Electronic Journal of Statistics
Vol. 17 (2023) 1457–1491
ISSN: 1935-7524
https://doi.org/10.1214/23-EJS2132

Bounds in L^1 Wasserstein distance on the normal approximation of general M-estimators*

François Bachoc†

Institut de Mathématiques de Toulouse; UMR5219. Université de Toulouse; CNRS. UT3, F-31062 Toulouse, France. e-mail: francois.bachoc@math.univ-toulouse.fr

Max Fathi

Université Paris Cité and Sorbonne Université, CNRS, Laboratoire Jacques Louis Lions & Laboratoire de Probabilités Statistique et Modélisation, F-75013 Paris, France. e-mail: mfathi@lpsm.paris

Abstract: We derive quantitative bounds on the rate of convergence in L^1 Wasserstein distance of general M-estimators, with an almost sharp (up to a logarithmic term) behavior in the number of observations. We focus on situations where the estimator does not have an explicit expression as a function of the data. The general method may be applied even in situations where the observations are not independent. Our main application is a rate of convergence for cross validation estimation of covariance parameters of Gaussian processes.

Keywords and phrases: Asymptotic normality, central limit theorem, Wasserstein distance, parametric estimation, logistic regression, cross validation.

Received January 2022.

1. Introduction

Our goal here is to derive quantitative bounds for approximate normality of parameter estimators that arise as minimizers of certain random functions. The main example to keep in mind is maximum likelihood estimation [56, Chapter 5.5], but other problems fit in the framework we shall consider, including least square estimators [52] and cross validation [8, 61].

Consider a fixed compact parameter space $\Theta \subset \mathbb{R}^p$ and a sequence of random functions $(M_n)_{n\in\mathbb{N}}$, where for $n \in \mathbb{N}$, $M_n : \Theta \to \mathbb{R}$. Throughout, $\mathbb{N}$ is the set of non-zero natural numbers. The variable n should be thought of as a sample size, and M_n the function for which a minimizer will be the M-estimator of interest, which is a (measurable) random vector $\hat{\theta}_n \in \Theta$ such that

$$\hat{\theta}_n \in \operatorname*{argmin}_{\theta\in\Theta} M_n(\theta). \tag{1}$$

*This work was supported by the Projects MESA (ANR-18-CE40-006) and GAP (ANR-21-CE40-0007) of the French National Research Agency (ANR).

†Corresponding author.

A classical family of M-estimators is given by functions of the form

$$M_n(\theta) = \frac{1}{n}\sum_{i=1}^{n} \rho(\theta, X_i) \tag{2}$$

where the X_i are the sample independent data, valued in a space $\mathcal{X}$, and $\rho : \Theta \times \mathcal{X} \to \mathbb{R}$ is a fixed function. We shall address in details this class in Sections 3.1 and 3.2, but investigation shall go beyond this framework, in particular to cover covariance estimation for Gaussian processes, addressed in Section 3.3.

Our goal will be to derive quantitative central limit theorems in L^1 Wasserstein (or optimal transport) distance for the fluctuations of $\hat{\theta}_n$ around a deterministic parameter $\theta_{0,n}$ (that is allowed to depend on n). The simplest example is when $\theta_{0,n} = \theta_0$ is fixed, typically when M_n stems from the likelihood function and there is a fixed data generating process characterized by the "true" parameter θ_0 [56, Chapter 5.5]. Nevertheless, we allow for a sample-size dependent $\theta_{0,n}$ which enables to address relevant situations such as misspecified models [14, 17, 36, 59]. In particular, in [14, 17], the parameter of interest $\theta_{0,n}$ that $\hat{\theta}_n$ estimates explicitly depends on sample size.

In the context of this paper, it is typically already known that the distribution of $n^{1/2}(\hat{\theta}_n - \theta_{0,n})$ converges to a Gaussian distribution. General techniques for showing this convergence are available in a wealth of contributions, see for instance [20, 51, 56] and references therein. Our goal is then to go beyond the convergence between these two distributions (for which, usually, no rates are available) by providing quantitative bounds on their L^1 Wasserstein distance. In this view, the main challenge is the M-estimation setting, which often entails that no explicit expression of $\hat{\theta}_n$ is available. Our main abstract result, Theorem 2, is a general statement about reducing the problem to a central limit theorem for an explicit function of the data. More precisely, the L^1 Wasserstein distance between the distribution of $n^{1/2}(\hat{\theta}_n - \theta_{0,n})$ and a Gaussian distribution is bounded by the sum of a term of order $O((\log n)n^{-1/2})$ and the distance between a Gaussian distribution and the normalized gradient of M_n at $\theta_{0,n}$.

Hence, Theorem 2 reduces the problem to quantifying the asymptotic normality of this normalized gradient. Since this quantity is explicit, there are many techniques in the literature that can be applied. We shall discuss this aspect of the problem in Section 2.3.

We shall illustrate the benefits of Theorem 2 with several examples of functions M_n: averages of independent functions in Section 3.1, maximum likelihood for logistic regression in Section 3.2 and cross validation estimation of covariance parameters of Gaussian processes in Section 3.3. This last example highlights the flexibility of our techniques, since the observations are dependent and the function M_n is not based on the likelihood. In all these three cases, eventually, we provide a bound, for the L^1 Wasserstein distance between the distribution of $n^{1/2}(\hat{\theta}_n - \theta_{0,n})$ and a Gaussian distribution, of order $O((\log n)n^{-1/2})$.

There has been a recent interest for bounding the normal approximation of M-estimators, as we do here. On connected topics, the normal approximation is quantified in [50] for the Delta method and in [3] for gradient descent. Considering now specifically M-estimators, a series of articles successfully addressed them: [1, 2, 4, 5, 6, 7, 15, 49, 54]. These articles address not only the univariate case (for θ) [1, 6, 7, 15, 49], but also the general multivariate one [2, 4, 5, 54]. In particular, some of these references exploit the characterization of the L^1 Wasserstein distance as a supremum of expectation differences, over Lipschitz functions. This enables to decompose the target Wasserstein distance into several terms that can be addressed independently with different approaches. This idea appears for instance in equations (9), (10) and (20) in the reference [1], as well as some of the other articles above. We also rely on it, see (22) and (24) (also, Remark 7 discusses the extension of the results of this paper to general L^p Wasserstein distances, $p > 1$).

We shall now highlight the novelty of our results compared to the above articles. First, the references [2, 4, 6, 7, 15, 49, 54] do not address the L^1 Wasserstein distance as we do. Only [1, 5] do. In [54], the distance is the supremum probability difference over convex sets, which is of the Berry-Esseen type. Earlier and similarly, [15, 49] considered the Kolmogorov distance in the univariate case. Also, [6, 7] address Zolotarev-type distances based on supremums of expectation differences over absolutely continuous bounded test functions (and Lipschitz in [7], yielding the bounded-Wasserstein distance). Similarly, [2, 4] consider test functions that are bounded with bounded derivatives of various orders. Remark that while the L^1 Wasserstein and Kolmogorov distances can be compared under regularity conditions and a priori moment bounds, using general comparison results typically worsens the quantitative estimates. Note also that bounding the L^1 Wasserstein distance is stronger than in [2, 4, 7], as it allows for a larger class of test functions. Remark furthermore that Berry-Esseen-type and Kolmogorov distances may be less sensitive than Wasserstein distances to, for instance, the moments of $\hat{\theta}_n - \theta_{0,n}$. Thus, the Wasserstein distances necessitate specific treatments compared to them (for instance, see the proof and use of Lemma 7 here, or the terms in Theorem 2.1 in [7] involving the moments of $\hat{\theta}_n - \theta_{0,n}$).

In addition, we allow for general functions M_n, while most of the above references focus on maximum likelihood. Some arguments provided for maximum likelihood do carry over to general functions M_n, but it is not clear that this is the case for all of them. Also, most of the above references focus on independent observations (often also identically distributed) defining the function M_n (with the exception of [1]), while we allow for M_n stemming from dependent observations. Again, some but not all arguments for independent observations can be extended to dependent observations. In the case of independent observations, as in [5] we shall rely on a result of Bonis [18] to bound the rate of convergence in the multivariate central limit theorem.

Furthermore, in comparison to [1, 2, 4, 5, 6, 7], our general bound in Theorem 2 only depends on M_n and its derivatives, and does not feature $\hat{\theta}_n - \theta_{0,n}$. In contrast, most of the general bounds in these references contain moments of $\hat{\theta}_n - \theta_{0,n}$ (see for instance Theorem 2.1 in [7]). Hence, our general bound seems more convenient to apply to examples, particularly when $\hat{\theta}_n$ does not have an explicit expression, which is often the case. In agreement with this, in most of the examples provided by [1, 2, 4, 5, 6, 7], $\hat{\theta}_n$ has an explicit expression. As an exception, [2, 7] address maximum likelihood estimation of the shape parameters of the Beta distribution. Finally, [1, 2, 4, 5, 6, 7] usually make the assumption that there is a unique $\hat{\theta}_n$ satisfying (1), while Theorem 2 here holds for any $\hat{\theta}_n$ satisfying (1). In many statistical models of interest, there is no guarantee that M_n has a unique minimizer over Θ, almost surely.

The examples we address are representative of the flexibility of Theorem 2. In particular we address general averages of independent functions in Section 3.1. We treat logistic regression in Section 3.2, with a simple proof once Theorem 2 is established, which illustrates that this theorem is efficient even when $\hat{\theta}_n$ does not have an explicit expression, and is not necessarily unique. Finally, in Section 3.3 we address cross validation estimation of covariance parameters of Gaussian processes. This last example highlights our flexibility to dependent observations and to M_n not stemming from a likelihood and even not being an average of functions of individual observations (most of the discussed references above consider these averages of functions for M_n). Again, $\hat{\theta}_n$ has no explicit expression in this cross validation example.

Note that a price we pay, so to speak, for the wide class of M-estimators we can address, is that our bounds are not exactly of order $O(n^{-1/2})$, but rather of order $O((\log n)n^{-1/2})$. In contrasts, the bounds given in the examples studied in [1, 2, 4, 5, 6, 7] are of order exactly $O(n^{-1/2})$. Remark 3 in Section 2 discusses the obstacles, in our setting and compared to [1, 2, 4, 5, 6, 7], for establishing bounds of order exactly $O(n^{-1/2})$.

Similarly, our bonds feature non-explicit constants (that do not depend on n but typically depend, for instance, on p). On the contrary, the bounds given in the examples studied in [1, 2, 4, 5, 6, 7] are fully explicit. In particular, they can be computed numerically, which can be beneficial in applications. Remark 4 provides further discussion on this point.

The rest of the paper is organized as follows. Section 2 provides the general technical conditions and the general bound of Theorem 2, reducing the problem to the asymptotic normality of the normalized gradient. It also discusses many references to address this asymptotic normality in the probabilistic literature. Section 3 addresses the three examples discussed above. Some of the proofs are postponed to the appendix.

2. General bounds

For an $\ell \times \ell$ matrix A, we write $\rho_\ell(A) \leq \cdots \leq \rho_1(A)$ for its singular values, and for a symmetric matrix, we write $\lambda_\ell(A) \leq \cdots \leq \lambda_1(A)$ for its eigenvalues.

2.1. Technical conditions

For $u, v \in \mathbb{R}^p$, $u \neq v$, we write $[u, v] = \{tu + (1-t)v; t \in [0, 1]\}$ and $(u, v) = \{tu + (1-t)v; t \in (0, 1)\}$. We also write $[u, u] = \{u\}$ and $(u, u) = \varnothing$. We write $\mathring{\Theta}$

for the interior of the parameter space Θ. The next condition means that Θ is, so to speak, well-behaved. It can be checked that this condition holds for most common compact parameter spaces, in particular hypercubes, balls, ellipsoids and polyhedral sets. Typically we expect Condition 1 not to be restrictive in practice.

Condition 1. There exist two constants $0 < C_\Theta < \infty$ and $0 < c'_\Theta < \infty$ such that for each $0 < \epsilon \leq c'_\Theta$, there exist $N \leq C_\theta \epsilon^{-p}$ and $\theta_1, \dots, \theta_N \in \mathring{\Theta}$ satisfying the following. For each $\theta \in \Theta$, there exists $i \in \{1, \dots, N\}$ such that $(\theta, \theta_i) \subseteq \mathring{\Theta}$ and $||\theta - \theta_i|| \leq \epsilon$.

Then, the next condition basically consists in asking for enough integrability on the derivatives of M_n to be able to commute expectation and derivation, which is usually established using the dominated convergence theorem. Remark that the conditions on the first two derivative orders will actually be implied by some of our later conditions, but we state them here independently for convenience of writing.

Condition 2. Consider $n \in \mathbb{N}$. For $\theta \in \Theta$, the random variable $M_n(\theta)$ is absolutely summable. Almost surely, the function M_n is three times differentiable on $\mathring{\Theta}$. For $i, j, k \in \{1, \dots, p\}$ and $\theta \in \mathring{\Theta}$, the random variables $\partial M_n(\theta)/\partial\theta_i$, $\partial^2 M_n(\theta)/\partial\theta_i\partial\theta_j$ and $\partial^3 M_n(\theta)/\partial\theta_i\partial\theta_j\partial\theta_k$ are absolutely summable. Furthermore,

$$\mathbb{E}\left(\frac{\partial M_n(\theta)}{\partial\theta_i}\right) = \frac{\partial \mathbb{E}(M_n(\theta))}{\partial\theta_i}, \qquad \mathbb{E}\left(\frac{\partial^2 M_n(\theta)}{\partial\theta_i\partial\theta_j}\right) = \frac{\partial^2 \mathbb{E}(M_n(\theta))}{\partial\theta_i\partial\theta_j}$$

and

$$\mathbb{E}\left(\frac{\partial^3 M_n(\theta)}{\partial\theta_i\partial\theta_j\partial\theta_k}\right) = \frac{\partial^3 \mathbb{E}(M_n(\theta))}{\partial\theta_i\partial\theta_j\partial\theta_k}.$$

Note that assuming that M_n is almost surely differentiable is also done in many of the references discussed in the introduction. This assumption is indeed satisfied in many practical cases. Nevertheless, this assumption does exclude some cases, among which, importantly, the median estimator in dimension one (defined as a minimizer of sums of absolute values) and estimators stemming from L^1 penalizations, for instance the lasso [55]. Asymptotic normality results exist for non-differentiable functions M_n, see for instance [56, Theorem 5.21]. In future work, providing quantitative bounds on the Wasserstein distance for these asymptotic normality results would definitely be relevant.

The next condition means that, for a fixed θ, $M_n(\theta)$ and $\partial M_n(\theta)/\partial\theta_i$, $i \in \{1, \dots, p\}$, concentrate around their expectations at rate $n^{-1/2}$, with an exponential decay for deviations of order larger than $n^{-1/2}$. Many tools from concentration inequalities (for instance [19, 23]) enable to check this condition in specific settings (see for instance those of Section 3). The rate n in the exponential is sharp in general for averages of i.i.d. random variables. For a function $f : \mathring{\Theta} \to \mathbb{R}$ and for $\theta \in \mathring{\Theta}$, we write $\nabla f(\theta)$ the gradient column vector of f at θ and we write $\nabla^2 f(\theta)$ the Hessian matrix of f at θ.

Condition 3. There are constants $0 < c_M < \infty$, $0 < c'_M < \infty$ and $0 < C_M < \infty$ such that for $n \in \mathbb{N}$ and $0 < \epsilon \leq c'_M$,

$$\sup_{\theta \in \Theta} \mathbb{P}(|M_n(\theta) - \mathbb{E}(M_n(\theta))| \geq \epsilon) \leq C_M e^{-n c_M \epsilon^2}$$

and

$$\sup_{\theta \in \mathring{\Theta}} \mathbb{P}(||\nabla M_n(\theta) - \mathbb{E}(\nabla M_n(\theta))|| \geq \epsilon) \leq C_M e^{-n c_M \epsilon^2}.$$

We expect Condition 3 to hold in many practical situations, except perhaps for instance when considering long range dependence in time series [16].

The next condition is a control on the deviations of the derivatives of M_n of order 1 and 2, that is uniform over $\mathring{\Theta}$. Remark that the deviations that are controlled are of larger order than those in Condition 3. Hence, again, the condition can be checked in many settings.

Condition 4. There are constants $0 < c_{d,1} < \infty$, $0 < C_{d,1} < \infty$ and $0 < C'_{d,1} < \infty$ such that for $n \in \mathbb{N}$ and $K \geq C'_{d,1}$,

$$\mathbb{P}\left(\sup_{\theta \in \mathring{\Theta}} ||\nabla M_n(\theta)|| \geq K \right) \leq C_{d,1} n e^{-c_{d,1} K}$$

and

$$\mathbb{P}\left(\sup_{\theta \in \mathring{\Theta}} \max_{i,j=1}^{p} \left| \frac{\partial^2 M_n(\theta)}{\partial \theta_i \partial \theta_j} \right| \geq K \right) \leq C_{d,1} n e^{-c_{d,1} K}.$$

We then require the derivatives of order 1, 2 and 3 of M_n to have certain bounded moments, respectively of order 1, 1 and 2.

Condition 5. There is a constant $C_{d,2}$ such that for $n \in \mathbb{N}$,

$$\sup_{\theta \in \mathring{\Theta}} \mathbb{E}\left(||\nabla M_n(\theta)|| \right) \leq C_{d,2}, \quad \sup_{\theta \in \mathring{\Theta}} \max_{i,j=1}^{p} \mathbb{E}\left(\left| \frac{\partial^2 M_n(\theta)}{\partial \theta_i \partial \theta_j} \right| \right) \leq C_{d,2} \tag{3}$$

and

$$\max_{j,k,\ell=1}^{p} \mathbb{E}\left(\sup_{\theta \in \mathring{\Theta}} \left| \frac{\partial^3 M_n(\theta)}{\partial \theta_j \partial \theta_k \partial \theta_\ell} \right|^2 \right) \leq C_{d,2}. \tag{4}$$

Above, the moments are for fixed θ for the order 1 and 2. The moments for the order 3 are uniform over $\mathring{\Theta}$. Note that it can be seen from the proof of Theorem 2 that assuming uniformity only locally around $\theta_{0,n}$ (see Condition 7) would be sufficient. For instance, [4] has a similar locally uniform moment bound on the third-derivatives of the log-likelihood function (see (R.C.3) there). Overall, for applications where M_n is already assumed to be three-times differentiable, Condition 5 is arguably not overly restrictive.

The next condition requires the variances of the derivatives of order 1 and 2 of M_n to be of order $1/n$. This condition is natural and easy to check in many settings, for example for i.i.d. random variables. Nevertheless, similarly as for Condition 3, Condition 6 could exclude some relevant settings, for instance long range dependence in time series.

Condition 6. There is a constant C_{Var} such that for $n \in \mathbb{N}$, $j,k \in \{1,\dots,p\}$,

$$\sup_{\theta \in \mathring{\Theta}} \max_{j=1}^{p} \mathrm{Var}\left(\frac{\partial M_n(\theta)}{\partial \theta_j}\right) \leq \frac{C_{\mathrm{Var}}}{n}$$

and

$$\sup_{\theta \in \mathring{\Theta}} \max_{j,k=1}^{p} \mathrm{Var}\left(\frac{\partial^2 M_n(\theta)}{\partial \theta_j \partial \theta_k}\right) \leq \frac{C_{\mathrm{Var}}}{n}.$$

Remark 1. In Condition 6, it is actually sufficient that the second inequality holds only for $\theta = \theta_{0,n}$. We state Condition 6 as it is only for convenience of writing, and because checking the inequality uniformly over θ in the bounded $\mathring{\Theta}$ usually brings no additional difficulty.

For $x \in \mathbb{R}^p$ and $r \geq 0$, we let $B(x,r)$ be the closed Euclidean ball in $\mathbb{R}^p$ with center x and radius r. The next condition introduces the sequence of deterministic parameters $(\theta_{0,n})_{n\in\mathbb{N}}$, to which $(\hat{\theta}_n)_{n\in\mathbb{N}}$ is asymptotically close. In the applications of Sections 3.2 and 3.3, $\theta_{0,n} = \theta_0$ does not depend on the sample size and determines the fixed unknown data generating process. Nevertheless, it is beneficial to allow for an n-dependent $\theta_{0,n}$, to cover general cases of misspecified models, for instance as in [14, 17, 36, 59].

Condition 7. There exists a sequence $(\theta_{0,n})_{n\in\mathbb{N}}$ and a constant $0 < c_{\theta_0} < \infty$ such that for each $n \in \mathbb{N}$, $B(\theta_{0,n}, c_{\theta_0}) \subseteq \mathring{\Theta}$. Additionally, for each $n \in \mathbb{N}$, $\mathbb{E}(\nabla M_n(\theta_{0,n})) = 0$. Finally, for each $r > 0$ such that $\Theta \backslash B(\theta_{0,n}, r) \neq \varnothing$, there exist constants $N_r \in \mathbb{N}$ and $0 < c_r < \infty$ such that for $n \geq N_r$,

$$\inf_{\substack{\theta \in \Theta \\ ||\theta - \theta_{0,n}|| \geq r}} \left(\mathbb{E}(M_n(\theta)) - \mathbb{E}(M_n(\theta_{0,n}))\right) \geq c_r.$$

Condition 7 is a usual one in M-estimation: $\theta_{0,n}$ cancels out the expected gradient of M_n and is asymptotically the minimizer of $\mathbb{E}(M_n)$, so to speak. In practice, Condition 7 will of course exclude statistical models that are not identifiable, a common example being the estimation of the mean vectors, covariance matrices and proportions in a mixture of Gaussian distributions (even with a known upper bound on the number of classes). We refer for instance to [24, 35, 37] on estimation problems with mixtures. For statistical models that are identifiable, Condition 7 is arguably not overly restrictive in practice.

Then, define the covariance matrix of the normalized gradient

$$\bar{C}_{n,0} = \mathrm{Cov}(\sqrt{n} \nabla M_n(\theta_{0,n})) \tag{5}$$

and the expected Hessian

$$\bar{H}_{n,0} = \mathbb{E}(\nabla^2 M_n(\theta_{0,n})). \tag{6}$$

The next condition requires the expected Hessian matrix of M_n at $\theta_{0,n}$ to be asymptotically strictly positive definite. Similarly to Condition 7, this is a usual requirement for $\theta_{0,n}$ and $\hat{\theta}_n$ to be close at asymptotic rate $n^{-1/2}$.

Condition 8. There are constants $0 < c_{\theta_0,H} < \infty$ and $N_{\theta_0,H} \in \mathbb{N}$ such that for $n \geq N_{\theta_0,H}$

$$\lambda_p(\bar{H}_{n,0}) \geq c_{\theta_0,H}.$$

We finally require the covariance matrix of the normalized gradient to be asymptotically strictly positive definite, so that the Gaussian limit in the central limit theorem is non-degenerate.

Condition 9. There are constants $c_{\theta_0,\nabla} > 0$ and $N_{\theta_0,\nabla} \in \mathbb{N}$ such that for $n \geq N_{\theta_0,\nabla}$,

$$\lambda_p(\bar{C}_{n,0}) \geq c_{\theta_0,\nabla}.$$

In practice, similarly as Condition 7, Conditions 8 and 9 are arguably not overly restrictive for identifiable statistical models, while they will typically not hold for non-identifiable ones.

2.2. Reduction to the normal approximation of the normalized gradient

For a symmetric non-negative definite matrix A, we write $A^{1/2}$ for its unique symmetric non-negative definite square root. When A is also invertible, we write $A^{-1/2} = (A^{1/2})^{-1} = (A^{-1})^{1/2}$.

Consider the normalized gradient $\bar{C}_{n,0}^{-1/2} n^{1/2} \nabla M_n(\theta_{0,n})$. If this normalized gradient (that has identity covariance matrix) converges to a standard Gaussian distribution, then the conditions of Section 2.1 imply that $n^{1/2}(\hat{\theta}_n - \theta_{0,n})$ is asymptotically normally distributed, with asymptotic covariance matrix taking the "sandwich" form $\bar{H}_{n,0}^{-1} \bar{C}_{n,0} \bar{H}_{n,0}^{-1}$. Equivalently, $\bar{C}_{n,0}^{-1/2} \bar{H}_{n,0} n^{1/2}(\hat{\theta}_n - \theta_{0,n})$ converges to a standard Gaussian distribution. We state this result formally as follows.

Theorem 1. *Assume that Conditions 1 to 9 hold. Assume also that*

$$\bar{C}_{n,0}^{-1/2} \sqrt{n} \nabla M_n(\theta_{0,n}) \xrightarrow[n\to\infty]{\mathcal{L}} \mathcal{N}(0, I_p). \tag{7}$$

Consider $\hat{\theta}_n$ as in (1). Then, we have

$$\bar{C}_{n,0}^{-1/2} \bar{H}_{n,0} \sqrt{n}(\hat{\theta}_n - \theta_{0,n}) \xrightarrow[n\to\infty]{\mathcal{L}} \mathcal{N}(0, I_p). \tag{8}$$

Theorem 1 is a direct consequence of (24) going to zero in the proof of Theorem 2 below. Theorem 1 can also be checked without referring to Theorem 2, by using standard arguments. The condition (7) holds in many situations, see in particular the references provided in Section 2.3. Many results similar to (8) are stated in the literature (although typically not using the exact same set of assumptions), with for instance [36] as one of the earliest ones in this vein.

We are interested in the Wasserstein distance between the two distributions in (8). We now introduce this distance. We let $\mathcal{L}_1$ be the set of 1-Lipschitz continuous functions from $\mathbb{R}^p$ to $\mathbb{R}$, that is the set of functions g such that, for all $x_1, x_2 \in \mathbb{R}^p$,

$$|g(x_1) - g(x_2)| \leq ||x_1 - x_2||.$$

Then, for two random vectors U and V in $\mathbb{R}^p$, the L^1 Wasserstein distance between the distributions of U and V is

$$\mathcal{W}_1(U, V) = \sup_{f \in \mathcal{L}_1} |\mathbb{E}(f(U)) - \mathbb{E}(f(V))|.$$

Equivalently, $\mathcal{W}_1(U, V)$ is also the well known L^1 optimal transport cost, according to the Kantorovitch-Rubinstein duality formula:

$$\mathcal{W}_1(U, V) = \inf_{(\tilde{U}, \tilde{V}) \sim \Pi(U,V)} \mathbb{E}(||\tilde{U} - \tilde{V}||),$$

where $\Pi(U, V)$ is the set of pairs of random vectors for which the first one is distributed as U and the second one as V. For $p > 1$, we can also define the L^p Wasserstein distance as minimizing the p-power of the distance over all possible couplings. We shall discuss the extension of our results to those stronger distances in Remark 7.

The next theorem is the main result of this paper. We show that the Wasserstein distance between the two distributions in (8) is bounded by the sum of a term of order $O((\log n)n^{-1/2})$ and the distance between $\bar{C}_{n,0}^{-1/2} n^{1/2} \nabla M_n(\theta_{0,n})$ and the standard Gaussian distribution. The benefit on Theorem 2 is then that $\bar{C}_{n,0}^{-1/2} n^{1/2} \nabla M_n(\theta_{0,n})$ is usually much easier to analyze than $\bar{C}_{n,0}^{-1/2} \bar{H}_{n,0} n^{1/2}$ $(\hat{\theta}_n - \theta_{0,n})$, since it takes an explicit form and is not defined as a minimizer. In Section 2.3, we discuss many existing possibilities to quantify the asymptotic normality of $\bar{C}_{n,0}^{-1/2} n^{1/2} \nabla M_n(\theta_{0,n})$.

Theorem 2. *Assume that Conditions 1 to 9 hold. Consider $\hat{\theta}_n$ as in (1). Then there are constants $0 < C_\mathcal{W} < \infty$ and $N_\mathcal{W} \in \mathbb{N}$ such that for $n \geq N_\mathcal{W}$, with Z following the standard Gaussian distribution on $\mathbb{R}^p$,*

$$\begin{aligned} \mathcal{W}_1(\bar{C}_{n,0}^{-1/2} \bar{H}_{n,0} \sqrt{n}(\hat{\theta}_n - \theta_{0,n}), Z) \leq & \mathcal{W}_1 \left(\bar{C}_{n,0}^{-1/2} \sqrt{n} \nabla M_n(\theta_{0,n}), Z \right) \\ & + C_\mathcal{W} \frac{\log n}{\sqrt{n}}. \end{aligned} \tag{9}$$

Remark 2. In Theorem 2, the bound on $\mathcal{W}_1(\bar{C}_{n,0}^{-1/2} \bar{H}_{n,0} n^{1/2} (\hat{\theta}_n - \theta_{0,n}), Z)$ directly provides a similar bound on $\mathcal{W}_1(n^{1/2}(\hat{\theta}_n - \theta_{0,n}), Z_n)$, where Z_n follows the centered Gaussian distribution with covariance matrix $\bar{H}_{n,0}^{-1} \bar{C}_{n,0} \bar{H}_{n,0}^{-1}$. Indeed the matrix $\bar{H}_{n,0}^{-1} \bar{C}_{n,0}^{1/2}$ is bounded and we can apply the well-known Lemma 1 below. The same remark applies to Theorems 3, 4 and 5, since the matrix $\bar{H}_{n,0}^{-1} \bar{C}_{n,0}^{1/2}$ is also bounded in these latter contexts (as is shown in the proofs).

Lemma 1. *Let U, V be two random vectors of $\mathbb{R}^p$ and $h : \mathbb{R}^p \to \mathbb{R}^p$ be such that for $u, v \in \mathbb{R}^p$, $||h(u) - h(v)|| \leq C||u - v||$ with $0 < C < \infty$. Then $\mathcal{W}_1(h(U), h(V)) \leq C\mathcal{W}_1(U, V)$.*

Remark 3. The bound in (9) is not exactly of order $O(n^{-1/2})$ as one may have wished; there is the extra factor $\log n$. With the proof techniques we have used, this extra factor can be seen in particular in (29) which itself follows from bounding the probability that the error $||\hat{\theta}_n - \theta_{0,n}||$ exceeds a threshold t in (28). For this probability bound to vanish, we need t to be of order $n^{-1/2}$ times $(\log n)^{1/2}$. This happens because the probability bound follows from a union bound with a diverging number of terms, see Lemma 7 and the proof of Lemma 2.

In future work, aiming at removing the extra factor $\log n$ in (9) is definitely of interest. The main obstacle for this is, in our opinion, that we typically need to bond the moments of $||\hat{\theta}_n - \theta_{0,n}||$, in cases where $\hat{\theta}_n$ does not have an explicit expression. Note that these moments also occur in the references [1, 2, 4, 5, 6, 7] discussed in the introduction, where they are bounded in several examples where $\hat{\theta}_n$ has an explicit expression. In these examples, in the end, bounds of order exactly $O(n^{-1/2})$ are obtained.

As Theorems 3 to 5 below rely on Theorem 2, extra factors $\log n$ appear there as well.

Remark 4. In Theorem 2, we do not provide an explicit expression of the constants $N_{\mathcal{W}}$ and $C_{\mathcal{W}}$, as a function, for instance, of p and of the constants given in Conditions 1 to 9. As a consequence, also Theorems 3 to 5 feature non-explicit bounds.

In specific examples, the references [1, 2, 4, 5, 6, 7] manage to provide explicit bounds, by exploiting in particular explicit expressions of $\hat{\theta}_n$. The numerical values of these explicit bounds also prove to be of practical use in simulations given in some of these references. In our case, given the level of generality of Theorem 2, providing explicit expressions for $N_{\mathcal{W}}$ and $C_{\mathcal{W}}$ appears to be difficult. One may also anticipate that, even if such explicit expressions were obtained and would yield fully explicit bounds in Theorems 3 to 5, these latter bounds would take too high numerical values to be of practical use.

2.3. *Background on approximate normality for functions of many random variables*

Theorem 2 reduces the problem of proving a quantitative bound on the distance to the Gaussian for a general M-estimator to proving the same statement for an explicit function of the data. We shall now describe some of the broad ideas for proving such statements, some of which will be used in the applications described in Section 3. We do not aim at being exhaustive, and other techniques can also be used in this context.

The abstract setting is to consider a random variable of the form $f(X_1, ..., X_n)$ where the X_i are random variables. The classical central limit theorem consists in taking the X_i to be i.i.d. and f to be a normalized sum.

When f is a sum, which arises for M-estimators of the form (2) (see Sections 3.1 and 3.2), there is a vast literature on quantitative central limit theorems, beyond the classical i.i.d. assumptions. For independent variables, we shall use here a very general result of Bonis [18], but many other results can be used in such a situation.

If f is not a sum, but is approximately affine, and all variables have some influence on the value, we still expect approximate normality. This heuristic has been made rigorous by second-order Poincaré inequalities, which bound distances to the Gaussian when certain functions of the first and second derivatives are small. They have been introduced in the Gaussian setting by Chatterjee [22], extended in [45], and analogues for general independent random variables via discrete second-order derivatives were studied in [21, 27, 29]. Second-order Poincaré inequalities for non-Gaussian, non-independent random variables do not seem to have been yet addressed in the literature, and warrant further investigation.

Another method for proving approximate normality in the Gaussian setting when the function f is a multivariate polynomial is via the quantitative fourth moment theorem of Nourdin and Peccati [42], which for example applies to U-statistics. When the polynomial is square-free and has low influences, it is possible to extend this phenomenon to more general i.i.d. random variables [46]. The approach extends to non-independent functions of Gaussian variables, a result known as the quantitative Breuer-Major theorem [41, 44]. We refer to the monograph [43] for a thorough discussion of this approach. We shall use a variant of it in Section 3.3.

For non-independent random variables, there have been successful implementations of variants of Stein's method, often in situations where there is some symmetry. Classical techniques include the exchangeable pairs method and the zero-bias transform, and we refer to [53] for a survey.

3. Applications

3.1. Minimization of averages of independent functions

We now show how Theorem 2 applies to estimators provided by

$$M_n(\theta) = \frac{1}{n}\sum_{i=1}^{n}\rho(\theta, X_i),$$

as in (2) with independent random vectors $X_1, \dots, X_n$.

We introduce the property of sub-Gaussianity, that holds for a large class of random variables, including Gaussian random variables, bounded random variables and uniformly log-concave random variables.

Definition 1. A real-valued random variable X is said to be sub-Gaussian with constant σ^2 if for any $t \in \mathbb{R}$ we have

$$\mathbb{E}\left(e^{t(X-\mathbb{E}[X])}\right) \leq e^{t^2\sigma^2/2}.$$

The next theorem, based on Theorem 2, provides a bound of order $O((\log n)\, n^{-1/2})$ in Wasserstein distance for the asymptotic normality of M-estimators based on (2), under uniform sub-Gaussianity for ρ and its derivatives with respect to θ.

Theorem 3. *Assume that $X_1, \dots, X_n$ are independent. Assume moreover that there are constants $0 < \sigma^2 < \infty$ and $0 < E_{\mathrm{sup}} < \infty$ such that for any $i \in \{1, \dots, n\}$, for any $j, k, \ell \in \{1, \dots, p\}$, for any $\theta_1 \in \Theta$, for any $\theta_2 \in \mathring{\Theta}$, and for any*

$$Y \in \left\{ \rho(\theta_1, X_i), \partial\rho(\theta_2, X_i)/\partial\theta_j, \partial^2\rho(\theta_2, X_i)/\partial\theta_j\partial\theta_k, \partial^3\rho(\theta_2, X_i)/\partial\theta_j\partial\theta_k\partial\theta_\ell \right\},\ Y \text{ is sub-Gaussian with constant } \sigma^2 \text{ and has absolute expectation bounded by } E_{\mathrm{sup}}. \tag{10}$$

Assume moreover that Conditions 1, 2 and 7 to 9 hold. Consider M_n, $\hat{\theta}_n$, $\bar{C}_{n,0}$ and $\bar{H}_{n,0}$ as in (2), (1), (5) and (6). Finally, assume that one of the two following conditions hold: either

- *Condition (O1): There exist fixed constants $\lambda > 0$ and $C < \infty$ such that*

$$\mathbb{E}\left(\exp\left(\lambda \sup_{\theta \in \mathring{\Theta}} ||\nabla\rho(\theta, X_k)||\right)\right) \le C;$$

$$\mathbb{E}\left(\exp\left(\lambda \sup_{\theta \in \mathring{\Theta}} \left|\frac{\partial^2\rho}{\partial\theta_i\partial\theta_j}(\theta, X_k)\right|\right)\right) \le C$$

  and

$$\mathbb{E}\left(\exp\left(\lambda \sup_{\theta \in \mathring{\Theta}} \left|\frac{\partial^3\rho}{\partial\theta_i\partial\theta_j\partial\theta_\ell}(\theta, X_k)\right|\right)\right) \le C$$

  for all $k \in \{1, \dots, n\}$ and $i, j, \ell \in \{1, \dots, p\}$.
  Or
- *Condition (O2): All the functions $||\nabla\rho(\cdot, x)||$, $\partial^2\rho(\cdot, x)/\partial\theta_i\partial\theta_j$ and $\partial^3\rho(\cdot, x)/\partial\theta_i\partial\theta_j\partial\theta_\ell$ have a modulus of continuity bounded by some function ω, uniformly in $x \in \mathcal{X}$ and in $i, j, \ell \in \{1, \dots, p\}$.*

Then, there are constants $0 < C_\rho < \infty$ and $N_\rho \in \mathbb{N}$ such that, for $n \ge N_\rho$, with Z following the standard Gaussian distribution,

$$\mathcal{W}_1(\bar{C}_{n,0}^{-1/2}\bar{H}_{n,0}\sqrt{n}(\hat{\theta}_n - \theta_{0,n}), Z) \le C_\rho \frac{\log n}{\sqrt{n}}.$$

Remark 5. The sub-Gaussianity assumption (10) of Theorem 3 on the partial derivatives of $\rho(\theta, X_i)$ with respect to θ can be checked based on the sub-Gaussianity of $X_1, \dots, X_n$ only and on regularity properties of ρ.

Indeed, it is known that if a random vector V with values in $\mathbb{R}^k$ has components that are sub-Gaussian with constant σ^2, then for any c-Lipschitz function $f : \mathbb{R}^k \to \mathbb{R}$, the variable $f(V)$ is sub-Gaussian with constant at most of order $kc^2\sigma^2$. The dimensional prefactor can be eliminated for example when the components are independent and satisfy Talagrand's L^2 transport-entropy inequality [33]. Consider then the case where $X_1, \dots, X_n$ are uniformly sub-Gaussian and for any $j, k, \ell \in \{1, \dots, p\}$, for any

$$f \in \left\{\rho\ , \partial\rho/\partial\theta_j,\ \partial^2\rho/\partial\theta_j\partial\theta_k,\ \partial^3\rho/\partial\theta_j\partial\theta_k\partial\theta_\ell\right\},$$

f is Lipschitz in its second variable, uniformly in θ, and $|f(\theta, x_{i,0})|$ is bounded, also uniformly in θ, for some reference values $x_{i,0}$ of X_i, $i = 1, \dots, n$. In this case then the uniform sub-Gaussianity assumption (10) of Theorem 3 holds.

Note also that these latter assumptions are not minimal. For example, we could relax the Lipschitz assumption on the second derivatives into some quadratic growth. The assumptions on the third derivatives are much stronger than what is necessary to ensure (4) to streamline applications: one can check essentially the same conditions on all derivatives up to order three, rather than single out a weaker condition for third derivatives.

Remark 6. The two possible conditions (O1) and (O2) in Theorem 3 are used to ensure that Condition 4 holds. There are other possible ways of verifying it, such as classical chaining techniques used to bound the suprema of stochastic processes when stochastic forms of continuity (in θ) hold, see for example [57, Chapter 8].

Proof of Theorem 3. First we must check that the conditions required by Theorem 2 are satisfied. By assumptions, this means checking conditions 3 to 6.

From the sub-Gaussianity and bounded expectation assumption (10), we uniformly control moments of all order, and the first two parts of Condition 5 hold. Condition 3 is an immediate consequence of the Gaussian concentration assumption and Chernoff's concentration bound. Condition 6 can be established using the fact that we wish to control the variances of averages of independent variables, and the uniform moment bounds.

Finally, we need to check that Condition 4 holds, assuming either (O1) or (O2) holds. If the first one holds, Condition 4 is just a consequence of Markov's inequality. If the second one holds, by continuity, Condition 1 and fixing some $\lambda > 0$, and some $\epsilon > 0$ small enough, we have for any $k \in \{1, \dots, n\}$,

$$\begin{aligned}
\mathbb{E}\left(\exp\left(\lambda \sup_{\theta \in \mathring{\Theta}} ||\nabla\rho(\theta, X_k)||\right)\right) &\leq \mathbb{E}\left(\exp\left(\lambda \sup_{\theta_i, i \leq N} ||\nabla\rho(\theta_i, X_k)|| + \lambda\omega(\epsilon)\right)\right) \\
&\leq e^{\lambda\omega(\epsilon)} \sum_{i \leq N} \mathbb{E}(\exp(\lambda||\nabla\rho(\theta_i, X_k)||)) \\
&\leq C',
\end{aligned}$$

for some constant $0 < C' < \infty$, where the final bound uses the Gaussian concentration of $||\nabla\rho(\theta, X_k)||$ for fixed θ and the uniform bound on its expectation. The same reasoning applies for the second derivatives, and therefore Condition 4 holds with the same argument as when (O1) holds. One can also check (4) with the same reasoning.

Since Theorem 2 applies, we are reduced to understanding the asymptotic behavior of

$$\sqrt{n}\nabla M_n(\theta_{0,n}) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \nabla\rho(\theta_{0,n}, X_i).$$

Hence we are in the setting of a quantitative central limit theorem for sums of independent random vectors. From the sub-Gaussianity assumption (10), we see that the fourth moments of $\nabla\rho(\theta_{0,n}, X_i)$, $i = 1, \dots, n$, are uniformly bounded. Moreover, by Condition 9, this is not modified by multiplying these vectors by $\bar{C}_{n,0}^{-1/2}$. Hence we are considering a sum of independent random vectors with covariances summing to the identity matrix I_p, and we can apply the following statement to conclude the proof, which is a particular case of a result of Bonis [18, Theorem 11].

Proposition 1. *Let $(Z_i)_{i=1,\dots,n}$ be a sequence of independent random vectors taking values in $\mathbb{R}^p$, each centered, and such that $\mathrm{Cov}(\sum_{i=1}^n Z_i) = nI_p$. Assume moreover that for any $i \in \{1, \dots, n\}$, $\mathbb{E}[||Z_i||^4] \leq \beta^2$, for a given $0 < \beta < \infty$. Then*

$$\mathcal{W}_1\left(\frac{1}{\sqrt{n}}\sum_{i=1}^n Z_i, Z\right) \leq \frac{26(\beta^{3/2} + p\beta)}{\sqrt{n}}$$

where Z is a standard Gaussian vector on $\mathbb{R}^p$. □

We stated here the version for independent but not identically distributed variables. For i.i.d. variables, the bound can be further improved, as discussed in [18].

3.2. Parameter estimation in logistic regression

We shall now present the simple example of logistic regression, where Theorem 3 is applied to a maximum likelihood estimator. We consider a deterministic sequence $(x_i)_{i\in\mathbb{N}}$ of vectors in $\mathbb{R}^p$. To match the assumptions of Theorem 3, we assume this sequence to be bounded.

Condition 10. There is a constant $0 < C_{x,1} < \infty$ such that for $i \in \mathbb{N}$,

$$||x_i|| \leq C_{x,1}.$$

As previously, we let Θ be a fixed compact subset of $\mathbb{R}^p$. We let $\theta_0 \in \mathring{\Theta}$ be fixed. We consider a sequence $(y_i)_{i\in\mathbb{N}}$ of independent random variables with, for $i \in \mathbb{N}$, $y_i \in \{0, 1\}$ and

$$P(y_i = 1) = \frac{e^{x_i^\top \theta_0}}{1 + e^{x_i^\top \theta_0}}. \tag{11}$$

We let, for $\theta \in \Theta$,

$$p_{i,\theta} = \frac{e^{x_i^\top \theta}}{1 + e^{x_i^\top \theta}}.$$

Hence, we are in the classical well-specified case where the parameter $\theta_0 \in \Theta$ characterizes the data generating process, or distribution, of $y_1, \dots, y_n$. The likelihood function of y_i is, for $\theta \in \Theta$,

$$\mathcal{L}(\theta, y_i) = p_{i,\theta}^{y_i}(1 - p_{i,\theta})^{1-y_i}.$$

Minus the logarithm of the likelihood of y_i is, for $\theta \in \Theta$,

$$\rho(\theta, x_i, y_i) = -y_i \log(p_{i,\theta}) - (1 - y_i)\log(1 - p_{i,\theta})$$

$$= -y_i x_i^\top \theta + \log\left(1 + e^{x_i^\top \theta}\right).$$

Hence minus the normalized log likelihood function is, for $\theta \in \Theta$,

$$M_n(\theta) = \frac{1}{n}\sum_{i=1}^{n}\left(-y_i x_i^\top \theta + \log\left(1 + e^{x_i^\top \theta}\right)\right). \tag{12}$$

Note that we do not have an explicit expression for the minimizer of M_n. We have, for $\theta \in \mathring{\Theta}$,

$$\nabla M_n(\theta) = \frac{1}{n}\sum_{i=1}^{n}\left(-y_i x_i + \frac{e^{x_i^\top \theta}}{1 + e^{x_i^\top \theta}} x_i\right) = \frac{1}{n}\sum_{i=1}^{n}\left(-y_i x_i + p_{i,\theta} x_i\right). \tag{13}$$

We also have, for $\theta \in \mathring{\Theta}$,

$$\nabla^2 M_n(\theta) = \frac{1}{n}\sum_{i=1}^{n}\frac{e^{x_i^\top \theta}(1 + e^{x_i^\top \theta}) - e^{x_i^\top \theta} e^{x_i^\top \theta}}{(1 + e^{x_i^\top \theta})^2} x_i x_i^\top = \frac{1}{n}\sum_{i=1}^{n}\frac{e^{x_i^\top \theta}}{(1 + e^{x_i^\top \theta})^2} x_i x_i^\top. \tag{14}$$

Hence we see that $M_n(\theta)$ is convex with respect to θ. Next, we assume that the empirical second moment matrix of the x_i's is asymptotically strictly positive definite. This type of condition is common for logistic regression [14, 30, 39] and ensures asymptotic identifiability (Condition 8).

Condition 11. There are constants $0 < c_{x,2} < \infty$ and $N_{x,2} \in \mathbb{N}$ such that, for $n \geq N_{x,2}$,

$$\lambda_p\left(\frac{1}{n}\sum_{i=1}^{n} x_i x_i^\top\right) \geq c_{x,2}.$$

We can now state the Wasserstein bound on the asymptotic normality of the maximum likelihood estimator, in logistic regression. To our knowledge, this is the first established rate of convergence of asymptotic normality in logistic regression.

Theorem 4. *Assume that Θ satisfies Condition 1. Assume that Conditions 10 and 11 hold. Consider M_n in (12), $\hat{\theta}_n$ as in (1), θ_0 as defined in (11), $\bar{C}_{n,0}$ as in (5) and $\bar{H}_{n,0}$ as in (6). Then, there are constants $0 < C_{log} < \infty$ and $N_{log} \in \mathbb{N}$ such that for $n \geq N_{log}$, with Z following the standard Gaussian distribution on $\mathbb{R}^p$,*

$$\mathcal{W}_1\left(\bar{C}_{n,0}^{-1/2}\bar{H}_{n,0}\sqrt{n}(\hat{\theta}_n - \theta_0), Z\right) \leq C_{log}\frac{\log n}{\sqrt{n}}.$$

3.3. Covariance parameter estimation for Gaussian processes by cross validation

Our last example stems from the field of spatial statistics [8, 9, 13, 25, 26, 34, 58, 60, 61]. The goal is to illustrate the benefit of Theorem 2 to a situation where the observations are dependent and where M_n does not correspond to a likelihood. We stress that $\hat{\theta}_n$ has no explicit expression.

We consider a sequence $(x_i)_{i\in\mathbb{N}}$ of deterministic vectors in $\mathbb{R}^d$, that we call observation points. Then, for $n \in \mathbb{N}$, the observed data consist in a vector $y^{(n)}$ of size $n \times 1$ which component i is $\xi(x_i)$, where $\xi : \mathbb{R}^d \to \mathbb{R}$ is a centered Gaussian process.

We are interested in the parametric estimation of the correlation function of ξ, based on a parametric set of stationary correlation functions $\{k_\theta; \theta \in \Theta\}$, where for $\theta \in \Theta$, $k_\theta : \mathbb{R}^d \to \mathbb{R}$ and $(u,v) \in \mathbb{R}^{2d} \mapsto k_\theta(u-v)$ is a correlation function. For an introduction to usual parametric sets of stationary correlation functions in spatial statistics, we refer for instance to [10, 25, 26, 32, 58].

As an estimator for θ, we consider the minimization of the average of square leave-one-out errors, letting, for $\theta \in \Theta$,

$$M_n(\theta) = \frac{1}{n}\sum_{i=1}^{n}\left(y_i^{(n)} - \mathbb{E}_\theta(y_i^{(n)}|y_{-i}^{(n)})\right)^2.$$

Above, $y_{-i}^{(n)}$ is obtained from $y^{(n)}$ by deleting the component i and $\mathbb{E}_\theta(\cdot|\cdot)$ means that the conditional expectation is computed as if the Gaussian process ξ had correlation function $(u,v) \in \mathbb{R}^{2d} \mapsto k_\theta(u-v)$. Now, for $\theta \in \Theta$, let $R_{n,\theta}$ be the $n \times n$ matrix with coefficient i,j equal to $k_\theta(x_i - x_j)$, that is, the correlation matrix of $y^{(n)}$ under correlation function given by k_θ. Then, from for instance [8, 28, 61] (to which we refer for more background and discussions on cross validation for Gaussian processes), we have

$$M_n(\theta) = \frac{1}{n} y^{(n)\top} R_{n,\theta}^{-1} \mathrm{diag}(R_{n,\theta}^{-1})^{-2} R_{n,\theta}^{-1} y^{(n)}, \tag{15}$$

where $\mathrm{diag}(M)$ is obtained by setting the off-diagonal elements of a square matrix M to zero.

For $n \in \mathbb{N}$, we let $\theta_{0,n} = \theta_0$, where θ_0 is a fixed element of $\mathring{\Theta}$ such that ξ has correlation function k_{θ_0}, which also implies that $y^{(n)}$ has correlation matrix R_{n,θ_0}. This corresponds to a well-specified parametric set of correlation functions. The next condition means that we consider the increasing-domain asymptotic framework, where the sequence of observation points is unbounded, with a minimal distance between any two distinct points [9, 26, 40].

Condition 12. There is a constant $c_x > 0$ such that for $i, j \in \mathbb{N}$, $i \neq j$,

$$||x_i - x_j|| \geq c_x.$$

The next condition is a lower bound on the smallest eigenvalues of the correlation matrices from the parametric model. Given the increasing-domain asymptotic framework (Condition 12), this lower bound indeed holds for a large class of families of stationary correlation functions [9, 12].

Condition 13. There is a constant $0 < c_{R,1} < \infty$ such that

$$\inf_{n\in\mathbb{N}} \inf_{\theta\in\Theta} \lambda_n(R_{n,\theta}) \geq c_{R,1}.$$

Next, we assume a third-order smoothness with respect to θ as well as a decay of the correlation at large distance. As before, many families of stationary correlation functions do satisfy this.

Condition 14. For any $x \in \mathbb{R}^d$, $k_\theta(x)$ is three times continuously differentiable with respect to θ on $\mathring{\Theta}$. There exist constants $0 < C_{R,2} < \infty$ and $0 < c_{R,2} < \infty$ such that for $\theta \in \Theta$, for $x \in \mathbb{R}^d$,

$$|k_\theta(x)| \leq \frac{C_{R,2}}{1 + ||x||^{d+c_{R,2}}}, \quad n \in \mathbb{N} \tag{16}$$

and for $\theta \in \mathring{\Theta}$, for $x \in \mathbb{R}^d$,

$$\max_{\substack{k \in \{1,2,3\} \\ i_1,\dots,i_k \in \{1,\dots,p\}}} \left| \frac{\partial^k}{\partial\theta_{i_1},\dots,\partial\theta_{i_k}} k_\theta(x) \right| \leq \frac{C_{R,2}}{1 + ||x||^{d+c_{R,2}}}, \quad n \in \mathbb{N}. \tag{17}$$

The next condition is interpreted as a global identifiability of the correlation parameter. This condition is already made in the increasing-domain asymptotic literature on cross validation and is not restrictive on the sequence $(x_i)_{i\in\mathbb{N}}$ and the set $\{k_\theta\}$ [9, 11].

Condition 15. For all $\mathcal{X} > 0$, there are constants $0 < c_{\mathcal{X}} < \infty$ and $N_{\mathcal{X}} \in \mathbb{N}$ such that for $n \geq N_{\mathcal{X}}$,

$$\inf_{\substack{\theta \in \Theta \\ ||\theta - \theta_0|| \geq \mathcal{X}}} \frac{1}{n} \sum_{i,j=1}^{n} \left(k_\theta(x_i - x_j) - k_{\theta_0}(x_i - x_j) \right)^2 \geq c_{\mathcal{X}}.$$

Finally, the last condition is interpreted as a local identifiability of the correlation parameter around θ_0. Its discussion is similar to the previous one.

Condition 16. For all $\alpha_1, \dots, \alpha_p \in \mathbb{R}$, with $\alpha_1^2 + \cdots + \alpha_p^2 > 0$, there are constants $0 < c_\alpha < \infty$ and $N_\alpha \in \mathbb{N}$ such that for $n \geq N_\alpha$,

$$\frac{1}{n} \sum_{i,j=1}^{n} \left(\sum_{\ell=1}^{p} \alpha_\ell \frac{\partial k_{\theta_0}(x_i - x_j)}{\partial \theta_\ell} \right)^2 \geq c_\alpha.$$

Under the above conditions, it is known from [9, 11] that $n^{1/2}(\hat{\theta}_n - \theta_0)$ converges in distribution to a centered Gaussian vector with covariance matrix $\bar{H}_{n,0}^{-1} \bar{C}_{n,0} \bar{H}_{n,0}^{-1}$, with the notation of (5) and (6). Based on Theorem 2, we can show that the rate of this convergence is $O((\log n) n^{-1/2})$ in Wasserstein distance. To the best of our knowledge, this is the first result of this kind for cross validation estimation for spatial Gaussian processes. We remark that Theorem 2 also enables to address maximum likelihood estimation of covariance parameters (see for instance [9, 26]), but we focus on cross validation for the sake of brevity and to highlight the benefits of Theorem 2 beyond maximum likelihood.

Theorem 5. *Assume that Θ satisfies Condition 1. Assume that Conditions 12 to 16 hold. Consider M_n in (15). Consider then $\hat{\theta}_n$ as in (1), θ_0 as defined* *after* (15)*,* $\bar{C}_{n,0}$ *as in* (5) *and* $\bar{H}_{n,0}$ *as in* (6)*. Then, there are constants* $0 < C_{CV} < \infty$ *and* $N_{CV} \in \mathbb{N}$ *such that for* $n \geq N_{CV}$*, with* Z *following the standard Gaussian distribution on* $\mathbb{R}^p$*,*

$$\mathcal{W}_1 \left(\bar{C}_{n,0}^{-1/2} \bar{H}_{n,0} \sqrt{n}(\hat{\theta}_n - \theta_0), Z \right) \leq C_{CV} \frac{\log n}{\sqrt{n}}.$$

Remark 7 (Extension to L^p Wasserstein distances). Most of our work can be fairly straightforwardly extended to L^p Wasserstein distances when $p > 1$, at the cost of appropriate stronger assumptions. More precisely, Theorem 2 uses a coupling argument (see (24)), and can be extended if we control higher moments, by replacing Conditions 5 and 6 with higher moment controls. Theorem 4 also extends, since we consider bounded variables and the quantitative central limit theorem of [18] holds for all L^p distances if all moments are bounded. As things stand, we lack a generalization of Proposition 2 to other distances to extend Theorem 5, but we expect that the techniques of [38] can be used to generalize it.

Quantitative central limit theorems for maximum likelihood in L^p Wasserstein distances for general p have been considered in [5] for i.i.d. random variables.

Appendix A: Proofs for Section 2

Lemma 2. *Assume that Conditions 1 to 5 hold. Then there are constants* $0 < c_{M,1} < \infty$*,* $0 < c'_{M,1} < \infty$*,* $0 < C_{M,1} < \infty$ *and* $0 < C'_{M,1} < \infty$ *such that, for* $0 < t \leq c'_{M,1}$ *and* $K \geq C'_{M,1}$*,*

$$\mathbb{P}\left(\sup_{\theta \in \Theta} |M_n(\theta) - \mathbb{E}(M_n(\theta))| \geq t \right) \leq \frac{C_{M,1} K^p e^{-n c_{M,1} t^2}}{t^p} + C_{M,1} n e^{-c_{M,1} K}.$$

Proof of Lemma 2. From Condition 1, and with c'_M and $C'_{d,1}$ from Conditions 3 and 4, there exists a constant $C_{\Theta,2}$ such that for $0 < r \leq c'_M / 2C'_{d,1}$, there exist $N \leq C_{\Theta,2} r^{-p}$ and $S_r = \{\theta_1, \ldots, \theta_N\} \subseteq \mathring{\Theta}$ such that for each $\theta \in \Theta$, there exists $i \in \{1, \ldots, N\}$ such that $(\theta, \theta_i) \subseteq \mathring{\Theta}$ and $||\theta - \theta_i|| \leq r$. We then have, for each $K \geq C'_{d,1}$, $0 < t \leq c'_M$, using the mean value theorem,

$$\mathbb{P}\left(\sup_{\theta \in \Theta} |M_n(\theta) - \mathbb{E}(M_n(\theta))| \geq t \right) \leq \mathbb{P}\left(\max_{\theta \in S_{t/2K}} |M_n(\theta) - \mathbb{E}(M_n(\theta))| \geq \frac{t}{2} \right) + \mathbb{P}\left(\sup_{\theta \in \mathring{\Theta}} ||\nabla M_n(\theta)|| \geq \frac{K}{2} \right) + \mathbb{P}\left(\sup_{\theta \in \mathring{\Theta}} ||\nabla \mathbb{E}(M_n(\theta))|| \geq \frac{K}{2} \right).$$

Hence, because $\nabla \mathbb{E}(M_n(\theta)) = \mathbb{E}(\nabla M_n(\theta))$ is bounded from Conditions 2 and 5, and using a union bound, there is a constant $C'_{d,1} \leq C_1 < \infty$ such that when $K \geq C_1$, $0 < t \leq c'_M$, we obtain

$$\mathbb{P}\left(\sup_{\theta \in \Theta} |M_n(\theta) - \mathbb{E}(M_n(\theta))| \geq t \right) \leq \tag{18}$$

$$\frac{C_{\Theta,2}2^pK^p}{t^p}\max_{\theta\in\Theta}\mathbb{P}\left(|M_n(\theta)-\mathbb{E}(M_n(\theta))|\geq\frac{t}{2}\right)+\mathbb{P}\left(\sup_{\theta\in\Theta}||\nabla M_n(\theta)||\geq\frac{K}{2}\right).$$

Hence, using Conditions 3 and 4, we obtain, for $0<t\leq c'_M$ and $K\geq C_1$,

$$\mathbb{P}\left(\sup_{\theta\in\Theta}|M_n(\theta)-\mathbb{E}(M_n(\theta))|\geq t\right)\leq\frac{C_{\Theta,2}2^pK^pC_Me^{-nc_Mt^2/4}}{t^p}+C_{d,1}ne^{-c_{d,1}K/2}.$$

This concludes the proof. □

Lemma 3. *Assume that Conditions 1 to 5 hold. Then there are constants* $0<c_{\nabla,1}<\infty$, $0<c'_{\nabla,1}<\infty$, $0<C_{\nabla,1}<\infty$ *and* $0<C'_{\nabla,1}<\infty$ *such that, for* $0<t\leq c'_{\nabla,1}$ *and* $K\geq C'_{\nabla,1}$,

$$\mathbb{P}\left(\sup_{\theta\in\Theta}||\nabla M_n(\theta)-\mathbb{E}(\nabla M_n(\theta))||\geq t\right)\leq\frac{C_{\nabla,1}K^pe^{-nc_{\nabla,1}t^2}}{t^p}+C_{\nabla,1}ne^{-c_{\nabla,1}K}.$$

Proof of Lemma 3. The proof is identical to that of Lemma 2. □

Lemma 4. *Assume that Conditions 1 to 5 and 7 hold. For any* $r>0$, *there are constants* $0<c_{\hat{\theta},r}<\infty$ *and* $0<C_{\hat{\theta},r}<\infty$ *such that*

$$\mathbb{P}(||\hat{\theta}_n-\theta_{0,n}||\geq r)\leq C_{\hat{\theta},r}ne^{-c_{\hat{\theta},r}n^{1/4}}.$$

Proof of Lemma 4. The event $||\hat{\theta}_n-\theta_{0,n}||\geq r$ implies

$$\inf_{\substack{\theta\in\Theta\\||\theta-\theta_{0,n}||\geq r}}(M_n(\theta)-M_n(\theta_{0,n}))\leq 0.$$

From Condition 7 and the triangle inequality, this implies, with a constant $0<c_1<\infty$, for n large enough,

$$\sup_{\theta\in\Theta}|M_n(\theta)-\mathbb{E}(M_n(\theta))|\geq c_1.$$

Hence

$$\mathbb{P}(||\hat{\theta}_n-\theta_{0,n}||\geq r)\leq\mathbb{P}\left(\sup_{\theta\in\Theta}|M_n(\theta)-\mathbb{E}(M_n(\theta))|\geq c_1\right).$$

Using now Lemma 2 with $K=n^{1/4}$ and n large enough, we obtain, for some constants $0<c_2<\infty$, $0<C_2<\infty$, $0<c_3<\infty$ and $0<C_3<\infty$, for n large enough,

$$\mathbb{P}(||\hat{\theta}_n-\theta_{0,n}||\geq r)\leq C_2n^{p/4}e^{-nc_2}+C_2ne^{-c_2n^{1/4}}\leq C_3ne^{-c_3n^{1/4}}.$$ □

Lemma 5. *Assume that Conditions 2, 5 and 8 hold. There exist constants* $0<c_{\nabla^2,1}<\infty$, $0<c'_{\nabla^2,1}<\infty$ *and* $N_{\nabla^2,1}\in\mathbb{N}$ *such that for* $n\geq N_{\nabla^2,1}$

$$\inf_{\substack{\theta\in\mathring{\Theta}\\||\theta-\theta_{0,n}||\leq c'_{\nabla^2,1}}}\lambda_p(\mathbb{E}(\nabla^2M_n(\theta)))\geq c_{\nabla^2,1}.$$

Proof of Lemma 5. Condition 5, together with the fact that we can exchange derivatives and expectation for M_n (Condition 2) imply that the derivatives of $\mathbb{E}(\nabla^2 M_n)$ are bounded uniformly in $\theta \in \mathring{\Theta}$. Hence, from Condition 8, we can conclude the proof. $\square$

Lemma 6. *Assume that Conditions 2, 5, 7 and 8 hold. Recall c_{θ_0} from Condition 7. There are constants $0 < c_{\nabla,2} < \infty$, $0 < c'_{\nabla,2} \leq c_{\theta_0}$ and $N_{\nabla,2} \in \mathbb{N}$ such that for $n \geq N_{\nabla,2}$, for $||\theta - \theta_{0,n}|| \leq c'_{\nabla,2}$,*

$$||\mathbb{E}(\nabla M_n(\theta))|| \geq c_{\nabla,2}||\theta - \theta_{0,n}||.$$

Proof of Lemma 6. Using Lemma 5 and $\mathbb{E}(\nabla^2 M_n(\theta)) = \nabla^2\mathbb{E}(M_n(\theta))$ (Condition 2), we have, for $||\theta - \theta_{0,n}|| \leq c'_{\nabla^2,1}$ and for n large enough,

$$\begin{aligned}&||\nabla\mathbb{E}(M_n(\theta)) - \nabla\mathbb{E}(M_n(\theta_{0,n}))||\ ||\theta - \theta_{0,n}|| \\ &\geq \left(\nabla\mathbb{E}(M_n(\theta)) - \nabla\mathbb{E}(M_n(\theta_{0,n}))\right)^\top (\theta - \theta_{0,n}) \geq c_{\nabla^2,1}||\theta - \theta_{0,n}||^2.\end{aligned}$$

From Conditions 2 and 7,

$$\nabla\mathbb{E}(M_n(\theta_{0,n})) = 0.$$

Hence we have, for $||\theta - \theta_{0,n}|| \leq c'_{\nabla^2,1}$ and for n large enough,

$$||\nabla\mathbb{E}(M_n(\theta))|| \geq c_{\nabla^2,1}||\theta - \theta_{0,n}||.$$

We conclude from Condition 2. $\square$

Lemma 7. *Assume that Conditions 1 to 5, 7 and 8 hold. Recall c_{θ_0} from Condition 7. For any constant $\gamma_1 > 0$, there are constants $0 < c_{\nabla,\hat{\theta},1} < \infty$, $0 < c'_{\nabla,\hat{\theta},1} \leq c_{\theta_0}$, $0 < C_{\nabla,\hat{\theta},1} < \infty$ and $N_{\nabla,\hat{\theta},1} \in \mathbb{N}$ such that for $n \geq N_{\nabla,\hat{\theta},1}$ and $t \leq c'_{\nabla,\hat{\theta},1}$,*

$$\begin{aligned}\mathbb{P}\left(\nabla M_n(\hat{\theta}_n) = 0, t \leq ||\hat{\theta}_n - \theta_{0,n}|| \leq c_{\theta_0}\right) &\leq C_{\nabla,\hat{\theta},1}\frac{\log(n)^{p\gamma_1}}{t^p}e^{-nc_{\nabla,\hat{\theta},1}t^2} \\ &\quad + C_{\nabla,\hat{\theta},1}ne^{-c_{\nabla,\hat{\theta},1}(\log n)^{\gamma_1}} + C_{\nabla,\hat{\theta},1}ne^{-c_{\nabla,\hat{\theta},1}n^{1/4}}.\end{aligned}$$

Proof of Lemma 7. Recall $c'_{\nabla,2}$ from Lemma 6. For $0 < t < c'_{\nabla,2}$, we have, using Lemmas 4 and 6, for n large enough,

$$\begin{aligned}&\mathbb{P}\left(\nabla M_n(\hat{\theta}_n) = 0, t \leq ||\hat{\theta}_n - \theta_{0,n}|| \leq c_{\theta_0}\right) \\ &\leq \mathbb{P}\left(\inf_{\theta \in B(\theta_{0,n},c'_{\nabla,2})\backslash B(\theta_{0,n},t)} ||\nabla M_n(\theta)|| = 0\right) + \mathbb{P}\left(||\hat{\theta}_n - \theta_{0,n}|| \geq c'_{\nabla,2}\right) \\ &\leq \mathbb{P}\left(\sup_{\theta\in\mathring{\Theta}} ||\nabla M_n(\theta) - \mathbb{E}(\nabla M_n(\theta))|| \geq c_{\nabla,2}t\right) + C_{\hat{\theta},c'_{\nabla,2}}ne^{-c_{\hat{\theta},c'_{\nabla,2}}n^{1/4}}. \qquad (19)\end{aligned}$$

For any constant $0 < \gamma_1 < \infty$, we can now use Lemma 3 with $K = (\log n)^{\gamma_1}$ to obtain, for $0 < t < \min(c'_{\nabla,2}, c'_{\nabla,1})$, for n large enough,

$$\mathbb{P}\left(\nabla M_n(\hat{\theta}_n) = 0, t \leq ||\hat{\theta}_n - \theta_{0,n}|| \leq c_{\theta_0}\right) \leq C_{\nabla,1} \frac{\log(n)^{p\gamma_1}}{c_{\nabla,2}^p t^p} e^{-nc_{\nabla,1}c_{\nabla,2}^2 t^2} \tag{20}$$

$$+ C_{\nabla,1} n e^{-c_{\nabla,1}(\log n)^{\gamma_1}} + C_{\hat{\theta},c'_{\nabla,2}} n e^{-c_{\hat{\theta},c'_{\nabla,2}} n^{1/4}}. \tag{21}$$

This concludes the proof. □

Proof of Theorem 2. From the triangle inequality, we have

$$\begin{aligned}
&\mathcal{W}_1\left(\bar{C}_{n,0}^{-1/2}\bar{H}_{n,0}\sqrt{n}(\hat{\theta}_n - \theta_{0,n}), Z\right) \\
\leq &\mathcal{W}_1\left(\bar{C}_{n,0}^{-1/2}\bar{H}_{n,0}\sqrt{n}(\hat{\theta}_n - \theta_{0,n}), -\bar{C}_{n,0}^{-1/2}\sqrt{n}\nabla M_n(\theta_{0,n})\right) \\
&+ \mathcal{W}_1\left(-\bar{C}_{n,0}^{-1/2}\sqrt{n}\nabla M_n(\theta_{0,n}), Z\right) \\
=: &W_1 + W_2.
\end{aligned} \tag{22}$$

Observe first that

$$\begin{aligned}
W_2 &= \mathcal{W}_1\left(-\bar{C}_{n,0}^{-1/2}\sqrt{n}\nabla M_n(\theta_{0,n}), Z\right) \\
&= \mathcal{W}_1\left(-\bar{C}_{n,0}^{-1/2}\sqrt{n}\nabla M_n(\theta_{0,n}), -Z\right) = \mathcal{W}_1\left(\bar{C}_{n,0}^{-1/2}\sqrt{n}\nabla M_n(\theta_{0,n}), Z\right).
\end{aligned} \tag{23}$$

Hence, it is sufficient to bound

$$W_1 = \mathcal{W}_1\left(\bar{C}_{n,0}^{-1/2}\bar{H}_{n,0}\sqrt{n}(\hat{\theta}_n - \theta_{0,n}), -\bar{C}_{n,0}^{-1/2}\sqrt{n}\nabla M_n(\theta_{0,n})\right),$$

which we now do. Using the trivial coupling (alternatively, using the definition of the L^1 Wasserstein distance as a supremum of expectation difference over 1-Lipschitz functions), we have

$$W_1 \leq \mathbb{E}\left(\left|\left|\bar{C}_{n,0}^{-1/2}\bar{H}_{n,0}\sqrt{n}(\hat{\theta}_n - \theta_{0,n}) + \bar{C}_{n,0}^{-1/2}\sqrt{n}\nabla M_n(\theta_{0,n})\right|\right|\right). \tag{24}$$

With c_{θ_0} as in Condition 7, observe that if $\hat{\theta}_n \in B(\theta_{0,n}, c_{\theta,0})$ then $\nabla M_n(\hat{\theta}_n) = 0$. Hence, applying Hölder's inequality, we obtain,

$$\begin{aligned}
W_1 \leq &\mathbb{E}\left(\left|\left|\bar{C}_{n,0}^{-1/2}\bar{H}_{n,0}\sqrt{n}(\hat{\theta}_n - \theta_{0,n}) + \bar{C}_{n,0}^{-1/2}\sqrt{n}\nabla M_n(\theta_{0,n})\right|\right|^2\right)^{1/2} \\
&\mathbb{P}\left(\hat{\theta}_n \notin B(\theta_{0,n}, c_{\theta,0})\right)^{1/2} \\
+ &\mathbb{E}\Big(\mathbb{1}_{\{\nabla M_n(\hat{\theta}_n)=0\}}\mathbb{1}_{\{\hat{\theta}_n \in B(\theta_{0,n}, c_{\theta,0})\}} \\
&\left|\left|\bar{C}_{n,0}^{-1/2}\bar{H}_{n,0}\sqrt{n}(\hat{\theta}_n - \theta_{0,n}) + \bar{C}_{n,0}^{-1/2}\sqrt{n}\nabla M_n(\theta_{0,n})\right|\right|\Big)
\end{aligned}$$

$$=\mathbb{E}(W_{1,1})^{1/2}\mathbb{P}(A_{1,1})^{1/2}+\mathbb{E}(W_{1,2}), \tag{25}$$

where we define

$$W_{1,1}=\left|\left|\bar{C}_{n,0}^{-1/2}\bar{H}_{n,0}\sqrt{n}(\hat{\theta}_n-\theta_{0,n})+\bar{C}_{n,0}^{-1/2}\sqrt{n}\nabla M_n(\theta_{0,n})\right|\right|^2,$$

$$A_{1,1}=\left\{\hat{\theta}_n\not\in B(\theta_{0,n},c_{\theta,0})\right\}$$

and

$$\begin{aligned}W_{1,2}=&\mathbb{1}_{\{\nabla M_n(\hat{\theta}_n)=0\}}\mathbb{1}_{\{\hat{\theta}_n\in B(\theta_{0,n},c_{\theta,0})\}}\\&\left|\left|\bar{C}_{n,0}^{-1/2}\bar{H}_{n,0}\sqrt{n}(\hat{\theta}_n-\theta_{0,n})+\bar{C}_{n,0}^{-1/2}\sqrt{n}\nabla M_n(\theta_{0,n})\right|\right|.\end{aligned}$$

Let us first bound $\mathbb{E}(W_{1,1})^{1/2}\mathbb{P}(A_{1,1})^{1/2}$. In $W_{1,1}$, $\bar{C}_{n,0}^{-1/2}$ is bounded from Condition 9 and $\bar{H}_{n,0}$ is bounded from Condition 5. Furthermore, $\sqrt{n}\nabla M_n(\theta_{0,n})$ has mean zero from Condition 7 and has bounded covariance matrix from Condition 6. Hence, since Θ is compact, with constants $0<C_1<\infty$ and $N_1\in\mathbb{N}$, we have for $n\geq N_1$,

$$\mathbb{E}(W_{1,1})^{1/2}\leq C_1\sqrt{n}.$$

Then Lemma 4 directly provides, for some constant $0<c_2<\infty$, $0<C_2<\infty$ and $N_2\in\mathbb{N}$, for $n\geq N_2$,

$$\mathbb{P}(A_{1,1})^{1/2}\leq C_2\sqrt{n}e^{-c_2n^{1/4}}.$$

Hence, eventually, for some constants $0<c_3<\infty$, $0<C_3<\infty$ and $N_3\in\mathbb{N}$, for $n\geq N_3$,

$$\mathbb{E}(W_{1,1})^{1/2}\mathbb{P}(A_{1,1})^{1/2}\leq C_3ne^{-c_3n^{1/4}}. \tag{26}$$

Let us now bound $\mathbb{E}(W_{1,2})$. When $\nabla M_n(\hat{\theta}_n)=0$ and $\hat{\theta}_n\in B(\theta_{0,n},c_{\theta,0})$, we have, since $B(\theta_{0,n},c_{\theta,0})\subset\mathring{\Theta}$,

$$0=\nabla M_n(\theta_{0,n})+\nabla^2M_n(\tilde{\theta}_1,\dots,\tilde{\theta}_p)(\hat{\theta}_n-\theta_{0,n}),$$

where $\tilde{\theta}_1,\dots,\tilde{\theta}_p$ are on the segment between $\hat{\theta}_n$ and $\theta_{0,n}$ and where $\nabla^2M_n(\tilde{\theta}_1,\dots,\tilde{\theta}_p)$ is $p\times p$ with line k equal to the line k of $\nabla^2M_n(\tilde{\theta}_k)$ for $k\in\{1,\dots,p\}$. This yields, when $\nabla M_n(\hat{\theta}_n)=0$ and $\hat{\theta}_n\in B(\theta_{0,n},c_{\theta,0})$,

$$\begin{aligned}&\bar{H}_{n,0}\sqrt{n}(\hat{\theta}_n-\theta_{0,n})+\sqrt{n}\nabla M_n(\theta_{0,n})\\&=\sqrt{n}\left(\mathbb{E}(\nabla^2M_n(\theta_{0,n}))-\nabla^2M_n(\tilde{\theta}_1,\dots,\tilde{\theta}_p)\right)(\hat{\theta}_n-\theta_{0,n}).\end{aligned} \tag{27}$$

Using Condition 9, we obtain, when $\nabla M_n(\hat{\theta}_n)=0$ and $\hat{\theta}_n\in B(\theta_{0,n},c_{\theta,0})$, for $n\geq N_{\theta_0,\nabla}$,

$$
\begin{aligned}
&W_{1,2}\\
&\leq \frac{1}{\sqrt{c_{\theta_0,\nabla}}}\sqrt{n}\left|\left|\left(\mathbb{E}(\nabla^2 M_n(\theta_{0,n})) - \nabla^2 M_n(\tilde{\theta}_1,\ldots,\tilde{\theta}_p)\right)(\hat{\theta}_n - \theta_{0,n})\right|\right|\\
&\leq \frac{1}{\sqrt{c_{\theta_0,\nabla}}}\sqrt{n}\rho_1\left(\mathbb{E}(\nabla^2 M_n(\theta_{0,n})) - \nabla^2 M_n(\tilde{\theta}_1,\ldots,\tilde{\theta}_p)\right)||\hat{\theta}_n - \theta_{0,n}||\\
&\leq C_4\sqrt{n}\max_{j,k=1}^{p}\left|\mathbb{E}(\nabla^2 M_n(\theta_{0,n}))_{j,k} - \nabla^2 M_n(\theta_{0,n})_{j,k}\right|||\hat{\theta}_n - \theta_{0,n}||\\
&+ C_4\sqrt{n}\max_{j,k,\ell=1}^{p}\sup_{\theta\in\mathring{\Theta}}\left|\frac{\partial^3 M_n(\theta)}{\partial\theta_j\partial\theta_k\partial\theta_\ell}\right|||\hat{\theta}_n - \theta_{0,n}||^2,
\end{aligned}
$$

where, in the last inequality, $0 < C_4 < \infty$ is a constant and we have used the mean value theorem. Using Hölder's inequality together with Conditions 5 and 6, we obtain, for some constants $0 < C_5 < \infty$, $0 < C_6 < \infty$ and $N_5 \in \mathbb{N}$, for $n \geq N_5$,

$$
\begin{aligned}
\mathbb{E}(W_{1,2}) \leq & C_5\sqrt{n}\max_{j,k,=1}^{p}\operatorname{Var}(\nabla^2 M_n(\theta_{0,n})_{j,k})^{1/2}\\
&\mathbb{E}\left(\mathbb{1}_{\{\nabla M_n(\hat{\theta}_n)=0\}}\mathbb{1}_{\{\hat{\theta}_n\in B(\theta_{0,n},c_{\theta,0})\}}||\hat{\theta}_n - \theta_{0,n}||^2\right)^{1/2}\\
&+ C_5\sqrt{n}\max_{j,k,\ell=1}^{p}\mathbb{E}\left(\sup_{\theta\in\mathring{\Theta}}\left|\frac{\partial^3 M_n(\theta)}{\partial\theta_j\partial\theta_k\partial\theta_\ell}\right|^2\right)^{1/2}\\
&\mathbb{E}\left(\mathbb{1}_{\{\nabla M_n(\hat{\theta}_n)=0\}}\mathbb{1}_{\{\hat{\theta}_n\in B(\theta_{0,n},c_{\theta,0})\}}||\hat{\theta}_n - \theta_{0,n}||^4\right)^{1/2}\\
\leq & C_6\mathbb{E}\left(\mathbb{1}_{\{\nabla M_n(\hat{\theta}_n)=0\}}\mathbb{1}_{\{\hat{\theta}_n\in B(\theta_{0,n},c_{\theta,0})\}}||\hat{\theta}_n - \theta_{0,n}||^2\right)^{1/2}\\
&+ C_6\sqrt{n}\mathbb{E}\left(\mathbb{1}_{\{\nabla M_n(\hat{\theta}_n)=0\}}\mathbb{1}_{\{\hat{\theta}_n\in B(\theta_{0,n},c_{\theta,0})\}}||\hat{\theta}_n - \theta_{0,n}||^4\right)^{1/2}.
\end{aligned}
$$

We now apply Lemma 7 with the choice of the constant $\gamma_1 = 2$ there. We obtain, with some constants $0 < c_7 < \infty$, $0 < c_7' < \infty$, $0 < C_7 < \infty$ and $N_7 \in \mathbb{N}$, for $n \geq N_7$ and $0 < t \leq c_7'$,

$$
\begin{aligned}
&\mathbb{P}\left(\nabla M_n(\hat{\theta}_n) = 0, ||\hat{\theta}_n - \theta_{0,n}|| \geq t, \hat{\theta}_n \in B(\theta_{0,n}, c_{\theta,0})\right) \leq C_7\frac{\log(n)^{2p}e^{-nc_7t^2}}{t^p}\\
&+ C_7 n e^{-c_7(\log n)^2} + C_7 n e^{-c_7 n^{1/4}}.
\end{aligned} \tag{28}
$$

Let $C_8 = (p/2+4)/c_7$. Using $\mathbb{E}(X) \leq A + X_{\max}\mathbb{P}(X \geq A)$ for a non-negative random variable X bounded by $X_{\max} > 0$ and for $A > 0$, we obtain, for a constant $0 < C_9 < \infty$, for $n \geq N_7$,

$$
\begin{aligned}
&\mathbb{E}(W_{1,2}) \leq C_9\Bigg(C_8\frac{\log(n)}{n} +\\
&\mathbb{P}\left(\nabla M_n(\hat{\theta}_n) = 0, ||\hat{\theta}_n - \theta_{0,n}|| \geq \frac{\sqrt{C_8\log(n)}}{\sqrt{n}}, \hat{\theta}_n \in B(\theta_{0,n}, c_{\theta,0})\right)\Bigg)^{1/2}
\end{aligned}
$$

$$+ C_9 \left(\frac{C_8^2 (\log n)^2}{n} + \right. \tag{29}$$

$$\left. n\mathbb{P}\left(\nabla M_n(\hat{\theta}_n) = 0, ||\hat{\theta}_n - \theta_{0,n}|| \geq \frac{\sqrt{C_8 \log(n)}}{\sqrt{n}}, \hat{\theta}_n \in B(\theta_{0,n}, c_{\theta,0}) \right) \right)^{1/2}.$$

Hence from (28), for a constant $N_8 \in \mathbb{N}$, for $n \geq N_8$,

$$\mathbb{E}(W_{1,2}) \leq$$

$$C_9 \left(C_8 \frac{\log(n)}{n} + C_7 \log(n)^{2p} \frac{n^{p/2} e^{-(p/2+4)\log n}}{(C_8 \log n)^{p/2}} \right. \tag{30}$$

$$\left. + C_7 n e^{-c_7 (\log n)^2} + C_7 n e^{-c_7 n^{1/4}} \right)^{1/2} \tag{31}$$

$$+ C_9 \left(\frac{C_8^2 (\log n)^2}{n} + C_7 \log(n)^{2p} \frac{n^{p/2+1} e^{-(p/2+4)\log n}}{(C_8 \log n)^{p/2}} \right. \tag{32}$$

$$\left. + C_7 n^2 e^{-c_7 (\log n)^2} + C_7 n^2 e^{-c_7 n^{1/4}} \right)^{1/2}. \tag{33}$$

As $n \to \infty$, the quantities

$$C_7 \log(n)^{2p} \frac{n^{p/2} e^{-(p/2+4)\log n}}{(C_8 \log n)^{p/2}} \quad \text{and} \quad C_7 \log(n)^{2p} \frac{n^{p/2+1} e^{-(p/2+4)\log n}}{(C_8 \log n)^{p/2}}$$

in (30) and (32) have smaller order than $n^{-5/2}$. Similarly, as $n \to \infty$, the quantities

$$C_7 n e^{-c_7 (\log n)^2} \quad \text{and} \quad C_7 n^2 e^{-c_7 (\log n)^2}$$

in (31) and (33) have smaller order than n^{-B} for any constant B. Hence, there are constants $N_{10} \in \mathbb{N}$ and $0 < C_{10} < \infty$ such that, when $n \geq N_{10}$,

$$\mathbb{E}(W_{1,2}) \leq C_{10} \frac{\log n}{\sqrt{n}}. \tag{34}$$

Hence from (22), (23), (25), (26) and (34), we obtain, for $n \geq N_3$ and $n \geq N_{10}$,

$$\mathcal{W}_1 \left(\bar{C}_{n,0}^{-1/2} \bar{H}_{n,0} \sqrt{n} (\hat{\theta}_n - \theta_{0,n}), Z \right)$$

$$\leq \mathcal{W}_1 \left(\bar{C}_{n,0}^{-1/2} \sqrt{n} \nabla M_n(\theta_{0,n}), Z \right) + C_3 n e^{-c_3 n^{1/4}} + C_{10} \frac{\log n}{\sqrt{n}}.$$

This concludes the proof. □

Appendix B: Proofs for Section 3.2

Proof of Theorem 4. As stated previously, the function M_n is given by

$$M_n(\theta) = \frac{1}{n}\sum_{i=1}^{n}\left(-y_i x_i^T\theta + \log(1+e^{x_i^T\theta})\right)$$

where the y_i are independent random variables with values in $\{0,1\}$. Defining $X_i = (x_i, y_i)$, we are in the framework of Theorem 3, so let us check that the required conditions indeed hold.

It can be checked that there is a constant $0 < C_1 < \infty$ such that for any Y as in (10), Y is almost surely bounded by C_1 (observe that Y only takes two values). Hence the assumption (10) of sub-Gaussianity and bounded expectation holds.

Condition 1 is already assumed to hold. Condition 2 can be shown simply. Let us show that Condition 7 holds. Indeed, $\nabla\mathbb{E}(M_n(\theta_0)) = 0$ can be seen directly from (13). Furthermore, from (14), we have, for $\theta \in \mathring{\Theta}$,

$$\nabla^2\mathbb{E}(M_n(\theta)) = \nabla^2 M_n(\theta) = \frac{1}{n}\sum_{i=1}^{n}\frac{e^{x_i^\top\theta}}{(1+e^{x_i^\top\theta})^2}x_i x_i^\top.$$

Hence, from Conditions 10 and 11, there are constants $N_2 \in \mathbb{N}$ and $0 < c_2 < \infty$ such that for $n \geq N_2$ and $\theta \in \mathring{\Theta}$,

$$\lambda_p\left(\nabla^2 M_n(\theta)\right) \geq c_2. \tag{35}$$

Hence, since $\nabla\mathbb{E}(M_n(\theta_0)) = 0$, by strong convexity, Condition 7 holds.

Condition 8 is a consequence of (35). Condition 9 holds because

$$\mathrm{Cov}(\sqrt{n}\nabla M_n(\theta_0)) = \nabla^2 M_n(\theta_0)$$

(this holds because we have a well-specified likelihood model and can also be checked directly).

Finally, since all the quantities involved are uniformly bounded, Condition (O1) for checking Condition 4 holds. Condition (O2) could also be used instead, since the functions involved are all uniformly globally Lipschitz.

Hence Theorem 3 can be applied, which concludes the proof. □

Appendix C: Proofs for Section 3.3

Lemma 8. *Assume that Conditions 12 and 14 hold. There is a constant C_R such that for $n \in \mathbb{N}$,*

$$\sup_{\theta\in\Theta}\rho_1\left(R_{n,\theta}\right) \leq C_R.$$

Proof of Lemma 8. The lemma follows from (16) and from Lemma 4 in [31]. □

Lemma 9. *Assume that Conditions 12 to 14 hold. Then, we have, for* $j \in \{1, \dots, p\}$*,* $\theta \in \mathring{\Theta}$ *and* $n \in \mathbb{N}$*,*

$$(\nabla M_n(\theta))_j = \frac{1}{n} y^{(n)\top} B_{n,\theta,j} y^{(n)} \tag{36}$$

with

$$B_{n,\theta,j} = \tag{37}$$
$$2R_{n,\theta}^{-1} \mathrm{diag}(R_{n,\theta}^{-1})^{-2} \left(\mathrm{diag} \left(R_{n,\theta}^{-1} \frac{\partial R_{n,\theta}}{\partial \theta_j} R_{n,\theta}^{-1} \right) \mathrm{diag}(R_{n,\theta}^{-1})^{-1} - R_{n,\theta}^{-1} \frac{\partial R_{n,\theta}}{\partial \theta_j} \right) R_{n,\theta}^{-1}.$$

For a constant and $0 < C_B < \infty$*, we have, for* $n \in \mathbb{N}$*,*

$$\max_{j=1,\dots,p} \sup_{\theta \in \mathring{\Theta}} \rho_1 (B_{n,\theta,j}) \leq C_B. \tag{38}$$

Proof of Lemma 9. The equation (36) is proved in [9, 11]. The equation (38) follows from Condition 13, Lemma 8 and (17) and from the arguments in the proof of Proposition D.7 in [9]. □

Lemma 10. *Assume that Conditions 12 to 14 hold. Then, we have, for* $j, k \in \{1, \dots, p\}$*, for* $\theta \in \mathring{\Theta}$*, for* $n \in \mathbb{N}$*,*

$$(\nabla^2 M_n(\theta))_{j,k} = \frac{1}{n} y^{(n)\top} C_{n,\theta,j,k} y^{(n)}, \tag{39}$$

where the matrices $C_{n,\theta,j,k}$ *satisfy, for a constant* $0 < C_C < \infty$*, for* $n \in \mathbb{N}$*,*

$$\max_{j,k=1,\dots,p} \sup_{\theta \in \mathring{\Theta}} \rho_1 (C_{n,\theta,j,k}) \leq C_C. \tag{40}$$

Proof of Lemma 10. Equation (39) is shown in [9], where the matrices $C_{n,\theta,j,k}$ are obtained from the matrices

$$R_{n,\theta}, R_{n,\theta}^{-1}, \partial R_{n,\theta}/\partial \theta_j, \partial R_{n,\theta}/\partial \theta_k \text{ and } \partial^2 R_{n,\theta}/\partial \theta_k \partial \theta_j = \partial^2 R_{n,\theta}/\partial \theta_j \partial \theta_k,$$

from sums and products and from the diag operator. The precise expressions of the matrices $C_{n,\theta,j,k}$ can be found in [9]. Equation (40) is then shown similarly to (38). □

Lemma 11. *Assume that Conditions 12 to 14 hold. Then, for* $j, k, \ell \in \{1, \dots, p\}$*, for* $\theta \in \mathring{\Theta}$*, for* $n \in \mathbb{N}$*, we have*

$$\frac{\partial^3 M_n(\theta)}{\partial \theta_j \partial \theta_k \partial \theta_\ell} = \frac{1}{n} y^{(n)\top} D_{n,\theta,j,k,\ell} y^{(n)}, \tag{41}$$

where the matrices $D_{n,\theta,j,k,\ell}$ *satisfy, for some constant* $0 < C_D < \infty$*, for* $n \in \mathbb{N}$*,*

$$\max_{j,k,\ell=1,\dots,p} \sup_{\theta \in \mathring{\Theta}} \rho_1 (D_{n,\theta,j,k,\ell}) \leq C_D. \tag{42}$$

Proof of Lemma 11. The proof is the same as for Lemma 10. □

Lemma 12. *Assume that Conditions 12 to 14 hold. Then, there is a constant* $0 < C_{\partial,y} < \infty$ *such that for* $n \in \mathbb{N}$,

$$\sup_{\theta \in \mathring{\Theta}} ||\nabla M_n(\theta)|| \leq C_{\partial,y} \frac{1}{n} ||y^{(n)}||^2, \tag{43}$$

$$\sup_{\theta \in \mathring{\Theta}} \rho_1(\nabla^2 M_n(\theta)) \leq C_{\partial,y} \frac{1}{n} ||y^{(n)}||^2 \tag{44}$$

and

$$\sup_{\theta \in \mathring{\Theta}} \max_{j,k,\ell=1,\dots,p} \left| \frac{\partial^3 M_n(\theta)}{\partial \theta_j \partial \theta_k \partial \theta_\ell} \right| \leq C_{\partial,y} \frac{1}{n} ||y^{(n)}||^2. \tag{45}$$

Proof of Lemma 12. Equations (43), (44) and (45) follow from Lemmas 9, 10 and 11. □

Lemma 13. *Assume that Conditions 12, 13, 14 and 16 hold. Then, Condition 8 holds with* M_n *as in* (15) *and* $\theta_{0,n} = \theta_0$ *as after* (15).

Proof of Lemma 13. Let $\alpha, \beta \in \mathbb{R}^p$ with $\alpha_1^2 + \cdots + \alpha_p^2 = 1$ and $\beta_1^2 + \cdots + \beta_p^2 = 1$. For a matrix M, let $||M||_F$ be its Frobenius norm. We have

$$\begin{aligned} &\frac{1}{\sqrt{n}} \left| \left| \sum_{\ell=1}^p \alpha_\ell \frac{\partial R_{n,\theta_0}}{\partial \theta_\ell} - \sum_{\ell=1}^p \beta_\ell \frac{\partial R_{n,\theta_0}}{\partial \theta_\ell} \right| \right|_F \\ &\leq ||\alpha - \beta|| \frac{1}{\sqrt{n}} \sum_{\ell=1}^p \left| \left| \frac{\partial R_{n,\theta_0}}{\partial \theta_\ell} \right| \right|_F \leq C_1 ||\alpha - \beta||, \end{aligned} \tag{46}$$

with a constant $0 < C_1 < \infty$, from (17) and Lemma 4 in [31]. Hence, Condition 16 implies that

$$\liminf_{n \to \infty} \inf_{\substack{\alpha_1,\dots,\alpha_p \in \mathbb{R} \\ \alpha_1^2 + \cdots + \alpha_p^2 = 1}} \frac{1}{n} \sum_{i,j=1}^n \left(\sum_{\ell=1}^p \alpha_\ell \frac{\partial (R_{n,\theta_0})_{i,j}}{\partial \theta_\ell} \right)^2 > 0. \tag{47}$$

The inequality (47) follows from (46) and Condition 16. Indeed, if (47) does not hold we can consider a convergent subsequence of unit norm vectors of $\mathbb{R}^p$, $(\alpha_n)_{n \in \mathbb{N}}$, for which the quantity in (47) goes to zero. Considering the limit of α_n and (46) yields a contradiction to Condition 16.

We have from the proof of Proposition 3.7 in [9] that there exists a constant $0 < c_2 < \infty$ such that, for all $\alpha \in \mathbb{R}^p$ with $\alpha_1^2 + \cdots + \alpha_p^2 = 1$,

$$\sum_{k,\ell=1}^p \alpha_k \alpha_\ell (\mathbb{E}(\nabla^2 M_n(\theta_0)))_{k,\ell} \geq c_2 \frac{1}{n} \sum_{i,j=1}^n \left(\sum_{\ell=1}^p \alpha_\ell \frac{\partial (R_{n,\theta_0})_{i,j}}{\partial \theta_\ell} \right)^2.$$

Hence from (47) we obtain

$$\liminf_{n \to \infty} \lambda_p \left(\mathbb{E}(\nabla^2 M_n(\theta_0)) \right) > 0.$$ □

Lemma 14. *Assume that Conditions 12, 13, 14 and 16 hold. Then, Condition 9 holds with M_n as in (15) and $\theta_{0,n} = \theta_0$ as after (15).*

Proof of Lemma 14. Assume that for all constants $0 < c_1 < \infty$ and $N_1 \in \mathbb{N}$, there is $n \geq N_1$ such that,

$$\lambda_p(\mathrm{Cov}(\sqrt{n}\nabla M_n(\theta_0))) \leq c_1. \tag{48}$$

Then, up to extracting a subsequence, there exists a sequence of unit vectors $(v_n)_{n\in\mathbb{N}}$ of $\mathbb{R}^p$ such that

$$v_n^\top \mathrm{Cov}(\sqrt{n}\nabla M_n(\theta_0))v_n \to_{n\to\infty} 0. \tag{49}$$

Let, for $t \geq 0$ such that $\theta_0 + tv_n \in \mathring{\Theta}$,

$$M_n(t) = M_n(\theta_0 + tv_n)$$

and let $M_n'(t)$ be the derivative at t of $t \mapsto M_n(t)$. We have

$$M_n'(0) = \nabla M_n(\theta_0)^\top v_n.$$

Hence (49) implies

$$\mathrm{Var}(\sqrt{n}M_n'(0)) \to_{n\to\infty} 0. \tag{50}$$

Consider the logarithm of the likelihood

$$L_n(t) = -\frac{1}{2}\log(\det(R_{n,t})) - \frac{1}{2}y^{(n)\top}R_{n,t}^{-1}y^{(n)},$$

where $R_{n,t} = R_{n,\theta_0+tv_n}$. Let $K > 0$ be fixed, to be selected later. Then, with $L_n'(t)$ and $L_n''(t)$ the first and second derivative of $t \mapsto L_n(t)$ at t, for n such that $B(\theta_0, K/\sqrt{n}) \subset \mathring{\Theta}$,

$$\begin{aligned}\left|L_n(0) - L_n(K/\sqrt{n})\right| &\leq \frac{K}{\sqrt{n}} \sup_{|t|\leq K/\sqrt{n}} |L_n'(t)| \\ &\leq \frac{K}{\sqrt{n}}|L_n'(0)| + \left(\frac{K}{\sqrt{n}}\right)^2 \sup_{|t|\leq K/\sqrt{n}} |L_n''(t)|. \end{aligned} \tag{51}$$

Let $\mathbb{P}_{n,t}$, $\mathbb{E}_{n,t}$ and $\mathrm{Var}_{n,t}$ be the Gaussian distribution of $y^{(n)}$, and the corresponding expectation and variance, assuming that $y^{(n)}$ has mean vector zero and covariance matrix $R_{n,t}$. From the arguments in [9], $|L_n''(t)|$ is bounded by $nC_1 + C_1||y^{(n)}||^2$ and $L_n'(0)$ has expectation under $\mathbb{P}_{n,0}$ equal to zero and variance under $\mathbb{P}_{n,0}$ bounded by $C_1 n$, where C_1 can be chosen independently of $t \in [0, K]$. Hence the quantity in (51) is bounded in $\mathbb{P}_{n,0}$ probability. We also have, for n such that $B(\theta_0, K/\sqrt{n}) \subset \mathring{\Theta}$,

$$\left|L_n(0) - L_n(K/\sqrt{n})\right| \leq \frac{K}{\sqrt{n}} \sup_{|t|\leq K/\sqrt{n}} |L_n'(t)|$$

$$\leq \frac{K}{\sqrt{n}}|L_n'(K/\sqrt{n})| + 2\left(\frac{K}{\sqrt{n}}\right)^2 \sup_{|t|\leq K/\sqrt{n}} |L_n''(t)| \tag{52}$$

and, similarly as before, the quantity in (52) is bounded in $\mathbb{P}_{n,K/\sqrt{n}}$ probability. Hence, from Le Cam's first lemma (see for instance [56, Lemma 6.4]), the measures $\mathbb{P}_{n,0}$ and $\mathbb{P}_{n,K/\sqrt{n}}$ are mutually contiguous.

Now (50) and $\mathbb{E}_{n,0}(M_n'(0)) = 0$ imply that

$$\sqrt{n}M_n'(0) \to_{n\to\infty}^{\mathbb{P}_{n,0}} 0. \tag{53}$$

Hence, we have, again from Le Cam's first lemma and from (53), that

$$\sqrt{n}M_n'(0) \to_{n\to\infty}^{\mathbb{P}_{n,K/\sqrt{n}}} 0. \tag{54}$$

We have, for $t \in [0, K/\sqrt{n}]$ and n such that $B(\theta_0, K/\sqrt{n}) \in \mathring{\Theta}$,

$$\begin{aligned}
&|\mathbb{E}_{n,0}(M_n''(0)) - \mathbb{E}_{n,t}(M_n''(t))| \\
&\leq |\mathbb{E}_{n,0}(M_n''(0)) - \mathbb{E}_{n,0}(M_n''(t))| + |\mathbb{E}_{n,0}(M_n''(t)) - \mathbb{E}_{n,t}(M_n''(t))| \\
&= |\mathbb{E}_{n,0}(M_n''(0)) - \mathbb{E}_{n,0}(M_n''(t))| + \frac{1}{n}\mathrm{Tr}\left((R_{n,0} - R_{n,t})Q_{n,t}\right),
\end{aligned}$$

with

$$Q_{n,t} = \sum_{j,k=1}^{p} (v_n)_j (v_n)_k C_{n,\theta_0+tv_n,j,k}$$

from (39). Hence from (45), (40), the Cauchy-Schwarz inequality and Lemma 8, we have

$$\sup_{t\in[0,K/\sqrt{n}]} |\mathbb{E}_{n,0}(M_n''(0)) - \mathbb{E}_{n,t}(M_n''(t))| \to_{n\to\infty} 0.$$

Hence, from Lemma 13, there exist $N_2 \in \mathbb{N}$ and $0 < c_2 < \infty$ such that, for $n \geq N_2$,

$$\inf_{t\in[0,K/\sqrt{n}]} \mathbb{E}_{n,t}(M_n''(t)) \geq c_2. \tag{55}$$

Note that c_2 can be chosen independently on K while N_2 depends on K (for instance, with $c_2 = c_{\theta_0,H}/2$ as in Condition 8). Similarly as for showing (55), we can change the values of c_2 and N_2 such that, for $n \geq N_2$,

$$\inf_{t_1,t_2\in[0,K/\sqrt{n}]} \mathbb{E}_{n,t_1}(M_n''(t_2)) \geq c_2. \tag{56}$$

Again, c_2 can be chosen independently on K while N_2 depends on K. Then, from the arguments of the proof of Lemma 6, together with (56), we obtain, for n larger than a constant $N_{K,1} \in \mathbb{N}$,

$$|\mathbb{E}_{n,K/\sqrt{n}}\sqrt{n}M_n'(0)| \geq \sqrt{n}c_2\frac{K}{\sqrt{n}}.$$

Furthermore, from (36), (38) and (16) we have, for n larger than a constant $N_{K,2} \in \mathbb{N}$, $\mathrm{Var}_{n,K/\sqrt{n}}(\sqrt{n}M_n'(0)) \leq C_3$ with a constant $0 < C_3 < \infty$ that does not depend on K. Hence, by taking K large enough, the lim inf of the $\mathbb{P}_{n,K/\sqrt{n}}$-probability that $|\sqrt{n}M'_n(0)|$ is larger than one can be made arbitrarily large. This is a contradiction to (54). Hence we have a contradiction to (48), which concludes the proof. □

Proposition 2. *Let $X = (Y^\top A_1 Y, ..., Y^\top A_p Y)$ be a random vector, with $A_1, ...,$ A_p symmetric $n \times n$ matrices, and Y a Gaussian vector with covariance matrix K. Let C be the $p \times p$ matrix with coefficients*

$$C_{i,j} = 2\,\mathrm{Tr}(KA_iKA_j)$$

and Z_C be a p-dimensional centered Gaussian vector with covariance matrix C. Assume moreover that X is centered, which is the same as assuming that

$$\mathrm{Tr}(A_iK) = 0, \ \ i = 1, \dots, p.$$

Then

$$\mathcal{W}_1(X, Z_C) \leq \frac{\sqrt{\lambda_1(C)}}{\lambda_p(C)} \sqrt{2 \sum_{i,j=1,\dots,p} \mathrm{Tr}((KA_iKA_j)^2)}.$$

Note that if all eigenvalues of the A_i are at most of order $1/\sqrt{n}$, if the eigenvalues of K are bounded from above and if $\lambda_p(C)$ is bounded from below (which will be the case for our application), this bound will be of order $p/\sqrt{n}$.

Proof of Proposition 2. The proposition is a direct consequence of [47, Proposition 4.3]. □

Proof of Theorem 5. Let us check that Conditions 1 to 9 hold in order to apply Theorem 2. Condition 1 is already assumed to hold. Condition 2 holds because of Lemmas 9 to 12. Let us check the first part of Condition 3. From (15), Condition 13, (16) and Lemma 8 and as in [9], we have

$$M_n(\theta) = \frac{1}{n} y^{(n)\top} A_{n,\theta} y^{(n)}$$

with $A_{n,\theta}$ symmetric and $\sup_{\theta\in\Theta} \rho_1(A_{n,\theta}) \leq C_1$ for a constant $0 < C_1 < \infty$. By diagonalization, for each fixed $\theta \in \Theta$, there exist independent standard Gaussian variables $z_{n,\theta,1}, \dots, z_{n,\theta,n}$ and scalars $\lambda_{n,\theta,1}, \dots, \lambda_{n,\theta,n}$, such that, with a constant $0 < C_2 < \infty$,

$$\sup_{n\in\mathbb{N}} \sup_{\theta\in\Theta} \max_{i=1}^{n} |\lambda_{n,\theta,i}| \leq C_2 \text{ and } M_n(\theta) = \frac{1}{n}\sum_{i=1}^{n} \lambda_{n,\theta,i} z_{n,\theta,i}^2.$$

Hence, we can apply Bernstein's inequality (for instance Theorem 2.8.1 in [57]) and we obtain, for $0 < \epsilon \leq 1$,

$$\sup_{\theta\in\Theta} \mathbb{P}(|M_n(\theta) - \mathbb{E}(M_n(\theta))| \geq \epsilon) \leq C_3 e^{-nc_3\epsilon^2},$$

with constants $0 < c_3 < \infty$ and $0 < C_3 < \infty$ that do not depend on ϵ. Hence the first part of Condition 3 indeed holds. The second part is shown in the same way, using Lemma 9.

Let us check the first part of Condition 4. From (43), we obtain

$$\mathbb{P}\left(\sup_{\theta\in\mathring{\Theta}}||\nabla M_n(\theta)|| \geq K\right) \leq \mathbb{P}\left(C_{\delta,y}\max_{i=1}^{n}\left(y_i^{(n)}\right)^2 \geq K\right)$$
$$\leq n\max_{i=1}^{n}\mathbb{P}\left(C_{\delta,y}(y_i^{(n)})^2 \geq K\right) \leq C_4 n e^{-c_4 K},$$

with constants $0 < c_4 < \infty$ and $0 < C_4 < \infty$, from, for instance, (A.2) in [23]. Hence the first part of Condition 4 holds. The second part is shown similarly.

Condition 5, (3) follows from (43) and (44). Condition 5, (4) holds using first (45), then observing that from for instance (A.6) and (A.7) in [48], we have

$$\mathbb{E}\left(\left(\frac{1}{n}||y^{(n)}||^2\right)^2\right) = \frac{1}{n^2}\operatorname{Tr}\left(R_{n,\theta_0}\right)^2 + \frac{2}{n^2}\operatorname{Tr}\left(R^2_{n,\theta_0}\right),$$

and finally using Lemma 8.

The first part of Condition 6 is shown from Lemma 9 and, e.g., (A.7) in [48]. The second part is shown similarly from Lemma 10. In Condition 7, the offline equation follows from Condition 15 and the proof of Proposition 3.4 in [9]. Furthermore, $\mathbb{E}(\nabla M_n(\theta_0)) = 0$ is shown for instance in [9] and can also be checked directly. Thus Condition 7 holds. Condition 8 holds from Lemma 13. Condition 9 holds from Lemma 14.

Hence Theorem 2 can be applied. From this theorem, in order to conclude the proof, it is sufficient to show that, with a constant $0 < C_5 < \infty$,

$$\mathcal{W}_1\left(\bar{C}_{n,0}^{-1/2}\sqrt{n}\nabla M_n(\theta_0), Z\right) \leq \frac{C_5}{\sqrt{n}}. \tag{57}$$

The quantity $\sqrt{n}\nabla M_n(\theta_0)$ satisfies the condition of Proposition 2, with $Y = y^{(n)}$ and, for $j = 1, \dots, p$,

$$A_j = \frac{1}{2\sqrt{n}}\left(B_{n,\theta_0,j} + B^\top_{n,\theta_0,j}\right),$$

from Lemma 9. From Condition 7, then indeed $\mathbb{E}(y^{(n)\top}A_j y^{(n)}) = 0$. Then Proposition 2 yields

$$\mathcal{W}_1(\sqrt{n}\nabla M_n(\theta_0), Z_n) \leq \frac{C_6}{\sqrt{n}}, \tag{58}$$

where Z_n is a Gaussian vector with mean zero and covariance matrix $\bar{C}_{n,0}$, for a constant $0 < C_6 < \infty$, from (38), Conditions 6 and 9 and Lemma 8. Then from Lemma 1 and Condition 9,

$$\mathcal{W}_1\left(\bar{C}_{n,0}^{-1/2}\sqrt{n}\nabla M_n(\theta_0), Z\right) \leq \frac{C_6}{\sqrt{n}\sqrt{c_{\theta_0,\nabla}}}.$$

Hence, (57) is shown, which concludes the proof. □

Acknowledgments

We are grateful to the editorial team for their comments that led to an improvement of this manuscript.

References

[1] A. Anastasiou. Bounds for the normal approximation of the maximum likelihood estimator from m-dependent random variables. *Statistics & Probability Letters*, 129:171–181, 2017. MR3688530
[2] A. Anastasiou. Assessing the multivariate normal approximation of the maximum likelihood estimator from high-dimensional, heterogeneous data. *Electronic Journal of Statistics*, 12(2):3794–3828, 2018. MR3881763
[3] A. Anastasiou, K. Balasubramanian, and M. A. Erdogdu. Normal approximation for stochastic gradient descent via non-asymptotic rates of martingale CLT. In *Conference on Learning Theory*, pages 115–137. PMLR, 2019.
[4] A. Anastasiou and R. E. Gaunt. Multivariate normal approximation of the maximum likelihood estimator via the delta method. *Brazilian Journal of Probability and Statistics*, 34(1):136–149, 2020. MR4058975
[5] A. Anastasiou and R. E. Gaunt. Wasserstein distance error bounds for the multivariate normal approximation of the maximum likelihood estimator. *Electronic Journal of Statistics*, 15(2):5758–5810, 2021. MR4355697
[6] A. Anastasiou and C. Ley. Bounds for the asymptotic normality of the maximum likelihood estimator using the delta method. *ALEA, Latin American Journal of Probability and Mathematical Statistics*, 14:153–171, 2017. MR3622464
[7] A. Anastasiou and G. Reinert. Bounds for the normal approximation of the maximum likelihood estimator. *Bernoulli*, 23(1):191–218, 2017. MR3556771
[8] F. Bachoc. Cross validation and maximum likelihood estimations of hyperparameters of Gaussian processes with model mispecification. *Computational Statistics and Data Analysis*, 66:55–69, 2013. MR3064023
[9] F. Bachoc. Asymptotic analysis of the role of spatial sampling for covariance parameter estimation of Gaussian processes. *Journal of Multivariate Analysis*, 125:1–35, 2014. MR3163828
[10] F. Bachoc. *Asymptotic analysis of maximum likelihood estimation of covariance parameters for Gaussian processes: an introduction with proofs*, pages 283–303. Springer International Publishing, Cham, 2021. MR4299285
[11] F. Bachoc, J. Bétancourt, R. Furrer, and T. Klein. Asymptotic properties of the maximum likelihood and cross validation estimators for transformed Gaussian processes. *Electronic Journal of Statistics*, 14(1):1962–2008, 2020. MR4091860
[12] F. Bachoc and R. Furrer. On the smallest eigenvalues of covariance matrices of multivariate spatial processes. *Stat*, 5(1):102–107, 2016. MR3478800
[13] F. Bachoc, M. G. Genton, K. Nordhausen, A. Ruiz-Gazen, and

J. Virta. Spatial blind source separation. *Biometrika*, 107(3):627–646, 2020. MR4138980

[14] F. Bachoc, D. Preinerstorfer, and L. Steinberger. Uniformly valid confidence intervals post-model-selection. *The Annals of Statistics*, 48(1):440–463, 2020. MR4065169

[15] V. Bentkus, M. Bloznelis, and F. Götze. A Berry–Esseen bound for M-estimators. *Scandinavian Journal of Statistics*, 24(4):485–502, 1997. MR1615335

[16] J. Beran. *Statistics for long-memory processes*. Routledge, 2017. MR1304490

[17] R. Berk, L. Brown, A. Buja, K. Zhang, and L. Zhao. Valid post-selection inference. *The Annals of Statistics*, pages 802–837, 2013. MR3099122

[18] T. Bonis. Stein's method for normal approximation in Wasserstein distances with application to the multivariate central limit theorem. *Probability Theory and Related Fields*, 178(3):827–860, 2020. MR4168389

[19] S. Boucheron, G. Lugosi, and P. Massart. *Concentration inequalities: a nonasymptotic theory of independence*. Oxford university press, 2013. MR3185193

[20] G. Casella and R. L. Berger. *Statistical inference*. Pacific Grove, CA: Duxbury Press, 2021. MR1051420

[21] S. Chatterjee. A new method of normal approximation. *Annals of Probability*, 36(4):1584–1610, 2008. MR2435859

[22] S. Chatterjee. Fluctuations of eigenvalues and second order Poincaré inequalities. *Probability Theory and Related Fields*, 143(1-2):1–40, 2009. MR2449121

[23] S. Chatterjee. *Superconcentration and related topics*, volume 15. Springer, 2014. MR3157205

[24] J. Chen, X. Tan, and R. Zhang. Inference for normal mixtures in mean and variance. *Statistica Sinica*, pages 443–465, 2008. MR2432278

[25] J.-P. Chiles and P. Delfiner. *Geostatistics: modeling spatial uncertainty*. John Wiley & Sons, 2009. MR2850475

[26] N. Cressie. *Statistics for spatial data*. John Wiley & Sons, 2015. MR3559472

[27] L. Decreusefond and H. Halconruy. Malliavin and Dirichlet structures for independent random variables. *Stochastic Processes and their Applications*, 129(8):2611–2653, 2019. MR3980139

[28] O. Dubrule. Cross validation of Kriging in a unique neighborhood. *Mathematical Geology*, 15:687–699, 1983. MR0720633

[29] M. Duerinckx. On the size of chaos via Glauber calculus in the classical mean-field dynamics. *Communications in Mathematical Physics*, 382(1):613–653, 2021. MR4223483

[30] L. Fahrmeir. Maximum likelihood estimation in misspecified generalized linear models. *Statistics*, 21(4):487–502, 1990. MR1087280

[31] R. Furrer, F. Bachoc, and J. Du. Asymptotic properties of multivariate tapering for estimation and prediction. *Journal of Multivariate Analysis*, 149:177–191, 2016. MR3507322

[32] M. G. Genton and W. Kleiber. Cross-covariance functions for multivariate

geostatistics. *Statistical Science*, 30(2):147–163, 2015. MR3353096

[33] N. Gozlan. A characterization of dimension free concentration in terms of transportation inequalities. *The Annals of Probability*, 37(6):2480–2498, 2009. MR2573565

[34] M. Hallin, Z. Lu, and K. Yu. Local linear spatial quantile regression. *Bernoulli*, 15(3):659–686, 2009. MR2555194

[35] P. Heinrich and J. Kahn. Strong identifiability and optimal minimax rates for finite mixture estimation. *The Annals of Statistics*, 46(6A):2844–2870, 2018. MR3851757

[36] P. J. Huber. The behavior of maximum likelihood estimates under non-standard conditions. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability: Weather modification*, volume 5, page 221. Univ of California Press, 1967. MR0216620

[37] B. Laurent, C. Marteau, and C. Maugis-Rabusseau. Multidimensional two-component Gaussian mixtures detection. *Annales de l'Institut Henri Poincaré, Probabilités et Statistiques*, 54(2):842–865, 2018. MR3795068

[38] M. Ledoux, I. Nourdin, and G. Peccati. Stein's method, logarithmic Sobolev and transport inequalities. *Geometric and Functional Analysis*, 25(1):256–306, 2015. MR3320893

[39] J. Lv and J. S. Liu. Model selection principles in misspecified models. *Journal of the Royal Statistical Society Series B*, 76:141–167, 2014. MR3153937

[40] K. Mardia and R. Marshall. Maximum likelihood estimation of models for residual covariance in spatial regression. *Biometrika*, 71:135–146, 1984. MR0738334

[41] I. Nourdin and D. Nualart. The functional Breuer-Major theorem. *Probability Theory and Related Fields*, 176(1-2):203–218, 2020. MR4055189

[42] I. Nourdin and G. Peccati. Stein's method on Wiener chaos. *Probability Theory and Related Fields*, 145(1-2):75–118, 2009. MR2520122

[43] I. Nourdin and G. Peccati. *Normal approximations with Malliavin calculus: from Stein's method to universality*, volume 192 of *Cambridge Tracts in Mathematics*. Cambridge University Press, Cambridge, 2012. MR2962301

[44] I. Nourdin, G. Peccati, and M. Podolskij. Quantitative Breuer-Major theorems. *Stochastic Processes and their Applications*, 121(4):793–812, 2011. MR2770907

[45] I. Nourdin, G. Peccati, and G. Reinert. Second order Poincaré inequalities and CLTs on Wiener space. *Journal of Functional Analysis*, 257(2):593–609, 2009. MR2527030

[46] I. Nourdin, G. Peccati, and G. Reinert. Invariance principles for homogeneous sums: universality of Gaussian Wiener chaos. *Annals of Probability*, 38(5):1947–1985, 2010. MR2722791

[47] I. Nourdin, G. Peccati, and A. Réveillac. Multivariate normal approximation using Stein's method and Malliavin calculus. *Annales de l'Institut Henri Poincaré Probabilités et Statistiques*, 46(1):45–58, 2010. MR2641769

[48] M. S. Paolella. *Linear models and time-series analysis: regression, ANOVA, ARMA and GARCH*. John Wiley & Sons, 2018. MR3839332

[49] I. Pinelis. Optimal-order uniform and nonuniform bounds on the rate of

convergence to normality for maximum likelihood estimators. *Electronic Journal of Statistics*, 11(1):1160–1179, 2017. MR3634332

[50] I. Pinelis and R. Molzon. Optimal-order bounds on the rate of convergence to normality in the multivariate delta method. *Electronic Journal of Statistics*, 10(1):1001–1063, 2016. MR3486424

[51] B. M. Pötscher and I. R. Prucha. *Dynamic nonlinear econometric models: asymptotic theory*. Springer Science & Business Media, 2013. MR1468737

[52] L. Pronzato and A. Pázman. Design of experiments in nonlinear models. *Lecture Notes in Statistics*, 212:1, 2013. MR3058804

[53] N. Ross. Fundamentals of Stein's method. *Probability Surveys*, 8:210–293, 2011. MR2861132

[54] Q.-M. Shao and Z.-S. Zhang. Berry-Esseen bounds for multivariate nonlinear statistics with applications to M-estimators and stochastic gradient descent algorithms. *Bernoulli*, 28(3):1548–1576, 2022. MR4411502

[55] R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society Series B*, 58:267–288, 1996. MR1379242

[56] A. W. Van der Vaart. *Asymptotic statistics*, volume 3. Cambridge University Press, 2000. MR1652247

[57] R. Vershynin. *High-dimensional probability: an introduction with applications in data science*, volume 47 of *Cambridge Series in Statistical and Probabilistic Mathematics*. Cambridge University Press, Cambridge, 2018. MR3837109

[58] H. Wackernagel. *Multivariate geostatistics: an introduction with applications*. Springer Science & Business Media, 2013. MR3074883

[59] H. White. Maximum likelihood estimation of misspecified models. *Econometrica: Journal of the Econometric Society*, 50(1):1–25, 1982. MR0640163

[60] H. Zhang. Inconsistent estimation and asymptotically equivalent interpolations in model-based geostatistics. *Journal of the American Statistical Association*, 99:250–261, 2004. MR2054303

[61] H. Zhang and Y. Wang. Kriging and cross validation for massive spatial data. *Environmetrics*, 21:290–304, 2010. MR2842244