Weierstraß-Institut
für Angewandte Analysis und Stochastik

Leibniz-Institut im Forschungsverbund Berlin e. V.

Preprint ISSN 2198-5855

Generalized bootstrap in the Bures–Wasserstein space

Alexey Kroshnin, Vladimir Spokoiny, Alexandra Suvorikova

submitted: November 28, 2024

Weierstrass Institute
Mohrenstr. 39
10117 Berlin
Germany
E-Mail: alexei.kroshnin@wias-berlin.de
vladimir.spokoiny@wias-berlin.de
alexandra.suvorikova@wias-berlin.de

No. 3145
Berlin 2024

2020 *Mathematics Subject Classification.* 62F40.

Key words and phrases. Bures-Wasserstein barycenter, bootstrap, generalized bootstrap, multiplier bootstrap.

Edited by
Weierstraß-Institut für Angewandte Analysis und Stochastik (WIAS)
Leibniz-Institut im Forschungsverbund Berlin e. V.
Mohrenstraße 39
10117 Berlin
Germany

Fax: +49 30 20372-303
E-Mail: preprint@wias-berlin.de
World Wide Web: http://www.wias-berlin.de/

Generalized bootstrap in the Bures–Wasserstein space

Alexey Kroshnin, Vladimir Spokoiny, Alexandra Suvorikova

Abstract

This study focuses on finite-sample inference on the non-linear Bures–Wasserstein manifold and introduces a generalized bootstrap procedure for estimating Bures–Wasserstein barycenters. We provide non-asymptotic statistical guarantees for the resulting bootstrap confidence sets. The proposed approach incorporates classical resampling methods, including the multiplier bootstrap highlighted as a specific example.

Additionally, the paper compares bootstrap-based confidence sets with asymptotic confidence sets obtained in the work of Kroshnin et al. [2021], evaluating their statistical performance and computational complexities. The methodology is validated through experiments on synthetic datasets and real-world applications.

1 Introduction

Optimal transport (OT) seeks the most efficient way to transform one distribution into another, given a transportation cost. This allows one to define geometrically meaningful distances between probability measures [Ambrosio et al., 2008, Villani, 2009, Santambrogio, 2015].

OT provides a powerful framework for modeling and analyzing objects and processes, with applications spanning diverse fields. These include machine learning [Arjovsky et al., 2017], information geometry [Khan and Zhang, 2022], image processing and computer vision [Bonneel and Digne, 2023], economics [Galichon, 2018], and bioinformatics [Schiebinger et al., 2019]. OT-based distances also play an important role in statistical inference [Del Barrio et al., 2015, Rippl et al., 2016, del Barrio et al., 2017, Bobkov and Ledoux, 2019, Panaretos and Zemel, 2020, Heinemann et al., 2022, Chewi et al., 2024].

Beyond their intrinsic interest, OT distances facilitate the definition of a new type of averaging, known as the Wasserstein barycenter, distinct from the classical notion of the mean. Barycenters have a broad spectrum of applications, such as image processing [Simon and Aberdam, 2020], time series modeling [Cheng et al., 2021], modern energy technologies [Larvaron et al., 2024], economics [Levantesi et al., 2024], machine learning [Mallasto and Feragen, 2017, Muzellec and Cuturi, 2018], among others.

Bhatia et al. [2019] established a connection between OT and quantum information theory and introduced the Bures–Wasserstein distance and corresponding barycenter that are the focus of the current study. Recent works [Haasler and Frossard, 2024, Maretic et al., 2022a,b] have demonstrated the applicability of this concept to graph alignment and averaging. In particular, Haasler and Frossard [2024] present a novel approach to analyzing graph-structured data and then show the usability of the Bures–Wasserstein barycenter of graphs.

This study considers the statistical framework for barycenters, assuming the observed data is random. Within this setting, numerous studies have addressed the consistency of barycenters and their variations [Bigot et al., 2012, Le Gouic and Loubes, 2017, Cazelles et al., 2017]. Additionally, explicit convergence rates, concentration inequalities, and large deviation results have been of significant interest [Ahidar-Coutrix et al., 2020, Brunel and Serres, 2024, Le Gouic et al., 2022, Jaffe and Santoro, 2024]. In some cases, the Central Limit Theorem has been established [Kroshnin et al., 2021, Carlier et al., 2021]. It is noteworthy that some classical results apply to the barycenter setting because barycenters are M-estimators [Van De Geer, 2006].

Some of the results mentioned above can be used to build asymptotic confidence sets for barycenters. A fundamentally different mechanism for constructing confidence sets is based on the bootstrap approach. Since their introduction in the seminal paper by Efron [1979], bootstrapping techniques have attracted much attention due to their algorithmic simplicity and computational tractability. Spokoiny and Zhilova [2015] apply multiplier bootstrap to construct likelihood-based confidence sets. Chen and Zhou [2020] investigate the case of heavy-tailed data. Naumov et al. [2019] validate bootstrap approximation for spectral projectors in the case of Gaussian data. Cheng and Huang [2010] provides approximation rates for multiplier bootstrap for M-estimators in semi-parametric models. Lee and Yang [2020] propose a resampling procedure for M-estimators for non-standard cases. For more examples, we recommend an excellent survey by Mammen and Nandi [2012].

The current study develops a generalized bootstrap procedure [Van Der Vaart et al., 1996] tailored for Bures–Wasserstein barycenters and provides non-asymptotic statistical guarantees for the resulting bootstrap confidence sets.

1.1 Brief introduction to optimal transport

We begin with a particularly important case, the 2-Wasserstein distance. It stands out due to its rich geometric structure. Let $\mathbb{R}^d$ be equipped with L^2-norm. Then the distance between distribution μ_1 and μ_2 on $\mathbb{R}^d$ with finite second moments is defined as

$$\mathcal{W}_2^2(\mu_1, \mu_2) = \inf_{\pi \in \varPi(\mu_1,\mu_2)} \int\limits_{\mathbb{R}^d \times \mathbb{R}^d} \|x-y\|_2^2 \, d\pi(x,y)$$

$$\varPi(\mu_1, \mu_2) = \left\{ \pi \in \mathcal{P}(\mathbb{R}^d \times \mathbb{R}^d) \,\middle|\, \int \pi(x,y)dy = \mu_1(x), \ \int \pi(x,y)dx = \mu_2(y) \right\},$$

with $\mathcal{P}(\mathbb{R}^d \times \mathbb{R}^d)$ being the set of all probability measures on $\mathbb{R}^d \times \mathbb{R}^d$. The 2-Wasserstein distance is of high practical interest; see, e.g., [Courty et al., 2016, Bistroń et al., 2022].

The case of the 2-Wasserstein distance for Gaussian distributions [Takatsu et al., 2011] is particularly interesting due to its analytical tractability. The 2-Wasserstein distance between Gaussian distributions, $\mu_1 = \mathcal{N}(m_1, \varSigma_1)$ and $\mu_2 = \mathcal{N}(m_2, \varSigma_2)$, is

$$\mathcal{W}_2^2(\mu_1, \mu_2) = \|m_1 - m_2\|_2^2 + \operatorname{tr} \varSigma_1 + \operatorname{tr} \varSigma_2 - 2 \operatorname{tr} \left(\varSigma_1^{1/2} \varSigma_2 \varSigma_1^{1/2} \right)^{1/2}. \tag{1.1}$$

It is important to note that a similar concept—the Bures distance [Bures, 1969]—arises in quantum information geometry. For any pair of positive-definite Hermitian operators $\varSigma_1$ and $\varSigma_2$, s.t. $\operatorname{tr} \varSigma_1 = \operatorname{tr} \varSigma_2 = 1$ (i.e., $\varSigma_1$ and $\varSigma_2$ are density operators), the Bures distance is defined as

$$\mathrm{B} = 2 \left(1 - \left(\varSigma_1^{1/2} \varSigma_2 \varSigma_1^{1/2} \right)^{1/2} \right).$$

Bhatia et al. [2019] combined the concept of the Bures distance with the 2-Wasserstein distance for Gaussian distributions (as defined in (1.1)) and introduced the Bures–Wasserstein distance. Let $\mathbb{H}(d)$ be the space of all $d \times d$ Hermitian matrices, with $\mathbb{H}_+(d)$ and $\mathbb{H}_{++}(d)$ representing its subspaces of positive semidefinite and positive definite matrices, respectively. The Bures–Wasserstein distance on $\mathbb{H}_{++}(d)$ is defined as

$$\mathcal{W}^2(Q,S) \overset{\text{def}}{=} \operatorname{tr} Q + \operatorname{tr} S - 2\operatorname{tr}\left(S^{1/2}QS^{1/2}\right)^{1/2}, \quad S,Q \in \mathbb{H}_+(d).$$

The 2-Wasserstein barycenter—introduced by Agueh and Carlier [2011]—is a Fréchet mean in the Wasserstein space. It aggregates probability measures in a geometrically meaningful way and reduces data variability. Given a set of probability distributions $\mu_1, \mu_2, \dots, \mu_n$ defined on $\mathbb{R}^d$, and a set of non-negative weights $w_1, w_2, \dots, w_n$ with $\sum_{i=1}^n w_i = 1$, the 2-Wasserstein barycenter $\overline{\mu}$ is defined as the probability measure that minimizes the weighted sum of squared 2-Wasserstein distances:

$$\overline{\mu} \overset{\text{def}}{=} \underset{\nu \in \mathcal{P}_2(\mathbb{R}^d)}{\operatorname{argmin}} \sum_{i=1}^n w_i \mathcal{W}_2^2(\nu, \mu_i),$$

where $\mathcal{P}_2(\mathbb{R}^d)$ is the set of probability measures on $\mathbb{R}^d$ with finite second moments.

In particular, this provides a geometrically meaningful method of averaging Gaussian distributions. Specifically, consider a set of d-dimensional Gaussian measures $\mu_1 = \mathcal{N}(m_1, \varSigma_1), \dots, \mu_n = \mathcal{N}(m_n, \varSigma_n)$. Agueh and Carlier [2011] showed that their 2-Wasserstein barycenter $\overline{\mu} = \mathcal{N}(\overline{m}, \overline{\varSigma})$ is Gaussian distribution as well with

$$\overline{m} = \frac{1}{n}\sum_{i=1}^n m_i, \quad \overline{\varSigma} = \sum_{i=1}^n w_i \left(\overline{\varSigma}^{1/2} \varSigma_i \overline{\varSigma}^{1/2}\right)^{1/2}. \tag{1.2}$$

Note that $\overline{\varSigma}$ is the unique solution of the fixed-point equation (1.2) [Agueh and Carlier, 2011]. The works by Álvarez-Esteban et al. [2016] and Chewi et al. [2020] discuss the computational aspects.

Bhatia et al. [2019] showed that the result similar to (1.2) holds for non-negative Hermitian operators. Namely, for fixed weights $w_1, \dots, w_n$, s.t. $\sum_i w_i = 1$, one can define the Bures–Wasserstein barycenter of positive semi-definite Hermitian operators $S_1, \dots, S_n \in \mathbb{H}_+(d)$ as

$$B = \underset{Q \in \mathbb{H}_{++}(d)}{\operatorname{argmin}} \sum_{i=1}^n w_i \mathcal{W}^2(S_i, Q), \quad B = \sum_{i=1}^n w_i \left(B^{1/2} S_i B^{1/2}\right)^{1/2}. \tag{1.3}$$

1.2 Generalized bootstrap

Let $\mathcal{M}\big(\mathbb{H}_{++}(d)\big)$ be the space of non-zero finite Borel measures on $\mathbb{H}_{++}(d)$ endowed with the Borel σ-algebra induced by the topology of weak convergence.

We define the barycenter mapping from $\mathcal{M}\big(\mathbb{H}_{++}(d)\big)$ to $\mathbb{H}_{++}(d)$ as

$$\mathcal{B}\colon \mu \mapsto B_\mu \overset{\text{def}}{=} \underset{Q \in \mathbb{H}_{++}(d)}{\operatorname{argmin}} \int\limits_{\mathbb{H}_{++}(d)} \mathcal{W}^2(Q,S)d\mu(S). \tag{1.4}$$

Recall that $\mathcal{B}$ is uniquely defined (see Theorem 2.1 in [Kroshnin et al., 2021]). Thus, according to Corollary 5 in [Le Gouic and Loubes, 2017] it is continuous w.r.t. the 2-Wasserstein metric on the subspace of probability measures $\mathcal{P}\big(\mathbb{H}_{++}(d)\big) \subset \mathcal{M}\big(\mathbb{H}_{++}(d)\big)$. Hence, by homogeneity $\mathcal{B}(\mu) = \mathcal{B}\left(\frac{\mu}{\mu(\mathbb{H}_{++}(d))}\right)$. Thus, $\mathcal{B}$ is measurable on $\mathcal{M}\big(\mathbb{H}_{++}(d)\big)$.

Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and $\mu \in \mathcal{M}\big(\mathbb{H}_{++}(d)\big)$ be a random measure with distribution $\mathcal{P}$. Respectively, B_μ is a random matrix. Let $B = \mathcal{B}(\mathbb{E}\,\mu)$. The goal is to approximate the law of $\mathcal{W}(B_\mu, B)$ to construct a confidence set around the observed B_μ. We use a generalized bootstrap approach: we will show that $\mathcal{W}(B, B_\mu) \overset{\mathrm{d}}{\approx} \mathcal{W}(B_\mu, B_{\hat{\mu}})$ for a properly selected random $\hat{\mu} \in \mathcal{M}\big(\mathbb{H}_{++}(d)\big)$ depending on μ.

The main result (Theorem 4.1) claims that under suitable assumptions on $\mathcal{P}$ and $\hat{\mu}$ with probability at least $1 - \mathtt{C}e^{-t}$ (with $\mathtt{C}$ being a generic constant), it holds

$$\sup_{z \ge 0} \left| \mathbb{P}\left\{\mathcal{W}(B_\mu, B) \le z\right\} - \mathbb{P}\left\{\mathcal{W}(B_{\hat{\mu}}, B_\mu) \le z \mid \mu\right\} \right| \le \Gamma(\mathrm{t}), \tag{1.5}$$

where $\mathbb{P}\{\cdot|\mu\}$ stands for probability conditioned on μ. This setting covers the classical resampling techniques, including the multiplier bootstrap.

1.3 Multiplier bootstrap

We consider an i.i.d. sample $S_1, \ldots, S_n$, where $S_i \overset{\mathrm{iid}}{\sim} P$, with P supported on $\mathbb{H}_{++}(d)$. The empirical distribution μ is defined as:

$$\mu := P_n = \frac{1}{n} \sum_i \delta_{S_i},$$

with δ_S being the Dirac measure. Correspondingly, $\mathcal{P}$ defined in 1.2 represents the law of empirical distributions constructed from i.i.d. samples of size n drawn from P.

We define $\hat{\mu}$ as a reweighted empirical distribution

$$\hat{\mu} := P_w = \frac{1}{n} \sum_i w_i \delta_{S_i},$$

where $w_1, \ldots, w_n$ are non-negative i.i.d. weights independent of the data. Specifically, we assume

$$w_i \overset{\mathrm{iid}}{\sim} W, \quad \mathbb{E}_w\, w_i = 1, \quad \mathrm{Var}_w\, w_i = 1.$$

It is noteworthy that the total mass $\hat{\mu}(\mathbb{H}_{++}(d))$ might be not equal 1. This flexibility is inherent to the generalized bootstrap framework

Finally, $B = \mathcal{B}(P)$. Collecting the results above, we have

$$B = \mathcal{B}(P), \quad B_n = \mathcal{B}(P_n), \quad B_w = \mathcal{B}(P_w), \tag{1.6}$$

where B, B_n, and B_w represent the population barycenter, the empirical barycenter, and the reweighted empirical barycenter, respectively.

The generalized bootstrap result ensures that the law of $\sqrt{n}\mathcal{W}(B_n, B_w)$ approximates the law of $\sqrt{n}\mathcal{W}(B, B_n)$. Consequently, one can utilize $\sqrt{n}\mathcal{W}(B_n, B_w)$ to construct non-asymptotic confidence sets for B_n. We will show in Theorem 5.1 that if P and W are sub-exponential, the approximation rate $\Gamma(\mathrm{t})$ (see (1.5)) is of order $n^{-1/2}$ up to a logarithmic factor.

1.4 Contribution of this paper

This study addresses the challenge of finite-sample inference on the non-linear Bures–Wasserstein manifold. The primary contribution is developing a generalized bootstrap procedure [Van Der Vaart et al., 1996] for the Bures–Wasserstein barycenters and providing non-asymptotic statistical guarantees for bootstrap confidence sets.

One of the central results of this study is the derivation of relative approximation bounds for the Bures–Wasserstein distance and the Bures–Wasserstein barycenters. These bounds, presented in Section 2, form the foundation for analyzing the generalized bootstrap procedure.

To validate the proposed bootstrap framework, the study establishes Gaussian approximation results for $\mathcal{W}(B_\mu, B)$ and $\mathcal{W}(B_{\hat{\mu}}, B_\mu)$ (see Section 3). Specifically, under mild assumptions and for properly chosen centered Gaussian vectors Z, and Z_μ, it holds

$$\mathcal{W}(B_\mu, B) \overset{\mathrm{d}}{\approx} \|\boldsymbol{A}Z\|_{\mathrm{F}}, \quad \mathcal{W}(B_{\hat{\mu}}, B_\mu) \overset{\mathrm{d}}{\approx} \|\boldsymbol{A}Z_\mu\|_{\mathrm{F}},$$

where $\boldsymbol{A}$ is a scaling operator encapsulating the geometric structure of the Bures–Wasserstein space (see (2.1)). These approximations are formulated as non-asymptotic bounds on the Kolmogorov distance between the corresponding distributions.

To illustrate the framework's versatility, the study considers the multiplier bootstrap as a specific case. In this context, we demonstrate how one can use model assumptions about the distribution of observed data to establish the necessary conditions for the generalized bootstrap's validity.

Furthermore, we show the procedure's applicability using graph-structured data, including a weighted stochastic block model and human brain connectomes (comprehensive maps of neural connections in the brain). Finally, the study compares the computational complexity of the proposed procedure with that of constructing asymptotic confidence sets as described in [Kroshnin et al., 2021]. The results demonstrate that the bootstrap procedure exhibits greater numerical efficiency than the asymptotic approach, making it a more practical choice for applications requiring computationally scalable inference methods, especially in high dimensions.

Organization of the paper and accepted notations

The paper is organized as follows. Section 2 presents approximation bounds in the Bures–Wasserstein space. In Section 3, we derive Gaussian approximation results, which are crucial for proving the generalized bootstrap. Section 4 presents the main theoretical result concerning non-asymptotic statistical guarantees for bootstrap confidence sets. Section 5 focuses on the case of multiplier bootstrap. In Section 6, we evaluate the performance of the proposed method on both synthetic and real datasets. In particular, we compare approximations constructed using the multiplier bootstrap with those derived from asymptotic results of Kroshnin et al. [2021]. Additionally, we analyze the computational complexities of both methods.

Table 1 lists the notations used throughout the text.

2 Approximation bounds in the Bures–Wasserstein space

We begin collecting some facts that are crucial for generalized bootstrap validation. We will often quantify the closeness of matrices or operators $S \succcurlyeq 0$ ($\boldsymbol{S} \succcurlyeq 0$) and $Q \succ 0$ ($\boldsymbol{Q} \succ 0$) as

$$r(Q, S) \overset{\mathrm{def}}{=} \left\|Q^{-1/2} S Q^{-1/2} - I\right\|, \quad r(\boldsymbol{Q}, \boldsymbol{S}) \overset{\mathrm{def}}{=} \left\|\boldsymbol{Q}^{-1/2} \boldsymbol{S} \boldsymbol{Q}^{-1/2} - \boldsymbol{I}\right\|.$$

with $\|\cdot\|$ being the operator norm, I standing for the $d \times d$ identity matrix and $\boldsymbol{I}$ being the identity operator.

$\mathbb{H}(d)$	$d \times d$ Hermitian operators
$\mathbb{H}_+(d), \mathbb{H}_{++}(d)$	$d \times d$ positive semi-definite and positive definite Hermitian operators
X	Matrices or vectors
$\boldsymbol{X}$	Operators
$\lambda_{\max}(X)$, $\lambda_{\min}(X)$	Largest and smallest eigenvalues
$\|X\|, \|X\|_{\mathrm{F}}, \|X\|_1, \|X\|_{\psi_\alpha}$	Operator, Frobenius, 1-Schatten, ψ_α-Orlicz norm
$\langle X, Y \rangle$	Inner product associated to Frobenius norm
$\kappa(X) = \|X\| \cdot \|X^{-1}\|$	Condition number of an operator or a matrix
$\otimes$	Tensor product
$\log(x)$	$\log(x) = \max\{1, \ln(x)\}$
$r(X, A)$	$r(X,A) = \|X^{-1/2} A X^{-1/2} - I\|$
$\mathtt{C}$	Generic constant
$\overset{\mathrm{d}}{\approx}$	Closeness in distribution

Table 1: List of accepted notations.

Now, we recall the concept of the optimal transportation (OT) map, one of the key concepts in OT. Of note, it is often referred to as the optimal push-forward.

For any $Q, S \in \mathbb{H}_{++}(d)$ we denote the OT map as $T_Q^S = Q^{-1/2}(Q^{1/2} S Q^{1/2})^{1/2} Q^{-1/2}$. It is differentiable in Fréchet sense (see Lemma A.2 by Kroshnin et al. [2021]),

$$T_{Q+X}^S = T_Q^S + \boldsymbol{dT}_Q^S(X) + o(\|X\|) \quad \text{as } \|X\| \to 0, \quad X \in \mathbb{H}(d),$$

where $\boldsymbol{dT}_Q^Q \colon \mathbb{H}(d) \to \mathbb{H}(d)$ is a negative semi-definite operator.

The first result in this section establishes a connection between $\mathcal{W}(Q,S)$ and the Frobenius norm of the difference $\|Q - S\|_{\mathrm{F}}$. From now on, we fix some $B \in \mathbb{H}_{++}(d)$ and introduce an auxiliary operator $\boldsymbol{A}$,

$$\boldsymbol{A} \overset{\mathrm{def}}{=} \left(-\frac{1}{2} \boldsymbol{dT}_B^B \right)^{1/2}. \tag{2.1}$$

The properties of $\boldsymbol{A}$ are investigated in Lemma A.2.

Lemma 2.1. *Let $Q, S \in \mathbb{H}_+(d)$ be s.t. $r(B,Q) \le 1/2$ and $r(B,S) \le 1/2$. Then*

$$\left| \frac{\mathcal{W}(Q,S)}{\|\boldsymbol{A}(Q-S)\|_{\mathrm{F}}} - 1 \right| \le 4r(B,Q) + 2r(B,S).$$

The proof is technical, so we postponed it to Appendix A.

Now, for a measure $\mu \in \mathcal{M}\big(\mathbb{H}_{++}(d)\big)$, by analogy with the barycenter mapping $\mathcal{B}(\mu)$ (see (1.4)), we introduce the $\mathcal{T}(\mu)$ mapping and $\mathcal{F}(\mu)$ mapping,

$$\mathcal{T} : \mu \mapsto T_\mu = \int\limits_{\mathbb{H}_{++}(d)} \left(T_B^S - I \right) d\mu(S); \tag{2.2}$$

$$\mathcal{F} : \mu \mapsto \boldsymbol{F}_\mu = - \int\limits_{\mathbb{H}_{++}(d)} \boldsymbol{dT}_B^S d\mu(S). \tag{2.3}$$

The following lemma connects B_μ and T_μ. Let $\boldsymbol{F}$ be some fixed positive-definite operator acting from $\mathbb{H}(d)$ to $\mathbb{H}(d)$. Denote

$$r \overset{\text{def}}{=} r(B, B_\mu) + r(\boldsymbol{F}, \boldsymbol{F}_\mu), \quad \rho \overset{\text{def}}{=} 2\sqrt{\kappa(\boldsymbol{F})}r, \tag{2.4}$$

with $\kappa(X) = \|X\| \cdot \|X^{-1}\|$ being the condition number of X.

Lemma 2.2. *Let $r \le \frac{1}{2}$, then the following approximations hold:*

$$\frac{\|B_\mu - B - \boldsymbol{F}^{-1}T_\mu\|_{\mathrm{F}}}{\|\boldsymbol{F}^{-1}T_\mu\|_{\mathrm{F}}} \le \rho, \tag{2.5}$$

$$\left|\frac{\mathcal{W}(B_\mu, B)}{\|\boldsymbol{A}\boldsymbol{F}^{-1}T_\mu\|_{\mathrm{F}}} - 1\right| \le 3\sqrt{\kappa(B)}\rho, \tag{2.6}$$

with $\boldsymbol{A}$ coming from (2.1)*.*

Proof. First, we introduce an auxiliary operator $\boldsymbol{D}_\mu$. Let $B_t = tB_\mu + (1-t)B$, $t \in [0,1]$. We set

$$\boldsymbol{D}_\mu \overset{\text{def}}{=} - \int\limits_{\mathbb{H}_{++}(d)} \left[\int\limits_0^1 \boldsymbol{d}\boldsymbol{T}^S_{B_t} dt\right] d\mu(S). \tag{2.7}$$

Proof of (2.5) We write the Taylor expansion for B_μ in the neighbourhood of B in integral form (see Theorem 2.2 by Kroshnin et al. [2021]), $B_\mu - B = \boldsymbol{D}_\mu^{-1}T_\mu$. This ensures

$$B_\mu - B - \boldsymbol{F}^{-1}T_\mu = \left(\boldsymbol{D}_\mu^{-1}\boldsymbol{F} - \boldsymbol{I}\right)\boldsymbol{F}^{-1}T_\mu,$$

with $\boldsymbol{I}$ being the identity operator. We set $B_\Delta := B_\mu - B$ and get

$$\frac{\|B_\Delta - \boldsymbol{F}^{-1}T_\mu\|_{\mathrm{F}}}{\|\boldsymbol{F}^{-1}T_\mu\|_{\mathrm{F}}} \le \left\|\boldsymbol{D}_\mu^{-1}\boldsymbol{F} - \boldsymbol{I}\right\|.$$

The bounds on $\boldsymbol{D}_\mu$ from Lemma A.4 yield

$$(1-r)\,\boldsymbol{F}^{-1} \preccurlyeq \boldsymbol{D}_\mu^{-1} \preccurlyeq (1+2r)\,\boldsymbol{F}^{-1}.$$

Therefore,

$$\left\|\boldsymbol{D}_\mu^{-1}\boldsymbol{F} - \boldsymbol{I}\right\| \le \sqrt{\kappa(\boldsymbol{F})}r(\boldsymbol{D}_\mu^{-1}, \boldsymbol{F}^{-1}) \le \rho.$$

The claim follows. The proof of (2.6) is similar. We postpone it to Appendix A. □

Now we fix some $\hat{\mu} \in \mathcal{M}\big(\mathbb{H}_{++}(d)\big)$ and define

$$\hat{r} \overset{\text{def}}{=} r(B, B_{\hat{\mu}}) + r(\boldsymbol{F}, \boldsymbol{F}_{\hat{\mu}}), \quad \hat{\rho} \overset{\text{def}}{=} 2\sqrt{\kappa(\boldsymbol{F})}\,\hat{r}. \tag{2.8}$$

Corollary 2.3. *If $r \le \frac{1}{2}$ and $\hat{r} \le \frac{1}{2}$, then the following bounds hold*

$$\big\|B_{\hat{\mu}} - B_\mu - \boldsymbol{F}^{-1}(T_{\hat{\mu}} - T_\mu)\big\|_{\mathrm{F}} \le \hat{\rho}\big\|\boldsymbol{F}^{-1}(T_{\hat{\mu}} - T_\mu)\big\|_{\mathrm{F}} + (\rho + \hat{\rho})\big\|\boldsymbol{F}^{-1}T_\mu\big\|_{\mathrm{F}}, \tag{2.9}$$

$$\begin{aligned} &\big|\mathcal{W}(B_{\hat{\mu}}, B_\mu) - \|\boldsymbol{A}\boldsymbol{F}^{-1}(T_{\hat{\mu}} - T_\mu)\|_{\mathrm{F}}\big| \\ &\le 6\kappa(\boldsymbol{A})\,(\hat{\rho} + \rho)\,\|\boldsymbol{A}\boldsymbol{F}^{-1}\,(T_{\hat{\mu}} - T_\mu)\|_{\mathrm{F}} + 4\,(\hat{\rho} + \rho)\,\|\boldsymbol{A}\|\|\boldsymbol{F}^{-1}T_\mu\|_{\mathrm{F}}. \end{aligned} \tag{2.10}$$

The proof is postponed to Appendix A. The next result estimates the proximity of B_μ and B in terms of $\|T_\mu\|_{\mathrm{F}}$.

Lemma 2.4. *For $X \in \mathbb{H}(d)$ we denote*

$$\boldsymbol{\xi}(X) \overset{\mathrm{def}}{=} B^{1/2}\boldsymbol{F}\left(B^{1/2}XB^{1/2}\right)B^{1/2},$$

and set

$$c_B \overset{\mathrm{def}}{=} \frac{4\|B\|}{\lambda_{\min}(\boldsymbol{\xi})}. \tag{2.11}$$

Assume that $r(\boldsymbol{F}, \boldsymbol{F}_\mu) \le \frac{1}{2}$ and $\|T_\mu\|_{\mathrm{F}} \le \frac{4}{3c_B}$. Then

$$\left\|B^{-1/2}B_\mu B^{-1/2} - I\right\|_{\mathrm{F}} \le c_B\|T_\mu\|_{\mathrm{F}}.$$

Proof. First, we set

$$\boldsymbol{\xi}_\mu(X) = B^{1/2}\boldsymbol{F}_\mu\left(B^{1/2}XB^{1/2}\right)B^{1/2}, \quad \zeta_\mu = \frac{1}{\lambda_{\min}(\boldsymbol{\xi}_\mu)}\left\|B^{1/2}T_\mu B^{1/2}\right\|_{\mathrm{F}}.$$

Provided that $\zeta_\mu \le \frac{2}{3}$, Lemma B.1 by Kroshnin et al. [2021] ensures

$$\left\|B^{-1/2}B_\mu B^{-1/2} - I\right\|_{\mathrm{F}} \le \frac{\zeta_\mu}{1 - \frac{3}{4}\zeta_\mu} \le 2\zeta_\mu.$$

Now we show that condition $\zeta_\mu \le \frac{2}{3}$ holds. Assumption $r(\boldsymbol{F}, \boldsymbol{F}_\mu) \le \frac{1}{2}$ implies $r(\boldsymbol{\xi}, \boldsymbol{\xi}_\mu) \le \frac{1}{2}$. This yields $\lambda_{\min}(\boldsymbol{\xi}_\mu) \ge \frac{\lambda_{\min}(\boldsymbol{\xi})}{2}$. Therefore, the assumptions of the lemma ensure

$$\zeta_\mu \le \frac{c_B}{2}\|T_\mu\|_{\mathrm{F}} \le \frac{2}{3}.$$

This finishes the proof. □

3 Gaussian approximation

This section presents the general Gaussian approximation result. It is the key ingredient for bootstrap validity. The first lemma contains an auxiliary term $\gamma(\cdot)$. To avoid breaking the logic of the presentation, we will define $\gamma(\cdot)$ immediately after the lemma. Moreover, from now on, we will denote generic absolute constants as $\mathtt{C}$.

Lemma 3.1 (GAR). *Let $X, Y \in \mathbb{R}_+$ be random variables satisfying the following assumptions:*

There exist constants $m, \delta > 0$, $\rho \in \left[0, \frac{1}{2}\right]$ s.t.

$$\mathbb{P}\left(|X - Y| \le \rho Y + m\right) \ge 1 - \delta. \tag{GAR-I}$$

There exists a centred Gaussian vector $G \sim \mathcal{N}(0, \boldsymbol{K})$ taking values in a Hilbert space H, and a constant $\varDelta \in (0, 1)$, s.t.

$$\sup_{z>0}\left|\mathbb{P}\left\{Y \le z\right\} - \mathbb{P}\left\{\|G\|_H \le z\right\}\right| \le \varDelta, \tag{GAR-II}$$

with $\|\cdot\|_H$ *denoting the norm induced by the scalar product in* H*. Then*

$$\sup_{z>0}|\mathbb{P}\left\{X \le z\right\} - \mathbb{P}\left\{\|G\|_H \le z\right\}| \le \Delta + \delta + \mathtt{C}\gamma(\boldsymbol{K})\left(\frac{m}{\sqrt{\operatorname{tr}(\boldsymbol{K})}} + \rho\right),$$

with $\gamma(\boldsymbol{K})$ *coming from* (3.2).

Now we define $\gamma(\cdot)$. Let $\boldsymbol{K}$ be a positive semi-definite Hilbert–Schmidt operator. We assume its eigenvalues $\{\lambda_k\}_k$ are arranged in non-increasing order. We define

$$\varkappa(\boldsymbol{K}) \overset{\text{def}}{=} (\varLambda_1\varLambda_2)^{-1/2} \quad \text{with } \varLambda_r^2 \overset{\text{def}}{=} \sum_{k\ge r}\lambda_k^2, \text{ where } r = 1, 2. \tag{3.1}$$

Lemma B.1 investigates the properties $\varkappa(\boldsymbol{K})$. Let

$$\gamma(\boldsymbol{K}) \overset{\text{def}}{=} \varkappa(\boldsymbol{K})\operatorname{tr}(\boldsymbol{K}). \tag{3.2}$$

Note that the function $\gamma(\boldsymbol{K})$ is dimension-free (i.e., scale-invariant). Moreover, $\gamma(\boldsymbol{K}) \ge 1$. This follows from the fact that for any $r \ge 1$ it holds $\varLambda_r^2 \le \left(\sum_{k\ge r}\lambda_k\right)^2 \le (\operatorname{tr}(\boldsymbol{K}))^2$.

Proof of Lemma 3.1. The union bound ensures

$$\mathbb{P}\left\{X \le z\right\} \le \mathbb{P}\left\{Y \le \tfrac{z+m}{1-\rho}\right\} + \mathbb{P}\left\{|X-Y| > \rho Y + m\right\} \le \mathbb{P}\left\{Y \le \tfrac{z+m}{1-\rho}\right\} + \delta,$$
$$\mathbb{P}\left\{Y \le \tfrac{z-m}{1+\rho}\right\} \le \mathbb{P}\left\{X \le z\right\} + \mathbb{P}\left\{|X-Y| > \rho Y + m\right\} \le \mathbb{P}\left\{X \le z\right\} + \delta.$$

Thus

$$\mathbb{P}\left\{Y \le \tfrac{z-m}{1+\rho}\right\} - \delta \le \mathbb{P}\left\{X \le z\right\} \le \mathbb{P}\left\{Y \le \tfrac{z+m}{1-\rho}\right\} + \delta.$$

Assumption (GAR-II) yields

$$\mathbb{P}\left\{\|G\|_H \le \tfrac{z-m}{1+\rho}\right\} - \delta - \Delta \le \mathbb{P}\left\{X \le z\right\} \le \mathbb{P}\left\{\|G\|_H \le \tfrac{z+m}{1-\rho}\right\} + \delta + \Delta.$$

Now one has to bound $\mathbb{P}\left\{\|G\|_H \le \tfrac{z-m}{1+\rho}\right\}$ and $\mathbb{P}\left\{\|G\|_H \le \tfrac{z+m}{1-\rho}\right\}$.

The assumption of the lemma $\rho \in \left[0, \frac{1}{2}\right]$ together with Lemma B.2 yield

$$\mathbb{P}\left\{\|G\|_H \le \tfrac{z-m}{1+\rho}\right\} \ge \mathbb{P}\left\{\|G\|_H \le \tfrac{z}{1+\rho}\right\} - \mathtt{C}\gamma(\boldsymbol{K})\frac{m}{\sqrt{\operatorname{tr}(\boldsymbol{K})}},$$
$$\mathbb{P}\left\{\|G\|_H \le \tfrac{z+m}{1-\rho}\right\} \le \mathbb{P}\left\{\|G\|_H \le \tfrac{z}{1-\rho}\right\} + \mathtt{C}\gamma(\boldsymbol{K})\frac{m}{\sqrt{\operatorname{tr}(\boldsymbol{K})}}.$$

Now we consider a Gaussian r.v. αG with some $\alpha > 0$. Note that by definition $\varkappa(\alpha^2\boldsymbol{K}) = \frac{1}{\alpha^2}\varkappa(\boldsymbol{K})$. To compare G and αG we use Corollary 2.3 by Götze et al. [2019]. This ensures for any $z > 0$

$$\begin{aligned}\left|\mathbb{P}\left\{\|G\|_H \le \tfrac{z}{\alpha}\right\} - \mathbb{P}\left\{\|G\|_H \le z\right\}\right| &\le \mathtt{C}\left(\varkappa(\boldsymbol{K}) + \varkappa(\alpha^2\boldsymbol{K})\right)\left\|\boldsymbol{K} - \alpha^2\boldsymbol{K}\right\|_1 \\ &= \mathtt{C}\left(1 + \tfrac{1}{\alpha^2}\right)|1-\alpha^2|\varkappa(\boldsymbol{K})\operatorname{tr}(\boldsymbol{K}).\end{aligned}$$

Setting $\alpha = 1 + \rho$ and taking into account that $\rho \in [0, \frac{1}{2}]$, we obtain

$$\mathbb{P}\left\{\|G\|_H \le \tfrac{z}{1+\rho}\right\} \ge \mathbb{P}\left\{\|G\|_H \le z\right\} - \mathtt{C}\gamma(\boldsymbol{K})\rho.$$

In a similar way,

$$\mathbb{P}\left\{\|G\|_H \le \tfrac{z}{1-\rho}\right\} \le \mathbb{P}\left\{\|G\|_H \le z\right\} + \mathtt{C}\gamma(\boldsymbol{K})\rho.$$

Collecting all the bounds, we get the result. □

3.1 Gaussian approximation for generalized bootstrap

We remind that both $\mu \sim \mathcal{P}$ ($\mathcal{P}$ is supported on $\mathcal{M}(\mathbb{H}_{++}(d))$) and $\hat{\mu} \in \mathcal{M}(\mathbb{H}_{++}(d))$ are random. So, Gaussian approximation is essential for validation of (1.5). Specifically, we will show that, given two independent centred Gaussian vectors Z and Z_μ,

$$\mathcal{W}(B_\mu, B) \overset{\mathrm{d}}{\approx} \|\boldsymbol{A}Z\|_{\mathrm{F}}, \quad \mathcal{W}(B_{\hat{\mu}}, B_\mu) \overset{\mathrm{d}}{\approx} \|\boldsymbol{A}Z_\mu\|_{\mathrm{F}},$$

with $\boldsymbol{A}$ coming from (2.1). To get these results, we impose some restrictions on μ, Z, $\hat{\mu}$ and Z_μ.

Assumptions on μ We assume there exist functions $\varepsilon_T(\mathrm{x}) > 0$ and $\varepsilon_{\mathrm{F}}(\mathrm{x}) > 0$, s.t.,

$$\mathbb{P}\left\{\|T_\mu\|_{\mathrm{F}} > \varepsilon_T(\mathrm{x})\right\} \le \mathtt{C}e^{-\mathrm{x}}, \tag{T}$$

$$\mathbb{P}\left\{r(\boldsymbol{F}, \boldsymbol{F}_\mu) > \varepsilon_{\mathrm{F}}(\mathrm{x})\right\} \le \mathtt{C}e^{-\mathrm{x}}, \tag{F}$$

Let $Z \sim \mathcal{N}(0, \Xi)$ be a centred Gaussian matrix s.t.,

$$\sup_{z>0}\left|\mathbb{P}\left\{\left\|\boldsymbol{F}^{-1}T_\mu\right\|_{\mathrm{F}} \le z\right\} - \mathbb{P}\left\{\|Z\|_{\mathrm{F}} \le z\right\}\right| \le \varepsilon_G. \tag{G}$$

We denote

$$\varepsilon(\mathrm{x}) \overset{\mathrm{def}}{=} 6\sqrt{\kappa(\boldsymbol{F})}\left(c_B\varepsilon_T(\mathrm{x}) + \varepsilon_{\mathrm{F}}(\mathrm{x})\right),$$

with c_B coming from (2.11).

Lemma 3.2 (Gaussian approximation for $\mathcal{W}(B_\mu, B)$). *Denote $\Xi' \overset{\mathrm{def}}{=} \boldsymbol{A}\Xi\boldsymbol{A}$. Let Assumptions ($T$), ($F$) and ($G$) be fulfilled. Then*

$$\sup_{z\ge 0}\left|\mathbb{P}\left\{\mathcal{W}(B_\mu, B) \le z\right\} - \mathbb{P}\left\{\|\boldsymbol{A}Z\|_{\mathrm{F}} \le z\right\}\right| \le \mathcal{E},$$

$$\mathcal{E} \overset{\mathrm{def}}{=} \varepsilon_G + \mathtt{C}\cdot\inf_{\mathrm{x}\in\mathcal{X}}\left\{e^{-\mathrm{x}} + \sqrt{\kappa(B)}\gamma(\Xi')\varepsilon(\mathrm{x})\right\}, \quad \mathcal{X} \overset{\mathrm{def}}{=} \left\{\mathrm{x}: \ \varepsilon(\mathrm{x}) \le \frac{1}{6\sqrt{\kappa(B)}}\right\}.$$

The proof is in Appendix B.1.

Assumptions on $\hat{\mu}$ We recall that $\hat{\mu}$ is a non-zero random measure that might depend on μ. We assume there exists a Borel set $\mathcal{A}_t \subset \mathcal{M}(\mathbb{H}_{++}(d))$, s.t. $\mathbb{P}\{\mu \in \mathcal{A}_t\} \ge 1 - \mathtt{C}e^{-t}$). The following assumptions hold on this event.

We assume there exist functions $\hat{\varepsilon}_T(\mathrm{x}, \mathrm{t}) > 0$ and $\hat{\varepsilon}_F(\mathrm{x}, \mathrm{t}) > 0$, s.t.,

$$\mathbb{P}\left\{\|T_{\hat{\mu}} - T_\mu\|_{\mathrm{F}} > \hat{\varepsilon}_T(\mathrm{x}, \mathrm{t}) \mid \mu\right\} \le \mathtt{C}e^{-\mathrm{x}}, \tag{$\hat{T}$}$$

$$\mathbb{P}\left\{r(\boldsymbol{F}, \boldsymbol{F}_{\hat{\mu}}) > \hat{\varepsilon}_F(\mathrm{x}, \mathrm{t}) \mid \mu\right\} \le \mathtt{C}e^{-\mathrm{x}}, \tag{$\hat{F}$}$$

Let $Z_\mu \sim \mathcal{N}(0, \Xi_\mu)$ be centred Gaussian matrix s.t.,

$$\sup_{z>0}\left|\mathbb{P}\left\{\left\|\boldsymbol{F}^{-1}\left(T_{\hat{\mu}} - T_\mu\right)\right\|_{\mathrm{F}} \le z \mid \mu\right\} - \mathbb{P}\left\{\|Z_\mu\|_{\mathrm{F}} \le z \mid \mu\right\}\right| \le \hat{\varepsilon}_G(\mathrm{t}). \tag{$\hat{G}$}$$

We denote

$$\hat{\varepsilon}(\mathrm{x}, \mathrm{t}) \overset{\mathrm{def}}{=} 6\sqrt{\kappa(\boldsymbol{F})}\left(c_B\hat{\varepsilon}_T(\mathrm{x}, \mathrm{t}) + \hat{\varepsilon}_F(\mathrm{x}, \mathrm{t})\right).$$

Lemma 3.3 (Gaussian approximation for $\mathcal{W}(B_\mu, B_{\hat{\mu}})$)**.** *Denote* $\Xi'_\mu \overset{\text{def}}{=} \boldsymbol{A}\Xi_\mu\boldsymbol{A}$*. Let Assumptions* $(\hat{T})$*,* $(\hat{F})$ *and* $(\hat{G})$ *be fulfilled. Then, on the event* $\left\{\mu \in \mathcal{A}_t,\ \rho \le \frac{1}{12\sqrt{\kappa(B)}}\right\}$*, it holds that*

$$\sup_{z\ge 0}\left|\mathbb{P}\left\{\mathcal{W}(B_{\hat{\mu}}, B_\mu) \le z \mid \mu\right\} - \mathbb{P}\left\{\|\boldsymbol{A}Z_\mu\|_{\mathrm{F}} \le z \mid \mu\right\}\right| \le \hat{\mathcal{E}}(\mathrm{t}),$$

$$\hat{\mathcal{E}}(\mathrm{t}) \overset{\text{def}}{=} \hat{\varepsilon}_G(\mathrm{t}) + \mathtt{C}\cdot \inf_{\mathrm{x}\in\hat{\mathcal{X}}(\mathrm{t})}\left\{e^{-\mathrm{x}} + \gamma(\Xi'_\mu)\sqrt{\kappa(B)}\left(\rho + \hat{\varepsilon}(\mathrm{x},\mathrm{t})\right)\left(\frac{\|\boldsymbol{A}\|\|\boldsymbol{F}^{-1}T_\mu\|_{\mathrm{F}}}{\sqrt{\mathrm{tr}(\Xi'_\mu)}} + 1\right)\right\},$$

$$\hat{\mathcal{X}}(\mathrm{t}) \overset{\text{def}}{=} \left\{\mathrm{x}:\ \hat{\varepsilon}(\mathrm{x},\mathrm{t}) \le \frac{1}{12\sqrt{\kappa(B)}}\right\}.$$

The proof is in Appendix B.1.

4 Bootstrap validity

To complete the proof, one has to ensure $\|\boldsymbol{A}Z\|_{\mathrm{F}} \overset{\mathrm{d}}{\approx} \|\boldsymbol{A}Z_\mu\|_{\mathrm{F}}$. The approximation relies on the following assumption,

$$\mathbb{P}\left\{\|\Xi - \Xi_\mu\|_1 > \varepsilon_\Xi(\mathrm{x})\right\} \le \mathtt{C}e^{-\mathrm{x}}, \tag{Ξ}$$

with $\|\cdot\|_1$ being 1-Schatten norm.

Theorem 4.1 (Bootstrap validity)**.** *Let all Assumptions* (T) – (Ξ) *be fulfilled. Denote* $\Xi' \overset{\text{def}}{=} \boldsymbol{A}\Xi\boldsymbol{A}$ *and let* $\mathrm{t} \ge 0$ *be s.t.*

$$\varepsilon_\Xi(\mathrm{t}) \le \mathtt{C}\frac{\varLambda_2^2\left(\Xi'\right)}{\|\boldsymbol{A}\|^2\|\Xi'\|}. \tag{4.1}$$

Then with probability at least $1 - \mathtt{C}e^{-\mathrm{t}}$*,*

$$\sup_{z\ge 0}\left|\mathbb{P}\left\{\mathcal{W}(B_\mu, B) \le z\right\} - \mathbb{P}\left\{\mathcal{W}(B_{\hat{\mu}}, B_\mu) \le z \mid \mu\right\}\right| \le \varGamma(\mathrm{t}),$$

$$\begin{aligned}\varGamma(\mathrm{t}) = {} & \mathtt{C}\varkappa(\Xi')\|\boldsymbol{A}\|^2\varepsilon_\Xi(\mathrm{t}) + \\ & + \varepsilon_G + \mathtt{C}\cdot\inf_{\mathrm{x}\in\mathcal{X}}\left\{e^{-\mathrm{x}} + \gamma(\Xi')\sqrt{\kappa(B)}\varepsilon(\mathrm{x})\right\} + \\ & + \hat{\varepsilon}_G(\mathrm{t}) + \mathtt{C}\cdot\inf_{\mathrm{x}\in\hat{\mathcal{X}}(\mathrm{t})}\left\{e^{-\mathrm{x}} + \gamma(\Xi')\sqrt{\kappa(B)}\left(\varepsilon(\mathrm{t}) + \hat{\varepsilon}(\mathrm{x},\mathrm{t})\right)\left(\tfrac{\|\boldsymbol{A}\|\|\boldsymbol{F}^{-1}\|}{\sqrt{\mathrm{tr}(\Xi')}}\varepsilon_T(\mathrm{t}) + 1\right)\right\},\end{aligned}$$

where

$$\mathcal{X} \overset{\text{def}}{=} \left\{\mathrm{x}:\ \varepsilon(\mathrm{x}) \le \frac{1}{6\sqrt{\kappa(B)}}\right\}, \quad \hat{\mathcal{X}}(\mathrm{t}) \overset{\text{def}}{=} \left\{\mathrm{x}:\ \hat{\varepsilon}(\mathrm{x},\mathrm{t}) \le \frac{1}{12\sqrt{\kappa(B)}}\right\}.$$

The proof is in Appendix C. We note that a similar result for $\|B - B_\mu\|_{\mathrm{F}}$ and $\|B_{\hat{\mu}} - B_\mu\|_{\mathrm{F}}$ can be easily obtained.

5 Multiplier bootstrap

To illustrate the method, we consider the setting presented in Section 1.3. We assume that both the data distribution P and the weight distribution W are sub-exponential.

Let $\|\cdot\|_{\psi_\alpha}$ be the Orlicz ψ_α-norm. In what follows, we consider only the cases $\alpha = 1, 2$. Recall that $\|X\|_{\psi_1} \le C$ and $\|X\|_{\psi_2} \le C$ characterize sub-exponential and sub-Gaussian random variables, respectively.

Let $S_1, \dots, S_n \in \mathbb{H}_{++}(d)$ be i.i.d. random observations, where $S_i \overset{\text{iid}}{\sim} P$. We assume that

$$\|\operatorname{tr} S_i\|_{\psi_1} = v_P < \infty, \tag{P}$$

We also assume that the bootstrap weights are independent of the data. Specifically, let $w_1, \dots, w_n$ be i.i.d. random variables, where $w_i \overset{\text{iid}}{\sim} W$ satisfying

$$w \sim W, \quad \|w - 1\|_{\psi_1} = v_w < \infty, \quad \mathbb{E}\, w = \operatorname{Var} w = 1. \tag{W}$$

Some specific examples are the exponential, the Poisson, or the Bernoulli weights.

Following the accepted notations, we define:

$$P_n \overset{\text{def}}{=} \mu = \frac{1}{n} \sum_i \delta_{S_i}, \quad P_w \overset{\text{def}}{=} \hat{\mu} = \frac{1}{n} \sum w_i \delta_{S_i}. \tag{5.1}$$

Consequently, we define:

$$B \overset{\text{def}}{=} \mathcal{B}(P), \quad B_n \overset{\text{def}}{=} B_\mu = \mathcal{B}(P_n), \quad B_w \overset{\text{def}}{=} B_{\hat{\mu}} = \mathcal{B}(P_w). \tag{5.2}$$

We note that in the case of Bernoulli or Poisson weights, it is possible for $\sum_i w_i = 0$, which would result in $B_w = 0$. This introduces an additional term in the approximation bound $\mathit{\Gamma}(\mathrm{t})$.

By definition (see (2.3)) $\boldsymbol{F} = \mathcal{F}(P) = -\mathbb{E}\, \boldsymbol{dT}_B^{S_1}$. We set

$$\boldsymbol{\Xi} \overset{\text{def}}{=} \frac{1}{n} \boldsymbol{F}^{-1} \left[\mathbb{E} \left(T_B^{S_1} - I \right) \otimes \left(T_B^{S_1} - I \right) \right] \boldsymbol{F}^{-1}.$$

Theorem 5.1. *Let Assumptions* (P) *and* (W) *be fulfilled. Let* p_0 *be the probability of observing* $w_i = 0$*, i.e.,* $p_0 = \mathbb{P}_w\{w_i = 0\}$*. Then, with h.p.*

$$\sup_{z \ge 0} \left| \mathbb{P}\left\{ \sqrt{n} \mathcal{W}(B_n, B) \le z \right\} - \mathbb{P}\left\{ \sqrt{n} \mathcal{W}(B_w, B_n) \le z \mid \mu \right\} \right| \le \mathit{\Gamma}(\mathrm{t}) + p_0^n.$$

Moreover, denote $\sigma_T^2 \overset{\text{def}}{=} \mathbb{E}\, \|T_1\|_{\mathrm{F}}^2$*,* $\sigma_F^2 \overset{\text{def}}{=} \left\| \mathbb{E} \left(\boldsymbol{dT}_B^{S_1} - \boldsymbol{F} \right)^2 \right\|$*, and let* $\hat{C}_\varepsilon, C_T, C_G > 0$ *be dimension-free constants. Then, for sufficiently large* n *(depending on* t*),*

$$\mathit{\Gamma}(\mathrm{t}) \lesssim d^3 \sqrt{\frac{C_G}{n}} + \kappa(\boldsymbol{\Xi}') \|\boldsymbol{A}\|^2 \left\| \boldsymbol{F}^{-1} \right\|^2 \sigma_T^2 \left(\sqrt{\frac{\hat{C}_\varepsilon}{n} \left(\mathrm{t} + \log \frac{nd}{\hat{C}_\varepsilon} \right)} + \sqrt{\frac{C_T}{n} (t + d^2)} \right).$$

Appendix D contains the proof. The explicit expressions of constants $\hat{C}_\varepsilon, C_T, C_G$ can be found in (D.10), (D.8) and (D.9), respectively. The explicit condition on the sample size n is in (D.11).

Remark 5.2. *The rate* $\frac{d^3}{\sqrt{n}}$ *is due to the technique used in the proof of the Gaussian approximation results (see Lemma D.6 and D.7). Specifically, we get* d^3 *instead of* $d^{3/2}$ *because the results are in the space of* $d \times d$ *matrices.*

5.1 Ideas behind the proof

The proof of the theorem relies on verifying all the assumptions outlined in Section 3 and Section 4. The following lemma plays a crucial role in this validation.

Lemma 5.3. *Let Assumption* (P) *be true. Then for fixed* $B \in \mathbb{H}_{++}(d)$ *and* $S \sim P$ *it holds for some constants* $v_S, v_T, v_F > 0$*, that*

$$\Big\| \|S\|^{1/2} \Big\|_{\psi_2} \le v_S, \quad \big\| \|T_B^S\|_{\mathrm{F}} \big\|_{\psi_2} \le v_T, \quad \big\| \|\boldsymbol{dT}_B^S\| \big\|_{\psi_2} \le v_F.$$

We explicitly estimate the bounding terms from all Assumptions (T) – (Ξ) and summarize them in Tab. 2 and Tab. 3. For the sake of simplicity, we omit the explicit constants. However, they can be tracked in the proofs in Appendix D.

	Assumption (P) **ensures that**	**Assumptions** (P) **and** (W) **ensure that**
Assumptions validating GAR	Assumption (T) holds due to Lemma D.4, $\varepsilon_T(\mathrm{x}) \lesssim \sigma_T \sqrt{\frac{\mathrm{x}}{n}}$	Assumption $(\hat{T})$ holds due to Lemma D.5, $\hat{\varepsilon}_T(\mathrm{x};\mathrm{t}) \lesssim \sigma_T \sqrt{\frac{\mathrm{x}}{n}}$
	Assumption (F) holds due to Lemma D.8, $\varepsilon_F(\mathrm{x}) \lesssim \|\boldsymbol{F}^{-1}\| \sigma_F \sqrt{\frac{\mathrm{x} + \log d}{n}}$	Assumption $(\hat{F})$ holds due to Lemma D.9, $\hat{\varepsilon}_F(\mathrm{x},\mathrm{t}) \lesssim (\|\boldsymbol{F}^{-1}\| \sigma_F + 1) \times$ $\times \sqrt{\frac{\mathrm{x} + \mathrm{t} + \log d}{n}}$
	Assumption (G) holds due Lemma D.6, $\varepsilon_G \lesssim d^3 \sqrt{\frac{C_G}{n}}$	Assumption $(\hat{G})$ holds due to Lemma D.7, $\hat{\varepsilon}_G(\mathrm{t}) \lesssim d^3 \sqrt{\frac{C_G}{n}}$
GAR	The result is due Lemma D.11, $\mathcal{E} \lesssim d^3 \sqrt{\frac{C_G}{n}} + \gamma(\boldsymbol{\Xi}') \sqrt{\frac{C_\varepsilon}{n} \log \frac{nd}{C_\varepsilon}}$	The result is due Lemma D.12, $\hat{\mathcal{E}}(\mathrm{t}) \lesssim d^3 \sqrt{\frac{C_G}{n}}$ $+ \gamma(\boldsymbol{\Xi}'_\mu) M_\mu \sqrt{\frac{\hat{C}_\varepsilon}{n} \left(\mathrm{t} + \log \frac{nd}{\hat{C}_\varepsilon} \right)},$ $M_\mu \overset{\mathrm{def}}{=} 1 + \sqrt{\frac{\mathrm{tr}(\boldsymbol{\Xi}')}{\mathrm{tr}(\boldsymbol{\Xi}'_\mu)}}$

Table 2: Assumptions for GAR and GAR

Cov. comp.	Assumption (Ξ) holds due Lemma D.10 $\varepsilon_{\Xi}(\mathrm{t}) \lesssim \sigma_T^2 \|\boldsymbol{F}^{-1}\|^2 \sqrt{C_T \frac{\mathrm{t}+d^2}{n}}$

Table 3: Covariance comparison

6 Experiments on graph-structured data

The aim of this section is twofold. First, drawing on the ideas from Haasler and Frossard [2024], we demonstrate how the multiplier bootstrap performs on both synthetic and real graph-structured data, specifically related to brain connectomes. Second, we compare the approximating distribution constructed via multiplier bootstrap with the asymptotic distribution presented in Corollary 2.1 of Kroshnin et al. [2021]. The code supporting our experiments is available at `https://github.com/asuvor/bw_paper/`.

6.1 Bures–Wasserstein barycenters of graphs

Haasler and Frossard [2024] propose a novel framework for defining and computing the mean of a set of graphs using the Bures–Wasserstein distance. In the following, we adhere to this setting and present it for completeness.

The authors focus on aligned graphs, meaning graphs with the same number of nodes, with each node corresponding to a specific node in the other graphs. For instance, each vertex might represent a specific area of the head where an electrode is placed to capture EEG signals. Section 6.2 introduces this setting in more detail.

Let G be an undirected weighted graph with d nodes without self-loops. In the following, we assume the weights to be positive. The adjacency matrix and degree matrix of G are denoted as A_G and D_G, respectively. The graph Laplacian of G is defined as

$$L \stackrel{\text{def}}{=} D_G - A_G.$$

Denote by $\mathcal{G}(d)$ the set of aligned positive-weighted and connected graphs with d nodes. The Bures–Wasserstein distance between $G_1 \in \mathcal{G}$ and $G_2 \in \mathcal{G}(d)$ is

$$\mathcal{W}_{\mathcal{G}}(G_1, G_2) = \mathcal{W}(L_1^{\dagger}, L_2^{\dagger}),$$

where $L_1^{\dagger}$, $L_2^{\dagger}$ are the pseudo-inverses of their graph Laplacians.

Consider a population of graphs $G_1, \ldots, G_n \in \mathcal{G}(d)$. Let the corresponding graph Laplacian be $L_1, \ldots, L_n$. The authors reduce the problem of finding the barycenter of the graphs to the problem of finding the barycenter of their inverted graph Laplacians.

Since all G_i are connected, $L_1, \ldots, L_n$ share the same kernel, $\operatorname{span}(\mathbf{1}_d)$, where $\mathbf{1}_d \in \mathbb{R}^d$ is the vector of all ones. Thus, it suffices to restrict the Laplacians to the orthogonal complement of the kernel and then compute the barycenter.

In what follows, we denote restricted inverse graph Laplacians as

$$S_i \stackrel{\text{def}}{=} U_{\mathbf{1}_d}^{\top} L_i^{\dagger} U_{\mathbf{1}_d} = \left(U_{\mathbf{1}_d}^{\top} L_i U_{\mathbf{1}_d}\right)^{-1}, \tag{6.1}$$

with $U_{\mathbf{1}_d} \in \mathbb{R}^{d\times(d-1)}$ being a matrix of an orthonormal basis on $\operatorname{span}(\mathbf{1}_d)^{\perp}$. By construction, $S_i \in \mathbb{H}_{++}(d-1)$.

In many practical problems, observed graphs are supposed to be i.i.d., $G_i \overset{\text{iid}}{\sim} P_G$. Consequently, the corresponding graph Laplacians L_i and their inverted restrictions S_i are i.i.d. ($S_i \overset{\text{iid}}{\sim} P$).

6.2 Data used in experiments

Brain connectomes are complex networks representing the connections between different brain regions. These connections include the physical links between neurons (nerve cells) and the functional connections between brain regions, which are determined by patterns of neuronal activity. Understanding its structure and function is crucial for uncovering how the brain processes information and produces behavior [Bullmore and Sporns, 2009, Fornito et al., 2016a].

A connectome can be represented as a graph. The nodes correspond to individual brain regions, and the edges correspond to connections, pathways, or interactions. This representation serves as the foundation for all subsequent investigations into the organization of the network [Fornito et al., 2016b]. To illustrate bootstrap performance, we use two data sets related to connectomes.

Synthetic data To generate the synthetic data, we use the weighted stochastic block model. It is well suited for describing natural phenomena. For example, Faskowitz et al. [2018] applied it to human connectomes.

Real data To illustrate the performance on real-world data, we use the EEGBCI dataset documented by Schalk et al. [2004]. This dataset contains electroencephalographic (EEG) recordings intended for brain-computer interface (BCI) research.

6.3 Experiments on synthetic data: bootstrap vs. asymptotic confidence sets

This section compares asymptotic confidence sets with the non-asymptotic ones introduced in Kroshnin et al. [2021]. We briefly recall the concept of asymptotic confidence sets, followed by a detailed presentation of the data-generating model and the simulation results.

Asymptotic confidence sets Let $S_1, \dots, S_n$ be i.i.d., $S_i \overset{\text{iid}}{\sim} P$, with P satisfying Assumption (P). Let P_n be an empirical counterpart of P (see (5.1)). Recall that

$$B = \mathcal{B}(P), \quad B_n = \mathcal{B}(P_n).$$

Further, we set the scaling operator $\boldsymbol{F}_n$ and the empirical covariance of $T_{B_n}^{S_i}$ to be

$$\boldsymbol{F}_n \overset{\text{def}}{=} -\frac{1}{n}\sum_i \boldsymbol{dT}_{B_n}^{S_i}, \quad \boldsymbol{\Sigma}_n \overset{\text{def}}{=} \frac{1}{n}\sum_i \left(T_{B_n}^{S_i} - I\right) \otimes \left(T_{B_n}^{S_i} - I\right). \tag{6.2}$$

And let Z be a centered Gaussian vector, $Z \sim \mathcal{N}(0, \Xi_n)$ with $\Xi_n \overset{\text{def}}{=} \boldsymbol{F}_n^{-1}\boldsymbol{\Sigma}_n\boldsymbol{F}_n^{-1}$. Corollary 2.1 by Kroshnin et al. [2021] ensures that

$$\sqrt{n}\mathcal{W}(B_n, B) \overset{\text{d}}{\approx} \left\|B_n^{1/2}\boldsymbol{dT}_{B_n}^{B_n}(Z)\right\|_{\mathrm{F}}, \text{ as } n \to \infty. \tag{6.3}$$

We aim to compare this confidence set with the one presented in Section 5. For completeness, we recall it here.

Given the distribution W of the bootstrap weights $w_1, \ldots, w_n$, $w_i \overset{\text{iid}}{\sim} W$, we construct $P_w = \frac{1}{n}\sum_i w_i \delta_{S_i}$ and the bootstrap barycenter $B_w = \mathcal{B}(P_w)$ (see (5.2)). Theorem 5.1 ensures

$$\sqrt{n}\mathcal{W}(B_n, B) \overset{\mathrm{d}}{\approx} \sqrt{n}\mathcal{W}(B_w, B_n). \tag{6.4}$$

Computational aspects. We use the iterative algorithm proposed by Álvarez-Esteban et al. [2016] to compute the barycenters. The computation complexity of a barycenter can be estimated as follows. Let I denote the average number of iterations in the iterative algorithm. Thus, computing the barycenter requires $O(I \cdot n \cdot \mathcal{K}(d))$ operations, where $\mathcal{K}(d)$ is the complexity of matrix operations (matrix inversion and matrix square root computations).

The best-known complexity for matrix inversion is approximately $O(d^{2.38})$. Moreover, to the best of our knowledge, the complexity of computing the square root of a matrix is $O(d^3)$. Therefore, $K(d) = O(d^3)$, resulting in a total computational complexity of $O(I \cdot n \cdot d^3)$.

Thus, the computational complexity of the multiplier bootstrap is $O(M \cdot I \cdot n \cdot d^3)$, where M is the number of resamplings.

To measure the computational complexity of estimating the asymptotic distribution, we note that the operator $\boldsymbol{dT}_B^S(X)$ admits an explicit representation; see Lemma A.2 by Kroshnin et al. [2021]. Specifically, computing each entry in its matrix representation requires $O(d^2)$ operations. Since the operator's dimension is $\frac{d(d+1)}{2}$, the total complexity of constructing the matrix representation is $O(d^2 \cdot (d^2)^2) = O(d^6)$.

Therefore, computing the representation of $\boldsymbol{F}_n$ (see (6.2)) requires $O(n \cdot d^6)$ operations. Additionally, sampling a Gaussian matrix Z involves $O(d^4)$ operations. Thus, the total complexity can be estimated as $O(n \cdot d^6 + M \cdot d^4)$, where M is the number of resamplings.

Therefore, for large d, estimating the asymptotic distribution can be significantly more resource-intensive compared to the bootstrap method.

Weighted stochastic block model (WSBM) We use WSBM data to compare non-asymptotic and asymptotic confidence sets.

Each generated graph G has d nodes divided into two non-overlapping groups (communities). The size of each group is random: the first group contains $d_1 = \frac{d}{2} - \text{Unif}\{-2, 2\}$ nodes, and the second group contains $d_2 = d - d_1$ nodes.

The corresponding adjacency matrix $A \in \mathbb{R}^{d \times d}$ has a block structure

$$A = \begin{pmatrix} C_{11} & C_{12} \\ C_{21} & C_{22} \end{pmatrix},$$

where C_{11} and C_{22} represent intra-group connections and $C_{12} = C_{21}^T$ represents inter-group connections.

The probabilities of observing a non-zero edge between each pair of nodes within the corresponding blocks are $p_{11} = 0.8, p_{22} = 0.7, p_{12} = p_{21} = 0.3$. The weights within each block are i.i.d. Poisson,

$$a \sim \begin{cases} \mathrm{Po}(20) & \text{for } a \in C_{11}, \\ \mathrm{Po}(15) & \text{for } a \in C_{22}, \\ \mathrm{Po}(6) & \text{for } a \in C_{12}, C_{21}. \end{cases}$$

To ensure that a generated graph is connected, we consider $A + \varepsilon E$, with E being $d \times d$ matrix of all ones. In the experiments, we set $\varepsilon = 1$.

Experiments We consider two scenarios with dimensions $d = 8$ and $d = 40$. We generated $N = 8000$ $d \times d$ WSBM adjacency matrices to model the population and constructed S_i as in (6.1). This experimental setup aligns with the framework introduced in Section 5.

The population barycenter B is computed using the entire dataset. We estimate the empirical barycenter B_n for different sample sizes $n \in \{10, 30, 100\}$.

To estimate the empirical cumulative distribution function (ECDF) of $\sqrt{n}\mathcal{W}(B_n, B)$, we subsample n observations with replacement from the entire population. For estimating the ECDFs of $\sqrt{n}\mathcal{W}(B_n, B_w)$, we employ the multiplier bootstrap method as described in (6.4) and set the bootstrap weights to be Poisson, $w_i \sim$ Po(1). To compute the asymptotic confidence sets, we utilize the procedure outlined in (6.3).

Figures 1 and 2 present the results for $d = 8$ and $d = 40$, respectively. The dark-blue ECDF represents the distribution of $\sqrt{n}\mathcal{W}(B, B_n)$. The light-blue ECDFs correspond to the distributions of $\sqrt{n}\mathcal{W}(B_n, B_w)$ (upper panel), while the orange ECDFs depict the distributions of $\|B_n^{1/2}\boldsymbol{dT}_{B_n}^{B_n}(Z)\|$ (middle panel). For each sample size n and each case, we generate 100 independent curves.

Finally, to evaluate the quality of the approximations provided by (6.4) and (6.3), we compute the Kolmogorov distance between the ECDF of $\sqrt{n}\mathcal{W}(B, B_n)$ and the realizations of $\sqrt{n}\mathcal{W}(B_n, B_w)$ and $\|B_n^{1/2}\boldsymbol{dT}_{B_n}^{B_n}(Z)\|$, respectively.

The lower panel illustrates the distributions of the Kolmogorov distances: the light-blue curves correspond to the bootstrap case, while the orange curves represent the asymptotic case.

For each dimension d, sample size n, and approximation method, the mean Kolmogorov distance and its standard deviation are displayed at the bottom right corner of the corresponding subplot.

6.4 Experiments on connectomes

The EEGBCI dataset contains EEG recordings from 64 electrodes from 109 participants. Each participant completed 14 sessions, corresponding to a distinct motor imagery task, i.e., a task associated with imagined movements.

Each electrode captures electrical activity from a particular region of the scalp and the underlying brain regions when a person fulfills an imagery task. From these recordings, functional connectomes were constructed.

Functional connectomes are networks that show how different brain regions connect and interact based on EEG data. Each node corresponds to a particular brain region. Edge weights represent the interactions between these regions, quantified using some chosen connectivity metric.

To construct the connectomes, we used EEG signals from 3 tasks (imagining moving the left hand and the feet). The edge weight between two nodes is the envelope correlation between the EEG signals from the corresponding pairs of electrodes.

Thus, we got 109 connectomes of size 64×64. We convert them to projected graph Laplacians as described in Section 6.2.

Using the entire population, we computed the true barycenter B of the observed restricted graph

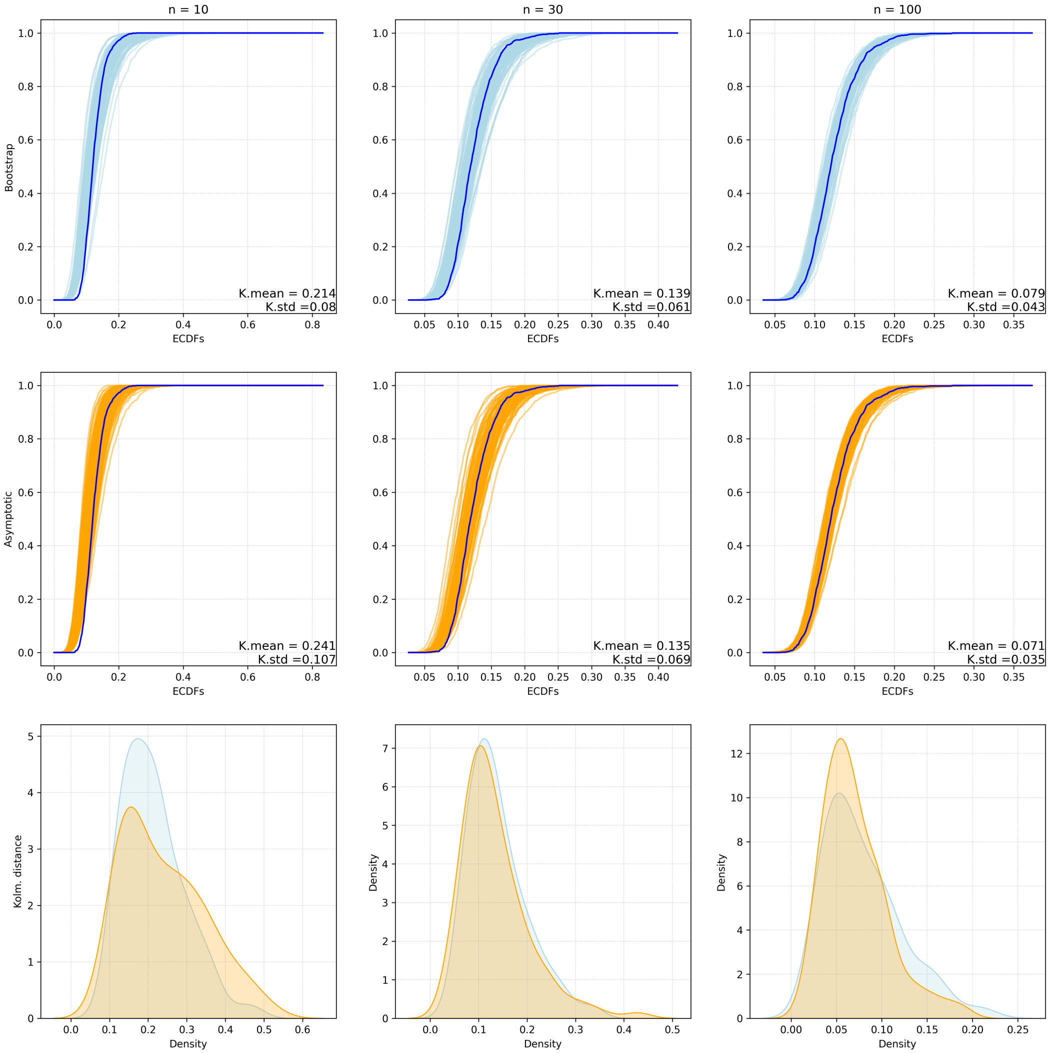

Figure 1: **WSBM data,** $d = 8$**.** Empirical cumulative distribution functions (ECDFs) of $\sqrt{n}\mathcal{W}(B, B_n)$ (dark blue), $\sqrt{n}\mathcal{W}(B_n, B_w)$ (light blue), and $\|B_n^{1/2}\boldsymbol{dT}_{B_n}^{B_n}(Z)\|$ (orange) are shown. All ECDFs are computed using 1000 observations. The lower panel shows the distribution of the Kolmogorov distance between the true ECDF and the bootstrap curves (light blue), as well as between the true ECDF and the “asymptotic” curves (orange).

Laplacians $S_1, \dots, S_{109}$. To estimate the distribution of $\sqrt{n}\mathcal{W}(B, B_n)$ (for $n = 10, 50, 70$), we sampled with replacement from the population. To estimate the distribution of $\sqrt{n}\mathcal{W}(B_n, B_w)$, we employed the multiplier bootstrap approach as outlined in (6.4) , using Poisson-distributed weights $w_i \sim \text{Po}(1)$.

Figure (3) presents the result.

For each sample size n, the mean Kolmogorov distance and its variance are displayed in the bottom-right corner of the corresponding subplot.

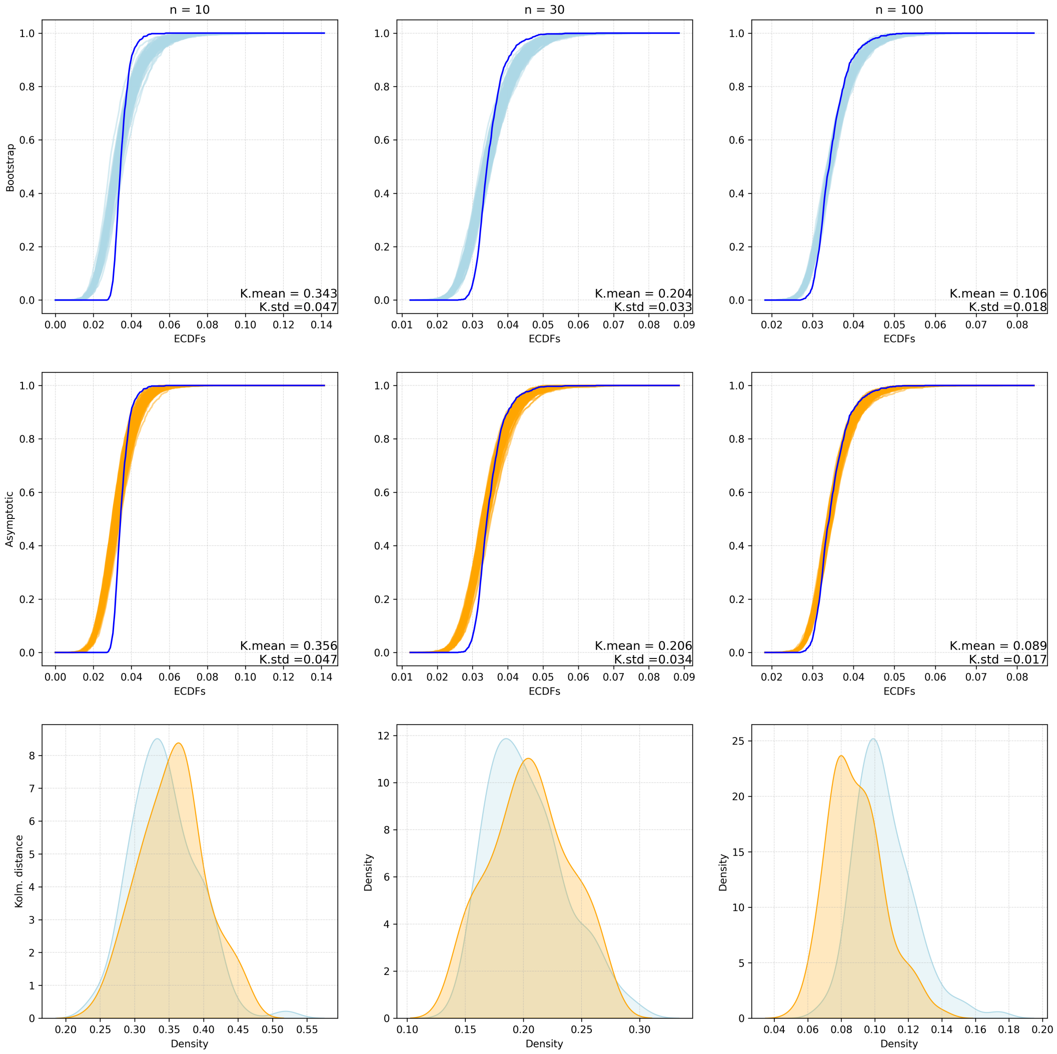

Figure 2: **WSBM data,** $d = 40$**.** Empirical cumulative distribution functions (ECDFs) of $\sqrt{n}\mathcal{W}(B, B_n)$ (dark blue), $\sqrt{n}\mathcal{W}(B_n, B_w)$ (light blue), and $\|B_n^{1/2}\boldsymbol{dT}_{B_n}^{B_n}(Z)\|$ (orange) are shown. All ECDFs are computed using 1000 observations. The lower panel shows the distribution of the Kolmogorov distance between the true ECDF and the bootstrap curves (light blue), as well as between the true ECDF and the “asymptotic” curves (orange).

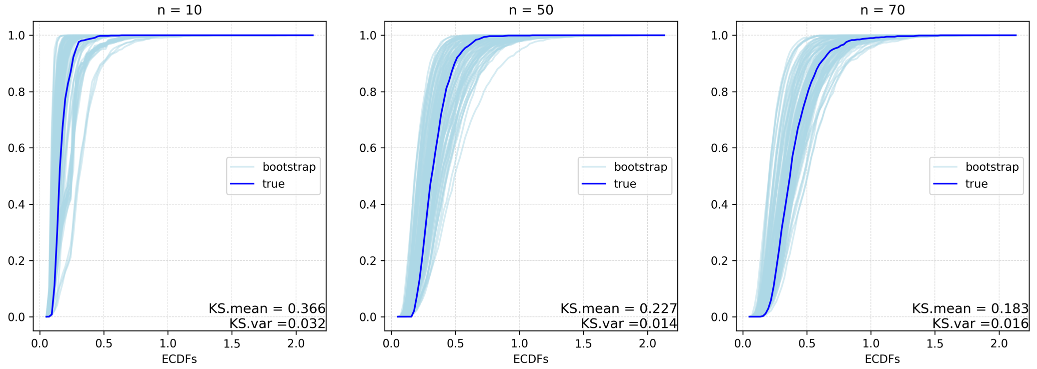

Figure 3: **EEGBCI data,** $d = 64$. Empirical distribution functions for $\sqrt{n}\mathcal{W}(B_n, B_w)$ (light-blue) and $\sqrt{n}\mathcal{W}(B_n, B)$ (dark-blue). To compute B_w, we use the Poisson weights, $w \sim$ Po(1).

References

Alexey Kroshnin, Vladimir Spokoiny, and Alexandra Suvorikova. Statistical inference for Bures–Wasserstein barycenters. *The Annals of Applied Probability*, 31(3):1264 – 1298, 2021.

Luigi Ambrosio, Nicola Gigli, and Giuseppe Savare. *Gradient Flows: In Metric Spaces and in the Space of Probability Measures*. Springer Science & Business Media, 2008. ISBN 978-3-7643-8722-8.

Cédric Villani. *Optimal Transport*, volume 338 of *Grundlehren der mathematischen Wissenschaften*. Springer Berlin Heidelberg, 2009. ISBN 978-3-540-71049-3 978-3-540-71050-9.

Filippo Santambrogio. Optimal transport for applied mathematicians. *Birkäuser, NY*, 55(58-63):94, 2015.

Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein generative adversarial networks. In *International conference on machine learning*, pages 214–223. PMLR, 2017.

Gabriel Khan and Jun Zhang. When optimal transport meets information geometry. *Information Geometry*, 5(1):47–78, 2022.

Nicolas Bonneel and Julie Digne. A survey of optimal transport for computer graphics and computer vision. In *Computer Graphics Forum*, volume 42, pages 439–460. Wiley Online Library, 2023.

Alfred Galichon. *Optimal transport methods in economics*. Princeton University Press, 2018.

Geoffrey Schiebinger, Jian Shu, Marcin Tabaka, Brian Cleary, Vidya Subramanian, Aryeh Solomon, Joshua Gould, Siyan Liu, Stacie Lin, Peter Berube, et al. Optimal-transport analysis of single-cell gene expression identifies developmental trajectories in reprogramming. *Cell*, 176(4):928–943, 2019.

Eustasio Del Barrio, Hélène Lescornel, and Jean-Michel Loubes. A statistical analysis of a deformation model with wasserstein barycenters: estimation procedure and goodness of fit test. *arXiv preprint arXiv:1508.06465*, 2015.

Thomas Rippl, Axel Munk, and Anja Sturm. Limit laws of the empirical wasserstein distance: Gaussian distributions. *Journal of Multivariate Analysis*, 151:90–109, 2016.

E del Barrio, JA Cuesta-Albertos, C Matrán, and A Mayo-Íscar. Robust clustering tools based on optimal transportation. *Statistics and Computing*, pages 1–22, 2017.

Sergey Bobkov and Michel Ledoux. *One-dimensional empirical measures, order statistics, and Kantorovich transport distances*, volume 261. American Mathematical Society, 2019.

Victor M Panaretos and Yoav Zemel. *An invitation to statistics in Wasserstein space*. Springer Nature, 2020.

Florian Heinemann, Axel Munk, and Yoav Zemel. Randomized wasserstein barycenter computation: Resampling with statistical guarantees. *SIAM Journal on Mathematics of Data Science*, 4(1):229–259, 2022.

Sinho Chewi, Jonathan Niles-Weed, and Philippe Rigollet. Statistical optimal transport. *arXiv preprint arXiv:2407.18163*, 2024.

Dror Simon and Aviad Aberdam. Barycenters of natural images constrained wasserstein barycenters for image morphing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7910–7919, 2020.

Kevin Cheng, Shuchin Aeron, Michael C Hughes, and Eric L Miller. Dynamical wasserstein barycenters for time-series modeling. *Advances in Neural Information Processing Systems*, 34:27991–28003, 2021.

Benjamin Larvaron, Marianne Clausel, Antoine Bertoncello, Sébastien Benjamin, Georges Oppenheim, and Clément Bertin. Conditional wasserstein barycenters to predict battery health degradation at unobserved experimental conditions. *Journal of Energy Storage*, 78:110015, 2024.

Susanna Levantesi, Andrea Nigri, Paolo Pagnottoni, and Alessandro Spelta. Wasserstein barycenter regression: application to the joint dynamics of regional gdp and life expectancy in italy. *AStA Advances in Statistical Analysis*, pages 1–24, 2024.

Anton Mallasto and Aasa Feragen. Learning from uncertain curves: The 2-wasserstein metric for gaussian processes. *Advances in Neural Information Processing Systems*, 30, 2017.

Boris Muzellec and Marco Cuturi. Generalizing point embeddings using the Wasserstein space of elliptical distributions. In *Advances in Neural Information Processing Systems*, pages 10237–10248, 2018.

Rajendra Bhatia, Tanvi Jain, and Yongdo Lim. On the bures–wasserstein distance between positive definite matrices. *Expositiones Mathematicae*, 37(2):165–191, 2019.

Isabel Haasler and Pascal Frossard. Bures-wasserstein means of graphs. In *International Conference on Artificial Intelligence and Statistics*, pages 1873–1881. PMLR, 2024.

Hermina Petric Maretic, Mireille El Gheche, Matthias Minder, Giovanni Chierchia, and Pascal Frossard. Wasserstein-based graph alignment. *IEEE Transactions on Signal and Information Processing over Networks*, 8:353–363, 2022a.

Hermina Petric Maretic, Mireille El Gheche, Giovanni Chierchia, and Pascal Frossard. Fgot: Graph distances based on filters and optimal transport. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 7710–7718, 2022b.

Jérémie Bigot, Thierry Klein, et al. Consistent estimation of a population barycenter in the wasserstein space. *ArXiv e-prints*, 49, 2012.

Thibaut Le Gouic and Jean-Michel Loubes. Existence and consistency of wasserstein barycenters. *Probability Theory and Related Fields*, 168:901–917, 2017.

Elsa Cazelles, Jérémie Bigot, and Nicolas Papadakis. Regularized barycenters in the wasserstein space. In *International Conference on Geometric Science of Information*, pages 83–90. Springer, 2017.

Adil Ahidar-Coutrix, Thibaut Le Gouic, and Quentin Paris. Convergence rates for empirical barycenters in metric spaces: curvature, convexity and extendable geodesics. *Probability theory and related fields*, 177(1):323–368, 2020.

Victor-Emmanuel Brunel and Jordan Serres. Concentration of empirical barycenters in metric spaces. In *International Conference on Algorithmic Learning Theory*, pages 337–361. PMLR, 2024.

Thibaut Le Gouic, Quentin Paris, Philippe Rigollet, and Austin J Stromme. Fast convergence of empirical barycenters in alexandrov spaces and the wasserstein space. *Journal of the European Mathematical Society*, 25(6):2229–2250, 2022.

Adam Quinn Jaffe and Leonardo V Santoro. Large deviations principle for bures-wasserstein barycenters. *arXiv preprint arXiv:2409.11384*, 2024.

Guillaume Carlier, Katharina Eichinger, and Alexey Kroshnin. Entropic-wasserstein barycenters: Pde characterization, regularity, and clt. *SIAM Journal on Mathematical Analysis*, 53(5):5880–5914, 2021.

Sara Van De Geer. *Empirical Processes in M-estimation*. Cambridge UP, 2006.

Bradley Efron. Bootstrap Methods: Another Look at the Jackknife. *The Annals of Statistics*, 7(1):1 – 26, 1979.

Vladimir Spokoiny and Mayya Zhilova. Bootstrap confidence sets under model misspecification. *The Annals of Statistics*, 43(6):2653 – 2675, 2015.

Xi Chen and Wen-Xin Zhou. Robust inference via multiplier bootstrap. *The Annals of Statistics*, 48(3): 1665–1691, 2020.

Alexey Naumov, Vladimir Spokoiny, and Vladimir Ulyanov. Bootstrap confidence sets for spectral projectors of sample covariance. *Probability Theory and Related Fields*, 174(3):1091–1132, 2019.

Guang Cheng and Jianhua Z Huang. Bootstrap consistency for general semiparametric m-estimation. *The Annals of Statistics*, 38(5):2884–2915, 2010.

Stephen MS Lee and Puyudi Yang. Bootstrap confidence regions based on m-estimators under nonstandard conditions. *The Annals of Statistics*, 48(1):274–299, 2020.

Enno Mammen and Swagata Nandi. Bootstrap and resampling. In *Handbook of computational statistics*, pages 499–527. Springer, 2012.

Aad W Van Der Vaart, Jon A Wellner, Aad W van der Vaart, and Jon A Wellner. *Weak convergence*. Springer, 1996.

Nicolas Courty, Rémi Flamary, Devis Tuia, and Alain Rakotomamonjy. Optimal transport for domain adaptation. *IEEE transactions on pattern analysis and machine intelligence*, 39(9):1853–1865, 2016.

Rafał Bistroń, Michał Eckstein, and Karol Życzkowski. Monotonicity of the quantum 2-wasserstein distance. *arXiv preprint arXiv:2204.07405*, 2022.

Asuka Takatsu et al. Wasserstein geometry of Gaussian measures. *Osaka Journal of Mathematics*, 48 (4):1005–1026, 2011.

Donald Bures. An extension of kakutani's theorem on infinite product measures to the tensor product of semifinite w*-algebras. *Transactions of the American Mathematical Society*, 135:199–212, 1969.

Martial Agueh and Guillaume Carlier. Barycenters in the Wasserstein space. *SIAM Journal on Mathematical Analysis*, 43(2):904–924, 2011.

Pedro C Álvarez-Esteban, E Del Barrio, JA Cuesta-Albertos, and C Matrán. A fixed-point approach to barycenters in wasserstein space. *Journal of Mathematical Analysis and Applications*, 441(2): 744–762, 2016.

Sinho Chewi, Tyler Maunu, Philippe Rigollet, and Austin J Stromme. Gradient descent algorithms for Bures–Wasserstein barycenters. *arXiv preprint arXiv:2001.01700*, 2020.

Friedrich Götze, Alexey Naumov, Vladimir Spokoiny, and Vladimir Ulyanov. Large ball probabilities, gaussian comparison and anti-concentration. *Bernoulli*, 25(4A):2538–2563, 2019.

Ed Bullmore and Olaf Sporns. Complex brain networks: graph theoretical analysis of structural and functional systems. *Nature reviews neuroscience*, 10(3):186–198, 2009.

Alex Fornito, Andrew Zalesky, and Edward Bullmore. *Fundamentals of brain network analysis*. Academic Press, 2016a.

Alex Fornito, Andrew Zalesky, and Edward T Bullmore. Chapter 3-connectivity matrices and brain graphs. *Fundamentals of brain network analysis*, pages 89–113, 2016b.

Joshua Faskowitz, Xiaoran Yan, Xi-Nian Zuo, and Olaf Sporns. Weighted stochastic block models of the human connectome across the life span. *Scientific reports*, 8(1):1–16, 2018.

Gerwin Schalk, Dennis J McFarland, Thilo Hinterberger, Niels Birbaumer, and Jonathan R Wolpaw. Bci2000: a general-purpose brain-computer interface (bci) system. *IEEE Transactions on biomedical engineering*, 51(6):1034–1043, 2004.

Alexey Kroshnin and Alexandra Suvorikova. Bernstein-type and bennett-type inequalities for unbounded matrix martingales. *arXiv preprint arXiv:2411.07878*, 2024.

V. Bentkus. On the dependence of the Berry–Esseen bound on dimension. *Journal of Statistical Planning and Inference*, 113(2):385–402, 2003. ISSN 03783758.

Vidmantas Bentkus. A lyapunov-type bound in rd. *Theory of Probability & Its Applications*, 49(2): 311–323, 2005.

Joel Tropp. User-friendly tail bounds for sums of random matrices. *Foundations of computational mathematics*, 12:389–434, 2012.

Appendix A Approximation bounds in the Bures–Wasserstein space

This Section relies on the results from Kroshnin et al. [2021]. For the sake of completeness, we provide Lemma A.3 from there.

Statement A.1. *For any $S \in \mathbb{H}_+(d)$, $Q \in \mathbb{H}_{++}(d)$, the properties of operator $\boldsymbol{dT}_Q^S$ are following:*

(I) it is self-adjoint;

(II) it is negative semi-definite;

(III) it enjoys the following bounds:

$$
\begin{aligned}
-\left\langle \boldsymbol{dT}_Q^S(X), X \right\rangle &\le \frac{\lambda_{\max}^{1/2}\left(S^{1/2}QS^{1/2}\right)}{2}\|Q^{-1/2}XQ^{-1/2}\|_{\mathrm{F}}^2,\\
-\left\langle \boldsymbol{dT}_Q^S(X), X \right\rangle &\ge \frac{\lambda_{\min}^{1/2}\left(S^{1/2}QS^{1/2}\right)}{2}\|Q^{-1/2}XQ^{-1/2}\|_{\mathrm{F}}^2;
\end{aligned}
$$

(IV) it is homogeneous w.r.t. Q with degree $-\frac{3}{2}$ and w.r.t. S with degree $\frac{1}{2}$, i.e. $\boldsymbol{dT}_{aQ}^S = a^{-3/2}\boldsymbol{dT}_Q^S$ and $\boldsymbol{dT}_Q^{aS} = a^{1/2}\boldsymbol{dT}_Q^S$ for any $a > 0$;

(V) it is monotone w.r.t. $S^{1/2}QS^{1/2}$ (once range S is fixed): $\boldsymbol{dT}_{Q_0}^{S_0} \preccurlyeq \boldsymbol{dT}_{Q_1}^{S_1}$ in the sense of self-adjoint operators on $\mathbb{H}(d)$ whenever $S_0^{1/2}Q_0S_0^{1/2} \preccurlyeq S_1^{1/2}Q_1S_1^{1/2}$ and $\mathrm{range}(S_0) = \mathrm{range}(S_1)$; in particular, $\boldsymbol{dT}_Q^S$ is monotone w.r.t. $Q \in \mathbb{H}_{++}(d)$ for fixed S.

Let $Q \in \mathbb{H}_{++}(d)$ and define

$$
\boldsymbol{A}_Q \overset{\mathrm{def}}{=} \left(-\frac{1}{2}\boldsymbol{dT}_Q^Q\right)^{1/2}.
$$

Lemma A.3 by Kroshnin et al. [2021] ensures its existence.

Lemma A.2 (Properties of of $\boldsymbol{A}_Q$)**.** *The following equalities hold*

$$
\|\boldsymbol{A}_Q\| = \frac{1}{2\sqrt{\lambda_{\min}(Q)}}, \quad \left\|\boldsymbol{A}_Q^{-1}\right\| = 2\sqrt{\lambda_{\max}(Q)}. \tag{A.1}
$$

Moreover, let $U(\mathbb{H}(d))$ be the set of unitary operators on $\mathbb{H}(d)$. There exists a unitary operator $\boldsymbol{U}_Q \in U(\mathbb{H}(d))$ s.t. for any $X \in \mathbb{H}(d)$ holds

$$
(\boldsymbol{U}_Q\boldsymbol{A}_Q)X = Q^{1/2}\boldsymbol{dT}_Q^Q(X). \tag{A.2}
$$

Proof. First we prove (A.2). Without loss of generality, let Q be a diagonal matrix, i.e. $Q = \mathrm{diag}(q_1, \dots, q_d)$. It is enough to consider diagonal Q, because for any unitary U

$$
T_{UQU^*}^{USU^*} = UT_Q^SU^*.
$$

Moreover, $X \in \mathbb{H}(d)$ is a matrix as well: $X = (X_{ij})$ with $i, j \in \{1, \dots, d\}$.

Using the explicit expression for $\boldsymbol{dT}_Q^Q(X)$ (see formula (A.2) by Kroshnin et al. [2021]), we get

$$
\begin{aligned}
-\left\langle \boldsymbol{dT}_Q^Q(X), X \right\rangle &= \sum_{i,j=1}^{d} \frac{X_{ij}}{q_i + q_j} X_{ij} = \sum_{i,j=1}^{d} (q_i + q_j) \left(\frac{X_{ij}}{q_i + q_j} \right)^2 \\
&= 2 \sum_{i,j=1}^{d} \left(\sqrt{q_i} \frac{X_{ij}}{q_i + q_j} \right)^2 = 2 \left\| Q^{1/2} \boldsymbol{dT}_Q^Q(X) \right\|_{\mathrm{F}}^2 .
\end{aligned}
$$

Then $\|\boldsymbol{A}_Q(X)\|_{\mathrm{F}} = \left\| Q^{1/2} \boldsymbol{dT}_Q^Q(X) \right\|_{\mathrm{F}}$. Thus, these operators are unitary equivalent.

Now we prove (A.1). The above chain of equations ensures

$$
\|\boldsymbol{A}_Q(X)\|_{\mathrm{F}}^2 = \frac{1}{2} \sum_{i,j=1}^{d} \frac{X_{ij}^2}{q_i + q_j} .
$$

This yields

$$
\frac{1}{4\lambda_{\max}(Q)} \|X\|_{\mathrm{F}}^2 \le \|\boldsymbol{A}_Q(X)\|_{\mathrm{F}}^2 \le \frac{1}{4\lambda_{\min}(Q)} \|X\|_{\mathrm{F}}^2 .
$$

One can show in the same way as in the proof of Corollary A.2 by Kroshnin et al. [2021] that these inequalities are sharp. The result follows immediately. □

Lemma A.3 (Local Lipschitz continuity of $\boldsymbol{A}_Q$)**.** *Let* $B, Q \in \mathbb{H}_{++}(d)$*. If* $r(B,Q) \le 1/2$*,*

$$
\|\boldsymbol{A}_B - \boldsymbol{A}_Q\| \le r(B,Q) \cdot \|\boldsymbol{A}_B\| .
$$

Proof. Let $Q' = B^{-1/2} Q B^{-1/2}$. Lemma A.1 ensures that the mapping $Q \mapsto \boldsymbol{dT}_Q^Q$ is monotone and (-1)-homogeneous. Then $Q \mapsto \boldsymbol{A}_Q$ is antimonotone and $(-\frac{1}{2})$-homogeneous. This entails

$$
\begin{aligned}
\left(1 - \frac{1}{2} r(B,Q)\right) \boldsymbol{A}_B \preccurlyeq \frac{1}{\sqrt{\lambda_{\max}(Q')}} \boldsymbol{A}_B \preccurlyeq \boldsymbol{A}_Q, \\
\boldsymbol{A}_Q \preccurlyeq \frac{1}{\sqrt{\lambda_{\min}(Q')}} \boldsymbol{A}_B \preccurlyeq (1 + r(B,Q))\, \boldsymbol{A}_B .
\end{aligned}
$$

This yields the result. □

Now we are ready to prove Lemma 2.1.

Proof of Lemma 2.1. For simplicity, we denote

$$
S'_Q := Q^{-1/2} S Q^{-1/2}, \quad Q'_B := B^{-1/2} Q B^{-1/2}, \quad S'_B := B^{-1/2} S B^{-1/2} .
$$

Lemma A.6 by Kroshnin et al. [2021] ensures

$$
\begin{aligned}
-\frac{2}{\left(1 + \lambda_{\max}^{1/2}(S'_Q)\right)^2} \left\langle \boldsymbol{dT}_Q^Q(S - Q), S - Q \right\rangle &\le \mathcal{W}^2(S,Q) \\
&\le -\frac{2}{\left(1 + \lambda_{\min}^{1/2}(S'_Q)\right)^2} \left\langle \boldsymbol{dT}_Q^Q(S - Q), S - Q \right\rangle .
\end{aligned}
$$

Due to the monotonicity and homogeneity of the operator $\boldsymbol{dT}_Q^S$ (see (IV) and (V) in Lemma A.1), it holds that

$$\boldsymbol{dT}_Q^Q \preccurlyeq \boldsymbol{dT}_{\lambda_{\max}(Q'_B)B}^{\lambda_{\max}(Q'_B)B} = \frac{1}{\lambda_{\max}(Q'_B)} \boldsymbol{dT}_B^B,$$
$$\boldsymbol{dT}_Q^Q \succcurlyeq \boldsymbol{dT}_{\lambda_{\min}(Q'_B)B}^{\lambda_{\min}(Q'_B)B} = \frac{1}{\lambda_{\min}(Q'_B)} \boldsymbol{dT}_B^B.$$

Combining these inequalities with (A.2), we get

$$\begin{aligned}\frac{4\|\boldsymbol{A}_B(S-Q)\|_F^2}{\lambda_{\max}(Q'_B)\left(1+\lambda_{\max}^{1/2}(S'_Q)\right)^2} &\le \mathcal{W}^2(S,Q) \\ &\le \frac{4\|\boldsymbol{A}_B(S-Q)\|_F^2}{\lambda_{\min}(Q'_B)\left(1+\lambda_{\min}^{1/2}(S'_Q)\right)^2}.\end{aligned} \tag{A.3}$$

The last step is to get the bounds on $\lambda_{\min}(Q'_B)$ and $\lambda_{\max}(Q'_B)$. Let

$$r_Q := r(B,Q), \quad r_S := r(B,S).$$

It holds

$$1 - r_Q \le \lambda_{\min}(Q'_B) \le \lambda_{\max}(Q'_B) \le 1 + r_Q.$$

Assumption $r_Q \le \frac{1}{2}$ yields

$$\lambda_{\max}^{-1/2}(Q'_B) \ge 1 - \frac{1}{2} r_Q, \quad \lambda_{\min}^{-1/2}(Q'_B) \le 1 + 2r_Q.$$

Further, assumptions $r_Q \le \frac{1}{2}$ and $r_S \le \frac{1}{2}$ yield

$$\lambda_{\min}(S'_Q) \ge \tfrac{\lambda_{\min}(S'_B)}{\lambda_{\max}(Q'_B)} \ge 1 - r_Q - r_S, \quad \lambda_{\max}(S'_Q) \le \tfrac{\lambda_{\max}(S'_B)}{\lambda_{\min}(Q'_B)} \le 1 + 2r_Q + 2r_S.$$

Then

$$\tfrac{2}{1+\lambda_{\max}^{1/2}(S'_Q)} \ge 1 - \frac{1}{2} r_Q - \frac{1}{2} r_S, \quad \tfrac{2}{1+\lambda_{\min}^{1/2}(S'_Q)} \le 1 + r_Q + r_S.$$

Thus, we obtain

$$2\lambda_{\max}^{-1/2}(Q'_B)\left(1+\lambda_{\max}^{1/2}(S'_Q)\right)^{-1} \ge 1 - r_Q - \frac{1}{2} r_S,$$
$$2\lambda_{\min}^{-1/2}(Q'_B)\left(1+\lambda_{\min}^{1/2}(S'_Q)\right)^{-1} \le 1 + 4r_Q + 2r_S.$$

Combining these inequalities with (A.3), we get the result. □

In the rest of this section, we will use the following notations,

$$r_B := r(B, B_\mu), \quad r_F := r(\boldsymbol{F}, \boldsymbol{F}_\mu), \quad r := r_B + r_F, \quad \rho := 2\sqrt{\kappa(\boldsymbol{F})}\, r.$$

The next lemma bounds the operator $\boldsymbol{D}_\mu$ defined in (2.7). This result is crucial for the proof of Lemma 2.2.

Lemma A.4 (Bounds on $\boldsymbol{D}_\mu$). *If $r \le \frac{1}{2}$, then*

$$\frac{1}{1+2r}\boldsymbol{F} \preccurlyeq \boldsymbol{D}_\mu \preccurlyeq \frac{1}{1-r}\boldsymbol{F}.$$

Proof. Let $B_t = (1-t)B + tB_\mu$. Lemma A.4 by Kroshnin et al. [2021] ensures

$$\frac{1}{1-r_B}\boldsymbol{dT}_B^S \preccurlyeq \int_0^1 \boldsymbol{dT}_{B_t}^S\, dt \preccurlyeq \frac{1}{1+\frac{3}{4}r_B}\boldsymbol{dT}_B^S.$$

Now recall the definition on the operator $\boldsymbol{F}_\mu$ (see (2.3)). Integrating the above inequality over $d\mu(S)$, we get

$$\frac{1}{1+\frac{3}{4}r_B}\boldsymbol{F}_\mu \preccurlyeq \boldsymbol{D}_\mu \preccurlyeq \frac{1}{1-r_B}\boldsymbol{F}_\mu.$$

Since $r_F = \|\boldsymbol{F}^{-1/2}\boldsymbol{F}_\mu\boldsymbol{F}^{-1/2} - I\|$, it holds

$$(1-r_F)\,\boldsymbol{F} \preccurlyeq \boldsymbol{F}_\mu \preccurlyeq (1+r_F)\,\boldsymbol{F}.$$

Combining these bounds, we obtain:

$$\frac{1}{1+2r}\boldsymbol{F} \preccurlyeq \frac{1-r_F}{1+\frac{3}{4}r_B}\boldsymbol{F} \preccurlyeq \boldsymbol{D}_\mu \preccurlyeq \frac{1+r_F}{1-r_B}\boldsymbol{F} \preccurlyeq \frac{1}{1-r}\boldsymbol{F}.$$

□

Proof of Lemma 2.2. To prove (2.6), we use Lemma 2.1 and set $Q = B$, $S = B_\mu$. This yields

$$\left|\frac{\mathcal{W}(B_\mu, B)}{\|\boldsymbol{A}B_\Delta\|_{\mathrm{F}}} - 1\right| \le 2r_B.$$

Combining the above line of reasoning with the triangle inequality, we get

$$\left|\frac{\|\boldsymbol{A}B_\Delta\|_{\mathrm{F}}}{\|\boldsymbol{A}\boldsymbol{F}^{-1}T_\mu\|_{\mathrm{F}}} - 1\right| \le \frac{\|\boldsymbol{A}\left(B_\Delta - \boldsymbol{F}^{-1}T_\mu\right)\|_{\mathrm{F}}}{\|\boldsymbol{A}\boldsymbol{F}^{-1}T_\mu\|_{\mathrm{F}}} \le \kappa(\boldsymbol{A})\frac{\|B_\Delta - \boldsymbol{F}^{-1}T_\mu\|_{\mathrm{F}}}{\|\boldsymbol{F}^{-1}T_\mu\|_{\mathrm{F}}} \overset{\text{by (2.5)}}{\le} \sqrt{\kappa(B)}\rho.$$

Note that the last inequality holds due to $\kappa(\boldsymbol{A}) = \sqrt{\kappa(B)}$ (see Lemma A.2). Combining the above bounds, we get

$$\begin{aligned}\left|\frac{\mathcal{W}(B_\mu, B)}{\|\boldsymbol{A}\boldsymbol{F}^{-1}T_\mu\|_{\mathrm{F}}} - 1\right| &\le 2r_B + (1+2r_B)\left|\frac{\|\boldsymbol{A}B_\Delta\|_{\mathrm{F}}}{\|\boldsymbol{A}\boldsymbol{F}^{-1}T_\mu\|_{\mathrm{F}}} - 1\right| \\ &\le 2r_B + 2\left|\frac{\|\boldsymbol{A}B_\Delta\|_{\mathrm{F}}}{\|\boldsymbol{A}\boldsymbol{F}^{-1}T_\mu\|_{\mathrm{F}}} - 1\right| \le 3\sqrt{\kappa(B)}\rho.\end{aligned}$$

The second and the third inequalities rely on $r \le \frac{1}{2}$ and $2r_B \le \rho$, respectively. □

Proof of Corollary 2.3. Claim (2.9) follows directly from (2.5). Next, we prove (2.10). For the moment we set

$$\varDelta := B_\mu - B_\nu - \boldsymbol{F}^{-1}\left(T_\mu - T_{\hat{\mu}}\right),$$

and

$$\hat{r}_B := r(B, B_{\hat{\mu}}), \quad \hat{r}_F := r(\boldsymbol{F}, \boldsymbol{F}_{\hat{\mu}}), \quad \hat{r} := \hat{r}_B + \hat{r}_{\mathrm{F}}, \quad \hat{\rho} := 2\sqrt{\kappa(\boldsymbol{F})}\, \hat{r}.$$

Lemma 2.1 combined with (2.9) yields

$$\begin{aligned}
&|\mathcal{W}(B_\mu, B_\nu) - \|\boldsymbol{A}\boldsymbol{F}^{-1}(T_\mu - T_{\hat{\mu}})\|_{\mathrm{F}}| \\
&\qquad \le (4r_B + 2\hat{r}_B)\,\|\boldsymbol{A}\,(B_\mu - B_\nu)\|_{\mathrm{F}} + \|\boldsymbol{A}\varDelta\|_{\mathrm{F}} \\
&\qquad = (4r_B + 2\hat{r}_B)\,\|\boldsymbol{A}\left(\varDelta + \boldsymbol{F}^{-1}(T_\mu - T_{\hat{\mu}})\right)\|_{\mathrm{F}} + \|\boldsymbol{A}\varDelta\|_{\mathrm{F}} \\
&\qquad \le (4r_B + 2\hat{r}_B)\,\|\boldsymbol{A}\boldsymbol{F}^{-1}(T_\mu - T_{\hat{\mu}})\|_{\mathrm{F}} + (1 + 4r_B + 2\hat{r}_B)\,\|\boldsymbol{A}\varDelta\|_{\mathrm{F}} \\
&\qquad \overset{\text{by } (2.9)}{\le} \ (4r_B + 2\hat{r}_B + \hat{\rho}\,(1 + 4r_B + 2\hat{r}_B))\,\|\boldsymbol{A}\|\|\boldsymbol{F}^{-1}(T_\mu - T_{\hat{\mu}})\|_{\mathrm{F}} \\
&\qquad + (1 + 4r_B + 2\hat{r}_B)\,(\rho + \hat{\rho}\,)\,\|\boldsymbol{A}\|\|\boldsymbol{F}^{-1}T_\mu\|_F \\
&\qquad \le 6\kappa(\boldsymbol{A})(\hat{\rho} + \rho)\|\boldsymbol{A}\boldsymbol{F}^{-1}(T_\mu - T_{\hat{\mu}})\| + 4(\hat{\rho} + \rho)\|\boldsymbol{A}\|\|\boldsymbol{F}^{-1}T_\mu\|.
\end{aligned}$$

□

Appendix B Proofs of Gaussian approximations

The first result in this section investigates the properties of $\varkappa(\cdot)$ introduced by (3.1).

Lemma B.1 (Bounds on $\varkappa(\cdot)$). *Let $\boldsymbol{\varPsi}$ and $\boldsymbol{\varPhi}$ be symmetric operators, s.t. $\|\boldsymbol{\varPhi} - \boldsymbol{\varPsi}\|_1 \le \frac{\varLambda_2^2(\boldsymbol{\varPsi})}{4\|\boldsymbol{\varPsi}\|}$, with $\|\cdot\|_1$ be 1-Schatten norm. Then the following bounds hold,*

$$\varkappa(\boldsymbol{\varPhi}) \le 2\varkappa(\boldsymbol{\varPsi}), \quad \operatorname{tr}\boldsymbol{\varPhi} \le \tfrac{5}{4}\operatorname{tr}\boldsymbol{\varPsi}.$$

Proof. Note, that $\varLambda_2^2(\boldsymbol{\varPsi}) \le \varLambda_1^2(\boldsymbol{\varPsi}) \le \|\boldsymbol{\varPsi}\|\operatorname{tr}(\boldsymbol{\varPsi})$ and therefore,

$$\operatorname{tr}(\boldsymbol{\varPhi}) \le \operatorname{tr}(\boldsymbol{\varPsi}) + \|\boldsymbol{\varPhi} - \boldsymbol{\varPsi}\|_1 \le \frac{5}{4}\operatorname{tr}(\boldsymbol{\varPsi}).$$

By the definition of $\varLambda_r^2(\cdot)$, $\varLambda_r^2(\boldsymbol{\varPhi}) \ge \varLambda_r^2(\boldsymbol{\varPsi}) - \|\boldsymbol{\varPsi}\|\|\boldsymbol{\varPhi} - \boldsymbol{\varPsi}\|_1$, then

$$\varLambda_1^2(\boldsymbol{\varPhi})\,\varLambda_2^2(\boldsymbol{\varPhi}) \ge \varLambda_1^2(\boldsymbol{\varPsi})\,\varLambda_2^2(\boldsymbol{\varPsi}) - \left(\varLambda_1^2(\boldsymbol{\varPsi}) + \varLambda_2^2(\boldsymbol{\varPsi})\right)\|\boldsymbol{\varPsi}\|\|\boldsymbol{\varPhi} - \boldsymbol{\varPsi}\|_1.$$

Then it follows that

$$\begin{aligned}
\varkappa(\boldsymbol{\varPhi}) &\le \varkappa(\boldsymbol{\varPsi})\left(1 - \tfrac{\varLambda_1^2(\boldsymbol{\varPsi}) + \varLambda_2^2(\boldsymbol{\varPsi})}{\varLambda_1^2(\boldsymbol{\varPsi})\varLambda_2^2(\boldsymbol{\varPsi})}\|\boldsymbol{\varPsi}\|\|\boldsymbol{\varPhi} - \boldsymbol{\varPsi}\|_1\right)^{-1} \\
&\le \varkappa(\boldsymbol{\varPsi})\left(1 - 2\tfrac{\|\boldsymbol{\varPsi}\|}{\varLambda_2^2(\boldsymbol{\varPsi})}\|\boldsymbol{\varPhi} - \boldsymbol{\varPsi}\|_1\right)^{-1} \le 2\varkappa(\boldsymbol{\varPsi}).
\end{aligned}$$

□

Lemma B.2 (Anti-concentration). *Let $G \sim \mathcal{N}(0, \boldsymbol{K})$ be a Gaussian vector taking values in some Hilbert space H. Then for any $\varepsilon, x \ge 0$ the following anti-concentration bound holds:*

$$\mathbb{P}\left\{x \le \|G\|_H \le x + \varepsilon\right\} \le \mathtt{C}\gamma(\boldsymbol{K})\frac{\varepsilon}{\sqrt{\operatorname{tr}(\boldsymbol{K})}}.$$

Proof. For any $x, h, \varepsilon > 0$ it holds that

$$(x+\varepsilon)^2 \le \begin{cases} x^2\left(1+\frac{\varepsilon}{h}\right)^2, & h \le x, \\ x^2 + 2h\varepsilon + \varepsilon^2, & h > x. \end{cases}$$

Thus, the union bound and Theorem 2.7 by Götze et al. [2019] yield

$$\begin{aligned} \mathbb{P}\left\{x \le \|G\|_H \le x+\varepsilon\right\} \le& \mathbb{P}\left\{x^2 \le \|G\|_H^2 \le x^2 + 2h\varepsilon + \varepsilon^2\right\} + \mathbb{P}\left\{x \le \|G\|_H \le x\left(1+\frac{\varepsilon}{h}\right)\right\} \\ \le& \mathtt{C}\varkappa(\boldsymbol{K})\left(h\varepsilon + \varepsilon^2 + \frac{\varepsilon}{h}\operatorname{tr}(\boldsymbol{K})\right) \le \mathtt{C}\varkappa(\boldsymbol{K})\left(\varepsilon\sqrt{\operatorname{tr}(\boldsymbol{K})} + \varepsilon^2\right), \end{aligned}$$

where the last inequality is ensured by $h = \sqrt{\operatorname{tr}\boldsymbol{K}}$. The above inequality can be rewritten as

$$\mathbb{P}\left\{x \le \|G\|_H \le x+\varepsilon\right\} \le \mathtt{C}\gamma(\boldsymbol{K})\left(\frac{\varepsilon}{\sqrt{\operatorname{tr}(\boldsymbol{K})}} + \frac{\varepsilon^2}{\operatorname{tr}(\boldsymbol{K})}\right).$$

Since $\gamma(\boldsymbol{K}) \ge 1$ and the probability on the l.h.s. is bounded by 1, it is enough to consider the case $\varepsilon \le \sqrt{\operatorname{tr}(\boldsymbol{K})}$. Thus, we obtain

$$\mathbb{P}\left\{x \le \|G\|_H \le x+\varepsilon\right\} \le \mathtt{C}\gamma(\boldsymbol{K})\frac{\varepsilon}{\sqrt{\operatorname{tr}(\boldsymbol{K})}}.$$

□

B.1 GAR for bootstrap validity

Before getting the main results, we write down some trivial but useful bounds.

Lemma B.3. *Let Assumptions* (T) *and* (F) *be fulfilled. Then with probability at least* $1 - \mathtt{C}e^{-\mathrm{x}}$ *it holds*

$$r(B, B_\mu) \le c_B\varepsilon_T(\mathrm{x}) \quad r(\boldsymbol{F}, \boldsymbol{F}_\mu) \le \varepsilon_F(\mathrm{x}), \quad \rho \le \varepsilon(\mathrm{x}). \tag{B.1}$$

If Assumptions $(\hat{T})$ *and* $(\hat{F})$ *hold as well, then, conditioned on any* $\mu \in \mathcal{A}_t$*, it holds with probability* $1 - \mathtt{C}e^{-\mathrm{x}}$*,*

$$r(B, B_{\hat{\mu}}) \le c_B\hat{\varepsilon}_T(\mathrm{x}, \mathrm{t}), \quad r(\boldsymbol{F}, \boldsymbol{F}_{\hat{\mu}}) \le \hat{\varepsilon}_{\mathrm{F}}(\mathrm{x}, \mathrm{t}), \quad \hat{\rho} \le \hat{\varepsilon}(\mathrm{x}, \mathrm{t}). \tag{B.2}$$

Proof. The proof is trivial and follows from Lemma 2.4. □

Proof of Lemma 3.2. We set

$$X = \mathcal{W}(B_\mu, B), \ \ Y = \left\|\boldsymbol{A}\boldsymbol{F}^{-1}T_\mu\right\|_{\mathrm{F}}, \ \ G = \boldsymbol{A}Z.$$

Assumption (GAR-I) holds due to (2.6) and (B.1):

$$|X - Y| \le 3\sqrt{\kappa(B)}\rho Y \le 3\sqrt{\kappa(B)}\varepsilon(\mathrm{x})Y,$$

the last inequality holds with probability at least $1 - \mathtt{C}e^{-\mathrm{x}}$ for all x, s.t. $\varepsilon(\mathrm{x}) \le \frac{1}{6\sqrt{\kappa(B)}}$.

Assumption (GAR-II) is fulfilled due to Assumption (G). The result follows immediately from Lemma 3.1.

□

Proof of Lemma 3.3. We have to check Assumptions (GAR-I) and (GAR-II). We set

$$X = \mathcal{W}(B_{\hat{\mu}}, B_{\mu}), \;\; Y = \left\| \boldsymbol{A}\boldsymbol{F}^{-1} \left(T_{\hat{\mu}} - T_{\mu}\right) \right\|_{\mathrm{F}}, \;\; G = \boldsymbol{A} Z_{\mu}.$$

Assumption (GAR-I) is valid due to Corollary 2.3 and assumptions $r \le \frac{1}{2}, \hat{r} \le \frac{1}{2}$,

$$|X - Y| \le 6\kappa(\boldsymbol{A}) \left(\hat{\rho} + \rho\right) Y + 4 \left(\hat{\rho} + \rho\right) \|\boldsymbol{A}\| \|\boldsymbol{F}^{-1} T_{\mu}\|_{\mathrm{F}}.$$

Note that Lemma (A.2) ensures $\kappa(\boldsymbol{A}) = \sqrt{\kappa(B)}$. Using Lemma B.3, we get

$$|X - Y| \le 6\sqrt{\kappa(B)} \left(\rho + \hat{\varepsilon}(\mathrm{x}, \mathrm{t})\right) Y + 4 \left(\rho + \hat{\varepsilon}(\mathrm{x}, \mathrm{t})\right) \|\boldsymbol{A}\| \|\boldsymbol{F}^{-1} T_{\mu}\|_{\mathrm{F}}.$$

The inequality holds with probability at least $1 - \mathtt{C} e^{-\mathrm{x}}$ for all x s.t. $\hat{\varepsilon}(\mathrm{x}, \mathrm{t}) + \rho \le \frac{1}{6\sqrt{\kappa(B)}}$. Since by assumption of the lemma $\rho \le \frac{1}{12\sqrt{\kappa(B)}}$, we get $\hat{\varepsilon}(\mathrm{x}, \mathrm{t}) \le \frac{1}{12\sqrt{\kappa(B)}}$.

Assumption (GAR-II) is valid due to Assumption ($\hat{G}$) with $\varDelta = \hat{\varepsilon}_G(\mathrm{t})$. The claim follows.

□

Appendix C Generalized bootstrap validity

Proof of Theorem 4.1. Lemma 3.2 and 3.3 ensure that for all $z \ge 0$ with probability at least $1 - \mathtt{C} e^{-\mathrm{t}}$, it holds

$$\left|\mathbb{P}\left\{\mathcal{W}(B_{\mu}, B) \le z\right\} - \mathbb{P}\left\{\|\boldsymbol{A} Z\|_{\mathrm{F}} \le z\right\}\right| \le \mathcal{E},$$
$$\left|\mathbb{P}\left\{\mathcal{W}(B_{\hat{\mu}}, B_{\mu}) \le z \mid \mu\right\} - \mathbb{P}\left\{\|\boldsymbol{A} Z_{\mu}\|_{\mathrm{F}} \le z \mid \mu\right\}\right| \le \hat{\mathcal{E}}(\mathrm{t}).$$

This yields

$$\begin{aligned} &\left|\mathbb{P}\left\{\mathcal{W}(B_{\mu}, B) \le z\right\} - \mathbb{P}\left\{\mathcal{W}(B_{\hat{\mu}}, B_{\mu}) \le z \mid \mu\right\}\right| \\ &\le \left|\mathbb{P}\left\{\|\boldsymbol{A} Z\|_{\mathrm{F}} \le z\right\} - \mathbb{P}\left\{\|\boldsymbol{A} Z_{\mu}\|_{\mathrm{F}} \le z \mid \mu\right\}\right| + \mathcal{E} + \hat{\mathcal{E}}(\mathrm{t}). \end{aligned} \tag{C.1}$$

First, we consider $\hat{\mathcal{E}}(\mathrm{t})$ coming from Lemma 3.3,

$$\hat{\mathcal{E}}(\mathrm{t}) = \hat{\varepsilon}_G(\mathrm{t}) + \mathtt{C} \cdot \inf_{\mathrm{x} \in \hat{\mathfrak{X}}(\mathrm{t})} \left\{ e^{-\mathrm{x}} + \sqrt{\kappa(B)} \gamma(\Xi'_{\mu}) \left(\rho + \hat{\varepsilon}(\mathrm{x}, \mathrm{t})\right) \left(\frac{\|\boldsymbol{A}\| \|\boldsymbol{F}^{-1} T_{\mu}\|_{\mathrm{F}}}{\sqrt{\mathrm{tr}(\Xi'_{\mu})}} + 1 \right) \right\}.$$

Lemma B.3 and Assumption (T) ensure with probability at least $1 - \mathtt{C} e^{-\mathrm{x}'}$, that

$$\rho \le \varepsilon(\mathrm{x}'), \quad \|\boldsymbol{F}^{-1} T_{\mu}\|_{\mathrm{F}} \le \|\boldsymbol{F}^{-1}\| \varepsilon_T(\mathrm{x}').$$

Further, condition (4.1) and Lemma B.1 ensure

$$\varkappa(\Xi'_{\mu}) \le 2\varkappa\left(\Xi'\right), \quad \mathrm{tr}(\Xi'_{\mu}) \le \frac{5}{4} \mathrm{tr}\left(\Xi'\right). \tag{C.2}$$

Taking into account the definition of $\gamma(\cdot)$ (3.1), we get that with probability at least $1 - e^{-\mathrm{x}'}$,

$$\hat{\mathcal{E}}(\mathrm{t}) \le \hat{\varepsilon}_G(\mathrm{t}) + \mathtt{C} \cdot \inf_{\mathrm{x} \in \hat{\mathfrak{X}}(\mathrm{t})} \left\{ e^{-\mathrm{x}} + \sqrt{\kappa(B)} \gamma(\Xi') \left(\varepsilon(\mathrm{x}') + \hat{\varepsilon}(\mathrm{x}, \mathrm{t})\right) \left(\frac{\|\boldsymbol{A}\| \|\boldsymbol{F}^{-1}\|}{\sqrt{\mathrm{tr}(\Xi')}} \varepsilon_T(\mathrm{x}') + 1 \right) \right\}.$$

Next, we have to bound

$$\left|\mathbb{P}\left\{\|\boldsymbol{A}Z\|_{\mathrm{F}} \le z\right\} - \mathbb{P}\left\{\|\boldsymbol{A}Z_\mu\|_{\mathrm{F}} \le z \mid \mu\right\}\right|.$$

Recall that $\boldsymbol{A}$ is self-adjoint. Corollary 2.3 by Götze et al. [2019] ensures,

$$\begin{aligned}\sup_{z\ge 0}&\left|\mathbb{P}\left\{\|\boldsymbol{A}Z\|_{\mathrm{F}} \le z\right\} - \mathbb{P}\left\{\|\boldsymbol{A}Z_\mu\|_{\mathrm{F}} \le z \mid \mu\right\}\right| \\ &\le \mathtt{C}\left(\varkappa\left(\Xi'\right) + \varkappa\left(\Xi'_\mu\right)\right)\left\|\Xi' - \Xi'_\mu\right\|_1.\end{aligned} \tag{C.3}$$

Taking into account (C.2) and Assumption (Ξ), we get with probability at least $1 - \mathtt{C}e^{-\mathrm{y}}$

$$\left\|\Xi' - \Xi'_\mu\right\|_1 \le \|\boldsymbol{A}\|^2 \|\Xi - \Xi_\mu\|_1 \le \|\boldsymbol{A}\|^2 \varepsilon_\Xi(\mathrm{y}).$$

Combining these bounds with (C.1) and (C.3) and setting $\mathrm{y} = \mathrm{x}' = \mathrm{t}$, we get the result. □

Appendix D Multiplier bootstrap validity

Proof of Lemma 5.3. First, Assumption (P) ensures,

$$\left\|\|S\|^{1/2}\right\|_{\psi_2} \le \left\|\sqrt{\operatorname{tr} S}\right\|_{\psi_2} < +\infty. \tag{D.1}$$

Now we recall that $T_B^S = B^{-1/2}\left(B^{1/2}SB^{1/2}\right)^{1/2} B^{-1/2}$. Using (D.1), we get

$$\left\|T_B^S\right\| \le \frac{\lambda_{\max}^{1/2}(B)}{\lambda_{\min}(B)}\|S\|^{1/2}.$$

Thus, $\|\|T_B^S\|_{\mathrm{F}}\|_{\psi_2} \le d \cdot v_S$. Finally, we use the result (III) in Lemma A.1 that ensures

$$\left\|\boldsymbol{dT}_B^S\right\| \le \frac{\lambda_{\max}^{1/2}\left(S^{1/2}BS^{1/2}\right)}{2\lambda_{\min}^2(B)} \le \frac{\lambda_{\max}^{1/2}(B)}{2\lambda_{\min}^2(B)}\|S\|^{1/2}.$$

Combining this fact with (D.1), we get the result. □

Before validating the bootstrap assumptions, we prove two auxiliary lemmas. The first lemma deals with concentrations of sub-exponential r.v. The first two results are well-known and we provide them for the sake of completeness.

In the following, we will often use the auxiliary concentration results. For the sake of completeness, we provide them below. Let $(x)_+ = \max\{0, x\}$ and $\log(x) = \max\{1, \ln x\}$.

Statement D.1 (Theorem 2.1 [Kroshnin and Suvorikova, 2024]). *Fix $\alpha \ge 1$. Let $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n \in \mathbb{H}(d)$ be independent and $\mathbb{E}\,\boldsymbol{X}_i = 0$ for all i. Define*

$$K = \max_i \|\|\boldsymbol{X}_i\|\|_{\psi_\alpha}, \quad U^2 \stackrel{\text{def}}{=} \sum_i \|\|\boldsymbol{X}_i\|\|_{\psi_\alpha}^2, \quad \sigma^2 \stackrel{\text{def}}{=} \left\|\sum_i \mathbb{E}\,\boldsymbol{X}_i^2\right\|, \quad z \stackrel{\text{def}}{=} \left(\log\frac{U}{\sigma}\right)^{1/\alpha}.$$

Then, with probability at least $1 - 2de^{-\mathrm{x}}$,

$$\left\|\frac{1}{n}\sum_i \mathbf{X}_i\right\| \lesssim \sigma\sqrt{\frac{\mathrm{x}}{n}} + Kz\frac{\mathrm{x}}{n},$$

with $\|\cdot\|$ being the operator norm.

Statement D.2 (Corollary 3.5 [Kroshnin and Suvorikova, 2024]). *Fix $\alpha \geq 1$. Let $(\mathcal{H}, \|\cdot\|_H)$ be a separable Hilbert space and assume $X_1, \ldots, X_n \in \mathcal{H}$ are independent random variables s.t. $\mathbb{E}\, X_i = 0$. Define*

$$K \overset{\text{def}}{=} \max_i \| \|X_i\|_H \|_{\psi_\alpha}, \quad U^2 \overset{\text{def}}{=} \sum_i \| \|X_i\|_H \|_{\psi_\alpha}^2, \quad \sigma^2 \overset{\text{def}}{=} \sum_i \mathbb{E} \|X_i\|_H^2, \quad z \overset{\text{def}}{=} \left(\log \frac{U}{\sigma} \right)^{1/\alpha}$$

Then for $\mathrm{x} \geq 1$ *with probability at least* $1 - e^{-\mathrm{x}}$

$$\left\| \frac{1}{n} \sum_i X_i \right\|_H \lesssim \sigma \sqrt{\frac{\mathrm{x}}{n}} + K z \frac{\mathrm{x}}{n}.$$

Statement D.3. *Fix $\alpha > 0$. Let $X_1, \ldots, X_n \geq 0$ be i.i.d. random variables, s.t. $\sigma^2 = \mathbb{E}\, X_1^2$, $v = \|X_1\|_{\psi_\alpha}$. Let $z \overset{\text{def}}{=} \left(\log \frac{v}{\sigma}\right)^{1/\alpha}$. Then for any $p \geq 2$ and $\mathrm{x} \geq 0$ it holds with probability at least $1 - 2e^{-\mathrm{x}}$*

$$\frac{1}{n} \sum_i X_i^p \lesssim \sigma^2 (vz)^{p-2} + v^p \left(z^p + (\mathrm{x} + \log n)^{\frac{p}{\alpha} - 1} \right) \frac{\mathrm{x}}{n}.$$

Moreover, with probability at least $1 - e^{-\mathrm{x}}$

$$\max_i X_i \leq v(\mathrm{x} + \ln 2n)^{1/\alpha}.$$

Proof. Theorem 2.1 from [Kroshnin and Suvorikova, 2024] ensures that

$$\begin{aligned}
\frac{1}{n} \sum_i X_i^p &\lesssim \mathbb{E}\, X_i^p + \sqrt{\frac{\mathrm{x}}{n} \mathbb{E}\, X_i^{2p}} + v^p \left(\log \frac{v^{2p}}{\mathbb{E}\, X_i^{2p}} \right)^{\frac{p}{\alpha}} \frac{\mathrm{x}}{n} \\
&\quad + v^p (\mathrm{x} + \ln n)^{\frac{p}{\alpha} - 1} \frac{\mathrm{x}}{n} \\
&\lesssim \sigma^2 \left(v \left(\log \frac{v}{\sigma} \right)^{\frac{1}{\alpha}} \right)^{p-2} + \sigma \left(v \left(\log \frac{v}{\sigma} \right)^{\frac{1}{\alpha}} \right)^{p-2} \sqrt{\frac{\mathrm{x}}{n}} + v^p \left(\left(\log \frac{v}{\sigma} \right)^{\frac{p}{\alpha}} + (\mathrm{x} + \ln n)^{\frac{p}{\alpha} - 1} \right) \frac{\mathrm{x}}{n} \\
&\lesssim \sigma^2 (vz)^{p-2} + v^p \left(z^p + (\mathrm{x} + \ln n)^{\frac{p}{\alpha} - 1} \right) \frac{\mathrm{x}}{n}.
\end{aligned}$$

To get the second result, we use a well-known line of reasoning,

$$\mathbb{P} \left\{ \max_i X_i \geq \mathrm{t} \right\} = \mathbb{P} \left\{ \bigcup_i \{ X_i \geq \mathrm{t} \} \right\} \leq 2n e^{-(\mathrm{t}/v)^\alpha} = e^{\ln(2n) - (\mathrm{t}/v)^\alpha}.$$

□

Throughout the rest of the text, we denote

$$T_i \overset{\text{def}}{=} T_B^{S_i} - I.$$

We also write down explicitly all the terms. The $\mathcal{T}$-mappings are written as

$$T_\mu := \mathcal{T}(P_n) = \frac{1}{n} \sum_i T_i, \quad T_{\hat{\mu}} := \mathcal{T}(P_w) = \frac{1}{n} \sum_i w_i T_i,$$

and the $\mathcal{F}$-mappings are

$$\boldsymbol{F} = \mathcal{F}(P) = -\,\mathbb{E}\,\boldsymbol{dT}_B^S, \quad \boldsymbol{F}_\mu = \mathcal{F}(P_n) = \frac{1}{n}\sum_i \boldsymbol{dT}_B^{S_i}, \quad \boldsymbol{F}_{\hat\mu} = \mathcal{F}(P_w) = \frac{1}{n}\sum_i w_i \boldsymbol{dT}_B^{S_i}.$$

The vectors used for Gaussian approximation are $Z \sim \mathcal{N}(0, \Xi)$ and $Z_\mu \sim \mathcal{N}(0, \Xi_\mu)$, where

$$\Xi \overset{\text{def}}{=} \frac{1}{n}\boldsymbol{F}^{-1}\left[\mathbb{E}\left(T_B^S - I\right) \otimes \left(T_B^S - I\right)\right]\boldsymbol{F}^{-1},$$
$$\Xi_\mu \overset{\text{def}}{=} \frac{1}{n}\boldsymbol{F}^{-1}\left[\frac{1}{n}\sum_i \left(T_B^{S_i} - I\right) \otimes \left(T_B^{S_i} - I\right)\right]\boldsymbol{F}^{-1},$$

with $\otimes$ denoting the tensor product. Throughout this section, we denote

$$\sigma_T^2 \overset{\text{def}}{=} \mathbb{E}\,\|T_1\|_{\mathrm{F}}^2, \quad C_T \overset{\text{def}}{=} \frac{v_T^2}{\sigma_T^2}\log\frac{v_T}{\sigma_T}.$$

Lemma D.4 (Assumption (T))**.** *Assumption* (P) *ensures that for all* $\mathrm{x} \geq 1$ *and* $n \gtrsim C_T \mathrm{x}$*,*

$$\varepsilon_T(\mathrm{x}) \lesssim \sigma_T\sqrt{\frac{\mathrm{x}}{n}}.$$

Proof. Let $\overline{T} := \frac{1}{n}\sum_i T_i$. We apply Statement D.2 with $\alpha = 2$ and get with probability at least $1 - e^{-\mathrm{x}}$,

$$\|\overline{T}\|_{\mathrm{F}} \lesssim \sigma_T\sqrt{\frac{\mathrm{x}}{n}} + v_T\sqrt{\log\left(\frac{v_T}{\sigma_T}\right)\frac{\mathrm{x}}{n}}.$$

By substituting the condition on n, we get the result. □

Now, we set

$$C_w \overset{\text{def}}{=} (v_w \log v_w)^2.$$

Lemma D.5 (Assumption $(\hat{T})$)**.** *Assumptions* (W) *and* (P) *ensure that for all* $\mathrm{x}, \mathrm{t} \geq 1$

$$\hat{\varepsilon}_T(\mathrm{x}; \mathrm{t}) \lesssim \sigma_T\sqrt{\frac{\mathrm{x}}{n}}$$

whenever $n \gtrsim C_w C_T \mathrm{x}(\mathrm{t} + \log n)$*.*

Proof. First, we denote

$$\overline{T} := T_{\hat\mu} - T_\mu = \frac{1}{n}\sum_i (w_i - 1)T_i.$$

Note that $\overline{T}$ is centred in the bootstrap world, i.e. $\mathbb{E}_w \overline{T} = 0$. Further, $\overline{T}$ sub-Gaussian due to Assumption (W).

We apply Statement D.2 and get with probability at least $1 - e^{-\mathrm{x}}$

$$\|\overline{T}\|_{\mathrm{F}} \lesssim \sqrt{\frac{1}{n}\sum_i \|T_i\|_{\mathrm{F}}^2 \frac{\mathrm{x}}{n}} + \max_i \|\|(w_i - 1)T_i\|_{\mathrm{F}}\|_{\psi_1} z^2 \frac{\mathrm{x}}{n}, \tag{D.2}$$

with

$$z^2 = \log \sqrt{\frac{\sum_i \| \| (w_i - 1) T_i \|_{\mathrm{F}} \|_{\psi_1}^2}{\sum_i \mathbb{E}_w \| (w_i - 1) T_i \|_{\mathrm{F}}^2}} = \log v_w .$$

Thus,

$$\| \overline{T} \|_{\mathrm{F}} \lesssim \sqrt{\frac{1}{n} \sum_i \| T_i \|_{\mathrm{F}}^2 \frac{\mathrm{x}}{n}} + v_w \log v_w \max_i \| T_i \|_{\mathrm{F}} \frac{\mathrm{x}}{n} .$$

Now we apply Lemma D.3 with $\alpha = p = 2$ and get with probability at least $1 - 2e^{-\mathrm{t}}$

$$\frac{1}{n} \sum_{i=1}^{n} \| T_i \|_{\mathrm{F}}^2 \lesssim \sigma_T^2 + v_T^2 \log \left(\frac{v_T}{\sigma_T} \right) \frac{\mathrm{t}}{n} \lesssim \sigma_T^2 . \tag{D.3}$$

Moreover, $\max_i \| T_i \|_{\mathrm{F}} \lesssim v_T \sqrt{t + \log n}$.

Thus, one can take

$$\hat{\varepsilon}(\mathrm{x}; \mathrm{t}) \lesssim \sigma_T \sqrt{\frac{\mathrm{x}}{n}} + v_w \log v_w v_T \sqrt{t + \log n} \frac{\mathrm{x}}{n} \lesssim \sigma_T \sqrt{\frac{\mathrm{x}}{n}} .$$

□

Now, we define the covariance of T_i and its empirical counterpart,

$$\varSigma \overset{\mathrm{def}}{=} \mathbb{E}\, T_1 \otimes T_1, \quad \boldsymbol{\varSigma}_\mu = \frac{1}{n} \sum_i T_i \otimes T_i,$$

with $\otimes$ being the tensor product. And set

$$K_T \overset{\mathrm{def}}{=} \left\| \left\| \boldsymbol{\varSigma}^{-1/2} T_i \right\|_{\mathrm{F}} \right\|_{\psi_2} \le \| \boldsymbol{\varSigma}^{-\mathbf{1/2}} \| v_T , \quad C_G \overset{\mathrm{def}}{=} \left(\frac{K_T}{d} \right)^2 \log \frac{K_T}{d} .$$

Lemma D.6 (Assumption (G))**.** *Under Assumption* (P) *it holds that*

$$\varepsilon_G \lesssim d^3 \sqrt{\frac{C_G}{n}} .$$

Proof. The result follows from Theorem 1.1 by Bentkus [2003] applied to $X_i = \boldsymbol{\varSigma}^{-1/2} T_i$ for all $i = 1, \ldots, n$. Namely,

$$\varepsilon_G \lesssim \frac{1}{\sqrt{n}} \mathbb{E} \left\| \boldsymbol{\varSigma}^{-1/2} T_i \right\|_{\mathrm{F}}^3 \lesssim \frac{1}{\sqrt{n}} d^2 K_T \sqrt{\log \frac{K_T}{d}} ,$$

by Lemma B.5 from [Kroshnin and Suvorikova, 2024], since $\mathbb{E} \left\| \boldsymbol{\varSigma}^{-1/2} T_i \right\|_{\mathrm{F}}^2 = \dim \boldsymbol{\varSigma} \le d^2$. □

Lemma D.7 (Assumption $(\hat{G})$)**.** *Let Assumptions* (P) *and* (W) *be true. For sufficiently large* n*, s.t.*

$$n \gtrsim \max \left\{ (\mathrm{t} + \log d) K_T^2 \log K_T , (\mathrm{t} + \log d)^{3/2} \left(\frac{K_T}{d} \right)^2 \right\} . \tag{D.4}$$

it holds

$$\hat{\varepsilon}_G(\mathrm{t}) \lesssim d^3 \sqrt{\frac{C_G}{n}} .$$

Proof. We denote $X_i = \frac{w_i-1}{\sqrt{n}}\boldsymbol{\Sigma}_\mu^{-1/2}T_i$. According to Bentkus [2005], $\hat{\varepsilon}_G(\mathrm{t})$ can be bounded with $(1-\mathtt{C}e^{-\mathrm{t}})$-quantile of

$$\mathbb{E}_w \sum_{i=1}^{n} \|X_i\|_{\mathrm{F}}^3 = \frac{1}{\sqrt{n}}\frac{1}{n}\sum_{i=1}^{n} \mathbb{E}_w\,|w_i-1|^3 \big\|\boldsymbol{\Sigma}_{\boldsymbol{\mu}}^{-1/2}T_i\big\|_{\mathrm{F}}^3 \lesssim \frac{v_w \log v_w}{\sqrt{n}}\frac{1}{n}\sum_{i=1}^{n}\big\|\boldsymbol{\Sigma}_{\boldsymbol{\mu}}^{-1/2}T_i\big\|_{\mathrm{F}}^3.$$

The last inequality is true because $\mathbb{E}\,|w_i-1|^3 \lesssim v_w \log v_w$.

Now, our goal is to estimate $\lambda_{\max}(\boldsymbol{I} - \boldsymbol{\Sigma}^{-1/2}\boldsymbol{\Sigma}_\mu\boldsymbol{\Sigma}^{-1/2})$. We will apply Bernstein inequality to random matrices $\boldsymbol{I} - (\boldsymbol{\Sigma}^{-1/2}T_i)\otimes(\boldsymbol{\Sigma}^{-1/2}T_i)$. Notice that

$$\begin{aligned}\Big\|\mathbb{E}\left(\boldsymbol{I} - (\boldsymbol{\Sigma}^{-1/2}T_i)\otimes(\boldsymbol{\Sigma}^{-1/2}T_i)\right)^2\Big\| &= \Big\|\mathbb{E}\left((\boldsymbol{\Sigma}^{-1/2}T_i)\otimes(\boldsymbol{\Sigma}^{-1/2}T_i)\right)^2 - \boldsymbol{I}\Big\| \\ &\le \Big\|\mathbb{E}\left((\boldsymbol{\Sigma}^{-1/2}T_i)\otimes(\boldsymbol{\Sigma}^{-1/2}T_i)\right)^2\Big\|.\end{aligned}$$

For simplicity, set $Y_i = \boldsymbol{\Sigma}^{-1/2}T_i$. Let $\mathit{\Pi}_{Y_i}$ be the orthogonal projector onto $\operatorname{span}(Y_i)$, so that $Y_i \otimes Y_i = \|Y_i\|_{\mathrm{F}}^2 \mathit{\Pi}_{Y_i}$. Since $\mathbb{E}\|Y_i\|_{\mathrm{F}}^2\mathit{\Pi}_{Y_i} = \mathbb{E}\,Y_i\otimes Y_i = \boldsymbol{I}$, by Lemma B.5 in [Kroshnin and Suvorikova, 2024] we obtain

$$\begin{aligned}&\Big\|\mathbb{E}\left((\boldsymbol{\Sigma}^{-1/2}T_i)\otimes(\boldsymbol{\Sigma}^{-1/2}T_i)\right)^2\Big\| = \big\|\mathbb{E}(Y_i\otimes Y_i)^2\big\| \\ &= \big\|\mathbb{E}(\|Y_i\|_{\mathrm{F}}\mathit{\Pi}_{Y_i})^4\big\| \lesssim \|I\| K_T^2 \log\frac{K_T}{\|I\|} = K_T^2 \log K_T.\end{aligned}$$

Bernstein inequality (Theorem 1.4 from Tropp [2012]) yields, with probability at least $1-e^{-\mathrm{t}}$,

$$\lambda_{\max}(\boldsymbol{I} - \boldsymbol{\Sigma}^{-1/2}\boldsymbol{\Sigma}_\mu\boldsymbol{\Sigma}^{-1/2}) \lesssim K_T\sqrt{\frac{\mathrm{t}+\log d}{n}\log K_T} + \frac{\mathrm{t}+\log d}{n}.$$

Condition (D.4) ensures that

$$\lambda_{\max}(\boldsymbol{I} - \boldsymbol{\Sigma}^{-1/2}\boldsymbol{\Sigma}_\mu\boldsymbol{\Sigma}^{-1/2}) \le \frac{1}{2}, \tag{D.5}$$

thus $\|\boldsymbol{\Sigma}_\mu^{-1}\| \le 2\|\boldsymbol{\Sigma}^{-1}\|$.

Finally, we have to estimate $\frac{1}{n}\sum_{i=1}^n \big\|\boldsymbol{\Sigma}^{-1/2}T_i\big\|_{\mathrm{F}}^3$. Note that $\mathbb{E}\big\|\boldsymbol{\Sigma}^{-1/2}T_i\big\|_{\mathrm{F}}^2 = \dim\boldsymbol{\Sigma} \le d^2$. Applying Statement D.3 with $p=3$ and $\alpha=2$, we get

$$\frac{1}{n}\sum_{i=1}^{n}\big\|\boldsymbol{\Sigma}^{-1/2}T_i\big\|_{\mathrm{F}}^3 \lesssim d^2 K_T\sqrt{\log\frac{K_T}{d}} + K_T^3\left(\left(\log\frac{K_T}{d}\right)^{3/2} + (\mathrm{t}+\log n)^{\frac{1}{2}}\right)\frac{\mathrm{t}}{n} \lesssim d^2 K_T\sqrt{\log\frac{K_T}{d}}.$$

□

Now, we set

$$\sigma_F^2 \overset{\text{def}}{=} \Big\|\mathbb{E}\left[\boldsymbol{dT}_B^{S_1} - \boldsymbol{F}\right]^2\Big\|, \quad C_F \overset{\text{def}}{=} \frac{v_F^2}{\sigma_F^2}\log\frac{v_F}{\sigma_F}.$$

Lemma D.8 (Assumption (F))**.** *Assumption* (P) *ensures that for all* $\mathrm{x}>0$ *it holds that for sufficiently large* n, $n \gtrsim C_F(\mathrm{x}+\log d)$

$$\varepsilon_F(\mathrm{x}) \lesssim \|\boldsymbol{F}^{-1}\|\sigma_F\sqrt{\frac{\mathrm{x}+\log d}{n}}.$$

Proof. We set $\boldsymbol{X}_i = \boldsymbol{dT}_B^{S_i} - \mathbb{E}\,\boldsymbol{dT}_B^{S_i} = \boldsymbol{dT}_B^{S_i} - \boldsymbol{F}$. By construction $\boldsymbol{X}_1, \dots, \boldsymbol{X}_n$ are i.i.d. Moreover, Lemma 5.3 ensures that $\|\boldsymbol{X}_1\|$ is sub-Gaussian with parameter v_F. Statement D.1 ensures that with probability at least $1 - e^{-\mathrm{x}}$,

$$\|\boldsymbol{F} - \boldsymbol{F}_\mu\| \lesssim \sigma_F \sqrt{\frac{\mathrm{x} + \ln d}{n}} + v_F \left(\log \frac{v_F}{\sigma_F}\right)^{1/2} \frac{\mathrm{x} + \ln d}{n} \lesssim \sigma_F \sqrt{\frac{\mathrm{x} + \log d}{n}}.$$

Taking into account that $r(A, B) \le \|B^{-1}\| \|A - B\|$, we get the result. □

Lemma D.9 (Assumption ($\hat{F}$))**.** *Let Assumptions* (P) *and* (W) *be true. For sufficiently large* $n \gtrsim C_w C_F (\mathrm{x} + \log d)(\mathrm{t} + \log n)$*, it holds*

$$\hat{\varepsilon}_F(\mathrm{x}, \mathrm{t}) \lesssim (\|\boldsymbol{F}^{-1}\| \sigma_F + 1) \sqrt{\frac{\mathrm{x} + \mathrm{t} + \log d}{n}}$$

Proof. We set again $\boldsymbol{X}_i = \boldsymbol{dT}_B^{S_i} - \mathbb{E}\,\boldsymbol{dT}_B^{S_i} = \boldsymbol{dT}_B^{S_i} - \boldsymbol{F}$ and consider

$$\boldsymbol{F}_{\hat{\mu}} - \boldsymbol{F} = \frac{1}{n}\sum_{i=1}^n w_i \boldsymbol{dT}_B^{S_i} - \boldsymbol{F} = \frac{1}{n}\sum_{i=1}^n (w_i - 1)\boldsymbol{X}_i + \boldsymbol{F} \cdot \frac{1}{n}\sum_{i=1}^n (w_i - 1) + \boldsymbol{F}_\mu - \boldsymbol{F}.$$

Thus,

$$r(\boldsymbol{F}, \boldsymbol{F}_{\hat{\mu}}) \le \|\boldsymbol{F}^{-1}\| \left\| \frac{1}{n}\sum_{i=1}^n (w_i - 1)\boldsymbol{X}_i \right\| + \left| \frac{1}{n}\sum_{i=1}^n (w_i - 1) \right| + r(\boldsymbol{F}, \boldsymbol{F}_\mu).$$

Next, since the weights w_i are sub-exponential with $\mathrm{Var}(w) = 1$, Statement D.1 yields

$$\left| \frac{1}{n}\sum_{i=1}^n (w_i - 1) \right| \le \sqrt{\frac{\mathrm{x}}{n}} + \frac{\mathrm{x}}{n} v_w \log v_w \lesssim \sqrt{\frac{\mathrm{x}}{n}}.$$

The last step is to bound $\frac{1}{n}\sum_{i=1}^n (w_i - 1)\boldsymbol{X}_i$. We apply Statement D.1 and get with probability $1 - e^{-\mathrm{x}}$

$$\left\| \frac{1}{n}\sum_{i=1}^n (w_i - 1)\boldsymbol{X}_i \right\| \lesssim \sqrt{\left\| \frac{1}{n}\sum_i \boldsymbol{X}_i^2 \right\| \frac{\mathrm{x} + \log d}{n}} + v_w \log w \cdot \max_i \|\boldsymbol{X}_i\| \frac{\mathrm{x} + \log d}{n}$$

Statement D.3 ensures that with probability at least $1 - 2e^{-\mathrm{t}}$, $\max_i \|\boldsymbol{X}_i\| \lesssim v_F \sqrt{\mathrm{t} + \log n}$.

Now we set $\boldsymbol{Y}_i = \boldsymbol{X}_i^2$ and notice that

$$\begin{aligned} \left\| \mathbb{E}\,(\boldsymbol{Y}_1 - \mathbb{E}\,\boldsymbol{Y}_1)^2 \right\| \le \left\| \mathbb{E}\,\boldsymbol{Y}_1^2 \right\| = \left\| \mathbb{E}\,\boldsymbol{X}_1^4 \right\| \lesssim \sigma_F^2 v_F \log \frac{v_F}{\sigma_F} \\ \left\| \lambda_{\max}\,(\boldsymbol{Y}_1 - \mathbb{E}\,\boldsymbol{Y}_1)_+ \right\|_{\psi_1} \le \left\| \|\boldsymbol{Y}_1\| \right\|_{\psi_1} = \left\| \|\boldsymbol{X}_1\|^2 \right\|_{\psi_1} = v_F^2 \end{aligned}$$

Consequently, Statement D.1 yields with probability at least $1 - e^{-\mathrm{t}}$

$$\left\| \frac{1}{n}\sum_{i=1}^n \boldsymbol{X}_i^2 \right\| \lesssim \left\| \mathbb{E}\,\boldsymbol{X}_i^2 \right\| + \sigma_F v_F \sqrt{\frac{\mathrm{t} + \log d}{n} \log \frac{v_F}{\sigma_F}} + v_F^2 \frac{\mathrm{t} + \log d}{n} \log \frac{v_F}{\sigma_F} \lesssim \sigma_F^2.$$

The last inequality holds due to the bound on n.

Consequently,

$$\left\|\frac{1}{n}\sum_{i=1}^{n}(w_i-1)\boldsymbol{X}_i\right\| \lesssim \sigma_F\sqrt{\frac{\mathrm{x}+\log d}{n}} + v_w \log w \cdot v_F\sqrt{\mathrm{t}+\log n}\frac{\mathrm{x}+\log d}{n} \lesssim \sigma_F\sqrt{\frac{\mathrm{x}+\log d}{n}}.$$

By Lemma D.8, $r(\boldsymbol{F},\boldsymbol{F}_\mu) \le \varepsilon_F(\mathrm{t})$, with probability at least $1-e^{-\mathrm{t}}$. Combining all the bounds, we get

$$\hat{\varepsilon}_F(\mathrm{x};\mathrm{t}) \lesssim \varepsilon_F(\mathrm{t}) + \|\boldsymbol{F}^{-1}\|\sigma_F\sqrt{\frac{\mathrm{x}+\log d}{n}} + \sqrt{\frac{\mathrm{x}}{n}} \lesssim (\|\boldsymbol{F}^{-1}\|\sigma_F+1)\sqrt{\frac{\mathrm{x}+\mathrm{t}+\log d}{n}}.$$

□

Lemma D.10 (Assumption (Ξ)). *Assumption* (P) *ensures for all sufficiently large* $n \gtrsim \mathrm{t}C_T$*, that*

$$\varepsilon_\Xi(\mathrm{t}) \lesssim \sigma_T^2\left\|\boldsymbol{F}^{-1}\right\|^2\sqrt{C_T\frac{\mathrm{t}+d^2}{n}}.$$

Proof. First, we notice that

$$\|\Xi-\Xi_\mu\|_1 \le \left\|\boldsymbol{F}^{-1}\right\|^2\|\boldsymbol{\Sigma}_\mu-\boldsymbol{\Sigma}\|_1.$$

Further, $\mathbb{E}\|T_1\otimes T_1\|_1^2 = \mathbb{E}\|T_1\|_{\mathrm{F}}^4$. Thus

$$\mathbb{E}\|T_1\otimes T_1-\boldsymbol{\Sigma}\|_1^2 \lesssim \mathbb{E}\|T_1\otimes T_1\|_1^2 = \mathbb{E}\|T_1\|_{\mathrm{F}}^4 \lesssim \sigma_T^2 v_T^2\log\frac{v_T}{\sigma_T}.$$

Moreover,

$$\|\|T_1\otimes T_1-\boldsymbol{\Sigma}\|_1\|_{\psi_1} \le \|\boldsymbol{\Sigma}\|_1 + \left\|\|T_1\|_{\mathrm{F}}^2\right\|_{\psi_1} \le 2\left\|\|T_1\|_{\mathrm{F}}^2\right\|_{\psi_1} \le 2v_T^2.$$

Consequently, Corollary 3.5 from [Kroshnin and Suvorikova, 2024] ensures that, with probability at least $1-e^{-\mathrm{t}}$,

$$\|\boldsymbol{\Sigma}_\mu-\boldsymbol{\Sigma}\|_1 \lesssim \mathbb{E}\|\boldsymbol{\Sigma}_\mu-\boldsymbol{\Sigma}\|_1 + \sigma_T v_T\sqrt{\frac{\mathrm{t}}{n}\log\frac{v_T}{\sigma_T}} + v_T^2 z\frac{\mathrm{t}}{n},$$

where

$$z = \log\frac{v_T^2}{\sigma_T v_T\sqrt{\log\frac{v_T}{\sigma_T}}} \le \log\frac{v_T}{\sigma_T}.$$

Further,

$$\begin{aligned}\mathbb{E}\|\boldsymbol{\Sigma}_\mu-\boldsymbol{\Sigma}\|_1 &\le d\,\mathbb{E}\|\boldsymbol{\Sigma}_\mu-\boldsymbol{\Sigma}\|_2 \le d\sqrt{\mathbb{E}\|\boldsymbol{\Sigma}_\mu-\boldsymbol{\Sigma}\|_2^2} = d\sqrt{\frac{1}{n}\mathbb{E}\|T_1\otimes T_1-\boldsymbol{\Sigma}\|_2^2}\\ &\le d\sqrt{\frac{1}{n}\mathbb{E}\|T_1\otimes T_1\|_2^2} = d\sqrt{\frac{1}{n}\mathbb{E}\|T_1\|_{\mathrm{F}}^4} \lesssim d\sigma_T v_T\sqrt{\frac{1}{n}\log\frac{v_T}{\sigma_T}}.\end{aligned}$$

Combining all the bounds, we get

$$\|\boldsymbol{\Sigma}_\mu-\boldsymbol{\Sigma}\|_1 \lesssim d\sigma_T v_T\sqrt{\frac{1}{n}\log\frac{v_T}{\sigma_T}} + \sigma_T v_T\sqrt{\frac{\mathrm{t}}{n}\log\frac{v_T}{\sigma_T}} + v_T^2\log\frac{v_T}{\sigma_T}\frac{\mathrm{t}}{n} \lesssim \sigma_T v_T\sqrt{\frac{\mathrm{t}+d^2}{n}\log\frac{v_T}{\sigma_T}}.$$

□

Denote $C_\varepsilon \overset{\text{def}}{=} \kappa(B)\kappa(\boldsymbol{F})\left(c_B\sigma_T + \|\boldsymbol{F}^{-1}\|\sigma_F\right)^2$.

Lemma D.11 (Gaussian approximation for $\mathcal{W}(B, B_n)$)**.** *Set*

$$N \overset{\text{def}}{=} \max\{C_T, C_F \log d, C_\varepsilon \log d\}.$$

Let $n \gtrsim N$*, then it holds that*

$$\mathcal{E} \lesssim d^3\sqrt{\frac{C_G}{n}} + \gamma(\Xi)\sqrt{\frac{C_\varepsilon}{n}\log\frac{nd}{C_\varepsilon}}$$

Proof. Recall that the GAR bounding term is

$$\mathcal{E} \lesssim \varepsilon_G + \inf_{\mathrm{x}\in\mathcal{X}}\left\{e^{-\mathrm{x}} + \gamma(\Xi')\sqrt{\kappa(B)}\varepsilon(\mathrm{x})\right\}, \quad \mathcal{X} \overset{\text{def}}{=} \left\{\mathrm{x} : \varepsilon(\mathrm{x}) \le \frac{1}{6\sqrt{\kappa(B)}}\right\}. \tag{D.6}$$

For the sake of completeness, we also recall that

$$\varepsilon(\mathrm{x}) \overset{\text{def}}{=} 6\sqrt{\kappa(\boldsymbol{F})}\left(c_B\varepsilon_T(\mathrm{x}) + \varepsilon_{\mathrm{F}}(\mathrm{x})\right),$$

with c_B coming from (2.11). Using Lemma D.4 and D.6, we get for any $x \ge 1$

$$\varepsilon(\mathrm{x}) \lesssim \sqrt{\kappa(\boldsymbol{F})}\left(c_B\sigma_T\sqrt{\frac{\mathrm{x}}{n}} + \|\boldsymbol{F}^{-1}\|\sigma_F\sqrt{\frac{\mathrm{x}+\log d}{n}}\right) \lesssim \sqrt{\frac{C_\varepsilon}{\kappa(B)n}(\mathrm{x}+\log d)}.$$

Taking $\mathrm{x} = \frac{1}{2}\log\frac{n}{C_\varepsilon}$ and using assumption on n, we ensure that

$$\kappa(B)\varepsilon^2(\mathrm{x}) \lesssim \frac{C_\varepsilon}{n}\left(\log\frac{n}{C_\varepsilon} + \log d\right) \lesssim 1.$$

Thus, the condition $\mathrm{x} \in \mathcal{X}$ is satisfied. Substituting ε_G from Lemma D.6 to (D.6), we get the result. □

Denote

$$\hat{C}_\varepsilon \overset{\text{def}}{=} \kappa(B)\kappa(\boldsymbol{F})\left(c_B\sigma_T + \|\boldsymbol{F}^{-1}\|\sigma_F + 1\right)^2, \quad \hat{C}_T \overset{\text{def}}{=} \kappa(B)\kappa^2(\boldsymbol{F}),$$
$$\hat{C}_G(\mathrm{t}) \overset{\text{def}}{=} \max\left\{(\mathrm{t}+\log d)K_T^2\log K_T, (\mathrm{t}+\log d)^{3/2}\left(\frac{K_T}{d}\right)^2\right\}.$$

Lemma D.12. *Let*

$$\hat{N}(\mathrm{t}) \overset{\text{def}}{=} \max\left\{C_w C_T \mathrm{t}, C_w C_F \mathrm{t}\log d, \hat{C}_\varepsilon(\mathrm{t}+\log d), \hat{C}_G(\mathrm{t}), \hat{C}_T\mathrm{t}\right\},$$

then for $n \gtrsim \hat{N}(\mathrm{t})$ *with probability* $1 - \mathtt{C}e^{-\mathrm{t}}$

$$\hat{\mathcal{E}}(\mathrm{t}) \lesssim \hat{\varepsilon}_G(\mathrm{t}) + \gamma(\Xi'_\mu)\left(1 + \sqrt{\frac{\mathrm{tr}(\Xi')}{\mathrm{tr}(\Xi'_\mu)}}\right)\sqrt{\frac{\hat{C}_\varepsilon}{n}\left(\mathrm{t}+\log\frac{nd}{\hat{C}_\varepsilon}\right)}.$$

Proof. To get the bound on the random variable $\hat{\mathcal{E}}(\mathrm{t})$, we note that Lemma 3.3 ensures with probability $1 - \mathtt{C}e^{-\mathrm{t}}$

$$\hat{\mathcal{E}}(\mathrm{t}) \lesssim \hat{\varepsilon}_G(\mathrm{t}) + \inf_{\mathrm{x} \in \hat{\mathfrak{X}}(\mathrm{t})} \left\{ e^{-\mathrm{x}} + \gamma(\boldsymbol{\Xi}'_\mu) \sqrt{\kappa(B)} \left(\varepsilon(\mathrm{t}) + \hat{\varepsilon}(\mathrm{x}, \mathrm{t}) \right) \left(\frac{\|\boldsymbol{A}\| \|\boldsymbol{F}^{-1}\|}{\sqrt{\mathrm{tr}(\boldsymbol{\Xi}'_\mu)}} \varepsilon_T(\mathrm{t}) + 1 \right) \right\}, \tag{D.7}$$

where

$$\hat{\mathfrak{X}}(\mathrm{t}) \stackrel{\mathrm{def}}{=} \left\{ \mathrm{x} : \; \hat{\varepsilon}(\mathrm{x}, \mathrm{t}) \le \frac{1}{12\sqrt{\kappa(B)}} \right\}.$$

First, we use Lemma D.5 and D.9 and get

$$\begin{aligned} \hat{\varepsilon}(\mathrm{x}, \mathrm{t}) &\stackrel{\mathrm{def}}{=} 6\sqrt{\kappa(\boldsymbol{F})} \left(c_B \hat{\varepsilon}_T(\mathrm{x}, \mathrm{t}) + \hat{\varepsilon}_F(\mathrm{x}, \mathrm{t}) \right) \\ &\lesssim \sqrt{\kappa(\boldsymbol{F})} \left(c_B \sigma_T \sqrt{\frac{\mathrm{x}}{n}} + (\sigma_F \|\boldsymbol{F}^{-1}\| + 1) \sqrt{\frac{\mathrm{x} + \mathrm{t} + \log d}{n}} \right) \lesssim \sqrt{\frac{\hat{C}_\varepsilon}{\kappa(B) n} (\mathrm{x} + \mathrm{t} + \log d)}. \end{aligned}$$

Condition on n yields

$$\frac{\|\boldsymbol{A}\| \|\boldsymbol{F}^{-1}\|}{\sqrt{\mathrm{tr}(\boldsymbol{\Xi}'_\mu)}} \varepsilon_T(\mathrm{t}) \lesssim \frac{\|\boldsymbol{A}\| \|\boldsymbol{F}^{-1}\|}{\sqrt{\mathrm{tr}(\boldsymbol{\Xi}'_\mu)}} \sigma_T \sqrt{\frac{\mathrm{t}}{n}} \lesssim \frac{1}{\|\boldsymbol{A}^{-1}\| \|\boldsymbol{F}\|} \sqrt{\frac{\mathrm{tr}\, \boldsymbol{\Sigma}}{\mathrm{tr}(\boldsymbol{\Xi}'_\mu)}} \lesssim \sqrt{\frac{\mathrm{tr}(\boldsymbol{\Xi}')}{\mathrm{tr}(\boldsymbol{\Xi}'_\mu)}}.$$

Next, according to the proof of Lemma D.11, $\varepsilon(\mathrm{t}) \lesssim \sqrt{\frac{C_\varepsilon}{\kappa(B) n} \mathrm{t}} \le \sqrt{\frac{\hat{C}_\varepsilon}{\kappa(B) n} \mathrm{t}}$.

Taking $\mathrm{x} = \frac{1}{2} \log \frac{n}{\hat{C}_\varepsilon}$, we obtain that

$$\kappa(B) (\varepsilon(\mathrm{t}) + \hat{\varepsilon}(\mathrm{x}; \mathrm{t}))^2 \lesssim \frac{\hat{C}_\varepsilon}{n} (\mathrm{x} + \mathrm{t} + \log d) \lesssim 1,$$

hence $\mathrm{x} \in \hat{\mathfrak{X}}(\mathrm{t})$, and

$$\hat{\mathcal{E}}(\mathrm{t}) \lesssim \hat{\varepsilon}_G(\mathrm{t}) + \gamma(\boldsymbol{\Xi}'_\mu) \left(1 + \sqrt{\frac{\mathrm{tr}(\boldsymbol{\Xi}')}{\mathrm{tr}(\boldsymbol{\Xi}'_\mu)}} \right) \sqrt{\frac{\hat{C}_\varepsilon}{n} \left(\mathrm{t} + \log \frac{nd}{\hat{C}_\varepsilon} \right)}.$$

□

Before proving the theorem, we collect some definitions used throughout the text below for completeness. The constants from lemmata that ensure GAR,

$$C_w \stackrel{\mathrm{def}}{=} (v_w \log v_w)^2, \quad C_T \stackrel{\mathrm{def}}{=} \frac{v_T^2}{\sigma_T^2} \log \frac{v_T}{\sigma_T}, \quad \hat{C}_T \stackrel{\mathrm{def}}{=} \kappa(B) \kappa^2(\boldsymbol{F}), \quad C_F \stackrel{\mathrm{def}}{=} \frac{v_F^2}{\sigma_F^2} \log \frac{v_F}{\sigma_F}, \tag{D.8}$$

$$K_T \stackrel{\mathrm{def}}{=} \left\| \left\| \boldsymbol{\Sigma}^{-1/2} T_i \right\|_{\mathrm{F}} \right\|_{\psi_2} \le \|\boldsymbol{\Sigma}^{\boldsymbol{-1/2}}\| v_T, \quad C_G \stackrel{\mathrm{def}}{=} \left(\frac{K_T}{d} \right)^2 \log \frac{K_T}{d}. \tag{D.9}$$

Moreover, the constants coming from Lemma D.11 and Lemma D.12

$$\begin{aligned} C_\varepsilon &= \kappa(B)\kappa(\boldsymbol{F})\left(c_B\sigma_T + \|\boldsymbol{F}^{-1}\|\sigma_F\right)^2, \\ \hat{C}_\varepsilon &\overset{\text{def}}{=} \kappa(B)\kappa(\boldsymbol{F})\left(c_B\sigma_T + \|\boldsymbol{F}^{-1}\|\sigma_F + 1\right)^2, \\ \hat{C}_G(\mathrm{t}) &\overset{\text{def}}{=} \max\left\{(\mathrm{t}+\log d)K_T^2\log K_T, (\mathrm{t}+\log d)^{3/2}\left(\frac{K_T}{d}\right)^2\right\}. \end{aligned} \tag{D.10}$$

In the following, we assume that

$$\begin{gathered} n \gtrsim \max\{N, \hat{N}(\mathrm{t}), \mathrm{t}C_T\} \qquad \text{(D.11)} \\ N \overset{\text{def}}{=} \max\{C_T, C_F\log d, C_\varepsilon\log d\} \\ \hat{N}(\mathrm{t}) \overset{\text{def}}{=} \max\left\{C_wC_T\mathrm{t}, C_wC_F\mathrm{t}\log d, \hat{C}_\varepsilon(\mathrm{t}+\log d), \hat{C}_G(\mathrm{t}), \hat{C}_T\mathrm{t}\right\}. \end{gathered}$$

Proof of Theorem 5.1. If W is s.t. $\mathbb{P}_w\{w=0\}=0$, the proof is trivial and reduces to validation of all assumptions in Theorem 4.1.

Now we consider the weight generating law W, s.t. $\mathbb{P}_w\{w=0\}=p_0$. Let an auxiliary measure $\widetilde{\mu}$ be

$$\widetilde{\mu} = \sum_i w_i\delta_{S_i}, \quad \text{s.t.} \sum_i w_i \neq 0,$$

and set, w.l.o.g., $\mathcal{B}(0) \overset{\text{def}}{=} B_0$ with $B_0 \in \mathbb{H}_{++}(d)$ being some fixed matrix.

We aim to show that

$$\left|\mathbb{P}\left\{\mathcal{W}(B_{\widetilde{\mu}}, B_\mu) \le z|\mu\right\} - \mathbb{P}\left\{\mathcal{W}(B_{\hat{\mu}}, B_\mu) \le z|\mu\right\}\right| \le p_0^n. \tag{D.12}$$

We will use the following facts,

$$\begin{gathered} \mathbb{P}\{A|B\} - \mathbb{P}\{A\} = \frac{\mathbb{P}\{A\cap B\}}{\mathbb{P}\{B\}} - \mathbb{P}\{A\} \le \mathbb{P}\{A\} + \left(\frac{1}{\mathbb{P}\{B\}} - 1\right)\mathbb{P}\{B\} - \mathbb{P}\{A\} \le 1 - \mathbb{P}\{B\}, \\ \mathbb{P}\{A|B\} - \mathbb{P}\{A\} \ge \mathbb{P}\{A\cap B\} - \mathbb{P}\{A\} \ge -\left(1-\mathbb{P}\{B\}\right). \end{gathered}$$

Thus, for a fixed set $S_1,\ldots,S_n$,

$$\left|\mathbb{P}_w\left\{\mathcal{W}(B_{\hat{\mu}}, B_\mu) \le z\middle|\sum_i w_i \neq 0\right\} - \mathbb{P}_w\left\{\mathcal{W}(B_{\widetilde{\mu}}, B_\mu) \le z\right\}\right| \le \mathbb{P}_w\left\{\sum_i w_i = 0\right\} = p_0^n.$$

Now, we notice that the condition $\sum_i w_i = 0$ is equivalent to $\hat{\mu}=0$. Thus, (D.12) follows from

$$\begin{aligned} &\left|\mathbb{P}\left\{\mathcal{W}(B_{\widetilde{\mu}}, B_\mu) \le z|\mu\right\} - \mathbb{P}\left\{\mathcal{W}(B_{\hat{\mu}}, B_\mu) \le z|\mu\right\}\right| \\ =&\left|\mathbb{P}\left\{\mathcal{W}(B_{\widetilde{\mu}}, B_\mu) \le z|\mu\right\} - \mathbb{P}\left\{\mathcal{W}(B_{\widetilde{\mu}}, B_\mu) \le z|\mu, \hat{\mu}\neq 0\right\}\right| \\ \le&\ \mathbb{P}\left\{\hat{\mu}=0|\mu\right\} = \mathbb{P}_w\left\{\sum_i w_i = 0\right\} = p_0^n. \end{aligned}$$

Further, Lemma 3.3, being applied to $\widetilde{\mu}$ (instead of $\hat{\mu}$) together with the above bound, yields for all $z>0$

$$\left|\mathbb{P}\left\{\mathcal{W}(B_{\hat{\mu}}, B_\mu) \le z \mid \mu\right\} - \mathbb{P}\left\{\|\boldsymbol{A}Z_\mu\|_{\mathrm{F}} \le z \mid \mu\right\}\right| \le \hat{\mathcal{E}}(\mathrm{t}) + p_0^n.$$

Thus, the resulting bound is written as

$$\sup_{z\geq 0}|\mathbb{P}\left\{\mathcal{W}(B_\mu, B)\leq z\right\}-\mathbb{P}\left\{\mathcal{W}(B_{\hat{\mu}}, B_\mu)\leq z\mid\mu\right\}|\leq\mathit{\Gamma}(\mathrm{t})+p_0^n.$$

Finally, to get the asymptotic bound on $\mathit{\Gamma}(\mathrm{t})+p_0^n$ for large n, we summarize all auxiliary results from this section.

To get the second result, we recall Theorem 4.1 and notice that

$$\mathit{\Gamma}(\mathrm{t})\lesssim\varkappa(\boldsymbol{\Xi}')\|\boldsymbol{A}\|^2\varepsilon_{\Xi}(\mathrm{t})+\mathcal{E}+\hat{\mathcal{E}}(\mathrm{t}).$$

First, we recall Lemma D.12,

$$\hat{\mathcal{E}}(\mathrm{t})\lesssim\hat{\varepsilon}_G(\mathrm{t})+\gamma(\boldsymbol{\Xi}'_\mu)\left(1+\sqrt{\frac{\mathrm{tr}(\boldsymbol{\Xi}')}{\mathrm{tr}(\boldsymbol{\Xi}'_\mu)}}\right)\sqrt{\frac{\hat{C}_\varepsilon}{n}\left(\mathrm{t}+\log\frac{nd}{\hat{C}_\varepsilon}\right)}$$

Assumption on n ensures $\gamma(\boldsymbol{\Xi}'_\mu)\lesssim\gamma(\boldsymbol{\Xi}')$, $\mathrm{tr}(\boldsymbol{\Xi}'_\mu)\lesssim\mathrm{tr}\,\gamma(\boldsymbol{\Xi}')$ (see C.2). Using Lemmata D.11, D.12 D.10, and the fact that by definition $\hat{C}_\varepsilon>C_\varepsilon$, we get

$$\mathit{\Gamma}(\mathrm{t})\lesssim d^3\sqrt{\frac{C_G}{n}}+\gamma(\boldsymbol{\Xi}')\sqrt{\frac{\hat{C}_\varepsilon}{n}(\mathrm{t}+\log\frac{nd}{\hat{C}_\varepsilon})}+\varkappa(\boldsymbol{\Xi}')\|\boldsymbol{A}\|^2\|\boldsymbol{F}^{-1}\|^2\sigma_T^2\sqrt{\frac{C_T}{n}(\mathrm{t}+d^2)}.$$

Finally,

$$\gamma(\boldsymbol{\Xi}')=\varkappa(\boldsymbol{\Xi}')\,\mathrm{tr}\,\boldsymbol{\Xi}'\leq\varkappa(\boldsymbol{\Xi}')\|\boldsymbol{A}\|^2\|\boldsymbol{F}^{-1}\|^2\,\mathrm{tr}\,\boldsymbol{\Sigma}=\varkappa(\boldsymbol{\Xi}')\|\boldsymbol{A}\|^2\|\boldsymbol{F}^{-1}\|^2\sigma_T^2.$$

Combining the bounds, we get the result. □