Weierstraß-Institut
für Angewandte Analysis und Stochastik
Leibniz-Institut im Forschungsverbund Berlin e. V.

Preprint ISSN 2198-5855

Statistical inference for Bures-Wasserstein barycenters

Alexey Kroshnin[1],

Vladimir Spokoiny[2], Alexandra Suvorikova[2]

submitted: November 10, 2020

[1] Institute for Information Transmission RAS
Bolshoy Karetny per. 19
127051 Moscow
Russia
E-Mail:kroshnin@math.univ-lyon1.fr

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

No. 2788
Berlin 2020

2010 *Mathematics Subject Classification.* 60F05.

Key words and phrases. Bures-Wasserstein barycenters, central limit theorem, optimal transport.

This work was partially supported by Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germanys Excellence Strategy The Berlin Mathematics Research Center MATH+ (EXC-2046/1, project ID: 390685689).

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/

Statistical inference for Bures-Wasserstein barycenters

Alexey Kroshnin,

Vladimir Spokoiny, Alexandra Suvorikova

Abstract

In this work we introduce the concept of Bures–Wasserstein barycenter Q_*, that is essentially a Fréchet mean of some distribution $\mathbb{P}$ supported on a subspace of positive semi-definite d-dimensional Hermitian operators $\mathbb{H}_+(d)$. We allow a barycenter to be constrained to some affine subspace of $\mathbb{H}_+(d)$, and we provide conditions ensuring its existence and uniqueness. We also investigate convergence and concentration properties of an empirical counterpart of Q_* in both Frobenius norm and Bures–Wasserstein distance, and explain, how the obtained results are connected to optimal transportation theory and can be applied to statistical inference in quantum mechanics.

1 Introduction

The space of finite-dimensional Hermitian operators is commonly applied for data representation. For instance, in quantum mechanics it is used for mathematical description of physical properties of a quantum system: the real-valued spectrum is associated to measurements observed in a physical experiment. Real-valued symmetric matrices are also widely used for description of systems in engineering applications, medical studies, neural sciences, evolutionary biology etc. Usually, one assumes a sample to be random, see, e.g., Goodnight and Schwartz [1997], Calsbeek and Goodnight [2009], Álvarez-Esteban et al. [2015], Del Barrio et al. [2017], Gonzalez et al. [2017]. Statistical characteristics of its distribution $\mathbb{P}$, such as mean and variance, are of interest for experimental design, and analysis of the results for further development of natural science models.

The current study focuses on the space of positive semi-definite Hermitian matrices $\mathbb{H}_+(d)$ and presents a theoretical analysis of aggregation methods of the relevant statistical information from data sets, for which the hypothesis of linearity might be violated. In this case, the widely-used Euclidean mean and variance are not sensitive enough to capture effects of interest. For instance, some data sets are described by probability measures belonging to some scale-location family, e.g. Álvarez-Esteban et al. [2018], Muzellec and Cuturi [2018]. The non-linearity assumption requires an adaptation of the tools of classical statistical analysis. In order to capture non-linear effects, we suggest to endow $\mathbb{H}_+(d)$ with the Bures–Wasserstein distance d_{BW}, originally introduced by Bhatia et al. [2018]. For a pair of positive semi-definite matrices $Q, S \in \mathbb{H}_+(d)$ the distance is defined as:

$$d^2_{BW}(Q,S) = \operatorname{tr} Q + \operatorname{tr} S - 2 \operatorname{tr}\left(Q^{1/2} S Q^{1/2}\right)^{1/2}. \tag{1.1}$$

It is worth noting that being restricted to the sub-space of symmetric positive semidefinite matrices $\mathrm{Sym}_+(d)$, d_{BW} turns into the 2-Wasserstein distance between two normal distributions. Let $\mathcal{N}(0,Q)$ and $\mathcal{N}(0,S)$ be two centred Gaussians. The 2-Wasserstein distance is

$$d^2_{W_2}\left(\mathcal{N}(0,Q), \mathcal{N}(0,S)\right) = \operatorname{tr} Q + \operatorname{tr} S - 2 \operatorname{tr}\left(Q^{1/2} S Q^{1/2}\right)^{1/2}.$$

DOI 10.20347/WIAS.PREPRINT.2788

It is worth noting that a natural extension of the Gaussian case is the case of distributions coming from the same scale-location family (see, e.g., Agueh and Carlier [2011], Section 6, or Álvarez-Esteban et al. [2018]).

In the last few years, the class of optimal transportation distances and in particular the 2-Wasserstein distance attract a lot of attention of the both mathematical and machine learning communities. The latter captures the geometrical similarities between objects coming from non-linear spaces, see, e.g., Courty et al. [2016], Montavon et al. [2016], Flamary et al. [2018], while the recent advances in computations make the distance useful for the real-world problems Cuturi [2013], Uribe et al. [2018], Gramfort et al. [2015]. For more information on the Wasserstein distance and optimal transportation theory in general, we recommend the excellent monograph by Villani [2009]. The book by Peyré et al. [2019] provides a state-of-the-art survey of numerical methods and their applications in data sciences.

This study focuses on the following statistical setting. Let $\mathbb{P}$ be a probability distribution supported on the set of non-negatively definite Hermitian matrices $\mathbb{H}_+(d)$. Two important characteristics of $\mathbb{P}$ are the Fréchet mean and variance. While the former is a “typical” representative of a data set in hand, the latter appears in the analysis of data variability, see, e.g., Del Barrio et al. [2015]. We briefly recall both concepts below. For an arbitrary point $Q \in \mathbb{H}_+(d)$, the Fréchet variance of $\mathbb{P}$ is defined as

$$\mathcal{V}(Q) \stackrel{\text{def}}{=} \int_{\mathbb{H}_+(d)} d^2_{BW}(Q, S) d\,\mathbb{P}(S).$$

The Fréchet mean of $\mathbb{P}$ is given by the set of global minimizers of the variance $\mathcal{V}(Q)$:

$$Q_* \in \operatorname*{argmin}_{Q \in \mathbb{H}_+(d)} \mathcal{V}(Q). \tag{1.2}$$

However, in some cases one might be interested in a minimizer belonging to an affine sub-space of Hermitian operators $\mathbb{H}(d)$, $\mathbb{A} \subset \mathbb{H}(d)$:

$$Q_* \in \operatorname*{argmin}_{Q \in \mathbb{H}_+(d) \cap \mathbb{A}} \mathcal{V}(Q). \tag{1.3}$$

For instance, such a necessity arises when considering a random set of quantum density operators. Section 3.2 discusses this example in more detail. Note that the setting (1.3) covers the setting (1.2). So, without loss of generality, we further address only (1.3).

Obviously, the first crucial question concerns existence and uniqueness of Q_*. Theorem 2.1 presents the positive answers to both issues. This immediately allows us to define the global Fréchet variance of $\mathbb{P}$:

$$\mathcal{V}_* \stackrel{\text{def}}{=} \mathcal{V}(Q_*).$$

Given an i.i.d. sample $S_1, \dots, S_n$, $S_i \sim \mathbb{P}$, one constructs an empirical version of $\mathcal{V}(Q)$, for an arbitrary Q, as follows:

$$\mathcal{V}_n(Q) \stackrel{\text{def}}{=} \frac{1}{n} \sum_{i=1}^{n} d^2_{BW}(Q, S_i).$$

The empirical Fréchet mean and the global empirical variance also exist and unique:

$$Q_n = \operatorname*{argmin}_{Q \in \mathbb{H}_+(d) \cap \mathbb{A}} \mathcal{V}_n(Q), \quad \mathcal{V}_n \stackrel{\text{def}}{=} \mathcal{V}_n(Q_n). \tag{1.4}$$

Both facts follow from Theorem 2.1.

This work studies the convergence of the estimators Q_n and $\mathcal{V}_n$ and investigates the concentration properties of both objects. A discussion of the practical applicability of the obtained results is postponed to Section 3. There we explain the connection to optimal transportation theory and present a possible application to statistical analysis in quantum mechanics.

1.1 Contribution of the present study

The central limit theorems for Q_n and $\mathcal{V}_n$ From now on we use bold symbols (e.g., $\boldsymbol{A}, \boldsymbol{B}$) to denote operators, whereas a classical font (e.g., A, B) stands for either matrices or vectors.

The first result of this study concerns the asymptotic normality of the approximation error of the population Fréchet mean by its empirical counterpart:

$$\sqrt{n}\,(Q_n - Q_*) \xrightarrow{\mathrm{w}} \mathcal{N}(0, \Xi),$$

where "$\xrightarrow{\mathrm{w}}$" stands for the weak convergence, and Ξ is a covariance operator acting on a linear subspace $\mathbb{M} \subset \mathbb{H}(d)$ associated with the affine subspace $\mathbb{A}$. The result is derived under some suitable assumptions on the distribution $\mathbb{P}$ introduced later, in Section 2.

At this point it is worth mentioning that the asymptotic normality of Q_n falls in the setting of the asymptotic normality of parametric M-estimations with a smooth and convex loss function $d_{BW}(\cdot,\cdot)$ defined over the convex set $\mathbb{A} \cap \mathbb{H}_+(d)$. The convexity and the smoothness of $d_{BW}(\cdot,\cdot)$ are validated by Lemma A.5 and Lemma A.6, respectively. However, the current study uses the proof techniques different from a verification of the standard assumptions on M-estimators. We discuss the issue in more detail in Section 2.4.

The above convergence result cannot be used directly for construction of asymptotic confidence sets, as it relies on the unknown covariance matrix Ξ. However, Theorem 2.2 ensures that the covariance operator Ξ can be replaced by an empirical counterpart $\hat{\boldsymbol{\Xi}}_{\boldsymbol{n}}$:

$$\sqrt{n}\,\hat{\boldsymbol{\Xi}}_{\boldsymbol{n}}^{-1/2}\,(Q_n - Q_*) \xrightarrow{\mathrm{w}} \mathcal{N}(0, \boldsymbol{I}),$$

where $\boldsymbol{I}$ denotes the identity operator. Along with the asymptotic normality of $\sqrt{n}(Q_n - Q_*)$, we are interested in the limiting distribution of $\mathcal{L}\left(\sqrt{n} d_{BW}(Q_n, Q_*)\right)$, where $\mathcal{L}(X)$ denotes the distribution of a random variable X. In what follows $\|A\|_F$ denotes the Frobenius norm of matrix A. Corollary 2.1 ensures

$$\mathcal{L}\left(\sqrt{n} d_{BW}(Q_n, Q_*)\right) \xrightarrow{\mathrm{w}} \mathcal{L}\left(\|\xi\|_F\right),$$

where ξ is some normally distributed vector. The data-driven asymptotic confidence sets for $\sqrt{n} d_{BW}(Q_n, Q_*)$ are obtained by replacing ξ by its empirical counterpart ξ_n:

$$d_{\mathrm{w}}\left(\mathcal{L}\left(\sqrt{n} d_{BW}(Q_n, Q_*)\right), \mathcal{L}\left(\|\xi_n\|_F\right)\right) \to 0,$$

where d_{w} is some metric inducing the weak convergence of measures. We also show the asymptotic normality of the approximation error of the variance $\mathcal{V}_*$ by its empirical analogue $\mathcal{V}_n$ (see Theorem 2.3):

$$\sqrt{n}\,(\mathcal{V}_n - \mathcal{V}_*) \xrightarrow{\mathrm{w}} \mathcal{N}\left(0, \operatorname{Var} d^2_{BW}(Q_*, S)\right).$$

All above-mentioned results are closely connected to the convergence of empirical 2-Wasserstein barycenters. For the sake of transparency we postpone a further discussion of this topic to Section 3.1.

The concentration of Q_n and $\mathcal{V}_n$ The technique of the proof of the central limit theorem, developed in the current study, appears to be suitable for an investigation of the concentration properties of Q_n and $\mathcal{V}_n$. To validate the concentration, we suppose the distribution $\mathbb{P}$ to be sub-Gaussian, see Assumption 3. This assumption ensures the following bounds which hold with high probability:

$$\|Q_*^{-1/2} Q_n Q_*^{-1/2} - I\|_F \le \frac{\mathtt{C}(\sqrt{m} + t)}{\sqrt{n}}, \quad d_{BW}(Q_n, Q_*) \le \frac{\mathtt{C}(\sqrt{m} + t)}{\sqrt{n}},$$

where m is the dimension of $\mathbb{M}$, $t \geq 0$, and $\mathtt{C}$ denotes a generic constant. For more details see Theorem 2.4 and Corollary 2.2, respectively. To the best of our knowledge, these results appear to be novel. Along with concentration of the empirical barycenter, we investigate the concentration of the empirical variance $\mathcal{V}_n$ which holds with high probability:

$$|\mathcal{V}_n - \mathcal{V}_*| \leq \max\left(\frac{\mu t^2}{n}, \frac{\nu t}{\sqrt{n}}\right) + \frac{\mathtt{C}(\sqrt{m}+t)^2}{n}$$

where μ and ν are some parameters depending on the distribution of $d^2_{BW}(Q_*, S)$, m is the dimension of $\mathbb{M}$, $\mathtt{C}$ stands for a generic constant, and the parameter is $t \geq 0$. The result is presented in Theorem 2.5. We discuss its relation to the existing results in Section 3.1.

The paper is organised as follows. Section 2 presents the obtained results in more detail. Section 3 illustrates the connection to other scientific problems. Finally, Section 4 suggests to use barycenters for replacement of the lost data. It contains a description of the idea, and experimental estimation of convergence rates for barycenters using both an artificial and a real data set. The latter one is related to the climate modelling.

2 Results

Following Bhatia et al. [2018], we continue to investigate properties of $d_{BW}(Q, S)$. Further we present an alternative analytical expression for the distance. The result is well-known for the case of real-valued symmetric matrices $Q, S \in \mathrm{Sym}_+(d)$, see Olkin and Pukelsheim [1982]. The proposition below extends it to the case of Hermitian matrices $\mathbb{H}_+(d)$.

Proposition 2.1. *Let $Q, S \in \mathbb{H}_+(d)$ and $Q \succ 0$. Then* (1.1) *can be rewritten as*

$$d^2_{BW}(Q,S) = \left\| \left(T_Q^S - I\right) Q^{1/2} \right\|_F^2 = \mathrm{tr}\left(T_Q^S - I\right) Q \left(T_Q^S - I\right),$$

where the optimal map from Q to S is

$$\begin{aligned} T_Q^S &\overset{\mathrm{def}}{=} \operatorname*{argmin}_{T : TQT^* = S} \left\|(T - I)Q^{1/2}\right\|_F \\ &= S^{1/2}\left(S^{1/2} Q S^{1/2}\right)^{-1/2} S^{1/2} = Q^{-1/2}\left(Q^{1/2} S Q^{1/2}\right)^{1/2} Q^{-1/2}. \end{aligned} \tag{2.1}$$

By $\left(S^{1/2}QS^{1/2}\right)^{-1/2}$ we denote the pseudo-inverse matrix $\left(\left(S^{1/2}QS^{1/2}\right)^{1/2}\right)^+$.

Note that being restricted to the sub-space $\mathrm{Sym}_{++}(d)$, T_Q^S coincides with the optimal push-forward (also known as the optimal map) between two centred normal distributions $\mathcal{N}(0, Q)$ and $\mathcal{N}(0, S)$: $T_Q^S \# \mathcal{N}(0, Q) = \mathcal{N}(0, S)$. For more details on general optimal transportation maps see Brenier [1991], for a particular case of scale-location families one may refer to Álvarez-Esteban et al. [2018], Takatsu et al. [2011]. The differentiability of T_Q^S is one of the key ingredients in the proofs. It is validated in Lemma A.2. Note that for a particular choice of $\mathbb{A}$ in (1.3), $\mathbb{A} = \mathrm{Sym}_{++}(d)$, the differentiability of optimal transportation maps are proved in Rippl et al. [2016]. Section A.2 is dedicated to the investigation of properties of T_Q^S and its differential $\boldsymbol{dT}_Q^S$. Section A.3 investigates properties of d_{BW}. We highly recommend to at least look through these two sections for a better understanding of the tools used in the proofs of central limit theorems and concentrations.

2.1 Existence and uniqueness of Q_* and Q_n

Along with investigation of properties of the distance in hand, and before moving to more general questions, one should ask her- or himself, whether the Fréchet mean Q_* exists and, if so, is it unique or not. We further assume that $\mathbb{A}$ has a non-empty intersection with the space of positive definite operators:

Assumption 1. Given the setting (1.3), we suppose an affine subspace $\mathbb{A} \subset \mathbb{H}(d)$ to be s.t. $\mathbb{H}_{++}(d) \cap \mathbb{A} \neq \emptyset$. By $\mathbb{M}$ we denote the linear subspace of $\mathbb{H}(d)$ associated with $\mathbb{A}$, i.e. the following representation holds: $\mathbb{A} = \{Q_0\} + \mathbb{M}$ for some $Q_0 \in \mathbb{H}(d)$.

Without loss of generality we assume that $\mathbb{P}$ assigns positive probability to the space of positive definite Hermitian matrices $\mathbb{H}_{++}(d)$. We also suppose $\mathbb{P}$ to be s.t. the spectrum of S is on average bounded away from infinity:

Assumption 2. Let data distribution $\mathbb{P}$, $S \sim \mathbb{P}$, be s.t.

$$\mathbb{P}\left(\mathbb{H}_{++}(d)\right) > 0, \quad \mathbb{E}\,\mathrm{tr}\, S < +\infty.$$

The next theorem ensures existence and uniqueness of the Fréchet mean introduced in (1.3).

Theorem 2.1 (Existence and uniqueness of Fréchet mean Q_*). *Under Assumptions 1 and 2, there exists unique positive-definite barycenter Q_* of $\mathbb{P}$, $Q_* \succ 0$. Moreover, it is characterised as the unique solution of the equation*

$$\boldsymbol{\mathit{\Pi}}_{\mathbb{M}}\, \mathbb{E}\, T_Q^S = \boldsymbol{\mathit{\Pi}}_{\mathbb{M}} I, \quad Q \in \mathbb{H}_{++}(d), \tag{2.2}$$

where $\boldsymbol{\mathit{\Pi}}_{\mathbb{M}}$ is the orthogonal projector onto $\mathbb{M}$.

Note that for any fixed $Q \in \mathbb{H}_{++}(d)$, T_Q^S is a random variable because it is a continuous function of the random variable S. The equation (2.2) generalises the result for scale-location families in 2-Wasserstein space, presented in Álvarez-Esteban et al. [2015], Theorem 3.10, and originally obtained for the Gaussian case in the seminal work Agueh and Carlier [2011], Theorem 6.1. Namely, if $\mathbb{A} = \mathrm{Sym}_{++}(d)$, then Q_* exists and is the unique solution of a fixed-point equation:

$$Q = \mathbb{E}\left(Q^{1/2} S Q^{1/2}\right)^{1/2}.$$

Note that it is similar to (2.2), as by multiplying the above equation from both sides by $Q^{-1/2}$ one obtains $\mathbb{E}\, T_Q^S = I$. Existence, uniqueness, and measurability of the estimator Q_n defined in (1.4) are a direct corollary of the above theorem. The proof of Theorem 2.1 is presented in Section A.4.

2.2 Limiting distributions of $\sqrt{n}(Q_n - Q_*)$, $\sqrt{n} d_{BW}(Q_n, Q_*)$, and $\sqrt{n}(\mathcal{V}_n - \mathcal{V}_*)$

Armed with the knowledge about properties of $d_{BW}(\cdot, \cdot)$, Q_*, and Q_n, we are now equipped enough to introduce the first main result of the current study. In what follows we denote the variance of optimal transportation map from the population barycenter Q_* to any $S \sim \mathbb{P}$ as

$$\mathrm{Var}\left(T_{Q_*}^S\right) = \mathbb{E}(T_{Q_*}^S - I) \otimes (T_{Q_*}^S - I), \text{ with } \mathbb{E}\, T_{Q_*}^S = I, \tag{2.3}$$

where $\otimes$ stands for the tensor product. Theorem 2.2 presents the asymptotic convergence of Q_n to Q_*.

Theorem 2.2 (Central limit theorem for the Fréchet mean)**.** *Under Assumptions 1 and 2 the approximation error rate of the Fréchet mean Q_* by its empirical counterpart Q_n is*

$$\sqrt{n}\,(Q_n - Q_*) \xrightarrow{\mathrm{w}} \mathcal{N}(0, \Xi)\,, \tag{A}$$

where Ξ is a self-adjoint linear operator acting from $\mathbb{M}$ to $\mathbb{M}$ defined in (A.7)*. Moreover, if $\operatorname{Var}\left(T^S_{Q_*}\right)$ is non-degenerated, then*

$$\sqrt{n}\,\hat{\Xi}_n^{-1/2}\,(Q_n - Q_*) \xrightarrow{\mathrm{w}} \mathcal{N}(0, (\boldsymbol{I})_{\mathbb{M}})\,, \tag{B}$$

where $\hat{\boldsymbol{\Xi}}_{\boldsymbol{n}}$ is a data-driven empirical counterpart of Ξ defined in (A.8)*.*

Remark 1. The notation $(\boldsymbol{A})_{\mathbb{M}}$ denotes a linear operator associated to the restriction of a quadratic form $\boldsymbol{A}$ to the subspace $\mathbb{M}$:

$$(\boldsymbol{A})_{\mathbb{M}} \colon \mathbb{M} \to \mathbb{M}, \quad X \mapsto \boldsymbol{\Pi}_{\mathbb{M}}\boldsymbol{A}(X).$$

We intentionally postpone the explicit definitions of Ξ and $\hat{\boldsymbol{\Xi}}_{\boldsymbol{n}}$, because they require an introduction of many technical details. This would make the description of the main results less transparent. It is worth noting that the result (B) enables construction of data-driven asymptotic confidence sets. However, inversion of the empirical covariance might be a problem. For instance, numerical simulations show that $\hat{\Xi}_n$ might be degenerated if $\mathbb{P}$ is supported on a set of diagonal matrices. This immediately raises a question concerning introduction of a resampling approach which would make the computations tractable. We consider this as a subject for the further research.

The proof of the central limit theorem relies on the Fréchet differentiablilty of T^S_Q by the lower argument Q at the point Q_*:

$$T^S_{Q_n} = S^{1/2}\left(S^{1/2}Q_nS^{1/2}\right)^{-1/2}S^{1/2}, \quad T^S_{Q_n} \approx T^S_{Q_*} + \boldsymbol{dT}^S_{Q_*}(Q_n - Q_*),$$

where $\boldsymbol{dT}^S_{Q_*}$ is a differential of T^S_Q at the point Q_*.

Since $\mathbb{H}_+(d)$ is endowed with the Bures–Wasserstein distance, the convergence properties of $d_{BW}(Q_n, Q_*)$ are also of great interest. The result is a corollary of the above theorem.

Corollary 2.1 (Asymptotic distribution of $d_{BW}(Q_n, Q_*)$)**.** *Under conditions of Theorem 2.2 it holds*

$$\mathcal{L}\left(\sqrt{n}d_{BW}(Q_n, Q_*)\right) \xrightarrow{\mathrm{w}} \mathcal{L}\left(\left\|Q_*^{1/2}\boldsymbol{dT}^{Q_*}_{Q_*}(Z)\right\|_F\right),$$

where $Z \in \mathbb{M} \subset \mathbb{H}(d)$ is a random matrix, $Z \sim \mathcal{N}(0, \Xi)$.

Moreover, replacing in the limiting distribution Q_* and Z by their empirical counterparts Q_n and $Z_n \sim \mathcal{N}\left(0, \hat{\boldsymbol{\Xi}}_{\boldsymbol{n}}\right)$, $Z_n \in \mathbb{M}$, respectively, one obtains the following convergence

$$d_{\mathrm{w}}\Big(\mathcal{L}\left(\sqrt{n}d_{BW}(Q_n, Q_*)\right), \mathcal{L}\left(\left\|Q_n^{1/2}\boldsymbol{dT}^{Q_n}_{Q_n}(Z_n)\right\|_F\right)\Big) \to 0,$$

where d_{w} is some metric inducing the weak convergence.

To illustrate the result, we consider the case of a diagonal $Q_* = \operatorname{diag}(q_1, \dots, q_d)$. This setting admits the explicit form of the limiting distribution:

$$\mathcal{L}\left(\sqrt{n}d_{BW}(Q_n, Q_*)\right) \xrightarrow{\mathrm{w}} \mathcal{L}\left(\sqrt{\sum_{i,j=1}^{d}\frac{Z_{ij}^2}{2(q_i + q_j)}}\right),$$

where $Z = (Z_{ij})_{i,j=1}^{d}$. This representation of the limiting distribution is derived in the proof of Corollary 2.1 which is based on the fact that

$$d^2_{BW}(Q_n, Q_*) = -\frac{1+o_P(1)}{2}\left\langle \boldsymbol{dT}^{Q_*}_{Q_*}(Q_n - Q_*), Q_n - Q_* \right\rangle,$$

with $o_P(\cdot)$ being o-small in probability, and an explicit formula for $\boldsymbol{dT}^S_Q$ from Lemma A.2. We discuss the above approximation of $d^2_{BW}(Q_n, Q_*)$ in more detail later in Section 2.4.

The last result concerning convergence of empirical barycenter is the central limit theorem for the empirical variance $\mathcal{V}_n$.

Theorem 2.3 (Central limit theorem for $\mathcal{V}_n$)**.** *Let Assumptions 1 and 2 be fulfilled and* $\mathbb{E}(\operatorname{tr} S)^2 < \infty$*. Then*

$$\sqrt{n}\left(\mathcal{V}_n - \mathcal{V}_*\right) \xrightarrow{\mathrm{w}} \mathcal{N}\left(0, \operatorname{Var} d^2_{BW}(Q_*, S)\right).$$

All proofs are collected in Section A.4. Section 4.1 illustrates the asymptotic behaviour of $\mathcal{L}\left(\sqrt{n}\|Q_n - Q_*\|_F\right)$, $\mathcal{L}\left(\sqrt{n}d_{BW}(Q_n, Q_*)\right)$, and $\mathcal{L}\left(\sqrt{n}|\mathcal{V}_* - \mathcal{V}_n|\right)$.

2.3 Concentration of Q_n and $\mathcal{V}_n$

This section discusses the concentration properties of Q_n under the assumption of sub-Gaussianity of $\mathbb{P}$:

Assumption 3 (Sub-Gaussianity of $\sqrt{\operatorname{tr} S}$). Let $\sqrt{\operatorname{tr} S}$ be sub-Gaussian:

$$\mathbb{P}\left\{\sqrt{\operatorname{tr} S} \geq t\right\} \leq Be^{-bt^2} \quad \text{for any } t \geq 0,$$

with some constants $B, b > 0$.

The first result concerns the concentration of $Q_*^{-1/2} Q_n Q_*^{-1/2}$ in Frobenius norm. This is a crucial step in the proof of concentration of $d_{BW}(Q_n, Q_*)$. From now on we denote the operator norm of a matrix A or an operator $\boldsymbol{A}$ as $\|A\|$, $\|\boldsymbol{A}\|$, respectively. The notations $\lambda_{\min}(A)$, $\lambda_{\min}(\boldsymbol{A})$ denote their smallest eigenvalues.

Theorem 2.4 (Concentration of $Q_*^{-1/2} Q_n Q_*^{-1/2}$ in F-norm)**.** *Let Assumptions 1, 2, and 3 be fulfilled, then*

$$\mathbb{P}\left\{\left\|Q_*^{-1/2} Q_n Q_*^{-1/2} - I\right\|_F \geq \frac{c_Q}{\sqrt{n}}(\sqrt{m} + t)\right\} \leq 2me^{-nt_F} + e^{-t^2/2} + (1-p)^n$$

for any $t \geq 0$ *and* $n \geq c_Q^2(\sqrt{m} + t)^2$*, where*

$$m \overset{\text{def}}{=} \dim(\mathbb{M}), \quad p \overset{\text{def}}{=} \mathbb{P}\big(\mathbb{H}_{++}(d)\big),$$
$$c_Q \overset{\text{def}}{=} \frac{4\|Q_*\|\sigma_T}{\lambda_{\min}(\boldsymbol{F'})}, \quad t_F \overset{\text{def}}{=} \mathtt{C}\min\left(\frac{\lambda_{\min}(\boldsymbol{F'})}{U\log^{1/2}(U/\sigma_F)}, \frac{\lambda^2_{\min}(\boldsymbol{F'})}{\sigma_F^2}\right),$$

where the operator $\boldsymbol{F'}$ *is defined in* (B.3)*, constant* σ_T *comes from auxiliary Proposition B.2, constants* σ_F *and* U *are defined in auxiliary Proposition B.1, and* $\mathtt{C}$ *denotes a generic constant.*

To make the result more transparent, we further discuss it in a less formal way. The proof is based on three steps, and each step yields a bounding term. The first step gives the term $2me^{-nt_F}$. It deals with the concentration of some auxiliary empirical operator $\boldsymbol{F_n'}$ defined in (B.2) in the vicinity of its population counterpart $\boldsymbol{F'}$. These two operators are essentially a price to pay for moving from the space of optimal transportation maps T_Q^S to the space of barycenters. The concentration of $\boldsymbol{F_n'}$ is derived from a result by Koltchinskii [2011] which is presented in Proposition B.1. The constants σ_F and U appear due to this concentration. Some prior bounds on σ_F and U are obtained in Lemma B.3. The second step yields the term $e^{-t^2/2}$. It ensures the concentration of $\left\|\frac{1}{n}\sum_i T_{Q_*}^{S_i} - I\right\|_F$, and relies on the result by Hsu et al. [2012]. To make the text self-contained, we introduce it in Proposition B.2. The constant σ_T comes from a bound on $\left\|\frac{1}{n}\sum_i T_{Q_*}^{S_i} - I\right\|_F$. The last step yields the term $(1-p)^n$. It comes from the requirement on non-degeneracy of Q_n. In other words, a high degeneracy leads to a smaller p and, thus, to worse bounds.

The next result deals with the concentration of $d_{BW}(Q_n, Q_*)$. It is a corollary of the above theorem.

Corollary 2.2 (Concentration of Q_n in d_{BW} distance)**.** *Under the conditions of Theorem 2.4 the following result holds:*

$$\mathbb{P}\left\{d_{BW}(Q_n, Q_*) \geq \tfrac{c_Q\|Q_*\|^{1/2}}{\sqrt{n}}(\sqrt{m}+t)\right\} \leq 2me^{-nt_F} + e^{-t^2/2} + (1-p)^n.$$

The last important result of the current study describes the concentration properties of the empirical Fréchet variance $\mathcal{V}_n$.

Theorem 2.5 (Concentration of $\mathcal{V}_n$)**.** *Let Assumptions 1, 2, and 3 be fulfilled, then, in the notation of Theorem 2.4,*

$$\mathbb{P}\left\{|\mathcal{V}_n - \mathcal{V}_*| \geq z(\mu,\nu,d,n,t)\right\} \leq 2me^{-nt_F} + 3e^{-t^2/2} + (1-p)^n$$

with

$$z(b,\nu,d,n,t) \overset{\text{def}}{=} \max\left(\tfrac{\mu t^2}{n}, \tfrac{\nu t}{\sqrt{n}}\right) + 3\tfrac{c_Q^2\|\boldsymbol{F'}\|}{n}(\sqrt{m}+t)^2.$$

A pair (ν,μ) is the parameters of a sub-exponential r.v. $d_{BW}^2(Q_*, S)$.

All the proofs are collected in Section B.

2.4 Central limit theorem and asymptotic normality of M-estimators

A possible approach to obtain the central limit theorem is to look at a more general result concerning the asymptotic normality of M-estimators. To make the text self-contained, we briefly recall the subject following Section 5.4 in the book by Van De Geer [2006]. Under the setting (1.4), $d_{BW}^2(Q,S)$ might be considered as a loss function parametrized by elements of the affine subspace, $Q \in \mathbb{A} \cap \mathbb{H}_+(d)$. Thus, the proof of the CLT for an empirical barycenter is equivalent to a validation of the following conditions.

(C1) There exists a function $\psi_Q \colon \mathbb{H}_+(d) \to \mathbb{H}(d)$ which is $L_2(\mathbb{P})$-integrable, s.t.

$$\lim_{Q\to Q_*} \frac{|d_{BW}^2(Q,S) - d_{BW}^2(Q_*,S) - \langle \psi_{Q_*}(S), Q - Q_*\rangle|}{\|Q - Q_*\|} = 0.$$

(C2) As $Q \to Q_*$, it holds

$$\int \bigl(d^2_{BW}(Q,S) - d^2_{BW}(Q_*,S)\bigr)\, d\,\mathbb{P}(S) \\ = \frac{1}{2}\,\langle Q - Q_*, \boldsymbol{V}(Q - Q_*)\rangle + o(\|Q - Q_*\|),$$

where $\boldsymbol{V}$ is some positive definite operator.

(C3) Let $Q \neq Q_*$, and define $g_Q(S) \overset{\text{def}}{=} \frac{d^2_{BW}(Q,S) - d^2_{BW}(Q_*,S)}{\|Q - Q_*\|}$. Suppose that for some $\varepsilon > 0$, the class $\{g_Q(S),\ Q : \|Q - Q_*\| \le \varepsilon\}$ has an envelope $G \in L_2(\mathbb{P})$ and that it is a Donsker class.

Lemma A.6 presents differentiability of the Bures–Wasserstein distance and ensures the following quadratic approximation. For any $Q \in \mathbb{H}_{++}(d)$ it holds:

$$-\frac{2}{\Bigl(1+\lambda^{1/2}_{\max}(Q_*^{-1/2} Q Q_*^{-1/2})\Bigr)^2}\,\bigl\langle \boldsymbol{dT}^S_{Q_*}(Q - Q_*), Q - Q_*\bigr\rangle \\ \le d^2_{BW}(Q,S) - d^2_{BW}(Q_*,S) + \langle T^S_{Q_*} - I, Q - Q_*\rangle \\ \le -\frac{2}{\Bigl(1+\lambda^{1/2}_{\min}(Q_*^{-1/2} Q Q_*^{-1/2})\Bigr)^2}\,\bigl\langle \boldsymbol{dT}^S_{Q_*}(Q - Q_*), Q - Q_*\bigr\rangle\,,$$

where $\lambda_{\max}(A)$ and $\lambda_{\min}(A)$ stand for maximal and minimal eigenvalues of a matrix A, respectively. This approximation ensures Condition (C1) and (C2) to be fulfilled with $\psi_{Q_*}(S) = I - T^S_{Q_*}$, and $\boldsymbol{V} = -\,\mathbb{E}\,\boldsymbol{dT}^S_{Q_*}$, respectively. However, it is not clear how one should proceed with the validation of Condition (C3). On the other hand, the direct proof of CLT introduced in the present study is suitable for the proof of the concentration results.

3 Connection to other problems

In this section we explain a connection of obtained results to some other problems. Section 3.1 investigates the relation between the Bures–Wasserstein barycenter and the 2-Wasserstein barycenter of some scale-location family. Section 3.2 illustrates the idea of a barycenter restricted to an affine subspace $\mathbb{A} \subset \mathbb{H}(d)$.

3.1 Connection to scale-location families of measures

We first present the concept of a scale-location family of absolutely continuous measures supported on $\mathbb{R}^d$.

Definition 3.1. Let $X \sim \mu$ be a random variable following a law $\mu \in \mathcal{P}^{ac}_2(\mathbb{R}^d)$, where $\mathcal{P}^{ac}_2(\mathbb{R}^d)$ is the set of absolutely continuous measures with a finite second moment. A set of all affine transformations of X is written as

$$\mathcal{SL}(\mu) \overset{\text{def}}{=} \Bigl\{\mathcal{L}\bigl(PX + p\bigr) : \ P \in \operatorname{Sym}_+(d),\ p \in \mathbb{R}^d\Bigr\}.$$

It is referred to as a scale-location family.

Scale-location families play an important role in modern data analysis and appear in many practical applications due to being user-friendly in terms of theoretical analysis and, at the same time, possessing high modelling power. For example, it is widely used in medical imaging Wassermann et al. [2010], modelling of molecular dynamic Gonzalez et al. [2017], clustering procedures Del Barrio et al. [2017], climate modelling Mallasto and Feragen [2017], embedding of complex objects in low dimensional spaces Muzellec and Cuturi [2018], and so on.

A possible metric that takes into account non-linearity of the underlying data-set is the 2-Wasserstein distance, d_{W_2}. Let μ_X, μ_Y be elements of $\mathcal{SL}(\mu)$, and let random variables sampled from μ_X and μ_Y be $X \sim \mu_X, Y \sim \mu_Y$, respectively. We denote their first and second moments as

$$\mathbb{E}\, X = m_X, \quad \mathbb{E}\, Y = m_Y, \quad \operatorname{Var}(X) = S_X, \quad \operatorname{Var}(Y) = S_Y. \tag{3.1}$$

It is a well-known fact that d_{W_2} between measures coming from the same scale-location family depends only on the first and second moments of the measures:

$$d^2_{W_2}(\mu_X, \mu_Y) = \|m_X - m_Y\|^2 + d^2_{BW}(S_X, S_Y).$$

For more details on a general class of optimal transportation distances we recommend excellent books Ambrosio and Gigli [2013], Villani [2009].

Distribution over a scale-location family In many cases we are interested in data sets coming from some scale-location family. Let $\mathbb{P}$ be a probability measure supported on some $\mathcal{SL}(\mu)$. And let $(\Omega, F, \mathbb{P})$ be a generic probability space, s.t. for any $\omega \in \Omega$ there exists an image $\mu_\omega \stackrel{\text{def}}{=} \mathcal{L}\bigl(P_\omega X + p_\omega\bigr)$, where $P_\omega \in \operatorname{Sym}_+(d)$ is a scaling parameter and $p_\omega \in \mathbb{R}^d$ is a shift parameter. A randomly sampled measure μ_ω belongs to $\mathcal{SL}(\mu)$ by construction, and its first and second moments (m_ω, S_ω) are written as

$$m_\omega \stackrel{\text{def}}{=} P_\omega r + p_\omega, \quad S_\omega \stackrel{\text{def}}{=} P_\omega Q P_\omega^\top,$$

where the pair (r, Q) denote the first and the second moments of the template measure μ. The Fréchet variance of $\mathbb{P}$ at any arbitrary point μ' is written as

$$\mathcal{V}(\mu') \stackrel{\text{def}}{=} \int_{\operatorname{supp}(\mathbb{P})} d^2_{W_2}(\mu', \nu_\omega)\, \mathbb{P}(d\omega).$$

Given an i.i.d. sample $\nu_1, \ldots, \nu_n$ from $\mathbb{P}$, we define the empirical counterpart of $\mathcal{V}(\mu')$:

$$\mathcal{V}_n(\mu') \stackrel{\text{def}}{=} \frac{1}{n} \sum_{i=1}^n d^2_{W_2}(\mu', \nu_i).$$

Then the population and the empirical barycenters μ_* and μ_n are

$$\mu_* = \operatorname*{argmin}_{\mu \in \mathcal{P}_2(\mathbb{R}^d)} \mathcal{V}(\mu), \quad \mu_n = \operatorname*{argmin}_{\mu \in \mathcal{P}_2(\mathbb{R}^d)} \mathcal{V}_n(\mu).$$

Note that μ_* and μ_n belong to $\mathcal{SL}(\mu)$ and are uniquely characterised by their first and second moments, (r_*, Q_*) and (r_n, Q_n), respectively, see Theorem 3.10 in Álvarez-Esteban et al. [2015]:

$$r_* = \int_{\operatorname{supp}(\mathbb{P})} m_\omega\, \mathbb{P}(d\omega), \quad Q_* = \int_{\operatorname{supp}(\mathbb{P})} \left(Q_*^{1/2} S_\omega Q_*^{1/2}\right)^{1/2} \mathbb{P}(d\omega), \tag{3.2}$$

$$r_n = \frac{1}{n} \sum_{i=1}^n m_i, \quad Q_n = \frac{1}{n} \sum_{i=1}^n \left(Q_n^{1/2} S_i Q_n^{1/2}\right)^{1/2}. \tag{3.3}$$

It is worth noting that the concept of Wasserstein barycenter presented originally in a seminal work by Agueh and Carlier [2011] becomes a topic of extensive scientific interest in the last few years. One of the main reasons is an introduction of computationally feasible procedures, see, e.g., Cuturi [2013], Peyré et al. [2019], Dvurechensky et al. [2018], Kroshnin et al. [2019]. There are also many works dedicated to investigation of theoretical properties of barycenters. A work of Bigot and Klein [2012] focuses on the convergence of a parametric class of barycenters, while Bigot et al. [2016] investigate the asymptotic properties of the regularised barycenters. The paper Le Gouic and Loubes [2017] ensures the convergence of the Wasserstein barycenters. To the best of our knowledge, the most state-of-the-art result concerning the rates of convergence of an empirical 2-Wasserstein barycenter is obtained as a particular illustration of a more general result by Le Gouic et al. [2019]. Namely, this work establishes fast rates of convergence for empirical barycenters over a large class of geodesic spaces with curvature bounds in the sense of Alexandrov. This work extends and completes the results by Ahidar-Coutrix et al. [2018]. The latter paper provides the rates of convergence for empirical barycenters of the Borel probability measure on a metric space either under assumptions on weak curvature constraint of the underlying space or for a case of a non-negatively curved space on which geodesics, emanating from a barycenter, can be extended. Corollary 2.1 extends the above results for $d_{W_2}(\mu_n, \mu_*)$ for the case of barycenters constrained to an affine sub-space for measures coming from some scale-location family. The rate is of order $n^{-1/2}$.

The paper Kroshnin [2018] obtains an analogue of the law of large numbers for the case of an arbitrary cost function on some affine sub-space $\mathbb{A}$.

The paper Agueh and Carlier [2017] introduces the CLT for an empirical barycenter for $\mathbb{P}$ supported on a finite set of Gaussian measures. It is worth noting that the idea of the proof also relies on the differentiability of optimal transportation maps.

At this point, it is worth mentioning that there are some other works dealing with the central limit theorem for the Wasserstein distance, e.g., Rippl et al. [2016], Del Barrio and Loubes [2019]. However, the setting in these works differs significantly from what is done in the present study. The paper Rippl et al. [2016] derives the central limit theorems for the p-Wasserstein distance, $p \geq 1$, between empirical distributions sampled from Gaussians supported on $\mathbb{R}^d$. The work Del Barrio and Loubes [2019] establishes the central limit theorem and the variance bounds for the 2-Wasserstein distance between an empirical measure and its true underlying counterpart on $\mathbb{R}^d$. A result, similar in spirit to Theorem 2.5 is obtained in Del Barrio et al. [2016]. However, the authors consider only the space of probability measures supported on the real line, $d = 1$, endowed with 2-Wasserstein distance. To the best of our knowledge, there are no results similar to the concentration Theorem 2.4 and Corollary 2.2 in the case of 2-Wasserstein distance.

3.2 Connection to quantum mechanics

The original Bures metric appears in quantum mechanics in relation to the fidelity measure between two quantum states and is used for the measurement of quantum entanglement Marian and Marian [2008], Dajka et al. [2011]. Let ρ and σ be two density operators. In essence a density matrix ρ is a Hermitian positive semi-definite operator with the unit trace, $\rho \in \mathbb{H}_+(d), \operatorname{tr} \rho = 1$. It is used as a possible way of description of statistical state of a quantum system. For an introduction to the density operators theory one may look Fano [1957]. Let ρ and σ be two quantum states:

$$\rho, \sigma \in \mathbb{H}_+(d), \quad \operatorname{tr} \rho = 1, \quad \operatorname{tr} \sigma = 1. \tag{3.4}$$

Fidelity of these states is defined as $\mathcal{F}(\rho, \sigma) = \left(\operatorname{tr}\sqrt{\rho^{1/2}\sigma\rho^{1/2}}\right)^2$. It quantifies "closeness" of ρ and σ, see Jozsa [1994]. It is obvious, that in case of (3.4) the Bures–Wasserstein distance turns into Bures distance:

$$d_B^2(\rho, \sigma) = 2\left(1 - \mathcal{F}^{1/2}(\rho, \sigma)\right). \tag{3.5}$$

The rest of this section illustrates the idea of the barycenter restricted to some affine sub-space $\mathbb{A}$. Given a random ensemble of density matrices, one is able to recovery its mean using averaging in the Euclidean sense. However, the Bures–Wasserstein barycenter suggests an alternative way to define the barycenter in terms of fidelity measure (3.5). We consider a following statistical setting. Let $(\Omega, F, \mathbb{P})$ be some mechanism which generates quantum states ρ_ω. Given an i.i.d. sample $\rho_1, \ldots, \rho_n$ we write a population and an empirical variance of $\mathbb{P}$ as

$$\mathcal{V}(\sigma) = \int_{\operatorname{supp}(\mathbb{P})} d_{BW}^2(\sigma, \rho_\omega)\, \mathbb{P}(d\omega), \quad \mathcal{V}_n(\sigma) = \frac{1}{n}\sum_{i=1}^n d_{BW}^2(\sigma, \rho_i).$$

Then the population and the empirical barycenters belonging the class of all $d \times d$-dimensional density operators are defined as

$$\rho_* = \operatorname*{argmin}_{\sigma\colon \operatorname{tr}\sigma = 1} \mathcal{V}(\sigma), \quad \rho_n = \operatorname*{argmin}_{\sigma\colon \operatorname{tr}\sigma = 1} \mathcal{V}_n(\sigma).$$

It can be easily shown, that by "taking the global Fréchet barycenter" or, in other words neglecting the condition $\operatorname{tr}\sigma = 1$, we end up with the global barycenter, which is the solution of the fixed point equation which is already mentioned in Section 2: $\rho = \int \left(\rho^{1/2}\rho_\omega\rho^{1/2}\right)^{1/2} \mathbb{P}(d\omega)$. However, this is a contraction mapping. Thus $\operatorname{tr}\rho_* < 1$, and ρ_* is not a density operator. In other words the condition $\operatorname{tr}\sigma = 1$ ensures, that ρ_* and ρ_n also belong to the class of density operators. Taking into account the results obtained in Section 2, ρ_n is in some sense a natural consistent estimator of ρ_* with the known rate of convergence and known deviation properties.

4 Interpolation using empirical BW barycenters

We suggest to use the empirical Bures–Wasserstein barycenters for filling in gaps in data sets consisting of either measures coming from the same scale-location family, or from a family of Hermitian matrices. As a motivation we consider a data set related to the climate dynamics collected in Siberia (Russia) between 1930 and 2009, Bulygina and Razuvaev [2012], Tatusko [1990], where observations for the years $1934, 1938$, 1942, 1948 and 1961 are lost. In this data set a behaviour of some quantities, such as a min/max daily temperatures during a year etc., is modelled using the Gaussian processes which parameters are estimated from the real measurements. We propose to replace the gaps in data with an an empirical barycenter constructed from available observations. To make the illustration more transparent, we consider a toy example for the case of two-dimensional covariance matrices. Each observed matrix is represented graphically by a two-dimensional ellipses. The upper panel at Fig. 1 depicts a family of i.i.d. covariance matrices sampled consecutively over the discrete time t. The eights observation is supposed to be missing. Three lower panels present a possible replacement constructed from two (2-d panel), six (3-d panel), and all available observations (4-th panel). The observations used for data reconstruction are coloured in the dark green. The red ellipses correspond to the Bures–Wasserstein mean, while the blue ones depict the Euclidean mean. The difference in obtained results presented by three lower panels raises a question of a proper choice of number of observations used for missing data completion. Though being very interesting, this question is beyond the scope of the current study. Another question concerns the construction of non-asymptotic confidence sets for the estimators. For instance the work by Ebert et al. [2017] suggests a suitable methodology based on multiplier bootstrap. However it considers only the case of commuting covariance matrices. Thus we consider this question as a matter for further research.

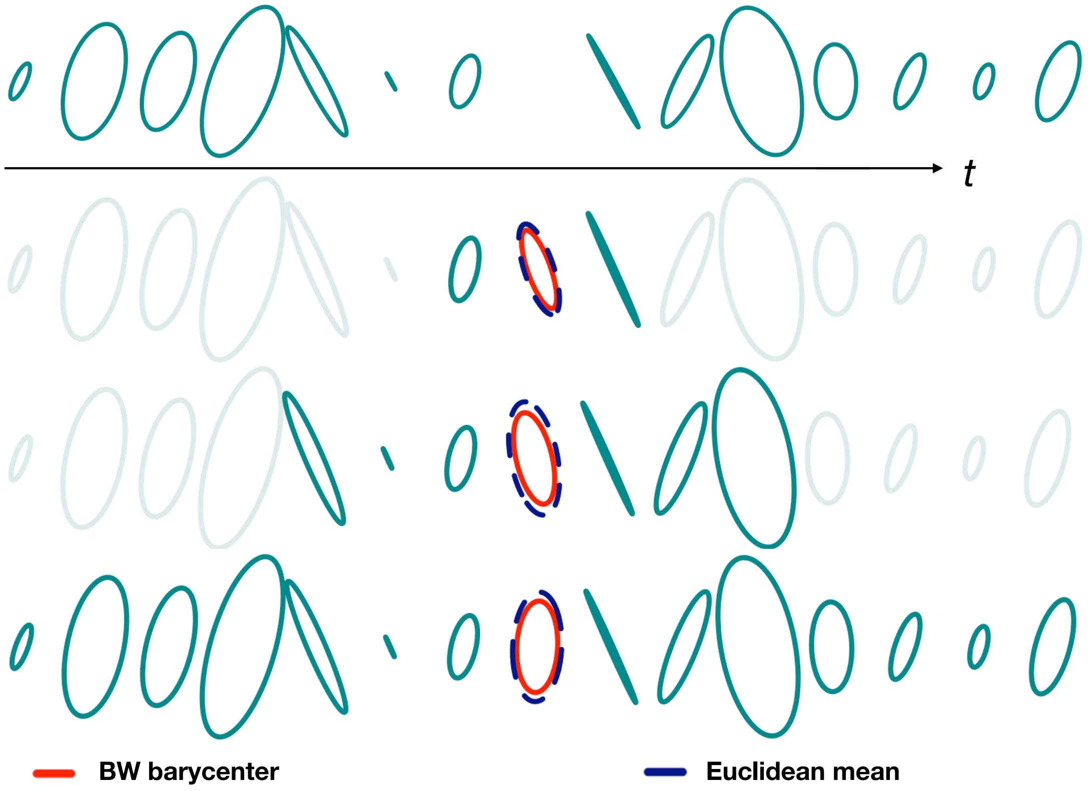

Figure 1: Interpolation of the lost data

The next two section provide some illustrations of the convergence rate of an empirical barycenter to the true one. To make the presentation complete, we also provide an illustration of the convergence of an empirical variance $\mathcal{V}_n$ to $\mathcal{V}_*$.

4.1 Simulated data

In this section we consider a simulated data set. Ccovariance matrices are generated as follows. A matrix $\widetilde{S}_k = A_k A_k^T$ is a d-dimensional matrix, where $A_k = (a_{ij}^k)$, $a_{ij}^k \overset{iid}{\sim} \mathrm{Unif}[0,1] + 1$ for all i, j. To ensure that $\widetilde{S}_k$ is non-degenerated, we consider the orthogonal decomposition $\widetilde{S}_k = \widetilde{U}_k^* \widetilde{\Lambda}_k \widetilde{U}_k$, and replace $\widetilde{\Lambda}_k$ by $\Lambda_k = \mathrm{diag}(\lambda_1^k, \dots, \lambda_d^k)$ s.t. $\lambda_i^k \sim \mathrm{Unif}[18, 22]$. Thus, an observed i.i.d. sample consists of matrices $S_k = \widetilde{U}_k^* \Lambda_k \widetilde{U}_k$, $k = 1, \dots, n$. In what follows, Q_n is a barycenter of the sample $S_1, \dots, S_n$.

Fig. 2 illustrates the convergence of $\mathcal{L}\left(\sqrt{n}\|Q_n - Q_*\|_F\right)$ to $\mathcal{L}(\|Z\|_F)$ with $Z \sim \mathcal{N}(0, \Xi)$ presented in Theorem 2.2. Fig. 3 depicts the convergence of the distribution $\mathcal{L}\left(\sqrt{n} d_{BW}(Q_n, Q_*)\right)$ to $\mathcal{L}\left(\left\|Q_*^{1/2} \boldsymbol{dT}_{Q_*}^{Q_*}(Z)\right\|_F\right)$ obtained in Corollary 2.1. Finally, Fig. 4 illustrates the convergence of density of $\mathcal{L}\left(\sqrt{n}(\mathcal{V}_* - \mathcal{V}_n)\right)$ to the density of the Gaussian distribution $\mathcal{N}\left(0, \mathrm{Var}\left(d_{BW}^2(Q_*, S)\right)\right)$ validated by Theorem 2.3. The numerical experiments are performed using R. The population barycenter Q_* was computed using a sample of 20000 observed covariance matrices. A solid line depicts the density of a respective limiting distributions, while the dashed lines correspond to the densities of $\mathcal{L}\left(\sqrt{n}\|Q_n - Q_*\|_F\right)$, $\mathcal{L}\left(\sqrt{n}d_{BW}(Q_n, Q_*)\right)$, or $\mathcal{L}\left(\sqrt{n}(\mathcal{V}_* - \mathcal{V}_n)\right)$, respectively. We consider different sample sizes for calculation of an empirical Bures–Wasserstein barycenter Q_n with $n \in \{3, 10, 100, 1000\}$. Simulation were carried out for the dimensions $d = 5$ and $d = 10$.

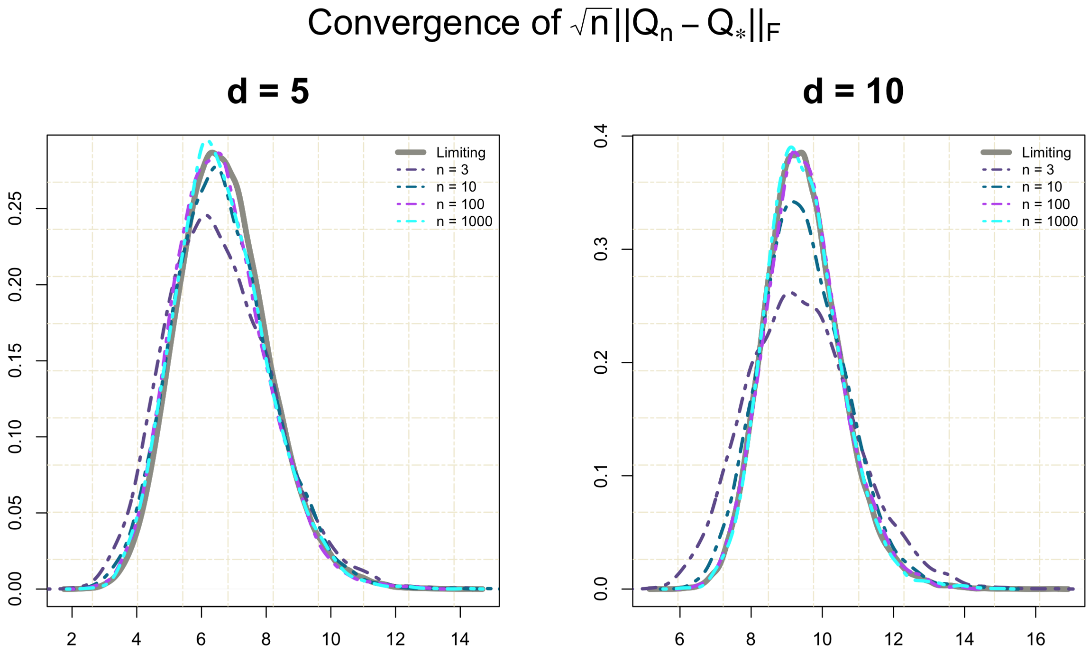

Figure 2: Densities of $\mathcal{L}(\sqrt{n}\|Q_n - Q_*\|_F)$

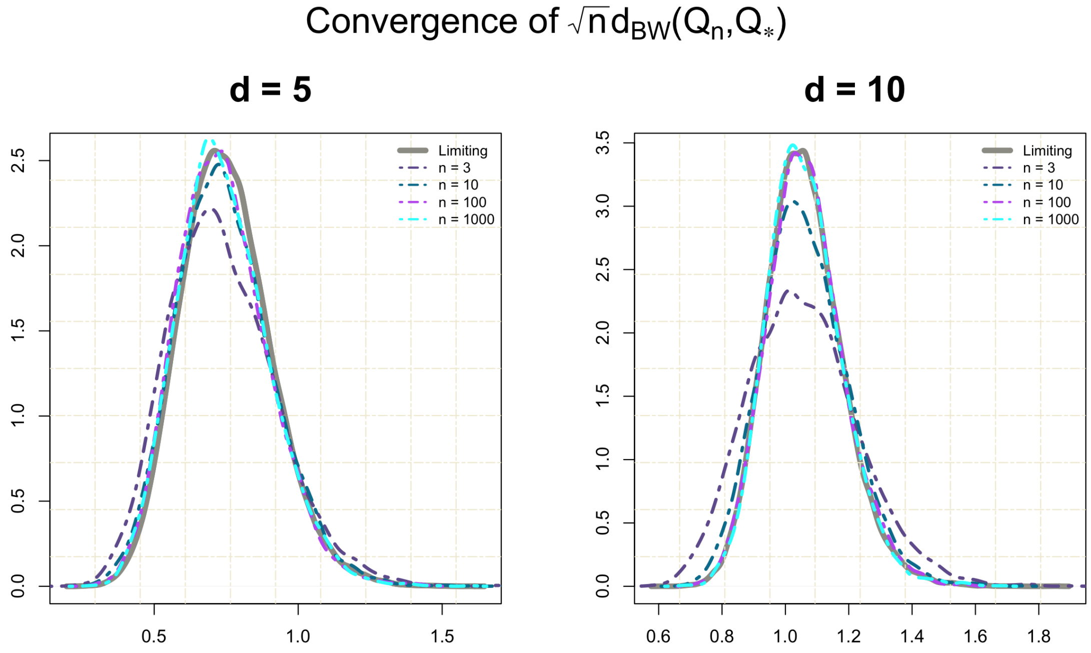

Figure 3: Densities of $\mathcal{L}\left(\sqrt{n}d_{BW}(Q_n, Q_*)\right)$

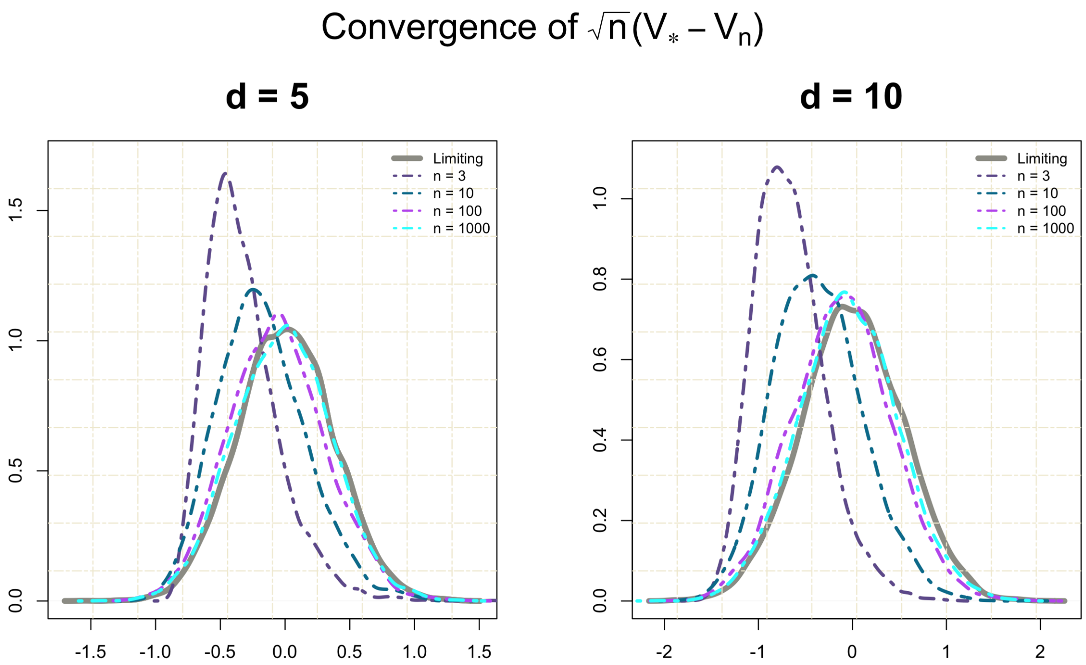

Figure 4: Densities of $\mathcal{L}\left(\sqrt{n}(\mathcal{V}_n - \mathcal{V}_*)\right)$

4.2 Data aggregation in climate modelling

In this section we demonstrate the convergence rates for the climate-related data set. At first, we discuss the set in more detail. Following the original setting, we assume that the daily minimum temperatures within a year is described by a class of Gaussian processes. The temperature is measured at a set of 30 randomly sampled meteorological stations located in Siberia. Each Gaussian curve is obtained through the regression, and the maximum likelihood estimation, and is sampled in 50 points Mallasto and Feragen [2017]. Thus, the observed data set $\mathcal{D}$ consists of 71 Gaussian distributions:

$$\mathcal{D} = \left\{\mathcal{N}(m_t, S_t),\ m_t \in \mathbb{R}^{50},\ S_t \in \mathrm{Sym}_{++}(50),\ t = 1, \ldots, 71\right\},$$

where $\mathcal{N}(m_t, S_t)$ is a Gaussian distribution related to a Gaussian process describing a t-th year, $t = 1933, \ldots, 2009$. The missing years in this data set are 1934, 1938, 1942, 1948, and 1961. This distribution is specified by a mean m_t and a covariance S_t. A Gaussian distribution $\mathcal{N}(r_*, Q_*)$ is the population Wasserstein barycenter of $\mathcal{D}$. It is characterised by the first and the second moments written as (r_*, Q_*)

$$r_* = \frac{1}{71}\sum_{t=1}^{71} m_t, \quad Q_* = \frac{1}{71}\sum_{t=1}^{71}\left(Q_*^{1/2} S_t Q_*^{1/2}\right)^{1/2}.$$

A family of empirical barycenters $\mathcal{N}(r_n, Q_n)$ with parameters (r_n, Q_n) coming from (3.3) is constructed by means of re-sampling with replacement of the original data set.

Fig. 5 and Fig. 6 present the convergence of densities of $\mathcal{L}\left(\sqrt{n}\|Q_n - Q_*\|_F\right)$ to $\mathcal{L}(\|Z\|_F)$ with $Z \sim \mathcal{N}(0, \Xi)$, and $\mathcal{L}\left(\sqrt{n} d_{BW}(Q_n, Q_*)\right)$ to $\mathcal{L}\left(\left\|Q_*^{1/2} \boldsymbol{dT}_{Q_*}^{Q_*}(Z)\right\|_F\right)$, respectively. The limiting distributions are depicted by solid lines, while the dashed ones stand for the densities computed for barycenters of n covariance matrices with $n \in \{3, 10, 70\}$.

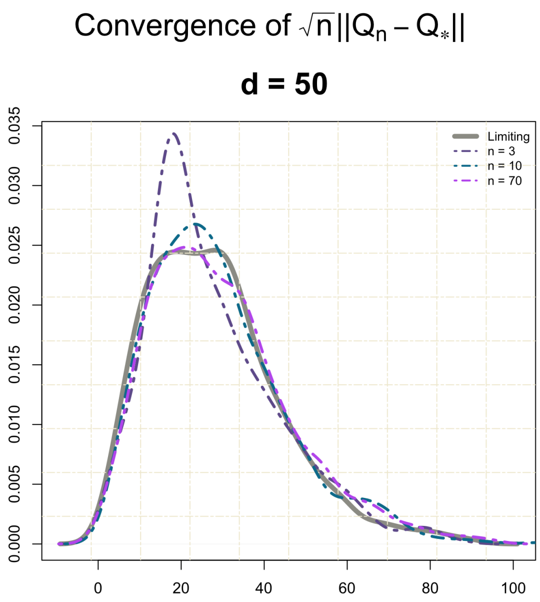

Figure 5: $\mathcal{L}\left(\sqrt{n}\|Q_n - Q_*\|_F\right)$

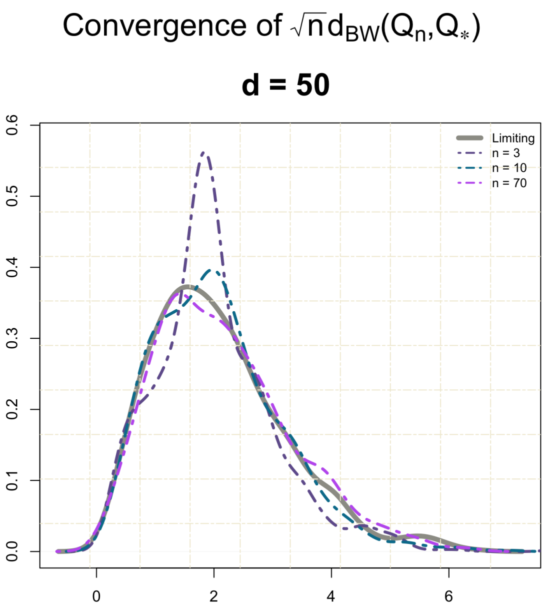

Figure 6: $\mathcal{L}\left(\sqrt{n} d_{BW}(Q_n, Q_*)\right)$

References

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

Martial Agueh and Guillaume Carlier. Vers un théorème de la limite centrale dans l'espace de Wasserstein? *Comptes Rendus Mathématique*, 355(7):812–818, 2017.

Adil Ahidar-Coutrix, Thibaut Le Gouic, and Quentin Paris. On the rate of convergence of empirical barycentres in metric spaces: curvature, convexity and extendible geodesics. *arXiv preprint arXiv:1806.02740*, 2018.

Pedro C. Álvarez-Esteban, Eustasio Del Barrio, Juan A. Cuesta-Albertos, and C. Matrán. Wide consensus for parallelized inference. *ArXiv e-prints*, November 2015.

Pedro C. Álvarez-Esteban, Eustasio Del Barrio, Juan A. Cuesta-Albertos, and Carlos Matrán. Wide consensus aggregation in the Wasserstein space. application to location-scatter families. *Bernoulli*, 24(4A):3147–3179, 2018.

Luigi Ambrosio and Nicola Gigli. A users guide to optimal transport. In *Modelling and optimisation of flows on networks*, pages 1–155. Springer, 2013.

Rajendra Bhatia, Tanvi Jain, and Yongdo Lim. On the Bures–Wasserstein distance between positive definite matrices. *Expositiones Mathematicae*, 2018.

Jérémie Bigot and Thierry Klein. Consistent estimation of a population barycenter in the Wasserstein space. *ArXiv e-prints*, 2012.

Jérémie Bigot, Elsa Cazelles, and Nicolas Papadakis. Penalized barycenters in the Wasserstein space. *arXiv e-prints*, art. arXiv:1606.01025, Jun 2016.

Stephen Boyd and Lieven Vandenberghe. *Convex optimization*. Cambridge university press, 2004.

Yann Brenier. Polar factorization and monotone rearrangement of vector-valued functions. *Communications on pure and applied mathematics*, 44(4):375–417, 1991.

ON Bulygina and VN Razuvaev. Daily temperature and precipitation data for 518 Russian meteorological stations. Technical report, ESS-DIVE (Environmental System Science Data Infrastructure for a Virtual Ecosystem); Oak Ridge National Lab.(ORNL), Oak Ridge, TN (United States), 2012.

Brittny Calsbeek and Charles J Goodnight. Empirical comparison of G matrix test statistics: finding biologically relevant change. *Evolution: International Journal of Organic Evolution*, 63(10):2627–2635, 2009.

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.

Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. In *Advances in neural information processing systems*, pages 2292–2300, 2013.

Jerzy Dajka, Jerzy Łuczka, and Peter Hänggi. Distance between quantum states in the presence of initial qubit-environment correlations: A comparative study. *Physical Review A*, 84(3):032120, 2011.

Eustasio Del Barrio and Jean-Michel Loubes. Central limit theorems for empirical transportation cost in general dimension. *The Annals of Probability*, 47(2):926–951, 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.

Eustasio Del Barrio, Paula Gordaliza, Hélène Lescornel, and Jean-Michel Loubes. Central limit theorem and bootstrap procedure for Wasserstein's variations with application to structural relationships between distributions. *ArXiv e-prints*, November 2016.

Eustasio Del Barrio, Juan A. Cuesta-Albertos, Carlos Matrán, and Agustín Mayo-Íscar. Robust clustering tools based on optimal transportation. *Statistics and Computing*, pages 1–22, 2017.

Pavel Dvurechensky, Darina Dvinskikh, Alexander Gasnikov, Csar A. Uribe, and Angelia Nedić. Decentralize and randomize: Faster algorithm for Wasserstein barycenters. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems 31*, NeurIPS 2018, pages 10783–10793. Curran Associates, Inc., 2018. arXiv:1806.03915.

Johannes Ebert, Vladimir Spokoiny, and Alexandra Suvorikova. Construction of non-asymptotic confidence sets in 2-Wasserstein space. *arXiv preprint arXiv:1703.03658*, 2017.

Ugo Fano. Description of states in quantum mechanics by density matrix and operator techniques. *Reviews of Modern Physics*, 29(1):74, 1957.

Rémi Flamary, Marco Cuturi, Nicolas Courty, and Alain Rakotomamonjy. Wasserstein discriminant analysis. *Machine Learning*, 107(12):1923–1945, 2018.

Oscar Gonzalez, Marco Pasi, Daiva Petkevičiūtė, Jaroslaw Glowacki, and JH Maddocks. Absolute versus relative entropy parameter estimation in a coarse-grain model of DNA. *Multiscale Modeling & Simulation*, 15(3):1073–1107, 2017.

Charles J Goodnight and James M Schwartz. A bootstrap comparison of genetic covariance matrices. *Biometrics*, pages 1026–1039, 1997.

Alexandre Gramfort, Gabriel Peyré, and Marco Cuturi. Fast optimal transport averaging of neuroimaging data. In *International Conference on Information Processing in Medical Imaging*, pages 261–272. Springer, 2015.

Daniel Hsu, Sham Kakade, and Tong Zhang. A tail inequality for quadratic forms of subgaussian random vectors. *Electronic Communications in Probability*, 17, 2012.

Richard Jozsa. Fidelity for mixed quantum states. *Journal of modern optics*, 41(12):2315–2323, 1994.

Vladimir Koltchinskii. Von Neumann entropy penalization and low-rank matrix estimation. *The Annals of Statistics*, 39(6):2936–2973, 2011.

Alexey Kroshnin. Fréchet barycenters in the Monge–Kantorovich spaces. *Journal of Convex Analysis*, 25(4):1371–1395, 2018.

Alexey Kroshnin, Nazarii Tupitsa, Darina Dvinskikh, Pavel Dvurechensky, Alexander Gasnikov, and Cesar Uribe. On the complexity of approximating Wasserstein barycenters. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 3530–3540, Long Beach, California, USA, 09–15 Jun 2019. PMLR. arXiv:1901.08686.

Thibaut Le Gouic and Jean-Michel Loubes. Existence and consistency of Wasserstein barycenters. *Probability Theory and Related Fields*, 168(3-4):901–917, 2017.

Thibaut Le Gouic, Quentin Paris, Philippe Rigollet, and Austin J Stromme. Fast convergence of empirical barycenters in Alexandrov spaces and the Wasserstein space. *arXiv preprint arXiv:1908.00828*, 2019.

Anton Mallasto and Aasa Feragen. Learning from uncertain curves: The 2-Wasserstein metric for Gaussian processes. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 5660–5670. Curran Associates, Inc., 2017.

Paulina Marian and Tudor A Marian. Bures distance as a measure of entanglement for symmetric two-mode Gaussian states. *Physical Review A*, 77(6):062319, 2008.

Grégoire Montavon, Klaus-Robert Müller, and Marco Cuturi. Wasserstein training of restricted Boltzmann machines. In *Advances in Neural Information Processing Systems*, pages 3718–3726, 2016.

Boris Muzellec and Marco Cuturi. Generalizing point embeddings using the Wasserstein space of elliptical distributions. *arXiv preprint arXiv:1805.07594*, 2018.

Ingram Olkin and Friedrich Pukelsheim. The distance between two random vectors with given dispersion matrices. *Linear Algebra and its Applications*, 48:257–263, 1982.

Gabriel Peyré, Marco Cuturi, et al. Computational optimal transport. *Foundations and Trends® in Machine Learning*, 11(5-6):355–607, 2019.

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

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

Rene Tatusko. Cooperation in climate research: An evaluation of the activities conducted under the US–USSR agreement for environmental protection since 1974. 1990.

César A Uribe, Darina Dvinskikh, Pavel Dvurechensky, Alexander Gasnikov, and Angelia Nedić. Distributed computation of Wasserstein barycenters over networks. In *2018 IEEE Conference on Decision and Control (CDC)*, pages 6544–6549. IEEE, 2018.

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

Cdric 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.

Demian Wassermann, Luke Bloy, Efstathios Kanterakis, Ragini Verma, and Rachid Deriche. Unsupervised white matter fiber clustering and tract probability map generation: Applications of a Gaussian process framework for white matter fibers. *NeuroImage*, 51(1):228–241, 2010.

A Proof of Central Limit Theorem

A.1 List of accepted notations

To make the presentation more transparent, we introduce a list of some used notations.

A, B	Matrices or vectors
$\boldsymbol{A}, \boldsymbol{B}$	Operators
$\lambda_{\max}(\square), \lambda_{\min}(\square)$	Largest and smallest eigenvalue of an operator or a matrix
$(\square)_{\mathbb{M}}$	Restriction of a quadratic form to a subspace $\mathbb{M}$
$\lVert\square\rVert$	Operator norm
$\lVert\square\rVert_F$	Frobenius norm
$\lVert\square\rVert_1$	Schatten norm
$\lVert\square\rVert_{\psi_1}$	ψ_1 Orlicz norm
$\lVert\square\rVert_{\psi_2}$	ψ_2 Orlicz norm
$\kappa(\square) = \lVert\square\rVert \cdot \lVert\square^{-1}\rVert$	Condition number of an operator or a matrix
$\langle\square, \square\rangle$	Inner product associated to Frobenius norm
$\otimes$	Tensor product
$\mathcal{L}(X)$	Distribution of a r.v. X
$\xrightarrow{\mathrm{w}}$	Weak convergence
$\xrightarrow{\mathrm{a.s.}}$	A.s. convergence
$d_{\mathrm{w}}(\square, \square)$	Metric inducing weak convergence
$o_P(\square)$	O-small in probability
$O_P(\square)$	O-big in probability

A.2 Properties of T_Q^S

Proof of Proposition 2.1. First, we prove that optimal T is self-adjoint. Indeed, assume the opposite, then

$$Q^{1/2}TQT^*Q^{1/2} = \left(Q^{1/2}TQ^{1/2}\right)\left(Q^{1/2}TQ^{1/2}\right)^* = Q^{1/2}SQ^{1/2}$$

and thus $\operatorname{tr} Q^{1/2}TQ^{1/2} < \operatorname{tr}\left(Q^{1/2}SQ^{1/2}\right)^{1/2}$. Therefore

$$\begin{aligned}\operatorname{tr}(T-I)Q(T^*-I) &= \operatorname{tr} S + \operatorname{tr} Q - 2\operatorname{tr} TQ = \operatorname{tr} S + \operatorname{tr} Q - 2\operatorname{tr} Q^{1/2}TQ^{1/2}\\ &> \operatorname{tr} S + \operatorname{tr} Q - 2\operatorname{tr}\left(Q^{1/2}SQ^{1/2}\right)^{1/2} = d^2_{BW}(Q,S).\end{aligned}$$

If T is Hermitian but not positive semi-definite, then $Q^{1/2}TQ^{1/2} \preccurlyeq \left(Q^{1/2}SQ^{1/2}\right)^{1/2}$, $Q^{1/2}TQ^{1/2} \neq \left(Q^{1/2}SQ^{1/2}\right)^{1/2}$, hence again $\operatorname{tr} Q^{1/2}TQ^{1/2} < \operatorname{tr}\left(Q^{1/2}SQ^{1/2}\right)^{1/2}$.

Finally, if $T \in \mathbb{H}_+(d)$, then it is straightforward to check that $T = T_Q^S$ given by (2.1) and

$$\operatorname{tr}(T-I)Q(T^*-I) = \operatorname{tr} S + \operatorname{tr} Q - 2\operatorname{tr}\left(Q^{1/2}SQ^{1/2}\right)^{1/2} = d^2_{BW}(Q,S).$$

□

The proof of the Central Limit Theorem mainly relies on the differentiability of the map (2.1). Lemma A.2 shows that T_Q^S can be linearised in the vicinity of Q:

$$T^S_{Q+X} = T^S_Q + \boldsymbol{dT}^S_Q(X) + o\big(\|X\|\big),$$

where $\boldsymbol{dT}^S_Q\colon \mathbb{H}(d) \to \mathbb{H}(d)$ is a self-adjoint negative-definite operator and $\|X\|$ stands for an operator norm of X. Properties of $\boldsymbol{dT}^S_Q$ are investigated in Lemma A.3. Let us introduce some notation: if $G(A)$ is a functional of a matrix A, then we denote its differential as $od_A G$.

Lemma A.1. *Map $Q \mapsto g(Q) = Q^{1/2}$ is differentiable on $\mathbb{H}_{++}(d)$, and its differential is given by*

$$\boldsymbol{d_Q}g(X) = U^*\left(\frac{(UXU^*)_{ij}}{\sqrt{q_i}+\sqrt{q_j}}\right)_{i,j=1}^d U, \quad X \in \mathbb{H}(d),$$

where $Q = U^*\operatorname{diag}(q)U$ is the eigenvalue decomposition.

Proof. First, let us consider the map $P \mapsto f(P) = P^2$. It is smooth and its differential

$$\boldsymbol{d_P}f(X) = PX + XP, \quad X \in \mathbb{H}(d)$$

is non-degenerated:

$$\langle \boldsymbol{d_P}f(X), X\rangle = 2\operatorname{tr} XPX > 0, \quad X \neq 0,$$

whenever $P \in \mathbb{H}_{++}(d)$. From now on $\langle\cdot,\cdot\rangle$ denotes a scalar product associated to Frobenius norm.

Now applying the inverse function theorem we obtain that the inverse map $g(\cdot)$ is also smooth and its differential enjoys the following equation

$$X = \left(\boldsymbol{d_P}f|_{P=Q^{1/2}}\right)\left(\boldsymbol{d_Q}g(X)\right) = Q^{1/2}\boldsymbol{d_Q}g(X) + \boldsymbol{d_Q}g(X)Q^{1/2},$$

thus

$$UXU^* = (\mathrm{diag}(q))^{1/2} U \boldsymbol{d_Q} g(X) U^* + U \boldsymbol{d_Q} g(X) U^* (\mathrm{diag}(q))^{1/2},$$

$$(UXU^*)_{ij} = (\sqrt{q_i} + \sqrt{q_j})(U \boldsymbol{d_Q} g(X) U^*)_{ij}, \quad 1 \le i, j \le d,$$

and

$$\boldsymbol{d_Q} g(X) = U^* \left(\frac{(UXU^*)_{ij}}{\sqrt{q_i} + \sqrt{q_j}} \right)_{i,j=1}^{d} U.$$

□

Lemma A.2 (Fréchet-differentiability of the map T^S_Q)**.** *For any $S \in \mathbb{H}_+(d)$ the map T^S_Q can be linearised in the vicinity of $Q \in \mathbb{H}_{++}(d)$ as*

$$T^S_{\widetilde{Q}} = T^S_Q + \boldsymbol{dT}^S_Q \left(\widetilde{Q} - Q \right) + o\left(\left\| \widetilde{Q} - Q \right\| \right), \quad \textit{as} \quad \widetilde{Q} \to Q,$$

where

$$\boldsymbol{dT}^S_Q(X) \stackrel{\text{def}}{=} -S^{1/2} U^* \varLambda^{-1/2} \delta \varLambda^{-1/2} U S^{1/2}, \quad X \in \mathbb{H}(d), \tag{A.1}$$

$U^* \varLambda U$ is an eigenvalue decomposition of $S^{1/2} Q S^{1/2}$

$$U^* \varLambda U = S^{1/2} Q S^{1/2}, \quad U^* U = U U^* = I, \quad \varLambda = \mathrm{diag}\left(\lambda_1, \ldots, \lambda_{\mathrm{rank}(S)}, 0, \ldots, 0 \right),$$

$$\varLambda^{-1/2} = \left(\varLambda^{1/2} \right)^+ = \mathrm{diag}(\lambda_1^{-1/2}, \ldots, \lambda_{\mathrm{rank}(S)}^{-1/2}, 0, \ldots, 0),$$

$$\delta = (\delta_{ij})_{i,j=1}^d, \quad \delta_{ij} = \begin{cases} \frac{\varDelta_{ij}}{\sqrt{\lambda_i} + \sqrt{\lambda_j}}, & i, j \le \mathrm{rank}(S) \\ 0, & \textit{otherwise} \end{cases}, \quad \varDelta = U S^{1/2} X S^{1/2} U^*.$$

Proof. The proof mainly relies on the differentiation of the pseudo-inverse term $\left(S^{1/2} Q S^{1/2} \right)^{-1/2}$, as long as

$$\boldsymbol{dT}^S_Q(X) = S^{1/2} \boldsymbol{d_Q} \left(S^{1/2} Q S^{1/2} \right)^{-1/2} (X) S^{1/2}.$$

Obviously we can consider only restriction to $\mathrm{range}(S)$ and therefore assume w.l.o.g. $S \succ 0$. As $\left(S^{1/2} (Q + X) S^{1/2} \right)^{-1/2} = U^* \left(\varLambda + \varDelta \right)^{-1/2} U$, by Lemma A.1 and von Neumann series expansion we obtain for infinitesimal $X \in \mathbb{H}(d)$ and corresponding $\varDelta$ that

$$\begin{aligned} (\varLambda + \varDelta)^{-1/2} &= \left(\varLambda^{1/2} + \delta + o(\|\varDelta\|) \right)^{-1} \\ &= \left(\varLambda^{1/4} \left(I + \varLambda^{-1/4} \delta \varLambda^{-1/4} + o(\|\varDelta\|) \right) \varLambda^{1/4} \right)^{-1} \\ &= \varLambda^{-1/4} \left(I - \varLambda^{-1/4} \delta \varLambda^{-1/4} + o(\|\varDelta\|) \right) \varLambda^{-1/4} \\ &= \varLambda^{-1/2} - \varLambda^{-1/2} \delta \varLambda^{-1/2} + o(\|\varDelta\|). \end{aligned}$$

Then the differential $\boldsymbol{d_Q} \left(S^{1/2} Q S^{1/2} \right)^{-1/2} (X)$ is written as

$$\boldsymbol{d_Q} \left(S^{1/2} Q S^{1/2} \right)^{-1/2} (X) = -U^* \varLambda^{-1/2} \delta \varLambda^{-1/2} U.$$

Therefore,

$$T^S_{Q+X} = T^S_Q + \boldsymbol{dT}^S_Q(X) + o(\|X\|),$$

where $\boldsymbol{dT}^S_Q(X)$ is defined by (A.1). □

Lemmas A.3 and A.4 are technical and explore properties of $\boldsymbol{dT}_Q^S$.

Lemma A.3. *For any $S \in \mathbb{H}_+(d)$, $Q \in \mathbb{H}_{++}(d)$, the properties of operator $\boldsymbol{dT}_Q^S$ defined in* (A.1) *are following:*

(I) it is self-adjoint;

(II) it is negative semi-definite;

(III) it enjoys the following bounds:

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

(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 $\operatorname{range}(S_0) = \operatorname{range}(S_1)$; in particular, $\boldsymbol{dT}_Q^S$ is monotone w.r.t. $Q \in \mathbb{H}_{++}(d)$ for fixed S.

Proof. Slightly changing notations, we rewrite (A.1) as

$$\boldsymbol{dT}_Q^S(X) = -S^{1/2}U^*\mathit{\Lambda}^{-1/2}\delta^X\mathit{\Lambda}^{-1/2}US^{1/2},$$

where matrices U and $\mathit{\Lambda}$ come from Lemma A.2 and

$$\delta^X = (\delta_{ij}^X)_{i,j=1}^d, \quad \delta_{ij}^X = \frac{\mathit{\Delta}_{ij}^X}{\sqrt{\lambda_i} + \sqrt{\lambda_j}}, \quad \mathit{\Delta}^X = US^{1/2}XS^{1/2}U^*.$$

(I) Self-adjointness

Consider a scalar product

$$\begin{aligned}\langle \boldsymbol{dT}_Q^S(X), Y\rangle = \operatorname{tr}\bigl(\boldsymbol{dT}_Q^S(X)Y\bigr) &= -\operatorname{tr}\bigl(S^{1/2}U^*\mathit{\Lambda}^{-1/2}\delta^X\mathit{\Lambda}^{-1/2}US^{1/2}Y\bigr)\\ &= -\operatorname{tr}\bigl(\mathit{\Lambda}^{-1/2}\delta^X\mathit{\Lambda}^{-1/2}US^{1/2}YS^{1/2}U^*\bigr).\end{aligned}$$

We now introduce a following notation

$$\mathit{\Delta}^Y \overset{\text{def}}{=} US^{1/2}YS^{1/2}U^*.$$

Then the above equality can be continued as follows:

$$\begin{aligned}-\operatorname{tr}\bigl(\mathit{\Lambda}^{-1/2}\delta^X\mathit{\Lambda}^{-1/2}US^{1/2}YS^{1/2}U^*\bigr) &= -\operatorname{tr}\bigl(\mathit{\Lambda}^{-1/2}\delta^X\mathit{\Lambda}^{-1/2}\mathit{\Delta}^Y\bigr)\\ = -\sum_{i,j=1}^r \frac{\delta_{ij}^X}{\sqrt{\lambda_i\lambda_j}}\mathit{\Delta}_{ij}^Y &= -\sum_{i,j=1}^r \frac{\mathit{\Delta}_{ij}^X\mathit{\Delta}_{ij}^Y}{\sqrt{\lambda_i\lambda_j}(\sqrt{\lambda_i} + \sqrt{\lambda_j})}\\ = \operatorname{tr}\bigl(\boldsymbol{dT}_Q^S(Y)X\bigr) &= \operatorname{tr}\bigl(X\boldsymbol{dT}_Q^S(Y)\bigr) = \langle X, \boldsymbol{dT}_Q^S(Y)\rangle,\end{aligned}$$

where $r = \operatorname{rank}(S)$. Thus the operator is self-adjoint.

(II) Boundedness and (III) eigenvalues

Denoting $\varDelta^X$ by $\varDelta$ (i.e. now $\varDelta = US^{1/2}XS^{1/2}U^*$) and taking into account the above expansion of an inner product, one obtains

$$-\langle \boldsymbol{dT}_Q^S(X), X\rangle = \sum_{i,j=1}^{r} \frac{\varDelta_{ij}^2}{\sqrt{\lambda_i\lambda_j}(\sqrt{\lambda_i}+\sqrt{\lambda_j})} = \sum_{i,j=1}^{r} \left(\frac{\varDelta_{ij}}{\sqrt{\lambda_i\lambda_j}}\right)^2 \frac{\sqrt{\lambda_i\lambda_j}}{\sqrt{\lambda_i}+\sqrt{\lambda_j}}. \tag{A.2}$$

Note, that the function $f(\lambda_i,\lambda_j) \overset{\text{def}}{=} \frac{\sqrt{\lambda_i\lambda_j}}{\sqrt{\lambda_i}+\sqrt{\lambda_j}}$ is monotonously increasing in both arguments λ_i and λ_j, thus

$$\max_{i,j} f(\lambda_i,\lambda_j) = \frac{\lambda_{\max}^{1/2}(\varLambda)}{2}, \quad \min_{i,j} f(\lambda_i,\lambda_j) = \frac{\lambda_{\min}^{1/2}(\varLambda)}{2}. \tag{A.3}$$

For the sake of simplicity we introduce a new variable

$$\zeta \overset{\text{def}}{=} Q^{-1/2}XQ^{-1/2},$$

its Frobenius norm is written as

$$\|\zeta\|_F^2 = \operatorname{tr}\bigl(XQ^{-1}XQ^{-1}\bigr).$$

Moreover, the following inequality for trace holds:

$$\begin{aligned} \operatorname{tr}\bigl(XQ^{-1}XQ^{-1}\bigr) &\geq \operatorname{tr}\bigl(\varPi_S X\varPi_S Q^{-1}\varPi_S X\varPi_S Q^{-1}\varPi_S\bigr) \\ &= \operatorname{tr}\bigl(\varDelta\varLambda^{+}\varDelta\varLambda^{+}\bigr) = \bigl\|\varLambda^{-1/2}\varDelta\varLambda^{-1/2}\bigr\|_F^2 = \sum_{i,j=1}^{r} \frac{\varDelta_{ij}^2}{\lambda_i\lambda_j}. \end{aligned}$$

Here $\varPi_S$ is the orthogonal projector onto the range of S.

Then combining (A.2) with (A.3), the upper and lower bounds can be obtained as follows:

$$\begin{aligned} -\langle \boldsymbol{dT}_Q^S(X), X\rangle &\leq \max_{i,j} f(\lambda_i,\lambda_j) \sum_{i,j=1}^{r} \left(\frac{\varDelta_{ij}}{\sqrt{\lambda_i\lambda_j}}\right)^2 \leq \frac{\lambda_{\max}^{1/2}(\varLambda)}{2}\|\zeta\|_F^2, \\ -\langle \boldsymbol{dT}_Q^S(X), X\rangle &\geq \min_{i,j} f(\lambda_i,\lambda_j) \sum_{i,j=1}^{r} \left(\frac{\varDelta_{ij}}{\sqrt{\lambda_i\lambda_j}}\right)^2 = \frac{\lambda_{\min}^{1/2}(\varLambda)}{2}\|\zeta\|_F^2. \end{aligned}$$

Note, that if S is degenerated, the lower bound becomes trivial.

(IV) Homogeneity and (V) monotonicity

Homogeneity follows directly from definition (A.1). Now we prove monotonicity. As range of $S^{1/2}QS^{1/2}$ is fixed, we may assume $S \succ 0$. Consider

$$\begin{aligned} \langle \boldsymbol{dT}_Q^S(X), X\rangle &= \operatorname{tr}\left(S^{1/2}U^*\varLambda^{-1/2}\delta\varLambda^{-1/2}US^{1/2}, X\right) \\ &= \left\langle U^*\varLambda^{-1/2}\delta\varLambda^{-1/2}U, S^{1/2}XS^{1/2}\right\rangle \\ &= \left\langle \boldsymbol{d_Q}\left(S^{1/2}QS^{1/2}\right)^{-1/2}(X), S^{1/2}XS^{1/2}\right\rangle \\ &= \left\langle \boldsymbol{d_M} M^{-1/2}\left(S^{1/2}XS^{1/2}\right), S^{1/2}XS^{1/2}\right\rangle, \end{aligned}$$

with replacement $M = S^{1/2}QS^{1/2}$ to be change of variables. As long as X is supposed to be fixed, it is enough to show that the differential $\boldsymbol{d_M} M^{-1/2}$ is monotone in M. Notice that the operator $\left(\boldsymbol{d_M} M^{-1/2}\right)^{-1}$ at point M is equal to the differential of the inverse map $P \mapsto P^{-2}$ at point $P = M^{-1/2}$:

$$\boldsymbol{d_M} M^{-1/2} = \left(\boldsymbol{d_P} P^{-2}\big|_{P=M^{-1/2}}\right)^{-1}.$$

In turn, $\boldsymbol{d_P} P^{-2}$ can be expressed as

$$\boldsymbol{d_P} P^{-2}(X) = -P^{-1}\left(P^{-1}X + XP^{-1}\right)P^{-1},$$

the right part of the above equation is self-adjoint, negative-definite and

$$\left\langle -P^{-1}\left(P^{-1}X + XP^{-1}\right)P^{-1}, X\right\rangle = -2\operatorname{tr} P^{-2}XP^{-1}X.$$

Choose $M_1 \succcurlyeq M_0 \succ 0$ (thus $M_1^{1/2} \succcurlyeq M_0^{1/2}$) and let $P_i = M_i^{-1/2}$ for $i = 0, 1$. Then for any fixed $X \in \mathbb{H}(d)$

$$-\operatorname{tr} P_1^{-2}XP_1^{-1}X = -\operatorname{tr} M_1XM_1^{1/2}X \le -\operatorname{tr} M_0XM_0^{1/2}X = -\operatorname{tr} P_0^{-2}XP_0^{-1}X,$$

i.e. $\boldsymbol{d_P} P^{-2}|_{P_1} \preccurlyeq \boldsymbol{d_P} P^{-2}|_{P_0}$ and hence for the differential of $M \mapsto M^{-1/2}$ the inverse inequality holds: $\boldsymbol{d_M} M^{-1/2}\big|_{M_0} \preccurlyeq \boldsymbol{d_M} M^{-1/2}\big|_{M_1}$. This entails monotonicity of $\boldsymbol{dT}_Q^S$. □

Corollary A.1. *Under conditions of Lemma A.3, it holds*

$$\lambda_{\max}(-\boldsymbol{dT}_Q^S) \le \frac{\lambda_{\max}^{1/2}(S^{1/2}QS^{1/2})}{2\lambda_{\min}^2(Q)}, \quad \lambda_{\min}(-\boldsymbol{dT}_Q^S) \ge \frac{\lambda_{\min}^{1/2}(S^{1/2}QS^{1/2})}{2\lambda_{\max}^2(Q)}.$$

Proof. Item (III) from the above lemma ensures that

$$-\left\langle \boldsymbol{dT}_Q^S(X), X\right\rangle \le \frac{\lambda_{\max}^{1/2}\left(S^{1/2}QS^{1/2}\right)}{2}\left\|Q^{-1/2}XQ^{-1/2}\right\|_F^2 \le \frac{\lambda_{\max}^{1/2}(S^{1/2}QS^{1/2})}{2\lambda_{\min}^2(Q)}\|X\|_F^2.$$

The second bound is proved in a similar way. □

Corollary A.2. *We define a following rescaled operator*

$$\boldsymbol{dt}_Q^S(\zeta) \overset{\text{def}}{=} Q^{1/2}\boldsymbol{dT}_Q^S\left(Q^{1/2}\zeta Q^{1/2}\right)Q^{1/2}, \quad \zeta \in \mathbb{H}(d). \tag{A.4}$$

Then a following bound on its eigenvalues hold:

$$\lambda_{\min}\left(-\boldsymbol{dt}_Q^S\right) = \frac{1}{2}\lambda_{\min}^{1/2}\left(S^{1/2}QS^{1/2}\right),$$
$$\lambda_{\max}\left(-\boldsymbol{dt}_Q^S\right) = \frac{1}{2}\lambda_{\max}^{1/2}\left(S^{1/2}QS^{1/2}\right).$$

Proof. Notice that inequalities

$$\lambda_{\min}\left(-\boldsymbol{dt}_Q^S\right) \ge \frac{1}{2}\lambda_{\min}^{1/2}\left(S^{1/2}QS^{1/2}\right),$$
$$\lambda_{\max}\left(-\boldsymbol{dt}_Q^S\right) \le \frac{1}{2}\lambda_{\max}^{1/2}\left(S^{1/2}QS^{1/2}\right),$$

are a trivial consequence of Lemma A.3 (III). Now defining for any $1 \le k \le \operatorname{rank}(S)$

$$\varDelta^k_{ij} = \begin{cases} 1, & i = j = k, \\ 0, & \text{otherwise}, \end{cases}, \quad X^k = S^{-1/2} U \varDelta^k U^* S^{-1/2}, \quad \zeta^k = Q^{-1/2} X^k Q^{-1/2}$$

we obtain from (A.2) that

$$-\left\langle \boldsymbol{dt}^S_Q(\zeta^k), \zeta^k \right\rangle = -\left\langle \boldsymbol{dT}^S_Q(X^k), X^k \right\rangle = \frac{\lambda_k^{1/2}}{2} \|\zeta^k\|_F^2.$$

Therefore, the above inequalities are sharp. □

Lemma A.4. *For any* $Q_0, Q_1 \in \mathbb{H}_{++}(d)$, $S \in \mathbb{H}_+(d)$ *consider*

$$Q_t \overset{\text{def}}{=} (1-t)Q_0 + tQ_1, \quad Q' \overset{\text{def}}{=} Q_0^{-1/2} Q_1 Q_0^{-1/2}. \tag{A.5}$$

Then

$$\begin{aligned} \frac{2}{\lambda_{\min}(Q') + \lambda_{\min}^{1/2}(Q')} \boldsymbol{dT}^S_{Q_0} \preccurlyeq \int_0^1 \boldsymbol{dT}^S_{Q_t} \, dt & \\ \preccurlyeq \frac{2}{\lambda_{\max}(Q') + \lambda_{\max}^{1/2}(Q')} \boldsymbol{dT}^S_{Q_0} & \\ \preccurlyeq \frac{1}{1 + 3\|Q' - I\|/4} \boldsymbol{dT}^S_{Q_0}. & \end{aligned} \tag{I}$$

Moreover, if $\|Q' - I\| < 1$*, then*

$$\int_0^1 \boldsymbol{dT}^S_{Q_t} \, dt \succcurlyeq \frac{1}{1 - \|Q' - I\|} \boldsymbol{dT}^S_{Q_0}. \tag{II}$$

Remark 2. The above inequality might seem confusing due to the fact that $\lambda_{\min}(\cdot) \le \lambda_{\max}(\cdot)$, however this is explained by the fact that $\boldsymbol{dT}^S_Q$ is *negative* definite.

Proof. Notice that

$$((1-t) + t\lambda_{\min}(Q'))\, Q_0 \preccurlyeq Q_t = Q_0^{1/2} \big((1-t)I + tQ' \big) Q_0^{1/2} \preccurlyeq ((1-t) + t\lambda_{\max}(Q'))\, Q_0.$$

Monotonicity and homogeneity with degree $-\frac{3}{2}$ of $\boldsymbol{dT}^S_Q$ (see Lemma A.3) yield

$$\begin{aligned} \boldsymbol{dT}^S_{Q_t} &\preccurlyeq \boldsymbol{dT}^S_{((1-t)+t\lambda_{\max}(Q'))Q_0} \\ &= \big((1-t) + t\lambda_{\max}(Q') \big)^{-3/2} \boldsymbol{dT}^S_{Q_0} \end{aligned}$$

and

$$\begin{aligned} \boldsymbol{dT}^S_{Q_t} &\succcurlyeq \boldsymbol{dT}^S_{((1-t)+t\lambda_{\min}(Q'))Q_0} \\ &= \big((1-t) + t\lambda_{\min}(Q') \big)^{-3/2} \boldsymbol{dT}^S_{Q_0}. \end{aligned}$$

Therefore,

$$\begin{aligned} \int_0^1 \boldsymbol{dT}^S_{Q_t} \, dt &\preccurlyeq \boldsymbol{dT}^S_{Q_0} \int_0^1 \big((1-t) + t\lambda_{\max}(Q') \big)^{-3/2} \, dt \\ &= \frac{2}{\lambda_{\max}(Q') + \lambda_{\max}^{1/2}(Q')} \boldsymbol{dT}^S_{Q_0} \end{aligned}$$

and respectively,

$$\int_0^1 \boldsymbol{dT}^S_{Q_t}\, dt \succcurlyeq \frac{2}{\lambda_{\min}(Q') + \lambda^{1/2}_{\min}(Q')} \boldsymbol{dT}^S_{Q_0}.$$

The inequality (II) follows from the fact that

$$\lambda_{\min}(Q') \ge 1 - \|Q' - I\|, \quad \lambda_{\max}(Q') \le 1 + \|Q' - I\|,$$

and inequalities

$$\sqrt{1+x} \le 1 + \frac{x}{2} \text{ for } x \ge 0,$$
$$\sqrt{1-x} \ge 1 - x \text{ for } 0 \le x \le 1.$$

□

A.3 Properties of $d_{BW}(Q, S)$

The next lemma ensures strict convexity of $d_{BW}(Q, S)$. In essence, the proof mainly relies on Theorem 7 in Bhatia et al. [2018].

Lemma A.5. *For any $S \in \mathbb{H}_+(d)$ a function $Q \mapsto d^2_{BW}(Q, S)$ is convex on $\mathbb{H}_+(d)$. Moreover, if $S \succ 0$, then it is strictly convex.*

Proof. According to [Bhatia et al., 2018, Theorem 7] a function $h(X) = \operatorname{tr} X^{1/2}$ is strictly concave on $\mathbb{H}_+(d)$, hence the function

$$Q \mapsto d^2_{BW}(Q, S) = \operatorname{tr} S + \operatorname{tr} Q - 2 \operatorname{tr} \left(S^{1/2} Q S^{1/2}\right)^{1/2}$$

is convex on $\mathbb{H}_+(d)$ for any positive semi-definite S. Moreover, if $S \succ 0$, then $Q \mapsto S^{1/2} Q S^{1/2}$ is an injective linear map, and therefore $d^2_{BW}(Q, S)$ is strictly convex. □

Further we introduce differentiability of $d^2_{BW}(Q, S)$ and provides its quadratic approximation.

Lemma A.6. *For any $Q \in \mathbb{H}_{++}(d)$, $S \in \mathbb{H}_+(d)$ the function $d^2_{BW}(Q, S)$ is twice differentiable in Q with*

$$\begin{aligned} \boldsymbol{d_Q} d^2_{BW}(Q, S)(X) &= \langle I - T^S_Q, X\rangle, && X \in \mathbb{H}(d), \\ \boldsymbol{d^2_Q} d^2_{BW}(Q, S)(X, Y) &= -\langle X, \boldsymbol{dT}^S_Q(Y)\rangle, && X, Y \in \mathbb{H}(d). \end{aligned}$$

Moreover, the following quadratic approximation holds: for any $Q_0, Q_1 \in \mathbb{H}_{++}(d)$

$$\begin{aligned} -\tfrac{2}{\left(1+\lambda^{1/2}_{\max}(Q')\right)^2} &\left\langle \boldsymbol{dT}^S_{Q_0}(Q_1 - Q_0), Q_1 - Q_0 \right\rangle \\ &\le d^2_{BW}(Q_1, S) - d^2_{BW}(Q_0, S) + \langle T^S_{Q_0} - I, Q_1 - Q_0\rangle \\ &\le -\tfrac{2}{\left(1+\lambda^{1/2}_{\min}(Q')\right)^2} \left\langle \boldsymbol{dT}^S_{Q_0}(Q_1 - Q_0), Q_1 - Q_0 \right\rangle . \end{aligned}$$

with Q' defined in (A.5)*.*

Proof. Note that

$$\boldsymbol{d}_{\boldsymbol{Q}} \left(S^{1/2} Q S^{1/2}\right)^{1/2}(X) = U^* \delta U,$$

where δ comes from Lemma A.2. Furthermore, Lemma A.1 implies that

$$\begin{aligned}
\boldsymbol{d}_{\boldsymbol{Q}} \operatorname{tr} \left(S^{1/2} Q S^{1/2}\right)^{1/2}(X) = \operatorname{tr} \boldsymbol{d}_{\boldsymbol{Q}} \left(S^{1/2} Q S^{1/2}\right)^{1/2}(X) &= \operatorname{tr} \delta \\
&= \sum_{i=1}^{\operatorname{rank}(S)} \frac{\mathit{\Delta}_{ii}}{2\sqrt{\lambda_i}} = \frac{1}{2} \operatorname{tr} \mathit{\Delta} \mathit{\Lambda}^{-1/2} \\
&= \frac{1}{2} \operatorname{tr} S^{1/2} X S^{1/2} \left(S^{1/2} Q S^{1/2}\right)^{-1/2} = \frac{1}{2} \left\langle T_Q^S, X \right\rangle .
\end{aligned}$$

Consequently, $d_{BW}^2(Q,S)$ is differentiable, and

$$\boldsymbol{d}_{\boldsymbol{Q}} d_{BW}^2(Q,S)(X) = \operatorname{tr} X - 2\boldsymbol{d}_{\boldsymbol{Q}} \operatorname{tr} \left(S^{1/2} Q S^{1/2}\right)^{1/2}(X) = \left\langle I - T_Q^S, X \right\rangle .$$

Applying Lemma A.2 one obtains

$$\boldsymbol{d}_{\boldsymbol{Q}}^{\boldsymbol{2}} d_{BW}^2(Q,S)(X,Y) = \boldsymbol{d}_{\boldsymbol{Q}} \left\langle I - T_Q^S, X \right\rangle = -\left\langle \boldsymbol{dT}_Q^S(Y), X \right\rangle (Y).$$

Quadratic approximation Let $Q_0, Q_1 \in \mathbb{H}_{++}(d)$, $Q_t \overset{\text{def}}{=} (1-t)Q_0 + tQ_1$, $t \in [0,1]$. The Taylor expansion in the integral form applied to $d_{BW}^2(Q_t,S)$ implies

$$\begin{aligned}
d_{BW}^2(Q_1,S) &= d_{BW}^2(Q_0,S) + \left\langle I - T_{Q_0}^S, Q_1 - Q_0 \right\rangle \\
&\quad + \int_0^1 (1-t) \left\langle -\boldsymbol{dT}_{Q_t}^S(Q_1 - Q_0), Q_1 - Q_0 \right\rangle \, dt \\
&= d_{BW}^2(Q_0,S) - \left\langle T_{Q_0}^S - I, Q_1 - Q_0 \right\rangle \\
&\quad - \left\langle \left[\int_0^1 (1-t) \boldsymbol{dT}_{Q_t}^S \, dt \right] (Q_1 - Q_0), Q_1 - Q_0 \right\rangle .
\end{aligned}$$

Following the same ideas as in the proof of Lemma A.4 one obtains that

$$\begin{aligned}
\int_0^1 (1-t) \boldsymbol{dT}_{Q_t}^S \, dt &\preccurlyeq \int_0^1 (1-t) \big((1-t) + t\lambda_{\max}(Q')\big)^{-3/2} \boldsymbol{dT}_{Q_0}^S \, dt \\
&= \tfrac{2}{\left(1+\lambda_{\max}^{1/2}(Q')\right)^2} \boldsymbol{dT}_{Q_0}^S
\end{aligned}$$

and

$$\int_0^1 (1-t) \boldsymbol{dT}_{Q_t}^S \, dt \succcurlyeq \tfrac{2}{\left(1+\lambda_{\min}^{1/2}(Q')\right)^2} \boldsymbol{dT}_{Q_0}^S .$$

Thus

$$\begin{aligned}
&-\tfrac{2}{\left(1+\lambda_{\max}^{1/2}(Q')\right)^2} \left\langle \boldsymbol{dT}_{Q_0}^S(Q_1 - Q_0), Q_1 - Q_0 \right\rangle \\
&\qquad \le d_{BW}^2(Q_1,S) - d_{BW}^2(Q_0,S) + \langle T_{Q_0}^S - I, Q_1 - Q_0 \rangle \\
&\qquad \le -\tfrac{2}{\left(1+\lambda_{\min}^{1/2}(Q')\right)^2} \left\langle \boldsymbol{dT}_{Q_0}^S(Q_1 - Q_0), Q_1 - Q_0 \right\rangle .
\end{aligned}$$

□

A.4 Central limit theorem for Q_n and $\mathcal{V}_n$

First let us prove uniqueness and positive-definiteness of Bures–Wasserstein barycenter.

Proof of Theorem 2.1. By Assumption 2 $\mathcal{V}(0)$ is bounded:

$$\mathcal{V}(0) = \mathbb{E}\, d^2_{BW}(0, S) = \mathbb{E}\operatorname{tr} S < \infty.$$

Moreover, $d_{BW}(Q, S) \to \infty$ as $\|Q\| \to \infty$. This implies $\mathcal{V}(Q) \to \infty$ as $\|Q\| \to \infty$. Thus, any minimizing sequence for $\mathcal{V}(\cdot)$ is bounded. This observation allows us to use the compactness argument. As $\mathcal{V}(\cdot)$ is continuous, this implies existence of a barycenter Q_* by the compactness argument.

In case $\mathbb{P}(\mathbb{H}_{++}(d)) > 0$ applying Lemma A.5 we obtain strict convexity of the integral

$$Q \mapsto \mathbb{E}\, d^2_{BW}(Q, S) = \mathcal{V}(Q), \quad Q \in \mathbb{H}_+(d),$$

and therefore, uniqueness of the minimizer Q_*.

To prove that $Q_* \succ 0$ consider arbitrary degenerated $Q_0 \in \mathbb{H}_+(d) \cap \mathbb{A}$, $Q_1 \in \mathbb{H}_{++}(d) \cap \mathbb{A}$ (which exists by Assumption 1) and $S \in \mathbb{H}_{++}(d)$. Let us define $Q_t = (1-t)Q_0 + tQ_1 \in \mathbb{A}$. We are going to show, that

$$\frac{d}{dt} d^2_{BW}(Q_t, S) = \langle I - T^S_{Q_t}, Q_1 - Q_0\rangle \to -\infty \quad \text{as} \quad t \to 0.$$

To prove this convergence, we consider the following eigen-decomposition $S^{1/2}Q_0S^{1/2} = U^*\mathit{\Lambda} U$, $\mathit{\Lambda} = \operatorname{diag}(\lambda_1, \dots, \lambda_r, 0, \dots, 0)$, where $r = \operatorname{rank}(Q_0)$. We denote as $C = US^{1/2}Q_1S^{1/2}U^*$, and write it in a block form:

$$C = \begin{pmatrix} C_{11} & C_{12} \\ C_{21} & C_{22} \end{pmatrix}, \quad C_{11} \in \mathbb{H}_{++}(r),\ C_{12} = C^*_{21} \in \mathbb{C}^{r\times(d-r)},\ C_{22} \in \mathbb{H}_{++}(d-r).$$

Thus, for all Q_t the following representation holds (see Section A.5.5, paragraph *Inverse of block matrix* in Boyd and Vandenberghe [2004]):

$$\begin{aligned} U\left(S^{1/2}Q_tS^{1/2}\right)^{-1}U^* &= \left((1-t)\mathit{\Lambda} + tC\right)^{-1} \\ &= \begin{pmatrix} E_t^{-1} + t^2E_t^{-1}C_{12}S_t^{-1}C_{21}E_t^{-1} & -tE_t^{-1}C_{12}S_t^{-1} \\ -tS_t^{-1}C_{21}E_t^{-1} & S_t^{-1} \end{pmatrix}, \end{aligned}$$

where $E_t = (1-t)\mathit{\Lambda}_{11} + tC_{11}$, $S_t = tC_{22} - t^2C_{21}E_t^{-1}C_{12}$, with $\mathit{\Lambda}_{11} = \operatorname{diag}(\lambda_1, \dots, \lambda_r)$. When $t \to 0$, $E_t \to \mathit{\Lambda}_{11} \succ 0$, $\frac{S_t}{t} \to C_{22} \succ 0$. This yields

$$tU\left(S^{1/2}Q_tS^{1/2}\right)^{-1}U^* \to \begin{pmatrix} 0 & 0 \\ 0 & C_{22}^{-1} \end{pmatrix},$$

and

$$\sqrt{t}U\left(S^{1/2}Q_tS^{1/2}\right)^{-1/2}U^* \to \begin{pmatrix} 0 & 0 \\ 0 & C_{22}^{-1/2} \end{pmatrix}.$$

Therefore,

$$\begin{aligned} \sqrt{t}\left\langle T^S_{Q_t}, Q_0\right\rangle &= \sqrt{t}\left\langle \left(S^{1/2}Q_tS^{1/2}\right)^{-1/2}, S^{1/2}Q_0S^{1/2}\right\rangle \\ &= \left\langle \sqrt{t}U\left(S^{1/2}Q_tS^{1/2}\right)^{-1/2}U^*, US^{1/2}Q_0S^{1/2}U^*\right\rangle \\ &\to \left\langle \begin{pmatrix} 0 & 0 \\ 0 & C_{22}^{-1/2} \end{pmatrix}, \begin{pmatrix} \mathit{\Lambda}_{11} & 0 \\ 0 & 0 \end{pmatrix}\right\rangle = 0. \end{aligned}$$

In the same way one can obtain

$$\sqrt{t}\left\langle T^S_{Q_t}, Q_1\right\rangle \to \left\langle \begin{pmatrix} 0 & 0 \\ 0 & C_{22}^{-1/2} \end{pmatrix}, \begin{pmatrix} C_{11} & C_{12} \\ C_{21} & C_{22} \end{pmatrix} \right\rangle = \operatorname{tr} C_{22}^{1/2} > 0 \text{ as } t \to 0.$$

Consequently,

$$\frac{d}{dt} d^2_{BW}(Q_t, S) = \langle I - T^S_{Q_t}, Q_1 - Q_0 \rangle = \operatorname{tr} Q_1 - \operatorname{tr} Q_0 - \frac{\operatorname{tr} C_{22}^{1/2} + o(1)}{\sqrt{t}} \to -\infty.$$

By Assumption 1 it holds $\mathbb{P}(\mathbb{H}_{++}(d)) > 0$. Further, since $d^2_{BW}(Q, S)$ is convex, its directional derivatives are bounded by difference quotients, thus one can apply Leibniz integral rule for a Lebesgue-integrable function. This yields the following equality:

$$\frac{d}{dt}\mathcal{V}(Q_t) = \mathbb{E}\,\frac{d}{dt} d^2_{BW}(Q_t, S) \to -\infty \quad \text{as} \quad t \to 0,$$

thus Q_0 cannot be a barycenter of $\mathbb{P}$. This yields $Q_* \succ 0$.

Since $\mathcal{V}(\cdot)$ is convex and barycenter of $\mathbb{P}$ is positive-definite and unique, it is characterized as a stationary point of Fréchet variation on subspace $\mathbb{A}$, i.e. as a solution to equation

$$\boldsymbol{\Pi}_{\mathbb{M}} \nabla \mathcal{V}(Q) = \boldsymbol{\Pi}_{\mathbb{M}}(I - \mathbb{E}\, T^S_Q) = 0, \quad Q \in \mathbb{A} \cap \mathbb{H}_{++}(d),$$

as required. The first equality follows from Lemma A.6. □

The proof of CLT relies on covariance operators on the space of optimal transportation maps and on the space of covariance matrices.

Covariance operator on the space of optimal maps Consider $T_i \overset{\text{def}}{=} T^{S_i}_{Q_*}$ with $\mathbb{E}\, T_i = I$, and $T_i^n \overset{\text{def}}{=} T^{S_i}_{Q_n}$. We define a covariance $\boldsymbol{\Sigma}$ of T_i, its empirical counterpart $\boldsymbol{\Sigma_n}$, and its data-driven estimator $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{n}}$ as follows:

$$\boldsymbol{\Sigma} \overset{\text{def}}{=} \mathbb{E}\,(T_i - I) \otimes (T_i - I)\,, \quad \boldsymbol{\Sigma_n} \overset{\text{def}}{=} \frac{1}{n}\sum_{i=1}^n (T_i - I) \otimes (T_i - I)\,,$$

$$\hat{\boldsymbol{\Sigma}}_{\boldsymbol{n}} \overset{\text{def}}{=} \frac{1}{n}\sum_{i=1}^n (T_i^n - I) \otimes (T_i^n - I)\,. \tag{A.6}$$

Covariance operators on the space of covariance matrices Let Q_n be an empirical barycenter. The covariance of Q_n and its empirical counterpart are defined as

$$\Xi \overset{\text{def}}{=} \boldsymbol{F}^{-1}(\boldsymbol{\Sigma})_{\mathbb{M}} \boldsymbol{F}^{-1}, \quad \Xi \colon \mathbb{M} \to \mathbb{M}, \tag{A.7}$$

$$\hat{\Xi}_{\boldsymbol{n}} \overset{\text{def}}{=} \hat{\boldsymbol{F}}_{\boldsymbol{n}}^{-1}(\hat{\boldsymbol{\Sigma}}_{\boldsymbol{n}})_{\mathbb{M}} \hat{\boldsymbol{F}}_{\boldsymbol{n}}^{-1}, \quad \hat{\Xi}_{\boldsymbol{n}} \colon \mathbb{M} \to \mathbb{M}, \tag{A.8}$$

where

$$\boldsymbol{F} \overset{\text{def}}{=} -\mathbb{E}\left(\boldsymbol{dT}^S_{Q_*}\right)_{\mathbb{M}} \quad \boldsymbol{F_n} \overset{\text{def}}{=} -\frac{1}{n}\sum_{i=1}^n \left(\boldsymbol{dT}^{S_i}_{Q_*}\right)_{\mathbb{M}}\,, \tag{A.9}$$

$$\hat{\boldsymbol{F}}_{\boldsymbol{n}} \overset{\text{def}}{=} -\frac{1}{n}\sum_{i=1}^n \left(\boldsymbol{dT}^{S_i}_{Q_n}\right)_{\mathbb{M}}\,. \tag{A.10}$$

Now we are almost ready to prove the central limit theorem for the empirical barycenter Q_n (Theorem 2.2). Another key object which appears in the proofs quite often is a rescaled empirical barycenter:

$$Q_n' \overset{\text{def}}{=} Q_*^{-1/2} Q_n Q_*^{-1/2}. \tag{A.11}$$

For the sake of transparency we provide below a complete statement.

Theorem (Central limit theorem for the covariance of empirical barycenter)**.** *The approximation error rate of the Fréchet mean Q_* by its empirical counterpart Q_n is*

$$\sqrt{n}\left(Q_n - Q_*\right) \xrightarrow{\text{w}} \mathcal{N}\left(0, \Xi\right), \tag{A}$$

Moreover, if $(\boldsymbol{\Sigma})_{\mathbb{M}}$ is non-degenerated, then

$$\sqrt{n}\hat{\boldsymbol{\Xi}}_n^{-1/2}\left(Q_n - Q_*\right) \xrightarrow{\text{w}} \mathcal{N}\left(0, (\boldsymbol{I})_{\mathbb{M}}\right). \tag{B}$$

Proof of Theorem 2.2. The proof consists of two parts: proofs of (A) and (B).

Proof of (A)

As $\mathcal{V}_n(\cdot)$ are convex functions, they a.s. uniformly converge to a strictly convex function $\mathcal{V}(\cdot)$ on any compact set by the uniform law of large numbers. Therefore, their minimizers also converge $Q_n \xrightarrow{\text{a.s.}} Q_*$, see, e.g., Van De Geer [2006], Lemma 5.2.2. In particular, $\mathbb{P}\left(Q_n \succ 0\right) \to 1$, with $n \to \infty$. The expansion from Lemma A.2 at Q_* implies

$$T_i^n = T_i + \int_0^1 \boldsymbol{dT}_{Q_t}^{S_i}(Q_n - Q_*)\, dt, \tag{A.12}$$

where $Q_t = (1-t)Q_* + tQ_n$. Note, that the condition for Q_n being a barycenter is $\boldsymbol{\Pi}_{\mathbb{M}}\left(\frac{1}{n}\sum_i T_i^n - I\right) = 0$. This fact together with averaging of (A.12) over i give:

$$\boldsymbol{\Pi}_{\mathbb{M}} I = \boldsymbol{\Pi}_{\mathbb{M}} \overline{T}_n - \boldsymbol{\alpha}_n\big(Q_n - Q_*\big), \tag{A.13}$$

where

$$\overline{T}_n \overset{\text{def}}{=} \frac{1}{n}\sum_{i=1}^n T_i, \quad \boldsymbol{\alpha}_n \overset{\text{def}}{=} -\frac{1}{n}\sum_i \int_0^1 \left(\boldsymbol{dT}_{Q_t}^{S_i}\right)_{\mathbb{M}} dt. \tag{A.14}$$

According to Lemma A.4

$$\frac{2}{\lambda_{\max}(Q_n') + \lambda_{\max}^{1/2}(Q_n')}\boldsymbol{F}_n \preccurlyeq \boldsymbol{\alpha}_n \preccurlyeq \frac{2}{\lambda_{\min}(Q_n') + \lambda_{\min}^{1/2}(Q_n')}\boldsymbol{F}_n$$

where $\boldsymbol{F}_n$ is defined in (A.9), and Q_n' comes from (A.11). Recall that $\boldsymbol{F}$ introduced in (A.9) is a population counterpart of $\boldsymbol{F_n}$. This operator is correctly defined since by Lemma A.3 one can show that it is self-adjoint, positive definite and bounded:

$$\|\boldsymbol{F}\| \le \mathbb{E}\left\|\boldsymbol{dT}_{Q_*}^S\right\| \le \mathbb{E}\,\frac{\left\|S^{1/2}Q_*S^{1/2}\right\|}{2\lambda_{\min}^2(Q_*)} < \infty.$$

This bound follows directly from Corollary A.1.

Since by the law of large numbers $\boldsymbol{F}_n \xrightarrow{\text{a.s.}} \boldsymbol{F}$ and $Q_n' \xrightarrow{\text{a.s.}} I$, it holds that $\lambda_{\min}(Q_n') \xrightarrow{\text{a.s.}} 1$ and $\lambda_{\max}(Q_n') \xrightarrow{\text{a.s.}} 1$, thus $\boldsymbol{\alpha}_n \xrightarrow{\text{a.s.}} \boldsymbol{F}$. Therefore we obtain from (A.13)

$$\begin{aligned} Q_n &= Q_* + \boldsymbol{\alpha}_n^{-1}\boldsymbol{\Pi}_{\mathbb{M}}\left(\overline{T}_n - I\right) && \text{(A.15)}\\ &= Q_* + \boldsymbol{F}^{-1}\boldsymbol{\Pi}_{\mathbb{M}}\left(\overline{T}_n - I\right) + o_P\left(\left\|\boldsymbol{\Pi}_{\mathbb{M}}\left(\overline{T}_n - I\right)\right\|\right), && \text{(A.16)}\end{aligned}$$

where $\boldsymbol{F}^{-1}$ is a bounded linear operator, because $\boldsymbol{dT}_{Q_*}^S$ is negative definite for any $S \succ 0$ by Lemma A.3. The result (A) follows immediately from the CLT for $\boldsymbol{\Pi}_{\mathbb{M}}\overline{T}_n$.

Proof of (B)

Note that result (A) is equivalent to the fact, that

$$\sqrt{n}\boldsymbol{\Xi}^{-1/2}\left(Q_n - Q_*\right) \xrightarrow{\text{w}} \mathcal{N}\left(0, (\boldsymbol{I})_{\mathbb{M}}\right).$$

To ensure convergence of $\hat{\boldsymbol{\Xi}}_{\boldsymbol{n}} \xrightarrow{\text{a.s.}} \boldsymbol{\Xi}$ we need to show that

a) $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{n}} \xrightarrow{\text{a.s.}} \boldsymbol{\Sigma}$ (follows from Lemma B.2, a.s. consistency of Q_n', and the LLN);

b) $\hat{\boldsymbol{F}}_{\boldsymbol{n}} \xrightarrow{\text{a.s.}} \boldsymbol{F}$.

Consider

$$\begin{aligned} \boldsymbol{dT}_{Q_n}^S &\preccurlyeq \boldsymbol{dT}_{\lambda_{\max}(Q_n')Q_*}^S = \left(\lambda_{\max}(Q_n')\right)^{-3/2}\boldsymbol{dT}_{Q_*}^S, \\ \boldsymbol{dT}_{Q_n}^S &\succcurlyeq \boldsymbol{dT}_{\lambda_{\min}(Q_n')Q_*}^S = \left(\lambda_{\min}(Q_n')\right)^{-3/2}\boldsymbol{dT}_{Q_*}^S, \end{aligned}$$

where the inequalities come from monotonicity of $\boldsymbol{dT}_Q^S$ (see (V) in Lemma A.3), and the fact that and bounds $\lambda_{\min}(Q_n')Q_* \preccurlyeq Q_n \preccurlyeq \lambda_{\max}(Q_n')Q_*$. The equalities hold due to homogeneity of $\boldsymbol{dT}_Q^S$ with degree $-\frac{3}{2}$ (see (IV) in Lemma A.3). This naturally leads to the following bounds:

$$\frac{1}{\lambda_{\max}^{3/2}(Q_n')}\boldsymbol{F_n} \preccurlyeq \hat{\boldsymbol{F}}_{\boldsymbol{n}} \preccurlyeq \frac{1}{\lambda_{\min}^{3/2}(Q_n')}\boldsymbol{F_n}.$$

Since $Q_n' \xrightarrow{\text{a.s.}} I$ and $\boldsymbol{F_n} \xrightarrow{\text{a.s.}} \boldsymbol{F}$, this implies $\hat{\boldsymbol{F}}_{\boldsymbol{n}} \xrightarrow{\text{a.s.}} \boldsymbol{F}$ due to the following continuity property: $\lambda_{\max}(Q_n') \le 1 + \|Q_n' - I\|$ and $\lambda_{\min}(Q_n') \ge 1 - \|Q_n' - I\|$.

The above results ensure the validity of substitution $\boldsymbol{\Xi}$ by $\hat{\boldsymbol{\Xi}}_{\boldsymbol{n}}$. This yields (B). □

The asymptotic convergence results for $d_{BW}(Q_n, Q_*)$ is a straightforward corollary of the above theorem. Here is the proof.

Proof of Corollary 2.1. Since $Q_n \xrightarrow{\text{a.s.}} Q_*$, Lemma A.6 implies

$$d_{BW}^2(Q_n, Q_*) = -\frac{1 + o_P(1)}{2}\left\langle \boldsymbol{dT}_{Q_*}^{Q_*}(Q_n - Q_*), Q_n - Q_* \right\rangle.$$

Without loss of generality we can consider case $Q_* = \operatorname{diag}(q_1, \ldots, q_d)$, thus Lemma A.3 implies (notice that $\Lambda = Q_*^2$ and $\Delta = Q_*^{1/2}XQ_*^{1/2}$)

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

By Theorem 2.2 $\sqrt{n}(Q_n - Q_*)$ is asymptotically normal and centred, therefore

$$\mathcal{L}\left(\sqrt{n}d_{BW}(Q_n, Q_*)\right) \xrightarrow{\text{w}} \mathcal{L}\left(\left\|Q_*^{1/2}\boldsymbol{dT}_{Q_*}^{Q_*}(Z)\right\|_F\right).$$

where $Z \in \mathbb{M} \subset \mathbb{H}(d)$ and $Z \sim \mathcal{N}(0, \Xi)$.

Note, that $Q_n \xrightarrow{\text{a.s.}} Q_*$, $\hat{\boldsymbol{\Xi}}_n \xrightarrow{\text{a.s.}} \boldsymbol{\Xi}$, and $\boldsymbol{dT}_{Q_n}^{Q_n} \xrightarrow{\text{a.s.}} \boldsymbol{dT}_{Q_*}^{Q_*}$. The last result follows from Lemma A.3 (IV, V), and can be validated using the same framework as in the proof of (B) in Theorem 2.2. Note, that $\lambda_{\min}(Q_n')Q_* \preccurlyeq Q_n \preccurlyeq \lambda_{\max}(Q_n')Q_*$, with Q_n' coming from (A.11). Then

$$\begin{aligned} \boldsymbol{dT}_{Q_n}^{Q_n} &\preccurlyeq \boldsymbol{dT}_{\lambda_{\max}(Q_n')Q_*}^{\lambda_{\max}(Q_n')Q_*} = \frac{1}{\lambda_{\max}(Q_n')}\boldsymbol{dT}_{Q_*}^{Q_*} \to \boldsymbol{dT}_{Q_*}^{Q_*}, \\ \boldsymbol{dT}_{Q_n}^{Q_n} &\succcurlyeq \boldsymbol{dT}_{\lambda_{\min}(Q_n')Q_*}^{\lambda_{\min}(Q_n')Q_*} = \frac{1}{\lambda_{\min}(Q_n')}\boldsymbol{dT}_{Q_*}^{Q_*} \to \boldsymbol{dT}_{Q_*}^{Q_*}, \end{aligned}$$

where the inequalities comes from monotonicity (see (V) in Lemma A.3). The equalities hold due to homogeneity (see (IV) in Lemma A.3). Furthermore, $\lambda_{\max}(Q_n') \le 1 + \|Q_n' - I\|$ and $\lambda_{\min}(Q_n') \ge 1 - \|Q_n' - I\|$. This yields

$$\mathcal{L}\left(\left\|Q_n^{1/2}\boldsymbol{dT}_{Q_n}^{Q_n}(Z_n)\right\|_F\right) \xrightarrow{\text{w}} \mathcal{L}\left(\left\|Q_*^{1/2}\boldsymbol{dT}_{Q_*}^{Q_*}(Z)\right\|_F\right),$$

where $Z_n \sim \mathcal{N}\left(0, \hat{\boldsymbol{\Xi}}_n\right)$. This, in turn, entails

$$d_{\text{w}}\Big(\mathcal{L}\left(\sqrt{n}d_{BW}(Q_n, Q_*)\right), \mathcal{L}\left(\left\|Q_n^{1/2}\boldsymbol{dT}_{Q_n}^{Q_n}(Z_n)\right\|_F\right)\Big) \to 0,$$

where d_{w} is some metric inducing the weak convergence of the measures. □

Finally, we are ready to prove Theorem 2.3.

Proof of Theorem 2.3. By definition empirical Fréchet variance is

$$\mathcal{V}_n(Q) = \frac{1}{n}\sum_{i=1}^n d_{BW}^2(Q, S_i).$$

Lemma A.6 ensures the following bound on $\mathcal{V}_n(Q_*) - \mathcal{V}_n(Q_n)$:

$$0 \le \mathcal{V}_n(Q_*) - \mathcal{V}_n(Q_n) \le \frac{2}{\left(1+\lambda_{\min}^{1/2}(Q_n')\right)^2}\langle \boldsymbol{F}_n(Q_n - Q_*), Q_n - Q_*\rangle$$

with $Q_n' \overset{\text{def}}{=} Q_*^{-1/2}Q_nQ_*^{-1/2}$. The above quadratic bound together with $Q_n \to Q_*$, $\boldsymbol{F}_n \to \boldsymbol{F}$ and $\sqrt{n}(Q_n - Q_*) \xrightarrow{\text{w}} \mathcal{N}(0, \boldsymbol{\Xi})$ yield:

$$\mathcal{V}_n(Q_n) - \mathcal{V}(Q_*) = \mathcal{V}_n(Q_*) - \mathcal{V}(Q_*) + O_P\left(\frac{1}{n}\right).$$

On the other hand, by the classical central limit theorem we obtain:

$$\begin{aligned} \sqrt{n}\left(\mathcal{V}_n(Q_*) - \mathcal{V}(Q_*)\right) &= \sqrt{n}\left(\frac{1}{n}\sum_i d_{BW}^2(Q_*, S_i) - \mathbb{E}\, d_{BW}^2(Q_*, S)\right) \\ &\xrightarrow{\text{w}} \mathcal{N}\left(0, \operatorname{Var} d_{BW}^2(Q_*, S)\right). \end{aligned}$$

□

B Concentrations of Q_n and $\mathcal{V}_n$

B.1 Concentration of Q_n

The next lemma is a key ingredient in the proof of the concentration result for Q_n.

Lemma B.1. *Consider*

$$\eta_n \overset{\text{def}}{=} \frac{1}{\lambda_{\min}(\boldsymbol{F}_n')} \bigl\| Q_*^{1/2} \boldsymbol{\Pi}_{\mathbb{M}} \left(\overline{T}_n - I\right) Q_*^{1/2} \bigr\|_F \tag{B.1}$$

where

$$\boldsymbol{F}_n'(X) \overset{\text{def}}{=} Q_*^{1/2} \boldsymbol{F}_n \left(Q_*^{1/2} X Q_*^{1/2}\right) Q_*^{1/2}, \quad X \in \left\{ Q_*^{-1/2} Y Q_*^{-1/2} \middle| Y \in \mathbb{M} \right\}. \tag{B.2}$$

Then

$$\bigl\| Q_*^{-1/2} Q_n Q_*^{-1/2} - I \bigr\|_F \le \frac{\eta_n}{1 - \frac{3}{4}\eta_n}$$

whenever $\eta_n < \frac{4}{3}$ *and* $Q_n \succ 0$.

Proof. Let us define $Q_t \overset{\text{def}}{=} tQ_n + (1-t)Q_*$ for $t \in [0,1]$, and Q_n' defined in (A.11). Due to Lemmas A.3 and A.4 we have for any $S \in \mathbb{H}_+(d)$

$$\begin{aligned}
\bigl\langle \boldsymbol{\Pi}_{\mathbb{M}} \left(T_{Q_*}^S - T_{Q_n}^S\right), Q_n - Q_* \bigr\rangle &= \langle T_{Q_*}^S - T_{Q_n}^S, Q_n - Q_* \rangle \\
&= \int_0^1 \bigl\langle -\boldsymbol{dT}_{Q_t}^S (Q_n - Q_*), Q_n - Q_* \bigr\rangle \, dt \\
&\ge \frac{1}{1 + \frac{3}{4}\|Q_n' - I\|} \bigl\langle -\boldsymbol{dT}_{Q_*}^S (Q_n - Q_*), Q_n - Q_* \bigr\rangle.
\end{aligned}$$

Therefore,

$$\begin{aligned}
\bigl\langle \boldsymbol{\Pi}_{\mathbb{M}} \left(\overline{T}_n - I\right), Q_n - Q_* \bigr\rangle &\ge \frac{1}{1 + \frac{3}{4}\|Q_n' - I\|} \bigl\langle \boldsymbol{F}_n (Q_n - Q_*), Q_n - Q_* \bigr\rangle \\
&= \frac{1}{1 + \frac{3}{4}\|Q_n' - I\|} \bigl\langle \boldsymbol{F}_n' (Q_n' - I), Q_n' - I \bigr\rangle \\
&\ge \frac{\lambda_{\min}(\boldsymbol{F}_n')}{1 + \frac{3}{4}\|Q_n' - I\|} \|Q_n' - I\|_F^2.
\end{aligned}$$

At the same time,

$$\begin{aligned}
\bigl\langle \boldsymbol{\Pi}_{\mathbb{M}} \left(\overline{T}_n - I\right), Q_n - Q_* \bigr\rangle &= \bigl\langle Q_*^{1/2} \boldsymbol{\Pi}_{\mathbb{M}} \left(\overline{T}_n - I\right) Q_*^{1/2}, Q_n' - I \bigr\rangle \\
&\le \bigl\| Q_*^{1/2} \boldsymbol{\Pi}_{\mathbb{M}} \left(\overline{T}_n - I\right) Q_*^{1/2} \bigr\|_F \|Q_n' - I\|_F.
\end{aligned}$$

Hence

$$\|Q_n' - I\|_F \le \frac{1 + \frac{3}{4}\|Q_n' - I\|}{\lambda_{\min}(\boldsymbol{F}_n')} \bigl\| Q_*^{1/2} \boldsymbol{\Pi}_{\mathbb{M}} \left(\overline{T}_n - I\right) Q_*^{1/2} \bigr\|_F = \left(1 + \tfrac{3}{4}\|Q_n' - I\|\right) \eta_n.$$

Rewriting the inequality above we obtain

$$\|Q_n' - I\|_F \le \frac{\eta_n}{1 - \frac{3}{4}\eta_n}$$

provided that $\eta_n < \frac{4}{3}$. □

Before proving concentration results, we define operator $\boldsymbol{F'}(X)$ as follows:

$$\boldsymbol{F'}(X) \overset{\text{def}}{=} Q_*^{1/2}\boldsymbol{F}\left(Q_*^{1/2}XQ_*^{1/2}\right)Q_*^{1/2} \text{ for } X \in \left\{Q_*^{-1/2}YQ_*^{-1/2}\middle| Y \in \mathbb{M}\right\}. \tag{B.3}$$

Proof of Theorem 2.4. Let t_n be s.t. the following upper bound on $\gamma_n(t_n)$ from Proposition B.1 holds:

$$\gamma_n(t_n) \le \frac{1}{2}\lambda_{\min}(\boldsymbol{F'}). \tag{B.4}$$

It is easy to see that this condition is fulfilled for $t_n = nt_F - \log(m)$ under a proper choice of generic constant in definition of t_F. Then with probability at least $1 - 2me^{-nt_F}$ the following bound holds:

$$\lambda_{\min}(\boldsymbol{F'_n}) \ge \lambda_{\min}(\boldsymbol{F'}) - \|\boldsymbol{F'_n} - \boldsymbol{F'}\| \ge \frac{1}{2}\lambda_{\min}(\boldsymbol{F'}),$$

with $\boldsymbol{F'_n}$ to be defined in (B.2). The above facts together with definition of η_n (B.1) yield

$$\begin{aligned}\eta_n &\overset{\text{def}}{=} \frac{\left\|Q_*^{1/2}\boldsymbol{\varPi}_{\mathbb{M}}\left(\overline{T}_n - I\right)Q_*^{1/2}\right\|_F}{\lambda_{\min}(\boldsymbol{F'_n})} \\ &\le \frac{2\|Q_*\|}{\lambda_{\min}(\boldsymbol{F'})}\left\|\boldsymbol{\varPi}_{\mathbb{M}}\left(\overline{T}_n - I\right)\right\|_F = \frac{c_Q}{2\sigma_T}\left\|\boldsymbol{\varPi}_{\mathbb{M}}\left(\overline{T}_n - I\right)\right\|_F.\end{aligned}$$

Combining the above bounds with Proposition B.2, we obtain:

$$\mathbb{P}\left\{\eta_n \ge \frac{c_Q}{2\sqrt{n}}(\sqrt{m} + t)\right\} \le 2me^{-nt_F} + e^{-t^2/2}.$$

Now it follows from Lemma B.1 that

$$\begin{aligned}\mathbb{P}\left\{\|Q'_n - I\|_F \ge \frac{c_Q}{\sqrt{n}}(\sqrt{m} + t)\right\} &\le \mathbb{P}\left\{2\eta_n \ge \frac{c_Q}{\sqrt{n}}(\sqrt{m} + t)\right\} + \mathbb{P}\{Q_n \not\succ 0\} \\ &\le 2me^{-nt_F} + e^{-t^2/2} + (1-p)^n,\end{aligned}$$

whenever $\frac{c_Q}{2\sqrt{n}}(\sqrt{m} + t) \le \frac{2}{3}$. Here we used that $Q_n \succ 0$ if at least one of matrices $S_1, \dots, S_n$ is non-degenerated. Here $Q \not\succ 0$ means that a matrix Q is not positive definite. □

Proof of Corollary 2.2. To prove this result we use Lemma A.6 and choose $Q_0 = S = Q_*$, $Q_1 = Q_n$. Thus we obtain

$$\begin{aligned}d^2_{BW}(Q_n, Q_*) &\le -\frac{2}{\left(1+\lambda_{\min}^{1/2}(Q'_n)\right)^2}\left\langle \boldsymbol{dT}^{Q_*}_{Q_*}(Q_n - Q_*), Q_n - Q_*\right\rangle \\ &\overset{Def.\ A.4}{=} \frac{2}{\left(1+\lambda_{\min}^{1/2}(Q'_n)\right)^2}\left\langle -\boldsymbol{dt}^{Q_*}_{Q_*}(Q'_n - I), Q'_n - I\right\rangle \\ &\le 2\lambda_{\max}\left(-\boldsymbol{dt}^{Q_*}_{Q_*}\right)\|Q'_n - I\|_F^2 \overset{C.A.2}{=} \lambda_{\max}(Q_*)\|Q'_n - I\|_F^2,\end{aligned}$$

with Q'_n coming from (A.11). Hence by Theorem 2.4

$$d_{BW}(Q_n, Q_*) \le \|Q_*\|^{1/2}\frac{c_Q}{\sqrt{n}}(\sqrt{m} + t)$$

with probability at least $1 - 2me^{-nt_F} - e^{-t^2/2} - (1-p)^n$. □

B.2 Concentration of $\mathcal{V}_n$

Proof of Theorem 2.5. Following the proof of Theorem 2.3 we consider $\mathcal{V}_n(Q_*) - \mathcal{V}_n(Q_n)$:

$$\begin{aligned} 0 \le \mathcal{V}_n(Q_*) - \mathcal{V}_n(Q_n) &\le \tfrac{2}{\left(1+\lambda_{\min}^{1/2}(Q_n')\right)^2}\langle \boldsymbol{F}_n(Q_n - Q_*), Q_n - Q_*\rangle \\ &= \tfrac{2}{\left(1+\lambda_{\min}^{1/2}(Q_n')\right)^2}\langle \boldsymbol{F}_n'(Q_n' - I), Q_n' - I\rangle \\ &\le 2\|\boldsymbol{F}_n'\| \cdot \|Q_n' - I\|_F^2, \end{aligned} \tag{B.5}$$

with $\boldsymbol{F}_n'$ to be defined in (B.2), and Q_n' in (A.11). Following the proof of Theorem 2.4, we obtain that with $\mathbb{P} \ge 1 - 2me^{-t_F n} - e^{-t^2/2} - (1-p)^n$ the following upper bounds hold:

$$\|Q_n' - I\|_F \le \frac{c_Q}{\sqrt{n}}(\sqrt{m} + t), \quad \|\boldsymbol{F}_n' - \boldsymbol{F}'\| \le \frac{1}{2}\lambda_{\min}(\boldsymbol{F}'),$$

with $\boldsymbol{F}'$ coming from (B.3). Thus

$$\|\boldsymbol{F}_n'\| \le \|\boldsymbol{F}'\| + \|\boldsymbol{F}_n' - \boldsymbol{F}'\| \le \frac{3}{2}\|\boldsymbol{F}'\|$$

and consequently

$$0 \le \mathcal{V}_n(Q_*) - \mathcal{V}_n(Q_n) \le 3\|\boldsymbol{F}'\|\frac{c_Q^2}{n}(\sqrt{m}+t)^2.$$

Now we consider a difference $\mathcal{V}_n(Q_*) - \mathcal{V}(Q_*)$. According to Assumption 3 S, and therefore $d_{BW}^2(Q_*, S)$, are sub-exponential r.v. with some parameters (ν, μ). Then Lemma B.4 ensures

$$|\mathcal{V}_n(Q_*) - \mathcal{V}(Q_*)| \le \max\left(\frac{2\mu t'}{n}, \nu\left(\frac{2t'}{n}\right)^{1/2}\right)$$

with probability $1 - 2e^{-t'}$. Combining two above bounds, we obtain:

$$|\mathcal{V}_n(Q_n) - \mathcal{V}(Q_*)| \le \max\left(\frac{2\mu t'}{n}, \nu\sqrt{\frac{2t'}{n}}\right) + 3\|\boldsymbol{F}'\|\frac{c_Q^2}{n}(\sqrt{m}+t)^2$$

with probability

$$\mathbb{P} \ge 1 - 2e^{-t'} - 2me^{-nt_F} - e^{-t^2/2} - (1-p)^n.$$

Choosing $t' = t^2/2$, we get

$$\begin{aligned} \mathbb{P}\left\{|\mathcal{V}_n(Q_n) - \mathcal{V}(Q_*)| \ge \max\left(\frac{\mu t^2}{n}, \frac{\nu t}{\sqrt{n}}\right) + 3\|\boldsymbol{F}'\|\frac{c_Q^2}{n}(\sqrt{m}+t)^2\right\} \\ \le 2me^{-nt_F} + 3e^{-t^2/2} + (1-p)^n. \end{aligned}$$

□

B.3 Auxiliary results

Lemma B.2. *Let $\|Q_n' - I\| \le \frac{1}{2}$, with Q_n' coming from* (A.11)*; then*

$$\left\| \hat{\boldsymbol{\Sigma}}_{\boldsymbol{n}} - \boldsymbol{\Sigma}_{\boldsymbol{n}} \right\|_1 \le \beta_n \left[2 \left(\frac{1}{n} \sum_i \|T_i - I\|_F^2 \right)^{1/2} + \beta_n \right],$$

where

$$\beta_n \overset{\text{def}}{=} \kappa(Q_*) \left(\frac{\frac{1}{n}\sum_i \|S_i\|}{\|Q_*\|} \right)^{1/2} \|Q_n' - I\|_F,$$

where $\kappa(Q_*) = \|Q_*\| \cdot \|Q_*^{-1}\|$ is the condition number of matrix Q_* and $\|\boldsymbol{A}\|_1$ is 1-Schatten (nuclear) norm of an operator $\boldsymbol{A}$.

Proof. Note, that for any $(T_i^n - I) \otimes (T_i^n - I)$ the following decomposition holds

$$\begin{aligned}
&(T_i^n - I) \otimes (T_i^n - I) \\
&\qquad = (T_i - I) \otimes (T_i - I) + (T_i^n - T_i) \otimes (T_i - I) \\
&\qquad + (T_i - I) \otimes (T_i^n - T_i) + (T_i^n - T_i) \otimes (T_i^n - T_i).
\end{aligned}$$

Summing over i yields

$$\begin{aligned}
\hat{\boldsymbol{\Sigma}}_{\boldsymbol{n}} - \boldsymbol{\Sigma}_{\boldsymbol{n}} &= \frac{1}{n} \sum_i (T_i^n - T_i) \otimes (T_i - I) \\
&+ \frac{1}{n} \sum_i (T_i - I) \otimes (T_i^n - T_i) + \frac{1}{n} \sum_i (T_i^n - T_i) \otimes (T_i^n - T_i).
\end{aligned} \tag{B.6}$$

Note, that each

$$\|(T_i^n - T_i) \otimes (T_i - I)\|_1 \le \|T_i^n - T_i\|_F \|T_i - I\|_F.$$

Lemmas A.3 (III) and A.4 yield

$$\begin{aligned}
\|T_i^n - T_i\|_F &\le \frac{1}{1 - \|Q_n' - I\|} \left\| \boldsymbol{dT}_{Q_*}^{S_i}(Q_n - Q_*) \right\|_F \\
&\le 2 \left\| Q_*^{-1/2} \boldsymbol{dt}_{Q_*}^{S_i} (Q_n' - I) Q_*^{-1/2} \right\|_F \le 2 \frac{\lambda_{\max}\left(\boldsymbol{dt}_{Q_*}^{S_i}\right)}{\lambda_{\min}(Q_*)} \|Q_n' - I\|_F \\
&\le \frac{\lambda_{\max}^{1/2}\left(S_i^{1/2} Q_* S_i^{1/2}\right)}{\lambda_{\min}(Q_*)} \|Q_n' - I\|_F \le \kappa(Q_*) \left(\frac{\|S_i\|}{\|Q_*\|} \right)^{1/2} \|Q_n' - I\|_F,
\end{aligned}$$

where $\boldsymbol{dt}_Q^S$ is defined in (A.4). Hence $\frac{1}{n}\sum_i \|T_i^n - T_i\|_F^2 \le \beta_n^2$. The above expression together with (B.6) and Cauchy–Schwarz inequality lead to the upper bound on $\left\| \hat{\boldsymbol{\Sigma}}_{\boldsymbol{n}} - \boldsymbol{\Sigma}_{\boldsymbol{n}} \right\|_1$:

$$\begin{aligned}
\left\| \hat{\boldsymbol{\Sigma}}_{\boldsymbol{n}} - \boldsymbol{\Sigma}_{\boldsymbol{n}} \right\|_1 &\le \frac{2}{n} \sum_i \|T_i - I\|_F \|T_i^n - T_i\|_F + \frac{1}{n} \sum_i \|T_i^n - T_i\|_F^2 \\
&\le 2\beta_n \left(\frac{1}{n} \sum_i \|T_i - I\|_F^2 \right)^{1/2} + \beta_n^2.
\end{aligned}$$

□

Further we present concentration of $\boldsymbol{F}_n$ around $\boldsymbol{F}$. Denote as $\|X\|_{\psi_2}$ an Orlicz norm with Young function $\psi_2(x) = e^{x^2} - 1$, i.e.

$$\|X\|_{\psi_2} \stackrel{\text{def}}{=} \inf\bigl\{c > 0 : \mathbb{E}\,\psi_2\left(|X|/c\right) \le 1\bigr\}.$$

Then sub-Gaussianity of a r.v. X is equivalent to $\|X\|_{\psi_2} < \infty$ and it ensures

$$\operatorname{Var}(X) \le \sqrt{2}\|X\|_{\psi_2}.$$

Proposition B.1 (Concentration of $\boldsymbol{F}_n'$, Proposition 2 in Koltchinskii [2011]). *Let $\boldsymbol{F_n'}$, $\boldsymbol{F'}$, and $\boldsymbol{dt}_Q^S$ be defined as* (B.2), (B.3)*, and* (A.4)*, respectively. There exists a constant* $\mathtt{C} > 0$*, s.t. for all* $t > 0$ *it holds with probability at least* $1 - e^{-t}$

$$\|\boldsymbol{F}_n' - \boldsymbol{F}'\| \le \gamma_n(t), \quad \gamma_n(t) \stackrel{\text{def}}{=} \mathtt{C}\max\left(\sigma_F\sqrt{\tfrac{t+\log(2m)}{n}}, U\sqrt{\log\left(\tfrac{U}{\sigma_F}\right)\tfrac{t+\log(2m)}{n}}\right),$$

where $\sigma_F^2 \stackrel{\text{def}}{=} \left\|\mathbb{E}\left(\boldsymbol{dt}_{Q_*}^S - \boldsymbol{F}'\right)^2\right\|$, $U \stackrel{\text{def}}{=} \left\|\|\boldsymbol{dt}_{Q_*}^S - \boldsymbol{F}'\|\right\|_{\psi_2}$.

Lemma B.3. *The size of the above constants can be estimated as follows:*

$$\sigma_F \le \frac{\|Q_*\|^{1/2}}{2}\left(\mathbb{E}\|S\|\right)^{1/2}, \quad U \le \frac{3}{2}\|Q_*\|^{1/2}\|\|S\|\|_{\psi_1}^{1/2},$$

where $\psi_1(x) = e^x - 1$ *is a Young function.*

Proof. By Corollary A.2 we obtain

$$\sigma_F^2 \stackrel{\text{def}}{=} \left\|\mathbb{E}\left(\boldsymbol{dt}_{Q_*}^S - \boldsymbol{F}'\right)^2\right\| \le \mathbb{E}\left\|\boldsymbol{dt}_{Q_*}^S\right\|^2 \le \frac{\|Q_*\|}{4}\,\mathbb{E}\|S\|$$

and (due to properties of Orlicz norm)

$$\begin{aligned} U \stackrel{\text{def}}{=} \left\|\|\boldsymbol{dt}_{Q_*}^S - \boldsymbol{F}'\|\right\|_{\psi_2} &\le \frac{\|\boldsymbol{F}'\|}{\sqrt{\ln 2}} + \left\|\|\boldsymbol{dt}_{Q_*}^S\|\right\|_{\psi_2} \\ &\le \frac{\|Q_*\|^{1/2}}{2}\left[2\,\mathbb{E}\|S\|^{1/2} + \left\|\|S\|^{1/2}\right\|_{\psi_2}\right] \\ &\le \frac{\|Q_*\|^{1/2}}{2}\left[2\left(\mathbb{E}\|S\|\right)^{1/2} + \|\|S\|\|_{\psi_1}^{1/2}\right] \\ &\le \frac{3}{2}\|Q_*\|^{1/2}\|\|S\|\|_{\psi_1}^{1/2}. \end{aligned}$$

□

The next proposition ensures the concentration of $\overline{T}_n$.

Proposition B.2 (Concentration of $\overline{T}_n$; Hsu et al. [2012], Theorem 1). *Under Assumption 3 it holds*

$$\mathbb{P}\left\{\left\|\boldsymbol{\Pi}_{\mathbb{M}}\left(\overline{T}_n - I\right)\right\|_F \ge \tfrac{\sigma_T}{\sqrt{n}}\left(\sqrt{m} + t\right)\right\} \le e^{-t^2/2} \quad \textit{for any} \quad t \ge 0.$$

Lemma B.4 (Sub-exponential tail bounds). *Suppose that X is sub-exponential with parameters ν, b. Then*

$$\mathbb{P}\left\{X \ge \mathbb{E}\,X + t\right\} \le \begin{cases} \exp\left(-\frac{t^2}{2\nu^2}\right), & \textit{if} \quad 0 \le t \le \frac{\nu^2}{b}, \\ \exp\left(-\frac{t}{2b}\right), & \textit{if} \quad t \ge \frac{\nu^2}{b}. \end{cases}$$