HIGH-ORDER STATISTICAL FUNCTIONAL EXPANSION AND ITS APPLICATION TO SOME NONSMOOTH PROBLEMS

BY FAN ZHOU\textsuperscript{1}, PING LI\textsuperscript{1} AND CUN-HUI ZHANG\textsuperscript{2}

Baidu Research\textsuperscript{1} and Rutgers University\textsuperscript{2}

Let \( x_j = \theta + \varepsilon_j, j = 1, \ldots, n \), be observations of an unknown parameter \( \theta \) in a Euclidean or separable Hilbert space \( \mathcal{H} \), where \( \varepsilon_j \) are noises as random elements in \( \mathcal{H} \) from a general distribution. We study the estimation of \( f(\theta) \) for a given functional \( f : \mathcal{H} \rightarrow \mathbb{R} \) based on \( x_j \)'s. The key element of our approach is a new method which we call High-Order Degenerate Statistical Expansion. It leverages the use of classical multivariate Taylor expansion and degenerate \( U \)-statistic and yields an elegant explicit formula. In the univariate case of \( \mathcal{H} = \mathbb{R} \), the formula expresses the error of the proposed estimator as a sum of order \( k \) degenerate \( U \)-products of the noises (with no tied index for each \( k \)-product) with coefficient \( f^{(k)}(\theta)/k! \) and an explicit remainder term in the form of the Riemann-Liouville integral as in the Taylor expansion around the true \( \theta \). For general \( \mathcal{H} \), the formula expresses the estimation error in terms of the inner product of \( f^{(k)}(\theta)/k! \) and the average of the tensor products of \( k \) noises with distinct indices and a parallel extension of the remainder term from the univariate case. This makes the proposed method a natural statistical version of the classical Taylor expansion. The proposed estimator can be viewed as a jackknife estimator of an ideal degenerate expansion of \( f(\cdot) \) around the true \( \theta \) with the degenerate \( U \)-product of the noises, and can be approximated by bootstrap. Thus, the jackknife, bootstrap and Taylor expansion approaches all converge to the proposed estimator. We develop risk bounds for the proposed estimator under proper moment conditions and a central limit theorem under a second moment condition (even in expansions of higher than the second order). We apply this new method to generalize several existing results with smooth and nonsmooth \( f \) to universal \( \varepsilon_j \)'s with only minimum moment constraints.

1. Introduction. We consider

\[
(1.1) \quad x_j = \theta + \varepsilon_j, \quad j = 1, \ldots, n
\]

where \( \theta \) is an unknown parameter and belongs to a finite-dimensional Euclidean or a separable Hilbert space \( \mathcal{H} \), and \( \varepsilon_j \) are mean zero random noises in \( \mathcal{H} \) with general distributions. The goal of this article is to study the estimation of \( f(\theta) \) for a given functional \( f : \mathcal{H} \rightarrow \mathbb{R} \) when the complexity parameter of the problem is large under only minimal moment constraints.

The study of the estimation of functionals of high-dimensional or infinite-dimensional parameter has a long history. Notable results include but not limited to [Lev76, Lev78, INH87, BR88, Nem91, BM95, Lau96, LNS99, Nem00, IH13]. Two types of special functionals are extensively studied, namely the linear and quadratic functionals. Results on the estimation of linear functionals include [DL87, DL91, KT01, CL05b] and the references therein. Results on the estimation of quadratic functionals include [DN90, LM00, BR03, CL05a, Kle06] and the references therein. Recently there is a noticeable surge of interest in efficient and minimax rate optimal estimation of functionals of parameter in high dimensional

\textsuperscript{1}The work of Cun-Hui Zhang was conducted as a consulting researcher at Baidu Research – 10900 NE 8th St. Bellevue, WA 98004, USA
models or models with growing dimension, see [CCT17, ZZ14, VdGBRD14, KZ21, ZL19]. However, these papers are mostly focused on the case of Gaussian noise which is quite restrictive in practice. Thus it is of great interest to develop robust inferential procedures that are less sensitive to the distributional assumptions. This is a major motivation for us to study the functional estimation of high-dimensional parameters under low moment constraints on the noise distribution.

A straightforward approach to this problem is to use the plug-in estimator \( f(\bar{x}) \) with the sample mean \( \bar{x} \) of the observations. One may think of \( f(\bar{x}) \) as a good estimator as \( \bar{x} \) is the maximum likelihood estimator of \( \theta \) when the noise \( \varepsilon_j \) are Gaussian. However, interestingly, our problem is more delicate. Even if in the Gaussian shift model, there are situations where the plug-in estimator is sub-optimal due to its large bias as soon as the complexity parameter \( \alpha \) is too large. Thus it is of great interest to develop robust inferential procedures that are less sensitive to the distributional assumptions. This is a major motivation for us to study the functional estimation of high-dimensional parameters under low moment constraints on the noise distribution.

### 1.1. Related works

Several methods were proposed recently to address the problem in the Gaussian shift model. One of them, developed in a series of works [Kol20, KZ21, KZ19], is to use an iterative bootstrap technique to correct the bias of the plug-in estimator. We briefly summarize the idea of iterative bootstrap as follows. Denote by \( T \) the linear operator given by

\[
Tg(\theta) := \mathbb{E}_\theta g(\theta) = \mathbb{E}g(\theta + \tilde{\varepsilon}),
\]

where \( \tilde{\varepsilon} = n^{-1} \sum_{j=1}^n \varepsilon_j \). Let \( I \) be the identity operator and \( B := T - I \). To create an estimator \( g(\bar{x}) \) of \( f(\theta) \) with small bias, it is tempting to solve the integral equation

\[
Tg(\theta) = (I + B)g(\theta) = f(\theta)
\]

as accurately as possible. However, solving such an integral equation itself is challenging. Instead, the authors used a natural finite approximation of the Neumann series \( (I + B)^{-1} = (I - B + B^2 - B^3 + \ldots) \) to create the following estimator

\[
f_k(\bar{x}) := \sum_{j=0}^k (-1)^j B^j f(\bar{x})
\]

with

\[
B^j f(x) := \mathbb{E}_{\varepsilon} f(j)(\bar{x} + \sum_{\ell=1}^j \tau_\ell \tilde{\varepsilon}_\ell)(\tilde{\varepsilon}_1, \ldots, \tilde{\varepsilon}_j)
\]

where \( \mathbb{E}_{\varepsilon} \) is the conditional expectation taking over the i.i.d. random variables \( \{ \tau_\ell \}_{i=1}^j \sim U[0, 1] \) and i.i.d. bootstrap samples \( \{ \tilde{\varepsilon}_i \}_{i=1}^j \) of \( \tilde{\varepsilon} \). It was shown that (1.3) holds exactly for \( B = T - I \) so that the bias of (1.2) is exactly \( (-1)^k B^{k+1} f(\theta) \). Intuitively, the main idea is that given the condition that the bootstrap operator \( B \) is small, relatively small noise \( \varepsilon \) and sufficiently smooth \( f \), the bias should also be small. Thus bias reduction can be achieved. In practice, the expectation in (1.3) can be replaced by averaging samples obtained from Monte-Carlo simulations, of which performance can be guaranteed by law of large numbers [ZL21].

Another work [JH20] compared multiple jackknife, iterative bootstrap and Taylor expansion approaches in a study of the estimation of \( f(\theta) \) for a given \( f \in C[0, 1] \) based on Bernoulli(\( \theta \)) observations. Although [JH20] studied the problem in a very specific classical
model, the methods they used to approach the problem are closely related to ours. Especially, in Sec. A of their paper, they proposed a multiple jackknife estimator

\[
\hat{f}_m := \sum_{k=1}^{m} C_k \frac{n_k!}{n!} \sum_{1 \leq i_1, \ldots, i_{n_k} \leq n} f\left(\frac{1}{n_k} \sum_{j=1}^{n_k} x_{i_j}\right),
\]

with sample sizes \(n_1 < n_2 < \cdots < n_m \leq n\) and constants \(C_k\) satisfying \(\sum_{k=1}^{m} C_k / n_k^p = I\{\rho = 0\}\) for \(\rho = 0, \ldots, m - 1\). This estimator makes use of \(U\)-statistics from sub-sampling data to cancel each other’s bias in order to achieve overall bias-reduction when the bias of \(f(\bar{x})\) is a sum of \(\sum_{\rho=1}^{m-1} \text{bias}_\rho / n^\rho\) and a higher order term. In Sec. C, they generalized the bias correction technique of first order Taylor expansion to higher orders in an iterative way. However, as they noted in [JH20], an iterative bias correction into Taylor series of higher orders can be tedious and computationally intensive even in the one dimensional case. As a result, they proposed to use the sample splitting technique and considered the Taylor expansion of \(f(\theta)\) at \(\theta\):

\[
\hat{f}_m := \sum_{k=0}^{2m-1} \frac{f(k)(\hat{\theta}^{(1)})}{k!} \sum_{j=0}^{k} \binom{k}{j} \hat{\theta}_j^{(2)} (-\hat{\theta}^{(1)})^{k-j}
\]

where \(\hat{\theta}_j^{(2)}\) is an unbiased estimator of \(\theta_j\), and \(\hat{\theta}^{(1)}\) and \(\hat{\theta}_j^{(2)}\) are obtained from split samples independent of each other.

In a series of works [CL11, CCT17, CCT20, CC19], the authors studied rate minimax estimation of \(f(\theta)\) under Gaussian shift model when \(f(\theta) = \|\theta\|_p\) or \(f(\theta) = \|\theta\|_p^2\). The main idea of their works is to use Hermite polynomials to approximate the given functional, possibly with some delicate truncation on the magnitude of each observational coordinate to address sparsity presented in \(\theta\).

Based on the seminal work on unbiased estimation [Kol50] and the fruitful idea of Littlewood-Paley theory, [ZL19] proposed a Fourier analytical estimator

\[
\hat{f}_R(\bar{x}) := \frac{1}{(2\pi)^{d/2}} \int_{\Omega} \mathcal{F}f(\zeta)e^{i\zeta \cdot \bar{x}} d\zeta,
\]

where \(\Omega := \{\zeta : \|\zeta\| \leq R\} \subset \mathbb{R}^d\) is a properly truncated region in the frequency domain. The main idea is that the factor \(e^{i\zeta \cdot \bar{x}}\) in (1.6) exactly cancels the characteristic function \(\mathbb{E}[e^{i\zeta \cdot \bar{x}}]\) of the noise which makes \(\hat{f}_R(\bar{x})\) an unbiased estimator of the analytical part of \(f\) and yet the remainder can be uniformly small. Thus, overall bias reduction is achieved.

In addition to the setting when \(\theta \in \mathbb{R}^d\) resides in a finite-dimensional space, there are also some progress recently in the related problems in nonparametric domain where \(\theta : \mathbb{R}^d \to \mathbb{R}\) belongs to some multivariate density class and \(f(\theta)\) is a known functional of \(\theta\). For instance, [BSY19] studied the estimation of differential entropy of \(\theta\): \(\hat{f}(\theta) = -\int_{\mathbb{R}^d} \theta(x) \log \theta(x) dx\) when \(\theta\) belongs to certain Hölder-type class. The authors generalized the idea of the Kozachenko-Leonenko estimator [KL87] by introducing a deliberately constructed weighted version, and proved its efficiency and asymptotic normality for arbitrary \(d\) while relaxed the support of \(\theta\) to be unbounded. In another work, [HJWW20] established the minimax optimal rate of the same problem over certain general Lipschitz balls with smoothness index \(s \in (0, 2]\). Their results hold when \(\theta\) doesn’t need to be bounded away from zero and can have unbounded support. Their approach was based on using kernel smoothing estimator \(\theta_h\) of \(\theta\), and then to estimate \(f(\theta)\) by constructing an estimator of \(f(\theta_h)\) with polynomial approximation, Taylor expansion and \(U\)-statistic techniques.
1.2. Our contribution. Although the methods we mentioned above are versatile and feasible in each of their settings, they are also either computationally intensive when it comes to implementation or limited by the form of the functional and/or the distribution type of the noise. In this article, inspired by (1.6), we go back to the classical multivariate Taylor expansion and leverage the use of degenerate $U$-statistic, and propose a new method which we call High-Order Degenerate Statistical Expansion (HODSE). The new method leads to a unified estimator of $f(\Theta)$ with a remarkably neat explicit formula for high-order expansions. What’s more important, the resulting new estimator miraculously and systematically cancels the bias and all other non-degenerate terms up to any preassigned order in estimating the intermediate derivatives when using classical Taylor series to reduce bias, with an explicit Riemann-Liouville integral formula for the remainder term as in the analytical Taylor expansion. This makes HODSE a natural statistical version of the classical Taylor expansion.

The paper is organized as follows. In Section 2, we introduce HODSE with the explicit formula for its degenerate expansion, explain its differences as well as connections to the classical Taylor series and leverage the use of degenerate $k$-linear forms in $H$. We denote by $\mathcal{H}$ to denote a bounded linear operator in $H$. Throughout this paper, we assume for simplicity that $\mathcal{H}$ is equipped with the inner-product $\langle \cdot, \cdot \rangle$ and the Hilbert-Schmidt norm $\| \cdot \|_{\text{HS}}$. We denote by $\langle \cdot, \cdot \rangle$ the tensor spectral norm $\| \cdot \|_{\text{T}}$, and the tensor product determined by the bilinear $\langle \cdot, \cdot \rangle_k$ determined by the bilinear extension of $\langle \otimes_{\ell=1}^k u_\ell, \otimes_{\ell=1}^k v_\ell \rangle_k = \prod_{\ell=1}^k \langle u_\ell, v_\ell \rangle$ in $H^{\otimes k}$, the Hilbert-Schmidt norm $\| \cdot \|_{\text{HS}}$ induced by the inner-product $\langle \cdot, \cdot \rangle_k$, and the tensor spectral norm $\| T^{(k)} \|_S := \max_{\| h_1 \| = \cdots = \| h_k \| = 1} \left| \langle T^{(k)}, \otimes_{j=1}^k h_j \rangle \right|$, $T^{(k)} \in H^{\otimes k}$.

A member $T^{(k)}$ in $H^{\otimes k}$ is symmetric if $\langle T^{(k)}, \otimes_{\ell=1}^k v_\ell \rangle_k = \langle T^{(k)}, \otimes_{\ell=1}^k v_\ell \rangle_k$ for all permutations $(i_1, \ldots, i_k)$ of $(1, \ldots, k)$. We denote by $\otimes$ the Kronecker product as the bilinear mapping from $(H^{\otimes j}, H^{\otimes (k-j)})$ to $H^{\otimes (j+k)}$ determined by $\langle \otimes_{\ell=1}^j u_\ell, \otimes_{\ell=1}^k v_\ell \rangle \rightarrow \otimes_{\ell=1}^j u_\ell \otimes_{\ell=1}^k v_\ell$. We denote by $\times_j$ the mode $j$ tensor product determined by $T^{(k)} \times_j \Sigma : \otimes_{\ell=1}^j v_\ell \rightarrow$...
\[ \langle T^{(k)} , \otimes_{\ell=1}^{j-1} v_\ell \otimes \Sigma v_{j} \otimes_{\ell=j+1}^{k} v_\ell \rangle_k \] as a \( k \)-linear form, and \( T^{(k)} \times J u : \otimes_{\ell=1}^{j-1} v_\ell \otimes_{\ell=j+1}^{k} v_\ell \to \langle T^{(k)} , \otimes_{\ell=1}^{j-1} v_\ell \otimes u \otimes_{\ell=j+1}^{k} v_\ell \rangle_k \) as a \((k - 1)\)-linear form.

Unless otherwise stated we assume \( f : \mathcal{H} \to \mathbb{R} \) is \( m \)-times Fréchet differentiable at \( \theta \) with \( f^{(k)}(\theta) \) being the \( k \)-th derivative as a symmetric \( k \)-linear form in \( \mathcal{H}^{\otimes k} \) with finite spectral norm \( \| f^{(k)}(\theta) \|_S \), \( 1 \leq k \leq m \). For any \( s > 0 \), the Hölder norm of order \( s \) is defined with \( m = \lceil s \rceil - 1 \) by

\[
\| f \|_{(s)} := \sup_{x, y \in \mathcal{H} : \|x - y\| \neq 0} \left\{ \frac{\| f^{(m)}(x) - f^{(m)}(y) \|_S}{\| x - y \|^{s - m}} \right\}.
\]

We may also use the Hölder smoothness of \( f \) induced by the Hilbert-Schmidt norm to

\[
\| f \|_{(s), HS} := \sup_{x, y \in \mathcal{H} : \|x - y\| \neq 0} \left\{ \frac{\| f^{(m)}(x) - f^{(m)}(y) \|_{HS}}{\| x - y \|^{s - m}} \right\}.
\]

For notational simplicity, we suppress the dependence of the spectral and Hilbert-Schmidt norms on the tensor order \( k \). We note that when \( \mathcal{H} = \mathbb{R} \), \( \| f^{(k)}(\theta) \|_S = \| f^{(k)}(\theta) \|_{HS} = | f^{(k)}(\theta) | \) and \( \| f \|_{(s), HS} = \| f \|_{(s)} \).

For \( p > 0 \), we denote the \( \ell_p \) norm by \( \| v \|_p = (\sum_{j=1}^{d} |v_j|^p)^{1/p} \) for \( v = (v_1, \ldots, v_d)^T \in \mathbb{R}^d \) and the \( L_p \) norm by \( \| X \|_{L_p} = (\mathbb{E} \| X \|^p)^{1/p} \) for random variables \( X \in \mathbb{R} \) and \( \| \varepsilon \|_{L_p} = (\mathbb{E} \| \varepsilon \|^p)^{1/p} \) for random elements \( \varepsilon \in \mathcal{H} \).

We use \( \mathcal{F} \) and \( \mathcal{F}^{-1} \) to denote the Fourier transform (FT) and inverse Fourier transform (IFT) respectively. We use the conventional notation \( \Rightarrow \) to denote weak convergence, i.e. convergence in distribution and use \( \Rightarrow_p \) to denote convergence in probability. Throughout the paper, given nonnegative numbers \( a \) and \( b \), \( a \leq b \) means that \( a \leq C b \) for a numerical constant \( C \), and \( a \times b \) means that \( a \leq b \) and \( b \leq a \), \( a \land b = \min\{a, b\} \) and \( a \lor b = \max\{a, b\} \). We use \( C_\alpha \) to denote a constant depending on \( \alpha \) only.

2. High-order degenerate statistical expansion. We are interested in the estimation of \( f(\theta) \) with a known function \( f \) and a high- or infinite-dimensional unknown parameter \( \theta \) based on observations \( x_j \) with mean \( \mathbb{E}[x_j] = \theta \). Throughout this section we assume that \( f(\cdot) : \mathcal{H} \to \mathbb{R} \) is \( m \) times Fréchet differentiable at \( \theta \) so that the following expansion holds,

\[
f(\theta + h) = f(\theta) + \sum_{k=1}^{m} \frac{\langle f^{(k)}(\theta), h^{\otimes k} \rangle_k}{k!} + o(\| h \|^m)
\]

for symmetric \( k \)-linear forms \( f^{(k)}(\theta) \) with finite spectral norm \( \| f^{(k)}(\theta) \|_S \), \( 1 \leq k \leq m \).

Consider the estimation of \( f(\theta) \) based on independent data points \( x_j \in \mathcal{H}, 1 \leq j \leq n \), with \( \mathbb{E}[x_j] = \theta \). Let \( \varepsilon_j = x_j - \theta, \bar{x} = \sum_{j=1}^{n} x_j/n \) and \( \bar{\varepsilon} = \bar{x} - \theta \). The plug-in estimator \( f(\bar{x}) \) has the Taylor expansion

\[
f(\bar{x}) = f(\theta) + \sum_{k=1}^{m} \frac{\langle f^{(k)}(\theta), \bar{\varepsilon}^{\otimes k} \rangle_k}{k!} + O(1)\| f \|_{(s)} \| \bar{\varepsilon} \|^s
\]

with the Hölder norm \( \| f \|_{(s)} \) in (1.7). When the remainder term is sufficiently small for \( 1 < s \leq 2 \), \( f(\bar{x}) \) is asymptotically linear and the CLT holds under the Lindeberg condition. We note that the Lindeberg condition is needed in our setting even when \( \varepsilon_j \) are i.i.d. with zero mean and finite variance because the distribution of \( \langle f^{(1)}(\theta), \varepsilon_j \rangle_1 \) typically changes when \( \theta \in \mathbb{R}^d \) and \( d = d_n \to \infty \).

Let \( \Sigma = \mathbb{E}[\sum_{j=1}^{n} \varepsilon_j \otimes \varepsilon_j], \sigma = \| \Sigma \|_S^{1/2} \) and \( r(\Sigma) = \text{trace}(\Sigma)/\sigma^2 \geq 1 \) be the effective rank of \( \Sigma \), where \( \text{trace}(\Sigma) = \langle I_\mathcal{H}, \Sigma \rangle_2 \) with the identity operator \( I_\mathcal{H} \) in \( \mathcal{H} \). The plug-in
estimator can be biased when we use higher order approximations, say $s > 2$ in (2.1), to control the remainder term. For example, by comparing the usual bounds for the standard deviation of the linear term and the bias in the second order expansion when $\varepsilon_j$’s are i.i.d. copies of an isotropic $\varepsilon$,

$$\begin{align*}
\{\mathbb{E}[\langle f^{(1)}(\theta), \varepsilon \rangle^2] \}^{1/2} &= \langle f^{(1)}(\theta) \otimes f^{(1)}(\theta), \Sigma/n \rangle^{1/2} \approx \|f^{(1)}(\theta)\|/n^{1/2}, \\
\mathbb{E}[\langle f^{(2)}(\theta), \varepsilon^2 \rangle] &= \langle f^{(2)}(\theta), \Sigma/n \rangle \leq \|f^{(2)}(\theta)\|_S \sigma^2 r(\Sigma)/n,
\end{align*}$$

we find that the bias is typically of smaller order than the linear term when

$$\|f^{(2)}(\theta)\|_S (\sigma r(\Sigma)/n^{1/2}) \ll \|f^{(1)}(\theta)\|,$$

which requires $r(\Sigma) \ll n^{1/2}$ when $\|f^{(2)}(\theta)\|_S$ and $\|f^{(1)}(\theta)\|$ are of the same order. Thus, bias correction may be needed when $r(\Sigma) \gg n^{1/2}$.

To remove the bias in the case of $s > 2$, one may subtract from $f(\hat{x})$ some estimator $\hat{\text{bias}}_2$ for bias $\text{bias}_2 = \langle f^{(2)}(\theta), \sum_{j=1}^n \varepsilon_j^2/(2n^2) \rangle_2$ in the expansion of $\langle f^{(2)}(\theta), \varepsilon^2 \rangle/2$. For $s \geq 3$, we also want to remove $\langle f^{(3)}(\theta), \sum_{j=1}^n \varepsilon_j^3/(6n^3) \rangle_3$ and the third order bias term in the expansion of $-\text{bias}_2$, so on and so forth. This approach was discussed in detail in [JH20] in the one dimensional case in the Bernoulli model. As the authors of [JH20] noted, even in the one dimensional case, such an ad hoc process can be tedious and hard to keep track of when dealing with higher-order correction terms. Yet our case is more delicate and can be high dimensional. One may also work with the Taylor expansion

$$f(\theta) = f(\bar{x}) + \sum_{k=1}^m \frac{\langle f^{(k)}(\theta), \varepsilon^{(k)} \rangle_k}{(-1)^k k!} + O(1)\|f\|_s \|\varepsilon\|^s$$

at the observable $\bar{x}$ but similar issues arise when we deal with the unobservable $\varepsilon^{(k)}$.

In this paper, we present a unified yet very neat formula for bias correction which was aimed at achieving

$$\hat{f} \approx f(\theta) + \sum_{k=1}^m \frac{\langle f^{(k)}(\theta), \varepsilon^{(k)} \rangle_k}{k!}, \quad (2.2)$$

with a remainder term of higher order than $m$, where

$$\varepsilon^{(k)} := \sum_{1 \leq j_1 \neq \cdots \neq j_k \leq n} \frac{\varepsilon_{j_1} \otimes \cdots \otimes \varepsilon_{j_k}}{n(n-1) \cdots (n-k+1)} \quad (2.3)$$

are completely degenerate $U$-tensors of order $k$. As $\varepsilon^{(k)}$ are completely degenerate, the bias of (2.2), $\mathbb{E}[\hat{f} - f(\theta)]$, is identical to the expectation of the remainder term in this expansion and thus is expected to be of higher order than the $m$-th order. An estimator of form (2.2) may also achieve variance reduction by removing terms like $\langle f^{(3)}(\theta), \sum_{j_1, j_2} \varepsilon_{j_1}^2 \otimes \varepsilon_{j_2}/(6n^3) \rangle_3$ in the expansion of the plug-in estimator $f(\bar{x})$.

Given the order $s$ of the Hölder smoothness of $f$, we propose to estimate $f(\theta)$ by

$$\hat{f} := f(\bar{x}) + \sum_{1 \leq k \leq m} \frac{\langle f^{(k)}(\bar{x}), \bar{u}^{(k)} \rangle_k}{k!} \quad (2.4)$$

with $m = \lceil s \rceil - 1$ and $U$-statistics

$$\bar{u}^{(k)} := \sum_{1 \leq j_1 \neq \cdots \neq j_k \leq n} \frac{(x_{j_1} - \bar{x}) \otimes \cdots \otimes (x_{j_k} - \bar{x})}{n(n-1) \cdots (n-k+1)}. \quad (2.5)$$
We note that $\bar{u}^{(1)} = 0$ so that the sum in \((2.4)\) actually runs from $k = 2$ to $k = m$.

We advocate the use of \((2.4)\), a plug-in estimator of the right-hand side of \((2.2)\), as the proper statistical expansion of the plug-in estimator $f(\bar{x})$ to higher orders. A remarkable feature of this natural formula is that the resulting estimator automatically cancels out the impacts of estimating $f^{(k)}(\theta)$ by $f^{(k)}(\bar{x})$ and $\varepsilon_j$ by $x_j - \bar{x}$ on the right-hand side of \((2.2)\) simultaneously for all orders $k = 1, \ldots, m$. Thus, the estimator \((2.4)\) can be viewed as a degenerate statistical expansion of function $f$. This claim is formally justified by the following proposition which gives an explicit formula for the remainder term in \((2.2)\) when the estimator \((2.4)\) is used.

**Proposition 2.1.** Let $\hat{f}$ be as in \((2.4)\) with $n \geq m \geq 2$. Then, \((2.2)\) holds,

\[
\hat{f} = f(\theta) + \sum_{k=1}^{m} \frac{\langle f^{(k)}(\theta), \varepsilon^{(k)} \rangle_{k}}{k!} - \text{Rem}_m,
\]

with the $\varepsilon^{(k)}$ in \((2.3)\) for the noise vectors $\varepsilon_j = x_j - \theta$ and with the remainder term

\[
\text{Rem}_m := \sum_{k=0}^{m} \frac{\sum_{\ell,k} \langle f^{(m-k)}h^{(\ell)}(t), \varepsilon^{(\ell)\otimes(k)} \rangle_{m-k}}{(-1)^{m-k}k!},
\]

where $J^0h = h(1)$ and $J^\alpha h = \int_0^1 h(t)(1 - t)^{\alpha - 1}dt / \Gamma(\alpha)$ give the Riemann-Liouville integral operator for $h : [0, \infty) \to \mathcal{H}^m$ and $\alpha > 0$, $\Delta^{(m)}(t) := f^{(m)}(\bar{x} + t(\theta - \bar{x})) - f^{(m)}(\bar{x})$ and $\varepsilon = n^{-1} \sum_{j=1}^{n} \varepsilon_j = \bar{x} - \theta$.

While the expansion \((2.6)\) looks similar to the expansion \((2.1)\), the difference is highly significant in high-order analysis because $\varepsilon^{(k)}$ in \((2.3)\) are completely degenerate $U$-tensors. For i.i.d. data, $E[\varepsilon^{(k)}] = 0$ for all integers $k \geq 1$ but $E[\varepsilon^{\otimes 2k}, \varepsilon^{\otimes 2k}] > 0$ for all $v \neq 0$. Moreover, $E[\langle f^{(k)}(\theta), \varepsilon^{(k)} \rangle^2_k]$ depends on $\varepsilon$ only through its covariance operator for all $k$ [See (3.8)].

We call \((2.4)\) High-Order Degenerate Statistical Expansion (HODSE) as it gives a natural statistical version of the Taylor expansion at the true parameter $\theta$ in the following sense: The $k$-th order term on the right-hand side of \((2.6)\) has zero mean and a standard deviation of the same form as the absolute value of the $k$-th order term in the analytical Taylor expansion. For $f : \mathbb{R} \to \mathbb{R}$ and independent noise with common standard deviation $n^{1/2}\sigma_n$, the standard deviation of the $k$-th order term in \((2.6)\) is $(1 + o(1))|f^{(k)}(\theta)|\sigma_n^k/k!$ whereas the absolute value of the $k$-th order term in the Taylor series of $f(\theta + t)$ is $|f^{(k)}(\theta)| \times |t|^k/k!$.

In the following lemma we provide some analytic upper bounds for the remainder term. For symmetric order $m$ tensors $T^{(m)}$, define norms

\[
\|T^{(m)}\|_{m-k,k} := \sup_{\|u\| = 1} \langle T^{(m)}, u^{\otimes (m-k)} \otimes v \rangle_m,
\]

where the supreme is taken over $u \in \mathcal{H}$ and order $k$ tensors $v$. In particular, $\|T^{(m)}\|_{0,0}$ is less than or equal to the Hilbert-Schmidt norm $\|T^{(m)}\|^2_{HS} = \langle T^{(m)}, T^{(m)} \rangle^{1/2}_{m}$.

**Lemma 2.2.** Let $n \geq m = |s| - 1 \geq 2$ and $\text{Rem}_m$ be the remainder term in \((2.7)\). Then,

\[
|\text{Rem}_m| \leq \sum_{k=0}^{m} \left( \max_{0 < t \leq 1} \frac{\|\Delta^{(m)}(t)\|_{m-k,k}}{(t\|\varepsilon\|)^{s-m}} \right) \|\varepsilon\| |\varepsilon^{(k)}|^{1/2} \Gamma(s - k + 1)k!}
\]
with the tensor norm \( \| \cdot \|_{m-k,k} \) in (2.8) and

\[
\text{(2.10)} \quad |\text{Rem}_m| \leq \left( \max_{0 < t \leq 1} \frac{\| \Delta^{(m)}(t) \|_S}{(t\| \epsilon \|)^{s-m}} \right) \sum_{k_1 + \ldots + k_b = t \leq m, k_a \geq 2 \forall a \geq 0} \frac{(C_{k,n}/C_{b,n})C_{k_1,\ldots,k_b}}{(s-k+1)!} \times \sum_{j_1 \neq \ldots \neq j_b} \| \epsilon_{j_1} \|^k \cdots \| \epsilon_{j_b} \|^k \| \epsilon \|^s - \ell \over n(n-1) \cdots (n-b+1)n^{t-\ell}
\]

with \( C_{k,n} = n^k(n-k)!/n! \) satisfying \( C_{k,n} \leq \exp((k-1)k/n) \) and certain positive integers \( C_{k_1,\ldots,k_b} \) satisfying \( r_k = \sum_{k_1 + \ldots + k_b = t \leq k} C_{k_1,\ldots,k_b} \leq k! \).

In the above lemma, (2.9) is typically sharper than (2.10) when \( \mathcal{H} \) is of low-dimension, and vice versa. For example, when \( \mathcal{H} = \mathbb{R}^d \), \( \Delta^{(m)}(t) \) is a \( d \times \cdots \times d \) tensor, so that \( \| \Delta^{(m)}(t) \|_{m-k,k} \leq d^{k-1} \| \Delta^{(m)}(t) \|_S \) and (2.9) could be sharper when \( d^{m-1} \leq m! \). In (2.10), the \( \mathcal{T} \)-tensor \( \bar{\epsilon}^{(k)} \) needs to be written as a sum of rank-one tensors of proper forms to apply the spectrum norm, and \( r_k \) is the number of different types of such rank-one tensors. The order of constants \( 1/\{(m-k)!k!\} \) and \( r_k \) in (2.9) and (2.10), arising from the explicit remainder formula, become crucial when we need to use the estimator (2.4) with diverging order \( m \); see for example, Subsections 5.1 and 5.2. In our study of the asymptotic normality in Theorem 4.1, both (2.9) and (2.10) are used respectively under conditions on the “effective rank” of \( f \) and the boundedness of the order \( m \) of the expansion.

2.1. Jackknife and bootstrap. The jackknife and bootstrap approaches converge to the Taylor expansion approach in HODSE. The following discussion casts the estimator (2.4) as a plug-in jackknife estimator of

\[
\text{(2.11)} \quad f(\theta) + \sum_{k=1}^{m} \frac{\langle f^{(k)}(\theta), \bar{\epsilon}^{(k)} \rangle_k}{k!}
\]

ideally with smaller estimation error for larger \( m \). Let \( \mathbb{E}^* \) be the conditional expectation given the data points \( \{x_1, \ldots, x_n\} \) in \( \mathcal{H} \), and \( \{x_1^*, \ldots, x_n^*\} \) be sampled without replacement from \( \{x_1, \ldots, x_n\} \) under \( \mathbb{E}^* \). Let \( \theta^* = \bar{x} \) be the mean of the resampled data and \( \epsilon_j^* = x_j^* - \theta^* \) the resampled noise. Because

\[
\mathbb{E}^* [\epsilon_1^* \otimes \cdots \otimes \epsilon_k^*] = \bar{u}^{(k)}
\]

with the \( \bar{u}^{(k)} \) in (2.5), the estimator (2.4) can be written as

\[
\text{(2.12)} \quad \hat{f} = f(\bar{x}) + \sum_{k=1}^{m} \frac{\langle f^{(k)}(\bar{x}), \bar{u}^{(k)} \rangle_k}{k}.
\]

This can be viewed a jackknife estimator of (2.11) as subsamples of size \( k \) are used to correct the biases of the estimates for \( \langle f^{(k)}(\theta), \bar{\epsilon}^{(k)} \rangle_k \) based on subsamples of sizes \( \ell = 1, \ldots, k-1 \), and the average \( \mathbb{E}^* \) over all permutations is used to achieve noise cancellation. It is different from the standard Jackknife as the same \( \theta^* = \bar{x} \) is plugged-in in the estimation of \( f^{(k)}(\theta) \) to achieve systematic bias cancellation. Bootstrap can be used to compute the estimator (2.4),

\[
\text{(2.13)} \quad \hat{f} \approx f(\bar{x}) + \sum_{k=1}^{m} \sum_{B=1}^{N_K} \frac{\langle f^{(k)}(\bar{x}), \epsilon_1^* \otimes \cdots \otimes \epsilon_k^*(B) \rangle_k}{k!N_K}.
\]
where \( \langle f^{(k)}(\bar{x}), \varepsilon_1^* \otimes \cdots \otimes \varepsilon_{k-1}^* \rangle, 1 \leq k \leq m \), are computed using \( \{\varepsilon_1^*, \ldots, \varepsilon_m^*\}^{(B)} \), and \( \{\varepsilon_1^*, \ldots, \varepsilon_m^*\}^{(B)} \) are i.i.d. copies of \( \{\varepsilon_1^*, \ldots, \varepsilon_m^*\} \) under \( \mathbb{E}^* \). We note that this bootstrap of the jackknife estimator (2.12) is different from the standard empirical bootstrap [Efr82] as \( \varepsilon_1^*, \ldots, \varepsilon_m^* \) are sampled without replacement for each bootstrap copy \( \{\varepsilon_1^*, \ldots, \varepsilon_m^*\}^{(B)} \) but the bootstrap copies \( \{\varepsilon_1^*, \ldots, \varepsilon_m^*\}^{(B)} \) for different \( B \) are sampled with replacement.

It is remarkable that the plug-in jackknife estimator (2.12) and its bootstrapped version (2.13), with the population mean \( \theta \) replaced by the jackknife/bootstrap mean \( \bar{x} \) simultaneously in the nonlinear \( f^{(k)}(\cdot) \), automatically yields the non-degenerate expansion (2.11) with a remainder term of higher order than \( m \), akin to the automatic Edgeworth expansion adjustments of resampling methods in the central limit theorem [Sin81, BF81, Wu86].

3. Risk bounds. We provide in this section upper bounds for the bias and \( L_1(\mathbb{P}) \) and \( L_2(\mathbb{P}) \) risks of the estimator (2.4). Unless otherwise stated, we assume in the rest of the paper that \( \varepsilon_j \) are independent but not necessarily identically distributed random elements in \( \mathcal{H} \). Recall that the covariance operator \( \Sigma \) for the noise, noise level \( \sigma \) and the effective rank of \( \Sigma \) are defined respectively as

\[
\Sigma = \mathbb{E} \left[ \frac{1}{n} \sum_{j=1}^{n} \varepsilon_j \otimes \varepsilon_j \right], \quad \sigma = \| \Sigma \|_S^{1/2}, \quad r = r(\Sigma) = \text{trace}(\Sigma) = \frac{\text{trace}(\Sigma)}{\sigma^2}.
\]

Because \( \mathbb{E}[\langle f^{(k)}(\theta), \bar{\varepsilon}^{(k)} \rangle] = 0 \) in the expansion (2.6), the bias of the estimator (2.4) is no greater in absolute value than the \( L_1(\mathbb{P}) \) norm of the remainder term in (2.7). In the following theorem, we develop such bias bounds under a Hölder smoothness condition on \( f \).

**Theorem 3.1.** Let \( \hat{f} \) be the estimator (2.4) with \( n \geq m = \lceil s \rceil - 1 \geq 2 \) and \( \text{Rem}_m \) be the remainder term in (2.7). Let \( p \geq 1 \) and \( C_{m,n}^* = \sum_{k=0}^{m} n^k (n-k)! / \{(m-k)!n!\} \). Then,

\[
\| \text{Rem}_m \|_{L_p(\mathbb{P})} \leq C_{m,n}^* \max \left\{ \| \bar{\varepsilon} \|_{L_p(\mathbb{P})}^{s}, \left\| \frac{1}{n^2} \sum_{j=1}^{n} \| \varepsilon_j \|_{L_p(\mathbb{P})}^{s/2} \right\|_{L_p(\mathbb{P})} \right\}
\]

for any (possibly with nonzero mean and dependent) random \( \varepsilon_j = x_j - \theta \). Moreover, if \( \varepsilon_j \) are independent with \( \mathbb{E}[\varepsilon_j] = 0 \), then for some constant \( C_{p,s}^* \) depending on \( p, s \) only

\[
\mathbb{E}[\hat{f} - f(\theta)] \leq \| \text{Rem}_m \|_{L_p(\mathbb{P})} \leq C_{p,s}^* \| f \|_{(s)} \left\{ \| \bar{\varepsilon} \|_{L_2(\mathbb{P})}^s + \left( \sum_{j=1}^{n} \| \varepsilon_j \|_{L_2(\mathbb{P})}^{p} \right)^{1/p} \right\}.
\]

We note that the \( L_1(\mathbb{P}) \) norm of the remainder in (2.7) is sufficient to bound the bias. The \( L_p(\mathbb{P}) \) bounds in Theorem 3.1 will be useful in the development below of bounds for the \( L_p(\mathbb{P}) \) risk of \( \hat{f} \). It follows from the upper bound for \( C_{k,n} \) in Lemma 2.2 that \( C_{m,n}^* \leq e^{1+((m-1)m/n)} \).

Theorem 3.1 requires \( s \)-th moment of the noise for bias correction of the same order. For i.i.d. \( \varepsilon_j \) and \( p = 1 \), \( \| \bar{\varepsilon} \|_{L_2(\mathbb{P})}^2 = \| \varepsilon_1 \|_{L_2(\mathbb{P})}^2 / n \) so that (3.3) can be written as

\[
\mathbb{E}[\hat{f} - f(\theta)] \leq \| \text{Rem}_m \|_{L_1(\mathbb{P})} \leq C_{p,s}^* \| f \|_{(s)} \left\{ \frac{\| \varepsilon_1 \|_{L_2(\mathbb{P})}^s}{n^{s/2}} + \frac{\| \varepsilon_1 \|_{L_2(\mathbb{P})}^s}{n^{s-1}} \right\},
\]

and the \( s \)-th moment term is subsumed into the second moment term \( \| \bar{\varepsilon} \|_{L_2(\mathbb{P})}^s \) iff

\[
\| \varepsilon_1 \|_{L_2(\mathbb{P})} = O(n^{1/2-1/s}).
\]
In the upper bounds in Theorem 3.1, we may replace \( \|f\|_s \) by \( 2\|f\|_s,\theta \) where
\[
\|f\|_s,\theta = \sup_{\|h\|>0} \|f^{(m)}(\theta + h) - f^{(m)}(\theta)\|_S/\|h\|^{s-m}
\]
with \( m = \lceil s \rceil - 1 \). As \( C^*_s \) is implicit, Theorem 3.1 is most useful when \( \mathcal{H} \) is high- or infinite-dimensional but \( s > 0 \) is fixed.

The \( L_p(\mathbb{P}) \) bounds for the remainder in (3.3) or (3.4) naturally lead to \( L_p(\mathbb{P}) \) risk bounds for \( \hat{f} \) given any \( L_p(\mathbb{P}) \) bound for the completely degenerate \( U \) variables \( \langle f^{(k)}(\theta), \varepsilon^{(k)} \rangle_k \) in (2.6). However, as we are interested in the case where bias correction with \( s > 2 \) is needed to improve the naive plug-in estimator \( f(\bar{x}) \), the following \( L_2(\mathbb{P}) \) bounds will be used.

Write the degenerate \( U \)-variables in the expansion (2.6) as a sum of uncorrelated terms,
\[
S_k = \langle f^{(k)}(\theta), \varepsilon^{(k)} \rangle_k = \sum_{1 \leq j_1 < \cdots < j_k \leq n} \frac{\langle f^{(k)}(\theta), \varepsilon_{j_1} \otimes \cdots \otimes \varepsilon_{j_k} \rangle_k}{n(n-1) \cdots (n-k+1)/k!}
\]
For \( k \geq 2 \) and i.i.d. \( \varepsilon_j \), the variance of \( S_k \) can be explicitly expressed as
\[
\text{Var}(S_k) = \mathbb{E}[S_k^2] = (C_{k,n} k!) V_k / n^k,
\]
with the constant \( C_{k,n} = (n^k - 1)/n! \leq e^{(k-1)k/n} \) in Lemma 2.2 and
\[
V_k = \langle f^{(k)}(\theta), f^{(k)}(\theta) \times_1 \Sigma \times_2 \cdots \times_k \Sigma \rangle_k,
\]
where \( \Sigma \) is the covariance operator in (3.1) and \( \times_\ell \) denotes the mode-\( \ell \) tensor product. For independent not identically distributed (i.n.i.d.) noise \( \varepsilon_j \),
\[
\text{Var}(S_k) = \mathbb{E}[S_k^2] \leq (C_{k,n} k!) V_k / n^k.
\]
With the tensor spectrum norm \( \| \cdot \|_S \) and the effective rank in (3.1), \( V_k \) is bounded by
\[
V_k \leq \|f^{(k)}(\theta)\|_S^2 \sigma^2 k (r(\Sigma))^{k-1}.
\]

The following theorem is based on the \( L_p(\mathbb{P}) \) bound (3.3) for the remainder and the \( L_2(\mathbb{P}) \) bounds for the degenerate \( U \)-tensors via (3.11).

**Theorem 3.2.** Let \( n \geq m = \lceil s \rceil - 1 \geq 2 \geq p \geq 1 \) and \( \hat{f} \) be as in (2.4) with independent observations \( x_j = \varepsilon_j + \theta \) with \( \mathbb{E}[\varepsilon_j] = 0 \). Then,
\[
\|\hat{f} - f(\theta)\|_{L_p(\mathbb{P})} \leq \left( \sum_{k=1}^m \frac{C_{k,n}^2 V_k}{n^k k!} \right)^{1/2} + C^*_s \|f\|_s \left( \frac{\sigma^2 r}{n} \right)^{s/2}
\]
\[
+ C^*_s \|f\|_s \left( \sum_{j=1}^n \frac{\|\varepsilon_j\|_{L_p(\mathbb{P})}}{n^{ps}} \right)^{1/p}
\]
with the \( \sigma \) and \( r = r(\Sigma) \) in (3.1), \( V_k \) in (3.9), \( C_{k,n} \leq e^{(k-1)k/n} \) and a constant \( C^*_s \) depending on \( s \) only. Moreover, if \( \max_{2 \leq k < s} \|f^{(k)}(\theta)\|_S \leq C_0 \) and \( \|f\|_s,\theta \leq C_0 \), then
\[
\|\hat{f} - f(\theta)\|_{L_p(\mathbb{P})} \leq \left( \frac{V_1}{n} \right)^{1/2} + C_0 C^*_s \max \left\{ \frac{\sigma^2 \sqrt{r}}{n}, \left( \frac{\sigma^2 r}{n} \right)^{s/2} \right\}
\]
\[
+ C_0 C^*_s \left( \sum_{j=1}^n \frac{\|\varepsilon_j\|_{L_p(\mathbb{P})}}{n^{ps}} \right)^{1/p}.
\]
Theorem 3.2 is proved in Section 6 along with proofs of (3.8), (3.10) and (3.11). Compared with [KZ21] in the literature where \( \varepsilon_j \) are assumed to be i.i.d. Gaussian, Theorem 3.2 requires the \( ps \)-th moment condition in the third component on the right-hand side of (3.12). In the i.i.d. Gaussian case, we may write \( \varepsilon_1 = \sum \lambda_{1/2}^j z_{\ell} v_\ell \) with i.i.d. \( z_\ell \sim \mathcal{N}(0,1) \) via the eigenvalue decomposition \( \Sigma = \sum \lambda_{1/2}^j v_\ell \otimes v_\ell \), so that for even integers \( ps \geq 2 \),

\[
(3.14) \quad \sum_{j=1}^{n} \left\| \varepsilon_j \right\|_{L_2(P)}^{ps/n} = n^{1-p/s} \mathbb{E} \left[ \left( \sum \lambda_{1/2}^j z_\ell^2 \right)^{ps/2} \right] \leq n^{1-ps/2} (\sigma^2 r/n)^{ps/2} (ps - 1)!!.
\]

Especially, when \( r = o(n) \) and we take \( p = 2 \) and \( s > 1 \), (3.14) is of an asymptotic order smaller than \( O((r/n)^{s/2}) \). As a result, bound (3.13) reproduces the bound on MSE: \( \left\| \hat{f} - f(\theta) \right\|_{L_2(P)}^2 \lesssim (n^{-1} \lor (r/n)^{s}) \), which was proved in [KZ21] under the i.i.d. Gaussian assumption and shown to be minimax optimal under the standard Gaussian shift model.

4. Asymptotic normality. In this section we develop asymptotic normality theory for the proposed estimator (2.4) of \( f(\theta) \).

We have developed upper bounds for the bias and \( L_p \) risk when the \( ps \)-moment of the noise does not grow too fast. The moment condition is natural for the \( L_p \) risk with expansions of order \( s \), and the risk bound can be used to remove higher order terms in the asymptotic normality analysis. However, such an approach would require higher moment condition than necessary. In Theorem 4.1 below, asymptotic normality requires only the Lindeberg condition on the linear term and a mild condition on the growth rate of the second moment, provided the Hölder smoothness of \( f(\cdot) \) at \( \theta \). In addition to the Hölder smoothness condition in the spectrum norm, we shall consider its counterpart in the Hilbert-Schmidt norm,

\[
\|f\|_{(s),HS,\theta} = \sup_{\|h\| > 0} \left\| f^{(m)}(\theta + h) - f^{(m)}(\theta) \right\|_{HS}/\|h\|^{s-m}.
\]

**Theorem 4.1.** Let \( n \geq m = [s] - 1 \geq 2 \) and \( \hat{f} \) be as in (2.4) based on independent observations \( x_j = \varepsilon_j + \theta \in \mathcal{H}, j \leq n \), with \( \mathbb{E}[\varepsilon_j] = 0 \). Let \( V_1 = \langle f^{(1)}(\theta) \otimes f^{(1)}(\theta), \Sigma \rangle_2 \) be as in (3.9), \( \sigma \) and \( r(\Sigma) \) be the noise level and effective rank as in (3.1), and \( \|f\|_{(s),\theta} \) be as in (3.6). Suppose \( \max_{2 \leq k < s} \|f^{(k)}(\theta)\|_S \leq C_0 \) and \( \|f\|_{(s),\theta} \leq C_0 \).

\[
(4.1) \quad C_0 \max \left\{ \sigma^2 r^{1/2}(\Sigma)/n, (\sigma^2 r(\Sigma)/n)^{s/2} \right\} \ll (V_1/n)^{1/2}, \quad s = O(1),
\]

and the Lindeberg condition holds for \( \{\langle f^{(1)}(\theta), \varepsilon_j \rangle, j \leq n \} \). Then,

\[
(4.2) \quad n^{1/2} (\hat{f} - f(\theta))/V_1^{1/2} \Rightarrow \mathcal{N}(0,1) \quad \text{as } n \to \infty.
\]

The asymptotic normality (4.2) still holds when the condition \( s = O(1) \) is weakened to \( s \leq n \) in (4.1) and the smoothness condition on \( f \) is replaced by \( \max_{2 \leq k < s} \|f^{(k)}(\theta)\|_S/d^{s/2} \leq C_0 \) in the spectrum norm and \( \|f\|_{(s),HS,\theta} \leq C_0 d^{m/2} \) in the Hilbert-Schmidt norm with any fixed positive real number \( d \).

In Theorem 4.1, all quantities, including \( \mathcal{H}, f : \mathcal{H} \to \mathbb{R}, \sigma, r(\Sigma) \) and \( C_0 \) are allowed to change with \( n \). While \( d \) is just a constant in Theorem 4.1, for \( \mathcal{H} = \mathbb{R}^d \) we have \( \|f^{(k)}(x)\|_{HS}^2 \leq \|f^{(k)}(x)\|_{S}^2 d^{k-1} \) and \( \|f\|_{(s),\theta}^2 \leq d^{m-1} \|f\|_{(s),\theta}^2 \) because \( f^{(m)}(x) \times_{k=1}^{m-2} e_{i_k} \) is a \( d \times d \) symmetric matrix given \( x \) and canonical basis vectors \( e_{i_1}, \ldots, e_{i_{m-2}}, 1 \leq i_k \leq d \). For general \( \mathcal{H}, d \) can be viewed as the effective rank of \( f(\cdot) \). Thus, the last statement of Theorem 4.1 asserts that the boundedness condition \( s = O(1) \) in (4.1) on the order of the expansion can
be replaced by the boundedness condition $d = O(1)$ on the effective rank of the functional $f(\cdot)$.

Both components on the left-hand side of the first condition in (4.1) are needed. For example, when $C_0 \asymp 1$ and $\sigma \asymp 1$, the component $r^{1/2}(\Sigma)/n$ can be omitted when $V_1 = o(1)$ but needed when $V_1 = o(1)$. Note that in a previous work [KZ21], asymptotic normality was proved under i.i.d. Gaussian assumption on $\varepsilon_j$’s. Here, only independence is required. Especially, an interesting result of [KZ21] is that when $V_1 \asymp 1$ and $r(\Sigma) = n^\alpha$, the condition $s > 1/(1 - \alpha)$ is sharp for (4.2) in the Gaussian shift model. In Theorem 4.1, (4.1) matches this sharp threshold level on the smoothness index $s$ under the first and second moment conditions on the noise without additional distributional assumption.

5. Estimation of non-smooth additive functions. In this section we consider the estimation of additive functions of the form

$$f(\theta) = \frac{1}{d} \sum_{a=1}^{d} f_0(\theta_a), \quad \theta = (\theta_1, \ldots, \theta_d)^T,$$

for a given but possibly non-smooth function $f_0(\cdot)$, based on independent observations $x_j = (x_{j,1}, \ldots, x_{j,d})^T$ with $E[x_j] = \theta$.

We allow general $f_0$ satisfying a smoothness condition of small order, including for example $f_0(x) = |x|^p$ with small $p > 0$. Such functions are called non-smooth in the literature as their order of smoothness is much smaller than the order of differentiability used in the analysis.

We assume the following conditions on the noise:

$$E[\varepsilon_{j,a}\varepsilon_{j,b}] = 0, \quad 1 \leq j \leq n, 1 \leq a < b \leq d,$$

$$E\left[|n^{-1} \sum_{j=1}^{n} \varepsilon_{j,a}^2 \varepsilon_{j,b}^{2k}| \right] \leq \sigma_n^2 k! 2^{k-1} k!, \quad 1 \leq a \leq d, 1 \leq k \leq \lceil s \rceil,$$

where $\varepsilon_{j,a} = x_{j,a} - \theta_a$ is the noise for the $a$-th component of the $j$-th observation, and $\{\varepsilon_{j,a}^2, 1 \leq j \leq n\}$ is a Rademacher sequence independent of $\{\varepsilon_{j,a}, 1 \leq j \leq n\}$.

We may set $\sigma_n = 1$ so that the model matches the Gaussian shift model where only a single instance of $\mathcal{N}(\theta, I_d)$ is observed. Typically in this setting, consistent estimation of a single $f_0(\theta_a)$ is infeasible as there is only one $\mathcal{N}(\theta_a, 1)$ observation, and the consistent estimation of $f(\theta)$ is achieved through high-order bias correction and noise averaging over data components $a = 1, \ldots, d$. Thus, for $\sigma_n \asymp 1$ and general $f_0$ we assume $d \to \infty$ and apply HODSE with $m \to \infty$ in (2.4). Two important features of our approach become more crucial when $m \to \infty$, namely (a): the explicit description of the remainder term in Proposition 2.1; and (b) sharp control of constant factors in Lemma 2.2 for the analysis of the estimator (2.4). Of course, our approach also allows $\sigma_n \to 0$ and $\sigma_n \to \infty$.

The first line of (5.2) asserts that the elements $\varepsilon_{j,a}$ of the noise vector $\varepsilon_j$ are uncorrelated for each $j$ in addition to the independence of $\varepsilon_1, \ldots, \varepsilon_n$. The second line of (5.2) can be viewed as a sub-Gaussian condition which holds when $E[\varepsilon_{j,a}^{2k}] \leq E[|\mathcal{N}(0, n\sigma_n^2)|^{2k}]$ for all positive integers $k \leq s$ due to the independence of $\varepsilon_1, \ldots, \varepsilon_n$. For example, (5.2) holds when $\varepsilon_{j,a}$ are i.i.d. with $P_{\{\varepsilon_{j,a} = \pm n^{1/2}\sigma_n\}} = 1/2$. The i.i.d. $\varepsilon_{j,a} \sim \mathcal{N}(0, n)$ assumption, which implies (5.2) with $\sigma_n = 1$, is equivalent to the standard Gaussian shift model due to the sufficiency of $x \sim \mathcal{N}(\theta, I_d)$. However, (5.2) is much weaker than the Gaussian shift model as the components of $x_j$ are only required to be uncorrelated, and non-Gaussian data and heteroscedasticity are allowed.

In [CL11] the authors proved that based on a single $\mathcal{N}(\theta, I_d)$ observation, the minimax rate for the estimation of the length normalized $\ell_1$ norm $f(\theta) = ||\theta||_1/d$ is

$$\inf_{\tilde{f}} \sup_{\theta \in \mathbb{R}^d} E \left[ (\tilde{f} - f(\theta))^2 \right] \asymp (\log d)^{-1}.$$
We note that the plug-in estimator is inconsistent here as there is only one observation available and \( f_0(x) = |x| \) is not analytic. As the \( \mathcal{N}(\theta, I_d) \) model is a special case of (5.2), our results in this section will extend their upper bound to more general \( f_0 \) and heteroscedastic non-Gaussian data.

Our idea is to first apply kernel smoothing to the function \( f_0 \).

(5.3) \[ f_h(x) := (K_h * f_0)(x) = \int K_h(x-t)f_0(t)dt, \] with \( K_h(x) := h^{-1}K(x/h) \),

with proper a kernel \( K(x) \) and a bandwidth \( h > 0 \), and then apply the proposed (2.4) to estimate the individual \( f_h(\theta_a) \), \( a = 1, \ldots, d \). This leads to the estimator

(5.4) \[ \hat{f} := \frac{1}{d} \sum_{a=1}^{d} \left\{ f_h(\bar{x}_a) + \sum_{k=2}^{m} \frac{f_h^{(k)}(\bar{x}_a)\bar{u}_a^{(k)}}{k!} \right\} \]

with \( \bar{x}_a = \frac{\sum_{j=1}^{n} x_{j,a}}{n} \) and \( \bar{u}_a^{(k)} = \{(n-k)!/n!\}\sum_{1 \leq j_1 \neq \cdots \neq j_k \leq n} \prod_{\ell=1}^{k} (x_{j_\ell,a} - \bar{x}_a) \).

Let \( m = s - 1 \leq n \) for some integer \( s > 2 \). By Proposition 2.1 and (2.9) of Lemma 2.2,

(5.5) \[ \hat{f} = \frac{1}{d} \sum_{a=1}^{d} f_h(\theta_a) + \sum_{k=1}^{m} \frac{1}{d} \sum_{a=1}^{d} \frac{f_h^{(k)}(\theta_a)\bar{\varepsilon}_a^{(k)}}{k!} - \frac{1}{d} \sum_{a=1}^{d} \text{Rem}_{m,a} \]

with completely degenerate \( \bar{\varepsilon}_a^{(k)} = \{(n-k)!/n!\}\sum_{1 \leq j_1 \neq \cdots \neq j_k \leq n} \prod_{\ell=1}^{k} \varepsilon_{j_\ell,a} \) and

(5.6) \[ |\text{Rem}_{m,a}| \leq \| f_h \| (s) \sum_{k=0}^{m} \frac{|\bar{\varepsilon}_a| s-k |\bar{\varepsilon}_a^{(k)}|}{(s-k)!k!} \]

where \( \bar{\varepsilon}_a = n^{-1} \sum_{j=1}^{n} \varepsilon_{j,a} \) is the average of the \( a \)-th noise component.

It can be seen from (5.5) that while the kernel smoothing (5.3) introduces some bias, correction of additional bias is achieved through the application of HODSE in the estimation of individual \( f_h(\theta_a) \), and de-noising is achieved through the averaging in the degenerate \( \bar{\varepsilon}_a^{(k)} \) and over the \( d \)-coordinates and terms of different orders in the second term.

By (5.2), \( \bar{\varepsilon}_a^{(k)} \), \( 1 \leq a \leq d \), \( 1 \leq k \leq m \), are uncorrelated with each other and

(5.7) \[ \mathbb{E} \left[ \left( \sum_{a=1}^{d} \sum_{k=1}^{m} \frac{f_h^{(k)}(\theta_a)\bar{\varepsilon}_a^{(k)}}{d(k!)} \right)^2 \right] \leq \sum_{a=1}^{d} \sum_{k=1}^{m} C_{k,n}^2 \left( f_h^{(k)}(\theta_a) \right)^2 \sigma_n^{2k} \]

with \( C_{k,n} \leq e^{(k-1)k/n} \). By the second line of (5.2), the remainder is bounded by

(5.8) \[ \mathbb{E} \left[ \left( \frac{1}{d} \sum_{a=1}^{d} |\text{Rem}_{m,a}| \right)^2 \right] \leq C_{s,n}^2 \| f_h \| (s)^2 \sigma_n^{2s} \square^{-1} / s! \]

The benefit of the neat formulas (2.6) and (2.7) is evident in this analysis in view of (5.8) in the case of \( \sigma_n \approx 1 \). It follows from (5.5), (5.7) and (5.8) that for integers \( s \) satisfying \( s > 2 \) and \( 2(s-1)s/n \leq \log 2 \) the error of the estimator (5.4) is bounded by

(5.9) \[ \| \hat{f} - f(\theta) \|_{L_2(\mathbb{P})} \leq \left\{ \text{bias}_{n,d}(\theta) + \kappa_{s,h,n,d}(\theta) \right\}^{1/2} + \| f_h \| (s) (2^{s/2} \sigma_n)^s / \sqrt{s!} \]

for the \( L_2(\mathbb{P}) \) risk given by \( \| \hat{f} - f(\theta) \|_{L_2(\mathbb{P})}^2 = \mathbb{E} \left[ (\hat{f} - f(\theta))^2 \right] \), where

(5.10) \[ \text{bias}_{h,d}(\theta) = \frac{1}{d} \sum_{a=1}^{d} (f_0 - f_h)(\theta_a), \quad \kappa_{t,h,n,d}(\theta) = \sum_{k=1}^{\lceil r \rceil - 1} \frac{\| f_h^{(k)}(\theta) \|_{L_2(\mathbb{P})}^2 \sigma_n^{2k}}{(d^2/2)k!} \]
are respectively the bias introduced by (5.3) and an upper bound for the variance in (5.7).

The convergence rate and sharpness of (5.9) would heavily depend on the smoothness properties of \( f_0 \) and \( f_h \) through the choice of the kernel \( K(\cdot) \) in (5.3). The following theorem provides an explicit error bound in an “\( \alpha \)-smooth” scenario, allowing arbitrarily small smoothness index \( \alpha \) for the function \( f_0(\cdot) \).

**Theorem 5.1.** Let \( s \) be an integer satisfying \( 2 < s \leq \sqrt{(n/2)\log 2} \). Let \( f(\theta) \) be as in (5.1) and \( \hat{f} \) as in (5.4) with \( m = s - 1 \). If (5.2) holds, then (5.9) holds. If (5.2) holds and

\[
\|f^{(k)}_0 - f^{(k)}_h\|_{L_\infty} \leq \eta_\alpha(h)/h^k, \quad 0 \leq k < \alpha,
\]

\[
\|f^{(k)}_h\|_{L_\infty} \leq \eta_\alpha(h)/h^k, \quad \alpha \leq k \leq s,
\]

for certain \( \alpha > 0 \) and \( \eta_\alpha(h) > 0 \) for the \( h \) and \( f_h \) in (5.3) with \( (\sigma_n/h)^2 \leq s/(2^7) \), then

\[
\|\hat{f} - f(\theta)\|_{L_2(\mathbb{P})} \leq \|\text{bias}_{h,d}(\theta)\| + \eta_\alpha(h)\left(s^{-1/4} + \sqrt{2/d\sigma_n/h^7/2}\right) + \kappa_{\alpha,0,n,d}(\theta)^{1/2}
\]

(5.12)

where \( \text{bias}_{h,d}(\theta) \) and \( \kappa_{\alpha,h,n,d}(\theta) \) are as defined in (5.10).

**Corollary 5.2.** Suppose (5.2) and (5.11) hold with \( \eta_\alpha(h) = C_\alpha h^\alpha \) for some constants \( \alpha \in (0, 1) \) and \( C_\alpha < \infty \). When \( (\sigma_n/h)^2 = \log(d/\log d) \) in (5.12),

\[
\|\hat{f} - f(\theta)\|_{L_2(\mathbb{P})} \leq \|\text{bias}_{h,d}(\theta)\| + s^{-1/4} + \sqrt{2/d\log d}C_\sigma \sigma_n^\alpha/(\log(d/\log d))^{\alpha/2}
\]

(5.13)

We notice that when \( \varepsilon_{j,a} \) are i.i.d. \( \mathcal{N}(0, \sigma^2) \), we have \( \sigma_n = \sigma/\sqrt{n} \), so that for \( \alpha = 1 \) the convergence rate in (5.13) is faster than the usual parametric rate due to the noise averaging over components \( a = 1, \ldots, d \). At high noise level \( \sigma_n \gg 1 \), consistent estimation of \( f(\theta) \) is still feasible but the convergence rate is logarithmic for general \( f_0 \).

Interestingly, in (5.13), the condition on the order of expansion \( m = s - 1 \) in the construction of the estimator (5.4), \( 2^7 \varepsilon \log(d/\log d) \leq s \leq \sqrt{(n/2)\log 2} \), depends on \( n \) and \( d \) only, not on the noise level \( \sigma_n \), while the choice of the bandwidth in (5.3), \( h = \sigma_n/\sqrt{\log(d/\log d)} \), is proportional to the noise level. We also note from the theorem that the risk is typically dominated by the bias in the kernel smoothing step. This phenomenon also presents in some other problems with logarithmic convergence rates such as deconvolution [Zha90, Fan91].

Theorem 5.1 is proved in Section 6 along with detailed derivations of (5.7) and (5.8). In the following two subsections, we verify condition (5.11) for specific \( f_0 \) and choice of the kernel \( K(x) \) in (5.3). We note that while our general solution achieves the existing optimal minimax rate in these two examples, the existing estimator uses polynomial approximation specifically constructed for the individual \( f_0 \).

### 5.1. Estimation of the \( \ell_1 \)-norm

For \( f_0(x) = |x| \), the function is 1-Lipschitz and \( f_0^{(2)}(x) = 2K(x) = 2h^{-1}K(x/h) \) in (5.3). Thus, condition (5.11) holds with \( \alpha = 1 \) and \( \eta_1(h) = C_1 h \) for all integers \( s > 2 \) when the kernel in (5.3) satisfies

\[
\int K(x)dx = 1, \quad \int |x^\ell K(x)|dx \leq C_1, \quad 2\|K^{(k-2)}\|_{L_\infty} \leq C_1,
\]

(5.14)

for \( \ell \in \{0, 1\} \) and \( 2 \leq k \leq s \). When \( K(\cdot) \) is the Fourier inversion of a twice continuously differentiable function \( Q(\cdot) \) with support \([-1, 1]\), these conditions on \( K(\cdot) \) hold when

\[
Q(0) = \frac{1}{\sqrt{2\pi}}, \quad \text{supp}(Q) = [-1, 1], \quad \|Q\|_{L_1} \leq \frac{C_1}{\sqrt{2/\pi}}, \quad \max_{k=0,2} \|Q^{(k)}\|_{L_2} \leq \frac{C_1}{\sqrt{2\pi}}
\]

(5.15)
The verification of the above claims is elementary but will be included in the proof of the following theorem for completeness.

**Theorem 5.3.** Let \( s = [2^7 e \log(d/\log d)] \) and \( m = s - 1 \). Suppose (5.2) holds and \( s \leq \sqrt{(n/2) \log 2} \). Let \( f_h \) be as in (5.3) with \( f_0(x) = |x|, h = \sigma_n/\sqrt{\log(d/\log d)} \) and a kernel \( K(\cdot) \) satisfying (5.14). Let \( f(\theta) = \|\theta\|_1/d \) and \( \hat{f} \) be as in (5.4). Then,
\[
\mathbb{E} \left[ |\hat{f} - f(\theta)|^2 \right] \leq \left( 1 + \sqrt{2/\log d} + s^{-1/4} \right)^2 C_2^2 \sigma_n^2 / \log(d/\log d).
\]

Moreover, (5.14) holds when \( K(\cdot) \) is the Fourier inversion of a function \( Q \) satisfying (5.15).

The above theorem demonstrates that at the high-noise level \( \sigma_n = 1 \) (e.g. \( \text{Var}(\varepsilon_{j,a}) = n \)), the degenerate statistical expansion (2.4) in high-order \( s \asymp \sqrt{\log d} \) yields the rate optimal minimax upper bound in [CL11]. The second conclusion of the theorem also holds when \( K(x) \) is taken as the real part of the Fourier inversion of a function \( Q \) satisfying (5.15).

#### 5.2. Estimation of the \( \ell_p \)-norm

For \( f_0(x) = |x|^p \) with \( p \in (0, 1) \), Theorem 5.1 is still applicable when \( K(\cdot) \) is the inverse Fourier transform of a function \( Q(\cdot) \) satisfying (5.15).

**Theorem 5.4.** Let \( f(\theta) = \|\theta\|_p^p/d \) with \( 0 < p < 1 \) and \( f_0(x) = |x|^p \) in (5.1). Let \( f_h \) be as in (5.3) with \( h = \sigma_n/\sqrt{\log(d/\log d)} \) and a kernel \( K(\cdot) \) being the inverse Fourier transform of \( Q(\cdot) \) satisfying (5.15). Let \( s = [2^7 e \log(d/\log d)] \) and \( \hat{f} \) be the estimator in (5.4) with \( m = s - 1 \). Suppose (5.2) holds and \( s \leq \sqrt{(n/2) \log 2} \). Then,
\[
\mathbb{E} \left[ |\hat{f} - f(\theta)|^2 \right] \leq \left( 1 + \sqrt{2/\log d} + s^{-1/4} \right)^2 C_1^2 (\sigma_n^2 / \log(d/\log d))^p.
\]

Again, the degenerate statistical expansion (2.4) in high-order \( s \asymp \sqrt{\log d} \) yields the upper bound \( O(\sigma_n^{2p} / (\log d)^p) \) on MSE which was shown to be minimax optimal by [CCT20] in the dense region. We note that the estimator in [CCT20] is implicitly defined through the optimal \((2m)\)-degree polynomial approximation of \(|x|^p \) under the \( L_\infty \) norm on \([-1, 1] \) with integer \( m \asymp \log d \).

### 6. Proofs

**Proof of Proposition 2.1.** An order \( k \) tensor \( T^{(k)} \) is symmetric if
\[
\langle T^{(k)}, v_1 \otimes \cdots \otimes v_k \rangle_k = \langle T^{(k)}, v_j, \otimes \cdots \otimes v_j \rangle_k
\]
for all permutations of indices \( \{j_1, \ldots, j_k\} = \{1, \ldots, k\} \). By algebra and the definition of \( \check{u}^{(k)} \), for any symmetric tensor \( T^{(k)} \)
\[
\langle T^{(k)}, \check{u}^{(k)} \rangle_k = \sum_{1 \leq j_1 \neq \cdots \neq j_k \leq n} \frac{\langle T^{(k)}, (x_{j_1} - \bar{x}) \otimes \cdots \otimes (x_{j_k} - \bar{x}) \rangle_k}{n(n-1) \cdots (n-k+1)}
\]
\[
= \sum_{1 \leq j_1 \neq \cdots \neq j_k \leq n} \frac{\langle T^{(k)}, (\varepsilon_{j_1} - \bar{\varepsilon}) \otimes \cdots \otimes (\varepsilon_{j_k} - \bar{\varepsilon}) \rangle_k}{n(n-1) \cdots (n-k+1)}
\]
\[
= \sum_{j=0}^{k} \binom{k}{j} (-1)^j \langle T^{(k)}, \bar{\varepsilon}^{\otimes j} \hat{\varepsilon}^{(k-j)} \rangle_k.
\]

Let \( h_k(y) = \langle f^{(k)}(\bar{x} + y(\theta - \bar{x})), \bar{\varepsilon}^{(k)} \rangle_k \). We have \( h_k(1) = \langle f^{(k)}(\theta), \bar{\varepsilon}^{(k)} \rangle_k \) and \( h_k(0) = \langle f^{(k)}(\bar{x}), \bar{\varepsilon}^{(k)} \rangle_k \). Let \( (J_\alpha h)(y) = \int_0^y h(t)(y - t)^{\alpha-1} dt / \Gamma(\alpha) \) be the Riemann–Liouville

integral, \(J^\alpha h = (J^\alpha h)(1)\), and \(\Delta_j^j(t) = h_j^j(t) - h_j^j(0)\) for any function \(h(t)\). We write the Taylor expansion of \(h_k(y)\) as

\[
h_k(y) = \sum_{j=0}^{m-k-1} \frac{h_k^j(0)}{j!} y^j + (J^{m-k}h_k^{m-k})(y) = \sum_{j=0}^{m-k} \frac{h_k^j(0)}{j!} y^j + (J^{m-k}\Delta_k^{m-k})(y).
\]

As \(\theta - \bar{x} = -\bar{e}\),

\[
h_k^j(y) = \langle f(k+j)(\bar{x} + y(\theta - \bar{x})), (-\bar{e})^\otimes j \otimes \bar{e}^{(k)} \rangle_{k+j}.
\]

Thus, taking \(y = 1\) in the expression for \(h_k(y)\), we find that the Taylor expansion for \(h_k(1)\) is

\[
\langle f(k)\theta, \bar{e}^{(k)} \rangle_k = \sum_{j=0}^{m-k} \frac{h_k^j(0)}{j!} + J^{m-k}\Delta_k^{m-k} = \sum_{j=0}^{m-k} \frac{\langle f(k+j)\bar{x}, \bar{e}^\otimes j \otimes \bar{e}^{(k)} \rangle_{k+j}}{(-1)^j j!} + \frac{\langle J^{m-k}\Delta_k^{m-k}, \bar{e}^\otimes (m-k) \otimes \bar{e}^{(k)} \rangle_m}{(-1)^{m-k}}.
\]

Summing the above expression over \(k = 0, \ldots, m\), we have

\[
\sum_{k=0}^m \frac{\langle f(k)\theta, \bar{e}^{(k)} \rangle_k}{k!} = \sum_{k=0}^m \sum_{j=0}^{m-k} \frac{\langle f(k+j)\bar{x}, \bar{e}^\otimes j \otimes \bar{e}^{(k)} \rangle_k}{(-1)^j j! k!} + \text{Rem}_m \tag{6.2}
\]

with the remainder term

\[
\text{Rem}_m = \sum_{k=0}^m \frac{\langle J^{m-k}\Delta_k^{m-k}, \bar{e}^\otimes (m-k) \otimes \bar{e}^{(k)} \rangle_m}{(-1)^{m-k} k!}
\]

in (2.7). Let \(\ell + j = k\). As \(f(k)\bar{x}\) are symmetric tensors,

\[
\hat{f} = \sum_{k=0}^m \frac{\langle f(k)\bar{x}, \bar{u}^{(k)} \rangle_k}{k!} = \sum_{0 \leq j \leq l \leq m} \binom{k}{j} \frac{\langle f(k)\bar{x}, \bar{e}^\otimes j \otimes \bar{e}^{(k-j)} \rangle_k}{(-1)^j j!} = \sum_{j=0}^{m-l} \frac{\langle f(l+j)\bar{x}, \bar{e}^\otimes j \otimes \bar{e}^{(l)} \rangle_{l+j}}{(-1)^j j! l!}
\]

by the formula for \(\bar{u}^{(k)}\) in (6.1). This and (6.2) yield (2.6). □

**Proof of Lemma 2.2.** By the definition of the norm \(\|T^{(m)}\|_{m-k,k}\),

\[
|\text{Rem}_m| \leq \sum_{k=0}^m \left( \max_{0 \leq t \leq 1} \frac{\|\Delta_k^{m}(t)\|_{m-k,k}}{(t\|\bar{e}\|)^{s-m}} \right) \int_0^1 \frac{t^{s-m}\|\bar{e}\|_{s-k,k}\|\bar{e}^{(k)}\|_{HS}}{(m-k-1)!k!} (1-t)^{m-k-1}dt.
\]

Then inequality (2.9) follows directly by applying the following fact to (6):

\[
\int_0^1 \frac{t^{s-m}(1-t)^{m-k-1}}{(m-k-1)!k!} dt = \frac{\Gamma(s-m+1)\Gamma(m-k)}{\Gamma(s-k+1)(m-k-1)!} = \frac{\Gamma(s-m+1)}{\Gamma(s-k)} \leq \frac{1}{\Gamma(s-k+1)}.
\]

The inequality follows from the convexity of \(\Gamma(x)\) in \([1, 2]\) and \(\Gamma(2) = \Gamma(1) = 1\).
To prove (2.10), we need to write $\bar{\varepsilon}^{(k)}$ as a sum of rank-one tensors before we can apply the spectrum norm on $\Delta^{(m)}(t)$. To this end we observe that for any order $k$ symmetric tensor $T^{(k)}$ and the completely degenerate tensor $\bar{\varepsilon}^{(k)}$ in (2.3)

$$
\left(\frac{n^k}{C_{k,n}}\right)\langle T^{(k)}, \bar{\varepsilon}^{(k)} \rangle_k
$$

$$
= \left\langle T^{(k)}, \sum_{1 \leq j_1 \neq \cdots \neq j_k \leq n} \varepsilon_{j_1} \otimes \cdots \otimes \varepsilon_{j_k} \right\rangle_k
$$

$$
= \left\langle T^{(k)}, \sum_{1 \leq j_1 \neq \cdots \neq j_{k-1} \leq n} \varepsilon_{j_1} \otimes \cdots \otimes \varepsilon_{j_{k-1}} \otimes \left(n\bar{\varepsilon} - \sum_{a=1}^{k-1} \varepsilon_{j_a} \right) \right\rangle_k
$$

$$
= \left\langle T^{(k)}, (n\bar{\varepsilon}) \otimes \sum_{1 \leq j_1 \neq \cdots \neq j_{k-1} \leq n} \varepsilon_{j_1} \otimes \cdots \otimes \varepsilon_{j_{k-1}} \right\rangle_k
$$

$$
- (k-1)\left\langle T^{(k)}, (n\bar{\varepsilon}) \otimes \sum_{1 \leq j_1 \neq \cdots \neq j_{k-1} \leq n} \varepsilon_{j_1}^{\otimes 2} \otimes \cdots \otimes \varepsilon_{j_{k-1}} \right\rangle_k.
$$

This turns a summation of $k$ indices into a sum of $(k-1)$ indices, counting multiplicity. We shall repeat this process until we write the above into a sum of terms of the following form

$$
(-1)^{(k_1-1)+\cdots+(k_b-1)}\left\langle T^{(k)}, \left(n\bar{\varepsilon}\right)^{(k-l)} \otimes \sum_{1 \leq j_1 \neq \cdots \neq j_b \leq n} \varepsilon_{j_1}^{\otimes k_1} \otimes \cdots \otimes \varepsilon_{j_b}^{\otimes k_b} \right\rangle_k
$$

for some integer $b \in [0, k/2]$ and $2 \leq k_1 \leq \cdots \leq k_b$ satisfying $\sum_{a=1}^{b} k_a = \ell$. By induction, the sum of the multiplicities of such terms, say $C_{k_1,\ldots,k_b}^{(k)}$, is bounded by $k!$. This means

$$
\left(\frac{n^k}{C_{k,n}}\right)\langle T^{(k)}, \bar{\varepsilon}^{(k)} \rangle_k
$$

$$
= \sum_{b \geq 0, 2 \leq k_1 \leq \cdots \leq k_b \atop k_1 + \cdots + k_b = k} \frac{C_{k_1,\ldots,k_b}^{(k)}}{(-1)^{\ell-b}} \left\langle T^{(k)}, \left(n\bar{\varepsilon}\right)^{(k-l)} \otimes \sum_{1 \leq j_1 \neq \cdots \neq j_b \leq n} \varepsilon_{j_1}^{\otimes k_1} \otimes \cdots \otimes \varepsilon_{j_b}^{\otimes k_b} \right\rangle_k
$$

where $C_{k_1,\ldots,k_b}^{(k)}$ are positive integers satisfying

$$
T_k = \sum_{b \geq 0, 2 \leq k_1 \leq \cdots \leq k_b \atop k_1 + \cdots + k_b = k} C_{k_1,\ldots,k_b}^{(k)} \leq k!.
$$

It follows by simple algebra that

$$
\left\langle T^{(k)}, \bar{\varepsilon}^{(k)} \right\rangle_k
$$

$$
= \sum_{b \geq 0, 2 \leq k_1 \leq \cdots \leq k_b \atop k_1 + \cdots + k_b = k} \frac{C_{k_1,\ldots,k_b}^{(k)}}{C_{b,n}(n(n-1)\cdots(n-b+1)n^{\ell-b})} \left\langle T^{(k)}, \sum_{1 \leq j_1 \neq \cdots \neq j_b \leq n} \varepsilon_{j_1}^{\otimes k_1} \otimes \cdots \otimes \varepsilon_{j_b}^{\otimes k_b} \otimes \bar{\varepsilon}^{(k-l)} \right\rangle_k.
$$

Applying the above identity to each term in the remainder formula (2.7) and the spectrum norm, we find that

$$
\left| \langle \Delta^{(m)}(t), \bar{\varepsilon}^{\otimes (m-k)} \otimes \bar{\varepsilon}^{(k)} \rangle_m \right|
$$

$$
\leq \|\Delta^{(m)}(t)\| \sum_{b \geq 0, 2 \leq k_1 \leq \cdots \leq k_b \atop k_1 + \cdots + k_b = k} \frac{C_{k_1,\ldots,k_b}^{(k)}}{C_{b,n}(n(n-1)\cdots(n-b+1)n^{\ell-b})} \sum_{1 \leq j_1 \neq \cdots \neq j_b \leq n} \|\varepsilon_{j_1}^{\otimes k_1} \otimes \cdots \otimes \varepsilon_{j_b}^{\otimes k_b} \otimes \bar{\varepsilon}^{(m-l)}\|.
$$

This gives (2.10) as in (6).
Finally we provide an upper bound for $C_{k,n}$ by a variation of the Stirling’s formula. Let $c_n = n!e^{n}/(n + 1/2)^{n+1/2}$ and $x_n = 1/(n + 1)$. We have

$$
\frac{c_{n+1}}{c_n} = \frac{(n+1)e(n + 1/2)^{n+1/2}}{(n + 3/2)^{n+3/2}} = e(1 - x_n/2)^{1/x_n - 1/2} = e\left(\frac{(1 - x_n/2)^{1-x_n/2}}{(1 + x_n/2)^{1+x_n/2}}\right)^{1/x_n}.
$$

Due to $-\log(1-t/2) - \log(1+t/2) \geq 0$,

$$
\frac{1}{x} \left((1 - \frac{x}{2}) \log\left(\frac{1 - x}{2}\right) - (1 + \frac{x}{2}) \log\left(\frac{1 + x}{2}\right)\right) = \int_0^x \left(-\frac{1}{2} - \frac{t}{2} \log\left(\frac{1 - t}{2}\right) - \frac{1}{2} \log\left(\frac{1 + t}{2}\right)\right) dt \geq -1,
$$

so that $c_{n+1} \geq c_n$. It follows that

$$
C_{k,n} = \frac{n^{k-1}(n-k)!}{(n-1)!} \leq n^{k-1}e^{k-n}(n-k+1/2)^{n-k+1/2} e^{1-n/(n-1/2)^{n-1/2}} \leq e^{k-1}\left(1 + \frac{1/2}{n-1/2}\right)^{k-1}\left(1 - \frac{k-1}{n-1/2}\right)^{n-k+1/2} \leq \exp\left(k - 1 + \frac{(k-1)/2}{n-1/2} - \frac{(k-1)(n-k+1/2)}{n-1/2}\right) = \exp\left((k-1)(k-1+1/2)/(n-1/2)\right).
$$

This gives $C_{k,n} \leq e^{(k-1)(2k-1)/(2n-1)} \leq e^{(k-1)k/n}$.

**Proof of Theorem 3.1.** By definition $\| \Delta^{(m)}(t) \|_S \leq \| f \|_S(t) \| \tilde{\varepsilon} \|^s - m$. By (2.10),

$$
\| \text{Rem}_m \|_{L_p(P)} \leq C'_{m,n} \| f \|_S \max_{k_1 + \ldots + k_b = \ell, m, k_a \geq 2, k_a, k_a \geq 0} \left\| \frac{c_{k_1} \ldots c_{k_b}}{c_{k_1+\ldots+k_b}} \right\|_L^p \leq \sum_{m \geq 0} \frac{C_{k,n}}{(m-k)!} = C^*_m n.
$$

Let $w = (\| \varepsilon_1 \|/n, \ldots, \| \varepsilon_n \|/n)^T$. Due to $\| w \|_k \leq \| w \|_2$ for $k \geq 2$,

$$
\| \text{Rem}_m \|_{L_p(P)} \leq C'_{m,n} \| f \|_S \max_{k_1 + \ldots + k_b = \ell, m, k_a \geq 2, k_a, k_a \geq 0} \left\| \frac{c_{k_1} \ldots c_{k_b}}{c_{k_1+\ldots+k_b}} \right\|_L^p \leq C'_{m,n} \| f \|_S \max_{0 \leq \ell \leq m} \| \| \tilde{\varepsilon} \|^s - \ell \| \| w \|^s \|_L^p(P)\).
$$

This gives (3.2) with an application of Hölder’s inequality,

$$
\| \text{Rem}_m \|_{L_p(P)} \leq C'_{m,n} \| f \|_S \max \left\{ \left\| \tilde{\varepsilon} \right\|^s_{L_p}, \left\| w \right\|^s_{L_p(P)} \right\} = C^*_m n \| f \|_S \max \left\{ \left\| \tilde{\varepsilon} \right\|^s_{L_p}, \left(\sum_{j=1}^n \left\| \varepsilon_j \right\|^2 \right)^{s/2}_{L_p,\parallel} \right\}.
$$
For independent \( \varepsilon_j \), \( \| \mathbb{E}[\hat{f}] - f(\theta) \| \leq \| \text{Rem}_m \|_{L_p(\mathbb{P})} \) and Rosenthal’s inequality applies:

\[
\| \mathbb{E}[\hat{f}] - f(\theta) \|_{L_p(\mathbb{P})} \leq C'_{p,s} \left( \left\| \mathbb{E}[\| \mathbb{E}[\hat{f}] - f(\theta) \|^2_{L_2(\mathbb{P})} \right\|^p_{L_p(\mathbb{P})} + \| \mathbb{E}[\hat{f}] - f(\theta) \|^2_{L_p(\mathbb{P})} \right)
\]

with \( C'_{p,s} = (1 + s / \log(s/2))^{s} \) for \( s \geq 4 \), and

\[
\left\| \left( \sum_{j=1}^{n} \frac{\| \varepsilon_j \|^2}{n^2} \right) - \| \mathbb{E}[\mathbb{E}[\hat{f}] - f(\theta)] \|^2_{L_2(\mathbb{P})} \right\|_{L_p(\mathbb{P})}^{p/2} \leq C''_{p,s} \left( \left( \sum_{j=1}^{n} \frac{\| \varepsilon_j \|^4}{n^4} \right)^{p/4} + \sum_{j=1}^{n} \frac{\| \varepsilon_j \|^p}{n^{p/2}} \right)
\]

\[
= C''_{p,s} \left( \| \mathbb{E}[\mathbb{E}[\hat{f}] - f(\theta)] \|^2_{L_4(\mathbb{P})} + \| \mathbb{E}[\hat{f}] - f(\theta) \|^2_{L_p(\mathbb{P})} \right)
\]

with \( C''_{p,s} \leq C''_{p,s} \). Because \( \| \mathbb{E}[\mathbb{E}[\hat{f}] - f(\theta)] \|^2_{L_4(\mathbb{P})} \leq \max\{ \| \mathbb{E}[\hat{f}] - f(\theta) \|^2_{L_2(\mathbb{P})}, \| \mathbb{E}[\hat{f}] - f(\theta) \|^2_{L_p(\mathbb{P})} \} \), the second inequality in \( (3.3) \) follows by inserting the above inequalities into \( (3.2) \).

PROOF OF THEOREM 3.2. With the degenerate \( U \)-tensors \( S_k \) in \( (3.7) \), we write

\[
\hat{f} - f(\theta) = \sum_{k=1}^{m} \frac{S_k}{k!} - \text{Rem}_m.
\]

We first bound the second moment of \( S_k \) in the sum as the leading term. For \( k = 1 \),

\[
\mathbb{E}[S_k^2] = \mathbb{E}\left[ \left\langle f^{(1)}(\theta), \varepsilon_1 \right\rangle^2 \right] / n = V_1 / n \leq C_0 \sigma^2 / n.
\]

For \( k = 2, \ldots, m \), \( \mathbb{E}\left[ \left\langle f^{(k)}(\theta), \varepsilon_j \times \cdots \times \varepsilon_j \right\rangle_k \times f^{(k)}(\theta), \varepsilon_j \times \cdots \times \varepsilon_j \right\rangle_k = 0 \) when \( j_1 < \cdots < j_k, j'_1 < \cdots < j'_k \) and \( j_i \neq j'_i \) for some \( i \). Thus, in the i.i.d. case

\[
\mathbb{E}[S_k^2] = \mathbb{E}\left[ \left\langle f^{(k)}(\theta), \varepsilon_j \times \cdots \times \varepsilon_j \right\rangle_k \right] = C_{k,n} k! / n^k V_k,
\]

which gives \( (3.8) \). In the i.n.i.d. case, \( \Sigma = n^{-1} \sum_{j=1}^{n} \Sigma_j \) with \( \Sigma_j = \mathbb{E}[\varepsilon_j \times \varepsilon_j] \), so that

\[
\mathbb{E}[S_k^2] = \left( \frac{C_{k,n} k!}{n^k} \right)^2 \sum_{1 \leq j_1 < \cdots < j_k \leq n} \left\langle f^{(k)}(\theta), f^{(k)}(\theta) \times \Sigma_j \times \cdots \times \Sigma_j \right\rangle_k \leq \frac{C_{k,n} k!}{n^{2k}} \sum_{1 \leq j_1 < \cdots < j_k \leq n} \left\langle f^{(k)}(\theta), f^{(k)}(\theta) \times \Sigma_j \times \cdots \times \Sigma_j \right\rangle_k = (C_{k,n} k! / n^k) V_k,
\]

which gives \( (3.10) \). Moreover, because \( \mathbb{E}[S_k, S_{k'}] = 0 \) for \( 1 \leq k_1 < k_2 \leq m \),

\[
\mathbb{E}\left[ \left( \sum_{k=1}^{m} \frac{S_k}{k!} \right)^2 \right] \leq \sum_{k=1}^{m} \frac{C_{k,n} k!}{n^k k!} V_k.
\]

This and \( (3.3) \) yield \( (3.12) \).

Next we verify the upper bound \( (3.11) \) for \( V_k \). For \( k = 2 \),

\[
V_2 = \left\langle f^{(2)}(\theta), f^{(2)}(\theta) \times \Sigma \times \Sigma \right\rangle_2 = \text{trace} \left( f^{(2)}(\theta) \times (\Sigma^{1/2} \times \Sigma^{1/2}) \otimes_2 \right) \leq \left\| f^{(2)}(\theta) \times (\Sigma^{1/2} \times \Sigma^{1/2}) \right\|_S \text{trace} \left( f^{(2)}(\theta) \times \Sigma^{1/2} \times \Sigma^{1/2} \right).
\]

HIGH-ORDER STATISTICAL FUNCTIONAL EXPANSION

19
For $k > 2$, we write $\Sigma = \sum \ell \lambda_{\ell} v_{\ell} \otimes v_{\ell}$ as the eigenvalue decomposition and use the bound

$$V_k = \langle f^{(k)}(\theta), f^{(k)}(\theta) \times \Sigma \times k \Sigma \rangle_k = \sum_{\ell_3, \ldots, \ell_k} \langle f^{(k)}(\theta), f^{(k)}(\theta) \times \Sigma \times k \Sigma \times j = 3 (\lambda_{\ell_3} v_{\ell_3}^{(2)}) \rangle_k$$

$$= \sum_{\ell_3, \ldots, \ell_k} \lambda_{\ell_3} \cdots \lambda_{\ell_k} \langle \{f^{(k)}(\theta) \times j = 3 v_{\ell_3} \}, \{f^{(k)}(\theta) \times j = 3 v_{\ell_3} \} \times \Sigma \times 2 \Sigma \rangle_2$$

$$\leq \sum_{\ell_3, \ldots, \ell_k} \lambda_{\ell_3} \cdots \lambda_{\ell_k} \|f^{(k)}(\theta) \times j = 3 v_{\ell_3}\|_2^4 \sigma^4 r$$

$$\leq \sum_{\ell_3, \ldots, \ell_k} \lambda_{\ell_3} \cdots \lambda_{\ell_k} \|f^{(k)}(\theta)\|_2^4 \sigma^4 r$$

$$= \|f^{(k)}(\theta)\|_2^4 \sigma^{2k} r^{k-1}.$$ 

Thus, (3.11) holds.

Inserting (3.11) and bounds $\max_{2 \leq k \leq s} \|f^{(k)}(\theta)\|_2 \leq C_0$ and $\|f_s(\theta)\|_2 \leq C_0$ into (3.12), we obtain (3.13) because $\max \{\sigma^2 r^{1/2} / n, \ldots, \sigma^m r^{m-1/2} / n^{m/2}, \sigma^r (s / n)^{s/2}\}$ is attained at max \{\sigma^2 r^{1/2} / n, \sigma^r (s / n)^{s/2}\}. 

PROOF OF THEOREM 4.1. We shall keep in mind that all quantities, including $\{H, f, s, \sigma, r, C_0\}$, are allowed to depend on $n$, so that $O(1)$ here is uniformly bounded by numerical constants. When $s^2 \leq n$, we have $C_{k, n} \leq C_{m, n} \leq e^{(m-1)m/n} \leq e$ and $C^*_{m, n} \leq e^2$ in Theorem 3.1.

We first prove that the remainder term is of smaller order than the standard deviation $(V_1 / n)^{1/2}$ of the linear term, for the $V_1$ defined in (3.9). Let $w = (\|\bar{e}_1\| / n, \ldots, \|\bar{e}_n\| / n)^T$. By the definition of the noise level $\sigma$ and effective rank $r$ in (3.1), $E[\|w\|^2] = \sum_{j = 1}^n \|\bar{e}_j\|^2 = \sum_{j = 1}^n \|\bar{e}_j\|^2 = 2 r / n$. Thus, by Chebyshev’s inequality, in a certain event $\Omega_M$ with $P\{\Omega_M\} \geq 1 - 2 / M^2$

$$\|w\|^2 \leq M^2 \sigma^2 r / n, \quad \|\bar{e}\|^2 \leq M^2 \sigma^2 r / n.$$ 

By (6.3) and (2.10) the remainder term is bounded in $\Omega_M$ by

$$|\text{Rem}_m| \leq 2 \|f\|_2 \sum_{k_1 + \ldots + k_b = \ell} \frac{(C_{k_1, n} / C_{b, n}) C_{k_1, \ldots, k_b}}{\Gamma(s - k + 1)k!} \times \sum_{j_1 \neq \ldots \neq j_b} \|e_{j_1}\|^{k_1} \cdots \|e_{j_b}\|^{k_b} (M^2 \sigma^2 r / n)^{(s - \ell)/2}$$

$$\leq 2 C^*_{m, n} C_0 \sum_{k_1 + \ldots + k_b = \ell} \max_{k_1, \ldots, k_b} \frac{\sum_{j_1 \neq \ldots \neq j_b} \|e_{j_1}\|^{k_1} \cdots \|e_{j_b}\|^{k_b} (M^2 \sigma^2 r / n)^{(s - \ell)/2} n^{k_1} \cdots n^{k_b}}{n^\ell}$$

as in the beginning of the proof of Theorem 3.2. Moreover, for $k_1, \ldots, k_b$ satisfying $k_1 \geq 2$ and $k_1 + \ldots + k_b = \ell$, $\|w\|_{k_a} \leq \|w\|_2$ and (6.3) gives

$$\sum_{j_1 \neq \ldots \neq j_b} \|e_{j_1}\|^{k_1} \cdots \|e_{j_b}\|^{k_b} \leq \prod_{a = 1}^b \|w\|_{k_1}^{k_a} \leq \|w\|_2^\ell \leq (M^2 \sigma^2 r / n)^{\ell/2}$$

It follows that under (4.1), in the event $\Omega_M$ and for $s = O(1)$

$$|\text{Rem}_m| \leq 2 C^*_{m, n} C_{m, n} C_0 (M^2 \sigma^2 r / n)^{s/2} \ll \sqrt{V_1} / n.$$
Next, we consider the \( U \)-variables \( S_k \). By (3.7), \( S_k \) are uncorrelated as they are completely degenerate \( U \)-variables of order \( k \), so that (3.10), (3.11) and (4.1) yield

\[
\mathbb{E} \left[ \left( \sum_{k=2}^{m} \frac{S_k}{k!} \right)^2 \right] \leq \sum_{k=2}^{m} \mathbb{E} \left[ \frac{(C_{k,n}^2)C_0^2 \sigma^{2k_r k-1}}{n^k k!} \right] \leq C_{m,n}^* \max_{k \in [2, s]} \frac{\sigma^{2k_r k-1}}{n^k} \lesssim V_1/n
\]

as the maximum over \( k \in [2, s] \) is attained at the endpoints. In view of the above bounds for \( |\text{Rem}_m| \) and \( \sum_{k=2}^{m} S_k/k! \) and the definition of \( S_k \) in (3.7), we find that the linear term with \( k = 1 \) dominates in the expansion (2.6). Thus, (4.2) follows from the Lindeberg central limit theorem.

Without the condition \( s = O(1) \), the factor \( M^s \) in (6.4) is unbounded but the constant factor \( C_{m,n}^* \) in (6.4) can be improved by using (2.9) instead of (2.10) as follows. In the event \( \Omega_M, (6.3) \) and the Hilbert-Schmidt smoothness condition yield

\[
|\text{Rem}_m| \leq 2\|f\|_{\text{HS}, \Theta} \sum_{k=0}^{m} \frac{\|\epsilon\|^{s-k} \langle \epsilon(k), \bar{\epsilon}(k) \rangle_{k}^{1/2}}{\Gamma(s - k + 1)k!}
\]

\[
\leq 2C_0 d^{m/2} \sum_{k=0}^{m} \frac{(M^2 \sigma^2 r/n)^{(s-k)/2} \langle \epsilon(k), \bar{\epsilon}(k) \rangle_{k}^{1/2}}{\Gamma(s - k + 1)k!}.
\]

Because \( \bar{\epsilon}(k) \) are degenerate \( U \)-tensors in (2.3),

\[
\mathbb{E} [\langle \epsilon(k), \bar{\epsilon}(k) \rangle_{k}] = k! \sum_{1 \leq j_1 < \ldots < j_k \leq n} \prod_{a=1}^{k} \mathbb{E} [\|\epsilon_{j_a}\|^2] \leq k! C_{k,n}(\sigma^2 r/n)^k.
\]

It follows that

\[
\mathbb{E} \left[ d^{m/2} \sum_{k=0}^{m} \frac{(M^2 \sigma^2 r/n)^{(s-k)/2} \langle \epsilon(k), \bar{\epsilon}(k) \rangle_{k}^{1/2}}{\Gamma(s - k + 1)k!} \right]
\]

\[
\leq d^{m/2} \sum_{k=0}^{m} \frac{(M^2 \sigma^2 r/n)^{s/2}}{(m - k)!k!}
\]

\[
= d^{m/2} (M^2 \sigma^2 r/n)^{s/2} \sqrt{C_{m,n}} \nu^{2m} / \sqrt{m!}
\]

\[
= M^{s-m} (\sigma^2 r/n)^{s/2} \sqrt{C_{m,n}} \nu \sqrt{4m d^m M^{2m}/m!}
\]

\[
\leq (1 + M) (\sigma^2 r/n)^{s/2} \sqrt{C_{m,n}} \nu e^{2dM^2}.
\]

Thus, for \( d = O(1) \) and in an event \( \Omega^*_M \) with at least probability \( 1 - 3/M^2 \),

\[
|\text{Rem}_m| \leq 2C_0 M^2 (1 + M) (\sigma^2 r/n)^{s/2} e^{1 + 2dM^2} \ll (V_1/n)^{1/2}
\]

as \( C_{m,n} \leq e^{m^2/n} \leq e \). Moreover, instead of (6.5) we have

\[
\mathbb{E} \left[ \left( \sum_{k=2}^{m} \frac{S_k}{k!} \right)^2 \right] \leq \sum_{k=2}^{m} \mathbb{E} \left[ \frac{(C_{k,n}^2)C_0^2 \sigma^{2k_r k-1}}{n^k k!} \right] \leq e^{d+2} C_0^2 \max_{k \in [2, s]} \frac{\sigma^{2k_r k-1}}{n^k} \ll V_1/n.
\]

As both the remainder and \( \sum_{k=2}^{m} S_k/k! \) are of smaller order than \( (V_1/n)^{1/2} \) under the first condition of (4.1) and \( d = O(1), (4.2) \) still holds under the Lindeberg condition.

**Proof of Theorem 5.1.** We first prove (5.7) and (5.8). Let \( C_{k,n} = n^k (n - k)!/n! \leq e^{(k-1)/k}/n \) as in Lemma 2.2. It follows respectively from the independence of \( \epsilon_1, \ldots, \epsilon_n \) and the first line of (5.2) that \( \mathbb{E} [\epsilon_a(k) \epsilon_b(k)] = 0 \) for \( k \neq k' \) and

\[
\mathbb{E} \left[ \frac{\epsilon_a(k) \epsilon_b(k)}{n^k} \right] = \left( \frac{C_{k,n} k!}{n^k} \right)^2 \sum_{1 \leq j_1 < \ldots < j_k \leq n} \prod_{l=1}^{k} \mathbb{E} [\epsilon_{j_l, a} \epsilon_{j_l, b}]
\]
By symmetrization with the Rademacher variables \( \varepsilon \), it follows from the above two inequalities and the second line of (5.2) with \( k = 1 \) give (5.7):

\[
E \left[ \left( \sum_{a=1}^{d} \sum_{k=1}^{m} \frac{f^{(k)}_{h}(\theta_{a})\varepsilon_{a}^{(k)}}{d(k!)} \right)^{2} \right] \leq \sum_{a=1}^{d} \sum_{k=1}^{m} C_{k,n}^{2} \frac{f^{(k)}_{h}(\theta_{a})^{2} \sigma_{n}^{2k}}{d^{2k}k!}.
\]

Let \( \varepsilon^{(k)}_{j} \) be i.i.d. random variables independent of \( \{ \varepsilon_{1}, \ldots, \varepsilon_{n} \} \) with \( P\{ \varepsilon^{(k)}_{j} = 2 \} = 1/3 \) and \( P\{ \varepsilon^{(k)}_{j} = -1 \} = 2/3 \). Because \( E[\varepsilon^{(k)}_{j}] = 0 \) and \( E[(\varepsilon^{(k)}_{j})^{k}] \geq 1 \) for all integers \( k \geq 2 \), we have

\[
E \left[ \prod_{t=1}^{2s} \varepsilon_{j,t,a} \right] \leq E \left[ \prod_{t=1}^{2s} (\varepsilon^{(k)}_{j,t}\varepsilon_{j,t,a}) \right]
\]

for all \( 1 \leq j_{1}, \ldots, j_{2s} \leq n \) and integers \( s \geq 1 \) regardless of multiplicity in the indices. Thus,

\[
C_{k,n}^{-2} E \left[ (\varepsilon^{s-k}\varepsilon^{(k)}_{a})^{2} \right] = n^{-2s} E \left[ \left( \sum_{1 \leq j_{1}, \ldots, j_{2s}} \prod_{t=1}^{2s} \varepsilon_{j,t,a} \right)^{2} \right]
\]

\[
\leq n^{-2s} E \left[ \sum_{1 \leq j_{1}, \ldots, j_{2s}} \prod_{t=1}^{2s} (\varepsilon^{(k)}_{j,t}\varepsilon_{j,t,a}) \right]
\]

\[
= E \left[ \left( \sum_{j=1}^{n} \varepsilon^{(k)}_{j}\varepsilon_{j,a} / n \right)^{2s} \right].
\]

By symmetrization with the Rademacher variables \( \varepsilon^{(k)}_{j} \) and moment comparison, we have

\[
E \left[ \left( \sum_{j=1}^{n} \varepsilon^{(k)}_{j}\varepsilon_{j,a} / n \right)^{2s} \right] \leq 2^{2s} E \left[ \left( \sum_{j=1}^{n} \varepsilon^{(k)}_{j}\varepsilon^{(k)}_{j,a} / n \right)^{2s} \right] \leq 2^{4s} E \left[ \left( \sum_{j=1}^{n} \varepsilon^{(k)}_{j}\varepsilon_{j,a} / n \right)^{2s} \right].
\]

It follows from the above two inequalities and the second line of (5.2) that

\[
C_{k,n}^{-1} \left( E \left[ (\varepsilon^{s-k}\varepsilon^{(k)}_{a})^{2} \right] \right)^{1/2} \leq 2^{2s} \left( E \left[ \left( \sum_{j=1}^{n} \varepsilon^{(k)}_{j}\varepsilon_{j,a} / n \right)^{2s} \right] \right)^{1/2} \leq 2^{2s} \sqrt{\frac{2^{2s}2^{s-1}s!}{\sigma_{n}^{2}2^{s-1}s!}}.
\]

Thus, by (5.6),

\[
E \left[ \left( \frac{1}{d} \sum_{a=1}^{d} \text{Rem}_{m,a} \right)^{2} \right] \leq \frac{\|f_{h}\|_{(a)}}{d} \sum_{a=1}^{d} \sum_{k=0}^{s} \frac{C_{m,n}2^{2s} \sqrt{\sigma_{n}^{2}2^{s-1}s!}}{(s-k)!k!} \leq C_{m,n}2^{s} \frac{\|f_{h}\|_{(a)}}{d} \sigma_{n}^{2}2^{s-1}s!.
\]

This gives (5.8).

Inserting (5.7) and (5.8) into (5.5), we find that (5.9) holds:

\[
\sqrt{E\left[ |\hat{f} - f(\theta)|^{2} \right]} \leq \left\{ \frac{1}{d} \sum_{a=1}^{d} (f_{0} - f_{h})(\theta_{a})^{2} + \sum_{a=1}^{d} \sum_{k=1}^{m} \frac{|f^{(k)}_{h}(\theta_{a})|^{2} \sigma_{n}^{2k}}{(d^{2}/2)k!} \right\}^{1/2} + \frac{\|f_{h}\|_{(a)}2^{7/2}s!}{\sigma_{n}^{8}2^{s-1}s!}.
\]
for integers $s$ satisfying $s > 2$ and $(s - 1)s \leq (n/2) \log 2$.

Under the smoothness condition (5.11), (5.9) yields (5.12) via

\[
\kappa_{s,h,n,d}'(\theta) - \kappa_{s,0,n,d}'(\theta) \leq \left\{ \sum_{k=1}^{s-1} \frac{\eta_{2}^{2}(h)(\sigma_{n}/h)^{2k}}{(d/2)k!} \right\}^{1/2} \leq \eta_{\alpha}(h) \sqrt{d/2} e^{(\sigma/h)^{2}/2},
\]

\[
\|f_{h}\|_{(s)} \leq \eta_{\alpha}(h)/h^{s}, \ s! \geq e^{-s}s^{s+1/2}\sqrt{2\pi} \text{ and } |\text{bias}_{\alpha,h}| \leq \eta_{\alpha}(h).
\]

PROOF OF THEOREM 5.3. In the following proof, we use the same $C_{1}$ to bound different quantities to simplify notation although this does not give the sharpest constant factors in the risk bounds. Let $y = (x - t)/h$ so that $t = x - hy$. By (5.3) and the condition \( \int K(x) dx = 1 \),

\[
(6.6) \quad f_{h}(x) = \int h^{-1} K((x - t)/h) f_{0}(t) dt = \int K(y) f_{0}(x - hy) dy.
\]

By the conditions \( \int K(x) dx = 1 \), \( |f_{0}(x) - f_{0}(y)| \leq |x - y| \) and \( \int |xK(x)| dx \leq C_{1} \),

\[
|f_{h}(x) - f_{0}(x)| = \left| \int K(y) (f_{0}(x - hy) - f_{0}(x)) dy \right| \leq h \int |yK(y)| dy \leq C_{1} h.
\]

Moreover, (6.6) gives \( f_{h}^{(1)}(x) = \int K_{h}(y) f_{0}^{(1)}(x - y) dy = \int_{-\infty}^{\infty} K_{h}(y) dy - \int_{-\infty}^{\infty} K_{h}(y) dy \) and \( f_{h}^{(2)}(x) = 2K_{h}(x) \), so that \( \|f_{h}^{(1)}\|_{L_{1}} \leq \|K_{h}\|_{L_{1}} = \|K\|_{L_{1}} \leq C_{1} \) and \( \|f_{h}^{(k)}\|_{L_{\infty}} = 2\|K^{(k-1)}\|_{L_{\infty}} \leq C_{1} \) for \( k \geq 2 \). Thus, (5.14) implies (5.11) with \( \alpha = 1 \) and \( \eta_{1}(h) = C_{1} h_{1} \), and the conclusion follows from (5.13) of Theorem 5.1.

Next, we verify (5.14) under (5.15). By the Fourier transform formula,

\[
Q(\zeta) = (2\pi)^{-1/2} \int_{-\infty}^{\infty} e^{-i\zeta x} K(x) dx, \ K(x) = (2\pi)^{-1/2} \int_{-\infty}^{\infty} e^{i\zeta x} Q(\zeta) d\zeta.
\]

It follows that \( \int K(x) dx = \sqrt{2\pi} Q(0) = 1 \) by the first condition in (5.15),

\[
2\|K^{(k-2)}\|_{L_{\infty}} \leq \sqrt{2/\pi} \int_{-1}^{1} |\zeta|^{k-2} Q(\zeta) d\zeta \leq \sqrt{2/\pi} \|Q\|_{L_{1}} \leq C_{1}
\]

for \( k \geq 2 \) by the second and third conditions in (5.15), and for \( \ell \in \{0, 1\} \)

\[
\int |x^{\ell} K(x)| dx \leq \left( \int (1 + x^{2})^{-1} dx \int (1 + x^{2}) |x^{\ell} K(x)|^{2} dx \right)^{1/2}
\]

\[
= \pi^{1/2} \left( \|Q^{(\ell)}\|_{L_{2}}^{2} + \|Q^{(\ell+1)}\|_{L_{2}}^{2} \right)^{1/2}
\]

is bounded by \( C_{1} \) by the Plancherel identity and the fourth conditions in (5.15). \( \square \)

PROOF OF THEOREM 5.4. Because \( K(x) \) is the Fourier inversion of \( Q(\zeta) \),

\[
K(x) = (2\pi)^{-1/2} \int_{-\infty}^{\infty} e^{i\zeta x} Q(\zeta) d\zeta, \ Q(\zeta) = (2\pi)^{-1/2} \int_{-\infty}^{\infty} e^{-i\zeta x} K(x) dx.
\]

Thus, \( \int K(x) dx = \sqrt{2\pi} Q(0) = 1 \) by the first condition in (5.15), and for \( p \in [0, 1] \)

\[
\int |x|^{p} K(x) dx \leq \left( \int (1 + x^{2})^{-1} dx \int (1 + x^{2}) |x|^{2p} |K(x)|^{2} dx \right)^{1/2}
\]

\[
\leq (2\pi)^{1/2} \max_{k=0,2} \|Q^{(k)}\|_{L_{2}}
\]
is bounded by $C_1$ by the Plancherel identity and the fourth conditions in (5.15). It follows from the above properties of the kernel $K(x)$ and the Hölder smoothness of $f_0$ that

$$|f_h(x) - f_0(x)| = \left| \int K(y) (f_0(x - hy) - f_0(x)) \, dy \right| \leq h^p \int |y|^p |K(y)| \, dy \leq C_1 h^p.$$  

By the Fourier inversion formula,

$$K_h(x - y) = h^{-1} (2\pi)^{-1/2} \int e^{i(x-y)\zeta/h} Q(\zeta) \, d\zeta = (2\pi)^{-1/2} \int e^{iy\zeta} (e^{-i\zeta} Q(-h\zeta)) \, d\zeta$$

For $0 < p < 1$, the Fourier transformation of $f_0^{(1)}(x) = p \cdot \sgn(x)|x|^{p-1}$ and

$$(2\pi)^{-1/2} \int f_0^{(1)}(x) e^{-i\zeta} \, dx = C_p |\zeta|^{1+p}$$

with $C_p = (2/\pi)^{1/2} \Gamma(p+1) \sin(\pi p/2) / \sqrt{2\pi}$. Thus, by the Plancherel identity,

$$f_h^{(1)}(x) = \int K_h(x - y) f_0^{(1)}(y) \, dy = \int_{-1/h}^{1/h} \left( e^{-i\zeta} Q(-h\zeta) \right) \left( C_p |\zeta|^{1+p} \right) \, d\zeta.$$  

We note that $f^{(1)}$ is real-valued so that we only need to apply the complex conjugate to its Fourier transformation in the application of the Plancherel identity. Moreover, $C_p \|Q\|_{L_1} \leq (2/\pi)^{1/2} \|Q\|_{L_1} \leq C_1$ by the third condition in (5.15). Consequently,

$$\|f_h^{(k)}\|_{L_\infty} \leq \int_{-1/h}^{1/h} |Q(-h\zeta)| C_p |\zeta|^{k-1-p} \, d\zeta \leq h^{p-k} C_p \|Q\|_{L_1} \leq C_1 h^{p-k}.$$  

for all $k \geq 1$ and $0 < p < 1$. Hence, (5.15) implies (5.11) with $0 < p < 1$ and $\eta_p(h) = C_1 h^p$. The conclusion follows from (5.13) of Theorem 5.1.  

REFERENCES

[BF81] Peter J. Bickel and David A. Freedman, Some asymptotic theory for the bootstrap, The Annals of Statistics (1981), 1196–1217.

[BM95] Lucien Birgé and Pascal Massart, Estimation of integral functionals of a density, The Annals of Statistics 23 (1995), no. 1, 11–29.

[BR88] 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 50 (1988), no. 3, 381–393.

[BR03] ______, Nonparametric estimators which can be plugged-in, The Annals of Statistics 31 (2003), no. 4, 1033–1053.

[BSY19] Thomas B. Berrett, Richard J. Samworth, and Ming Yuan, Efficient multivariate entropy estimation via $k$-nearest neighbour distances, The Annals of Statistics 47 (2019), no. 1, 288–318.

[CC19] Olivier Collier and Laëtitia Comminges, Minimax optimal estimators for general additive functional estimation, arXiv preprint arXiv:1908.11070 (2019).

[CCT17] Olivier Collier, Laëtitia Comminges, and Alexandre B. Tsybakov, Minimax estimation of linear and quadratic functionals on sparsity classes, The Annals of Statistics 45 (2017), no. 3, 923–958.

[CCT20] ______, On estimation of nonsmooth functionals of sparse normal means, Bernoulli 26 (2020), no. 3, 1989 – 2020.

[CL05a] T. Tony Cai and Mark G. Low, Nonquadratic estimators of a quadratic functional, The Annals of Statistics 33 (2005), no. 6, 2930–2956.
On necessary conditions for efficient estimation of functionals of a density.

BIAS CORRECTION WITH JACKNIFE, BOOTSTRAP, AND TAYLOR SERIES

Sharp adaptive estimation of quadratic functionals.

Efficient estimation of smooth functionals in normal models.

Bias correction with jackknife, bootstrap, and Taylor series.
[Wu86] Chien-Fu Jeff Wu, *Jackknife, bootstrap and other resampling methods in regression analysis*, the Annals of Statistics **14** (1986), no. 4, 1261–1295.

[Zha90] Cun-Hui Zhang, *Fourier methods for estimating mixing densities and distributions*, The Annals of Statistics (1990), 806–831.

[ZL19] Fan Zhou and Ping Li, *A fourier analytical approach to estimation of smooth functions in gaussian shift model*, arXiv preprint arXiv:1911.02010 (2019).

[ZL21] ______, *Optimal estimation of high dimensional smooth additive function based on noisy observations*, Proceedings of the 38th International Conference on Machine Learning (ICML), 2021, pp. 12813–12823.

[ZZ14] Cun-Hui Zhang and Stephanie S. Zhang, *Confidence intervals for low dimensional parameters in high dimensional linear models*, Journal of the Royal Statistical Society: Series B (Statistical Methodology) **76** (2014), no. 1, 217–242.