gaussian random variable. In Theorem 4.1, we derive concentration inequalities of the remainder $S_g(\boldsymbol{\theta}; \bar{\boldsymbol{\varepsilon}})$ around its mean $\mathbb{E} S_g(\boldsymbol{\theta}; \bar{\boldsymbol{\varepsilon}})$. These results are part of the major contributions of this work and new analysis is developed to adapt to the additive structure. Such concentration bounds are crucial for us to derive normal approximation bounds and establish asymptotic normality of our estimator.

Theorem 4.1. *Under model (1.1), assume that $f \in \mathcal{F}_d^\beta$ with $\beta > 1$. Consider the estimator $f_{\ell-1}(\bar{\mathbf{x}})$ with $\ell = \lfloor \beta \rfloor$ defined as in (3.3). Then there exists a numerical constant C_1^* such that for all $t \geq 1$, with probability at least $1 - e^{-t}$, for any $\beta \geq 2$,*

$$\begin{aligned} &\left| S_{f_{\ell-1}}(\boldsymbol{\theta}; \bar{\boldsymbol{\varepsilon}}) - \mathbb{E} S_{f_{\ell-1}}(\boldsymbol{\theta}; \bar{\boldsymbol{\varepsilon}}) \right| \leq \\ &C_1^*(L \vee L')\sigma^2 \Big(\sqrt{\frac{d}{n}} \bigvee \sqrt{\frac{t}{n}} \Big) \sqrt{\frac{t}{n}}, \end{aligned} \tag{4.2}$$

and with the same probability and some numerical constant C_2^*, for any $1 < \beta < 2$,

$$\begin{aligned} &\left| S_{f_{\ell-1}}(\boldsymbol{\theta}; \bar{\boldsymbol{\varepsilon}}) - \mathbb{E} S_{f_{\ell-1}}(\boldsymbol{\theta}; \bar{\boldsymbol{\varepsilon}}) \right| \leq \\ &C_2^*(L/\beta)\sigma^\beta \Big(\sqrt{\frac{d}{n^{\beta-1}}} \bigvee \sqrt{\frac{d^{(2-\beta)} t^{(\beta-1)}}{n^{\beta-1}}} \Big) \sqrt{\frac{t}{n}}. \end{aligned} \tag{4.3}$$

The proof of Theorem 4.1 is deferred to Appendix A.5. As a major result, the main idea of the proof is based on a truncation technique developed in Koltchinskii and Lounici (2016)

and tools in Gaussian isoperimetric inequality, see Giné and Nickl (2016). The key challenge in applying Gaussian concentration inequality is that $S_{f_{\ell-1}}(\boldsymbol{\theta};\bar{\boldsymbol{\varepsilon}})$ may not be a Lipschitz function of the random vector $\bar{\boldsymbol{\varepsilon}}$. As a result, we establish a key lemma with delicate analysis which is a major contribution of this proof that shows a modified version of $S_{f_{\ell-1}}(\boldsymbol{\theta};\bar{\boldsymbol{\varepsilon}})$ with truncation is Lipschitz continuous. Then standard tools in Gaussian concentration can be applied.

Remark 6. When $\beta \geq 2$, the concentration bound (4.2) is of the order $o_P(n^{-1/2})$ with $d = n^\alpha$ for all $\alpha \in (0,1)$. It shows that the remainder is well concentrated around its mean and has negligible effects in asymptotic sense as long as $d = o(n)$.

Remark 7. When $\beta \in (1,2)$, the concentration bound (4.3) indicates that rate $o(n^{-1/2})$ doesn't always hold. Especially, when $d = n^\alpha$ and take $t \asymp \log n$, bound (4.3) is of the order $O_P(d^{1/2}/n^{\beta/2})$. To make it of the order $o(n^{-1/2})$, we need $\beta > 1 + \alpha$. On the other hand, in the case $\|\nabla f(\boldsymbol{\theta})\| \asymp \sqrt{d}$, when both sides of bound (4.3) are scaled by $\sigma\|\nabla f(\boldsymbol{\theta})\|/\sqrt{n}$, the right hand side of (4.3) still goes to zero as $n \to \infty$ with $\beta \in (1,2)$. As we shall see in Section 5, the proposed estimator has an asymptotic standard deviation $\sigma\|\nabla f(\boldsymbol{\theta})\|/\sqrt{n}$. As long as $\|\nabla f(\boldsymbol{\theta})\| \asymp \sqrt{d}$, it is not worrisome.

5. Normal Approximation

In this section, we prove the normal approximation results. It is shown that the estimator (3.3) is normally distributed around the true parameter $f(\boldsymbol{\theta})$ when n is large enough. This kind of result is of vital importance in both theory and applications since it provides theoretical guarantee to build effective confidence intervals of the true parameter using the estimator. As we shall see in Section 7.3, our simulations show that confidence intervals built based on estimator (3.3) are very accurate at all dimension levels and noticeably better than those built based on the plug-in estimates.

For a given differentiable function $\psi : \mathbb{R}^d \mapsto \mathbb{R}$ and random vector $\boldsymbol{\xi} \sim \mathcal{N}(0, \sigma_\xi^2 \mathbf{I}_d)$, we denote by $\sigma^2_{\psi,\boldsymbol{\xi}}(\boldsymbol{\theta}) := \sigma_\xi^2 \|\nabla\psi(\boldsymbol{\theta})\|^2$. Note that based on one data point $\mathbf{x} = \boldsymbol{\theta} + \boldsymbol{\xi}$, the Fisher information is $I_{\boldsymbol{\theta}} = \mathbf{I}_d/\sigma_\xi^2$ for the estimation of $\boldsymbol{\theta}$ and $I_{\psi(\boldsymbol{\theta})} = 1/\sigma^2_{\psi,\boldsymbol{\xi}}(\boldsymbol{\theta})$ for the estimation of $\psi(\boldsymbol{\theta})$.

In Theorem 5.1 below, we use the results proved in previous sections to establish a normal approximation bound of the estimator (3.3) when $\beta \geq 2$. Such results characterize the convergence rate to a standard normal distribution of our estimator after proper centering and rescaling. The proof of Theorem 5.1 is deferred to Appendix A.7.

Theorem 5.1. *Under model (1.1), assume that $f \in \mathcal{F}_d^\beta$ with $\beta \geq 2$. Consider the estimator $f_{\ell-1}(\bar{\mathbf{x}})$ with $\ell = \lfloor\beta\rfloor$ defined as in (3.3). Then the following normal approximation bound holds:*

$$\sup_{x\in\mathbb{R}} \Big|\mathbb{P}_{\boldsymbol{\theta}}\Big\{\frac{\sqrt{n}\big(f_{\ell-1}(\bar{\mathbf{x}}) - f(\boldsymbol{\theta})\big)}{\sigma\|\nabla f(\boldsymbol{\theta})\|} \leq x\Big\} - \mathbb{P}\{Z \leq x\}\Big| \leq C_3^*(\beta, L)\|\nabla f(\boldsymbol{\theta})\|^{-1}\Big(\frac{\sigma^{\beta-1}d}{n^{(\beta-1)/2}} \bigvee \sigma\sqrt{\frac{d\log(n/\sigma^2)}{n}}\Big) \tag{5.1}$$

where $C_3^*(\beta, L) := \Big(L \vee L' \vee \frac{L\Gamma(\beta-\ell+3/2)(\sqrt{2})^\beta}{(\sqrt{\pi})^\ell(\beta-\ell+1)}\Big)$ and Z is a standard normal random variable $Z \sim \mathcal{N}(0,1)$. Especially,

$$\frac{\sqrt{n}\cdot\mathbb{E}_{\boldsymbol{\theta}}^{1/2}\big(f_{\ell-1}(\bar{\mathbf{x}}) - f(\boldsymbol{\theta})\big)^2}{\sigma\|\nabla f(\boldsymbol{\theta})\|} \lesssim_{\beta,L,L'} 1 + \|\nabla f(\boldsymbol{\theta})\|^{-1}\Big(\sigma\sqrt{\frac{d}{n}} \bigvee \frac{\sigma^{\beta-1}d}{n^{(\beta-1)/2}}\Big). \tag{5.2}$$

Remark 8. Bound (5.1) in Theorem 5.1 indicates that when the dimension $d = n^\alpha$ with all $0 < \alpha < 1$, and $\|\nabla f(\boldsymbol{\theta})\| = \sqrt{\sum_{j=1}^d f_j'^2(\theta_j)} \asymp \sqrt{d}$, then for $\beta \geq 2$,

$$\frac{\sqrt{n}\big(f_{\ell-1}(\bar{\mathbf{x}}) - f(\boldsymbol{\theta})\big)}{\sigma\|\nabla f(\boldsymbol{\theta})\|} \Rightarrow \mathcal{N}(0,1) \text{ as } n \to \infty,$$

which means that the estimator $f_{\ell-1}(\bar{\mathbf{x}})$ is always normally distributed around the true parameter $f(\boldsymbol{\theta})$ for any $\alpha \in (0,1)$ in an asymptotic sense. Note that such normal approximation results are not necessarily true for general smooth functionals where the smoothness needs to be above some threshold related to the dimension constraint, see Zhou and Li (2019); Koltchinskii and Zhilova (2021). In short, for smooth functionals with an additive structure, the constraint to achieve asymptotic normality is much looser. We recognize this as a benefit of the additive structure.

Remark 9. Given that $n/\sigma^2\|\nabla f(\boldsymbol{\theta})\|^2$ is the Fisher information for the estimation of $f(\boldsymbol{\theta})$, bound (5.2) indicates that the proposed estimator has an optimal asymptotic variance according to Cramér-Rao bound. This not only indicates that its bias reduction doesn't blow off its variance, but also shows that the proposed estimator is actually asymptotically efficient.

Remark 10. Recall that for any $f \in \mathcal{F}_d^\beta$, by definition $\|\nabla f(\boldsymbol{\theta})\| \lesssim_{L'} \sqrt{d}$. Then bound in (5.2) shows an upper bound on the MSE for $\beta \geq 2$.

$$\mathbb{E}_{\boldsymbol{\theta}}\big(f_{\ell-1}(\bar{\mathbf{x}}) - f(\boldsymbol{\theta})\big)^2 \lesssim_{\beta,L,L'} \sigma^2 d/n$$

In Section 6, we show that this bound is actually minimax optimal for $\beta \geq 2$. As we shall see in Section 7.2, the simulation results show this estimate is very accurate and validates our analysis.

Note that the first term $\sigma^{\beta-1}d/n^{(\beta-1)/2}$ on the right hand side in bound (5.1) comes from the bias. Recall from Corollary 3.1, if $f(\boldsymbol{\theta})$ is a polynomial of degree up to ℓ, $f_{\ell-1}(\bar{\mathbf{x}})$

is an unbiased estimator of $f(\boldsymbol{\theta})$. As a result, this term will disappear. Then for polynomials of degree up to order $\ell \geq 2$ bound (5.2) turns into

$$\mathbb{E}_{\boldsymbol{\theta}}\big(f_{\ell-1}(\bar{\mathbf{x}})-f(\boldsymbol{\theta})\big)^2 \lesssim_{\beta,L,L'} (\sigma^2\|\nabla f(\boldsymbol{\theta})\|^2/n \vee \sigma^4 d/n^2).$$

Especially, when $\beta = 2$, this bound reproduces the minimax optimal rate on MSE for quadratic functional estimation.

Corollary 5.1. Under model (1.1), assume that $f(\boldsymbol{\theta})$ is the quadratic functional, and $\boldsymbol{\theta} \in \boldsymbol{\Theta} := \{\boldsymbol{\theta} | \|\boldsymbol{\theta}\| \leq \tau\}$. Then there exists some absolute constant $C > 0$ such that

$$\mathbb{E}_{\boldsymbol{\theta}}\big(f_{\ell-1}(\bar{\mathbf{x}}) - f(\boldsymbol{\theta})\big)^2 \leq C\big(\sigma^2\tau^2/n \vee \sigma^4 d/n^2\big) \quad (5.3)$$

Similar result can be found in Collier et al. (2017) and it is showed in that paper this rate is indeed minimax optimal, so we omit the tedious proofs here.

Now we switch to the case $\beta \in (1, 2)$. Note that in this case $f_{\ell-1}(\bar{\mathbf{x}})$ becomes the trivial plug-in estimator $f(\bar{\mathbf{x}})$. The proof of Theorem 5.2 is deferred to Appendix A.8.

Theorem 5.2. *Under model (1.1), assume that $f \in \mathcal{F}_d^\beta$ with $\beta \in (1, 2)$. Consider the plug-in estimator $f(\bar{\mathbf{x}})$. Then the following normal approximation bound holds:*

$$\begin{aligned}&\sup_{x\in\mathbb{R}}\left|\mathbb{P}_{\boldsymbol{\theta}}\left\{\frac{\sqrt{n}\big(f(\bar{\mathbf{x}}) - f(\boldsymbol{\theta})\big)}{\sigma\|\nabla f(\boldsymbol{\theta})\|} \leq x\right\} - \mathbb{P}\{Z \leq x\}\right| \\ &\leq C_4^*(\beta, L)\|\nabla f(\boldsymbol{\theta})\|^{-1}\frac{\sigma^{\beta-1}d}{n^{(\beta-1)/2}}\end{aligned} \quad (5.4)$$

where $C_4^*(\beta, L) := \left(L \vee L' \vee \frac{L\Gamma(\beta-\ell+3/2)(\sqrt{2})^\beta}{(\sqrt{\pi})^\ell(\beta-\ell+1)}\right)$ and Z is a standard normal random variable $Z \sim \mathcal{N}(0, 1)$ and C_4^* is a numerical constant. Especially,

$$\begin{aligned}&\frac{\sqrt{n}\cdot\mathbb{E}_{\boldsymbol{\theta}}^{1/2}\big(f(\bar{\mathbf{x}}) - f(\boldsymbol{\theta})\big)^2}{\sigma\|\nabla f(\boldsymbol{\theta})\|} \\ &\lesssim_{\beta,L,L'} 1 + \|\nabla f(\boldsymbol{\theta})\|^{-1}\frac{\sigma^{\beta-1}d}{n^{(\beta-1)/2}}.\end{aligned} \quad (5.5)$$

Remark 11. Bound (5.4) in Theorem 5.2 indicates that when $d = n^\alpha$ and $\|\nabla f(\boldsymbol{\theta})\| \asymp \sqrt{d}$ the smoothness index $\beta > 1 + \alpha$ can guarantee asymptotic normality of $f(\bar{\mathbf{x}})$. Currently, we don't know whether this threshold $1 + \alpha$ on smoothness is sharp or not. However, we are inclined to believe it is necessary at least for $f(\bar{\mathbf{x}})$ since such threshold comes from the bias. In the situation each component f_j's are almost the same, it seems that the d factor in the bias is necessary.

Remark 12. Bound (5.5) provides an upper bound on the MSE. As we can see, because $\|\nabla f(\boldsymbol{\theta})\| \lesssim_{L'} \sqrt{d}$,

$$\mathbb{E}_{\boldsymbol{\theta}}\big(f(\bar{\mathbf{x}}) - f(\boldsymbol{\theta})\big)^2 \lesssim_{\beta,L,L'} \big(\sigma^2 d/n \vee \sigma^{2\beta} d^2/n^\beta\big).$$

Especially, when $d = n^\alpha$ and $\beta > 1 + \alpha$, it implies that $\mathbb{E}_{\boldsymbol{\theta}}\big(f(\bar{\mathbf{x}}) - f(\boldsymbol{\theta})\big)^2 \lesssim_{\beta,L,L'} \sigma^2 d/n$.

As a byproduct of Theorem 5.2, it can reproduce the minimax rate on MSE for linear functional estimation. Because the term $\sigma^{2\beta} d^2/n^\beta$ is introduced by the bias, and Corollary 3.1 indicates that when $f(\boldsymbol{\theta})$ is a linear functional, i.e., $\beta = \ell = 1$, it can be dropped. So we summarize it in the following corollary without repeating the tedious proof.

Corollary 5.2. Under model (1.1), assume that $f(\boldsymbol{\theta})$ is a linear functional. Then there exists some absolute constant $C' > 0$ such that

$$\mathbb{E}_{\boldsymbol{\theta}}\big(f(\bar{\mathbf{x}}) - f(\boldsymbol{\theta})\big)^2 \leq C'\sigma^2 d/n. \quad (5.6)$$

We refer to Collier et al. (2017) for a thorough study on this particular topic in case the reader is interested.

6. Minimax Lower Bound

We establish a minimax lower bound on MSE which shows that the proposed estimator $f_{\ell-1}(\bar{\mathbf{x}})$ is minimax optimal when non-trivial bias reduction is introduced, namely $\beta \geq 2$. Without loss of generality, we assume that the domain of $\mathcal{F}_d^\beta$ is $\boldsymbol{\Theta}$. Note that $\boldsymbol{\Theta}$ can be bounded or unbounded.

Theorem 6.1. *Under model (1.1), assume that $\beta \geq 1$. Then there exists an absolute constant $c_1 > 0$ such that for any integer $d \geq 1$, the following minimax lower bound holds:*

$$\sup_{f\in\mathcal{F}_d^\beta}\inf_{\widehat{T}}\sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}}\mathbb{E}_{\boldsymbol{\theta}}\big(\widehat{T} - f(\boldsymbol{\theta})\big)^2 \geq c_1\sigma^2 d/n. \quad (6.1)$$

where the infimum is taken among all estimators of $f(\boldsymbol{\theta})$.

The proof is deferred to Appendix A.9 and is based on the standard Le Cam's two point method, see LeCam (1973).

Remark 13. Note that in Theorem 5.1 we showed that under model (1.2), for a given $f \in \mathcal{F}_d^\beta$ with $\beta \geq 2$, the following bound holds on MSE

$$\mathbb{E}_{\boldsymbol{\theta}}\big(f_{\ell-1}(\bar{\mathbf{x}}) - f(\boldsymbol{\theta})\big)^2 \leq C'(\beta, L, L')\sigma^2 d/n.$$

which matches the bound in (6.1). This shows that the proposed estimator is actually minimax optimal under our model for all $\beta \geq 2$.

7. Numerical Simulation

We present numerical simulation results of estimator (3.3) and compare its performance with that of plug-in estimator under model (1.1). The unknown parameters $\boldsymbol{\theta} \in \mathbb{R}^d$ are randomly generated that yield a uniform distribution over $[0.2, 0.4]^d$ for different dimension parameter d. We set $\sigma = 1$ and $n = 10^4$. 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 distribution of $\boldsymbol{\varepsilon}$ we use is $\boldsymbol{\varepsilon} \sim \mathcal{N}(\mathbf{0}, \sigma^2\mathbf{I}_d)$. The additive function we use has a homogeneous Hölder

structure: $f(\boldsymbol{\theta}) := \sum_{j=1}^{d} \theta_j^{\beta}$. Note that $h(\theta) = \theta^{\beta}$ with $\theta \in [0,1]$ belongs to the Hölder class $\Sigma(\beta, L)$. We denote the estimator in (3.3) by IB-Estimator (iterative bootstrap).

7.1. Adaptation

As we briefly mentioned in Section 3, there are two practical issues when it comes to implementation. Firstly, there is an adaptive estimation issue. Namely, to adopt iterative bootstrap, one has to know the variance of the noise to generate new samples of $\boldsymbol{\xi}$. However, in reality we usually don't know σ^2 in advance. This issue can be solved by using the sample variance computed from the observations since all coordinates $x_j^{(i)}$'s of $\mathbf{x}_j$ still have the variance σ^2. To be more specific, one can compute the diagonal of sample covariance matrix using the observations $\mathbf{x}_j$, $j = 1, \ldots, n$ as estimates of σ^2. Namely,

$$\widehat{\boldsymbol{\Sigma}} = \mathrm{Diag}\Big\{ \frac{1}{n(n-1)} \sum_{j=1}^{n} \big(\mathbf{x}_j - \bar{\mathbf{x}}\big) \cdot \big(\mathbf{x}_j - \bar{\mathbf{x}}\big)^T \Big\}.$$

where $\mathrm{Diag}\{\cdot\}$ takes the diagonal part of a matrix. Denote by $\widehat{\sigma}_i^2$ as the sample variance computed from the i-th coordinate, each would serve as an estimate of σ^2. In fact, if all coordinates' variance are the same, one can further improve the estimate by taking the average of $\widehat{\sigma}_i^2$, $i = 1, \ldots, d$, i.e.

$$\widehat{\sigma}^2 := d^{-1} \sum_{i=1}^{d} \widehat{\sigma}_i^2. \tag{7.1}$$

Since under our model, all coordinates are still independent, this implies that $\widehat{\sigma}_i^2$, $i = 1, \ldots, d$ are independent estimators of σ^2. Together with $d = n^{\alpha}$, it implies the approximation error of σ^2 by $\widehat{\sigma}^2$ would be of a much smaller order than $O(n^{-1/2})$. In other words, replacing σ^2 by $\widehat{\sigma}^2$ results in an accurate estimate and should not be worrisome in asymptotic sense. As we shall see from the simulation results, performance of the proposed estimator using the sample variance is very similar to that when the variance σ^2 is given, both of which are better than the plug-in approach. To draw a conclusion, this issue can be addressed by using the sample variance computed from data. We didn't observe obvious performance degradation of this data-driven solution.

Secondly, we may not have an explicit formula of the expectation in (3.3), i.e., the term

$$\mathbb{E}_{\tau, \boldsymbol{\xi}} \Big[\sum_{j=1}^{d} f_j^{(k)} \Big(\bar{x}_j + \sum_{i=1}^{k} \tau_i \xi_{ij} \Big) \prod_{i=1}^{k} \xi_{ij} \Big] \tag{7.2}$$

is often difficult to compute. Again, we propose a simple solution to this. Recall that in (3.3), to compute (7.2) we need to sample $\tau(k) := \{\tau_i : i = 1, \ldots, k\}$ and $\boldsymbol{\xi}(k) := \{\boldsymbol{\xi}_i : i = 1, \ldots, k\}$. Then our idea is to sample N i.i.d. copies of $(\tau(k), \boldsymbol{\xi}(k))$ and use them to get N i.i.d. copies of $f_{IB} := \sum_{j=1}^{d} f_j^{(k)} \Big(\bar{x}_j + \sum_{i=1}^{k} \tau_i \xi_{ij} \Big) \prod_{i=1}^{k} \xi_{ij}$. Then we simply replace (7.2) by the sample average of f_{IB}. Put simply, we are taking the advantage of law of large numbers to replace the expectation by the sample mean. The next question is how large N should be used to give good approximations. Unlike our data driven solution to the previous issue where almost no performance degradation can be observed, this solution's performance varies with different choice of N. Intuitively, the larger the N is, the better the performance. Recall that Baum and Katz (1965) showed the convergence rate of law of large numbers of N square-integrable i.i.d. random variables is roughly $O(1/N^{1/2})$. Then we should expect the bias of our estimator using the sample mean is bounded by $O(N^{-1/2} \vee dn^{-\beta/2})$. Thus in general, this means when we take $N = n$, it should be good enough to give accurate approximations. We will illustrate this through simulation results in Section 7.4.

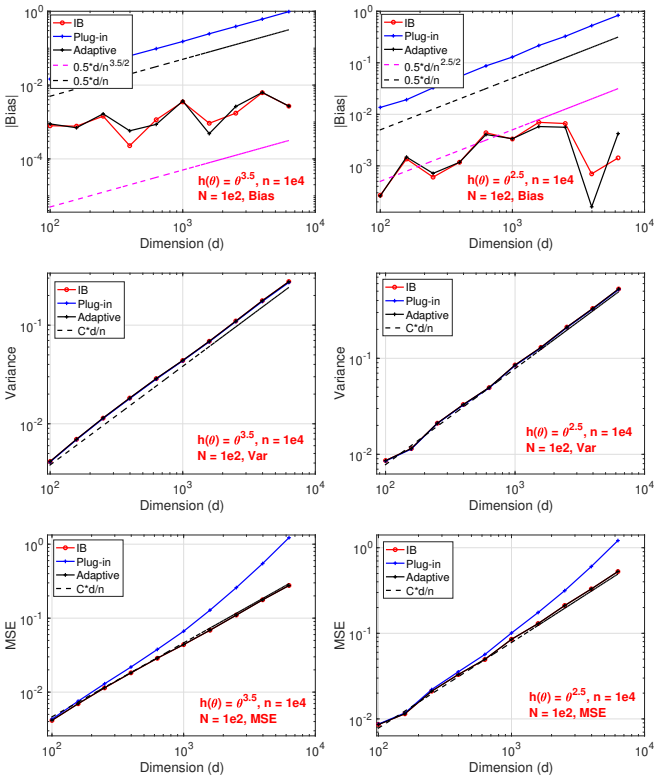

Figure 1. Comparison between IB and Plug-in estimator on bias, variance, and MSE.

7.2. Bias, variance, and MSE comparison

In this section, we compare the performance of estimator (3.3) (IB) and adaptive estimator using sample variance (Adaptive) with plug-in estimator $f(\bar{\mathbf{x}})$ (Plug-in). We set sample size $n = 10^4$ and $N = 10^2$ to approximate (7.2).

We choose two different base functions of f to test the performance. One is $\beta_1 = 2.5$ with $\ell_1 = 2$ and the other is $\beta_2 = 3.5$ with $\ell_2 = 3$. One should notice that $h_2(\theta) = \theta^{3.5}$ belongs to the Hölder class with smoothness $\beta = 3.5$ which has higher order smoothness than $h_1(\theta) = \theta^{2.5}$. The difference reflects on the implementation of estimator (3.3) where one extra step is added in the case $\beta_2 = 3.5$. The bias comparison are plotted in the first row of Figure 1. The expectation of each estimator is simulated by averaging the outcome of 10^4 independent runs. As we can see, for both cases IB and Adaptive are very effective in bias reduction compared with the Plug-in. The purple dash lines are plotted as the upper bounds on the bias derived in Theorem 3.1. As we can see, even we only used $N = 100$, the estimators still work well. Especially, when for the case $\beta = 2.5$, the bias already matches the bound. As for the case $\beta = 3.5$, one needs to increase N to further increase bias reduction to an intended level. Another observation is that when the true variance σ^2 is replaced by the sample variance $\widehat{\sigma}^2$, there is almost no performance degradation by comparing IB and Adaptive.

Another aspect we are interested in is the performance of variance and MSE of estimator (3.3). We simulated the variance and MSE from 10^4 independent runs. The variance comparisons are plotted in the middle row of Figure 1 for both cases. As we have shown in Section 5, the proposed estimator has optimal variance in asymptotic sense. As we can see, both IB and Adaptive almost has the same variance as Plug-in's and aligns the optimal variance line . Given this and its advantages of bias reduction over Plug-in, it shows estimator (3.3) is superior. This is further reflected in the bottom row of Figure 1 which show that when the dimension d is large, estimator (3.3) has very obvious reduction in MSE contributed by improvements in bias reduction. Meanwhile, one should notice that the black dash lines are plotted as $\sigma\|\nabla f(\boldsymbol{\theta})\|^2/n$ in both Figures in the right column which is the optimal variance according to Cramér-Rao bound. As we can see, for both cases the bound on MSE we computed in Theorem 5.1 which is $O(d/n)$ matches MSE of IB and Adaptive. Recall from Section 6, this bound is also the minimax lower bound. This result experimentally verified that the proposed estimator is minimax optimal.

7.3. Normal approximation and confidence interval

To test normal approximation, we collect the estimates of a fixed underlying parameter $f(\boldsymbol{\theta})$ from 10^4 independent runs. We use the MATLAB built-in function $\mathbf{histfit}()$ to draw the histograms and use $\mathbf{fitdist}()$ to fit the histograms into a normal distribution. The corresponding data can be found in supplementary material. Figure 2 displays the histograms with a fitted normal curve for both $\beta_1 = 2.5$ and $\beta_2 = 3.5$ with large value of α which makes the dimension d large. As we shall see for both β, even when α is large

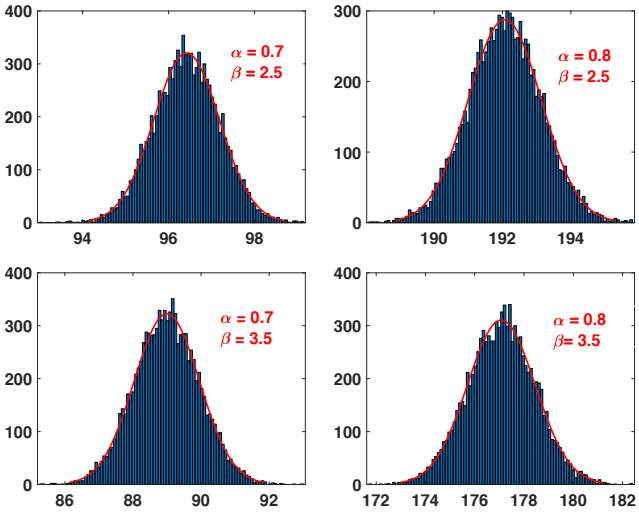

Figure 2. Histogram fit of normal curve for $f_{\ell-1}(\bar{\mathbf{x}})$ with large α

and close to 1, the normal approximations are still very accurate. This aligns well with our theory in Theorem 5.1 which suggests that asymptotic normality always holds for any $\alpha \in (0, 1)$ as long as $\beta \geq 2$ and $\|\nabla f(\boldsymbol{\theta})\| \asymp \sqrt{d}$. Note that these results don't hold for general functions without an additive structure. Typically, for a general f with the same level of smoothness, normal approximation will fail when α stays above certain threshold, see Zhou and Li (2019); Koltchinskii and Zhilova (2021).

One important application of normal approximation in practice is to use the estimator to build confidence intervals of the true parameter. Thus we show the 95% confidence intervals for estimation of $f(\boldsymbol{\theta})$ from the fitted normal models in Table 1 and Table 2 with $\beta = 3.5$ and $\beta = 2.5$, respectively. As we can see, confidence intervals based on estimator (3.3) are accurate and always better than the ones built based on the plug-in approach at all levels of dimension. In fact, the true parameters always fall outside the ones built based on the plug-in estimators.

Table 1. 95% Confidence Interval with $\beta = 3.5$

α	$f(\boldsymbol{\theta})$	$f(\bar{\mathbf{x}})$	$f_{\ell-1}(\bar{\mathbf{x}})$
0.40	11.715	[11.770, 11.784]	[11.709, 11.723]
0.45	16.142	[16.217, 16.234]	[16.134, 16.150]
0.50	23.663	[23.776, 23.796]	[23.654, 23.673]
0.55	30.831	[30.986, 31.008]	[30.819, 30.842]
0.60	46.363	[46.588, 46.616]	[46.348, 46.376]
0.65	61.296	[61.592, 61.623]	[61.262, 61.294]
0.70	89.001	[89.448, 89.486]	[88.976, 89.013]
0.75	123.860	[124.513, 124.558]	[123.85, 123.895]
0.80	177.093	[177.99, 178.044]	[177.05, 177.103]

Table 2. 95% Confidence Interval with $\beta = 2.5$

α	$f(\boldsymbol{\theta})$	$f(\bar{\mathbf{x}})$	$f_{\ell-1}(\bar{\mathbf{x}})$
0.40	12.582	[12.604, 12.615]	[12.576, 12.586]
0.45	16.655	[16.691, 16.703]	[16.652, 16.664]
0.50	25.677	[25.725, 25.741]	[25.668, 25.683]
0.55	34.418	[34.489, 34.507]	[34.409, 34.427]
0.60	49.122	[49.223, 49.245]	[49.111, 49.133]
0.65	67.261	[67.409, 67.434]	[67.252, 67.277]
0.70	96.422	[96.627, 96.657]	[96.403, 96.433]
0.75	140.515	[140.832, 140.868]	[140.514, 140.550]
0.80	192.063	[192.488, 192.530]	[192.043, 192.085]

7.4. The effect of different choices of N

As we have explained in Section 7.1, using different N to get the sample mean to replace the expectation (7.2) will affect of the performance of the estimator. In this section, we illustrate this through numerical simulations. In Figure 3, we plot bias and variance of Plug-in, IB and Adaptive against different sample size n. We fix the dimension $d = 256$ while using three different $N = 10, 10^2, 10^3$. The bias and variance are simulated using the result of 10^4 independent trials. As one would expect, when $N = 10$ is too small, the performance is the worst. Firstly, the variance of both IB and Adaptive are already worse than Plug-in's, which should have been the same. Secondly, the bias is also not as good as $N = 10^2, 10^3$. For $N = 10$, the bias is around $10^{-2.5}$ while for the other two is around 10^{-3}. On the other hand, for $N = 10^2, 10^3$, the variance is already ideal. However, as for the bias, when n is large the discrepancy between the actual bias and its bound (blue dash line) is also large. To fill this gap, one would need large N comparable to n.

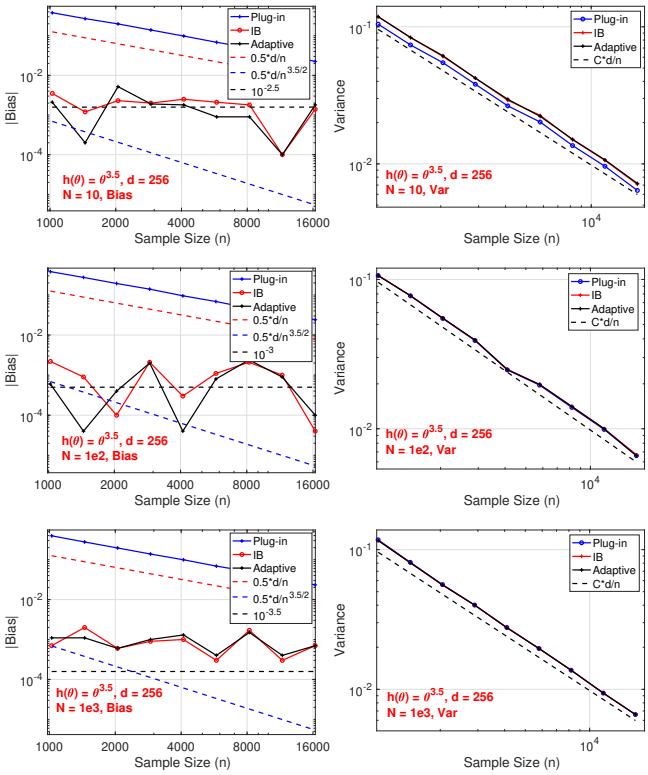

Figure 3. Comparison of different choices of N

8. Conclusion and Discussion

In this article, we studied the estimation of $f(\boldsymbol{\theta})$ for a given smooth additive function f based on noisy observations of $\boldsymbol{\theta}$. The major motivation is that when dimension d is large compared with the sample size n, the bias of the plug-in estimator $f(\bar{\mathbf{x}})$ can be large, which makes it sub-optimal. Thus non-trivial bias reduction is needed. We adopt the idea of iterative bootstrap, and applied this approach to each component function to derive a bias-reducing estimator. By establishing upper bounds on the bias and normal approximation results, we showed that the additive structure of smooth function can contribute to a looser constraint on smoothness to achieve asymptotic normality for the proposed estimator. Meanwhile, those results also imply that the proposed estimator is asymptotically efficient and can be minimax optimal. We also addressed several adaptation issues on the computational aspect. Numerical simulations validate our analysis and show the new estimator's advantage over the plug-in approach in terms of bias reduction and confidence interval construction.

As we have shown both analytically and experimentally, the proposed estimator can be minimax optimal when $\beta \geq 2$. However, one interesting question to ask is that when $\beta \in (1, 2)$, whether the plug-in estimator is optimal or not. As for this case, the estimator (3.3) is simply $f(\bar{\mathbf{x}})$. If $f(\bar{\mathbf{x}})$ is not minimax optimal, then it implies that the threshold we discussed on smoothness $\beta > 1 + \alpha$ is sharp. Then how to construct one would be interesting.

Another interesting future work direction is on the theoretical guarantee of adaptive estimation. As we have already mentioned in Section 7, the implementation of iterative bootstrap step in estimator (3.3) requires one to resample the random noise. Thus one needs to know the variance of the noise in advance, which is not so realistic in practice. We proposed a simple solution by using the sample variance computed from the observations as a surrogate. This works well experimentally and we believe theoretical results established in this paper should still hold in the Gaussian case since the approximation error introduced by using sample variance should be of a much smaller order intuitively. Nevertheless, theoretical justification in adaptive estimation requires new techniques and methods for rigorous proofs which are different from what we used here. We will leave this as an interesting future research topic.

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.

Leonard E Baum and Melvin Katz. Convergence rates in the law of large numbers. *Transactions of the American Mathematical Society*, 120(1):108–123, 1965.

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, Latitia Comminges, Alexandre B. Tsybakov, and Nicolas Verzelen. Optimal adaptive estimation of linear functionals under sparsity. *The Annals of Statistics*, 46(6A):3130–3150, 12 2018.

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

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

David L. Donoho and Richard C. 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.

Jerome H. Friedman. Greedy function approximation: A gradient boosting machine. *The Annals of Statistics*, 29 (5):1189–1232, 10 2001.

Evarist Giné and Richard Nickl. *Mathematical foundations of infinite-dimensional statistical models*, volume 40. Cambridge University Press, 2016.

Yanjun Han, Jiantao Jiao, and Rajarshi Mukherjee. On estimation of l_r-norms in gaussian white noise models. *Probability Theory and Related Fields*, 177(3):1243–1294, 2020.

Yi Hao and Alon Orlitsky. The broad optimality of profile maximum likelihood. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 10991–11003, Vancouver, Canada, 2019.

Trevor Hastie and Robert Tibshirani. Generalized additive models. *Statistical Science*, 1(3):297–310, 08 1986.

Il'dar A. Ibragimov and Rafail Z. KHas'Minskii. *Statistical estimation: asymptotic theory*, volume 16. Springer Science & Business Media, 2013.

Il'dar A. Ibragimov, Arkadi Nemirovski, and Rafail Z. 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 Transactions on Information Theory*, 66(7):4392–4418, 2020.

Jiantao Jiao, Kartik Venkat, Yanjun Han, and Tsachy Weissman. Minimax estimation of functionals of discrete distributions. *IEEE Transactions on Information 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.

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

Vladimir Koltchinskii and Karim Lounici. Asymptotics and concentration bounds for bilinear forms of spectral projectors of sample covariance. In *Annales de l'Institut Henri Poincaré, Probabilités et Statistiques*, volume 52, pages 1976–2013. Institut Henri Poincaré, 2016.

Vladimir Koltchinskii and Karim Lounici. Concentration inequalities and moment bounds for sample covariance operators. *Bernoulli*, 23(1):110–133, 02 2017.

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, 2021.

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

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

L. LeCam. Convergence of estimates under dimensionality restrictions. *The Annals of Statistics*, 1(1):38–53, 01 1973.

Oleg Lepski, Arkadi Nemirovski, and Vladimir Spokoiny. On estimation of the l_r 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 nonparametric 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 Nemirovski. 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 Nemirovski. Topics in non-parametric. *Ecole d'Eté de Probabilités de Saint-Flour*, 28:85, 2000.

Alexandre B. Tsybakov. *Introduction to Nonparametric Estimation*. Springer, New York, NY, 2009. ISBN 978-0-387-79051-0.

Roman Vershynin. *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge University Press, 2018.

Yihong Wu and Pengkun Yang. Minimax rates of entropy estimation on large alphabets via best polynomial approximation. *IEEE Transactions on Information 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.