Submitted to the Annals of Statistics

ESTIMATION OF SMOOTH FUNCTIONALS IN HIGH-DIMENSIONAL MODELS: BOOTSTRAP CHAINS AND GAUSSIAN APPROXIMATION

BY VLADIMIR KOLTCHINSKII*

Georgia Institute of Technology§

School of Mathematics
Georgia Institute of Technology
Atlanta, GA 30332-0160
vlad@math.gatech.edu

April 16, 2022

Let $X^{(n)}$ be an observation sampled from a distribution $P_\theta^{(n)}$ with an unknown parameter θ, θ being a vector in a Banach space E (most often, a high-dimensional space of dimension d). We study the problem of estimation of $f(\theta)$ for a functional $f : E \mapsto \mathbb{R}$ of some smoothness $s > 0$ based on an observation $X^{(n)} \sim P_\theta^{(n)}$. Assuming that there exists an estimator $\hat{\theta}_n = \hat{\theta}_n(X^{(n)})$ of parameter θ such that $\sqrt{n}(\hat{\theta}_n - \theta)$ is sufficiently close in distribution to a mean zero Gaussian random vector in E, we construct a functional $g : E \mapsto \mathbb{R}$ such that $g(\hat{\theta}_n)$ is an asymptotically normal estimator of $f(\theta)$ with $\sqrt{n}$ rate provided that $s > \frac{1}{1-\alpha}$ and $d \le n^\alpha$ for some $\alpha \in (0,1)$. We also derive general upper bounds on Orlicz norm error rates for estimator $g(\hat{\theta})$ depending on smoothness s, dimension d, sample size n and the accuracy of normal approximation of $\sqrt{n}(\hat{\theta}_n - \theta)$. In particular, this approach yields asymptotically efficient estimators in high-dimensional log-concave exponential models.

1. Introduction. The problem of estimation of a smooth functional $f(\theta)$ of parameter θ of a high-dimensional statistical model is studied in this paper in the case when there exists an estimator $\hat{\theta}$ of θ for which normal approximation holds as both the dimension d and the sample size are reasonably large.

Estimation of functionals of parameters of non-parametric and, more recently, high-dimensional statistical models has been studied by many authors since the 70s [46, 47, 31, 5, 32, 26, 27, 19, 20, 21, 51, 6, 7, 44, 48, 52, 10, 11, 34, 59, 62, 17, 60, 28, 50]. Most of the results have been obtained for special statistical models (Gaussian sequence model, Gaussian white noise model, density estimation model) and special functionals (linear and quadratic functionals, norms in classical Banach spaces, certain classes of integral functionals of unknown density). Estimation of general smooth functionals was studied in [32, 51, 52] for the model of an unknown infinite-dimensional function (signal) observed in a Gaussian white noise. Sharp thresholds on the smoothness of the functional depending on the complexity (smoothness) of the signal that guarantee efficient estimation of the functional were studied in these papers.

*Supported in part by NSF grants DMS-1810958 and DMS-2113121
AMS 2000 subject classifications: Primary 62H12; secondary 62G20, 62H25, 60B20.
Keywords and phrases: Efficiency, Smooth functionals, Bootstrap chain, Concentration inequalities, Normal approximation.

Our approach is based on a bias reduction method that goes back to the idea of iterated bootstrap (see [30, 29]). This method has been recently studied in the case of high-dimensional normal models (see [36, 37, 40, 41]). In particular, it was shown that it yields efficient estimation of functionals of smoothness s of unknown mean and covariance with parametric $\sqrt{n}$ convergence rate provided that $s > \frac{1}{1-\alpha}$ and $d \le n^{\alpha}$ for some $\alpha \in (0,1)$, d being the dimension of the space. Moreover, the smoothness threshold $\frac{1}{1-\alpha}$ is sharp in the sense that for $s < \frac{1}{1-\alpha}$ the minimax optimal convergence rate is slower than $\sqrt{n}$. Our goal is to extend some of these results to more general high-dimensional models under an assumption that the model admits statistical estimators of unknown parameter for which normal approximation holds for large n and sufficiently high dimension of the parameter.

1.1. *Bias reduction.* Let $X^{(n)}$ be an observation sampled from a probability distribution $P_{\theta}^{(n)}$ in a measurable space $(S^{(n)}, \mathcal{A}^{(n)})$ with unknown parameter $\theta \in T$. A particular example of interest is $X^{(n)} = (X_1, \dots, X_n)$, where $X_1, \dots, X_n$ are i.i.d. observations in a measurable spaces $(S, \mathcal{A})$. It will be assumed in what follows that the parameter space T is an open subset of a separable Banach space E (which could be a high-dimensional or infinite-dimensional space). Let $\hat{\theta} = \hat{\theta}_n = \hat{\theta}(X^{(n)}) \in T$ be an estimator of θ based on the observation $X^{(n)}$. We will be especially interested in estimators $\hat{\theta}$ that could be approximated in distribution by a Gaussian random vector in E (whose distribution, of course, depends on unknown parameter $\theta \in T$ provided that $X^{(n)} \sim P_{\theta}^{(n)}$). More precisely, it will be assumed in what follows that, for all $\theta \in T$ (or in properly chosen subsets of T), $\sqrt{n}(\hat{\theta} - \theta)$ is close in distribution to a mean zero Gaussian random vector $\xi(\theta)$ in E. In Section 2, it will be described more precisely in which sense this approximation should hold.

Given a smooth functional $f : T \mapsto \mathbb{R}$, our main goal is to construct an estimator of $f(\theta)$ based on $X^{(n)}$. It is well known that in high-dimensional and infinite-dimensional models the plug-in estimator $f(\hat{\theta})$ is often sub-optimal even when the base estimator $\hat{\theta}$ is optimal. This is largely due to the fact that for non-linear functionals f the plug-in estimator $f(\hat{\theta})$ has a large bias even when $\hat{\theta}$ is unbiased, or has a small bias. Thus, the bias reduction becomes a crucial part of the design of estimators of $f(\theta)$ with optimal error rates. To construct an unbiased estimator of $f(\theta)$ (which is not always possible) one has to solve an integral equation $\mathcal{T}g = f$ for the following integral operator:

$$(\mathcal{T}g)(\theta) := \mathbb{E}_{\theta} g(\hat{\theta}) = \int_T g(t) P(\theta; dt), \theta \in T, \tag{1.1}$$

where

$$P(\theta; A) = \mathbb{P}_{\theta}\{\hat{\theta} \in A\}, A \subset T \tag{1.2}$$

is a Markov kernel on the parameter space T (the distribution of estimator $\hat{\theta}$). Recall that, by the definition of Markov kernel, it is assumed that $T \ni \theta \mapsto P(\theta; A)$ is a Borel measurable function for all Borel subsets $A \subset T$.

Note that $\mathcal{T}f$ is well defined for all functions $f \in L_{\infty}(T)$ and, moreover, $\mathcal{T} : L_{\infty}(T) \mapsto L_{\infty}(T)$ is a contraction. Most often, we will deal with operator $\mathcal{T}$ acting on uniformly bounded Lipschitz functions (or even on sufficiently smooth functions).

Finding an estimator of $f(\theta)$ with a small bias then reduces to an approximate solution of equation $\mathcal{T}g = f$. If $\mathcal{B} := \mathcal{T} - \mathcal{I}$ is a "small operator" (which is the case when the estimator $\hat{\theta}$ is "close" to θ with a high probability), then the solution of this equation could be written (at least, formally) as the sum of Neumann series

$$g = (\mathcal{I} + \mathcal{B})^{-1} f = f - \mathcal{B}f + \mathcal{B}^2 f - \mathcal{B}^3 f + \dots$$

and one can try to use the following function $f_k(\theta)$ (with a properly chosen k),

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

as an approximate solution of equation $\mathcal{T}g = f$. This yields an estimator $f_k(\hat{\theta})$ with a reduced bias

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

Another way to look at this bias reduction procedure is to observe that the bias of the plug-in estimator $f(\hat{\theta})$ is equal to

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

To reduce the bias of $f(\hat{\theta})$, one could subtract from it the plug-in estimator of the function $(\mathcal{B}f)(\theta)$ yielding the estimator $f_1(\hat{\theta}) = f(\hat{\theta}) - (\mathcal{B}f)(\hat{\theta})$. The bias of $f_1(\hat{\theta})$ is equal to $-(\mathcal{B}^2 f)(\theta)$. To further reduce the bias, we have to add its plug-in estimator $(\mathcal{B}^2 f)(\hat{\theta})$ yielding the estimator $f_2(\hat{\theta}) = f(\hat{\theta}) - (\mathcal{B}f)(\hat{\theta}) + (\mathcal{B}^2 f)(\hat{\theta})$, and so on.

This higher order bias reduction method has been studied in [36, 37, 40, 41] in the case of various high-dimensional normal models and in [33] in the case of the classical binomial model. In particular, the approach to the analysis of this method initiated in [36, 37] and further developed in [41] is based on the derivation of integral representation formulas for functions $(\mathcal{B}^k f)(\theta)$ in terms of so called smooth random homotopies. These formulas provide a way to obtain sharp bounds on the bias of estimator $f_k(\hat{\theta})$ and to establish smoothness properties of functions f_k needed to develop concentration inequalities for this estimator (see Section 4 for more details). However, the construction of random homotopies for a given estimator $\hat{\theta}$ relies on certain coupling techniques. In particular, it is based on the existence of a smooth stochastic process $G(\theta), \theta \in \Theta$ with values in Θ such that $G(\theta) \stackrel{d}{=} \hat{\theta}(X^{(n)}), X^{(n)} \sim P_\theta$. The bounds on the bias of estimator $f_k(\hat{\theta})$ obtained in [41] rely on the existence of such a coupling and the Hölder norms of process G are involved in these bounds. Such a coupling trivially exists in the case of random shift models [40, 42] and it is easy to construct in the case of general Gaussian models [41] as well as some other exponential transformation families. However, it is much harder to develop smooth random homotopies for MLE and other relevant estimators in the case of more general high-dimensional parametric models. A possible approach could rely on general coupling methods developed in the literature such as optimal transport maps and Moser's coupling (see, e.g., [63]). However, the bounds on Hölder norms for such coupling maps with explicit dependence on the dimension have not been developed in the literature and their development leads to difficult questions concerning smoothness of solutions of PDEs (in particular, Monge-Ampère and Poisson type equations) in high dimensions. Another serious difficulty is the need to develop tight concentration bounds for estimators $f_k(\hat{\theta})$ that are also not readily available for general high-dimensional models (with Gaussian, log-concave and some closely related models being exceptions). Due to these difficulties, the higher order bias reduction method described above has been so far fully studied only in the case of Gaussian models as well as some random shift models with Poincaré type noise [41].

In this paper, we study the problem under an additional assumption that the estimator $\hat{\theta}$ admits sufficiently accurate normal approximation. More precisely, we assume that $\sqrt{n}(\hat{\theta} - \theta)$ can be approximated in distribution by a Gaussian r.v. $\xi(\theta)$ in E. This assumption allows us to define an approximating Gaussian model, an "estimator" $\tilde{\theta} = \theta + \frac{\xi(\theta)}{\sqrt{n}}$ of parameter θ for

this model and the corresponding operators $\tilde{\mathcal{T}}, \tilde{\mathcal{B}}$ and functions $\tilde{f}_k, k \geq 0$. We show that functions $\tilde{f}_k$ provide a reasonable approximation of functions f_k and one can reduce the bounds on estimator $f_k(\hat{\theta})$ of $f(\theta)$ to the bounds on estimator $\tilde{f}_k(\tilde{\theta})$ in the corresponding approximating Gaussian model. This approach allows us to circumvent the difficulties with the direct analysis of estimator $f_k(\hat{\theta})$ since both the technique of random homotopies and concentration inequalities are applicable to the approximating model. As a result, we prove *"reduction theorems"* (stated in Section 2) showing that the risk bounds and normal approximation properties established earlier in the Gaussian case hold also for general models, provided that the normal approximation of estimator $\hat{\theta}$ is sufficiently accurate.

1.2. *Smoothness classes and distances between random variables.* Let F be a Banach space and let $U \subset E$. For a function $g : U \mapsto F$, denote

$$\|g\|_{L_\infty(U)} := \sup_{x \in U} \|g(x)\|, \quad \|g\|_{\mathrm{Lip}(U)} := \sup_{x,x' \in U, x \neq x'} \frac{\|g(x) - g(x')\|}{\|x - x'\|}$$

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

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

We will now introduce Hölder spaces $C^s(U;F)$ of functions of smoothness $s > 0$ from an open subset $U \subset E$ into a Banach space F (most often, either $F = \mathbb{R}$, or $F = E$). Given a function $g : U \mapsto F$, let $g^{(j)}$ denote its Fréchet derivative of order j (in particular, $g^{(0)} = g$). Note that, for all $x \in U$, $g^{(j)}(x)$ is a symmetric bounded j-linear form (with values in F). For such forms $M[u_1, \dots, u_j], u_1, \dots, u_j \in E$, we will use the operator norm

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

and $g^{(j)}$ will be always viewed as a mapping from U into the space of symmetric bounded j-linear forms equipped with the operator norm. Let $s = m + \rho$, $m \geq 0, \rho \in (0,1]$. For an m-times Fréchet differentiable function g from U into F, define

$$\|g\|_{C^s(U;F)} := \max\Big(\|g\|_{L_\infty(U)}, \max_{0 \leq j \leq m-1} \|g^{(j)}\|_{\mathrm{Lip}(U)}, \|g^{(m)}\|_{\mathrm{Lip}_\rho(U)}\Big).$$

The space $C^s(U,F)$ is then defined as the set of all m-times Fréchet differentiable functions g from U into F such that $\|g\|_{C^s(U,F)} < \infty$. When the space F is clear from the context (in particular, when $F = \mathbb{R}$), we will write simply $C^s(U)$ and $\|\cdot\|_{C^s(U)}$ instead of $C^s(U,F)$ and $\|\cdot\|_{C^s(U,F)}$.

REMARK 1.1. The definition of the space $C^s(U)$ used here is not quite standard. In particular, the space $C^1(U)$ consists of all uniformly bounded Lipschitz functions in U rather than continuously differentiable functions. Note also that, for a j times Fréchet differentiable function g, $\|g^{(j)}\|_{L_\infty(U)} \leq \|g^{(j-1)}\|_{\mathrm{Lip}(U)}$, with the equality holding when U is convex (which would lead to a more standard definition of Hölder norms).

We will also use the following notation. Let $s = m + \rho$, $m \geq 0, \rho \in (0,1]$. For $l = 0, \dots, m$, denote

$$\|g\|_{C^{l,s}(U;F)} := \max\Big(\max_{l \leq j \leq m-1} \|g^{(j)}\|_{\mathrm{Lip}(U)}, \|g^{(m)}\|_{\mathrm{Lip}_\rho(U)}\Big).$$

and let $C^{l,s}(U;F)$ be the set of all m-times Fréchet differentiable functions g from U into F such that $\|g\|_{C^{l,s}(U,F)} < \infty$. In particular, $\|\cdot\|_{C^{0,1}(U)} = \|\cdot\|_{\mathrm{Lip}(U)}$.

REMARK 1.2. Note that, by McShane-Whitney extension theorem, any Lipschitz function $g : U \mapsto \mathbb{R}$ could be extended to a Lipschitz function defined on the whole space E with preservation of its Lipschitz norm $\|g\|_{\mathrm{Lip}(U)}$ (in fact, this theorem applies to general metric spaces, not just to Banach spaces). Moreover, any function $g \in C^1(U)$ (a uniformly bounded Lipschitz function) could be extended to the whole space E with preservation of its C^1-norm. In what follows, it will be convenient to assume that bounded Lipschitz functions (in particular, functions from the space $C^s(U)$ for $s \geq 1$) and Lipschitz functions (in particular, functions from the space $C^{0,s}(U)$ for $s \geq 1$) are indeed extended to the whole space this way. Similarly, any function from space $C^s(U), U \subset E, s \in (0,1]$ could be extended to the whole space E with preservation of its norm (again by the application of McShane-Whitney extension theorem to the metric space (E,d), $d(x,y) := \|x-y\|^s, x, y \in E$). Note that the problem of extension of smooth functions (from space $C^s(U)$ with $s > 1$) to the whole space with preservation of the norm is much more complicated and such extensions do not always exist in general Banach spaces.

We will need to quantify the accuracy of normal approximation for random variable $\sqrt{n}(\hat{\theta} - \theta)$ by $\xi(\theta)$ (as well as for other random variables), and, for this purpose, we will introduce below certain distances between distributions of random variables.

Let η_1, η_2 be random variables defined on a probability space $(\Omega, \Sigma, \mathbb{P})$ with values in a measurable space $(S, \mathcal{A})$, and let $\mathcal{F}$ be a set of measurable functions on S. Define

$$\Delta_{\mathcal{F}}(\eta_1, \eta_2) := \sup_{f \in \mathcal{F}} |\mathbb{E}f(\eta_1) - \mathbb{E}f(\eta_2)|.$$

REMARK 1.3. Note that, in fact, $\Delta_{\mathcal{F}}(\eta_1, \eta_2)$ is a distance between the laws $\mathcal{L}(\eta_1), \mathcal{L}(\eta_2)$ of random variables η_1, η_2 (so, it does not matter whether η_1, η_2 are defined on the same probability space or not; however, it is always possible to assume that they are and it will be convenient for our purposes).

Let now $\psi : \mathbb{R} \mapsto \mathbb{R}_+$ be an even convex function with $\psi(0) = 0$ and such that ψ is increasing in $\mathbb{R}_+$. The Orlicz ψ-norm of real valued r.v. ζ is defined as

$$\|\zeta\|_{\psi} := \inf\Big\{c > 0 : \mathbb{E}\psi\Big(\frac{|\zeta|}{c}\Big) \leq 1\Big\}.$$

Denote $L_{\psi}(\mathbb{P}) := \{\zeta : \|\zeta\|_{\psi} < +\infty\}$. We will also write $\|\zeta\|_{L_{\psi}(\mathbb{P})} = \|\zeta\|_{\psi}$ (to emphasize the dependence of the Orlicz norm on the underlying probability measure $\mathbb{P}$). If $\psi(u) := |u|^p, u \in \mathbb{R}, p \geq 1$, then $\|\cdot\|_{\psi} = \|\cdot\|_{L_p}$ and $L_{\psi}(\mathbb{P}) = L_p(\mathbb{P})$. Other common choices of ψ are $\psi_1(u) = e^{|u|} - 1$ (subexponential Orlicz norm) and $\psi_2(u) = e^{u^2} - 1$ (subgaussian Orlicz norm).

We will need another distance between random variables η_1, η_2 in a space $(S, \mathcal{A})$ defined as follows:

$$\Delta_{\mathcal{F},\psi}(\eta_1, \eta_2) := \sup_{f \in \mathcal{F}} |\|f(\eta_1)\|_{\psi} - \|f(\eta_2)\|_{\psi}|.$$

If (S, d) is a metric space, one can also define the following Wasserstein type distance:

$$W_{\psi}(\eta_1, \eta_2) := \inf\Big\{\|d(\eta_1', \eta_2')\|_{\psi} : \eta_1' \stackrel{d}{=} \eta_1, \eta_2' \stackrel{d}{=} \eta_2\Big\},$$

where the infimum is taken over all random variables η_1', η_2' on $(\Omega, \Sigma, \mathbb{P})$ such that η_1' has the same distribution as η_1 and η_2' has the same distribution as η_2. If $\psi(u) = |u|^p$ this becomes

a usual definition of the Wasserstein p-distance W_p. For this choice of ψ, we also use the notation $\Delta_{\mathcal{F},p}$ instead of $\Delta_{\mathcal{F},\psi}$.

We will also use the notations $\Delta_{\mathcal{F},\mathbb{P}}(\eta_1,\eta_2)$, $\Delta_{\mathcal{F},\psi,\mathbb{P}}(\eta_1,\eta_2)$ and $W_{\psi,\mathbb{P}}(\eta_1,\eta_2)$ whenever it is needed to emphasize the dependence of these distances on $\mathbb{P}$.

Since, for $\eta_1' \stackrel{d}{=} \eta_1, \eta_2' \stackrel{d}{=} \eta_2$,

$$|\|f(\eta_1)\|_\psi - \|f(\eta_2)\|_\psi| = |\|f(\eta_1')\|_\psi - \|f(\eta_2')\|_\psi| \le \|f(\eta_1') - f(\eta_2')\|_\psi,$$

we could conclude that

$$\Delta_{\mathcal{F},\psi}(\eta_1,\eta_2) \le \sup_{f\in\mathcal{F}} W_\psi(f(\eta_1),f(\eta_2)) \tag{1.3}$$

(for r.v. η_1,η_2 with values in an arbitrary measurable space S). If (S,d) is a complete separable metric space and $\mathcal{F}$ is the set of all contractions (Lipschitz functions on S with constant 1), then, for all $f\in\mathcal{F}$, $|f(\eta_1)-f(\eta_2)| \le d(\eta_1,\eta_2)$ implying that

$$\Delta_{\mathcal{F},\psi}(\eta_1,\eta_2) \le \sup_{f\in\mathcal{F}} W_\psi(f(\eta_1),f(\eta_2)) \le W_\psi(\eta_1,\eta_2). \tag{1.4}$$

Note also that for $\psi(u)=|u|$ (the L_1-norm), we have

$$\Delta_{\mathcal{F},1}(\eta_1,\eta_2) \le W_1(\eta_1,\eta_2) = \Delta_{\mathcal{F}}(\eta_1,\eta_2),$$

where $\mathcal{F}$ is the set of all real valued contractions on S (follows from Kantorovich-Rubinstein duality).

Most often, we will deal with random variables in a Banach space F (in particular, $F=E$ and $F=\mathbb{R}$) and the set $\mathcal{F}$ will usually be a Hölder ball of certain smoothness, such as $\mathcal{F} = \{f : \|f\|_{C^s(E)} \le 1\}$ for $s>0$, or $\mathcal{F} := \{f : \|f\|_{C^s(U)} \le 1\}$, or $\mathcal{F} := \{f : \|f\|_{C^{l,s}(U)} \le 1\}$ for some $0 \le l < s$ and for $U \subset F$. In particular, we will use the notations

$$\Delta_s(\eta_1,\eta_2) = \Delta_{\mathcal{F}}(\eta_1,\eta_2) \text{ and } \Delta_{s,\psi}(\eta_1,\eta_2) = \Delta_{\mathcal{F},\psi}(\eta_1,\eta_2)$$

for $\mathcal{F} = \{f : \|f\|_{C^s(E)} \le 1\}$.

Other distances that will be used in the future include:

- Kolmogorov's distance between random variables η_1,η_2 in $\mathbb{R}$ (more precisely, between their laws $\mathcal{L}(\eta_1),\mathcal{L}(\eta_2)$) defined as

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

  where $\mathcal{F} := \{I_{(-\infty,x]} : x\in\mathbb{R}\}$.
- For $s = k+\rho,\ k\ge 0, \rho\in(0,1]$ and random variables η_1,η_2 in a Banach space E, let

$$\zeta_s(\eta_1,\eta_2) := \sup_{\|f^{(k)}\|_{\mathrm{Lip}_\rho(E)}\le 1} |\mathbb{E}f(\eta_1) - \mathbb{E}f(\eta_2)| = \Delta_{\mathcal{F}}(\eta_1,\eta_2), \tag{1.5}$$

  where $\mathcal{F} := \{f : \|f^{(k)}\|_{\mathrm{Lip}_\rho(E)} \le 1\}$. Note that, for $s=1$, $\zeta_1(\eta_1,\eta_2) = W_1(\eta_1,\eta_2)$.

Finally, in a statistical framework, we have to deal with a family of probability measures $\mathbb{P}_\theta, \theta\in\Theta$ (that generates different distributions of the data) and we will use uniform versions of the distances defined above:

$$\Delta_{\mathcal{F},\Theta}(\eta_1,\eta_2) := \sup_{\theta\in\Theta} \Delta_{\mathcal{F},\mathbb{P}_\theta}(\eta_1,\eta_2),$$

$$\Delta_{\mathcal{F},\psi,\Theta}(\eta_1,\eta_2) := \sup_{\theta\in\Theta} \Delta_{\mathcal{F},\psi,\mathbb{P}_\theta}(\eta_1,\eta_2) \ \text{ and } \ W_{\psi,\Theta}(\eta_1,\eta_2) := \sup_{\theta\in\Theta} W_{\psi,\mathbb{P}_\theta}(\eta_1,\eta_2).$$

We will also use the distance

$$\Delta^{+}_{\mathcal{F},\psi,\Theta}(\eta_1,\eta_2) := \Delta_{\mathcal{F},\Theta}(\eta_1,\eta_2) + \Delta_{\mathcal{F},\psi,\Theta}(\eta_1,\eta_2).$$

Throughout the paper, the following notations will be used. For real non-negative variables A, B, $A \lesssim B$ means that there exists a universal constant $C > 0$ such that $A \le CB$, $A \gtrsim B$ means that $B \lesssim A$ and $A \asymp B$ means that $A \lesssim B$ and $B \lesssim A$. If constant C in the above inequalities depends on additional parameters, the corresponding relationships will be provided with subscripts: say, $A \lesssim_{s,\psi} B$ means that $A \le CB$ with $C = C_{s,\psi} > 0$ depending on s and ψ.

2. Main results: bounds on L_ψ-errors and normal approximation of $f_k(\hat\theta)$. In this section, we study the error rates of estimator $f_k(\hat\theta)$ depending on the smoothness of functional f and show that they coincide with the rates known to be optimal in the Gaussian case provided that normal approximation error for $\hat\theta$ is negligible.

In [40], the following Gaussian shift model $X^{(n)} = \theta + \frac{\xi}{\sqrt{n}}, \theta \in E$ was studied, where ξ is a Gaussian r.v. in E with mean zero and covariance operator Σ. It was assumed that θ is an unknown parameter and Σ is known and the goal is to estimate $f(\theta)$ for a given smooth functional f. The complexity of this estimation problem could be characterized by two parameters: the "weak variance" of the noise ξ, $\|\Sigma\| = \sup_{\|u\|\le 1}\mathbb{E}\langle \xi, u\rangle^2$, and its "strong variance" $\mathbb{E}\|\xi\|^2 = \mathbb{E}\sup_{\|u\|\le 1}\langle \xi, u\rangle^2$. Note that, in the case of Euclidean space $E = \mathbb{R}^d$ and $\xi \sim N(0, \sigma^2 I_d)$, $\|\Sigma\| = \sigma^2$ and $\mathbb{E}\|\xi\|^2 = \sigma^2 d$.

The following result was proved.

THEOREM 2.1. *Let $s > 0$. For $s \in (0,1]$, set $k := 0$ and for $s > 1$, let $s = k + 1 + \rho$ for some $k \ge 0$ and $\rho \in (0,1]$. Let $\hat\theta = \hat\theta(X^{(n)}) = X^{(n)}$. Then*

$$\sup_{\|f\|_{C^s(E)}\le 1}\sup_{\theta\in E}\|f_k(\hat\theta) - f(\theta)\|_{L_2(\mathbb{P}_\theta)} \lesssim_s \Big(\frac{\|\Sigma\|^{1/2}}{n^{1/2}} \bigvee \Big(\sqrt{\frac{\mathbb{E}\|\xi\|^2}{n}}\Big)^s\Big) \bigwedge 1.$$

Note that the term $\frac{\|\Sigma\|^{1/2}}{\sqrt{n}}$ of the error bound of Theorem 2.1 controls the concentration of estimator $f_k(\hat\theta)$ around its expectation whereas the term $\Big(\sqrt{\frac{\mathbb{E}\|\xi\|^2}{n}}\Big)^s$ controls the bias of this estimator. Moreover, it was also shown in [40] that, for $E = \mathbb{R}^d$ equipped with the standard Euclidean norm and $\xi \sim N(0, \sigma^2 I_d)$,

$$\sup_{\|f\|_{C^s(\mathbb{R}^d)}\le 1}\inf_T\sup_{\theta\in\mathbb{R}^d}\|T(X^{(n)}) - f(\theta)\|_{L_2(\mathbb{P}_\theta)} \asymp \Big(\frac{\|\Sigma\|^{1/2}}{n^{1/2}} \bigvee \Big(\sqrt{\frac{\mathbb{E}\|\xi\|^2}{n}}\Big)^s\Big) \bigwedge 1,$$

where the infimum is taken over all estimators $T(X^{(n)})$, implying the minimax optimality of the L_2 error rates in the case of Gaussian shift model in the Euclidean space $E = \mathbb{R}^d$.

Note also that the convergence rate is of the order $O(n^{-1/2})$ if $\|\Sigma\| \lesssim 1$, $\mathbb{E}\|\xi\|^2 \lesssim n^\alpha$ for $\alpha \in (0,1)$ and $s \ge \frac{1}{1-\alpha}$ and it is slower than $n^{-1/2}$ if $s < \frac{1}{1-\alpha}$. For $s > \frac{1}{1-\alpha}$, it was proved in [40] that $\sqrt{n}(f_k(\hat\theta) - f(\theta))$ could be approximated in distribution by $\sigma_f(\theta)Z, Z \sim N(0,1)$ as $n \to \infty$, where $\sigma_f^2(\theta) := \langle \Sigma f'(\theta), f'(\theta)\rangle$, and, moreover, it was shown that $f_k(\hat\theta)$ is an asymptotically efficient estimator.

We will try to extend some of these results to general models and general estimators $\hat\theta$ for which Gaussian approximation holds.

To describe Gaussian approximation properties for estimator $\hat{\theta}$ more precisely, let

$$G(\theta) := \theta + \frac{\xi(\theta)}{\sqrt{n}}, \theta \in T,$$

where $\xi : T \mapsto E$ is a Gaussian stochastic process. In what follows, $\tilde{\theta} := G(\theta), \theta \in T$ will be viewed as a Gaussian approximation of estimator $\hat{\theta}$. In other words, the estimator $\hat{\theta}$ in the initial model is approximated by the "estimator" $\tilde{\theta}$ in a Gaussian shift model with unknown parameter θ and small Gaussian noise $\frac{\xi(\theta)}{\sqrt{n}}$. For simplicity, we also assume that $\mathbb{E}\xi(\theta) = 0, \theta \in T$ and let $\Sigma(\theta)$ denote the covariance operator of random variable $\xi(\theta)$. As a typical example, consider the case when $E := \mathbb{R}^d$ and $\xi(\theta) := A(\theta)Z, Z \sim N(0, I_d)$, where $A(\theta) : \mathbb{R}^d \mapsto \mathbb{R}^d$ is a bounded linear operator. Even more specifically, when $X^{(n)} = (X_1, \dots, X_n)$, $X_1, \dots, X_n$ being i.i.d. $\sim P_\theta, \theta \in T \subset \mathbb{R}^d$, one can think of the maximum likelihood estimator $\hat{\theta}$ and $A(\theta) = I(\theta)^{-1/2}$, where $I(\theta)$ is the Fisher information matrix (since in the case of regular statistical models $\sqrt{n}(\hat{\theta} - \theta)$ is close in distribution to $I(\theta)^{-1/2}Z$).

In the results stated below, the L_ψ-error of estimator $f_k(\hat{\theta})$ and its normal approximation will be controlled uniformly in a subset Θ of parameter space T. It will be assumed that $\xi(\theta)$ is bounded or even sufficiently smooth in a small neighborhood

$$\Theta_\delta := \{\theta \in E : \mathrm{dist}(\theta; \Theta) < \delta\} \subset T$$

of set Θ for some $\delta > 0$, and, moreover, that the normal approximation of $\hat{\theta}$ by $\tilde{\theta}$, or of $\sqrt{n}(\hat{\theta} - \theta)$ by $\xi(\theta)$ holds in proper distances uniformly in $\theta \in \Theta_\delta$. The behavior of the process $\xi(\theta)$ outside of Θ_δ will be of no importance for us, and, without loss of generality, we can and will set $\xi(\theta) := 0, \theta \in E \setminus \Theta_\delta$. With this definition, we still have that $\|\xi\|_{L_\infty(E)} = \|\xi\|_{L_\infty(\Theta_\delta)}$.

In what follows, we will deal with loss functions $\psi : \mathbb{R} \mapsto \mathbb{R}_+$. It will be assumed that ψ is convex with $\psi(0) = 0$. Moreover, ψ is even, increasing on $\mathbb{R}_+$ and satisfies the condition

$$c'u \le \psi(u) \le c''\psi_1(u), u \ge 0$$

with some constants $c', c'' > 0$, where $\psi_1(u) = e^u - 1$. Let Ψ be the set of such loss functions. Given $\psi \in \Psi$, denote

$$\tilde{\psi}(u) := \frac{1}{\psi^{-1}\left(\frac{1}{u}\right)}, u \ge 0.$$

For instance, in the case of $\psi(u) = |u|^p, p \ge 1$, we have $\tilde{\psi}(u) = u^{1/p}, u \ge 0$, and in the case of $\psi(u) = \psi_1(u) = e^{|u|} - 1$, we have $\tilde{\psi}(u) = \frac{1}{\log(1+\frac{1}{u})}, u \ge 0$.

For $\psi \in \Psi$, we will study Orlicz norm error rates $\|f_k(\hat{\theta}) - f(\theta)\|_{L_\psi(\mathbb{P}_\theta)}$ of estimator $f_k(\hat{\theta})$ depending on the smoothness of functional f. We will also study normal approximation of r.v. $\sqrt{n}(f_k(\hat{\theta}) - f(\theta))$. The choice of k depends on the degree of smoothness of functional f. Namely, if f is C^s-smooth with $s = k + 1 + \rho$, $k \ge 0, \rho \in (0, 1]$, we will use estimator $f_k(\hat{\theta})$. Note that, for $k = 0$, we have $f_0 = f$ and one can use a standard plug-in estimator $f(\hat{\theta})$ for all $s \in (0, 2]$. First, we will state the results in this simple case.

Given $\Theta \subset T$, denote $\mathfrak{v}_\xi(\Theta) := \sup_{\theta \in \Theta} \mathbb{E}\|\xi(\theta)\|^2$.

THEOREM 2.2. *Let $\Theta \subset T$ be an open subset and let $\psi \in \Psi$. The following statements hold:*
(i) For all $s \in (0, 1]$,

$$\sup_{\|f\|_{C^s(E)} \le 1} \sup_{\theta \in \Theta} \|f(\hat{\theta}) - f(\theta)\|_{L_\psi(\mathbb{P}_\theta)} \lesssim_{s,\psi} \left(\sqrt{\frac{\mathfrak{v}_\xi(\Theta)}{n}}\right)^s + \Delta_{\mathcal{H},\psi,\Theta}(\hat{\theta}, \tilde{\theta}), \tag{2.1}$$

where $\mathcal{H} := \{g : \|g\|_{C^s(E)} \le 1\}$.
(ii) Let $\delta > 0$ *be such that* $\Theta_\delta \subset T$. *For* $s = 1 + \rho$ *with* $\rho \in (0,1]$, *there exists a constant* $c_s \in (0,1)$ *such that*

$$\sup_{\|f\|_{C^s(\Theta_\delta)} \le 1} \sup_{\theta \in \Theta} \|f(\hat\theta) - f(\theta)\|_{L_\psi(\mathbb{P}_\theta)} \lesssim_{s,\psi} \left[\frac{\|\Sigma\|^{1/2}_{L_\infty(\Theta)}}{n^{1/2}} + \left(\sqrt{\frac{\mathfrak{v}_\xi(\Theta)}{n}}\right)^s + \Delta^+_{\mathcal{H},\psi,\Theta_\delta}(\hat\theta, \tilde\theta) \right.$$

$$\left. + \sup_{\theta \in \Theta} \tilde\psi^{1/2}\Big(\mathbb{P}\{\|\xi(\theta)\| \ge c_s \delta \sqrt{n}\}\Big) \right] \bigwedge 1, \tag{2.2}$$

where $\mathcal{H} := \{g : \|g\|_{C^s(\Theta_{c_s\delta})} \le 1\}$.

REMARK 2.1. For $\Theta = T = E$, bound (2.2) of Theorem 2.2 simplifies as follows:

$$\sup_{\|f\|_{C^s(E)} \le 1} \sup_{\theta \in E} \|f(\hat\theta) - f(\theta)\|_{L_\psi(\mathbb{P}_\theta)} \lesssim_{s,\psi} \left[\frac{\|\Sigma\|^{1/2}_{L_\infty(E)}}{n^{1/2}} + \left(\sqrt{\frac{\mathfrak{v}_\xi(E)}{n}}\right)^s + \Delta^+_{\mathcal{H},\psi,E}(\hat\theta, \tilde\theta) \right] \bigwedge 1,$$

where $\mathcal{H} := \{g : \|g\|_{C^s(E)} \le 1\}$.

In the general case, there are additional terms in the bounds depending on tail probabilities of $\|\xi(\theta)\|$. Note that under the assumption that $\mathfrak{v}_\xi(E) \le c_1' \delta^2 n$ for small enough constant $c_1' > 0$, it easily follows from the Gaussian concentration inequality that

$$\mathbb{P}\{\|\xi(\theta)\| \ge c_s \delta \sqrt{n}\} \le \exp\Big\{ - \frac{c_1'' \delta^2 n}{\|\Sigma\|_{L_\infty(\Theta)}} \Big\}, \theta \in \Theta.$$

Since for $\psi \in \Psi$, $\psi(u) \lesssim \psi_1(u), u \ge 0$ and $\tilde\psi_1(u) = \frac{1}{\log(1+\frac{1}{u})}$, it is easy to conclude that

$$\sup_{\theta \in \Theta} \tilde\psi^{1/2}\Big(\mathbb{P}\{\|\xi(\theta)\| \ge c_s \delta \sqrt{n}\}\Big) \lesssim \frac{1}{\delta} \frac{\|\Sigma\|^{1/2}_{L_\infty(\Theta)}}{n^{1/2}}.$$

Thus, in the worst case, the additional term in bound (2.2) is of the same order (up to a factor $\frac{1}{\delta}$) as the term $\frac{\|\Sigma\|^{1/2}_{L_\infty(\Theta)}}{n^{1/2}}$ present in the optimal bounds in the Gaussian case. For slower growing losses, this additional term becomes negligible. For instance, for the loss $\psi(u) = u^p, u > 0, p \ge 1$, it is dominated by $\exp\Big\{ - \frac{c''}{p} \frac{\delta^2 n}{\|\Sigma\|_{L_\infty(\Theta)}} \Big\}$ for some constant $c'' > 0$, so, it decays exponentially fast as $n \to \infty$. Note that constants c_1', c_1'', c'' might depend on s.

The next result provides bounds on normal approximation of the error $f(\hat\theta) - f(\theta)$ for functionals f of smoothness $s \in (1,2]$. Recall that for a Fréchet differentiable functional f,

$$\sigma_f^2(\theta) := \langle \Sigma(\theta) f'(\theta), f'(\theta) \rangle.$$

THEOREM 2.3. *Let* $s = 1 + \rho$ *with* $\rho \in (0,1]$ *and let* $\delta > 0$. *Let* Θ *be a subset of* E *such that* $\Theta_\delta \subset T$. *Suppose, for some sufficiently small constant* $c_1 > 0$, $\mathfrak{v}_\xi(\Theta) \le c_1 n$. *Then, for some constant* $c_s \in (0,1)$, *the following bounds hold.*

(i) For all $\psi \in \Psi$,

$$\sup_{\|f\|_{C^s(\Theta_\delta)} \le 1} \sup_{\theta \in \Theta} \Big| \|f(\hat\theta) - f(\theta)\|_{L_\psi(\mathbb{P}_\theta)} - n^{-1/2} \sigma_f(\theta) \|Z\|_{L_\psi(\mathbb{P})} \Big|$$

$$\lesssim_{s,\psi} \left(\sqrt{\frac{\mathfrak{v}_\xi(\Theta)}{n}}\right)^s + \Delta^+_{\mathcal{H},\psi,\Theta_\delta}(\hat\theta, \tilde\theta) + \sup_{\theta \in \Theta} \tilde\psi^{1/2}\Big(\mathbb{P}\{\|\xi(\theta)\| \ge c_s \delta \sqrt{n}\}\Big), \tag{2.3}$$

where $\mathcal{H} := \{g : \|g\|_{C^s(\Theta_{c_s\delta})} \le 1\}$.

(ii) For all $s' \in [1, s]$, [1]

$$\sup_{\|f\|_{C^s(\Theta_\delta)}\le 1}\sup_{\theta\in\Theta}\Delta_{s'}(\sqrt{n}(f(\hat\theta)-f(\theta)),\sigma_f(\theta)Z)$$

(2.4)

$$\lesssim_s \left[\sqrt{n}\Big(\sqrt{\frac{\mathfrak{v}_\xi(\Theta)}{n}}\Big)^s + \Delta_{\mathcal{F},\Theta_\delta}(\sqrt{n}(\hat\theta-\theta),\xi(\theta)) + \sqrt{n}\sup_{\theta\in\Theta}\mathbb{P}^{1/4}\{\|\xi(\theta)\|\ge c_s\delta\sqrt{n}\}\right],$$

where $\mathcal{F} := \{g : \|g\|_{C^{0,s'}(U_{c_s\delta\sqrt{n}})} \le 1\}$.

REMARK 2.2. For $\Theta = T = E$, under condition $\mathfrak{v}_\xi(E) \le c_1 n$ for a small enough constant $c_1 > 0$, the bounds of Theorem 2.3 simplify as follows:

$$\sup_{\|f\|_{C^s(E)}\le 1}\sup_{\theta\in E}\Big|\|f(\hat\theta)-f(\theta)\|_{L_\psi(\mathbb{P}_\theta)} - n^{-1/2}\sigma_f(\theta)\|Z\|_{L_\psi(\mathbb{P})}\Big|$$

$$\lesssim_{s,\psi} \Big(\sqrt{\frac{\mathfrak{v}_\xi(E)}{n}}\Big)^s + \Delta^+_{\mathcal{H},\psi,E}(\hat\theta,\tilde\theta)$$

with $\mathcal{H} := \{h : \|h\|_{C^s(E)} \le 1\}$, and

$$\sup_{\|f\|_{C^s(E)}\le 1}\sup_{\theta\in E}\Delta_{s'}(\sqrt{n}(f(\hat\theta)-f(\theta)),\sigma_f(\theta)Z)$$

$$\lesssim_s \left[\sqrt{n}\Big(\sqrt{\frac{\mathfrak{v}_\xi(E)}{n}}\Big)^s + \Delta_{\mathcal{F},E}(\sqrt{n}(\hat\theta-\theta),\xi(\theta))\right],$$

where $\mathcal{F} := \{g : \|g\|_{C^{0,s'}(E)} \le 1\}$.

The problem becomes much more difficult in the case when $s = k + 1 + \rho > 2$ ($k \ge 1, \rho \in (0, 1]$). In this case, $f_k(\hat\theta)$ is no longer a standard plug-in estimator and a non-trivial analysis of its bias is needed (see Section 4). This analysis requires some smoothness assumptions on the Gaussian stochastic process $\xi(\theta)$. Namely, instead of quantity $\mathfrak{v}_\xi(\Theta)$, we will use such quantities as

$$\mathfrak{d}_\xi(\Theta; s) := \mathbb{E}\|\xi\|^2_{C^s(\Theta)}.$$

Note that, if $\mathfrak{d}_\xi(\Theta; s) < \infty$, then, for all $p \ge 1$, $\mathbb{E}^{1/p}\|\xi\|^p_{C^s(\Theta)} \lesssim_p \sqrt{\mathfrak{d}_\xi(\Theta; s)}$, which easily follows from Gaussian concentration. Note also that, if $\xi(\theta) = A(\theta)Z$, where Z is a given Gaussian vector in E and $\Theta \ni \theta \mapsto A(\theta) \in L(E)$ is a C^s function with values in the space $L(E)$ of bounded linear operators in E, then $\mathfrak{d}_\xi(\Theta; s) \le \|A\|^2_{C^s(\Theta)}\mathbb{E}\|Z\|^2$. In particular, if $E = \mathbb{R}^d$ (equipped with the Euclidean norm) and $Z \sim N(0, I_d)$, then

$$\mathfrak{d}_\xi(\Theta; s) \le \|A\|^2_{C^s(\Theta)} d. \tag{2.5}$$

In such cases, the conditions in terms of $\mathfrak{d}_\xi(\Theta; s)$ can be reduced to smoothness assumptions on the "scaling" operator $A(\theta)$ (which is related to regularity properties of covariance $\Sigma(\theta)$ as a function of θ).

If $\Theta = T = E$, we will use the notation $\mathfrak{d}_\xi(s) := \mathfrak{d}_\xi(E; s)$. In what follows, such quantities will be used as complexity parameters in our problem.

[1] Here and in what follows, $U_r := \{x \in E : \|x\| < r\}$.

We are now ready to state the main results of the paper. The next theorem provides a bound on the L_ψ-error of estimator $f_k(\hat\theta)$ for $\psi \in \Psi$. Recall that it is assumed that $\xi(\theta) = 0$ outside of the neighborhood Θ_δ specified in the theorems.

THEOREM 2.4. *Let $\Theta \subset T$, let $\delta > 0$ and let $s = k + 1 + \rho$ with $k \geq 1$ and $\rho \in (0, 1]$. Suppose that $\Theta_\delta \subset T$. Then, for all $\psi \in \Psi$ and for some constant $c_s \in (0, 1)$,*

$$\begin{aligned}
&\sup_{\|f\|_{C^s(\Theta_\delta)} \leq 1} \sup_{\theta \in \Theta} \|f_k(\hat\theta) - f(\theta)\|_{L_\psi(\mathbb{P}_\theta)} \\
&\lesssim_s \left[\frac{\|\Sigma\|^{1/2}_{L_\infty(E)}}{n^{1/2}} + \left(\sqrt{\frac{\mathfrak{d}_\xi(\Theta_\delta; s-1)}{n}} \right)^s + \Delta^+_{\mathcal{H},\psi,\Theta_\delta}(\hat\theta, \tilde\theta) \right. \\
&\left. + \sup_{\theta \in \Theta_\delta} \tilde\psi\Big(\mathbb{P}_\theta\{\|\hat\theta - \theta\| \geq c_s\delta\} \Big) + \tilde\psi^{1/2}\Big(\mathbb{P}\{\|\xi\|_{L_\infty(E)} \geq c_s\delta\sqrt{n}\} \Big) \right] \bigwedge 1, \qquad (2.6)
\end{aligned}$$

where $\mathcal{H} := \{g : \|g\|_{C^s(\Theta_{c_s\delta})} \leq 1\}$.

Next we study normal approximation of estimator $f_k(\hat\theta)$.

THEOREM 2.5. *Let $s = k + 1 + \rho$ with $k \geq 1$ and $\rho \in (0, 1]$, and let $\delta > 0$. Let Θ be a subset of T such that $\Theta_\delta \subset T$. Suppose that, for some sufficiently small constant $c_1 > 0$,*

$$\mathfrak{d}_\xi(\Theta_\delta; s) \leq c_1 n.$$

Then, the following statements hold.

(i) For all $\psi \in \Psi$ and some constant $c_s \in (0, 1)$,

$$\begin{aligned}
&\sup_{\|f\|_{C^s(\Theta_\delta)} \leq 1} \sup_{\theta \in \Theta} \Big| \|f_k(\hat\theta) - f(\theta)\|_{L_\psi(\mathbb{P}_\theta)} - n^{-1/2}\sigma_f(\theta)\|Z\|_{L_\psi(\mathbb{P})} \Big| \\
&\lesssim_{s,\psi} \left(\sqrt{\frac{\mathfrak{d}_\xi(\Theta_\delta; s-1)}{n}} \right)^s + \frac{\|\Sigma\|^{1/2}_{L_\infty(E)}}{n^{1/2}} \sqrt{\frac{\mathfrak{d}_\xi(\Theta_\delta; s-1)}{n}} + \Delta^+_{\mathcal{H},\psi,\Theta_\delta}(\hat\theta, \tilde\theta) \\
&+ \sup_{\theta \in \Theta_\delta} \tilde\psi\Big(\mathbb{P}_\theta\{\|\hat\theta - \theta\| \geq c_s\delta\} \Big) + \tilde\psi^{1/2}\Big(\mathbb{P}\{\|\xi\|_{L_\infty(E)} \geq c_s\delta\sqrt{n}\} \Big), \qquad (2.7)
\end{aligned}$$

where $\mathcal{H} := \{g : \|g\|_{C^s(\Theta_{c_s\delta})} \leq 1\}$.

(ii) For all $s' \in [1, s]$ and some constant $c_s \in (0, 1)$,

$$\begin{aligned}
&\sup_{\|f\|_{C^s(\Theta_\delta)} \leq 1} \sup_{\theta \in \Theta} \Delta_{s'}(\sqrt{n}(f_k(\hat\theta) - f(\theta)), \sigma_f(\theta)Z) \\
&\lesssim_s \left[\sqrt{n} \left(\sqrt{\frac{\mathfrak{d}_\xi(\Theta_\delta; s-1)}{n}} \right)^s + \|\Sigma\|^{1/2}_{L_\infty(E)} \sqrt{\frac{\mathfrak{d}_\xi(\Theta_\delta; s-1)}{n}} + \Delta_{\mathcal{F},\Theta_\delta}(\sqrt{n}(\hat\theta - \theta), \xi(\theta)) \right. \\
&\left. + \sqrt{n} \sup_{\theta \in \Theta_\delta} \mathbb{P}_\theta\{\|\hat\theta - \theta\| \geq c_s\delta\} + \sqrt{n}\mathbb{P}^{1/4}\{\|\xi\|_{L_\infty(E)} \geq c_s\delta\sqrt{n}\} \right], \qquad (2.8)
\end{aligned}$$

where $\mathcal{F} := \{g : \|g\|_{C^{0,s'}(U_{c_s\delta\sqrt{n}})} \leq 1\}$.

REMARK 2.3. For $\Theta = T = E$, the bounds of Theorem 2.4 simplify as follows:

$$\sup_{\|f\|_{C^s(E)} \leq 1} \sup_{\theta \in E} \|f_k(\hat\theta) - f(\theta)\|_{L_\psi(\mathbb{P}_\theta)}$$

$$\lesssim_{s,\psi} \left[\frac{\|\Sigma\|^{1/2}_{L_\infty(E)}}{n^{1/2}} + \left(\sqrt{\frac{\mathfrak{d}_\xi(s-1)}{n}}\right)^s + \Delta^+_{\mathcal{H},\psi,E}(\hat\theta,\tilde\theta)\right] \wedge 1, \tag{2.9}$$

where $\mathcal{H} := \{g : \|g\|_{C^s(E)} \le 1\}$.

Assume that, for some sufficiently small constant $c_1 > 0$, $\mathfrak{d}_\xi(s) \le c_1 n$. Then, for all $\psi \in \Psi$, the following versions of bounds of Theorem 2.5 hold:

$$\sup_{\|f\|_{C^s(E)}\le 1}\sup_{\theta\in E}\Big|\|f_k(\hat\theta) - f(\theta)\|_{L_\psi(\mathbb{P}_\theta)} - n^{-1/2}\sigma_f(\theta)\|Z\|_{L_\psi(\mathbb{P})}\Big|$$

$$\lesssim_{s,\psi} \left(\sqrt{\frac{\mathfrak{d}_\xi(s-1)}{n}}\right)^s + \frac{\|\Sigma\|^{1/2}_{L_\infty(E)}}{n^{1/2}}\sqrt{\frac{\mathfrak{d}_\xi(s-1)}{n}} + \Delta^+_{\mathcal{H},\psi,E}(\hat\theta,\tilde\theta) \tag{2.10}$$

with $\mathcal{H} := \{g : \|g\|_{C^s(E)} \le 1\}$, and, for all $s' \in [1, s]$,

$$\sup_{\|f\|_{C^s(E)}\le 1}\sup_{\theta\in E}\Delta_{s'}(\sqrt{n}(f_k(\hat\theta) - f(\theta)), \sigma_f(\theta)Z)$$

$$\lesssim_s \left[\sqrt{n}\left(\sqrt{\frac{\mathfrak{d}_\xi(s-1)}{n}}\right)^s + \|\Sigma\|^{1/2}_{L_\infty(E)}\sqrt{\frac{\mathfrak{d}_\xi(s-1)}{n}} + \Delta_{\mathcal{F},E}(\sqrt{n}(\hat\theta-\theta), \xi(\theta))\right],$$

where $\mathcal{F} := \{g : \|g\|_{C^{0,s'}(E)} \le 1\}$.

In the general case, there are additional terms depending on tail probabilities of $\|\xi\|_{L_\infty(E)}$ and $\|\hat\theta - \theta\|$. The term $\tilde\psi^{1/2}\Big(\mathbb{P}\{\|\xi\|_{L_\infty(E)} \ge c_s\delta\sqrt{n}\}\Big)$ is negligible (smaller than $n^{-1/2}$) for losses ψ that grow slower than sub-exponential loss ψ_1 (see Remark 2.1). For the term

$$\sup_{\theta\in\Theta_\delta}\tilde\psi\Big(\mathbb{P}_\theta\{\|\hat\theta - \theta\| \ge c_s\delta\}\Big)$$

to be of the order $O(n^{-1/2})$, some conditions on the tail probabilities

$$\sup_{\theta\in\Theta_\delta}\mathbb{P}_\theta\{\|\hat\theta - \theta\| \ge c_s\delta\},$$

ranging from polynomial decay in the case of L_p-losses $\psi(u) = u^p$ to exponential decay in the case of sub-exponential losses, are needed. In some cases, it is possible to reduce these conditions to the conditions on the tails of $\|\xi\|_{L_\infty(E)}$ using normal approximation (see Corollary 2.2).

REMARK 2.4. The bounds of theorems 2.2, 2.3, 2.4 and 2.5 show that the estimator $f_k(\hat\theta)$ of $f(\theta)$ exhibits the same type of behavior as in the case of Gaussian shift model studied in [40] (see also Theorem 2.1 at the beginning of this section and the discussion that follows) provided that normal approximation of $\hat\theta$, quantified by such parameters as

$$\Delta^+_{\mathcal{H},\psi,\Theta_\delta}(\hat\theta,\tilde\theta) \ \text{ and } \ \Delta_{\mathcal{F},\Theta_\delta}(\sqrt{n}(\hat\theta-\theta),\xi(\theta)),$$

is sufficiently accurate.

1. The *"Gaussian parts"* of the bounds of these theorems, such as the part

$$\frac{\|\Sigma\|^{1/2}_{L_\infty(E)}}{n^{1/2}} + \Big(\sqrt{\frac{\mathfrak{d}_\xi(\Theta_\delta; s-1)}{n}}\Big)^s$$

of bound (2.6), are similar to the main part $\frac{\|\Sigma\|^{1/2}}{n^{1/2}} \bigvee \left(\sqrt{\frac{\mathbb{E}\|\xi\|^2}{n}}\right)^s$ of the bound of Theorem 2.1. The "Gaussian part" of bound (2.6) consists of two terms: the concentration term $\frac{\|\Sigma\|^{1/2}_{L_\infty(E)}}{n^{1/2}}$ controlling the random error of estimator $f_k(\hat\theta)$ and the bias term

$\left(\sqrt{\frac{\mathfrak{d}_\xi(\Theta_\delta;s-1)}{n}}\right)^s$ controlling the bias of the estimator. In fact, the Gaussian parts are exactly the same as in the case of Gaussian shift model when $\xi(\theta)$ does not depend on θ and it is possible to obtain Theorem 2.1 for Gaussian shift models as a corollary of our general results, see Corollary 2.1 below.

2. In typical examples, such as $\xi(\theta)=A(\theta)Z, Z\sim N(0,I_d)$, complexity parameters $\mathfrak{v}_\xi(\Theta)$ and $\mathfrak{d}_\xi(\Theta_\delta;s-1)$ could be easily controlled in terms of some dimension type parameter d (see, e.g., bound (2.5)), and the Gaussian parts of the bounds are controlled by the expression $\frac{1}{\sqrt{n}}+\left(\sqrt{\frac{d}{n}}\right)^s$. If the normal approximation terms of bounds of theorems 2.2, 2.3, 2.4 and 2.5 are negligible comparing with the Gaussian part, there is a phase transition from the classical $\frac{1}{\sqrt{n}}$ error rate when the smoothness s of functional f is sufficiently large to slower rates when the smoothness is not sufficient (similarly to the case of Gaussian shift models [40]). More precisely, if $d\le n^\alpha$ for some $\alpha\in(0,1)$, then $\frac{1}{\sqrt{n}}$ error rate for estimators $f_k(\hat\theta)$ holds for all $s\ge\frac{1}{1-\alpha}$ and slower rates hold for $s<\frac{1}{1-\alpha}$ (which is known to be a sharp threshold in the case of Gaussian shift models). Moreover, if $s>\frac{1}{1-\alpha}$, then the bounds of theorems 2.3 and 2.5 also imply normal approximation of estimator $f_k(\hat\theta)$. However, for Gaussian type bounds on estimator $f_k(\hat\theta)$ to hold in the whole range of values of $\alpha\in(0,1)$, the normal approximation of $\sqrt{n}(\hat\theta-\theta)$ by $\xi(\theta)$ should hold for $d=o(n)$ (see further discussion in Section 3).
3. Finally, note that, for $s\in(0,2]$, there is no need in bias reduction to achieve the optimal (in the Gaussian case) error rates and plug-in estimator $f(\hat\theta)$ could be used for this purpose (see theorems 2.2 and 2.3). For $s>2$, the bias of the plug-in estimator is too large and estimators with reduced bias, such as $f_k(\hat\theta)$, are needed to achieve the optimal rate (see theorems 2.4 and 2.5).

It is not hard to obtain a generalization of results of [40] to more general Gaussian shift models as a corollary of the results of the current paper. Namely, suppose that $X^{(n)}$ satisfies the following Gaussian shift model $X^{(n)}=\theta+\frac{\xi(\theta)}{\sqrt{n}},\theta\in E$, where $\xi(\theta)$ is a Gaussian random variable in E with mean 0 and covariance operator $\Sigma(\theta),\theta\in E$. In particular, this includes the Gaussian shift models studied in [40] in which the noise $\xi(\theta)=\xi$ did not depend on θ. Let $\hat\theta=\hat\theta(X^{(n)})=X^{(n)}$. The next corollary is immediate since $\hat\theta=\tilde\theta$ and $\sqrt{n}(\hat\theta-\theta)=\xi(\theta)$, implying that

$$\Delta_{\mathcal{F},\Theta_\delta}(\sqrt{n}(\hat\theta-\theta),\xi(\theta))=\Delta^+_{\mathcal{H},\psi,E}(\hat\theta,\tilde\theta)=0.$$

COROLLARY 2.1. *Let $s=k+1+\rho$ with $k\ge0$ and $\rho\in(0,1]$. For all $\psi\in\Psi$,*

$$\sup_{\|f\|_{C^s(E)}\le1}\sup_{\theta\in E}\|f_k(\hat\theta)-f(\theta)\|_{L_\psi(\mathbb{P}_\theta)}\lesssim_{s,\psi}\left[\frac{\|\Sigma\|^{1/2}_{L_\infty(E)}}{n^{1/2}}+\left(\sqrt{\frac{\mathfrak{d}_\xi(s-1)}{n}}\right)^s\right]\bigwedge 1.$$

Moreover, for $k\ge1$ under the assumption that $\mathfrak{d}_\xi(s)\le c_1n$ for a small enough constant $c_1>0$,

$$\sup_{\|f\|_{C^s(E)}\le1}\sup_{\theta\in E}\Big|\|f_k(\hat\theta)-f(\theta)\|_{L_\psi(\mathbb{P}_\theta)}-n^{-1/2}\sigma_f(\theta)\|Z\|_{L_\psi(\mathbb{P})}\Big|$$

$$\lesssim_{s,\psi}\left(\sqrt{\frac{\mathfrak{d}_\xi(s-1)}{n}}\right)^s+\frac{\|\Sigma\|^{1/2}_{L_\infty(E)}}{n^{1/2}}\sqrt{\frac{\mathfrak{d}_\xi(s-1)}{n}}$$

and, for all $s' \in [1,s]$,

$$\sup_{\|f\|_{C^s(E)}\le 1}\sup_{\theta\in E}\Delta_{s'}(\sqrt{n}(f_k(\hat\theta)-f(\theta)),\sigma_f(\theta)Z)$$
$$\lesssim_s \left[\sqrt{n}\left(\sqrt{\frac{\mathfrak{d}_\xi(s-1)}{n}}\right)^s + \|\Sigma\|^{1/2}_{L_\infty(E)}\sqrt{\frac{\mathfrak{d}_\xi(s-1)}{n}}\right].$$

For $k=0$, *the last two bounds hold without the terms involving* $\|\Sigma\|^{1/2}_{L_\infty(E)}\sqrt{\frac{\mathfrak{d}_\xi(s-1)}{n}}$.

In the case when the noise $\xi(\theta)=\xi$ does not depend on θ, we have $\mathfrak{d}_\xi(s-1)=\mathbb{E}\|\xi\|^2$, and the above bounds immediately imply the main results of paper [40].

The bounds of theorems 2.4 and 2.5 show that the "Gaussian error rates" would hold for other models provided that the additional terms related to the accuracy of normal approximation and to the tails of random variables $\|\hat\theta-\theta\|$ and $\|\xi\|_{L_\infty(E)}$ are negligible comparing with the Gaussian terms. To ensure this (and, in particular, to ensure that $\sqrt{n}$ convergence rate is attainable for estimator $f_k(\hat\theta)$ if f is sufficiently smooth), one needs the conditions

$$\Delta_{\mathcal{F},\Theta_\delta}\Big(\sqrt{n}(\hat\theta-\theta),\xi(\theta)\Big)\to 0 \ \text{ and } \ \Delta^+_{\mathcal{H},\psi,\Theta_\delta}(\hat\theta,\tilde\theta)=o(n^{-1/2}) \text{ as } n\to\infty.$$

The following proposition provides useful upper bounds on the distances $\Delta_{\mathcal{H},\Theta_\delta}(\hat\theta,\tilde\theta)$, $\Delta_{\mathcal{H},\psi,\Theta_\delta}(\hat\theta,\tilde\theta)$ and $\Delta^+_{\mathcal{H},\psi,\Theta_\delta}(\hat\theta,\tilde\theta)$.

PROPOSITION 2.1. *Let* $s\ge 1$. *For* $\mathcal{H}:=\{g:\|g\|_{C^s(\Theta_\delta)}\le 1\}$,

$$\Delta_{\mathcal{H},\Theta_\delta}(\hat\theta,\tilde\theta)\le \frac{\Delta_{\mathcal{F},\Theta_\delta}\Big(\sqrt{n}(\hat\theta-\theta),\xi(\theta)\Big)}{\sqrt{n}},\ \Delta_{\mathcal{H},\psi,\Theta_\delta}(\hat\theta,\tilde\theta)\le \frac{\Delta_{\mathcal{F},\psi,\Theta_\delta}\Big(\sqrt{n}(\hat\theta-\theta),\xi(\theta)\Big)}{\sqrt{n}}$$
$$\text{and } \ \Delta^+_{\mathcal{H},\psi,\Theta_\delta}(\hat\theta,\tilde\theta)\le \frac{\Delta^+_{\mathcal{F},\psi,\Theta_\delta}\Big(\sqrt{n}(\hat\theta-\theta),\xi(\theta)\Big)}{\sqrt{n}},$$

where $\mathcal{F}:=\{g:\|g\|_{C^{0,s}(U_{\delta\sqrt{n}})}\le 1\}$.

It follows that the condition $\Delta^+_{\mathcal{H},\psi,\Theta_\delta}(\hat\theta,\tilde\theta)=o(n^{-1/2})$ holds if

$$\Delta^+_{\mathcal{F},\psi,\Theta_\delta}\Big(\sqrt{n}(\hat\theta-\theta),\xi(\theta)\Big)=o(1).$$

Next we state corollaries of theorems 2.4 and 2.5 (and, for $s\in(1,2]$, of theorems 2.2 and 2.3) in the case of quadratic loss $\psi(u)=u^2$. In these corollaries, we will use Wasserstein distances W_1, W_2 to quantify the accuracy of normal approximation and to obtain a simpler form of the results.

COROLLARY 2.2. *Let* $s=k+1+\rho$ *with* $k\ge 1$ *and* $\rho\in(0,1]$, *and let* $\delta>0$. *Let* Θ *be a subset of* E *such that* $\Theta_\delta\subset T$. *Suppose that, for some sufficiently small constant* $c_1>0$,

$$\mathfrak{d}_\xi(\Theta_\delta;s)\le c_1\delta^2 n.$$

Then, for all $s'\in[1,s]$ *and for some constant* $c_2>0$,

$$\sup_{\|f\|_{C^s(\Theta_\delta)}\le 1}\sup_{\theta\in\Theta}\Delta_{s'}(\sqrt{n}(f_k(\hat\theta)-f(\theta)),\sigma_f(\theta)Z)$$

$$\lesssim_{s,\delta} \left[\sqrt{n}\left(\sqrt{\frac{\mathfrak{d}_\xi(\Theta_\delta; s-1)}{n}}\right)^s + \|\Sigma\|_{L_\infty(E)}^{1/2}\sqrt{\frac{\mathfrak{d}_\xi(\Theta_\delta; s-1)}{n}} \right.$$
$$\left. + W_{1,\Theta_\delta}(\sqrt{n}(\hat\theta-\theta), \xi(\theta)) + \sqrt{n}\exp\left\{-\frac{c_2\delta^2 n}{\|\Sigma\|_{L_\infty(E)}}\right\}\right].$$

COROLLARY 2.3. *Let $\Theta \subset T$, let $\delta > 0$ and let $s = k + 1 + \rho$ with $k \geq 0$ and $\rho \in (0,1]$. Suppose that $\Theta_\delta \subset T$. Then, for some constant $c_2 > 0$,*

$$\sup_{\|f\|_{C^s(\Theta_\delta)}\leq 1}\sup_{\theta\in\Theta}\|f_k(\hat\theta) - f(\theta)\|_{L_2(\mathbb{P}_\theta)}$$
$$\lesssim_{s,\delta} \left[\frac{\|\Sigma\|_{L_\infty(E)}^{1/2}}{n^{1/2}} + \left(\sqrt{\frac{\mathfrak{d}_\xi(\Theta_\delta; s-1)}{n}}\right)^s \right.$$
$$\left. + \frac{W_{2,\Theta_\delta}(\sqrt{n}(\hat\theta-\theta), \xi(\theta))}{\sqrt{n}} + \exp\left\{-\frac{c_2\delta^2 n}{\|\Sigma\|_{L_\infty(E)}}\right\}\right] \bigwedge 1. \tag{2.11}$$

Moreover, if for some sufficiently small constant $c_1 > 0$, $\mathfrak{d}_\xi(\Theta_\delta; s) \leq c_1\delta^2 n$, then

$$\sup_{\|f\|_{C^s(\Theta_\delta)}}\sup_{\theta\in\Theta}\left|\|f_k(\hat\theta) - f(\theta)\|_{L_2(\mathbb{P}_\theta)} - n^{-1/2}\sigma_f(\theta)\right|$$
$$\lesssim_{s,\delta} \left(\sqrt{\frac{\mathfrak{d}_\xi(\Theta_\delta; s-1)}{n}}\right)^s + \frac{\|\Sigma\|_{L_\infty(E)}^{1/2}}{n^{1/2}}\sqrt{\frac{\mathfrak{d}_\xi(\Theta_\delta; s-1)}{n}}$$
$$+ \frac{W_{2,\Theta_\delta}(\sqrt{n}(\hat\theta-\theta), \xi(\theta))}{\sqrt{n}} + \exp\left\{-\frac{c_2\delta^2 n}{\|\Sigma\|_{L_\infty(E)}}\right\}. \tag{2.12}$$

REMARK 2.5. Bounds of corollaries 2.2 and 2.3 also holds for $k = 0$. In this case, the terms involving $\|\Sigma\|_{L_\infty(E)}^{1/2}\sqrt{\frac{\mathfrak{d}_\xi(\Theta_\delta; s-1)}{n}}$ could be dropped.

As a simple consequence, we get the following result that shows asymptotic normality of estimator $f_k(\hat\theta)$ with $\sqrt{n}$ rate and provides an exact limit of its mean squared error if normal approximation holds and the functional f is sufficiently smooth.

COROLLARY 2.4. *Let $\Theta = \Theta_n \subset T$, let $\delta > 0$ and let $s = k + 1 + \rho$ with $k \geq 0$ and $\rho \in (0,1]$. Suppose that $\Theta_\delta \subset T$ and, for some $\alpha \in (0,1)$, $\mathfrak{d}_\xi(\Theta_\delta; s) \lesssim n^\alpha$. Suppose also that $\|\Sigma\|_{L_\infty(E)} \lesssim 1$. Assume that $s > \frac{1}{1-\alpha}$. Finally, suppose that*

$$W_{2,\Theta_\delta}\Big(\sqrt{n}(\hat\theta-\theta), \xi(\theta)\Big) \to 0 \text{ as } n\to\infty. \tag{2.13}$$

Then

$$\sup_{\|f\|_{C^s(\Theta_\delta)}\leq 1}\sup_{\theta\in\Theta}\left|n\mathbb{E}_\theta(f_k(\hat\theta) - f(\theta))^2 - \sigma_f^2(\theta)\right| \to 0 \text{ as } n\to\infty, \tag{2.14}$$

and, for all $\sigma_0 > 0$,

$$\sup_{\|f\|_{C^s(\Theta_\delta)}\leq 1}\sup_{\theta\in\Theta, \sigma_f(\theta)\geq\sigma_0} d_K\Big(\frac{\sqrt{n}(f_k(\hat\theta) - f(\theta))}{\sigma_f(\theta)}, Z\Big) \to 0 \text{ as } n\to\infty,$$

where $Z \sim N(0,1)$.[2]

[2]Of course, it is assumed here and in Theorem 2.5 that $\hat\theta = \hat\theta(X^{(n)}), X^{(n)} \sim P_\theta^{(n)}$.

REMARK 2.6. Let $\Theta = T = E$ and assume that, for some sufficiently small constant $c_1 > 0$, $\mathfrak{d}_\xi(s) \le c_1 n$. Then, for all $s' \in [1, s]$, the following version of the bound of Corollary 2.2 holds:

$$\sup_{\|f\|_{C^s(E)}\le 1} \sup_{\theta\in E} \Delta_{s'}(\sqrt{n}(f_k(\hat\theta) - f(\theta)), \sigma_f(\theta)Z)$$

$$\lesssim_s \left[\sqrt{n}\left(\sqrt{\frac{\mathfrak{d}_\xi(s-1)}{n}}\right)^s + \|\Sigma\|^{1/2}_{L_\infty(E)}\sqrt{\frac{\mathfrak{d}_\xi(s-1)}{n}} + W_{1,E}(\sqrt{n}(\hat\theta-\theta), \xi(\theta))\right]. \tag{2.15}$$

The bounds of Corollary 2.3 simplify as follows:

$$\sup_{\|f\|_{C^s(E)}\le 1} \sup_{\theta\in E} \|f_k(\hat\theta) - f(\theta)\|_{L_2(\mathbb{P}_\theta)}$$

$$\lesssim_s \left[\frac{\|\Sigma\|^{1/2}_{L_\infty(E)}}{n^{1/2}} + \left(\sqrt{\frac{\mathfrak{d}_\xi(s-1)}{n}}\right)^s + \frac{W_{2,E}(\sqrt{n}(\hat\theta-\theta), \xi(\theta))}{\sqrt{n}}\right] \wedge 1, \tag{2.16}$$

and, under the condition $\mathfrak{d}_\xi(s) \le c_1 n$ for a small enough constant $c_1 > 0$,

$$\sup_{\|f\|_{C^s(E)}\le 1} \sup_{\theta\in E} \Big| \|f_k(\hat\theta) - f(\theta)\|_{L_2(\mathbb{P}_\theta)} - n^{-1/2}\sigma_f(\theta)\Big|$$

$$\lesssim_s \left(\sqrt{\frac{\mathfrak{d}_\xi(s-1)}{n}}\right)^s + \frac{\|\Sigma\|^{1/2}_{L_\infty(E)}}{n^{1/2}}\sqrt{\frac{\mathfrak{d}_\xi(s-1)}{n}} + \frac{W_{2,E}(\sqrt{n}(\hat\theta-\theta), \xi(\theta))}{\sqrt{n}}. \tag{2.17}$$

3. Examples and applications: estimation of functionals and normal approximation in high-dimensional spaces. To apply the results of Section 2 to concrete statistical models, one needs to use sharp bounds on the accuracy of normal approximation over classes of smooth functions for typical statistical estimators (such as maximum likelihood estimators) in a high-dimensional setting. Ideally, in the case of a d-dimensional parameter θ, bounds on such distances as $\Delta_{\mathcal{F},\Theta_\delta}(\sqrt{n}(\hat\theta-\theta), \xi(\theta))$ with $\mathcal{F} := \{g : \|g\|_{C^{0,s'}(U_{\delta\sqrt{n}})} \le 1\}$ of the order $\sqrt{\frac{d}{n}}$, or $\Delta_{\mathcal{H},\psi,\Theta_\delta}(\hat\theta, \tilde\theta)$ with $\mathcal{H} := \{h : \|h\|_{C^s(\Theta_\delta)} \le 1\}$ of the order $n^{-1/2}\sqrt{\frac{d}{n}}$ are needed to ensure that the normal approximation holds for $d = o(n)$. This would allow us to deduce from Theorem 2.4 and Theorem 2.5 the results known to be optimal in the Gaussian case. Unfortunately, such bounds are, in our view, underdeveloped in the literature, not only in the case of general classes of estimators for high-dimensional models, such as MLE (see, e.g., [1, 2]), but even in the case of classical central limit theorems (CLT) in high-dimensional spaces (see, e.g., [56] where there are counterexamples showing that CLT could fail for some reasonable distributions in $\mathbb{R}^d$ unless $d^2 = o(n)$). The main difficulties involved in these problems are purely probabilistic: identifying classes of distributions in high-dimensional spaces with a reasonably good dependence of the normal approximation bounds on the dimension. The importance of these problems in high-dimensional statistics goes far beyond their applications to functional estimation discussed in the current paper. In this section, we will provide a very brief review of some approaches to high-dimensional CLT (including, very recent ones) and discuss several applications to the problem of functional estimation. A more detailed development of this approach is beyond the scope of the paper.

3.1. *High-dimensional CLT.* The rates of convergence in CLT in $\mathbb{R}^d$ and in infinite-dimensional spaces have been studied for over fifty years (see [8], [53], [61] and references therein) with a goal to obtain the bounds on the accuracy of normal approximation in various distances in the spaces of probability distributions often represented by sup-norms over

classes of sets (for instance, convex sets), or classes of functions (for instance, Lipschitz functions).

The distances ζ_s defined by (1.5) are particularly useful for our purposes. Such distances occur very naturally in connection to the Lindeberg's proof of CLT, they are used as a tool in bounding other distances (the sup-norms over convex sets, bounded Lipschitz distance, etc) and they were advocated in [64]. In particular, the following fact is straightforward: if $X_1, \dots, X_n$ are i.i.d. r.v. in $\mathbb{R}^d$ (equipped with the Euclidean norm) with mean zero and identity covariance and Z is a standard normal r.v. in $\mathbb{R}^d$, then

$$\zeta_3\Big(\frac{X_1+\dots+X_n}{\sqrt{n}}, Z\Big) \lesssim \frac{\mathbb{E}\|X\|^3}{\sqrt{n}}.$$

Typically, $\mathbb{E}\|X\|^3$ could be of the order $d^{3/2}$, yielding the bound on ζ_3-distance of the order $\frac{d^{3/2}}{\sqrt{n}}$. Thus, normal approximation holds when $d = o(n^{1/3})$. This is not good enough for our purposes since an interesting regime in functional estimation problem is $d \sim n^{\alpha}$ for $\alpha \geq 1/2$, which leads to non-trivial bias reduction problems. However, in some special cases, in particular, in the case of random vectors with independent components, one can improve bounds on ζ_s-distance rather substantially. The following fact is very simple and well known (see [64] for similar statements).

PROPOSITION 3.1. *Let $Y = (Y_1, \dots, Y_d)$, $Y' = (Y'_1, \dots, Y'_d)$ be two random vectors with independent components. Then*

$$\zeta_s(Y, Y') \leq \sum_{j=1}^{d} \zeta_s(Y_j, Y'_j).$$

As a consequence, in the case when r.v. $X = (X^{(1)}, \dots, X^{(d)})$ has independent components,

$$\zeta_3\Big(\frac{X_1+\dots+X_n}{\sqrt{n}}, Z\Big) \lesssim \max_{1\leq j\leq d} \mathbb{E}|X^{(j)}|^3 \frac{d}{\sqrt{n}}.$$

and if, in addition, $\mathbb{E}(X^{(j)})^3 = 0$, then it is easy to see that

$$\zeta_4\Big(\frac{X_1+\dots+X_n}{\sqrt{n}}, Z\Big) \lesssim \max_{1\leq j\leq d} \mathbb{E}|X^{(j)}|^4 \frac{d}{n}.$$

The last bound is of the order $\frac{d}{n}$, which is already sufficient for our purposes.

In Subsection 3.2 below, we use this very simple approach to study estimation of smooth functionals for some statistical models with independent components.

In the recent years, there has been a lot of interest in studying normal approximation bounds in high-dimensional CLT in optimal transport distances (in particular, Wasserstein type distances). A recent result in [23], provides the following bound on the Wasserstein W_2-distance in normal approximation: assuming that $\|X\| \leq \beta$ a.s.,

$$W_2\Big(\frac{X_1+\dots+X_n}{\sqrt{n}}, Z\Big) \lesssim \frac{\beta\sqrt{d\log n}}{\sqrt{n}}.$$

Thus, for convergence of W_2-distance to 0, this bound requires the condition $d = o\Big(\sqrt{\frac{n}{\log n}}\Big)$ in typical situations when $\beta \sim \sqrt{d}$. This is again too restrictive for our purposes.

In recent papers [18, 24], another approach to high-dimensional normal approximation has been developed. It is based on the technique of Stein kernels and it applies to probability distributions in $\mathbb{R}^d$ with bounded Poincaré constants, in particular, to some log-concave distributions (see also [3] for more general results).

A probability measure μ on $\mathbb{R}^d$ is said to satisfy Poincaré inequality iff there exists a constant $C>0$ such that for all locally Lipschitz functions $g:\mathbb{R}^d\mapsto\mathbb{R}$ and for $X\sim\mu$,

$$\mathrm{Var}_\mu(g(X))\le C\mathbb{E}_\mu\|(\nabla g)(X)\|^2.$$

Let $C_P(\mu)$ denote the infimum of all constants $C>0$ for which the inequality holds. It is called the Poincaré constant of probability measure μ.

A probability measure (distribution) μ on $\mathbb{R}^d$ with density p is called *log-concave* if p is a log-concave function, that is, $\log p$ is concave. Among the examples of log-concave distributions are Gaussian measures and uniform distributions in convex bodies of $\mathbb{R}^d$. It is known that log-concave distributions satisfy Poincaré inequality.

REMARK 3.1. The following facts are well known:

1. For a standard Gaussian measure μ on $\mathbb{R}^d$, $C_P(\mu)=1$. Moreover, if $\mu(dx)=e^{-V(x)}dx$ with $V:\mathbb{R}^d\mapsto\mathbb{R}$ such that $V''(x)\succeq C^{-1}$ for a symmetric positively definite matrix C, then $C_P(\mu)\le\|C\|$, and, if B is a symmetric positively definite matrix and
$$\mu(dx)=\exp\Big\{-\frac{1}{2}\langle B^{-1}x,x\rangle-V(x)\Big\}dx,$$
where V is a convex function on $\mathbb{R}^d$, then $C_P(\mu)\le\|B\|$ (see [9]).
2. There are also ways to control the value of Poincaré constant under certain perturbations of probability measure. For instance, if μ,ν are two probability measures 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 $C_P(\mu)\le\frac{A}{a}C_P(\nu)$. Also, if μ,ν are log-concave measures on $\mathbb{R}^d$ and, for some $\varepsilon\in(0,1)$,
$$d_{TV}(\mu,\nu):=\sup_{A\subset\mathbb{R}^d}|\mu(A)-\nu(A)|\le 1-\varepsilon,$$
then $C_P(\mu)\lesssim_\varepsilon C_P(\nu)$ (see [49]).
3. Let μ be an arbitrary log-concave distribution with covariance Σ. According to the Kannan-Lovàsz-Simonovits (KLS) conjecture, $C_P(\mu)\lesssim\|\Sigma\|$. Although this conjecture still remains open, it was recently shown in [13] (building upon earlier results of [23, 45]) that for some constant $c>0$
$$C_P(\mu)\le d^{c\left(\frac{\log\log d}{\log d}\right)^{1/2}}\|\Sigma\|.$$

It was proved in [18] that, if $X_1,\dots,X_n$ are i.i.d. mean zero random variables with identity covariance sampled from a distribution μ on $\mathbb{R}^d$ such that $C_P(\mu)<\infty$, then

$$W_2\Big(\frac{X_1+\cdots+X_n}{\sqrt{n}},Z\Big)\le\sqrt{C_P(\mu)-1}\sqrt{\frac{d}{n}}, \tag{3.1}$$

where $Z\sim N(0;I_d)$. Thus, the convergence in high-dimensional CLT in the W_2-distance holds provided that $d=o(n)$ for all the distributions with bounded Poincaré constants. If distribution μ is log-concave, then it follows from the bound on Poincaré constant proved in [13] (see Remark 3.1) that the CLT holds provided that $d\le n^{1-\delta}$ for an arbitrary $\delta>0$.

This approach will be used in Subsection 3.3 below to study smooth functional estimation for some classes of log-concave and related models.

REMARK 3.2. Another interesting approach to high-dimensional normal approximation was initiated in [14]. In this paper, the authors were trying to overcome the "curse of dimensionality" in CLT by sacrificing the convergence rate with respect to n. Namely, they proved the bound on the accuracy of normal approximation in sup-norm over the class of hyperrectangles of the order $O\Big(\Big(\frac{\log^7(dn)}{n}\Big)^{1/6}\Big)$, implying that normal approximation holds provided that $\log^7 d = o(n)$. More recently, this result was improved in [43, 35, 15, 16]. In particular, it was shown in [16] that the accuracy of normal approximation over hyperrectangles is of the order $O\Big(\frac{\log^{3/2} d}{\sqrt{n}}\log n\Big)$, which is optimal up to a $\log n$ factor. Thus, the normal approximation holds when $\log^3 d = o\Big(\frac{n}{\log^2 n}\Big)$. In principle, the results of this type could be adapted for our purposes in the case when E is the space $\mathbb{R}^d$ equipped with the ℓ_∞-norm. However, in this case $\mathbb{E}\|\xi\|^2_{\ell_\infty}$ would be typically of the order $\log d$ and this would be also a typical size of such parameters as $\mathfrak{d}_\xi(s)$ involved in our bounds. Thus, the Gaussian part of the error bounds in functional estimation (see Remark 2.4) would be of the order $\frac{1}{\sqrt{n}} + \Big(\sqrt{\frac{\log d}{n}}\Big)^s$. If $\log d = o(n^\alpha)$ for some $\alpha < 1/3$, the classical rate $n^{-1/2}$ dominates the bias term $\Big(\sqrt{\frac{\log d}{n}}\Big)^s$ for all $s \geq 3/2$ and there is no improvement of the rate when the degree of smoothness s is above $3/2$. Moreover, for such low values of $\log d$, the bias reduction is not required and the optimal rates would be attained for plug-in estimators. One would need to have normal approximation in high-dimensional CLT in the distances relevant in our paper for $\log d = o(n)$ to take the full advantage of the bias reduction method in the whole range of smoothness of the functionals. However, such normal approximation results do not seem to be available in the current literature.

3.2. *Independent components.* We will start this section with an application of Corollary 2.2 and Corollary 2.3 to statistical models with many independent components.

Let $X^{(n)} = (X_1^{(n)}, \dots, X_d^{(n)})$ be an observation with values in the space $S^{(n)} := S_1^{(n)} \times \cdots \times S_d^{(n)}$, where $(S_j^{(n)}, \mathcal{A}_j^{(n)}), j = 1, \dots, d$ are measurable spaces and $S^{(n)}$ is equipped with the product σ-algebra $\mathcal{A}^{(n)} := \mathcal{A}_1^{(n)} \times \cdots \times \mathcal{A}_d^{(n)}$. We will assume that the components $X_1^{(n)}, \dots, X_d^{(n)}$ of $X^{(n)}$ are independent r.v. and $X_j^{(n)} \sim P_{\theta_j}^{(n)}$ with parameter θ_j taking values in a Banach space E_j, $j = 1, \dots, d$. Let $E := E_1 \times \cdots \times E_d$ be equipped with a standard structure of linear space (the direct sum of linear spaces $E_1, \dots, E_d$) and with the norm $\|x\| = \Big(\sum_{j=1}^d \|x_j\|^2\Big)^{1/2}$, $x = (x_1, \dots, x_d) \in E$. Then, clearly, $X^{(n)} \sim P_\theta^{(n)}, \theta \in E$, where $P_\theta^{(n)} := P_{\theta_1}^{(n)} \times \cdots \times P_{\theta_d}^{(n)}$, $\theta = (\theta_1, \dots, \theta_d) \in E$. In the problems we have in mind, $\{P_{\theta_j}^{(n)} : \theta_j \in E_j\}, j = 1, \dots, d$ are low dimensional models and the complexity of combined model $\{P_\theta^{(n)} : \theta = (\theta_1, \dots, \theta_d) \in E\}$ depends only on the number d of independent components.

Let $\hat{\theta}_j = \hat{\theta}_j(X_j^{(n)})$ be estimators of parameters $\theta_j, j = 1, \dots, d$ and let $\hat{\theta} := (\hat{\theta}_1, \dots, \hat{\theta}_d)$ be the estimator of θ. Assume that $\sqrt{n}(\hat{\theta}_j - \theta_j)$ could be approximated in distribution by a centered Gaussian r.v. $\xi_j(\theta_j)$ with values in E_j and with covariance operator $\Sigma_j(\theta_j)$. Since $\hat{\theta}_j, j = 1, \dots, d$ are independent r.v., we assume that $\xi_j, j = 1, \dots, d$ are also independent and $\xi(\theta) := (\xi_1(\theta_1), \dots, \xi_d(\theta_d)), \theta = (\theta_1, \dots, \theta_d) \in E$ can be used to approximate $\sqrt{n}(\hat{\theta} - \theta)$ in

distribution. The following formula holds for the covariance operator $\Sigma(\theta)$ of $\xi(\theta)$:

$$\langle \Sigma(\theta)u, v\rangle = \sum_{j=1}^d \langle \Sigma_j(\theta_j)u_j, v_j\rangle,$$

$$u = (u_1, \dots, u_d), v = (v_1, \dots, v_d) \in E^* = E_1^* \times \cdots \times E_d^*. \tag{3.2}$$

Moreover, we will view $\theta_j \mapsto \xi_j(\theta_j)$ as a stochastic process and use the following characteristic of ξ :

$$\mathfrak{q}_\xi(s) := \begin{cases} \frac{1}{d}\sum_{i=1}^d \mathbb{E}\|\xi_i\|^2_{C^s(E_i)} & \text{for } s \in (0,1] \\ \frac{1}{d}\sum_{i=1}^d \mathbb{E}\|\xi_i\|^2_{C^1(E_i)} + \frac{\log(2d)}{d}\max_{1\le i\le d}\mathbb{E}\|\xi_i\|^2_{C^{1,s}(E_i)} & \text{for } s > 1. \end{cases}$$

Based on estimator $\hat\theta := (\hat\theta_1, \dots, \hat\theta_d)$, define operators $\mathcal{T}, \mathcal{B}$ and functions f_k. Note that

$$\sigma_f^2(\theta) = \langle \Sigma(\theta)f'(\theta), f'(\theta)\rangle = \sum_{j=1}^d \langle \Sigma_j(\theta_j)f'_{\theta_j}(\theta), f'_{\theta_j}(\theta)\rangle,$$

where $f'_{\theta_j}(\theta) \in E_j^*$ denotes the partial Fréchet derivative of $f(\theta) = f(\theta_1, \dots, \theta_d)$ w.r.t. θ_j.

The following result holds.

COROLLARY 3.1. *Suppose that $\mathfrak{q}_\xi(s) \lesssim 1$ and, for some sufficiently small constant $c_1 > 0$, $d \le c_1 n$. Let $s = k + 1 + \rho$ with $k \ge 1$ and $\rho \in (0,1]$. Then, for all $s' \in [1, s]$,*

$$\sup_{\|f\|_{C^s(E)}\le 1}\sup_{\theta\in E}\Delta_{s'}(\sqrt{n}(f_k(\hat\theta) - f(\theta)), \sigma_f(\theta)Z)$$

$$\lesssim_s \Bigg[\sqrt{n}\mathfrak{q}_\xi^{s/2}(s-1)\left(\sqrt{\frac{d}{n}}\right)^s + \max_{1\le j\le d}\|\Sigma_j\|^{1/2}_{L_\infty(E_j)}\mathfrak{q}_\xi^{1/2}(s-1)\sqrt{\frac{d}{n}}$$

$$+ \Big(\sum_{j=1}^d W^2_{2,E_j}(\sqrt{n}(\hat\theta_j - \theta_j), \xi_j(\theta_j))\Big)^{1/2}\Bigg] \tag{3.3}$$

and

$$\sup_{\|f\|_{C^s(E)}\le 1}\sup_{\theta\in E}\Big|\|f_k(\hat\theta) - f(\theta)\|_{L_2(\mathbb{P}_\theta)} - n^{-1/2}\sigma_f(\theta)\Big|$$

$$\lesssim_s \mathfrak{q}_\xi^{s/2}(s-1)\left(\sqrt{\frac{d}{n}}\right)^s + \frac{\max_{1\le j\le d}\|\Sigma_j\|^{1/2}_{L_\infty(E_j)}}{n^{1/2}}\mathfrak{q}_\xi^{1/2}(s-1)\sqrt{\frac{d}{n}}$$

$$+ \frac{1}{\sqrt{n}}\Big(\sum_{j=1}^d W^2_{2,E_j}(\sqrt{n}(\hat\theta_j - \theta_j), \xi_j(\theta_j))\Big)^{1/2}. \tag{3.4}$$

In particular, bound (3.4) implies that

$$\sup_{\|f\|_{C^s(E)}\le 1}\sup_{\theta\in E}\|f_k(\hat\theta) - f(\theta)\|_{L_2(\mathbb{P}_\theta)} \lesssim_s \frac{\max_{1\le j\le d}\|\Sigma_j\|^{1/2}_{L_\infty(E_j)}}{n^{1/2}} + \mathfrak{q}_\xi^{s/2}(s-1)\left(\sqrt{\frac{d}{n}}\right)^s$$

$$+ \frac{1}{\sqrt{n}}\Big(\sum_{j=1}^d W^2_{2,E_j}(\sqrt{n}(\hat\theta_j - \theta_j), \xi_j(\theta_j))\Big)^{1/2}. \tag{3.5}$$

REMARK 3.3. The bounds also hold for $k = 0$. In this case, the terms

$$\max_{1\le j\le d} \|\Sigma_j\|^{1/2}_{L_\infty(E_j)} \mathfrak{q}_\xi^{1/2}(s-1)\sqrt{\frac{d}{n}}$$

of (3.3) and

$$\frac{\max_{1\le j\le d} \|\Sigma_j\|^{1/2}_{L_\infty(E_j)}}{n^{1/2}} \mathfrak{q}_\xi^{1/2}(s-1)\sqrt{\frac{d}{n}}$$

of (3.4) could be dropped.

REMARK 3.4. Suppose $\mathfrak{q}_\xi(s) \lesssim 1$. In particular, for $s > 1$, this holds if $\max_{1\le i\le d} \mathbb{E}\|\xi_i\|^2_{C^1(E_i)} \lesssim 1$ and $\max_{1\le i\le d} \mathbb{E}\|\xi_i\|^2_{C^{1,s}(E_i)} \lesssim \frac{d}{\log d}$. Suppose, in addition, that $\max_{1\le j\le d} \|\Sigma_j\|^{1/2}_{L_\infty(E_j)} \lesssim 1$. If the models $\{P_{\theta_j} : \theta_j \in E_j\}$ are low-dimensional and sufficiently regular, the assumptions above hold for maximum likelihood estimators $\hat\theta_j$ of $\theta_j, j = 1,\dots,d$. In fact, in this case, we would have $\max_{1\le i\le d} \mathbb{E}\|\xi_i\|^2_{C^s(E_i)} \lesssim 1$ (if the Fisher information matrices $I_j(\theta_j)$ of low dimensional models are sufficiently smooth). If, in addition, the following normal approximation bound holds for the estimators $\hat\theta_j$ of the components θ_j

$$\max_{1\le j\le d} W_{2,E_j}(\sqrt{n}(\hat\theta_j - \theta_j), \xi_j(\theta_j)) \lesssim n^{-1/2}, \tag{3.6}$$

then we have

$$\Big(\sum_{j=1}^d W^2_{2,E_j}(\sqrt{n}(\hat\theta_j - \theta_j), \xi_j(\theta_j))\Big)^{1/2} \lesssim \sqrt{\frac{d}{n}},$$

which guarantees normal approximation of $\sqrt{n}(\hat\theta - \theta)$ by $\xi(\theta)$ for $d = o(n)$. In this case, (3.3) implies

$$\sup_{\|f\|_{C^s(E)}\le 1} \sup_{\theta\in E} \Delta_{s'}(\sqrt{n}(f_k(\hat\theta) - f(\theta)), \sigma_f(\theta)Z) \lesssim_s \sqrt{n}\left(\sqrt{\frac{d}{n}}\right)^s + \sqrt{\frac{d}{n}},$$

(3.4) implies

$$\sup_{\|f\|_{C^s(E)}\le 1} \sup_{\theta\in E} \Big| \|f_k(\hat\theta) - f(\theta)\|_{L_2(\mathbb{P}_\theta)} - n^{-1/2}\sigma_f(\theta)\Big| \lesssim_s \left(\sqrt{\frac{d}{n}}\right)^s + \frac{1}{\sqrt{n}}\sqrt{\frac{d}{n}}$$

and (3.5) implies that

$$\sup_{\|f\|_{C^s(E)}\le 1} \sup_{\theta\in E} \|f_k(\hat\theta) - f(\theta)\|_{L_2(\mathbb{P}_\theta)} \lesssim_s \frac{1}{\sqrt{n}} + \left(\sqrt{\frac{d}{n}}\right)^s.$$

If $d \le n^\alpha$ for some $\alpha \in (0,1)$ and $s > \frac{1}{1-\alpha}$, the above bounds imply the asymptotic normality of estimator $f_k(\hat\theta)$ with $\sqrt{n}$ rate as well as the convergence of $\sqrt{n}\|f_k(\hat\theta) - f(\theta)\|_{L_2(\mathbb{P}_\theta)}$ to $\sigma_f(\theta)$. Note also that, if $d \le n^\alpha$ for some $\alpha \in (0,1)$, then it is sufficient for asymptotic normality of $f_k(\hat\theta)$ and for convergence of its normalized risk to $\sigma_f(\theta)$ to have normal approximation error in (3.6) of the order $o(n^{-\alpha/2})$ instead of $n^{-1/2}$.

REMARK 3.5. In the low-dimensional case, bounds of the order $n^{-1/2}$ on the accuracy of normal approximation of MLE and more general M-estimators in Kolmogorov's distance (Berry-Esseen type bounds) could be found in [54, 4, 55] and in Wasserstein's W_1-distance in [2]. We are not aware of similar published results for Wassertein's W_2-distance. However,

it is possible to adapt the approach of these papers in combination with known bounds on the accuracy of normal approximation in CLT (see, e.g., [58, 23]) to obtain bounds for the W_2-distance suitable in the framework of Corollary 3.1.

Additional examples of models with independent components are provided in the supplement [38].

3.3. *Poincaré constants and log-concave models.* Let $E=\mathbb{R}^d$ be equipped with the Euclidean norm and let $X\sim P_\theta, \theta\in T, T\subset\mathbb{R}^d$ be a statistical model with the sample space $\mathbb{R}^d$. As before, we assume that T is an open subset. Also assume that $\mathbb{E}_\theta\|X\|^2<\infty, \theta\in T$ and let

$$\Psi(\theta):=\mathbb{E}_\theta X,\quad \Sigma(\theta):=\mathbb{E}_\theta(X-\Psi(\theta))\otimes(X-\Psi(\theta)), \theta\in\Theta.$$

Moreover, let us assume that $\Psi: T\mapsto \Psi(T)$ is a homeomorphism between open sets T and $\Psi(T)$. This assumption would allow us to re-parametrize our model by setting $\vartheta:=\Psi(\theta)=\mathbb{E}_\theta X, \theta\in T$ and using parameter $\vartheta\in\Psi(T)$ instead of θ. For this new parameter, we simply have $\mathbb{E}_\vartheta X=\vartheta, \vartheta\in\Psi(T)$.

Given i.i.d. observations $X_1,\dots,X_n$ of X, let

$$\bar X:=\frac{X_1+\cdots+X_n}{n},\qquad \hat\theta=\hat\theta(X_1,\dots,X_n)=\begin{cases}\Psi^{-1}(\bar X) & \text{if } \bar X\in\Psi(T)\\ \theta_0 & \text{if } \bar X\notin\Psi(T),\end{cases}$$

where $\theta_0\in T$ is an arbitrary point, and

$$\hat\vartheta=\hat\vartheta(X_1,\dots,X_n)=\begin{cases}\bar X & \text{if } \bar X\in\Psi(T)\\ \Psi(\theta_0) & \text{if } \bar X\notin\Psi(T)\end{cases}=\Psi(\hat\theta).$$

It is easy to check that $\mathcal{T}(f\circ\Psi^{-1})=(\mathcal{T}f)\circ\Psi^{-1}$, $\mathcal{B}(f\circ\Psi^{-1})=(\mathcal{B}f)\circ\Psi^{-1}$ and $(f\circ\Psi^{-1})_k=f_k\circ\Psi^{-1}$, where, with a little abuse of notation, we keep the same letters $\mathcal{T}$ and $\mathcal{B}$ to denote the operators based on estimator $\hat\vartheta$. This allows us to reduce the problem of estimation of functional $f(\theta)$ to the problem of estimation of functional $(f\circ\Psi^{-1})(\vartheta)$ under its proper smoothness and to use for this purpose the estimator

$$f_k(\hat\theta)=(f_k\circ\Psi^{-1})(\hat\vartheta)=(f\circ\Psi^{-1})_k(\hat\vartheta).$$

Of course, one can expect that $\sqrt{n}(\hat\vartheta-\vartheta)$ could be approximated by Gaussian random variable $\xi(\theta)$ with mean zero and covariance operator $\Sigma(\theta)$ (for $\vartheta=\Psi(\theta)$).

We will assume that P_θ satisfies Poincaré inequality, so, $C_P(P_\theta)<\infty$. Let

$$\sigma^2_{f\circ\Psi^{-1}}(\vartheta)=\langle\Sigma(\Psi^{-1}(\vartheta))(f\circ\Psi^{-1})'(\vartheta),(f\circ\Psi^{-1})'(\vartheta)\rangle.$$

PROPOSITION 3.2. *Let $d=d_n$ and $\Theta=\Theta_n\subset\mathbb{R}^d$ with $\mathrm{Diam}(\Theta)\lesssim n^A$ for some $A>0$. Let $\delta>0$ and let $s=k+1+\rho$ with $k\geq 0$ and $\rho\in(0,1]$. Suppose that $\Theta_\delta\subset T$ and*

$$\|\Sigma\|_{C^s(\Theta_\delta)}\lesssim 1 \text{ and } \|\Sigma^{-1}\|_{L_\infty(\Theta_\delta)}\lesssim 1. \tag{3.7}$$

Suppose that, for some $\alpha\in(0,1)$, $d\lesssim n^\alpha$ and assume that $s>\frac{1}{1-\alpha}$. Finally, suppose that

$$\sup_{\theta\in\Theta_\delta} C_P(P_\theta)=o(n^{1-\alpha}) \text{ as } n\to\infty. \tag{3.8}$$

Let θ_0 in the definition of $\hat\theta$ be a point from Θ. Then

$$\sup_{\|f\circ\Psi^{-1}\|_{C^s((\Psi(\Theta))_\delta)}\leq 1}\sup_{\theta\in\Theta}\left|n\mathbb{E}_\theta(f_k(\hat\theta)-f(\theta))^2-\sigma^2_{f\circ\Psi^{-1}}(\Psi(\theta))\right|\to 0 \tag{3.9}$$

and, for all $\sigma_0 > 0$,

$$\sup_{\|f\circ\Psi^{-1}\|_{C^s((\Psi(\Theta))_\delta)}\le 1}\ \sup_{\theta\in\Theta,\sigma_{f\circ\Psi^{-1}}(\Psi(\theta))\ge\sigma_0} d_K\Big(\frac{\sqrt{n}(f_k(\hat\theta)-f(\theta))}{\sigma_{f\circ\Psi^{-1}}(\Psi(\theta))}, Z\Big)\to 0 \tag{3.10}$$

as $n\to\infty$.

REMARK 3.6. Suppose that, for small $\delta>0$, Ψ is a C^s-diffeomorphism between Θ_δ and $\Psi(\Theta_\delta)$ (with bounded C^s-norms of Ψ and Ψ^{-1}). Then, for a small enough $\delta>0$, there exists $\delta'>0$ such that $\Psi^{-1}((\Psi(\Theta))_{\delta'})\subset\Theta_\delta$ and the first supremum in (3.10) and (3.9) could be taken over the set $\|f\|_{C^s(\Theta_\delta)}\lesssim 1$.

REMARK 3.7. The properties of Poincaré constants discussed in Remark 3.1 provide a way to check condition (3.8). In particular, the claim of the corollary obviously holds in the Gaussian case. Moreover, if P_θ is absolutely continuous with respect to a measure ν_θ for which $C_P(\nu_\theta)$ is controlled by a numerical constant (for instance, a Gaussian measure) and the densities $\frac{dP_\theta}{d\nu_\theta}$ are bounded from above by a constant $A>0$ and bounded from below by a constant $a>0$, then $C_P(P_\theta)\lesssim 1$ and condition (3.8) holds. Thus, the claim of Proposition 3.2 also holds (under the rest of its conditions).

REMARK 3.8. Suppose that measures $P_\theta,\theta\in\Theta_\delta$ are log-concave. It follows from a recent result of [13] (see Remark 3.1) that $\sup_{\theta\in\Theta_\delta} C_P(P_\theta)\lesssim_\nu d^\nu$ for an arbitrary $\nu>0$. Thus, in this case, condition (3.8) holds for all $\alpha\in(0,1)$ (and $d\lesssim n^\alpha$) and so are the claims of Proposition 3.2.

In the simplest case, $X=\theta+\eta$, where η is a mean zero noise sampled from some distribution μ_θ in $\mathbb{R}^d$, depending on the parameter θ. In this case, $\vartheta=\Psi(\theta)=\theta$ and it is easy to state a simplified version of Proposition 3.2. If the distribution $\mu_\theta=\mu$ of the noise does not depend on θ and $C_P(\mu)<\infty$, similar problems were studied in a recent paper [42]. The approach was based on a more direct analysis of estimator $f_k(\hat\theta)$ in the case of such Poincaré random shift models without using normal approximation. However, this approach could not be extended to more general models with distribution μ_θ of the noise depending on θ since, in this case, the construction of random homotopies between estimator $\bar X$ and parameter ϑ leads to rather challenging coupling problems (see also the discussion in Section 1.1).

A slightly more complicated example, is an exponential family[3]

$$P_\theta(dx)=\frac{1}{Z(\theta)}\exp\{\langle\theta,x\rangle\}h(x)dx,\theta\in T, \tag{3.11}$$

where $h:\mathbb{R}^d\mapsto[0,+\infty)$ is a Borel function and $Z(\theta):=\int_{\mathbb{R}^d}\exp\{\langle\theta,x\rangle\}h(x)dx<\infty,\theta\in T$. Note that the set $\{\theta\in\mathbb{R}^d: Z(\theta)<+\infty\}$ is convex and T is a subset of this set. Assume that T is convex, too. It is well known that $T\ni\theta\mapsto\log Z(\theta)$ is a strictly convex smooth function and

$$\vartheta=\Psi(\theta)=\mathbb{E}_\theta X=(\nabla\log Z)(\theta),\theta\in T.$$

Moreover, $\Psi=\nabla\log Z$ is a strictly monotone vector field on T (as a gradient of a strictly convex smooth function) and, therefore, it is a one-to-one mapping from T onto $\Psi(T)$ (as before, it is also assumed to be a homeomorphism). Following the terminology of [12] (which

[3] All the facts about exponential families used below could be found, for instance, in [12]

is not quite standard), θ is called the *canonical parameter* of the exponential family and ϑ is called its *natural parameter*.

Note also that $(\log Z)''(\theta) = \Psi'(\theta) = \Sigma(\theta)$ is the covariance of X. It is also the Fisher information matrix $I(\theta)$ for this model with respect to the canonical parameter θ and the inverse Fisher information matrix $\mathcal{I}^{-1}(\vartheta)$ with respect to the natural parameter $\vartheta = \Psi(\theta)$. Let now $X_1, \dots, X_n$ be i.i.d. $\sim P_\theta, \theta \in T$. If $\bar{X} \in \Psi(T)$, then $\hat{\theta} = \Psi^{-1}(\bar{X})$ is the unique maximum likelihood estimator for this exponential model.

We will call exponential family (3.11) log-concave iff the function h is log-concave. Clearly, in this case the distributions $P_\theta, \theta \in T$ are log-concave. Proposition (3.2) and the above discussion yield the following corollary.

COROLLARY 3.2. *Let $d = d_n$ and let $P_\theta, \theta \in T = T_n \subset \mathbb{R}^d$ be a log-concave exponential family. Let $\Theta = \Theta_n \subset T$ with $\mathrm{Diam}(\Theta) \lesssim n^A$ for some $A > 0$. Let $\delta > 0$ and let $s = k + 1 + \rho$ with $k \geq 0$ and $\rho \in (0, 1]$. Suppose that $\Theta_\delta \subset T$ and conditions (3.7) hold. Suppose that, for some $\alpha \in (0,1)$, $d \lesssim n^\alpha$ and assume that $s > \frac{1}{1-\alpha}$. Let θ_0 in the definition of $\hat{\theta}$ be a point from Θ. Then asymptotic relationships (3.9) and (3.10) hold for estimator $f_k(\hat{\theta})$ of $f(\theta)$.*

REMARK 3.9. Note that, in the case of exponential model, the limit variance $\sigma_{f\circ\Psi^{-1}}(\Psi(\theta))$ in Proposition 3.2 is equal to $\langle \mathcal{I}^{-1}(\vartheta)(f\circ\Psi^{-1})'(\vartheta), (f\circ\Psi^{-1})'(\vartheta)\rangle$ with $\vartheta = \Psi(\theta)$. It is possible to prove local minimax lower bounds showing optimality of this variance and the asymptotic efficiency of estimator of $f_k(\hat{\theta})$ (for instance, using Van Trees inequality [25], see [36], [40] for similar results).

REMARK 3.10. The result of Corollary 3.2 also holds under more general assumption that function h in the definition of exponential model (3.11) satisfies the condition $c^{-1}g(x) \leq h(x) \leq cg(x), x \in \mathbb{R}^d$ for a non-negative log-concave function g and for a constant $c \geq 1$.

REMARK 3.11. It was shown in [57], Theorem 3.1 that, under some moment assumptions on d-dimensional exponential families with MLE $\hat{\theta}$, $\hat{\theta} - \theta$ could be approximated by a sample mean with accuracy $O_{\mathbb{P}}(\frac{d}{n})$. Together with a high-dimensional CLT proved in [56], this implies that normal approximation of $\sqrt{n}(\hat{\theta} - \theta)$ holds if $d = o(\sqrt{n})$. It was also shown in [57], Proposition 3.1 that, if d is larger than $\sqrt{n}$, the normal approximation of $\sqrt{n}(\hat{\theta} - \theta)$ could fail even for linear functionals. Thus, additional conditions on exponential family (for instance, shape constraints such as log-concavity) are needed to justify normal approximation for MLE when $d \geq \sqrt{n}$ (which is an interesting regime for functional estimation requiring the bias reduction).

4. Outline of the proofs: bootstrap chains and random homotopies. Let $\hat{\theta}^{(k)}, k \geq 0$[4] be the Markov chain in the space T with transition probability kernel $P(\theta, A), \theta \in T, A \subset T$, defined by (1.2), and with $\hat{\theta}^{(0)} = \theta$. For this chain, $\hat{\theta}^{(1)}$ has the same distribution as $\hat{\theta}$; conditionally on $\hat{\theta}^{(1)}$, $\hat{\theta}^{(2)}$ is sampled from the distribution $P(\hat{\theta}^{(1)}; \cdot)$; conditionally on $\hat{\theta}^{(2)}$, $\hat{\theta}^{(3)}$ is sampled from the distribution $P(\hat{\theta}^{(2)}, \cdot)$, etc. Thus, the Markov chain $\hat{\theta}^{(k)}, k \geq 0$ is constructed by an iterative application of parametric bootstrap to the estimator $\hat{\theta}$ and it was called in [36] *the bootstrap chain* of this estimator. Bootstrap chains are involved in representations of functionals $\mathcal{B}^k f, k \geq 1$ needed to control the bias of estimator $f_k(\hat{\theta})$. Namely

[4] In this section, k denotes the time index of bootstrap chains and is not related to smoothness parameter s unless it is stated otherwise (as in Proposition 4.1).

(see [36, 37, 40, 41]),

$$(\mathcal{B}^k f)(\theta) = \mathbb{E}_\theta \sum_{j=0}^{k} (-1)^{k-j} \binom{k}{j} f(\hat{\theta}^{(j)}),$$

which is the expectation of the k-th order difference of function f along the sample path of the bootstrap chain. It is well known that for a k times continuously differentiable function f in the real line, its k-th order difference

$$\Delta_h^k f(x) = \sum_{j=0}^{k} (-1)^{k-j} \binom{k}{j} f(x + jh) = f^{(k)}(x) h^k + o(h^k) \text{ as } h \to 0.$$

If, for a small $\delta > 0$, $\sup_{\theta \in T} \mathbb{P}_\theta \{\|\hat{\theta} - \theta\| \geq \delta\}$ is also small, we would have that $\|\hat{\theta}^{(j+1)} - \hat{\theta}^{(j)}\| < \delta$ with a high probability. In this case, one could expect that, for a k times continuously differentiable function $f : E \mapsto \mathbb{R}$, $(\mathcal{B}^k f)(\theta)$ is of the order δ^k, and, if f is $k+1$ times continuously differentiable function, then the bias of estimator $f_k(\hat{\theta})$ of $f(\theta)$

$$\mathbb{E}_\theta f_k(\hat{\theta}) - f(\theta) = (-1)^k (\mathcal{B}^{k+1} f)(\theta) = O(\delta^{k+1}).$$

This heuristic was justified in [41] (with some ideas developed earlier in [36, 37, 40]) using representations of bootstrap chains as superpositions of so called *random homotopies*.

A random homotopy between parameter θ and its estimator $\hat{\theta}$ is an a.s. continuous stochastic process $H : T \times [0,1] \times \Omega \mapsto T$ defined on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ such that, for all $\theta \in \Theta$,

$$H(\theta; 0) \coloneqq \theta, \; H(\theta; 1) \stackrel{d}{=} \hat{\theta}, \text{ where } \hat{\theta} \sim P(\theta; \cdot).$$

In addition, random homotopy $H(\theta, t), \theta \in T, t \in [0,1]$ will be assumed to be sufficiently smooth. In other words, random homotopy is a coupling that provides a smooth path between parameter θ and a random variable in the parameter space with the same distribution as the estimator $\hat{\theta}$. Given i.i.d. copies $H_1, H_2, \ldots$, one can define their superpositions $G_k := H_k \bullet \cdots \bullet H_1$ as follows:

$$G_k(\theta; t_1, \ldots, t_k) := H_k(G_{k-1}(\theta; t_1, \ldots, t_{k-1}), t_k), (t_1, \ldots, t_k) \in [0,1]^k$$

with $G_0 \equiv \theta$. One can also define a Markov chain $\tilde{\theta}^{(k)} := G_k(\theta; 1, \ldots, 1)$ with $\tilde{\theta}^{(0)} = \theta$ and show that $(\hat{\theta}^{(k)} : k \geq 0) \stackrel{d}{=} (\tilde{\theta}^{(k)} : k \geq 0)$, see Lemma 4.1 in [41]. Moreover, it is also shown in the same lemma that

$$\hat{\theta}_l \stackrel{d}{=} G_k(\theta; t_1, \ldots, t_k), (t_1, \ldots, t_k) \in \{0,1\}^k, \sum_{i=1}^{k} t_i = l.$$

Using these facts, it is easy to derive the following representation of $(\mathcal{B}^k f)(\theta)$

$$(\mathcal{B}^k f)(\theta) = \mathbb{E} \Delta^{(1)} \ldots \Delta^{(k)} f(G_k(\theta; t_1, \ldots, t_k)),$$

where

$$\Delta^{(i)} \varphi(t_1, \ldots, t_k) = \varphi(t_1, \ldots, t_i, \ldots, t_k)_{|t_i = 1} - \varphi(t_1, \ldots, t_i, \ldots, t_k)_{|t_i = 0}, i = 1, \ldots, k.$$

Under proper smoothness assumptions on f and on random homotopies, this yields the following formula:

$$(\mathcal{B}^k f)(\theta) = \int_0^1 \cdots \int_0^1 \mathbb{E} \frac{\partial^k f(G_k(\theta; t_1, \ldots, t_k))}{\partial t_1 \ldots \partial t_k} dt_1 \ldots dt_k,$$

This approach and other analytic techniques developed in [41] led to the bounds on the Hölder norms of functions $\mathcal{B}^j f$ and f_k as well as the bounds on the bias of estimator $f_k(\hat{\theta})$ of $f(\theta)$.

For a function $V : T \times [0,1] \mapsto F$ with values in a Banach space F and such that $V(\cdot,t) \in C^s(T), t \in [0,1]$, denote

$$\|V\|_{\widetilde{C}^s(T\times[0,1])} := \sup_{t\in[0,1]} \|V(\cdot,t)\|_{C^s(T)} \text{ and } \|V\|_{\widetilde{C}^{0,s}(T\times[0,1])} := \sup_{t\in[0,1]} \|V(\cdot,t)\|_{C^{0,s}(T)}.$$

We will summarize some facts proved in [41] (see, in particular, Theorem 3.1, Theorem 3.2, Proposition 7.1).

PROPOSITION 4.1. *Let $s = k+1+\rho$, $k \geq 1, \rho \in (0,1]$. Assume that $H(\theta;t)$ is $k+1$ times continuously differentiable in $T \times [0,1]$ and let $\dot{H}(\theta;t) := \frac{d}{dt}H(\theta;t)$. Then, the following statements hold:*

1. *If*

$$\mathbb{E}(\|H\|^{s-1}_{\widetilde{C}^{0,s-1}(T\times[0,1])})\|\dot{H}\|_{\widetilde{C}^{s-1}(T\times[0,1])} < +\infty, \tag{4.1}$$

then

$$\|\mathcal{B}\|_{C^s(T)\mapsto C^{s-1}(T)} \leq 4(k+1)^{k+2}\mathbb{E}(\|H\|^{s-1}_{\widetilde{C}^{0,s-1}(T\times[0,1])})\|\dot{H}\|_{\widetilde{C}^{s-1}(T\times[0,1])}.$$

2. *Moreover, under the same assumption, for some constant D_s and for all $j = 1,\dots,k$,*

$$\|\mathcal{B}^j\|_{C^s(T)\mapsto C^{1+\rho}(T)} \leq D_s\Big(\mathbb{E}(\|H\|^{s-1}_{\widetilde{C}^{0,s-1}(T\times[0,1])})\|\dot{H}\|_{\widetilde{C}^{s-1}(T\times[0,1])}\Big)^j.$$

3. *If $D_s\mathbb{E}(\|H\|^{s-1}_{\widetilde{C}^{0,s-1}(T\times[0,1])})\|\dot{H}\|_{\widetilde{C}^{s-1}(T\times[0,1])} \leq 1/2$, then*

$$\|f_k\|_{C^{1+\rho}(T)} \leq 2\|f\|_{C^s(T)}.$$

4. *If assumption (4.1) holds, then for all $\theta \in T$,*

$$|\mathbb{E}_\theta f_k(\hat{\theta}) - f(\theta)| \lesssim_s \|f\|_{C^s(T)}\Big(\mathbb{E}(\|H\|^{s-1}_{\widetilde{C}^{0,s-1}(T\times[0,1])})\|\dot{H}\|_{\widetilde{C}^{s-1}(T\times[0,1])}\Big)^k$$
$$\times\Big(\Big\|\mathbb{E}\int_0^1 \dot{H}(\theta;t)dt\Big\| + \mathbb{E}\|\dot{H}\|^{1+\rho}_{L_\infty(T\times[0,1])}\Big).$$

These facts provide a way to control the bias of estimator $f_k(\hat{\theta})$ and, using the smoothness of function f_k, to study the concentration of $f_k(\hat{\theta})$ around its expectation (in the case of normal models where Gaussian concentration could be used). However, both the construction of random homotopies and the development of concentration bounds for more general statistical models than Gaussian are challenging problems.

We get around this difficulty by using the normal approximation of estimator $\hat{\theta}$ and reducing the problem to the Gaussian case. More precisely, instead of developing random homotopies directly for the estimator $\hat{\theta}$, we use a very simple random homotopy

$$H(\theta;t) := \theta + \frac{t\xi(\theta)}{\sqrt{n}}$$

for the "estimator" $\tilde{\theta} = G(\theta) = \theta + \frac{\xi(\theta)}{\sqrt{n}}$, or for a slightly modified "estimator" $\tilde{\theta}_\delta$, defined as follows:

$$\tilde{\theta}_\delta := G_\delta(\theta) := \theta + \frac{\xi_\delta(\theta)}{\sqrt{n}} \in \Theta_\delta,$$

where $\xi_\delta(\theta) := \xi(\theta) I\Big(\|\xi\|_{L_\infty(E)} < \delta\sqrt{n}\Big), \theta \in E$. Using $\tilde\theta_\delta$ instead of $\tilde\theta$ would allow us to prove our results under smoothness assumptions on the process ξ and functional f locally in a neighborhood of Θ (if ξ and f are smooth in the whole space, using $\tilde\theta$ would suffice). For these estimators, we construct the corresponding bootstrap chains $\tilde\theta^{(k)}$ and $\tilde\theta_\delta^{(k)}$, defined as superpositions

$$\tilde\theta^{(k)} = (G_k \circ \cdots \circ G_1)(\theta) \text{ and } \tilde\theta_\delta^{(k)} = (G_{k,\delta} \circ \cdots \circ G_{1,\delta})(\theta)$$

of i.i.d. copies $G_1, G_2, \dots$ and $G_{1,\delta}, G_{2,\delta}, \dots$ of stochastic processes G and G_δ. We show that these chains approximate in distribution the bootstrap chain $\hat\theta^{(k)}$ of the initial estimator $\hat\theta$ (see Theorem 5.1 in Section 5). We also approximate operator $\mathcal{T}$ by the operators $\tilde{\mathcal{T}}$ and $\tilde{\mathcal{T}}_\delta$:

$$(\tilde{\mathcal{T}} f)(\theta) := \mathbb{E}_\theta f(\tilde\theta) = \mathbb{E} f(G(\theta)), \quad (\tilde{\mathcal{T}}_\delta f)(\theta) := \mathbb{E}_\theta f(\tilde\theta_\delta) = \mathbb{E} f(G_\delta(\theta)),\ \theta \in E, f \in \mathrm{Lip}(E)$$

and define $\tilde{\mathcal{B}} := \tilde{\mathcal{T}} - \mathcal{I}$, $\tilde{\mathcal{B}}_\delta := \tilde{\mathcal{T}}_\delta - \mathcal{I}$. This allows us to approximate the function f_k by similar functions $\tilde f, \tilde f_{\delta,k}$ defined as follows

$$\tilde f_k(\theta) := \sum_{j=0}^k (-1)^j (\tilde{\mathcal{B}}^j f)(\theta), \tilde f_{\delta,k}(\theta) := \sum_{j=0}^k (-1)^j (\tilde{\mathcal{B}}_\delta^j f)(\theta).$$

Namely, we prove the following bound (see Theorem 5.2, [38]): for all $s \ge 1, k \ge 1$ and $\delta > 0$ such that $\Theta_{k\delta} \subset T$,

$$\|f_k - \tilde f_{\delta,k}\|_{L_\infty(\Theta)}$$
$$\lesssim_{s,k} \|f\|_{C^s(\Theta_{k\delta})} \left(1 + \frac{\mathbb{E}\|\xi\|^s_{C^s(\Theta_{(k-1)\delta})}}{n^{s/2}}\right)^{k-1} \Big[\Delta_{\mathcal{F},\Theta}(\eta_1,\eta_2), (\hat\theta,\tilde\theta) + \mathfrak{Q}_n(\Theta_{(k-1)\delta}; \delta)\Big],$$

where $\mathcal{F} := \{f : \|f\|_{C^s(\Theta_\delta)} \le 1\}$ and

$$\mathfrak{Q}_n(\Theta,\delta) := \sup_{\theta\in\Theta} \mathbb{P}\{\|\hat\theta - \theta\| \ge \delta\} + \mathbb{P}\{\|\xi\|_{L_\infty(E)} \ge \delta\sqrt{n}\}.$$

Proposition 4.1 allows us to control the bias $\mathbb{E}\tilde f_{\delta,k}(\tilde\theta_\delta) - f(\theta)$. Moreover, in Section 6, [38], we use Gaussian concentration (more precisely, Maurey-Pisier type inequalities) to obtain bounds on the error of "estimator" $f_{\delta,k}(\theta_\delta)$. This yields the following inequaity (see Theorem 6.1, [38]) that holds, for $s = k + 1 + \rho, k \ge 1, \rho \in (0,1]$, under the assumption that $\Theta_{(k+3)\delta} \subset T$:

$$\sup_{\theta\in\Theta} \Big\| \tilde f_{\delta,k}(\tilde\theta_\delta) - f(\theta) - n^{-1/2}\langle f'(\theta), \xi(\theta)\rangle \Big\|_{L_\psi(\mathbb{P})}$$
$$\lesssim_{s,\psi} \|f\|_{C^s(\Theta_{(k+3)\delta})} \Bigg[\left(\sqrt{\frac{\mathfrak{d}_\xi(\Theta_{(k+2)\delta}; s-1)}{n}}\right)^s + \frac{\|\Sigma\|^{1/2}_{L_\infty(E)}}{n^{1/2}} \sqrt{\frac{\mathfrak{d}_\xi(\Theta_{(k+2)\delta}; s-1)}{n}}$$
$$+ \sqrt{\frac{\mathfrak{d}_\xi(\Theta_{(k+2)\delta}; s-1)}{n}} \tilde\psi^{1/2}(\mathbb{P}\{\|\xi\|_{L_\infty(E)} \ge \delta\sqrt{n}\}) \Bigg].$$

We combine all these pieces together in Section 7, [38] to complete the proofs of main results.

REMARK 4.1. It easily follows from the proofs of the main results that they also hold for estimators $\tilde f_k(\hat\theta)$ and $\tilde f_{\delta,k}(\hat\theta)$, based on the functionals related to the Gaussian approximation of estimator $\hat\theta$.

Acknowledgment. The author is very thankful to Clément Deslandes for careful reading of the manuscript and suggesting a number of corrections and to the referees for helpful comments.

SUPPLEMENTARY MATERIAL

Supplement to "Estimation of smooth functionals in high-dimensional models: bootstrap chains and Gaussian approximation"

In the supplementary material [38], we develop all the necessary tools and provide the detailed proofs of the main results. In particular, we develop a method of approximation of bootstrap chains by the Markov chains for Gaussian model and prove concentration bounds for this model. We also state and prove some additional results.

REFERENCES

[1] A. Anastasiou. Assessing the multivariate normal approximation of the maximum likelihood estimator from high-dimensional, heterogeneous data. *Electronic J. of Statistics*, 2018, 12, 2, 3794–3828.

[2] A. Anastasiou and R. Gaunt. Wasserstein distance error bounds for the multivariate normal approximation of the maximum likelihood estimator. 2020, *arXiv:2005.0520*.

[3] B. Arras and C. Houdré. On Stein's Method for Multivariate Self-Decomposable Laws. *Electron. J. Probab.*, 2019, 24, 128, 1–63.

[4] V. Bentkus, M. Bloznelis and F. Götze. A Berry-Esseen Bound for M-estimators. *Scandinavian Journal of Statistics*, 1997, 24, 4, 485–502.

[5] P. Bickel and Y. Ritov. Estimating integrated square density derivatives: sharp best order of convergence estimates. *Sankhya*, 1988, 50, 381–393.

[6] P.J. Bickel, C.A.J. Klaassen, Y. Ritov and J.A. Wellner. *Efficient and Adaptive Estimation for Semiparametric Models*. Johns Hopkins University Press, Baltimore, 1993.

[7] L. Birgé and P. Massart. Estimation of integral functionals of a density. *Annals of Statistics*, 1995, 23, 11-29.

[8] R.N. Bhattacharya and R. Ranga Rao. *Normal Approximation and Asymptotic Expansions*. John Willey &Sons, New York, 1976.

[9] S. Bobkov and M. Ledoux. From Brunn-Minkowski to Brascamp-Lieb and to logarithmic Sobolev inequalities. *Geom. Funct. Anal.*, 2000, 10(5), 1028–1052.

[10] T.T. Cai and M. Low. On adaptive estimation of linear functionals. *Annals of Statistics*, 2005, 33, 2311–2343.

[11] T.T. Cai and M. Low. Non-quadratic estimators of a quadratic functional. *Annals of Statistics*, 2005, 33, 2930–2956.

[12] N.N. Čencov. *Statistical decision rules and optimal inference.* American Mathematical Society, 1982.

[13] Y. Chen. An Almost Constant Lower Bound of the Isoperimetric Coefficient in the KLS Conjecture. *Geometric and Functional Analysis*, 2021, 31, 34–61.

[14] V. Chernozhukov, D. Chetverikov and K. Kato. Central limit theorems and bootstrap in high dimensions. *Annals of Probability*, 2017, 45, 4, 2309–2352.

[15] V. Chernozhukov, D. Chetverikov and Y. Koike. Nearly optimal central limit theorem and bootstrap approximation in high dimensions. *arXiv: 2012.09513*

[16] V. Chernozhukov, D. Chetverikov, K. Kato and Y. Koike. Improved central limit theorem and bootstrap approximation in high dimensions. *Annals of Statistics*, 2022, to appear.

[17] 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.

[18] T.A. Courtade, M. Fathi and A. Pananjadi. Existence of Stein Kernels under a Spectral Gap, and Discrepancy Bounds. *Ann. Inst. H. Poincaré,* 2019, 55, 2, 777–790.

[19] D. Donoho and R. Liu. On minimax estimation of linear functionals. Technical Report N 105. Department of Statistics, UC Berkeley, August 1987.

[20] D. Donoho and R. Liu. Geometrizing rates of convergence, II. *Annals of Statistics*, 1991, 19, 2, 633-667.

[21] D. Donoho and M. Nussbaum. Minimax quadratic estimation of a quadratic functional. *J. Complexity*, 1990, 6, 290–323.

[22] R. Eldan. Thin shell implies spectral gap up to polylog via a stochastic localization scheme. *Geometric and Functional Analysis*, 2013, 23, 2, 532–569.

[23] R. Eldan, D. Mikulincer and A. Zhai. The CLT in high dimensions: quantitative bounds via martingale embedding. *Annals of Probability*, 2020, 48, 5, 2494–2524.

[24] M. Fathi. Higher order Stein Kernels for Gaussian approximation. *Studia Mathematica*, 2019, 256, 241–258.

[25] R.D. Gill and B.Y. Levit. Applications of the van Trees inequality: a Bayesian Cramér-Rao bound. *Bernoulli,* 1995, 1(1-2), 59–79.

[26] V.L. Girko. Introduction to general statistical analysis. *Theory Probab. Appl.*, 1987, 32, 2: 229–242.

[27] V.L. Girko. *Statistical analysis of observations of increasing dimension*. Springer, 1995.

[28] Y. Han, J. Jiao and R. Mukherjee. On estimation of L_r-norms in Gaussian white noise model. *Probability Theory and Related Fields*, 2020, 177, 1243–1294.

[29] P. Hall. *The Bootstrap and Edgeworth Expansion.* Springer-Verlag, New York, 1992.

[30] P. Hall and M.A. Martin. On Bootstrap Resampling and Iteration. *Biometrika*, 1988, 75, 4, 661–671.

[31] I. A. Ibragimov and R.Z. Khasminskii. *Statistical Estimation: Asymptotic Theory*. Springer-Verlag, New York, 1981.

[32] 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.

[33] J. Jiao and Y. Han. Bias correction with Jackknife, Bootstrap and Taylor Series. *IEEE Transactions on Information Theory*, 2020, 66, 7, 4392–4418.

[34] J. Klemelä. Sharp adaptive estimation of quadratic functionals. *Probability Theory and Related Fields*, 2006, 134, 539–564.

[35] Y. Koike. Notes on the dimension dependence in high-dimensional central limit theorems for hyperrectangles. *Japanese J. of Statistics and Data Science*, 2021, 4(1), 257–297.

[36] V. Koltchinskii. Asymptotically Efficient Estimation of Smooth Functionals of Covariance Operators. *J. European Mathematical Society*, 2021, 23, 3, 765–843.

[37] V. Koltchinskii. Asymptotic Efficiency in High-Dimensional Covariance Estimation. *Proc. ICM 2018*, Rio de Janeiro, 2018, vol. 3, 2891–2912.

[38] V. Koltchinskii. Supplement to "Estimation of smooth functionals in high-dimensional models: bootstrap chains and Gaussian approximation", 2021.

[39] V. Koltchinskii, M. Löffler and R. Nickl. Efficient Estimation of Linear Functionals of Principal Components. *Annals of Statistics*, 2020, 48, 1, 464–490.

[40] V. Koltchinskii and M. Zhilova. Efficient estimation of smooth functionals in Gaussian shift models. *Ann. Inst. H. Poincarè - Probab. et Statist.*, 2021, 57, 1, 351–386.

[41] V. Koltchinskii and M. Zhilova. Estimation of Smooth Functionals in Normal Models: Bias Reduction and Asymptotic Efficiency. *Annals of Statistics*, 2021, to appear. *arXiv:1912.08877*.

[42] V. Koltchinskii and M. Zhilova. Estimation of smooth functionals of location parameter in Gaussian and Poincaré random shift models. *Sankhya*, 2021, v. 83, issue 2, no. 4, 569–596.

[43] A.K Kuchibhotla, S. Mukherjea and D. Banerjee. High-dimensional CLT: Improvements, Non-uniform Extensions and Large Deviations. *Bernoulli*, 2021, 27, 1, 192–217.

[44] B. Laurent. Efficient estimation of integral functionals of a density. *Annals of Statistics*, 1996, 24, 659–681.

[45] Y.-T. Lee and S. Vempala. Eldan's Stochastic Localization and the KLS Hyperplane Conjecture: An Improved Lower Bound for Expansion. *58th Annual IEEE Symposium on Foundations of Computer Science FOCM 2017*.

[46] B. Levit. On the efficiency of a class of non-parametric estimates. *Theory of Prob. and applications*, 1975, 20(4), 723–740.

[47] B. Levit. Asymptotically efficient estimation of nonlinear functionals. *Probl. Peredachi Inf. (Problems of Information Transmission)*, 1978, 14(3), 65–72.

[48] 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.

[49] E. Milman. On the role of convexity in isoperimetry, spectral gap and concentration. *Invent. Math.*, 2009, 177(1), 1-43.

[50] R. Mukherjee, W. Newey and J. Robins. Semiparametric Efficient Empirical Higher Order Influence Function Estimators. 2017, *arXiv:1705.07577*.

[51] 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.

[52] 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.

[53] V. Paulauskas and A. Rachkauskas. *Approximation theory in central limit theorems. Exact results in Banach spaces*. Kluwer Academic Publishers, 1989.

[54] J. Pfanzagl. The Berry-Esseen bound for minimum contrast estimates. *Metrika*, 1971, 17, 82–91.

[55] I. Pinelis. Optimal-order uniform and nonuniform bounds on the rate of convergence to normality for maximum likelihood estimators. *Electronic Journal of Statistics*, 2017, 11, 1160–1179.

[56] S. Portnoy. On the central limit theorem in $\mathbb{R}^p$ when $p \to \infty$. *Probability Theory and Related Fields*, 1986, 73, 581–583.

[57] S. Portnoy. Asymptotic behavior of likelihood methods for exponential families when the number of parameters tends to infinity. *Annals of Statistics*, 1988, 16, 1, 356–366.

[58] E. Rio. Upper bounds for minimal distances in the central limit theorem. *Ann. Inst. H. Poincaré - Probab. et Statist.*, 2009, 45, 3, 802–817.

[59] 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.

[60] 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.

[61] V. Senatov. *Normal Approximation: New Results, Methods and Problems*. VSP, Utrecht, The Netherlands, 1998.

[62] A. van der Vaart. Higher order tangent spaces and influence functions. *Statistical Science*, 2014, 29, 4, 679–686.

[63] C. Villani. *Optimal Transport. Old and New*. Springer, 2009.

[64] V. M. Zolotarev. Metric distances in spaces of random variables and their distributions. *Mat. Sb. (N.S.)*, 1976, 101(143), 3(11), 416–454.