Noname manuscript No.
(will be inserted by the editor)

Estimation of smooth functionals of location parameter in Gaussian and Poincaré random shift models

Estimation of smooth functionals

Vladimir Koltchinskii · Mayya Zhilova

Received: date / Accepted: date

Abstract Let E be a separable Banach space and let $f : E \mapsto \mathbb{R}$ be a smooth functional. We discuss a problem of estimation of $f(\theta)$ based on an observation $X = \theta + \xi$, where $\theta \in E$ is an unknown parameter and ξ is a mean zero random noise, or based on n i.i.d. observations from the same random shift model. We develop estimators of $f(\theta)$ with sharp mean squared error rates depending on the degree of smoothness of f for random shift models with distribution of the noise ξ satisfying Poincaré type inequalities (in particular, for some log-concave distributions). We show that for sufficiently smooth functionals f these estimators are asymptotically normal with a parametric convergence rate. This is done both in the case of known distribution of the noise and in the case when the distribution of the noise is Gaussian with covariance being an unknown nuisance parameter.

Keywords Smooth functionals · Efficiency · Random shift model · Poincaré inequality · Normal approximation

1 Introduction

One of the important lines of work of C.R. Rao is related to the development of concepts of optimality of statistical methods. As a student of R.A. Fisher

V. Koltchinskii
School of Mathematics
Georgia Institute of Technology
Atlanta, GA 30332-0160
E-mail: vlad@math.gatech.edu

M. Zhilova
School of Mathematics
Georgia Institute of Technology
Atlanta, GA 30332-0160
E-mail: mzhilova@math.gatech.edu

and a creator of information lower bounds, he has made seminal contributions to the theory of efficiency and asymptotic efficiency of statistical estimators.

The notion of asymptotic efficiency goes back to Fisher [8, 9], who conjectured that maximum likelihood estimators (MLE) of parameters of regular statistical models would have the smallest limit variance among all the asymptotically normal estimators ("Fisher's program"). A well-known counterexample by Hodges of superefficient estimators made apparent serious difficulties with implementing Fisher's program and led to the development by several leading statisticians of the 20th century (R. R. Bahadur, J. Hàjek, L. Le Cam, C.R. Rao, J. Wolfowitz) of contemporary view of asymptotic efficiency. In particular, in order to overcome the difficulties with Fisher's initial definition, C.R. Rao in [29] came up with an important idea of uniformity (with respect to the parameter) of convergence in distribution of properly normalized estimators. A widely used concept of asymptotic efficiency as "local asymptotic minimaxity" and the complete proof of asymptotic efficiency of MLE for regular statistical models are due to Hàjek and Le Cam ([10, 19, 20]).

Let $X_1, \dots, X_n$ be i.i.d. random variables sampled from the distribution with density $p_\theta, \theta \in \Theta$, $\Theta \subset \mathbb{R}^d$ being an open subset, and let $\hat\theta = \hat\theta_n$ be the MLE. If $f : \Theta \mapsto \mathbb{R}$ is a continuously differentiable function and the model $\{p_\theta : \theta \in \Theta\}$ is "regular", then the asymptotic normality of MLE easily implies that the sequence of r.v. $\sqrt{n}(f(\hat\theta_n) - f(\theta))$ converges in distribution to $N(0; \sigma_f^2(\theta))$, where $\sigma_f^2(\theta) := \langle I(\theta)^{-1} f'(\theta), f'(\theta)\rangle$ and $I(\theta)$ is the Fisher information matrix. Moreover, Hàjek–Le Cam's local asymptotic minimax lower bound

$$\lim_{c\to\infty} \limsup_{n\to\infty} \inf_{T_n} \sup_{\|\theta-\theta_0\|\le cn^{-1/2}} n\mathbb{E}_\theta (T_n - f(\theta))^2 \ge \sigma_f^2(\theta_0), \theta_0 \in \Theta$$

with the infimum taken over all estimators $T_n = T_n(X_1, \dots, X_n)$ implies the optimality of both the rate and the limit variance of the simple plug-in estimator $f(\hat\theta)$ of $f(\theta)$.

The same problem is considerably harder when the parameter of the model is high-dimensional or infinite-dimensional (nonparametric models). In this case, the naive plug-in estimator often fails partly due to its large bias and more sophisticated estimators have to be designed to achieve optimal error rates (in particular, to achieve $\sqrt{n}$-rate in the problems where the optimal rates of estimation of the whole parameter θ are slower than $\sqrt{n}$). Such problems have been studied by many authors since the seventies, most often, for specific models and for special functionals. A very incomplete list of references includes [2–4, 6, 7, 11–13, 16, 23–28, 30–32].

In this paper, we study the problem of estimation of smooth functionals of location parameters in random noise models in Banach spaces. This framework includes a number of high-dimensional models for vector, matrix or functional data. Following [13–15, 17, 18, 27, 28], we study how the mean squared error rates in these problems depend on the smoothness of the functionals and determine how much smoothness is needed for efficient estimation. More specifically, we extend the results of [17] from Gaussian shift models to more general random shift models with noise having a distribution satisfying a

Poincaré inequality (in particular, log-concave distribution). We do it both in the case of known noise distribution and also in the case when the distribution of the noise is Gaussian and the covariance is an unknown nuisance parameter.

2 Random shift models

Let E be a separable Banach space and let E^* be its dual space. In what follows, we often use the inner product notation $\langle x, u\rangle$ for the value of a linear functional $u \in E^*$ on a vector $x \in E$. We will use a generic norm notation $\|\cdot\|$ for the norms of E, E^* and other Banach spaces (providing it with subscripts only when the space it is not clear from the context).

Let $X = \theta + \xi$ be a random variable in E with $\theta \in E$ being *an unknown location parameter* and ξ being a random noise with $\mathbb{E}\|\xi\| < \infty, \mathbb{E}\xi = 0$. This simple model will be called *a random shift model.* In particular, if ξ is a mean zero Gaussian r.v. in E, it is called *a Gaussian shift model.*

Given a smooth functional $f : E \mapsto \mathbb{R}$, the goal is to estimate $f(\theta)$ based on an observation of X. The main difficulty in this problem is that a trivial plug-in estimator $f(X)$ could have a large bias (at least, when the norm of the noise is sufficiently large). We will try to overcome this difficulty by replacing f with another functional $g : E \mapsto \mathbb{R}$, such that the bias of estimator $g(X)$ is small. To this end, define an operator (acting on uniformly bounded Borel functions $g : E \mapsto \mathbb{R}$)

$$(\mathcal{T}g)(\theta) := \mathbb{E}_\theta g(X) = \mathbb{E}g(\theta + \xi), \theta \in E$$

and let $\mathcal{B} := \mathcal{T} - \mathcal{I}$. Note that the plug-in estimator $f(X)$ has bias

$$\mathbb{E}_\theta f(X) - f(\theta) = (\mathcal{B}f)(\theta), \theta \in E.$$

We can estimate this bias with a plug-in estimator $(\mathcal{B}f)(X)$ and define an estimator $f_1(X)$ with the first order bias correction as

$$f_1(X) := f(X) - (\mathcal{B}f)(X).$$

This estimator has bias

$$\mathbb{E}_\theta f_1(X) - f(\theta) = -(\mathcal{B}^2 f)(\theta), \theta \in E.$$

This yields the second order bias correction:

$$f_2(X) := f(X) - (\mathcal{B}f)(X) + (\mathcal{B}^2 f)(X).$$

Iterating these bias corrections k times yields the estimator

$$f_k(X) := \sum_{j=0}^{k} (-1)^j (\mathcal{B}^j f)(X),$$

whose bias is equal to

$$\mathbb{E}_\theta f_k(X) - f(\theta) = (-1)^k (\mathcal{B}^{k+1} f)(\theta), \theta \in E.$$

In other words, to find an estimator $g(X)$ of $f(\theta)$ with a small bias, one has to find an approximate solution of the equation $\mathcal{T}g = f$. If $\mathcal{T} = \mathcal{I} + \mathcal{B}$ is a small perturbation of the identity operator $\mathcal{I}$, one can formally write

$$g = (\mathcal{I} + \mathcal{B})^{-1} = \mathcal{I} - \mathcal{B} + \mathcal{B}^2 - \dots$$

Truncating this Neumann's series yields the function

$$f_k(\theta) := \sum_{j=0}^{k} (-1)^j (\mathcal{B}^j f)(\theta), \theta \in E$$

and the estimator $f_k(X)$.

Of course, it is assumed here that the distribution of the noise ξ is known and, hence, the functions $\mathcal{B}^j f, j = 1, \dots, k$ and f_k are also known. As a result, the estimator $f_k(X)$ is well defined. Later on, we will consider a different version of the problem in which the distribution of the noise is an unknown nuisance parameter and the function f_k has to be replaced by its estimator $\hat{f}_k$. It will be said in what follows that the functions f_k defined above are *associated with the model* $X = \theta + \xi$.

Before discussing the results of the paper, we will provide a couple of examples (already considered in [17]) that illustrate the bias reduction method in the case of two classes of functionals, polynomials and trigonometric polynomials (although, the analysis of bias reduction problem in Section 3 is based on different ideas and does not rely on approximation by polynomials or trigonometric polynomials).

Example 1 Consider the case of $E = \mathbb{R}$ and $\xi \sim \mathcal{N}(0, \sigma^2)$. Let $f(\theta) = \theta^2$. It is easy to see that $(\mathcal{T}f)(\theta) = \theta^2 + \sigma^2$, $(\mathcal{B}f)(\theta) = \sigma^2$ and $(\mathcal{B}^2 f)(\theta) = 0$ for all $\theta \in \mathbb{R}$. In this case, we have $f_1(\theta) = \theta^2 - \sigma^2$ and $f_1(X)$ is an unbiased estimator of $f(\theta)$. More generally, if $f(\theta) = \theta^m$ for some $m \geq 1$, it is well known that the unbiased estimator of $f(\theta)$ is $\sigma^m H_m(\frac{X}{\sigma})$, where $H_j, j \geq 0$ are Hermite polynomials (see, e.g., [13]). Moreover, this estimator coincides with $f_k(X)$ for all $k \geq [\frac{m}{2}]$. If now $E = \mathbb{R}^d$ for $d \geq 1$, $\xi \sim \mathcal{N}(0, \sigma^2 I_d)$ and

$$f(\theta) := \sum_{k_1, \dots, k_d} c_{k_1, \dots, k_d} \theta_1^{k_1} \dots \theta_d^{k_d}, \ \theta = (\theta_1, \dots, \theta_d) \in \mathbb{R}^d$$

is a polynomial of degree m, then, for all $k \geq [\frac{m}{2}]$,

$$f_k(\theta) = \sum_{k_1, \dots, k_d} c_{k_1, \dots, k_d} \sigma^{k_1 + \dots k_d} H_{k_1}\Big(\frac{\theta_1}{\sigma}\Big) \dots H_{k_d}\Big(\frac{\theta_d}{\sigma}\Big)$$

and $f_k(X)$ is an unbiased estimator of $f(\theta)$ (see [17], Example 2.1 and Corollary 3.1 for more details).

Example 2 As another example, consider an arbitrary Banach space E (for instance, $E = \mathbb{R}^d$) and $\xi \sim \mathcal{N}(0, \Sigma)$. By a simple computation, for all $u \in E^*$

$$\mathcal{T} \cos(\langle \cdot, u\rangle) = \exp\{-\langle \Sigma u, u\rangle\} \cos(\langle \cdot, u\rangle),\ \mathcal{T} \sin(\langle \cdot, u\rangle) = \exp\{-\langle \Sigma u, u\rangle\} \sin(\langle \cdot, u\rangle).$$

If now

$$f(\theta) = \sum_i [c_i \cos(\langle \theta, u_i\rangle) + d_i \sin(\langle \theta, u_i\rangle)], \theta \in E$$

is a trigonometric polynomial, then

$$f_k(\theta) = \sum_i \lambda_k(\Sigma, u_i)[c_i \cos(\langle \theta, u_i\rangle) + d_i \sin(\langle \theta, u_i\rangle)], \theta \in E,$$

where

$$\lambda_k(\Sigma, u) := \exp\{\langle \Sigma u, u\rangle/2\}[1 - (1 - \exp\{-\langle \Sigma u, u\rangle/2\})^{k+1}],$$

and

$$\begin{aligned}\mathbb{E}_\theta f_k(X) - f(\theta) &= (-1)^k(\mathcal{B}^{k+1} f)(\theta)\\ &= -\sum_i (1 - \exp\{-\langle \Sigma u_i, u_i\rangle/2\})^{k+1}[c_i \cos(\langle \theta, u_i\rangle) + d_i \sin(\langle \theta, u_i\rangle)]\end{aligned}$$

(see [17], Example 2.2). If, for all i, $\langle \Sigma u_i, u_i\rangle$ is bounded away from 0, then the coefficients $(1 - \exp\{-\langle \Sigma u_i, u_i\rangle/2\})^{k+1}$ become small for a sufficiently large k ensuring the reduction of the bias.

For Banach spaces E, F and a function $g : E \mapsto F$, let

$$\|g\|_{L_\infty} := \sup_{x\in E} \|g(x)\|,$$

$$\|g\|_{\rm Lip} := \sup_{x,x'\in E, x\neq x'} \frac{\|g(x) - g(x')\|}{\|x - x'\|}$$

and, for $\rho \in (0, 1]$,

$$\|g\|_{{\rm Lip}_\rho} := \sup_{x,x'\in E, x\neq x'} \frac{\|g(x) - g(x')\|}{\|x - x'\|^\rho}.$$

We will now introduce Hölder spaces $C^s(E)$ of functionals $g : E \mapsto \mathbb{R}$ of smoothness $s > 0$. Let $g^{(j)}$ denote the Fréchet derivative of g of order j (in particular, $g^{(0)} = g$). [1] Note that, for all $x \in E$, $g^{(j)}(x)$ is a symmetric bounded j-linear form. For such a form $M[u_1, \dots, u_j], u_1, \dots, u_j \in E$, define its operator norm as

$$\|M\| := \sup_{\|u_1\|\leq 1,\dots,\|u_j\|\leq 1} |M[u_1, \dots, u_j]|.$$

[1] The definition and relevant properties of j-th order Fréchet derivatives could be found, e.g., in [5], sections 2.1, 5.1, 5.3.

In particular, for $j = 1$, $M[x] = \langle x, u\rangle$ for some $u \in E^*$ and $\|M\| = \|u\|$. The derivative $g^{(j)}$ will be always viewed as a mapping from E into the space of symmetric bounded j-linear forms equipped with the operator norm. Let $s = k + \rho$, $k \geq 0, \rho \in (0, 1]$. For a k-times Fréchet differentiable functional g, define

$$\|g\|_{C^s} := \max\Big(\|g\|_{L_\infty}, \max_{1\leq j\leq k} \|g^{(j)}\|_{L_\infty}, \|g^{(k)}\|_{\mathrm{Lip}_\rho}\Big).$$

The space $C^s(E)$ is defined as the set of all k-times Fréchet continuously differentiable functionals g such that $\|g\|_{C^s} < \infty$.

Recall that the covariance operator $\Sigma_\xi : E^* \mapsto E$ of a mean zero r.v ξ in E (with a finite weak second moment) is defined as

$$\Sigma_\xi u := \mathbb{E}\langle \xi, u\rangle \xi, u \in E^*.$$

The operator norm of Σ_ξ is

$$\|\Sigma_\xi\| := \sup_{\|u\|\leq 1} \|\Sigma_\xi u\|.$$

It is easy to see that $\mathbb{E}\|\xi\|^2 \geq \|\Sigma_\xi\|$.

We use the following notations. We write $\xi \sim \mu$ if a random variable ξ is sampled from a distribution μ. The notation $\eta_1 \overset{d}{=} \eta_2$ means that random variables η_1, η_2 have the same distribution. For $A, B \geq 0$, $A \lesssim B$ means that there exists a numerical constant $C > 0$ such that $A \leq CB$, $A \gtrsim B$ means that $B \lesssim A$ and $A \asymp B$ means that $A \lesssim B$ and $A \gtrsim B$. If a constant C in the above relationships is allowed to depend on some parameters, say, α, β, we will write $A \lesssim_{\alpha,\beta} B$, $A \gtrsim_{\alpha,\beta} B$ and $A \asymp_{\alpha,\beta} B$.

2.1 Gaussian shift models

We start with an overview of the results on estimation of smooth functionals in Gaussian shift models obtained in [17]. It is assumed that the distribution of the noise ξ is known. The following theorem provides an upper bound on the mean squared error of estimator $f_k(X)$ (see Theorem 2.1 in [17]).[2]

Theorem 1 *Suppose ξ is a Gaussian r.v. in E with mean zero and covariance operator Σ_ξ. Then, for all $s = k + 1 + \rho$, $k \geq 0, \rho \in (0, 1]$,*

$$\sup_{\|f\|_{C^s}\leq 1} \sup_{\theta\in E} \mathbb{E}_\theta(f_k(X) - f(\theta))^2 \lesssim_s \Big(\|\Sigma_\xi\| \vee (\mathbb{E}\|\xi\|^2)^s\Big) \wedge 1.$$

[2] In fact, a more general result was proved in [17] for spaces $C^{s,\gamma}(E)$ of functionals of smoothness s whose derivatives are allowed to grow as $\|\theta\|^\gamma$. For simplicity, we consider here only the case of $\gamma = 0$ (both f and its derivatives are uniformly bounded).

It was also shown in [17] (Theorem 2.2) that the above bound is sharp in a minimax sense in the case of the classical Gaussian shift model: $E = \mathbb{R}^d$ with the standard Euclidean norm and $\xi \sim N(0, \sigma^2 I_d)$. In this case, $\|\Sigma_\xi\| = \sigma^2$, $\mathbb{E}\|\xi\|^2 = \sigma^2 d$, and we have

$$\sup_{\|f\|_{C^s}\le 1} \inf_T \sup_{\theta\in E} \mathbb{E}_\theta (T(X) - f(\theta))^2 \asymp_s \Big(\|\Sigma_\xi\| \vee (\mathbb{E}\|\xi\|^2)^s\Big) \wedge 1 \tag{1}$$

with the minimax optimal rate attained (up to a constant) for the estimator $f_k(X)$.

Note that the bound of Theorem 1 could be also applied to a sample of i.i.d. copies $X_1, \dots, X_n$ of a Gaussian shift observation $X = \theta + \xi$ since $\bar X = \frac{X_1+\dots+X_n}{n}$ also satisfies a Gaussian shift model $\bar X = \theta + \bar\xi$, $\bar\xi \sim N(0, n^{-1}\Sigma_\xi)$ and one can define functions f_k associated with this model. In this case, we get

$$\sup_{\|f\|_{C^s}\le 1} \sup_{\theta\in E} \mathbb{E}_\theta (f_k(\bar X) - f(\theta))^2 \lesssim_s \Big(\frac{\|\Sigma_\xi\|}{n} \vee \Big(\frac{\mathbb{E}\|\xi\|^2}{n}\Big)^s\Big) \wedge 1. \tag{2}$$

Suppose $\|\Sigma_\xi\| \lesssim 1$ and $\mathbb{E}\|\xi\|^2 \lesssim d$, where d is the dimension or other relevant complexity parameter. In this case, if $d \lesssim n^\alpha$ for some $\alpha \in (0,1)$ and $s \ge \frac{1}{1-\alpha}$, then

$$\sup_{\|f\|_{C^s}\le 1} \sup_{\theta\in E} \mathbb{E}_\theta (f_k(\bar X) - f(\theta))^2 = O(n^{-1}) \text{ as } n \to \infty.$$

It follows from the minimax bound (1) that the threshold $\frac{1}{1-\alpha}$ on the smoothness of f needed for estimation of $f(\theta)$ with mean squared error rate $O(n^{-1})$ is sharp (at least, in the case of the classical Gaussian shift model): if $d \gtrsim n^\alpha$ for some $\alpha \in (0,1)$ and $s < \frac{1}{1-\alpha}$, then the minimax optimal mean squared error rate is of the order $n^{-(1-\alpha)s}$, which is slower than n^{-1}.

Moreover, it was also proved in [17] that, for $s > \frac{1}{1-\alpha}$, the estimator $f_k(\bar X)$ is asymptotically normal with $\sqrt{n}$-rate and limit variance $\sigma_f^2(\theta) := \langle \Sigma_\xi f'(\theta), f'(\theta)\rangle$. Namely, the following fact follows from the bound of Theorem 2.3 in [17]. For r.v. η_1, η_2 in $\mathbb{R}$, denote

$$d_K(\eta_1, \eta_2) := \sup_{x\in\mathbb{R}} |\mathbb{P}\{\eta_1 \le x\} - \mathbb{P}\{\eta_2 \le x\}|.$$

Then, under the assumptions $\|\Sigma_\xi\| \lesssim 1$, $\mathbb{E}\|\xi\|^2 \lesssim d \lesssim n^\alpha$ for some $\alpha \in (0,1)$ and $s > \frac{1}{1-\alpha}$,[3]

$$\sup_{\|f\|_{C^s}\le 1} \sup_{\theta\in E} \Big| n\mathbb{E}_\theta (f_k(\bar X) - f(\theta))^2 - \sigma_f^2(\theta)\Big| \to 0 \tag{3}$$

[3] Note that these asymptotic relationships hold in the case when $E = E_n$ depends on n, for instance, when $E = \mathbb{R}^d$ with $d = d_n$.

and, for all $\sigma_0 > 0$,

$$\sup_{\|f\|_{C^s}\le 1} \sup_{\theta\in E, \sigma_f(\theta)\ge \sigma_0} d_K\Big(\frac{\sqrt{n}(f_k(\bar X) - f(\theta))}{\sigma_f(\theta)}, Z\Big) \to 0 \text{ as } n\to\infty. \qquad (4)$$

Finally, note that the following Hàjek–Le Cam type local asymptotic minimax lower bound

$$\lim_{c\to\infty} \limsup_{n\to\infty} \inf_{T_n} \sup_{\|\theta-\theta_0\|\le cn^{-1/2}} \frac{n\mathbb{E}_\theta(T_n - f(\theta))^2}{\sigma_f^2(\theta)} \ge 1 \qquad (5)$$

holds for all $s > 1$, all $\sigma_0 > 0$ and all functionals f with $\|f\|_{C^s} \le 1$ under the assumptions $\|\Sigma_\xi\| \lesssim 1$, $\sigma_f(\theta_0) \ge \sigma_0$. This follows from a non-asymptotic local minimax lower bound of Theorem 2.4 in [17] and it implies asymptotic efficiency of $f_k(\bar X)$ as an estimator of $f(\theta)$.

We now turn to the problem of estimation of $f(\theta)$ in the case when the noise ξ is a mean zero Gaussian r.v. with *unknown* covariance operator Σ_ξ, which could be viewed as a *nuisance parameter*. Note that in the case of known Σ_ξ the upper bound of Theorem 1 also holds for estimator $f_k(\bar X)$, where $\bar X$ is the sample mean based on n i.i.d. observations of $X = \theta + \xi$, with a constant that depends on n. Thus, in the case of a small fixed n (say, $n = 2$), we have

$$\sup_{\|f\|_{C^s}\le 1} \sup_{\theta\in E} \mathbb{E}_\theta(f_k(\bar X) - f(\theta))^2 \lesssim_{s,n} \Big(\|\Sigma_\xi\| \vee \Big(\mathbb{E}\|\xi\|^2\Big)^s\Big) \wedge 1.$$

This follows from bound (2) (note also that the function f_k is now associated with the Gaussian shift model $\bar X = \theta + \bar\xi$, $\bar\xi \sim N(0; n^{-1}\Sigma_\xi)$).

First, we show how to construct an estimator of $f(\theta)$ based on two independent observations X_1, X_2 of X for which the same upper bound hods in the case of unknown covariance of the noise. To this end, define $\bar X := \frac{X_1+X_2}{2}$ and let ξ be a version of the noise defined as follows: $\xi := \frac{X_1-X_2}{\sqrt 2}$. Note that $\bar X$ and ξ are independent r.v., $\xi \sim N(0; \Sigma_\xi)$ and $\bar X \overset{d}{=} \theta + \frac{\xi}{\sqrt 2}$. Define

$$\hat V_j(\theta;\xi) := \sum_{l=0}^{j} (-1)^{j-l}\binom{j}{l} f\Big(\theta + \sqrt{\frac{l}{2}}\xi\Big)$$

and

$$\hat f_k(\theta) := \hat f_k(\theta;\xi) := \sum_{j=0}^{k} (-1)^j \hat V_j(\theta;\xi).$$

It turns out that $\hat f_k(\theta;\xi)$ with $\xi = \frac{X_1-X_2}{\sqrt 2}$ is an unbiased estimator of $f_k(\theta)$ (where the function f_k is associated with the Gaussian shift model $\bar X = \theta + \bar\xi$, $\bar\xi \sim N(0; \Sigma_\xi/2)$). Due to this fact, it becomes natural to use $\hat f_k(\bar X)$ as an estimator of $f(\theta)$. The following proposition will be proved:

Proposition 1 *Suppose ξ is a Gaussian r.v. in E with mean zero and unknown covariance operator Σ_ξ. Then, for all $s = k+1+\rho$, $k \geq 0, \rho \in (0,1]$,*

$$\sup_{\|f\|_{C^s}\leq 1} \sup_{\theta\in E} \mathbb{E}_\theta(\hat{f}_k(\bar{X}) - f(\theta))^2 \lesssim_s \Big(\|\Sigma_\xi\| \vee (\mathbb{E}\|\xi\|^2)^s\Big) \wedge 1.$$

A similar idea can be used in the case of n i.i.d. observations $X_1, \dots, X_n$ of r.v. $X = \theta + \xi$ when the sample size n is large. Assume that $n \geq 2m$ for some $m \geq 1$. Let us write $X_1, \dots, X_{2m}$ as $X_1, X_1', \dots, X_m, X_m'$. Define $\tilde{\xi}_i := \frac{X_i - X_i'}{\sqrt{2}}, i = 1, \dots, m$, which are i.i.d. random variables with the same distribution as ξ. Note also that $\{\tilde{\xi}_i : i = 1, \dots, m\}$ and

$$\bar{X} = \frac{X_1 + \dots X_n}{n} = \frac{X_1 + X_1' + \dots X_m + X_m' + X_{2m+1} + \dots + X_n}{n}$$

are independent r.v.. Define

$$\tilde{V}_j(\theta; \tilde{\xi}_i) := \sum_{l=0}^{j} (-1)^{j-l} \binom{j}{l} f\Big(\theta + \sqrt{\frac{l}{n}}\tilde{\xi}_i\Big),$$

$$\tilde{f}_k(\theta; \tilde{\xi}_i) := \sum_{j=0}^{k} (-1)^j \tilde{V}_j(\theta; \tilde{\xi}_i), i = 1, \dots, m$$

and

$$\hat{f}_{m,k}(\theta) := m^{-1} \sum_{i=1}^{m} \tilde{f}_k(\theta; \tilde{\xi}_i).$$

We will see that $\hat{f}_{m,k}(\theta)$ is an unbiased estimator of $f_k(\theta)$ with f_k being associated with the Gaussian shift model $\bar{X} = \theta + \bar{\xi}$, $\bar{\xi} \sim N(0; n^{-1}\Sigma_\xi)$. To estimate $f(\theta)$, we now use the estimator $\hat{f}_{m,k}(\bar{X})$.

Proposition 2 *Suppose ξ is a Gaussian r.v. in E with mean zero and unknown covariance operator Σ_ξ and let $X_1, \dots, X_n$ be i.i.d. copies of $X = \theta + \xi$. Then, for all $s = k+1+\rho$, $k \geq 0, \rho \in (0,1]$,*

$$\sup_{\|f\|_{C^s}\leq 1} \sup_{\theta\in E} \mathbb{E}_\theta(\hat{f}_{m,k}(\bar{X}) - f(\theta))^2 \lesssim_s \Big(\frac{\|\Sigma_\xi\|}{n} \vee \Big(\frac{\mathbb{E}\|\xi\|^2}{n}\Big)^s\Big) \wedge 1. \tag{6}$$

Moreover, suppose that, for some d and for some $\alpha \in (0,1)$, $\|\Sigma_\xi\| \lesssim 1$, $\mathbb{E}\|\xi\|^2 \lesssim d \lesssim n^\alpha$ and $s > \frac{1}{1-\alpha}$. If $m = m_n \to \infty$ as $n \to \infty$, then

$$\sup_{\|f\|_{C^s}\leq 1} \sup_{\theta\in E} \Big| n\mathbb{E}_\theta(\hat{f}_{m,k}(\bar{X}) - f(\theta))^2 - \sigma_f^2(\theta)\Big| \to 0 \tag{7}$$

and, for all $\sigma_0 > 0$,

$$\sup_{\|f\|_{C^s}\leq 1} \sup_{\theta\in E, \sigma_f(\theta)\geq\sigma_0} d_K\Big(\frac{\sqrt{n}(\hat{f}_{m,k}(\bar{X}) - f(\theta))}{\sigma_f(\theta)}, Z\Big) \to 0 \text{ as } n \to \infty. \tag{8}$$

Remark 1 Note that limit relationships (7) and (8) along with local asymptotic minimax bound (5) show that estimator $\hat{f}_{m,k}(\bar{X})$ is asymptotically efficient.

Remark 2 In [34], other estimators of $f(\theta)$ were proposed in the case when the covariance is an unknown nuisance parameter of a Gaussian shift model in $\mathbb{R}^d$ equipped with the standard Euclidean norm. However, it was done under much stronger smoothness assumptions (Besov smoothness instead of Hölder smoothness). Since the Besov norms for functions on high-dimensional spaces could differ from the Hölder norms (in the sense of our definition) by a dimension dependent constant, the error rates obtained in [34] are sub-optimal (see also Remark 2.1 in [17]).

2.2 Poincaré random shift models

In this section, we extend the bound of Theorem 1 and some other results of [17] to more general random shift models $X = \theta + \xi, \theta \in E$, namely, to the models with the noise ξ satisfying Poincaré inequality. Assume that $\mathbb{E}\|\xi\|^2 < \infty$ and that the distribution of ξ is known.

It will be said that a random variable ξ in E satisfies the Poincaré inequality iff there exists a constant $C > 0$ such that for all Fréchet continuously differentiable functions $g : E \mapsto \mathbb{R}$ [4]

$$\mathrm{Var}(g(\xi)) \leq C\mathbb{E}\|g'(\xi)\|^2. \tag{9}$$

Denote by $C_P(\xi)$ the smallest possible value of C in the above inequality. It will be called the Poincaré constant of ξ (or, more precisely, of its distribution $\mathcal{L}(\xi)$). The following properties of $C_P(\xi)$ are obvious:

1. $C_P(\xi + a) = C_P(\xi), a \in E$;
2. $C_P(b\xi) = b^2 C_P(\xi), b \in \mathbb{R}$;
3. for any linear operator $A : E \mapsto E$, $C_P(A\xi) \leq \|A\|^2 C_P(\xi)$.

Note that if $g_u(x) = \langle x, u\rangle, x \in E, u \in E^*$ is a linear functional, then $\mathrm{Var}(g_u(\xi)) = \langle \Sigma_\xi u, u\rangle$ and $(Lg_u)(x) = \|u\|, x \in E$. This easily implies that

$$C_P(\xi) \geq \|\Sigma_\xi\|.$$

It is also well known that Gaussian random variables in E satisfy Poincaré inequality with $C_P(\xi) = \|\Sigma_\xi\|$.

Let $\xi = (\xi_1, \dots, \xi_k) \in E_1 \times \cdots \times E_k$ be a random vector, where $E_1, \dots, E_k$ are separable Banach spaces and $E := E_1 \times \cdots \times E_k$ is equipped with the norm

$$\|x\| := \Big(\sum_{j=1}^k \|x_j\|^2\Big)^{1/2}, x = (x_1, \dots, x_k) \in E.$$

[4] A standard way to write Poincaré inequality for r.v. in $\mathbb{R}^d$ is:

$$\mathrm{Var}(g(\xi)) \leq C\mathbb{E}\|\nabla g(\xi)\|^2,$$

where ∇g is the gradient of function g that, by Rademacher's theorem, exists a.s. for all locally Lipschitz functions and $\|\cdot\|$ is the standard Euclidean norm.

If $\xi = (\xi_1, \ldots, \xi_k)$ has independent components, then

$$C_P(\xi) = \max_{1\leq j\leq k} C_P(\xi_j)$$

(see, e.g., [21], Corollary 5.7).

Another well known class of random variables with a finite Poincaré constant are r.v. sampled from *log-concave* distributions in $\mathbb{R}^d$. It will be said that ξ in $\mathbb{R}^d$ is log-concave iff $\xi \sim \mu$, where

$$\mu(dx) = \exp\{-V(x)\}dx$$

with a convex function $V : \mathbb{R}^d \mapsto \bar{\mathbb{R}}$. In addition to normal distributions, the class of log-concave distributions also includes, for instance, uniform distributions on convex bodies of finite volume. This class of probability distributions has been intensively studied in high-dimensional probability and convex geometry and it is becoming an important modeling tool in high-dimensional statistics. It is well known that log-concave random variables have finite Poincaré constants. Moreover, according to one of the forms of well known Kannan-Lovász-Simonovits (KLS) conjecture in convex geometry, for any log-concave r.v. ξ with covariance Σ_ξ, $C_P(\xi) \lesssim \|\Sigma_\xi\|$ with a dimension free constant (this conjecture remains open, see, e.g. [22]).

Some other well known properties of Poincaré constants of r.v. in $\mathbb{R}^d$ (including partial results related to KLS conjecture) are stated below (see, e.g., [1]):

- If μ, ν are two probability measures, $\xi \sim \mu, \eta \sim \nu$, and μ is absolutely continuous with respect to ν with the density $\frac{d\mu}{d\nu}$ bounded from above by a constant $A > 0$ and bounded from below by a constant $a > 0$, then $C_P(\xi) \leq \frac{A}{a} C_P(\eta)$.
- If $\xi \sim \mu$, $\mu(dx) = e^{-V(x)}dx$ with $V : \mathbb{R}^d \mapsto \mathbb{R}$ such that $V^{''}(x) \succeq C^{-1}$ for a symmetric positively definite matrix C, then $C_P(\xi) \leq \|C\|$.
- If C is a symmetric positively definite matrix and $\xi \sim \mu$,

$$\mu(dx) = \exp\Big\{-\frac{1}{2}\langle C^{-1}x, x\rangle - V(x)\Big\}dx,$$

  where V is a convex function on $\mathbb{R}^d$, then $C_P(\xi) \leq \|C\|$.
- If $\xi \sim \mu$, $\mu(dx) = \frac{1}{Z}\exp\{-\|x\|_{\ell_p}^p - V(x)\}dx$, where $V : \mathbb{R}^d \mapsto \mathbb{R}$ is convex and $V(-x) = V(x), x \in \mathbb{R}^d$, $p \in [1, 2]$, then $C_P(\xi) \lesssim (\log d)^{\frac{2-p}{p}}$
- If μ, ν are log-concave measures on $\mathbb{R}^d$, $\xi \sim \mu, \eta \sim \nu$, and, for some $\varepsilon \in (0, 1)$,

$$d_{TV}(\mu, \nu) := \sup_{A\subset\mathbb{R}^d, A \text{ Borel}} |\mu(A) - \nu(A)| \leq 1 - \varepsilon,$$

  then $C_P(\xi) \lesssim_\varepsilon C_P(\eta)$.

– If ξ is log-concave with covariance Σ_ξ, then

$$C_P(\xi) \lesssim \|\Sigma_\xi\|_{HS} \lesssim \|\Sigma_\xi\|\sqrt{d}$$

(see [22]).

We will prove the following theorem.

Theorem 2 *Suppose ξ is a r.v. in E with mean zero, $\mathbb{E}\|\xi\|^2 < \infty$, covariance operator Σ_ξ and $C_P(\xi) < \infty$. Then, for all $s = k + 1 + \rho$, $k \geq 0, \rho \in (0,1]$*

$$\sup_{\|f\|_{C^s} \leq 1} \sup_{\theta \in E} \mathbb{E}_\theta (f_k(X) - f(\theta))^2 \lesssim_s \Big(\|\Sigma_\xi\| \vee C_P(\xi)(\mathbb{E}\|\xi\|^2)^\rho \vee (\mathbb{E}\|\xi\|^2)^s\Big) \wedge 1. \tag{10}$$

Moreover, if $\mathbb{E}^{1/2}\|\xi\|^2 \leq 1/2$, then

$$\sup_{\|f\|_{C^s} \leq 1} \sup_{\theta \in E} \|f_k(X) - f(\theta) - \langle \xi, f'(\theta)\rangle\|_{L_2(\mathbb{P}_\theta)} \leq 2\Big(3C_P^{1/2}(\xi)(\mathbb{E}^{1/2}\|\xi\|^2)^\rho + (\mathbb{E}^{1/2}\|\xi\|^2)^s\Big). \tag{11}$$

For r.v. η_1, η_2 in $\mathbb{R}$ and $p \geq 1$, define the Wasserstein p-distance between η_1, η_2 as

$$W_p(\eta_1, \eta_2) := \inf\{\mathbb{E}^{1/p}|\eta_1' - \eta_2'|^p : \eta_1' \overset{d}{=} \eta_1, \eta_2' \overset{d}{=} \eta_2\},$$

where the infimum is taken over all the copies (in distribution) η_1' of η_1 and η_2' of η_2. Note that W_p is, in fact, the distance between the distributions of η_1 and η_2. It is easy to see that

1. for all $c \in \mathbb{R}$
$$W_p(\eta_1 + c, \eta_2 + c) = W_p(\eta_1, \eta_2); \tag{12}$$
2. for all $c \in \mathbb{R}$
$$W_p(c\eta_1, c\eta_2) = |c| W_p(\eta_1, \eta_2). \tag{13}$$

Also, it is well known that

$$W_1(\eta_1, \eta_2) = \sup_{\|g\|_{\mathrm{Lip}(\mathbb{R})} \leq 1} |\mathbb{E}g(\eta_1) - \mathbb{E}g(\eta_2)|.$$

The last formula allows one to show that, for an arbitrary r.v. η and $Z \sim N(0,1)$,

$$d_K(\eta, Z) \leq 2W_1^{1/2}(\eta, Z) \leq 2W_2^{1/2}(\eta, Z). \tag{14}$$

Let

$$\sigma_f^2(\theta) := \sigma_f^2(\theta, \xi) := \mathbb{E}\langle \xi, f'(\theta)\rangle^2.$$

The following corollary is obvious.

Corollary 1 *Suppose ξ is a r.v. in E with mean zero, $\mathbb{E}^{1/2}\|\xi\|^2 \le 1/2$, covariance operator Σ_ξ and $C_P(\xi) < \infty$. Then, for all $s = k+1+\rho$, $k \ge 0, \rho \in (0,1]$*

$$\sup_{\|f\|_{C^s}\le 1} \sup_{\theta\in E} \Big| \mathbb{E}_\theta^{1/2}(f_k(X) - f(\theta))^2 - \sigma_f(\theta) \Big| \le 2\Big(3C_P^{1/2}(\xi)(\mathbb{E}^{1/2}\|\xi\|^2)^\rho + (\mathbb{E}^{1/2}\|\xi\|^2)^s\Big).$$

and

$$\sup_{\|f\|_{C^s}\le 1} \sup_{\theta\in E} W_2(f_k(X) - f(\theta), \sigma_f(\theta)Z) \le 2\Big(3C_P^{1/2}(\xi)(\mathbb{E}^{1/2}\|\xi\|^2)^\rho + (\mathbb{E}^{1/2}\|\xi\|^2)^s\Big).$$

As an example, consider the following linear model:

$$X = A\beta + n^{-1/2}\xi,$$

where $\beta \in \mathbb{R}^p$ is the vector of unknown coefficients, $A : \mathbb{R}^p \mapsto \mathbb{R}^m$ is a linear operator and ξ is a mean zero random vector in $\mathbb{R}^m$. Let $L := \mathrm{Im}(A) = A\mathbb{R}^p \subset \mathbb{R}^m$ with $d := \dim(L)$ and let $\theta = A\beta \in L$. Denote $\hat\theta = A\hat\beta = P_L X$, where $\hat\beta$ is a least squares estimator of β and P_L denotes the orthogonal projection onto subspace L. Then $\hat\theta = \theta + P_L(n^{-1/2}\xi)$ and we can use this random shift model to define the associated functions f_k. The goal is to estimate a smooth functional $f(\theta)$ of unknown parameter $\theta = A\beta$ based on an observation X when the noise level tends to zero (or $n \to \infty$).

Corollary 2 *Suppose that $d \lesssim n^\alpha$ for some $\alpha \in (0,1)$ and $s > \frac{1}{1-\alpha}$. Suppose also that $C_P(\xi) \lesssim 1$. Then*

$$\sup_{\|f\|_{C^s(L)}\le 1} \sup_{\theta\in E} \Big| n\mathbb{E}_\theta(f_k(\hat\theta) - f(\theta))^2 - \sigma_f^2(\theta) \Big| \to 0$$

and, for all $\sigma_0 > 0$,

$$\sup_{\|f\|_{C^s(L)}\le 1} \sup_{\theta\in E, \sigma_f(\theta)\ge\sigma_0} d_K\Big(\frac{\sqrt{n}(f_k(\hat\theta) - f(\theta))}{\sigma_f(\theta)}, Z\Big) \to 0 \text{ as } n \to \infty.$$

Consider now the case of i.i.d. observations $X_1, \dots, X_n$ of r.v. $X = \theta+\xi, \theta \in E$ with mean zero noise ξ such that $C_P(\xi) < \infty$. We will use $\bar X = \frac{X_1+\cdots+X_n}{n}$ as an estimator of θ and define the functions f_k associated with the model $\bar X = \theta + \bar\xi$.

Suppose the following assumption holds: for some parameter $d \ge 1$, for all $n \ge 1$ and for i.i.d. copies $\xi_1, \dots, \xi_n$ of ξ,

$$\mathbb{E}\Big\|\frac{\xi_1 + \cdots + \xi_n}{n}\Big\|^2 \lesssim C_P(\xi)\frac{d}{n}. \tag{15}$$

In this assumption, d is typically a dimension, or some other complexity parameter of the problem (such as the effective rank of ξ, see [17]). For instance, if $E = \mathbb{R}^d$ is equipped with the standard Euclidean norm, then

$$\mathbb{E}\Big\|\frac{\xi_1 + \cdots + \xi_n}{n}\Big\|^2 = \frac{\mathrm{tr}(\Sigma_\xi)}{n} \le \|\Sigma_\xi\|\frac{d}{n} \le C_P(\xi)\frac{d}{n},$$

so, assumption (15) holds.

More generally, the following simple proposition provides a sufficient condition for assumption (15) to hold.

Proposition 3 *Let $U^* := \{u \in E^* : \|u\| \le 1\}$ be the unit ball of the dual space E^*. Suppose there exists a set $\mathcal{U} \subset U^*$ of cardinality* $\mathrm{card}(\mathcal{U}) \le C^d$ *for some numerical constant C such that*

$$\|x\| \lesssim_C \max_{u \in \mathcal{U}} |\langle x, u\rangle|.$$

Then condition (15) *holds.*

In particular, this proposition could be used in the case when $E = \mathbb{M}_d$ is the space of $d \times d$ matrices with real entries equipped with the operator norm. The dual space E^* is equipped with the nuclear norm and, for all $A \in \mathbb{M}_d$,

$$\|A\| = \sup_{\|u\|\le 1, \|v\|\le 1, u,v \in \mathbb{R}^d} |\langle Au, v\rangle| \le 2 \max_{u,v \in M} |\langle Au, v\rangle|,$$

where M is a 1/4-net for $\{u \in \mathbb{R}^d : \|u\| \le 1\}$. Note that one can choose M such that $\mathrm{card}(M) \le 9^d$ (see [33], Chapter 4 for more details on such discretization arguments). Thus, one can choose $\mathcal{U} := \{u \otimes v : u, v \in M\}$ for which $\mathrm{card}(\mathcal{U}) \le 81^d$.

Under assumption (15), the following result will be proved (that can be applied, for instance, to the examples of $E = \mathbb{R}^d$ or $E = \mathbb{M}_d$).

Corollary 3 *Suppose assumption* (15) *holds for some $d \ge 1$. Then, for all $s = k + 1 + \rho$, $k \ge 0, \rho \in (0, 1]$,*

$$\sup_{\|f\|_{C^s}\le 1} \sup_{\theta \in E} \mathbb{E}_\theta (f_k(\bar{X}) - f(\theta))^2 \lesssim_s \Big(\frac{\|\Sigma_\xi\|}{n} \vee \frac{C_P^{1+\rho}(\xi)}{n} \Big(\frac{d}{n}\Big)^\rho \vee C_P^s(\xi) \Big(\frac{d}{n}\Big)^s \Big) \wedge 1. \tag{16}$$

Moreover, if $C_P(\xi) \lesssim 1$, $d \lesssim n^\alpha$ for some $\alpha \in (0,1)$ and $s > \frac{1}{1-\alpha}$, then

$$\sup_{\|f\|_{C^s}\le 1} \sup_{\theta \in E} \Big| n\mathbb{E}_\theta (f_k(\bar{X}) - f(\theta))^2 - \sigma_f^2(\theta)\Big| \to 0$$

and, for all $\sigma_0 > 0$,

$$\sup_{\|f\|_{C^s}\le 1} \sup_{\theta \in E, \sigma_f(\theta) \ge \sigma_0} d_K\Big(\frac{\sqrt{n}(f_k(\bar{X}) - f(\theta))}{\sigma_f(\theta)}, Z\Big) \to 0 \text{ as } n \to \infty.$$

3 Proofs

The analysis of estimator $f_k(X)$ is based on simple representation formulas for functions $\mathcal{B}^j f$. Note that, by Newton's binomial formula,

$$\mathcal{B}^j = (\mathcal{T} - \mathcal{I})^j = \sum_{i=0}^{j} (-1)^{j-i} \binom{j}{i} \mathcal{T}^i.$$

Let $\xi_1, \xi_2, \dots$ be i.i.d. copies of ξ. It follows from the definition of operator $\mathcal{T}$ that

$$(\mathcal{T}^i f)(\theta) = \mathbb{E} f\Big(\theta + \sum_{l=1}^{i} \xi_l\Big). \tag{17}$$

Thus, one can write

$$(\mathcal{B}^j f)(\theta) = \mathbb{E} \sum_{i=0}^{j} (-1)^{j-i} \binom{j}{i} f\Big(\theta + \sum_{l=1}^{i} \xi_l\Big). \tag{18}$$

For a k times continuously differentiable function f, its j-th order difference

$$(\Delta_h^j f)(\theta) = \Delta_h \dots \Delta_h f(\theta) = \sum_{i=0}^{j} (-1)^{j-i} \binom{j}{i} f(\theta + ih) = f^{(j)}(\theta)[h, \dots, h] + o(\|h\|^j),$$

where $(\Delta_h f)(\theta) := f(\theta + h) - f(\theta)$. In particular, this implies that

$$|(\Delta_h^j f)(\theta)| = O(\|h\|^j) \text{ as } h \to 0.$$

These facts are well known for smooth functions in the real line and could be easily extended to an arbitrary Banach space. The function $(\mathcal{B}^j f)(\theta)$ is the expectation of the "j-th order difference" of f with respect to i.i.d. random translations $\xi_1, \dots, \xi_j$ (rather than a single translation h) and we need similar properties of j-th order differences in this case.

Proposition 4 *Let f be a j times Fréchet continuously differentiable function on E. Suppose that $\mathbb{E}\|\xi\| < \infty$. Let $\tau_1, \tau_2, \dots$ be i.i.d. r.v. uniformly distributed in $[0, 1]$ and independent of $\xi_1, \xi_2, \dots$. Then*

$$(\mathcal{B}^j f)(\theta) = \mathbb{E} f^{(j)}\Big(\theta + \sum_{l=1}^{j} \tau_l \xi_l\Big)[\xi_1, \dots, \xi_j], \theta \in E. \tag{19}$$

Proof Define

$$\psi(t_1, \dots, t_j) := f\Big(\theta + \sum_{l=1}^{j} t_l \xi_l\Big), (t_1, \dots, t_j) \in [0, 1]^j.$$

Then, for all $(t_1, \dots, t_j) \in \{0, 1\}^j$ such that $\sum_{l=1}^{j} t_l = i$, we have

$$(\mathcal{T}^i f)(\theta) = \mathbb{E}\psi(t_1, \dots, t_j).$$

Therefore,

$$\begin{aligned}
(\mathcal{B}^j f)(\theta) &= \sum_{i=0}^{j} (-1)^{j-i} \binom{j}{i} (\mathcal{T}^i f)(\theta) \\
&= \sum_{i=0}^{j} (-1)^{j-i} \sum_{(t_1, \dots, t_j) \in \{0,1\}^j, \sum_{l=1}^{j} t_l = i} \mathbb{E}\psi(t_1, \dots, t_j) \\
&= \mathbb{E} \sum_{(t_1, \dots, t_j) \in \{0,1\}^j} (-1)^{j - \sum_{l=1}^{j} t_l} \psi(t_1, \dots, t_j). \qquad (20)
\end{aligned}$$

It is also easy to observe that

$$\sum_{(t_1,\dots,t_j)\in\{0,1\}^j} (-1)^{j-\sum_{l=1}^j t_l}\psi(t_1,\dots,t_j) = \Delta^{(1)}\dots\Delta^{(j)}\psi(t_1,\dots,t_j), \tag{21}$$

where

$$\Delta^{(l)}\psi(t_1,\dots,t_j) := \psi(t_1,\dots,t_j)_{|t_l=1} - \psi(t_1,\dots,t_j)_{|t_l=0}.$$

Moreover, if f is j times continuously differentiable, then so is the function $\psi(t_1,\dots,t_j)$ and

$$\frac{\partial^j\psi(t_1,\dots,t_j)}{\partial t_1\dots\partial t_j} = f^{(k)}\Big(\theta+\sum_{l=1}^j t_l\xi_l\Big)[\xi_1,\dots,\xi_j]. \tag{22}$$

By Newton-Leibnitz formula,

$$\Delta^{(1)}\dots\Delta^{(j)}\psi(t_1,\dots,t_j) = \int_0^1\cdots\int_0^1 \frac{\partial^j\psi(t_1,\dots,t_j)}{\partial t_1\dots\partial t_j}dt_1\dots dt_j. \tag{23}$$

It follows from (20), (21), (22) and (23) that

$$(\mathcal{B}^j f)(\theta) = \int_0^1\cdots\int_0^1 \mathbb{E}f^{(j)}\Big(\theta+\sum_{l=1}^j t_l\xi_l\Big)[\xi_1,\dots,\xi_j],$$

which implies (19).

□

This simple integral representation of functions $\mathcal{B}^j f$ (obtained first in [17]) provides a way to study various smoothness properties of these functions under further smoothness assumptions on f. These properties are summarized in the following proposition (it could be proved similarly to Proposition 3.3, Theorem 3.3. and Proposition 3.4 in [17], where the proof is given in the Gaussian case).

Proposition 5 *Suppose $\mathbb{E}\|\xi\| < \infty$. The following statements hold:*

1. *If $f \in C^j(E)$, then*

$$\|B^j f\|_{L_\infty} \le \|f^{(j)}\|_{L_\infty}(\mathbb{E}\|\xi\|)^j.$$

2. *If $f \in C^{j+1}(E)$, then the function $\mathcal{B}^j f$ is Fréchet continuously differentiable with derivative*

$$\langle h, (\mathcal{B}^j f)'(\theta)\rangle = (\mathcal{B}^j f)'(\theta)[h] = \mathbb{E}f^{(j+1)}\Big(\theta+\sum_{l=1}^j \tau_l\xi_l\Big)[\xi_1,\dots,\xi_j,h], h\in E.$$

3. *If $s = k+1+\rho$ for some $k\ge 0$, $\rho\in(0,1]$, and $f\in C^s(E)$, then, for all $j=1,\dots,k$,*

$$\|\mathcal{B}^j f\|_{C^{1+\rho}} \le 2\|f\|_{C^s}(\mathbb{E}\|\xi\|)^j.$$

It immediately follows from Proposition 5 and the definition of function f_k that

$$\|f_k\|_{C^{1+\rho}} \leq 4\|f\|_{C^s} \tag{24}$$

provided that $\mathbb{E}\|\xi\| \leq 1/2$.

It is now easy to derive an upper bound on the bias of estimator $f_k(X)$ of functional $f(\theta)$. For this, the following simple lemma will be useful.

Lemma 1 *For a Fréchet differentiable function $g : E \mapsto \mathbb{R}$, denote by*

$$S_g(\theta; h) := g(\theta + h) - g(\theta) - \langle h, g'(\theta)\rangle, \theta, h \in E$$

the remainder of its first order Taylor expansion. Then

$$|S_g(\theta; h)| \leq \|g\|_{C^{1+\rho}}\|h\|^{1+\rho}.$$

In addition, the function $E \ni h \mapsto S_g(\theta; h)$ is continuously differentiable with derivative

$$(S_g(\theta; h))'_h = g'(\theta + h) - g'(\theta)$$

and

$$\|(S_g(\theta; h))'_h\| \leq \|g\|_{C^{1+\rho}}\|h\|^{\rho}.$$

Proposition 6 *Suppose $s = k + 1 + \rho$ for some $k \geq 0$, $\rho \in (0, 1]$, and $f \in C^s(E)$. Then*

$$|\mathbb{E}_\theta f_k(X) - f(\theta)| \leq 2\|f\|_{C^s}(\mathbb{E}\|\xi\|)^k \mathbb{E}\|\xi\|^{1+\rho}, \theta \in E.$$

Proof Recall that

$$\mathbb{E}_\theta f_k(X) - f(\theta) = (-1)^k(\mathcal{B}^{k+1}f)(\theta), \theta \in E.$$

We have

$$(\mathcal{B}^{k+1}f)(\theta) = \mathcal{B}(\mathcal{B}^k f)(\theta) = \mathbb{E}(\mathcal{B}^k f)(\theta + \xi) - (\mathcal{B}^k f)(\theta)$$
$$= \mathbb{E}\langle \xi, (\mathcal{B}^k f)'(\theta)\rangle + \mathbb{E}S_{\mathcal{B}^k f}(\theta; \xi) = \mathbb{E}S_{\mathcal{B}^k f}(\theta; \xi).$$

Thus, using statement 3 of Proposition 5 and the first bound of Lemma 1, we get

$$|(\mathcal{B}^{k+1}f)(\theta)| \leq \mathbb{E}|S_{\mathcal{B}^k f}(\theta; \xi)| \leq \|\mathcal{B}^k f\|_{C^{1+\rho}}\mathbb{E}\|\xi\|^{1+\rho}$$
$$\leq 2\|f\|_{C^s}(\mathbb{E}\|\xi\|)^k \mathbb{E}\|\xi\|^{1+\rho},$$

implying the claim.

□

We are now ready to prove Theorem 2.

Proof We will use the following decomposition:

$$\begin{aligned}
&f_k(X) - f(\theta) = \mathbb{E}_\theta f_k(X) - f(\theta) + f_k(X) - \mathbb{E}_\theta f_k(X)\\
&= \mathbb{E}_\theta f_k(X) - f(\theta) + f_k(\theta+\xi) - \mathbb{E} f_k(\theta+\xi)\\
&= \mathbb{E}_\theta f_k(X) - f(\theta) + \langle \xi, f_k^{'}(\theta)\rangle + S_{f_k}(\theta;\xi) - \mathbb{E}S_{f_k}(\theta;\xi)\\
&= \mathbb{E}_\theta f_k(X) - f(\theta) + \langle \xi, f'(\theta)\rangle + \langle \xi, f_k^{'}(\theta) - f'(\theta)\rangle + S_{f_k}(\theta;\xi) - \mathbb{E}S_{f_k}(\theta;\xi).
\end{aligned}$$

It implies that

$$\begin{aligned}
&\|f_k(X) - f(\theta) - \langle \xi, f'(\theta)\rangle\|_{L_2(\mathbb{P}_\theta)}\\
&\le |\mathbb{E}_\theta f_k(X) - f(\theta)| + \|\langle \xi, f_k^{'}(\theta) - f'(\theta)\rangle\|_{L_2(\mathbb{P})} + \|S_{f_k}(\theta;\xi) - \mathbb{E}S_{f_k}(\theta;\xi)\|_{L_2(\mathbb{P})}.
\end{aligned} \tag{25}$$

To control the bias of $f_k(X)$, we use Proposition 6 and observe that

$$(\mathbb{E}\|\xi\|)^k \mathbb{E}\|\xi\|^{1+\rho} \le (\mathbb{E}\|\xi\|^2)^{k/2}(\mathbb{E}\|\xi\|^2)^{(1+\rho)/2} = (\mathbb{E}\|\xi\|^2)^{s/2}.$$

This yields

$$|\mathbb{E}_\theta f_k(X) - f(\theta)| \le 2\|f\|_{C^s}(\mathbb{E}\|\xi\|^2)^{s/2}, \theta \in E. \tag{26}$$

Note also that

$$\begin{aligned}
&\|\langle \xi, f_k^{'}(\theta) - f'(\theta)\rangle\|_{L_2(\mathbb{P})} = \langle \Sigma_\xi(f_k^{'}(\theta) - f'(\theta)), f_k^{'}(\theta) - f'(\theta)\rangle^{1/2}\\
&\le \|\Sigma_\xi\|^{1/2}\|f_k^{'}(\theta) - f'(\theta)\|
\end{aligned}$$

and, by statement 2 of Proposition 5,

$$\|f_k^{'}(\theta) - f'(\theta)\| \le \sum_{j=1}^k \|(\mathcal{B}^j f)'(\theta)\| \le \sum_{j=1}^k \|f^{(j+1)}\|_{L_\infty}(\mathbb{E}\|\xi\|)^j \le 2\|f\|_{C^s}\mathbb{E}\|\xi\|$$

(under the assumption that $\mathbb{E}\|\xi\| \le 1/2$). Therefore,

$$\|\langle \xi, f_k^{'}(\theta) - f'(\theta)\rangle\|_{L_2(\mathbb{P})} \le 2\|f\|_{C^s}\|\Sigma_\xi\|^{1/2}\mathbb{E}^{1/2}\|\xi\|^2. \tag{27}$$

Finally, let

$$g(h) := S_{f_k}(\theta;h), h \in E.$$

By the second bound of Lemma 1,

$$\|g'(h)\| \le \|f_k\|_{C^{1+\rho}}\|h\|^\rho.$$

Using Poincaré inequality, we get

$$\|S_{f_k}(\theta;\xi) - \mathbb{E}S_{f_k}(\theta;\xi)\|^2_{L_2(\mathbb{P})} = \mathrm{Var}(g(\xi)) \le \|f_k\|^2_{C^{1+\rho}} C_P(\xi)\mathbb{E}\|\xi\|^{2\rho},$$

and, using (24), this implies

$$\|S_{f_k}(\theta;\xi) - \mathbb{E}S_{f_k}(\theta;\xi)\|_{L_2(\mathbb{P})} \le 4\|f\|_{C^s} C_P^{1/2}(\xi)(\mathbb{E}^{1/2}\|\xi\|^2)^\rho. \tag{28}$$

It follows from bounds (25), (26), (27) and (28) that

$$\|f_k(X) - f(\theta) - \langle \xi, f'(\theta)\rangle\|_{L_2(\mathbb{P}_\theta)} \tag{29}$$
$$\le 2\|f\|_{C^s}(\mathbb{E}\|\xi\|^2)^{s/2} + 2\|f\|_{C^s}\|\Sigma_\xi\|^{1/2}\mathbb{E}^{1/2}\|\xi\|^2 + 4\|f\|_{C^s}C_P^{1/2}(\xi)(\mathbb{E}^{1/2}\|\xi\|^2)^\rho. \tag{30}$$

Since $\|\Sigma_\xi\|^{1/2}\mathbb{E}^{1/2}\|\xi\|^2 \le C_P^{1/2}(\xi)(\mathbb{E}^{1/2}\|\xi\|^2)^\rho$, this completes the proof of bound (11). To complete the proof of bound (10), observe that

$$\|\langle \xi, f'(\theta)\rangle\|_{L_2(\mathbb{P}_\theta)} = \langle \Sigma_\xi f'(\theta), f'(\theta)\rangle^{1/2} \le \|\Sigma_\xi\|^{1/2}\|f'(\theta)\| \le \|\Sigma_\xi\|^{1/2}$$

for all f satisfying $\|f\|_{C^s} \le 1$, and combine this with bound (29). It now remains to show that

$$\sup_{\|f\|_{C^s}\le 1}\sup_{\theta\in E}\mathbb{E}_\theta(f_k(X) - f(\theta))^2 \lesssim_s 1,$$

which follows from the facts that, for $\|f\|_{C^s} \le 1$, we have $\|f\|_{L_\infty} \le 1$ and $\|f_k\|_{L_\infty} \lesssim_k 1$ (the last fact easily follows from the definition of f_k and representation (18)).

□

We give the proof of Proposition 3.

Proof For $\bar{\xi} = \frac{\xi_1+\cdots+\xi_n}{n}$, we have

$$\|\bar{\xi}\| \lesssim_C \max_{u\in\mathcal{U}}|\langle\bar{\xi}, u\rangle| = \max_{u\in\mathcal{U}}\Big|n^{-1}\sum_{j=1}^n\langle\xi_j, u\rangle\Big|.$$

It is well known and easy to check that, for all $u \in U^*$, $\langle \xi, u\rangle$ is a sub-exponential r.v. with[5]

$$\|\langle\xi, u\rangle\|_{\psi_1} \lesssim C_P^{1/2}(\xi)$$

(see [21], Theorem 3.1). Using Bernstein's inequality for sums of i.i.d. sub-exponential r.v. (see [33], Theorem 2.8.1), we conclude that, for all $u \in \mathcal{U} \subset U^*$, with probability at least $1 - e^{-t}$

$$\Big|n^{-1}\sum_{j=1}^n\langle\xi_j, u\rangle\Big| \lesssim C_P^{1/2}(\xi)\Big(\sqrt{\frac{t}{n}} \vee \frac{t}{n}\Big).$$

Recall that $\mathrm{card}(\mathcal{U}) \le C^d$. Replacing t by $t + d\log C$ and using the union bound, we get that with probability at least $1 - e^{-t}$

$$\|\bar{\xi}\| \lesssim_C \max_{u\in\mathcal{U}}\Big|n^{-1}\sum_{j=1}^n\langle\xi_j, u\rangle\Big| \lesssim C_P^{1/2}(\xi)\Big(\sqrt{\frac{t+d}{n}} \vee \frac{t+d}{n}\Big).$$

[5] Here $\|\cdot\|_{\psi_1}$ is the Orlicz ψ-norm corresponding to the sub-exponential tails; see [33], Chapter 2 for the definitions.

Integrating out the tail probabilities in the above inequality, we easily get that, for $d \leq n$,

$$\mathbb{E}\|\bar{\xi}\|^2 \lesssim C_P(\xi)\frac{d}{n}.$$

□

The proof of Corollary 2 is straightforward. We will now prove corollary 3.

Proof Note that $\bar{X} = \theta + \bar{\xi}$, where $\bar{\xi} = \frac{\xi_1+\cdots+\xi_n}{n}$, $\xi_1, \dots, \xi_n$ being i.i.d. copies of ξ and, by assumption (15),

$$\mathbb{E}\|\bar{\xi}\|^2 \leq C_P(\xi)\frac{d}{n}.$$

Since

$$C_P((\xi_1, \dots, \xi_n)) \leq \max_{1\leq j\leq n} C_P(\xi_j) = C_P(\xi),$$

it is easy to check that

$$C_P(\bar{\xi}) \leq \frac{C_P(\xi)}{n}.$$

Indeed, denote $\tilde{g}(x_1, \dots, x_n) := g(n^{-1}(x_1+\cdots+x_n))$, $x_j \in E$. For $C = C_P(\xi)$, we have

$$\mathrm{Var}(g(\bar{\xi})) = \mathrm{Var}(\tilde{g}(\xi_1, \dots, \xi_n)) \leq C\mathbb{E}\|\tilde{g}'(\xi_1, \dots, \xi_n)\|^2.$$

Since

$$\tilde{g}'(\xi_1, \dots, \xi_n)[h] = \sum_{j=1}^n \langle h_j, \tilde{g}'_{x_j}(\xi_1, \dots, \xi_n)\rangle, h = (h_1, \dots, h_n) \in E \times \cdots \times E$$

and

$$\tilde{g}'_{x_j}(\xi_1, \dots, \xi_n) = n^{-1}g'(\bar{\xi}),$$

we easily get

$$\|\tilde{g}'(\xi_1, \dots, \xi_n)\|^2 = \sum_{j=1}^n \|\tilde{g}'_{x_j}(\xi_1, \dots, \xi_n)\|^2 = n^{-1}\|g'(\bar{\xi})\|^2.$$

Therefore,

$$\mathrm{Var}(g(\bar{\xi})) \leq \frac{C}{n}\mathbb{E}\|g'(\bar{\xi})\|^2,$$

implying the claim.

It is now easy to derive bound (16) from bound (10).

To prove other claims, note that $\sigma_f(\theta,\bar{\xi}) = \frac{\sigma_f(\theta,\xi)}{\sqrt{n}}$. It follows from Corollary 1 that

$$\sup_{\|f\|_{C^s}\le 1}\sup_{\theta\in E}\Big|\sqrt{n}\mathbb{E}_\theta^{1/2}(f_k(\bar{X})-f(\theta))^2-\sigma_f(\theta)\Big|$$
$$\le 2\Big(3C_P^{(1+\rho)/2}(\xi)\Big(\frac{d}{n}\Big)^{\rho/2}+\sqrt{n}C_P^{s/2}(\xi)\Big(\frac{d}{n}\Big)^{s/2}\Big)$$

and

$$\sup_{\|f\|_{C^s}\le 1}\sup_{\theta\in E}W_2(\sqrt{n}(f_k(\bar{X})-f(\theta)),\sigma_f(\theta)Z)$$
$$\le 2\Big(3C_P^{(1+\rho)/2}(\xi)\Big(\frac{d}{n}\Big)^{\rho/2}+\sqrt{n}C_P^{s/2}(\xi)\Big(\frac{d}{n}\Big)^{s/2}\Big)$$

with $\sigma_f(\theta)=\sigma_f(\theta,\xi)$. Assuming that $d\lesssim n^\alpha$ for some $\alpha\in(0,1)$ and $s>\frac{1}{1-\alpha}$, the right-hand sides of the above bounds tend to zero as $n\to\infty$, implying that

$$\sup_{\|f\|_{C^s}\le 1}\sup_{\theta\in E}\Big|\sqrt{n}\mathbb{E}_\theta^{1/2}(f_k(\bar{X})-f(\theta))^2-\sigma_f(\theta)\Big|\to 0$$

and

$$\sup_{\|f\|_{C^s}\le 1}\sup_{\theta\in E}W_2(\sqrt{n}(f_k(\bar{X})-f(\theta)),\sigma_f(\theta)Z)\to 0$$

(where we also used properties (12) and (13) of metric W_2). In view of bound (14) and property (13), the last relationship also implies that, for all $\sigma_0>0$,

$$\sup_{\|f\|_{C^s}\le 1}\sup_{\theta\in E,\sigma_f(\theta)\ge\sigma_0}d_K\Big(\frac{\sqrt{n}(f_k(\bar{X})-f(\theta))}{\sigma_f(\theta)},Z\Big)\to 0 \text{ as } n\to\infty,$$

which completes the proof.

□

We turn to the proof of Proposition 1.

Proof Note that, for the model $\bar{X}=\theta+\bar{\xi},\bar{\xi}\sim N(0;\Sigma_\xi/2)$, formula (18) becomes

$$(\mathcal{B}^j f)(\theta)=\mathbb{E}\sum_{l=0}^{j}(-1)^{j-l}\binom{j}{l}f\Big(\theta+\frac{1}{\sqrt{2}}\sum_{i=1}^{l}\xi_i\Big).$$

Note also that $\frac{1}{\sqrt{2}}\sum_{i=1}^l\xi_i\overset{d}{=}\sqrt{\frac{l}{2}}\xi$. Thus, for

$$\hat{V}_j(\theta;\xi)=\sum_{l=0}^{j}(-1)^{j-l}\binom{j}{l}f\Big(\theta+\sqrt{\frac{l}{2}}\xi\Big),$$

we have $\mathbb{E}\hat{V}_j(\theta;\xi) = (\mathcal{B}^j f)(\theta)$, which implies that $\mathbb{E}\hat{f}_k(\theta;\xi) = f_k(\theta)$. Note also that the function $g(x) := f_k(\theta;x), x \in E$ is continuously differentiable with

$$\|g'\|_{L_\infty} = \Bigl\|\sum_{j=0}^{k}(-1)^j (V_j)'_x(\theta;\cdot)\Bigr\|_{L_\infty} \le \sum_{j=0}^{k}\|(V_j)'_x(\theta;\cdot)\|_{L_\infty}$$
$$\le \sum_{j=0}^{k}\sum_{l=0}^{j}\binom{j}{l}\sqrt{\frac{l}{2}}\|f'\|_{L_\infty} \le \sqrt{k}2^{k+1/2}\|f'\|_{L_\infty}.$$

Therefore, by Poincaré inequality,

$$\begin{aligned}\mathrm{Var}(\hat{f}_k(\theta;\xi)) &= \mathrm{Var}(g(\xi)) \le \|\Sigma_\xi\|\|g'\|^2_{L_\infty}\\ &\le k2^{2k+1}\|f'\|^2_{L_\infty}\|\Sigma_\xi\| \le k2^{2k+1}\|f\|^2_{C^s}\|\Sigma_\xi\|. \end{aligned} \tag{31}$$

Since $\bar{X}$ and ξ are independent, we can now condition on $\bar{X}$ to get

$$\begin{aligned}&\mathbb{E}_\theta(\hat{f}_k(\bar{X}) - f(\theta))^2 = \mathbb{E}_\theta(\hat{f}_k(\bar{X}) - f_k(\bar{X}) + f_k(\bar{X}) - f(\theta))^2\\ &\mathbb{E}_\theta\mathbb{E}((\hat{f}_k(\bar{X}) - f_k(\bar{X}))^2|\bar{X}) + 2\mathbb{E}_\theta\mathbb{E}((\hat{f}_k(\bar{X}) - f_k(\bar{X})|\bar{X})(f_k(\bar{X}) - f(\theta))\\ &+ \mathbb{E}_\theta(f_k(\bar{X}) - f(\theta))^2. \end{aligned} \tag{32}$$

Note that

$$\mathbb{E}((\hat{f}_k(\bar{X};\xi) - f_k(\bar{X})|\bar{X}) = 0$$

since $\hat{f}_k(\theta;\xi)$ is an unbiased estimator of $f_k(\theta)$. Moreover, by bound (31),

$$\mathbb{E}((\hat{f}_k(\bar{X}) - f_k(\bar{X}))^2|\bar{X}) \le k2^{2k+1}\|f\|^2_{C^s}\|\Sigma_\xi\|.$$

Therefore,

$$\mathbb{E}_\theta(\hat{f}_k(\bar{X}) - f(\theta))^2 \le k2^{2k+1}\|f\|^2_{C^s}\|\Sigma_\xi\| + \mathbb{E}_\theta(f_k(\bar{X}) - f(\theta))^2,$$

and the result follows from the bound of Theorem 1.

□

We will now prove Proposition 2.

Proof As in the proof of Proposition 1,

$$\mathbb{E}\tilde{V}_j(\theta;\tilde{\xi}_i) = \sum_{l=0}^{j}(-1)^{j-l}\binom{j}{l}\mathbb{E}f\Bigl(\theta + \sqrt{\frac{l}{n}}\tilde{\xi}_i\Bigr) = (B^j f)(\theta),$$

which implies that $\mathbb{E}\tilde{f}_k(\theta;\tilde{\xi}_i) = f_k(\theta)$ and $\mathbb{E}\hat{f}_{m,k}(\theta) = f_k(\theta), \theta \in E$. Moreover, using Poincaré inequality, we get (similarly to (31)) that

$$\mathrm{Var}(\tilde{f}_k(\theta;\tilde{\xi}_i)) \le k2^{2k+1}\|f\|^2_{C^s}\frac{\|\Sigma_\xi\|}{n}.$$

Therefore, for all $\theta \in E$,

$$\mathrm{Var}(\hat{f}_{m,k}(\theta)) = \frac{1}{m^2}\sum_{i=1}^{m} \mathrm{Var}(\tilde{f}_k(\theta;\tilde{\xi}_i)) \leq k2^{2k+1}\|f\|_{C^s}^2 \frac{\|\Sigma_\xi\|}{nm}. \tag{33}$$

Since $\{\tilde{\xi}_i : i = 1,\dots,m\}$ and $\bar{X}$ are independent, we can use the same conditioning argument as in (32) to get

$$\mathbb{E}_\theta(\hat{f}_k(\bar{X}) - f(\theta))^2 = \mathbb{E}_\theta\mathbb{E}((\hat{f}_k(\bar{X}) - f_k(\bar{X}))^2|\bar{X}) + \mathbb{E}_\theta(f_k(\bar{X}) - f(\theta))^2,$$

which, in view of (33), implies that

$$|\mathbb{E}_\theta(\hat{f}_k(\bar{X}) - f(\theta))^2 - \mathbb{E}_\theta(f_k(\bar{X}) - f(\theta))^2| \leq k2^{2k+1}\|f\|_{C^s}^2 \frac{\|\Sigma_\xi\|}{nm}. \tag{34}$$

Taking into account (2), this immediately implies (6).

If $\|\Sigma_\xi\| \lesssim 1$ and $m \to \infty$ as $n \to \infty$, bound (34) also implies that

$$\sup_{\|f\|_{C^s}\leq 1}\sup_{\theta\in\Theta}|n\mathbb{E}_\theta(\hat{f}_k(\bar{X}) - f(\theta))^2 - n\mathbb{E}_\theta(f_k(\bar{X}) - f(\theta))^2| \to 0$$

as $n \to \infty$. Together with (3), this yields (7).

Finally, it also follows from bound (33) (again, by conditioning on $\bar{X}$) that

$$\|\hat{f}_{m,k}(\bar{X}) - f_k(\bar{X})\|_{L_2(\mathbb{P}_\theta)}^2 = \mathbb{E}_\theta\mathbb{E}((\hat{f}_{m,k}(\bar{X}) - f_k(\bar{X}))^2|\bar{X}) \leq k2^{2k+1}\|f\|_{C^s}^2 \frac{\|\Sigma_\xi\|}{nm}.$$

Therefore, using (12) and (13), we get

$$\begin{aligned} &W_2(\sqrt{n}(\hat{f}_{m,k}(\bar{X}) - f(\theta)), \sqrt{n}(f_k(\bar{X}) - f(\theta))) \\ &\leq \sqrt{n}\|\hat{f}_{m,k}(\bar{X}) - f_k(\bar{X})\|_{L_2(\mathbb{P}_\theta)} \leq k^{1/2}2^{k+1/2}\|f\|_{C^s}\frac{\|\Sigma_\xi\|^{1/2}}{\sqrt{m}}. \end{aligned}$$

If $\|\Sigma_\xi\| \lesssim 1$ and $m \to \infty$ as $n \to \infty$, the last bound and (13) imply that, for all $\sigma_0 > 0$,

$$\sup_{\|f\|_{C^s}\leq 1}\ \sup_{\theta\in\Theta,\sigma_f(\theta)\geq\sigma_0} W_2\Big(\frac{\sqrt{n}(\hat{f}_{m,k}(\bar{X}) - f_k(\theta))}{\sigma_f(\theta)}, \frac{\sqrt{n}(f_k(\bar{X}) - f(\theta))}{\sigma_f(\theta)}\Big) \to 0$$

as $n \to \infty$, and the same limit relationship also holds for the distance d_K (in view of inequality (14)). It remains to combine it with (4) to complete the proof.

□

Acknowledgements V. Koltchinskii was supported in part by NSF Grant DMS-1810958. M. Zhilova was supported in part by NSF Grant DMS-1712990.

References

1. F. Barthe and B. Klartag. Spectral gaps, symmetries and log-concave perturbations. 2019, *arXiv:1907.01823.*
2. P. Bickel and Y. Ritov. Estimating integrated square density derivatives: sharp best order of convergence estimates. *Sankhya*, 1988, 50, 381–393.
3. L. Birgé and P. Massart. Estimation of integral functionals of a density. *Annals of Statistics*, 1995, 23, 11-29.
4. T.T. Cai and M. Low. Non-quadratic estimators of a quadratic functional. *Annals of Statistics*, 2005, 33, 2930–2956.
5. H. Cartan. *Differential Calculus*, Houghton Mifflin Co, 1971.
6. O. Collier, L. Comminges and A. Tsybakov. Minimax estimation of linear and quadratic functionals on sparsity classes. *Annals of Statistics*, 2017, 45, 3, 923–958.
7. D. Donoho and M. Nussbaum. Minimax quadratic estimation of a quadratic functional. *J. Complexity*, 1990, 6, 290–323.
8. R.A. Fisher. On the mathematical foundation of theoretical statistics. *Philosophical Transactions of the Royal Society of London, Series A*, 1922, 222, 309–368.
9. R.A. Fisher. Theory of statistical estimation. *Proceedings of the Cambridge Philosophical Society*, 1925, 22, 700–725.
10. J. Hàjek. Local asymptotic minimax and admissibility in estimation. *Proceedings of the Sixth Berkeley Symposium on Mathematical Statistics and Probability*, 1972, 1, 175–194.
11. Y. Han, J. Jiao and R. Mukherjee. On estimation of L_r-norms in Gaussian white noise model. *Probability Theory and Related Fields*, 2020, to appear. *arXiv:1710.03863.*
12. I. A. Ibragimov and R.Z. Khasminskii. *Statistical Estimation: Asymptotic Theory.* Springer-Verlag, New York, 1981.
13. I.A. Ibragimov, A.S. Nemirovski and R.Z. Khasminskii. Some problems of nonparametric estimation in Gaussian white noise. *Theory of Probab. and Appl.*, 1987, 31, 391–406.
14. V. Koltchinskii. Asymptotically Efficient Estimation of Smooth Functionals of Covariance Operators. *J. European Math. Society*, 2020, to appear. *arxiv:1710.09072*
15. V. Koltchinskii. Asymptotic Efficiency in High-Dimensional Covariance Estimation. *Proc. ICM 2018*, Rio de Janeiro, 2018, vol. 3, pp. 2891–2912.
16. V. Koltchinskii, M. Löffler and R. Nickl. Efficient Estimation of Linear Functionals of Principal Components. *Annals of Statistics*, 2020, 48,1, 464–490.
17. V. Koltchinskii and M. Zhilova. Efficient Estimation of Smooth Functionals in Gaussian Shift Models. *Annales de l'Institut Henri Poincaré, Probabilités et Statistiques*, 2020, to appear. *arXiv:1810.02767.*
18. V. Koltchinskii and M. Zhilova. Estimation of Smooth Functionals in Normal Models: Bias Reduction and Asymptotic Efficiency. 2019. *arXiv:1912.08877*
19. L. Le Cam. On some asymptotic properties of maximum likelihood estimates and related Bayes estimates. *University of California Publications in Statistics*, 1953, 1, 277–330.
20. L. Le Cam. Limits of experiments. Proc. Sixth Berkeley Symp. Math. Statist. Prob., I, 1972, L. Le Cam, J. Neyman and E.L. Scott (Eds.), University of California Press, 245-261.
21. M. Ledoux. *The Concentration of Measure Phenomenon.* American Mathematical Society. 2001.
22. Y.-T. Lee and S. Vempala. Eldan's Stochastic Localization and the KLS Hyperplane Conjecture: An Improved Lower Bound for Expansion. In: *58th Annual IEEE Symposium on Foundations of Computer Science FOCM 2017.*
23. B. Levit. On the efficiency of a class of non-parametric estimates. *Theory of Prob. and applications*, 1975, 20(4), 723–740.
24. B. Levit. Asymptotically efficient estimation of nonlinear functionals. *Probl. Peredachi Inf. (Problems of Information Transmission)*, 1978, 14(3), 65–72.
25. O. Lepski, A. Nemirovski and V. Spokoiny. On estimation of the L_r norm of a regression function. *Probab. Theory Relat. Fields*, 1999, 113, 221–253.
26. R. Mukherjee, W. Newey and J. Robins. Semiparametric Efficient Empirical Higher Order Influence Function Estimators. 2017, *arXiv:1705.07577.*
27. A. Nemirovski. On necessary conditions for the efficient estimation of functionals of a nonparametric signal which is observed in white noise. *Theory of Probab. and Appl.*, 1990, 35, 94–103.

28. A. Nemirovski. *Topics in Non-parametric Statistics.* Ecole d'Ete de Probabilités de Saint-Flour. Lecture Notes in Mathematics, v. 1738, Springer, New York, 2000.
29. C.R. Rao. Criteria of estimation in large samples. *Sankhya*, Ser. A., 1963, 25, 2, 189–206.
30. J. Robins, L. Li, E. Tchetgen and A. van der Vaart. Higher order influence functions and minimax estimation of nonlinear functionals. *IMS Collections Probability and Statistics: Essays in Honor of David. A. Freedman*, 2008, vol. 2, 335-421.
31. J. Robins, L. Li, E. Tchetgen and A. van der Vaart. Asymptotic Normality of Quadratic Estimators. *Stochastic Processes and Their Applications.* 2016, 126(12), 3733–3759.
32. A. van der Vaart. Higher order tangent spaces and influence functions. *Statistical Science*, 2014, 29, 4, 679–686.
33. R. Vershynin. *High-Dimensional Probability. An Introduction with Applications to Data Science.* Cambridge University Press, 2018.
34. F. Zhou and P. Li. A Fourier Analytical Approach to Estimation of Smooth Functions in Gaussian Shift Model. 2019, *arXiv:1911.02010.*