Proceedings of Machine Learning Research vol 178:1–23, 2022 35th Annual Conference on Learning Theory

Strong Gaussian Approximation for the Sum of Random Vectors

Nazar Buzun BUZUN@AIRI.NET
Artificial Intelligence Research Institute. 32 Kutuzovsky Ave., Moscow, 121170, Russian Federation.

Nikolay Shvetsov N.SHVETSOV@SKOLTECH.RU and **Dmitry V. Dylov** D.DYLOV@SKOLTECH.RU
Skolkovo Institute of Science and Technology, Bolshoi blvd., 30/1, Moscow, 121205, Russian Federation

Editors: Po-Ling Loh and Maxim Raginsky

Abstract

This paper derives a new strong Gaussian approximation bound for the sum of independent random vectors. The approach relies on the optimal transport theory and yields *explicit* dependence on the dimension size p and the sample size n. This dependence establishes a new fundamental limit for all practical applications of statistical learning theory. Particularly, based on this bound, we prove approximation in distribution for the maximum norm in a high-dimensional setting ($p > n$).

Keywords: Gaussian approximation, Central Limit Theorem, Wasserstein distance, High-dimensional statistics.

1. Introduction

Gaussian approximation for the sum of random vectors attracts the attention of mathematicians because of the uncertain dependence of the outcome on the dimension size p, Bentkus (2003); Zaitsev (2013); Chernozhukov et al. (2014). In high dimensions, it is more meaningful to search for a strong Gaussian approximation rather than estimate the proximity of distributions, because in the latter case, each random vector and their sum have to be defined in the same probability space. As such, finding an accurate strong bound for the sum is recognized as a preferred alternative, becoming one of the most important tasks in the field of limit theorems of the probability theory.

In (Zaitsev, 2013), the uncertain property of the approximation was discerned, yielding the following finite-sample bound $\forall t \geq 0$:

$$\mathbb{P}\left(\sup_{1\leq h\leq n}\left\|\frac{1}{\sqrt{n}}\sum_{i=1}^{h}\boldsymbol{\xi}_i-\frac{1}{\sqrt{n}}\sum_{i=1}^{h}\boldsymbol{\gamma}_i\right\|>\frac{C_1(\alpha)p^{\frac{23}{4}+\alpha}\log p\log n+t\,C_2p^{\frac{7}{2}}\log p}{\sqrt{n}}\right)\leq e^{-t}, \quad (1)$$

for some constants $C_1(\alpha)$, C_2, and under conditions that the vectors $\{\boldsymbol{\xi}_i\}_{i=1}^n$ are centered and independent, $\sum_{i=1}^n \boldsymbol{\xi}_i$ and $\sum_{i=1}^n \boldsymbol{\gamma}_i \in \mathcal{N}(0, n\Sigma)$ have the same variance matrix and finite exponential moments over the same domain in $\mathbb{R}^p$. The handicap of this result is the power of p, which requires extremely large sample sizes n for the approximation to be of any practical value. Asymptotically, one needs $n > p^{11}$ samples to apply the approximation.

On the other hand, the general-case multivariate Gaussian approximation in distribution, studied in Bentkus (2003), has been known to yield an alternative error bound that favors smaller sample size. Namely, for all convex events $\mathcal{A}$:

$$\left|\mathbb{P}\left(\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\boldsymbol{\xi}_i\in\mathcal{A}\right)-\mathbb{P}\left(\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\boldsymbol{\gamma}_i\in\mathcal{A}\right)\right|\leq O\left(\frac{p^{7/4}}{\sqrt{n}}\right). \quad (2)$$

For a particular case of Euclidean norm, it was proven in Buzun (2019) that the upper bound approaches the asymptotic value $p/\sqrt{n}$. Naturally, the observed gap in inequalities (1) and (2) suggests that the dependence on p in the strong Gaussian approximation can be improved.

In this paper, we narrow this gap by deriving a new type of strong Gaussian approximation bound for the sum of independent random vectors. Our approach is based on the optimal transport theory and the Gaussian approximation with the Wasserstein distance. We derive the approximation in probability in a simplified form (without the maximum by h, refer to expression (1)) and the found convergence rate is comparable with the approximation in distribution (Refs. Bentkus (2003); Buzun (2019)). Specifically, given the sub-Gaussian assumption for vectors $\{\boldsymbol{\xi}_i\}_{i=1}^n$, we prove that

$$\mathbb{P}\left(\left\|\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\boldsymbol{\xi}_i - \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\boldsymbol{\gamma}_i\right\| > O\left(\frac{p^{3/2}\log p \log n}{\sqrt{n}}\right) e^{t/\log(np)}\right) \le e^{-t}, \quad \forall t \ge 0$$

.

Related work. In Erdös and Kac (1946), it was observed that the limit distributions of some functional of the growing sums of independent identically distributed random variables (with a finite variance) do not depend on the distribution of the individual terms and, therefore, an approximation can be computed if the distribution of the terms has a specific simple form. In Einmahl (1989), the work reports new multidimensional results for the accuracy of the strong Gaussian approximation for infinite sequences of sums of independent random vectors. Consequently, Gaussian approximation of the sums of independent random vectors with finite moments Zaitsev (2007), including multidimensional approaches Zaitsev (2001); Sakhanenko (2006); Götze and Zaitsev (2009); Zaitsev (2013), have been reported.

Notably, the strong Gaussian approximation can help approximate the maximum sum of random vectors in distribution for the high-dimensional case ($p > n$). Gaussian approximation of the maximum function is very useful for justifying the Bootstrap validity and for approximating the distributions with different statistics in high-dimensional models. Besides, the aforementioned papers Chernozhukov et al. (2014, 2017), some relevant results can also be found in works Koike et al. (2019) and Sun (2020), where the authors rely on Malliavin calculus and high-order moments to assess the corresponding bounds. In Fang and Koike (2020), the authors go after the same approximation as reported herein; however, using a completely different tool: the Stein method instead of the Wasserstein distance, yielding a result that is only valid under the constraint that the measure of i.i.d. vectors $\boldsymbol{\xi}_i$ has the Stein kernel (a consequence of the log-concavity).

In the most recent works Chernozhukov et al. (2019), the result of Chernozhukov et al. (2014) was superseded by considering specific distribution of the max statistic in high dimensions. This statistic takes the form of the maximum over the components of the sum of independent random vectors and its distribution plays a key role in many high-dimensional econometric problems. The new iterative randomized Lindeberg method allowed the authors to derive new bounds for the distributional approximation errors. Specifically, in Chernozhukov et al. (2020), new nearly optimal bounds ($B_n \log^{3/2} p/\sqrt{n}$) for the Gaussian approximation over the class of rectangles were obtained, functional in the case when all components of vectors $\boldsymbol{\xi}_i$ are bounded by B_n and the covariance matrix of the scaled average is non-degenerate. The authors also demonstrated that the bounds can be further improved in some special smooth and zero-skewness cases.

Practical value. The demand for strong Gaussian approximation can frequently be encountered in a variety of statistical learning problems, being very important for developing efficient approximation algorithms. Likewise, strong Gaussian approximation is sought after in many applications across different domains (physics, biology, economy, *etc*.), making it one of the most desired tools of modern probability theory. Practical demand for the approximation are widespread and range from change point detection Buzun and Avanesov (2017), to variance matrix estimation Avanesov and Buzun (2018), to selection of high-dimensional sparse regression models Chernozhukov et al. (2013), to adaptive specification testing Chernozhukov et al. (2013), to anomaly detection in periodic signals Shvetsov et al. (2020), and to many others.

In structuring the article, we followed the common rule: the main proofs representing our contributions are kept in the main text, and the rest of the auxiliary technical proofs are moved to the appendix. The only exception is our Theorem 10, the proof of which is very long. Hence, we included it in the main part in a shortened version, while placing the complete rigorous proof to the Appendix. We highlighted the dependence between all theorems and lemmas in the visual diagram in the beginning of the Appendix.

2. Main Results

In this work, we prove the Theorem that finds the upper bound for the strong Gaussian approximation. Herein, we consider a sum of independent zero-mean random vectors $\boldsymbol{\xi} = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \boldsymbol{\xi}_i$ in $\mathbb{R}^p$ that has a covariance matrix

$$\varSigma = \mathbb{E} \boldsymbol{\xi} \boldsymbol{\xi}^T.$$

Its norm $\|\varSigma\|$ will be used in the expressions below, implying the largest eigenvalue of the matrix. A Gaussian random vector $\boldsymbol{\gamma} \in \mathcal{N}(0, \varSigma)$ is assumed to have the same 1-st and 2-nd moments.

A very useful tool in approximation in probability is the Wasserstein distance. It reveals the joint distribution and minimises the difference between two random variables making them dependent with some fixed marginal distributions. Denote by $\pi(\boldsymbol{\xi}, \boldsymbol{\gamma})$ a joint distribution of $\boldsymbol{\xi}$ and $\boldsymbol{\gamma}$. By definition, the Wasserstein distance is

$$W_L^L(\boldsymbol{\xi}, \boldsymbol{\gamma}) = \min_{\pi \in \varPi[\boldsymbol{\xi}, \boldsymbol{\gamma}]} \left\{ \mathbb{E} \|\boldsymbol{\xi} - \boldsymbol{\gamma}\|^L \right\}. \tag{3}$$

Notation $\pi \in \varPi[\boldsymbol{\xi}, \boldsymbol{\gamma}]$ means that $\int \pi(\boldsymbol{\xi}, d\boldsymbol{\gamma})$ yields the distribution of $\boldsymbol{\xi}$ and $\int \pi(d\boldsymbol{\xi}, \boldsymbol{\gamma})$ yields the distribution of $\boldsymbol{\gamma}$. Throughout the text, we will say that a vector $\boldsymbol{\xi}$ has restricted exponential or sub-Gaussian moments if $\exists\, \mathrm{g} > 0$:

$$\log \mathbb{E} \exp\left(\boldsymbol{\lambda}^\top \boldsymbol{\xi}\right) \le \|\boldsymbol{\lambda}\|^2/2, \quad \forall \boldsymbol{\lambda} \in \mathbb{R}^p, \quad \|\boldsymbol{\lambda}\| \le \mathrm{g}. \tag{4}$$

Below, we obtain Gaussian approximation bound for W_L, with its proof being the most difficult part of this manuscript. Then, we will use it to prove a new strong Gaussian approximation result under l_2-norm with an improved dependence on the dimension p. Additionally, we will propose a new method to derive the high-dimensional Central Limit Theorem under the $\max$-norm.

Theorem 1 *For the random vectors $\{\boldsymbol{\xi}_i\}_{i=1}^n$ defined above, assume that $\varSigma$ is non-singular and $\forall i : \frac{1}{\nu_0} \varSigma^{-1/2} \boldsymbol{\xi}_i$ is sub-Gaussian (4) with $0.3\, \mathrm{g} \ge \sqrt{p}$. The Wasserstein distance between $\boldsymbol{\xi}$ and $\boldsymbol{\gamma}$ with the cost function $\|x - y\|^L$, where $L \ge 2$, has the upper bound*

$$W_L(\boldsymbol{\xi}, \boldsymbol{\gamma}) \le \frac{C\, L^{3/2} \nu_0^2 \|\varSigma\|^{1/2}}{\log L} \left(\frac{p}{\sqrt{n}} + \frac{L + L^2/g^2}{n^{1-1/L}} \right) + \frac{5 L \nu_0^3 p^{3/2} \|\varSigma\|^{1/2} \log(2n)}{\sqrt{n}},$$

for some absolute constant C.

Remark 2 *The main advantage of this theorem is the explicit sharp dependence on parameters n, p, and L. The sharpness of the formula is numerically validated in the simulation studies as a function of n (ref. Section 5). Of particular practical value is the simplified asymptotic of this upper bound:*

$$O\left(\frac{L^{3/2}p + Lp^{3/2}\log n}{\sqrt{n}}\right).$$

Here, we reduced the component $(L + L^2/g^2)/n^{1-1/L}$, taking into account that $p/\sqrt{n}$ is greater than it when $L \le pn^{1-1/L-1/2}$. This agrees with the result of the Proposition 5.1 in Bobkov (2018), where the same bound was derived for the one-dimensional case ($p = 1$). The asymptotic by L is better than $L^{3(d+1)/2}$ from the Proposition 5.1 in Bobkov (2018), where a separate parameter $d \ge 1$ was needed depending on the distribution of ξ.

We assume the condition $L \ge 2$, however, the Jensen's inequality guarantees that $W_1 \le W_2$, which serves as an asymptotic for W_1. This bound also remains sharp up to the $\log n$ factor when $L = 1$, supporting the asymptotic bound of $O(p^{3/2}/\sqrt{n})$ obtained in Bentkus (2003) and Buzun (2019) for the Gaussian approximation in W_1. Bonis (2019) also studied the upper bound for the multi-dimensional W_L, without deriving the explicit dependence on L and considering only a particular case of $L = 2$.

Theorem 3 *Under the conditions from Theorem 1, there exists a Gaussian vector $\boldsymbol{\gamma} \in \mathcal{N}(0, \Sigma)$, dependent on $\boldsymbol{\xi}$, such that $\forall t \ge 0$*

$$I\!\!P\left(\|\boldsymbol{\xi} - \boldsymbol{\gamma}\| > C(n,p)\|\Sigma\|^{1/2}\, e^{t/\log(np)}\right) \le e^{-t},$$

where

$$C(n,p) = C\,\nu_0^2\,\frac{p\log(np)^{3/2}}{\sqrt{n}} + 5\nu_0^3\,\frac{p^{3/2}\log(np)\log(2n)}{\sqrt{n}},$$

for some absolute constant C.

Remark 4 *Note that the convergence in probability in the Central Limit Theorem does not exist, i.e., there is no random vector $\boldsymbol{\gamma}$, such that $\boldsymbol{\xi}(n)$ converges to it when $n \to \infty$ (it follows from the Kolmogorov's 0-1 law). However, at the same time, one could take a new Gaussian vector that depends on n to yield*

$$\|\boldsymbol{\xi}(n) - \boldsymbol{\gamma}(n)\| \xrightarrow{p} 0, \quad n \to \infty,$$

which follows from Theorem 3.

Proof Using the notation $\pi \in \varPi[\boldsymbol{\xi}, \boldsymbol{\gamma}]$ and the Markov's inequality

$$\min_{\pi \in \varPi[\boldsymbol{\xi},\boldsymbol{\gamma}]} I\!\!P\left(\|\boldsymbol{\xi} - \boldsymbol{\gamma}\| > \varDelta\right) \le \frac{1}{\varDelta^L} \min_{\pi \in \varPi[\boldsymbol{\xi},\boldsymbol{\gamma}]} I\!\!E\,\|\boldsymbol{\xi} - \boldsymbol{\gamma}\|^L = \frac{W_L^L(\boldsymbol{\xi},\boldsymbol{\gamma})}{\varDelta^L},$$

we can approximate W_L from Theorem 1 by setting parameters

$$L = \log(np) \ \text{ and } \ \varDelta = C(n,p)\|\Sigma\|^{1/2}\, e^{t/\log(np)}.$$

■

Gaussian approximation in Wasserstein distance also gives us a new instrumentation to approach the high-dimensional Central Limit Theorem. In this case, we will compare the distributions of $\max \boldsymbol{\xi}$ and $\max \boldsymbol{\gamma}$, keeping in mind that one can always derive an approximation in distribution from the approximation in probability. Henceforth, we will apply a component-wise bound of the one-dimensional variant of Theorem 1 to derive the result in high-dimensions.

Theorem 5 *Let $\{\boldsymbol{\xi}_i\}_{i=1}^n$ be i.i.d random vectors. Additionally, assume that $\forall i, k : 1 \le i \le n$, $1 \le k \le p$, each random variable $\xi_{ik}/(\nu_0 \Sigma_{kk})$ is sub-Gaussian (ref. Expression 4 with $p = 1$ and $\mathtt{g} \ge 3.4$). Denote $\xi_k = \sum_i \xi_{ik}$. Restrict the values of the diagonal elements of the covariance matrix, such that $\forall k : \underline{\sigma}^2 \le \Sigma_{kk} \le \overline{\sigma}^2$, and denote $\lambda_{\min}$ to be the minimal eigenvalue of the matrix Σ. Then, $\forall t \in I\!R$*

$$\left| I\!P\left(\max_k \xi_k < t\right) - I\!P\left(\max_k \gamma_k < t\right)\right| \le O\left(\nu_0^3 \frac{\overline{\sigma}}{\underline{\sigma}} \frac{\log^2(np)}{\sqrt{n}} + \nu_0^3 \frac{\overline{\sigma}^2}{\lambda_{\min}} \frac{\log(np)^{5/2} \log n}{\sqrt{n}}\right).$$

Remark 6 *To the best of our knowledge, $O(\log^2 p/\sqrt{n})$ is the most accurate upper bound for the high-dimensional Central Limit Theorem reported thus far (Theorem 2.2 in Chernozhukov et al. (2020)), proved under assumptions $p > n$ with some absolute constants ν_0 and $\lambda_{\min}$). Theorem 5 agrees with this approximation but does not make it more accurate. However, we will show that Theorems 5 and 1 allow for an effortless derivation of this asymptotic rate. Corresponding experimental verification is given in Section 5.*

3. Gaussian Approximation in Wasserstein Distance

This section is devoted to the proof of Theorem 1 and the auxiliary Theorems and Lemmas.

3.1. Technical Notation

We will use the *multi-indices* $\alpha \in \mathbb{N}^p$. Define the multi-index power, the derivative, the absolute value, and the factorial, $\forall x \in I\!R^p$ and a function $f : I\!R^p \to I\!R$, by the following:

$$x^\alpha = \prod_{k=1}^p x_k^{\alpha_k}, \quad \partial^\alpha f(x) = \frac{\partial^{\alpha_1}}{\partial x_1^{\alpha_1}} \dots \frac{\partial^{\alpha_p}}{\partial x_p^{\alpha_p}} f(x), \quad |\alpha| = \sum_{k=1}^p \alpha_k, \quad \alpha! = \prod_{k=1}^p \alpha_k!\,.$$

The summation over the multi-index means the sum is taken over all of its components, having an additional condition about the possible summation region. For example, with indicator function $1\!\!1$

$$\sum_{\alpha \ge 0} = \sum_{\alpha_1=0}^\infty \dots \sum_{\alpha_p=0}^\infty, \quad \sum_{|\alpha|=1} = \sum_{\alpha_1=0}^\infty \dots \sum_{\alpha_p=0}^\infty 1\!\!1[|\alpha| = 1]$$

Given the definitions of the multi-indices, the Taylor series expansion $\forall x, x_0 \in I\!R^p$ and an infinitely differentiable f, takes the following form:

$$f(x) = \sum_{\alpha \ge 0} \frac{\partial^\alpha f(x_0)}{\alpha!}(x - x_0)^\alpha = f(x_0) + \sum_{|\alpha|=1} \frac{\partial^\alpha f(x_0)}{\alpha!}(x - x_0)^\alpha + \sum_{|\alpha|=2} \dots\,.$$

Denote by $H_\alpha(y,\Sigma)$ the multivariate Hermite polynomials, such that $\forall y \in \mathbb{R}^p$

$$H_\alpha(y,\Sigma) = (-1)^{|\alpha|} e^{\frac{y^T\Sigma^{-1}y}{2}} \partial^\alpha e^{-\frac{y^T\Sigma^{-1}y}{2}}. \tag{5}$$

Note that in the Taylor expansion with the Gaussian random vectors $\boldsymbol{\gamma}$, we may compute the derivatives by using the formula $\partial^\alpha \mathbb{E} f(x+\boldsymbol{\gamma}) = \mathbb{E} H_\alpha(\boldsymbol{\gamma},\Sigma) f(x+\boldsymbol{\gamma})$.

3.2. Proof of Theorem 1

In this Section, we will use a special version of Stein's method Chen et al. (2010), proposed in Bonis (2019), that was adjusted for the Wasserstein distance, borrowing the upper bound for the transport mapping from Otto and Villani (2000). We first prove the approximation of the derivative of $W_L(\boldsymbol{\xi},\boldsymbol{\gamma})$. For that, define a smooth transition from $X_0 = \boldsymbol{\xi}$ to $X_\infty = \boldsymbol{\gamma}$, parametrized by t:

$$X_t = e^{-t}\boldsymbol{\xi} + \sqrt{1-e^{-2t}}\boldsymbol{\gamma}, \quad t \in [0,\infty]. \tag{6}$$

This definition, with the vectors $\boldsymbol{\xi}$ and $\boldsymbol{\gamma}$ being independent, determines the distribution of X_t.

Remark 7 *The choice of specific transition in Eq. (6) is justified by the fact that it corresponds to Markov semi-group $\{P_t\}_{t\geq 0}$, where $P_t f(x) = \mathbb{E} f(e^{-t}x + \sqrt{1-e^{-2t}}\boldsymbol{\gamma})$ for all bounded measurable functions f, also known as the Ornstein–Uhlenbeck semi-group. Various applications of the Stein's method use this widespread tool, because of its suitable generator $Af(x) = \nabla^T \Sigma \nabla f(x) - x^T \nabla f(x)$, $\forall f \in \mathbb{C}^2$ and a Gaussian stationary measure Bakry et al. (2014).*

Let us note one important property of this process, which will help us find an upper bound for the derivative of W_L by t. There exists a function $\mathcal{G}_t(x)$ that affects the transition of a measure $\mu_t(x)$ of X_t in the following way $\forall f \in \mathbb{C}^1$:

$$\frac{d}{dt}\int f(x)d\mu_t(x) = \int \nabla^T f(x) \nabla \mathcal{G}_t(x) d\mu_t(x), \quad t \in [0,\infty), \tag{7}$$

and its gradient has explicit representation $\nabla \mathcal{G}_t(X_t) = \mathbb{E}\left[\frac{d}{dt}X_t \middle| X_t\right]$. The last claim follows from the following expression:

$$\begin{aligned}\frac{d}{dt}\int f(x)d\mu_t(x) &= \lim_{\Delta\to 0} \frac{\mathbb{E} f(X_{t+\Delta}) - \mathbb{E} f(X_t)}{\Delta} \\ &= \lim_{\Delta\to 0} \mathbb{E}\nabla^T f(X_t) \frac{\mathbb{E}[X_{t+\Delta} - X_t | X_t]}{\Delta} = \mathbb{E}\nabla^T f(X_t) \mathbb{E}\left[\frac{d}{dt}X_t \middle| X_t\right].\end{aligned}$$

Differentiating Eq. (6), we find the gradient

$$\nabla \mathcal{G}_t(X_t) = -e^{-t}\mathbb{E}\left(\boldsymbol{\xi} - \frac{e^{-t}}{\sqrt{1-e^{-2t}}}\boldsymbol{\gamma} \middle| X_t\right). \tag{8}$$

The next Lemma allows us to incrementally bound the Wasserstein distance in a small interval of the transition parameter t.

Lemma 8 (proof in Appendix B) *Let $\nabla \mathcal{G}_t(X_t) \in \mathbb{R}^p$, $t \in [0, \infty]$ be uniformly continuous and bounded random process. For an infinitely small time shift ds and any $a, b, L \geq 0$, the following inequalities are valid:*

$$W_L^L(X_t, X_{t+ds}) \leq \mathbb{E}\|ds\nabla \mathcal{G}_t(X_t)\|^L,$$

$$W_L(X_a, X_b) \leq \int_a^b \left\{\mathbb{E}\|\nabla \mathcal{G}_t(X_t)\|^L\right\}^{1/L} dt.$$

According to the base procedure of the Stein's method Chen et al. (2010), we should consequently replace the random vectors $\boldsymbol{\xi}_i$ by their independent copies $\boldsymbol{\xi}_i'$, while drawing a random index at each step. This allows to estimate a change of some function of interest that depends on the transition process X_t (6). In our case, the function is $\nabla \mathcal{G}_t(X_t)$ (8). For a formal description, consider two additional random vectors with random uniform index $I \in \{1, \dots, n\}$

$$\boldsymbol{\xi}'(t) = \boldsymbol{\xi} + \frac{\boldsymbol{\xi}_I' - \boldsymbol{\xi}_I}{\sqrt{n}} \, \mathbb{I}_I(t), \quad \mathbb{I}_I(t) = \mathbb{I}\left[\frac{\|\Sigma^{-1/2}(\boldsymbol{\xi}_I' - \boldsymbol{\xi}_I)\|}{\sqrt{n}} \leq \left(\frac{e^{2t}-1}{L}\right)^{1/2}\right]$$

and

$$\boldsymbol{\tau}(t) = \mathbb{E}\left[\frac{n}{2}(\boldsymbol{\xi}'(t) - \boldsymbol{\xi})\left(1 + \sum_{\alpha \geq 0} \frac{(\boldsymbol{\xi}'(t) - \boldsymbol{\xi})^\alpha H_\alpha(\boldsymbol{\gamma}, \Sigma)}{\alpha!(e^{2t}-1)^{|\alpha|/2}}\right)\middle|\boldsymbol{\xi}, \boldsymbol{\gamma}\right]. \tag{9}$$

The intuition behind the previous definition is that, if $E(\boldsymbol{\tau}(t)|X_t) = 0$, then, the following bound holds for X_t and for any jointly-measurable random vector $\boldsymbol{\eta}$:

$$\mathbb{E}\|\mathbb{E}(\boldsymbol{\eta}|X_t)\| = \mathbb{E}\|\mathbb{E}(\boldsymbol{\eta} + \boldsymbol{\tau}(t)|X_t)\| \leq \mathbb{E}\|\boldsymbol{\eta} + \boldsymbol{\tau}(t)\|.$$

Hence, we chose $\tau(t)$ that minimizes the last expression under the condition $E(\tau(t)|X_t) = 0$. We restrict the term $\|\Sigma^{-1/2}(\boldsymbol{\xi}_I' - \boldsymbol{\xi}_I)\|$ from above in order to avoid irregularity in the random vector $\boldsymbol{\tau}(t)$ when $t \to 0$. Note that for any smooth function f, according to the construction of $\boldsymbol{\xi}'(t)$ (ref. Lemma 10 in Bonis (2019)),

$$2\mathbb{E}\boldsymbol{\tau}(t)f(X_t) = n\mathbb{E}\left(X_t' - X_t\right)\left(f(X_t') + f(X_t)\right) = 0,$$

where $X_t' = e^{-t}\boldsymbol{\xi}'(t) + \sqrt{1 - e^{-2t}}\boldsymbol{\gamma}$. And as a consequence, indeed $\mathbb{E}(\boldsymbol{\tau}(t)|X_t) = 0$. Hence, one may note that $\boldsymbol{\tau}(t)$ merely shifts the process X_t without changing its measure. Oppositely, $\nabla \mathcal{G}_t(X_t)$ shifts X_t outside its measure towards $\boldsymbol{\gamma}$. To reduce the unconditional expectation of $\nabla \mathcal{G}_t(X_t)$, we will substitute $\boldsymbol{\tau}(t)$ within its expression. Accounting to the definition of $\nabla \mathcal{G}_t(X_t)$ (8), we obtain:

$$\nabla \mathcal{G}_t(X_t) = \nabla \mathcal{G}_t(X_t) - e^{-t}\mathbb{E}(\boldsymbol{\tau}(t)|X_t) = -e^{-t}\mathbb{E}\left(\boldsymbol{\xi} - \frac{1}{\sqrt{e^{2t}-1}}\boldsymbol{\gamma} + \boldsymbol{\tau}(t)\middle|X_t\right). \tag{10}$$

We, thus, added the zero component to the initial value of $\nabla \mathcal{G}_t(X_t)$. Below, we will use Jensen's inequality, moving the conditional expectation outside the norm, with $\boldsymbol{\tau}(t)$ decreasing its expected value. The vector $\boldsymbol{\tau}(t)$ includes convolutions with H_α, the evaluation of which requires the next technical Lemma.

Lemma 9 (proof in Appendix C) *Let $\boldsymbol{\gamma} \in \mathcal{N}(0, I)$ and for all $\alpha \in \mathbb{N}^p$: $\boldsymbol{\zeta}(\alpha) \in \mathcal{P}(I\!R^p)$ and $H_\alpha(y, I)$ be the multi-variate Hermite polynomials (5); and let $\boldsymbol{\zeta}(\alpha) = \sum_i \boldsymbol{\zeta}_i(\alpha)$, where $\{\boldsymbol{\zeta}_i(\alpha)\}$ are independent random vectors. Then, for $2 \le q < \infty$,*

$$\left[I\!E \left\| \sum_{i,\alpha \ge 0} \boldsymbol{\zeta}_i(\alpha) H_\alpha(\boldsymbol{\gamma}, I) \right\|^q \right]^{1/q} \le C(q) \left[\sum_{i,\alpha \ge 0} \alpha!(q-1)^{|\alpha|} I\!E \left\| \boldsymbol{\zeta}_i(\alpha) \right\|^2 \right]^{1/2}$$
$$+ \left[\sum_{\alpha \ge 0} \alpha!(q-1)^{|\alpha|} \left\| I\!E \boldsymbol{\zeta}(\alpha) \right\|^2 \right]^{1/2}$$
$$+C(q) \left[\sum_i I\!E \left(\sum_{\alpha \ge 0} \alpha!(q-1)^{|\alpha|} \left\| \boldsymbol{\zeta}_i(\alpha) \right\|^2 \right)^{q/2} \right]^{1/q},$$

where $C(q) = Cq/\log q$, for some constant C.

Now, Lemma 9 and the Expression (10) allow us to find the upper bound of $I\!E \|\nabla \mathcal{G}_t(X_t)\|^L$. Applying this bound to Lemma 8, we will obtain the final Gaussian approximation in Wasserstein distance, as demonstrated below.

Theorem 10 *Let $\boldsymbol{\gamma} \in \mathcal{N}(0, \Sigma)$, $\boldsymbol{\xi} = \sum_{i=1}^n \boldsymbol{\xi}_i / \sqrt{n}$, and $\{\boldsymbol{\xi}_i\}_{i=1}^n$ be independent random vectors with zero mean and covariance $I\!E \boldsymbol{\xi} \boldsymbol{\xi}^T = \Sigma$. The Wasserstein distance between $\boldsymbol{\xi}$ and $\boldsymbol{\gamma}$, with the cost function $\|x - y\|^L$, where $L \ge 2$, has the following upper bound:*

$$\frac{W_L(\boldsymbol{\xi}, \boldsymbol{\gamma})}{\|\Sigma\|^{1/2}} \le \frac{C\, L^{3/2}}{\log L} \left(\mu_4^{1/2} + \mu_{2L}^{1/L} \right) + \frac{\sqrt{2} e^{1/2} L^{1/2} \mu_2}{\sqrt{n}} + \frac{e^{1/2} L \mu_3}{2} \log(2n),$$

where C is some absolute constant and $\boldsymbol{\xi}_i'$ are independent copies of $\boldsymbol{\xi}_i$ and

$$\mu_k = \frac{1}{n^{k/2}} \sum_{i=1}^n I\!E \left\| \Sigma^{-1/2} (\boldsymbol{\xi}_i - \boldsymbol{\xi}_i') \right\|^k.$$

Proof [Only the major steps are shown; ref. Appendix D for the complete proof] From expression $W_L(\boldsymbol{\xi}, \boldsymbol{\gamma}) \le \|\Sigma\|^{1/2} W_L(\Sigma^{-1/2} \boldsymbol{\xi}, \Sigma^{-1/2} \boldsymbol{\gamma})$ follows that without loss of generality we may assume that $\Sigma = I$. Denote $\Sigma_i = I\!E \boldsymbol{\xi}_i \boldsymbol{\xi}_i^T / n$, with the evident property $\sum_{i=1}^n \Sigma_i = I$. Use a short notation $I\!E_{X_t}$ for the expectation operator with the X_t condition. Then, unwrap the expression (10):

$$I\!E \left(\boldsymbol{\xi} - \frac{1}{\sqrt{e^{2t} - 1}} \boldsymbol{\gamma} + \boldsymbol{\tau}(t) \middle| X_t \right) = \frac{1}{\sqrt{n}} \sum_{i=1}^n I\!E_{X_t} (\boldsymbol{\xi}_i - \boldsymbol{\xi}_i')(1 - \mathbb{I}_i(t))$$
$$+ \frac{1}{\sqrt{e^{2t} - 1}} I\!E_{X_t} \sum_{i=1}^n \left(\frac{(\boldsymbol{\xi}_i' - \boldsymbol{\xi}_i)(\boldsymbol{\xi}_i' - \boldsymbol{\xi}_i)^T \, \mathbb{I}_i(t)}{2n} - \Sigma_i \right) \boldsymbol{\gamma} \tag{11}$$
$$+ \sum_{\alpha \ge 2} \frac{1}{2\alpha!(e^{2t} - 1)^{|\alpha|/2}} I\!E_{X_t} H_\alpha(\boldsymbol{\gamma}, I) \sum_{i=1}^n \frac{(\boldsymbol{\xi}_i' - \boldsymbol{\xi}_i)(\boldsymbol{\xi}_i' - \boldsymbol{\xi}_i)^\alpha}{n^{1/2 + |\alpha|/2}} \, \mathbb{I}_i(t).$$

Using notation from Lemma 9, we rewrite the last expression as $\mathbb{E}_{X_t} \sum_{i,\alpha\geq 0} \boldsymbol{\zeta}_i(\alpha) H_\alpha(\gamma, I)$ and set

$$\text{For} \quad |\alpha| = 0: \quad \boldsymbol{\zeta}_i(0) = \frac{1}{\sqrt{n}}(\boldsymbol{\xi}_i - \boldsymbol{\xi}_i')(1 - \mathbb{I}_i(t)).$$

$$\text{For} \quad |\alpha| = 1: \quad \boldsymbol{\zeta}_i(\alpha) = \frac{1}{\sqrt{e^{2t}-1}}\left(\frac{(\boldsymbol{\xi}_i' - \boldsymbol{\xi}_i)(\boldsymbol{\xi}_i' - \boldsymbol{\xi}_i)^\alpha\, \mathbb{I}_i(t)}{2n} - \frac{\mathbb{E}(\boldsymbol{\xi}_i' - \boldsymbol{\xi}_i)(\boldsymbol{\xi}_i' - \boldsymbol{\xi}_i)^\alpha}{2n}\right).$$

$$\text{For} \quad |\alpha| > 1: \quad \boldsymbol{\zeta}_i(\alpha) = \frac{1}{2\alpha!(e^{2t}-1)^{|\alpha|/2}} \frac{(\boldsymbol{\xi}_i' - \boldsymbol{\xi}_i)(\boldsymbol{\xi}_i' - \boldsymbol{\xi}_i)^\alpha}{n^{1/2+|\alpha|/2}}\, \mathbb{I}_i(t).$$

Next, we estimate the power L of the norm of $\sum_{i,\alpha\geq 0} \boldsymbol{\zeta}_i(\alpha) H_\alpha(\gamma, I)$. We move $\mathbb{E}_{X_t}$ outside the norm, relying on the Jensen's inequality: $\mathbb{E}\|\mathbb{E}_{X_t}[\ldots]\| \leq \mathbb{E}\|[\ldots]\|$, and define a vector $\overline{\boldsymbol{\xi}}_i = \boldsymbol{\xi}_i' - \boldsymbol{\xi}_i$. We will also need to use the following inequalities $\forall \alpha > 0$:

$$\mathbb{I}_i(t) \leq \frac{n^{|\alpha|/2}}{\|\overline{\boldsymbol{\xi}}_i\|^{|\alpha|}} \left(\frac{e^{2t}-1}{L}\right)^{|\alpha|/2}, \quad 1 - \mathbb{I}_i(t) \leq \frac{\|\overline{\boldsymbol{\xi}}_i\|^{|\alpha|}}{n^{|\alpha|/2}} \left(\frac{L}{e^{2t}-1}\right)^{|\alpha|/2}.$$

Omitting a few intermediate steps (ref. Appendix D for the entire derivation), we get the following upper bounds (*):

$$\left[\sum_{i,\alpha\geq 0} \alpha! L^{|\alpha|} \mathbb{E}\|\boldsymbol{\zeta}_i(\alpha)\|^2\right]^{1/2} \leq \frac{(2L)^{1/2}}{(e^{2t}-1)^{1/2}} \left(\frac{1}{n^2} \sum_{i=1}^n \mathbb{E}\|\boldsymbol{\xi}_i' - \boldsymbol{\xi}_i\|^2 \|\overline{\boldsymbol{\xi}}_i\|^2\right)^{1/2}$$

$$\left[\sum_i \mathbb{E}\left(\sum_{\alpha\geq 0} \alpha! L^{|\alpha|} \|\boldsymbol{\zeta}_i(\alpha)\|^2\right)^{L/2}\right]^{1/L} \leq \frac{(2L)^{1/2}}{(e^{2t}-1)^{1/2}} \left(\frac{1}{n^L} \sum_{i=1}^n \mathbb{E}\|\boldsymbol{\xi}_i' - \boldsymbol{\xi}_i\|^L \|\overline{\boldsymbol{\xi}}_i\|^L\right)^{1/L}$$

$$\sum_{\alpha\geq 0} \alpha! L^{|\alpha|} \|\mathbb{E}\boldsymbol{\zeta}(\alpha)\|^2 \leq \min\left\{\frac{e^{1/2}L}{n^{3/2}(e^{2t}-1)} \sum_{i=1}^n \mathbb{E}\|\boldsymbol{\xi}_i' - \boldsymbol{\xi}_i\| \|\overline{\boldsymbol{\xi}}_i\|^2, \frac{\|\Sigma\|^{1/2}\mu_2 e^{1/2} L^{1/2}}{(e^{2t}-1)^{1/2}}\right\}^2$$

The last term requires two bounds because the integration of function $e^{-t}/(e^{2t}-1)$ diverges near the zero. Recall that $\nabla \mathcal{G}_t(X_t) = -e^{-t} \sum_{i,\alpha\geq 0} \boldsymbol{\zeta}_i(\alpha) H_\alpha(\gamma, I)$. From Lemma 9 and expressions (*), it follows that

$$\begin{aligned} \left[\mathbb{E}\|\nabla \mathcal{G}_t(X_t)\|^L\right]^{1/L} &\leq \frac{C\, L^{3/2} e^{-t}}{\log(L)(e^{2t}-1)^{1/2}} \left(\mu_4^{1/2} + \mu_{2L}^{1/L}\right) \\ &\quad + \min\left[\frac{e^{1/2} L e^{-t}}{e^{2t}-1}\mu_3, \frac{e^{1/2} L^{1/2} e^{-t}}{(e^{2t}-1)^{1/2}}\mu_2\right]. \end{aligned}$$

Lastly, integrating over t and using Lemma 8, we obtain the final bound for $W_L(\boldsymbol{\xi}, \boldsymbol{\gamma})$. ■

Remark 11 *We reduced the proof of the previous theorem to a particular case of $\Sigma = I$. However, generally speaking, one can make the bound in this theorem slightly tighter, with the moments μ_{2k} including the product of $\|\Sigma^{-1/2}(\boldsymbol{\xi}_i - \boldsymbol{\xi}_i')\|^k$ and $\|\boldsymbol{\xi}_i - \boldsymbol{\xi}_i'\|^k$. That will not affect the asymptotic in the main theorems, however.*

To derive the statement of Theorem 1 from Theorem 10, we have to estimate the moments μ_2, μ_3, μ_4 and μ_{2L}, $L > 2$, taking into account the sub-Gaussian condition from Theorem 1. Refer to Appendix E for the details of the complete step-by-step derivation.

4. Central Limit Theorem for Max Norm in High Dimension

In this section, we will prove Theorem 5. For that, we must combine Theorem 1 (keeping in mind Markov's inequality) with the anti-concentration property (Lemma 12), which will yield the needed approximation in distribution.

Lemma 12 *Chernozhukov et al. (2017). Let $\boldsymbol{\gamma}$ be a centered Gaussian random vector in $\mathbb{R}^p$, such that $\mathbb{E}[\boldsymbol{\gamma}_k^2] \geq \underline{\sigma}^2$ for all $k \in \{1,\dots,p\}$ and some constants $\underline{\sigma}^2 > 0$, $\varDelta > 0$, then*

$$\mathbb{P}\left(\max_{1\leq k\leq p} \gamma_k \leq t + \varDelta\right) - \mathbb{P}\left(\max_{1\leq k\leq p} \gamma_k \leq t\right) \leq \frac{\varDelta}{\underline{\sigma}}(\sqrt{2\log p} + 2).$$

Below, we will reduce the Gaussian approximation task to the distribution comparison of two Gaussian vectors with different covariance matrices. The following Lemma 13 will be exploited.

Lemma 13 *Fang and Koike (2020) Let $\boldsymbol{\gamma} \in \mathcal{N}(0, \varSigma)$, $\boldsymbol{\gamma}' \in \mathcal{N}(0, \varSigma')$ be centered Gaussian random vectors in $\mathbb{R}^p$, such that $\overline{\sigma}^2 \geq \mathbb{E}[\boldsymbol{\gamma}_k^2] \geq \underline{\sigma}^2$ for all $k \in \{1,\dots,p\}$. Then,*

$$\max_t \left| \mathbb{P}\left(\max_{1\leq k\leq p} \gamma_k \leq t\right) - \mathbb{P}\left(\max_{1\leq k\leq p} \gamma'_k \leq t\right)\right| \leq C\delta \log p \, \log\left(\frac{\overline{\sigma}}{\delta\underline{\sigma}}\right),$$

where C is some absolute constant and $\delta = \|\varSigma - \varSigma'\|_\infty / \lambda_{\min}(\varSigma)$.

Yet, another Lemma 14 is instrumental as it helps prove that the convergence in distribution follows from the convergence in probability. Afterwards, we can proceed to the proof of Theorem 5.

Lemma 14 (proof in Appendix F) *For the random variables ξ and η, and a shift of size $\varDelta > 0$,*

$$\mathbb{P}(\eta < t - \varDelta) - \mathbb{P}(|\xi| \geq \varDelta) \leq \mathbb{P}(\eta + \xi < t) \leq \mathbb{P}(\eta < t + \varDelta) + \mathbb{P}(|\xi| \geq \varDelta).$$

Proof of Theorem 5. Consider a sequence of Gaussian random variables $\gamma'_1, \dots, \gamma'_p$, such that all its elements are distributed as the elements of vector $\boldsymbol{\gamma}$, but their variance matrix $\varSigma'$ is unknown. We will approximate in probability each random variable $\xi_k = \sum_i \xi_{ik}$ by γ'_k independently from the other components. Next, we will estimate $\|\varSigma - \varSigma'\|_\infty$ and, finally, by means of Lemma 13, we will compare the distributions of $\boldsymbol{\gamma}'$ and the original $\boldsymbol{\gamma}$. From Lemma 14, we get that for an arbitrary shift $\varDelta > 0$

$$\mathbb{P}\left(\max_k \gamma'_k < t - \varDelta\right) - \mathbb{P}_\varDelta \leq \mathbb{P}\left(\max_k \xi_k < t\right) \leq \mathbb{P}\left(\max_k \gamma'_k < t + \varDelta\right) + \mathbb{P}_\varDelta,$$
$$\mathbb{P}_\varDelta = \mathbb{P}\left(|\max_k \gamma'_k - \max_k \xi_k| \geq \varDelta\right).$$

Next, the anti-concentration Lemma 12 allows us to "move" $\varDelta$ out of the probability function:

$$\mathbb{P}\left(\max_k \gamma'_k < t + \varDelta\right) \leq \mathbb{P}\left(\max_k \gamma'_k < t\right) + \varDelta C_A,$$
$$\mathbb{P}\left(\max_k \gamma'_k < t - \varDelta\right) \geq \mathbb{P}\left(\max_k \gamma'_k < t\right) - \varDelta C_A,$$

where $C_A = (\sqrt{2\log p} + 2)/\underline{\sigma}$. Now, we have to minimize the following expression by $\varDelta$:

$$\left| I\!P\left(\max_k \xi_k < t\right) - I\!P\left(\max_k \gamma_k < t\right)\right| \le \min_{\pi \in \varPi[\gamma'_1, \dots, \gamma'_p, \boldsymbol{\xi}]} I\!P_\varDelta + \varDelta C_A, \tag{12}$$

where π is a common distribution of $\boldsymbol{\gamma}'$ and $\boldsymbol{\xi}$ and notation $\pi \in \varPi[\gamma'_1, \dots, \gamma'_p, \boldsymbol{\xi}]$ means that the distributions of each gamma component and $\boldsymbol{\xi}$ are fixed, without including the common distribution of $\boldsymbol{\gamma}$ (opposite to $\varPi[\boldsymbol{\gamma}', \boldsymbol{\xi}]$). From the Boole's inequality, it follows that

$$\min_{\pi \in \varPi[\gamma'_1, \dots, \gamma'_p, \boldsymbol{\xi}]} I\!P_\varDelta \le \min_{\pi \in \varPi[\gamma'_1, \dots, \gamma'_p, \boldsymbol{\xi}]} \left\{ \sum_{k=1}^p I\!P(|\gamma'_k - \xi_k| > \varDelta) \right\} = \sum_{k=1}^p \min_{\pi_k \in \varPi[\gamma'_k, \xi_k]} I\!P(|\gamma'_k - \xi_k| > \varDelta).$$

In the last step, we set the argmin $\pi = \pi(\boldsymbol{\xi})\pi_1(\gamma'_1|\xi_1) \dots \pi_p(\gamma'_p|\xi_p)$ and this allowed us to swap the sum and the min operations. From the Markov's inequality and Theorem 1, we obtain that $\forall k$

$$\min_{\pi_k \in \varPi[\gamma'_k, \xi_k]} I\!P(|\gamma'_k - \xi_k| > \varDelta) \le \frac{1}{\varDelta^L} W_L^L(\xi_k, \gamma'_k),$$

$$W_L(\xi_k, \gamma'_k) \le \frac{C \varSigma_{kk}^{1/2}}{\sqrt{n}} (L^{3/2} \nu_0^2 + L \nu_0^3 \log n). \tag{13}$$

In Eq. (13), we reduced the component $L/n^{1-1/L}$, assuming that $L \le n^{1/2 - 1/L}$. Therefore, from (12) and the previous upper bound, we obtain:

$$\left| I\!P\left(\max_k \xi_k < t\right) - I\!P\left(\max_k \gamma_k < t\right)\right| \le \varDelta C_A + \frac{p}{\varDelta^L} \left(\frac{C\overline{\sigma}(L^{3/2}\nu_0^2 + L\nu_0^3 \log n)}{\sqrt{n}} \right)^L.$$

We, then, choose $L = \log(np)$ and $\varDelta$ to satisfy the condition

$$\frac{p}{\varDelta^L} \left(\frac{C\overline{\sigma}(L^{3/2}\nu_0^2 + L\nu_0^3 \log n)}{\sqrt{n}} \right)^L \le \frac{1}{n}, \qquad \varDelta = \frac{Ce\overline{\sigma}(L^{3/2}\nu_0^2 + L\nu_0^3 \log n)}{\sqrt{n}}.$$

Setting these values gives the required bound in the Gaussian approximation. So, we have

$$\left| I\!P\left(\max_k \xi_k < t\right) - I\!P\left(\max_k \gamma'_k < t\right)\right| \le O\left(\nu_0^3 \frac{\overline{\sigma}}{\underline{\sigma}} \frac{\log^2(np)}{\sqrt{n}} \right).$$

Now, we compare $\boldsymbol{\gamma}'$ with $\boldsymbol{\gamma}$ by estimating $\|\varSigma - \varSigma'\|_\infty$. Considering arbitrary element $\varSigma'_{kl}$ and using the property $W_2(\cdot,\cdot) \le W_L(\cdot,\cdot)$ (Jensen's inequality), along with the Cauchy—Bunyakovsky inequality for the correlation of two random variables, we obtain:

$$\begin{aligned} \left| I\!E \xi_k \xi_l - I\!E \gamma'_k \gamma'_l \right| &= \left| I\!E(\xi_k \xi_l - \xi_k \gamma'_l + \xi_k \gamma'_l - \gamma'_k \gamma'_l) \right| \\ &\le \sqrt{I\!E \xi_k^2} \sqrt{I\!E(\gamma'_l - \xi_l)^2} + \sqrt{I\!E(\gamma'_l)^2} \sqrt{I\!E(\gamma'_k - \xi_k)^2} \\ &\le 2\overline{\sigma} W_L(\xi_k, \gamma'_k) \le \frac{C\overline{\sigma}^2 (L^{3/2}\nu_0^2 + L\nu_0^3 \log n)}{\sqrt{n}}. \end{aligned}$$

In the last inequality, we used the upper bound expressed by inequality (13).
Denoting $\delta_\Sigma = \max_{kl} |\mathbb{E}\xi_k\xi_l - \mathbb{E}\gamma'_k\gamma'_l| / \lambda_{\min}(\Sigma)$ and using Lemma 13, we arrive at the final approximation, $\forall t$:

$$\left| \mathbb{P}\left(\max_{1\le k\le p} \gamma_k \le t \right) - \mathbb{P}\left(\max_{1\le k\le p} \gamma'_k \le t \right) \right| \le O\left(\delta_\Sigma \log p \, \log\left(\frac{\overline{\sigma}}{\delta_\Sigma \underline{\sigma}} \right) \right)$$
$$\le O\left(\frac{\overline{\sigma}^2 \log(np)^{5/2} \nu_0^3}{\lambda_{\min}(\Sigma)\sqrt{n}} \log n \right).$$

5. Experiments

In this section, we present the results of experimental validation of the derived asymptotic of the strong Gaussian approximation. It includes two experiments: assessment with the Wasserstein distance W_L (ref. Equation (3) and Theorem 1) and approximation in distribution with the maximum norm (Theorem 5). Theorem 1 states that the convergence rate of the Gaussian approximation for W_L is $O(L^{3/2}p/\sqrt{n} + Lp^{3/2}/\sqrt{n})$, which is relatively close to the empirical asymptotic $O(p^{3/4}L^{1/2}/\sqrt{n})$. More precisely, the computed power of n is $0.498 \approx 0.5$, the power of p is $0.748 \approx 0.75$, and power of L is $0.439 \approx 0.5$. The experiment is presented in Figure 1.

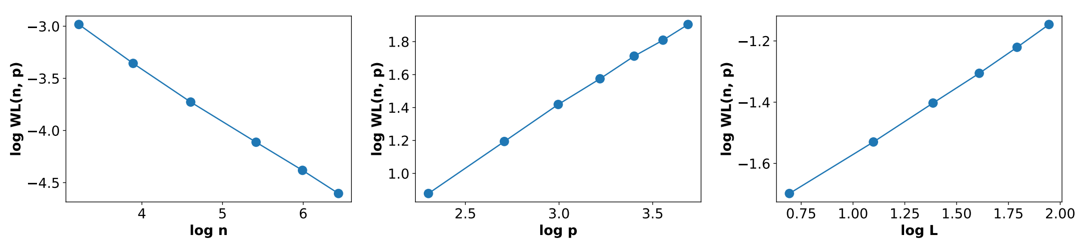

Figure 1: Asymptotic dependence of Wasserstein distance W_L on parameters n (the number of elements in the sum), p (the dimension), and L (the power of the ground distance). The empirical convergence rate is $O(p^{3/4}L^{1/2}/n^{1/2})$.

The second experiment relates to Theorem 5, which states that the convergence rate of the Gaussian approximation for the maximum norm is $O(\log^{5/2} p/\sqrt{n})$. This is also relatively close to the empirical asymptotic $O(\log p/\sqrt{n})$. More precisely, the computed power of n is $0.496 \approx 0.5$, the power of p is $0.997 \approx 1.0$. We observe that the logarithmic factor may be improved further in this theorem and that it also depends on the power of L, similarly to the previous asymptotic (Figure 2).
In both experiments, we sampled i.i.d. random variables $\xi_{ik} \sim (\mathcal{B}e(0.5) - 0.5)$ for $1 \le i \le n$ and $1 \le k \le p$ from the Bernoulli distribution. The number of samples for W_L is 10k and for the maximum norm is 100k. The simulation uses random generators from `pytorch`[1] and a GPU device (Tesla V100). To compute Wasserstein distance, we have used Sinkhorn algorithm Cuturi (2013) implemented in `geomloss`[2] with hyperparameters blur $= 0.01$ and scaling $= 0.99$.

1. https://pytorch.org/
2. https://www.kernel-operations.io/geomloss/

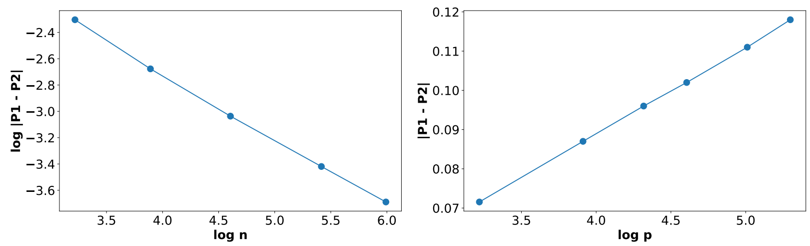

Figure 2: Dependence of the convergence rate on $\{p, n\}$ of the Gaussian approximation with the maximum norm. Left: $p = 5$, $n \in \{25, 49, 100, 225, 400\}$. Right: $n = 50$, $p \in \{25, 50, 75, 100, 150, 200\}$. The empirical convergence rate is $O(\log p/\sqrt{n})$. P_1 and P_2 are the distributions of $\max_k \xi_k$ and $\max_k \gamma_k$.

6. Conclusion

One of the most fundamental problems in the probability theory – the Central Limit Theorem – is now enhanced with a strong error bound using the apparatus of the modern optimal transport theory. The new paradigm allows us to derive a more accurate convergence rate and to study its explicit dependence on the sum size n, the parameter L, and the dimension of the vectors p. This work establishes a new fundamental limit for all practical applications of probability and statistical learning theories.

Acknowledgments

The authors thank Prof. Andrei Zaitsev for the comments that greatly improved the paper.

References

Valeriy Avanesov and Nazar Buzun. Change-point detection in high-dimensional covariance structure. *Electron. J. Statist.*, pages 3254–3294, 2018. doi: 10.1214/18-EJS1484.

Dominique Bakry, Ivan Gentil, and Michel Ledoux. *Analysis and Geometry of Markov Diffusion operators*. Grundlehren der mathematischen Wissenschaften, Vol. 348. Springer, January 2014. URL https://hal.archives-ouvertes.fr/hal-00929960.

Vidmantas Bentkus. On the dependence of the berry–esseen bound on dimension. *Journal of Statistical Planning and Inference*, 2003.

Sergey G. Bobkov. Berry–esseen bounds and edgeworth expansions in the central limit theorem for transport distances. *Probability Theory and Related Fields*, pages 229–262, February 2018. ISSN 0178-8051. doi: 10.1007/s00440-017-0756-2.

Thomas Bonis. Stein's method for normal approximation in wasserstein distances with application to the multivariate central limit theorem. *arXiv 1905.13615*, 2019.

Nazar Buzun. Gaussian approximation for empirical barycenters. *arXiv 1904.00891*, 2019.

Nazar Buzun and Valeriy Avanesov. Bootstrap for change point detection. *arXiv 1710.07285*, 2017.

L.H.Y. Chen, L. Goldstein, and Q.M. Shao. *Normal Approximation by Stein's Method*. Springer Verlag, 2010. ISBN 978-3-642-15006-7. URL http://www.springer.com/mathematics/probability/book/978-3-642-15006-7.

Victor Chernozhukov, Denis Chetverikov, and Kengo Kato. Gaussian approximations and multiplier bootstrap for maxima of sums of high-dimensional random vectors. *The Annals of Statistics*, page 2786–2819, Dec 2013. ISSN 0090-5364. doi: 10.1214/13-aos1161.

Victor Chernozhukov, Denis Chetverikov, and Kengo Kato. Central limit theorems and bootstrap in high dimensions. *The Annals of Probability*, 2014.

Victor Chernozhukov, Denis Chetverikov, and Kengo Kato. Detailed proof of nazarov's inequality. *arXiv 1711.10696*, 2017.

Victor Chernozhukov, Denis Chetverikov, Kengo Kato, and Yuta Koike. Improved central limit theorem and bootstrap approximations in high dimensions. *arXiv 1912.10529*, 2019.

Victor Chernozhukov, Denis Chetverikov, and Yuta Koike. Nearly optimal central limit theorem and bootstrap approximations in high dimensions. *arXiv 2012.09513*, 2020.

Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems*. Curran Associates, Inc., 2013. URL https://proceedings.neurips.cc/paper/2013/file/af21d0c97db2e27e13572cbf59eb343d-Paper.pdf.

Uwe Einmahl. Extensions of results of komlós, major, and tusnády to the multivariate case. *Journal of multivariate analysis*, pages 20–68, 1989.

P. Erdös and M. Kac. On certain limit theorems of the theory of probability. *Bulletin of the American Mathematical Society*, pages 292–302, 1946.

Xiao Fang and Yuta Koike. High-dimensional central limit theorems by stein's method. *arXiv 2001.10917*, 2020.

Friedrich Götze and A Yu Zaitsev. Bounds for the rate of strong approximation in the multidimensional invariance principle. *Theory of Probability & Its Applications*, pages 59–80, 2009.

Pawel Hitczenko. Best constants in martingale version of rosenthal's inequality. *Ann. Probab.*, pages 1656–1668, 10 1990. doi: 10.1214/aop/1176990639. URL https://doi.org/10.1214/aop/1176990639.

Yuta Koike et al. Gaussian approximation of maxima of wiener functionals and its application to high-frequency data. *The Annals of Statistics*, pages 1663–1687, 2019.

F. Otto and C. Villani. Generalization of an inequality by talagrand and links with the logarithmic sobolev inequality. *Journal of Functional Analysis*, pages 361 – 400, 2000. ISSN 0022-1236. doi: https://doi.org/10.1006/jfan.1999.3557. URL http://www.sciencedirect.com/science/article/pii/S0022123699935577.

A. I. Sakhanenko. Estimates in the principle of invariance in terms of truncated power moments. *Siberian Mathematical Journal*, pages 1355–1371, 2006.

Nikolay Shvetsov, Nazar Buzun, and Dmitry V. Dylov. Unsupervised non-parametric change point detection in quasi-periodic signals. *arXiv 2002.02717*, 2020.

Vladimir Spokoiny. Penalized maximum likelihood estimation and effective dimension. *Ann. Inst. H. Poincaré Probab. Statist.*, pages 389–429, 02 2017. doi: 10.1214/15-AIHP720. URL https://doi.org/10.1214/15-AIHP720.

Qiang Sun. Gaussian approximations for maxima of random vectors under (2+i)-th moments. *Statistics and Probability Letters*, page 108523, 2020. ISSN 0167-7152. doi: https://doi.org/10.1016/j.spl.2019.05.022.

A Zaitsev. Estimates of the accuracy of the strong gaussian approximation of sums of independent identically distributed random vectors. *Notes of scientific seminars POMI*, pages 141–157, 2007.

A Yu Zaitsev. Multidimensional version of a result of sakhanenko in the invariance principle for vectors with finite exponential moments. i. *Theory of Probability & Its Applications*, pages 624–641, 2001.

A. Yu. Zaitsev. The accuracy of strong gaussian approximation for sums of independent random vectors. *Russian Math. Surveys*, pages 129–172, 2013.

Appendix A. Links between major results

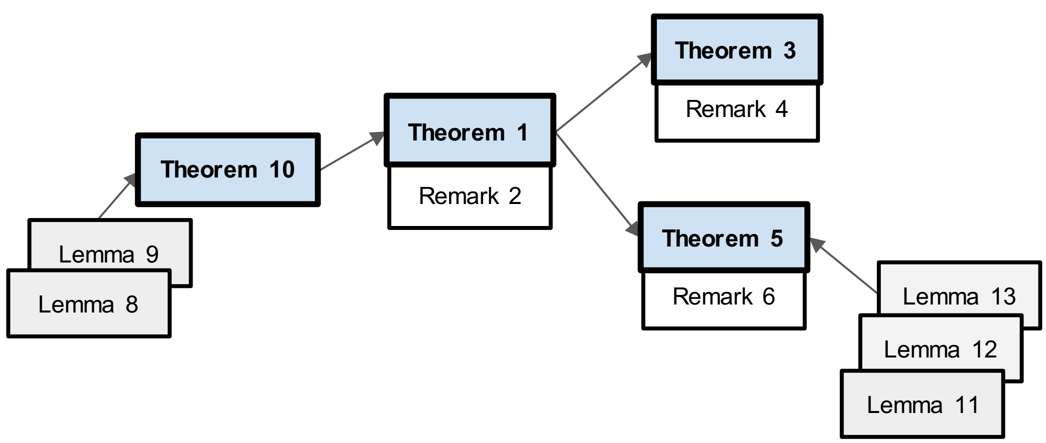

Figure 3: Links between major results of this manuscript to facilitate following the proof.

Appendix B. Proof of Lemma 8

Fix an arbitrary point t. Define a new mapping T_s on the space of the random variable X_t by equation

$$\frac{\partial}{\partial s} T_s(x) = \nabla \mathcal{G}_{t+s}(T_s(x)), \quad T_0(x) = x, \quad s \geq 0. \tag{14}$$

It turns out that function $x \to T_s(x)$ is the push-forward transport mapping, such that $X_{t+s} = T_s(X_t)$ in distribution (see proof of Lemma 2 in Otto and Villani (2000)). In order to be consistent with our notation we provide below another proof of this proposition. Note that for an arbitrary functions $h \in \mathbb{C}^1$ and $h_s = h(T_s^{-1})$, we can write

$$\frac{d}{ds} h_s(T_s) = \frac{\partial}{\partial s} h_s(T_s) + \nabla^T h_s(T_s) \frac{\partial}{\partial s} T_s = 0.$$

And using the differential Eq. (14) for T_s, one gets $\forall x$ from the domain of X_t

$$\frac{\partial}{\partial s} h_s(x) = -\nabla^T \mathcal{G}_{t+s}(x) \nabla h_s(x).$$

Then, involving property (7) leads to

$$\frac{d}{ds} \int h_s(x) d\mu_{t+s}(x) = \int \frac{\partial}{\partial s} h_s(x) + \nabla^T \mathcal{G}_{t+s}(x) \nabla h_s(x) d\mu_{t+s} = 0.$$

It proves that $T_s^{-1}(X_{t+s}) = X_t$ and subsequently $X_{t+s} = T_s T_s^{-1}(X_{t+s}) = T_s(X_t)$. By the definition of Wasserstein distance and the push-forward transport mapping T_s (14)

$$W_L^L(X_t, X_{t+ds}) \le \mathbb{E} \big\| T_{ds}(X_t) - T_0(X_t) \big\|^L = \mathbb{E} \big\| ds \nabla \mathcal{G}_t(X_t) \big\|^L$$

and consequently by triangle inequality

$$\frac{d}{dt} W_L(X_0, X_t) \le \frac{1}{ds} W_L(X_t, X_{t+ds}) \le \frac{1}{ds} \{ \mathbb{E} \| ds \nabla \mathcal{G}_t(X_t) \|^L \}^{1/L}.$$

After the integration we obtain the second statement of this lemma.

Appendix C. Proof of Lemma 9

Involve a sequence of useful lemmas that will help us to estimate term $\| \sum_{i,\alpha \ge 0} \boldsymbol{\zeta}_i(\alpha) H_\alpha(\boldsymbol{\gamma}, I) \|$.

Lemma 15 (Rosenthal's inequality Hitczenko (1990)) *For all martingales $S_k = \sum_{i=1}^k \boldsymbol{\zeta}_i$ and $2 \le q < \infty$:*

$$\left[\mathbb{E} \max_{0 \le k < \infty} \| S_k \|^q \right]^{1/q} \le \frac{Cq}{\log q} \left[\mathbb{E} \left(\sum_{i=1}^{\infty} \mathbb{E}_{i-1} \| \boldsymbol{\zeta}_i \|^2 \right)^{q/2} + \mathbb{E} \max_{0 \le i < \infty} \| \boldsymbol{\zeta}_i \|^q \right]^{1/q}.$$

By $\mathbb{E}_{i-1}$ we denote conditional expectation on previous values of the martingale $S_0, \dots, S_{i-1}$ and by definition $\boldsymbol{\zeta}_i$ are increments of the martingale which may be dependent in general case.

Lemma 16 *Bonis (2019) Let $\boldsymbol{\gamma} \in \mathcal{N}(0, I)$ and for all $\alpha \in \mathbb{N}^p$: $x_\alpha \in \mathbb{R}^p$ and $H_\alpha(y, I)$ be the multivariate Hermite polynomials, defined by*

$$H_\alpha(y, \Sigma) = (-1)^{|\alpha|} e^{\frac{y^T \Sigma^{-1} y}{2}} \partial^\alpha e^{-\frac{y^T \Sigma^{-1} y}{2}}, \quad H_\alpha(y) = H_\alpha(y, I). \tag{15}$$

Then,

$$\left[\mathbb{E} \left\| \sum_\alpha x_\alpha H_\alpha(\boldsymbol{\gamma}) \right\|^q \right]^{2/q} \le \sum_\alpha \max(1, q-1)^{|\alpha|} \alpha! \, \| x_\alpha \|^2 .$$

Proof [of Lemma 9] Note that for a random variable ξ, the operator $[\mathbb{E}\xi^q]^{1/q}$ is the probability norm L_q. Also, denote

$$\delta = \left\|\left\| \sum_{i,\alpha \ge 0} \boldsymbol{\zeta}_i(\alpha) H_\alpha(\boldsymbol{\gamma}) \right\|\right\|_{L_q}.$$

To construct a martingale, one has to substitute the expectation from each $\boldsymbol{\zeta}_i(\alpha)$. For that, one can use the triangle inequality

$$\delta \le \left\|\left\| \sum_{i,\alpha \ge 0} (\boldsymbol{\zeta}_i(\alpha) - \mathbb{E}[\boldsymbol{\zeta}_i(\alpha)]) H_\alpha(\boldsymbol{\gamma}) \right\|\right\|_{L_q} + \left\|\left\| \sum_{i,\alpha \ge 0} \mathbb{E}[\boldsymbol{\zeta}_i(\alpha)] H_\alpha(\boldsymbol{\gamma}) \right\|\right\|_{L_q}.$$

Now the sum by i consists of independent centered random vectors and we can use the martingale inequality for it from Lemma 15, setting

$$S_k = \sum_{i=1}^{k} \sum_{\alpha \ge 0} (\boldsymbol{\zeta}_i(\alpha) - \mathbb{E}[\boldsymbol{\zeta}_i(\alpha)]) H_\alpha(\boldsymbol{\gamma}).$$

Without any loss of generality, we can also assume that $\mathbb{E}[\boldsymbol{\zeta}_i(\alpha)] = 0$, because $\|\boldsymbol{\zeta}_i(\alpha) - \mathbb{E}\boldsymbol{\zeta}_i(\alpha)\|_{L_q} \le 2\|\boldsymbol{\zeta}_i(\alpha)\|_{L_q}$. Lemma 16, then, yields:

$$\delta \le \left\| \left(\sum_{\alpha \ge 0} \alpha!(q-1)^{|\alpha|} \left\| \sum_i \boldsymbol{\zeta}_i(\alpha) \right\|^2 \right)^{1/2} \right\|_{L_q}.$$

Finally, one can apply Lemma 15 to this hybrid norm and upper-bound the maximum by the sum of components with the power of q, such that $\forall x_\alpha \in \mathbb{R}^p$

$$\|x_\alpha\|_* = \left(\sum_{\alpha \ge 0} \alpha!(q-1)^{|\alpha|} \|x_\alpha\|^2 \right)^{1/2}, \quad \delta \le \mathbb{E}\left[\left\| \sum_i \boldsymbol{\zeta}_i(\alpha) \right\|_*^q \right]^{1/q}.$$

Using definition of $C(q) = Cq/\log q$ we derive that

$$\frac{\delta}{C(q)} \le \left[\sum_i \mathbb{E} \|\boldsymbol{\zeta}_i(\alpha)\|_*^2 \right]^{1/2} + \left[\mathbb{E} \max_i \left(\|\boldsymbol{\zeta}_i(\alpha)\|_*^2 \right)^{q/2} \right]^{1/q},$$

$$\frac{\delta}{C(q)} \le \left[\sum_{i,\alpha \ge 0} \alpha!(q-1)^{|\alpha|} \mathbb{E} \|\boldsymbol{\zeta}_i(\alpha)\|^2 \right]^{1/2} + \left[\mathbb{E} \max_i \left(\sum_{\alpha \ge 0} \alpha!(q-1)^{|\alpha|} \|\boldsymbol{\zeta}_i(\alpha)\|^2 \right)^{q/2} \right]^{1/q}$$

and

$$\max_i \left(\sum_{\alpha \ge 0} \alpha!(q-1)^{|\alpha|} \|\boldsymbol{\zeta}_i(\alpha)\|^2 \right)^{q/2} \le \sum_i \left(\sum_{\alpha \ge 0} \alpha!(q-1)^{|\alpha|} \|\boldsymbol{\zeta}_i(\alpha)\|^2 \right)^{q/2}.$$

■

Appendix D. Proof of Theorem 10

Remind that according to expression (11)

$$\nabla \mathcal{G}_t(X_t) = \nabla \mathcal{G}_t(X_t) - e^{-t} I\!E(\boldsymbol{\tau}(t)|X_t) = -e^{-t} I\!E\left(\boldsymbol{\xi} - \frac{1}{\sqrt{e^{2t}-1}}\boldsymbol{\gamma} + \boldsymbol{\tau}(t)\middle| X_t\right).$$

We have just added zero component to the initial value of $\nabla \mathcal{G}_t(X_t)$. Below we will use Jensen's inequality moving the conditional expectation outside the norm and $\boldsymbol{\tau}(t)$ will decrease the expected value of the norm. For the sake of brevity, let's assume in this proof that the factor $1/\sqrt{n}$ is included into random vectors $\{\boldsymbol{\xi}_i\}$ and make redefinition $\forall i$:

$$\boldsymbol{\xi}_i := \frac{1}{\sqrt{n}}\boldsymbol{\xi}_i, \quad \boldsymbol{\xi}_i' := \frac{1}{\sqrt{n}}\boldsymbol{\xi}_i'.$$

From expression

$$W_L(\boldsymbol{\xi}, \boldsymbol{\gamma}) \le \|\varSigma\|^{1/2} W_L(\varSigma^{-1/2}\boldsymbol{\xi}, \varSigma^{-1/2}\boldsymbol{\gamma})$$

follows that without loss of generality we may assume that $\varSigma = I$. Add notation $\varSigma_i = I\!E\boldsymbol{\xi}_i\boldsymbol{\xi}_i^T$ with evident property $\sum_{i=1}^n \varSigma_i = I$. Use a short notation $I\!E_{X_t}$ for the expectation operator with X_t condition. Unwrap the last expression with $\nabla \mathcal{G}_t(X_t)$ referring the definition of $\boldsymbol{\tau}(t)$ (9):

$$\begin{aligned} & I\!E\left(\boldsymbol{\xi} - \frac{1}{\sqrt{e^{2t}-1}}\boldsymbol{\gamma} + \boldsymbol{\tau}(t)\middle| X_t\right) \\ &= \sum_{i=1}^n I\!E_{X_t}(\boldsymbol{\xi}_i - \boldsymbol{\xi}_i')(1 - 1\!\!\mathrm{I}_i(t)) \\ &+ \frac{1}{\sqrt{e^{2t}-1}} I\!E_{X_t} \sum_{i=1}^n \left(\frac{(\boldsymbol{\xi}_i' - \boldsymbol{\xi}_i)(\boldsymbol{\xi}_i' - \boldsymbol{\xi}_i)^T\, 1\!\!\mathrm{I}_i(t)}{2} - \varSigma_i\right)\boldsymbol{\gamma} \\ &+ \sum_{\alpha \ge 2} \frac{1}{2\alpha!(e^{2t}-1)^{|\alpha|/2}} I\!E_{X_t} H_\alpha(\gamma, I) \sum_{i=1}^n (\boldsymbol{\xi}_i' - \boldsymbol{\xi}_i)(\boldsymbol{\xi}_i' - \boldsymbol{\xi}_i)^\alpha\, 1\!\!\mathrm{I}_i(t). \end{aligned} \tag{16}$$

Using notation from Lemma 9, we rewrite the last expression as

$$I\!E\left(\boldsymbol{\xi} - \frac{1}{\sqrt{e^{2t}-1}}\boldsymbol{\gamma} + \boldsymbol{\tau}(t)\middle| X_t\right) = I\!E_{X_t} \sum_{i,\alpha \ge 0} \boldsymbol{\zeta}_i(\alpha) H_\alpha(\gamma, I)$$

and set

$$\boldsymbol{\zeta}_i(0) = (\boldsymbol{\xi}_i - \boldsymbol{\xi}_i')(1 - 1\!\!\mathrm{I}_i(t)).$$

For $|\alpha| = 1$

$$\boldsymbol{\zeta}_i(\alpha) = \frac{1}{\sqrt{e^{2t}-1}}\left(\frac{(\boldsymbol{\xi}_i' - \boldsymbol{\xi}_i)(\boldsymbol{\xi}_i' - \boldsymbol{\xi}_i)^\alpha\, 1\!\!\mathrm{I}_i(t)}{2} - \frac{I\!E(\boldsymbol{\xi}_i' - \boldsymbol{\xi}_i)(\boldsymbol{\xi}_i' - \boldsymbol{\xi}_i)^\alpha}{2}\right).$$

For $|\alpha| > 1$

$$\boldsymbol{\zeta}_i(\alpha) = \frac{1}{2\alpha!(e^{2t}-1)^{|\alpha|/2}}(\boldsymbol{\xi}_i' - \boldsymbol{\xi}_i)(\boldsymbol{\xi}_i' - \boldsymbol{\xi}_i)^\alpha\, 1\!\!\mathrm{I}_i(t).$$

Next we want no estimate the power L of $\sum_{i,\alpha\geq 0}\boldsymbol{\zeta}_i(\alpha)H_\alpha(\gamma,I)$. Apply Jensen's inequality in order to move the conditional expectation outside the norm

$$\left\|\mathbb{E}_{X_t}\sum_{i,\alpha\geq 0}\boldsymbol{\zeta}_i(\alpha)H_\alpha(\gamma,I)\right\|^L \leq \mathbb{E}_{X_t}\left\|\sum_{i,\alpha\geq 0}\boldsymbol{\zeta}_i(\alpha)H_\alpha(\gamma,I)\right\|^L.$$

Define a vector

$$\overline{\boldsymbol{\xi}}_i = \boldsymbol{\xi}_i' - \boldsymbol{\xi}_i.$$

In the expressions below we are going to use the following inequalities $\forall\alpha > 0$

$$\mathbb{I}_i(t) \leq \frac{1}{\|\overline{\boldsymbol{\xi}}_i\|^{|\alpha|}}\left(\frac{e^{2t}-1}{L}\right)^{|\alpha|/2}, \quad 1-\mathbb{I}_i(t) \leq \|\overline{\boldsymbol{\xi}}_i\|^{|\alpha|}\left(\frac{L}{e^{2t}-1}\right)^{|\alpha|/2},$$

and equality for the norm of some $x\in\mathbb{R}^p$ and $m\geq 0$

$$\sum_{|\alpha|=m}\frac{1}{\alpha!}|x^\alpha|^2 = \frac{1}{m!}\|x\|^{2m}.$$

Note that $\forall l > 0$ and $m > 1$

$$\|\boldsymbol{\zeta}_i(0)\| \leq \|\boldsymbol{\xi}_i' - \boldsymbol{\xi}_i\|\|\overline{\boldsymbol{\xi}}_i\|^l\left(\frac{L}{e^{2t}-1}\right)^{l/2},$$

$$\sum_{|\alpha|=1}\|\boldsymbol{\zeta}_i(\alpha)\|^2 \leq \frac{\|\boldsymbol{\xi}_i'-\boldsymbol{\xi}_i\|^2\|\overline{\boldsymbol{\xi}}_i\|^2}{4(e^{2t}-1)} + \frac{(\mathbb{E}\|\boldsymbol{\xi}_i'-\boldsymbol{\xi}_i\|\|\overline{\boldsymbol{\xi}}_i\|)^2}{4(e^{2t}-1)},$$

$$\begin{aligned}\sum_{|\alpha|=m}\alpha!\|\boldsymbol{\zeta}_i(\alpha)\|^2 &= \sum_{|\alpha|=m}\frac{1}{4\alpha!(e^{2t}-1)^m}\|\boldsymbol{\xi}_i'-\boldsymbol{\xi}_i\|^2(\overline{\boldsymbol{\xi}}_i)^{2\alpha}\,\mathbb{I}_i(t)\\ &\leq \frac{1}{4m!L^l(e^{2t}-1)^{(m-l)}}\|\boldsymbol{\xi}_i'-\boldsymbol{\xi}_i\|^2\|\overline{\boldsymbol{\xi}}_i\|^{2(m-l)}.\end{aligned}$$

Then, setting $l = m-1$, one gets:

$$\sum_{\alpha\geq 0}^{\infty}\alpha!L^{|\alpha|}\|\boldsymbol{\zeta}_i(\alpha)\|^2 \leq \frac{L}{e^{2t}-1}\left(\left(1+\frac{e}{4}\right)\|\boldsymbol{\xi}_i'-\boldsymbol{\xi}_i\|^2\|\overline{\boldsymbol{\xi}}_i\|^2 + \frac{(\mathbb{E}\|\boldsymbol{\xi}_i'-\boldsymbol{\xi}_i\|\|\overline{\boldsymbol{\xi}}_i\|)^2}{4}\right).$$

Apply expectation operator and make summation by i

$$\left[\sum_{i,\alpha\geq 0}\alpha!L^{|\alpha|}\mathbb{E}\|\boldsymbol{\zeta}_i(\alpha)\|^2\right]^{1/2} \leq \frac{(2L)^{1/2}}{(e^{2t}-1)^{1/2}}\left(\sum_{i=1}^n\mathbb{E}\|\boldsymbol{\xi}_i'-\boldsymbol{\xi}_i\|^2\|\overline{\boldsymbol{\xi}}_i\|^2\right)^{1/2}. \tag{17}$$

Analogical expression holds in power $L/2$:

$$\left[\sum_i\mathbb{E}\left(\sum_{\alpha\geq 0}\alpha!L^{|\alpha|}\|\boldsymbol{\zeta}_i(\alpha)\|^2\right)^{L/2}\right]^{1/L} \leq \frac{(2L)^{1/2}}{(e^{2t}-1)^{1/2}}\left(\sum_{i=1}^n\mathbb{E}\|\boldsymbol{\xi}_i'-\boldsymbol{\xi}_i\|^L\|\overline{\boldsymbol{\xi}}_i\|^L\right)^{1/L}. \tag{18}$$

For the other part of bound in Lemma 9 one has to evaluate term $\|\sum_i \mathbb{E}\boldsymbol{\zeta}_i(\alpha)\|^2$. Note that

$$\sum_{\alpha\geq 0}\left\|\sum_i \mathbb{E}\boldsymbol{\zeta}_i(\alpha)\right\|^2 = \sum_{i,j}\sum_{\alpha\geq 0}(\mathbb{E}\boldsymbol{\zeta}_i(\alpha))^T\mathbb{E}\boldsymbol{\zeta}_j(\alpha)$$

and $\forall l > 0$:

$$\sum_{|\alpha|=1}(\mathbb{E}\boldsymbol{\zeta}_i(\alpha))^T\mathbb{E}\boldsymbol{\zeta}_j(\alpha) \leq \frac{\mathbb{E}\|\boldsymbol{\xi}_i' - \boldsymbol{\xi}_i\|\|\overline{\boldsymbol{\xi}}_i\|^{1+l}\mathbb{E}\|\boldsymbol{\xi}_j' - \boldsymbol{\xi}_j\|\|\overline{\boldsymbol{\xi}}_j\|^{1+l}}{4}\frac{L^l}{(e^{2t}-1)^{l+1}}$$

and for $m > 1$:

$$\sum_{|\alpha|=m}\alpha!(\mathbb{E}\boldsymbol{\zeta}_i(\alpha))^T\mathbb{E}\boldsymbol{\zeta}_j(\alpha) \leq \frac{\mathbb{E}\|\boldsymbol{\xi}_i' - \boldsymbol{\xi}_i\|\|\overline{\boldsymbol{\xi}}_i\|^{(m-l)}\mathbb{E}\|\boldsymbol{\xi}_j' - \boldsymbol{\xi}_j\|\|\overline{\boldsymbol{\xi}}_j\|^{(m-l)}}{4m!L^l(e^{2t}-1)^{(m-l)}}.$$

Then, setting $l = m-2$, one obtains

$$\left[\sum_{\alpha\geq 0}\alpha!L^{|\alpha|}\|\mathbb{E}\boldsymbol{\zeta}(\alpha)\|^2\right]^{1/2} \leq \frac{e^{1/2}L}{e^{2t}-1}\sum_{i=1}^n \mathbb{E}\|\boldsymbol{\xi}_i' - \boldsymbol{\xi}_i\|\|\overline{\boldsymbol{\xi}}_i\|^2. \tag{19}$$

Moreover, setting $l = m-1$ we additionally derive that

$$\left[\sum_{\alpha\geq 0}\alpha!L^{|\alpha|}\|\mathbb{E}\boldsymbol{\zeta}(\alpha)\|^2\right]^{1/2} \leq \frac{e^{1/2}L^{1/2}}{(e^{2t}-1)^{1/2}}\sum_{i=1}^n \mathbb{E}\|\boldsymbol{\xi}_i' - \boldsymbol{\xi}_i\|\|\overline{\boldsymbol{\xi}}_i\| = \frac{\mu_2 e^{1/2}L^{1/2}}{(e^{2t}-1)^{1/2}}. \tag{20}$$

The last term requires two bounds because integration of function $e^{-t}/(e^{2t}-1)$ diverges near zero point. We just have shown above in Eq.(16) that

$$\nabla\mathcal{G}_t(X_t) = -e^{-t}\sum_{i,\alpha\geq 0}\boldsymbol{\zeta}_i(\alpha)H_\alpha(\boldsymbol{\gamma}, I).$$

From Lemma 9 and expressions (17), (18), (19), (20) it follows that

$$\begin{aligned}\left[\mathbb{E}\|\nabla\mathcal{G}_t(X_t)\|^L\right]^{1/L} &\leq \frac{C\,L^{3/2}e^{-t}}{\log(L)(e^{2t}-1)^{1/2}}\left(\mu_4^{1/2} + \mu_{2L}^{1/L}\right)\\ &\quad + \min\left[\frac{e^{1/2}Le^{-t}}{e^{2t}-1}\mu_3,\ \frac{e^{1/2}L^{1/2}e^{-t}}{(e^{2t}-1)^{1/2}}\mu_2\right].\end{aligned}$$

Now in order to obtain the bound for $W_L(\boldsymbol{\xi}, \boldsymbol{\gamma})$ by means of Lemma 8 one has to integrate the previous expression over t. Apply the following integrals:

$$\int_0^\infty \frac{e^{-t}}{(e^{2t}-1)^{1/2}}dt = 1.$$

For $t_0 = 1/n$

$$
\begin{aligned}
&A\int_0^{t_0}\frac{e^{-t}}{(e^{2t}-1)^{1/2}}dt + B\int_{t_0}^{\infty}\frac{e^{-t}}{e^{2t}-1}dt \\
&= A(1-e^{-2t_0})^{1/2} + B\left(-e^{-t_0} + \frac{1}{2}\log\frac{e^{-t_0}+1}{1-e^{-t_0}}\right) \\
&\le \frac{\sqrt{2}A}{\sqrt{n}} + \frac{B}{2}\log(2n).
\end{aligned}
$$

Set $A = e^{1/2}L^{1/2}\mu_2$ and $B = e^{1/2}L\mu_3$, then finally

$$
\int_0^{\infty}\left[\mathbb{E}\|\nabla\mathcal{G}_t(X_t)\|^L\right]^{1/L}dt \le \frac{C\,L^{3/2}}{\log L}\left(\mu_4^{1/2}+\mu_{2L}^{1/L}\right) + \frac{\sqrt{2}e^{1/2}L^{1/2}\mu_2}{\sqrt{n}} + \frac{e^{1/2}L\mu_3}{2}\log(2n).
$$

Appendix E. Proof of Theorem 1

Consider a random vector $\boldsymbol{\xi}$ which has restricted exponential or sub-Gaussian moments: $\exists\, \mathtt{g} > 0$

$$
\log\mathbb{E}\exp\bigl(\boldsymbol{\gamma}^\top\boldsymbol{\xi}\bigr) \le \|\boldsymbol{\gamma}\|^2/2, \quad \boldsymbol{\gamma}\in\mathbb{R}^p, \quad \|\boldsymbol{\gamma}\|\le\mathtt{g}. \tag{21}
$$

For ease of presentation, assume below that $\mathtt{g}$ is sufficiently large.

Lemma 17 (Spokoiny (2017)) *Define $\mathtt{x}_c = \mathtt{g}^2/4$. Let (21) hold and $0.3\mathtt{g}\ge\sqrt{p}$. Then, for each $\mathtt{x}>0$*

$$
\mathbb{P}\bigl(\|\boldsymbol{\xi}\|\ge z(p,\mathtt{x})\bigr) \le 2\mathrm{e}^{-\mathtt{x}} + 8.4\mathrm{e}^{-\mathtt{x}_c}\,\mathbb{1}(\mathtt{x}<\mathtt{x}_c),
$$

where $z(p,\mathtt{x})$ is defined by

$$
z(p,\mathtt{x}) = \begin{cases}\bigl(p+2\sqrt{p\mathtt{x}}+2\mathtt{x}\bigr)^{1/2}, & \mathtt{x}\le\mathtt{x}_c\\ \mathtt{g}+2\mathtt{g}^{-1}(\mathtt{x}-\mathtt{x}_c), & \mathtt{x}>\mathtt{x}_c.\end{cases}
$$

Lemma 18 *With conditions from Lemma 17 for each $k\ge 2$*

$$
\mathbb{E}\,\|\boldsymbol{\xi}\|^k \le 4(\sqrt{p}+\sqrt{2k})^k + \left(\frac{4k}{g}\right)^k.
$$

Proof First separate events $\{\|\boldsymbol{\xi}\|\le\sqrt{p}\}$ and $\{\|\boldsymbol{\xi}\|>\sqrt{p}\}$. Then, we obtain:

$$
\mathbb{E}\|\boldsymbol{\xi}\|^k \le p^{k/2} + k\int_{\sqrt{p}}^{\infty} t^{k-1}\mathbb{P}(\|\boldsymbol{\xi}\|>t)dt.
$$

Apply Lemma 17 that is valid in the region $\{t>\sqrt{p}\}$

$$
\begin{aligned}
&\int_{\sqrt{p}}^{\infty} t^{k-1}\mathbb{P}(\|\boldsymbol{\xi}\|>t)dt \\
&\le 2\int_0^{\infty} z(p,\mathtt{x})^{k-1}(\mathrm{e}^{-\mathtt{x}})z'(p,\mathtt{x})d\mathtt{x} \\
&+ 8.4e^{-\mathtt{x}_c}\int_0^{\mathtt{x}_c} z(p,\mathtt{x})^{k-1}dz(p,\mathtt{x}).
\end{aligned}
$$

Through integration by parts we find that

$$k\int_0^\infty z(p,\mathtt{x})^{k-1}(\mathrm{e}^{-\mathtt{x}})z'(p,\mathtt{x})d\mathtt{x} = -z^k(p,0) + \int_0^\infty z(p,\mathtt{x})^k(\mathrm{e}^{-\mathtt{x}})d\mathtt{x}.$$

The function $z(p,\mathtt{x})$ behaves differently before and after the point $\mathtt{x}_c$. So we will integrate it separately.

$$\begin{aligned}
&\int_0^{\mathtt{x}_c} z(p,\mathtt{x})^k(\mathrm{e}^{-\mathtt{x}})d\mathtt{x} = \int_0^{\mathtt{x}_c}(\sqrt{p}+\sqrt{2x})^k(\mathrm{e}^{-\mathtt{x}})d\mathtt{x}\\
&\leq \int_0^\infty (\sqrt{p}+y)^k e^{-\frac{y^2}{4}}e^{-\frac{y^2}{4}}d\frac{y^2}{2}\\
&\leq \{\text{set } y = \operatorname{argmax}(\sqrt{p}+y)^k e^{-\frac{y^2}{4}}\} \leq \frac{(\sqrt{p}+\sqrt{2k})^k}{2}\int_0^\infty e^{-\frac{y^2}{4}}d\frac{y^2}{2}\\
&\leq (\sqrt{p}+\sqrt{2k})^k.
\end{aligned}$$

$$\begin{aligned}
&\int_{\mathtt{x}_c}^\infty z(p,\mathtt{x})^k(\mathrm{e}^{-\mathtt{x}})d\mathtt{x} = \int_{\mathtt{x}_c}^\infty \left(g+\frac{2}{g}(\mathtt{x}-\mathtt{x}_c)\right)^k(\mathrm{e}^{-\mathtt{x}})d\mathtt{x}\\
&= e^{-\mathtt{x}_c}\int_0^\infty \left(g+\frac{2}{g}y\right)^k e^{-\frac{y}{2}}e^{-\frac{y}{2}}dy\\
&\leq \left\{\left(g+\frac{2}{g}y\right)^k e^{-\frac{y}{2}} \leq \max\left\{g, \frac{4k}{g}\right\}^k\right\} \leq 2e^{-\mathtt{x}_c}\max\left\{g,\frac{4k}{g}\right\}^k\\
&\leq 2e^{-k/2}(k)^{k/2} + \left(\frac{4k}{g}\right)^k e^{-g^2/4} \leq k^{k/2} + \left(\frac{4k}{g}\right)^k.
\end{aligned}$$

Compute the last integral part:

$$8.4e^{-\mathtt{x}_c}k\int_0^{\mathtt{x}_c} z(p,\mathtt{x})^{k-1}dz(p,\mathtt{x}) = 8.4e^{-\mathtt{x}_c}(\sqrt{p}+\sqrt{2\mathtt{x}_c})^k \leq (\sqrt{p}+\sqrt{k})^k.$$

■

Using this lemma one may evaluate moments in Theorem 10 and derive result from Theorem 1, such that

$$I\!E\left\|\frac{1}{\sqrt{2}\nu_0}\varSigma^{-1/2}(\boldsymbol{\xi}_i-\boldsymbol{\xi}_i')\right\|^k \leq 4(\sqrt{p}+\sqrt{2k})^k + \left(\frac{4k}{g}\right)^k$$

and

$$\mu_k = \frac{1}{n^{k/2}}\sum_{i=1}^n I\!E\left\|\varSigma^{-1/2}(\boldsymbol{\xi}_i-\boldsymbol{\xi}_i')\right\|^k \leq \frac{2^{k/2+2}\nu_0^k\left\{(\sqrt{p}+\sqrt{2k})^k+\left(\frac{4k}{g}\right)^k\right\}}{n^{k/2-1}}.$$

Appendix F. Proof of Lemma 14

Consider two events $H_1 = \{|\xi| \geq \Delta\}$ and $H_2 = \{|\xi| < \Delta\}$, they compose a full group:

$$\mathbb{P}(\xi + \eta < t) = \mathbb{P}(\xi + \eta < t, H_1) + \mathbb{P}(\xi + \eta < t, H_2) = \mathbb{P}_1 + \mathbb{P}_2.$$

Lets find bounds for $\mathbb{P}_1 + \mathbb{P}_2$. For $\mathbb{P}_1$, the bounds are

$$0 \leq \mathbb{P}_1 \leq \mathbb{P}(H_1).$$

For $\mathbb{P}_2$, the upper bound will be

$$\mathbb{P}_2 = \mathbb{P}(\xi + \eta < t, -\Delta < \xi < \Delta) \leq \mathbb{P}(-\Delta + \eta < t) = \mathbb{P}(\eta < t + \Delta),$$

and the lower bound

$$\begin{aligned}&\mathbb{P}_2 = \mathbb{P}(\xi + \eta < t, -\Delta < \xi < \Delta) \geq \mathbb{P}(\Delta + \eta < t, -\Delta < \xi < \Delta) \geq \\ &\mathbb{P}(\Delta + \eta < t) - \mathbb{P}(|\xi| \geq \Delta) = \mathbb{P}(\eta < t - \Delta) - \mathbb{P}(H_1).\end{aligned}$$

We have found the bounds for $\mathbb{P}_1 + \mathbb{P}_2$ and consequently for $\mathbb{P}(\xi + \eta < t)$.