

Wasserstein F-tests for Fréchet regression on Bures-Wasserstein manifolds

Haoshu Xu

HAOSHUXU@SAS.UPENN.EDU

*Graduate Group in Applied Mathematics and Computational Science
University of Pennsylvania
Philadelphia, PA 19104, USA*

Hongzhe Li

HONGZHE@UPENN.EDU

*Department of Biostatistics, Epidemiology and Informatics
University of Pennsylvania
Philadelphia, PA 19104, USA*

Editor: Xiaotong Shen

Abstract

This paper addresses regression analysis for covariance matrix-valued outcomes with Euclidean covariates, motivated by applications in single-cell genomics and neuroscience where covariance matrices are observed across many samples. Our analysis leverages Fréchet regression on the Bures-Wasserstein manifold to estimate the conditional Fréchet mean given covariates x . We establish a non-asymptotic uniform \sqrt{n} -rate of convergence (up to logarithmic factors) over covariates with $\|x\| \lesssim \sqrt{\log n}$ and derive a pointwise central limit theorem to enable statistical inference. For testing covariate effects, we devise a novel test whose null distribution converges to a weighted sum of independent chi-square distributions, with power guarantees against a sequence of contiguous alternatives. Simulations validate the accuracy of the asymptotic theory. Finally, we apply our methods to a single-cell gene expression dataset, revealing age-related changes in gene co-expression networks.

Keywords: Fréchet regression; Functional calculus; Hypothesis testing; Optimal transport; Wasserstein distance.

1. Introduction

In modern data analysis, positive definite matrices frequently arise across various fields, including medical imaging (Dryden et al., 2009; Fillard et al., 2007), neuroscience (Friston, 2011; Kong et al., 2020; Hu et al., 2021), signal processing (Arnaudon et al., 2013), and computer vision (Caseiro et al., 2012). For example, in large-scale single-cell RNA-seq data, individual-specific covariance matrices can be estimated and interpreted as co-expression networks among a set of genes. In neuroimaging, covariance matrices (or correlation matrices after standardization) of multiple brain regions are used to summarize functional connectivity, providing insights into brain network organization. A central challenge in these applications is performing regression analysis where the outcome is a covariance matrix and the covariates are Euclidean.

Several regression models have been proposed for covariance matrix outcomes. Chiu et al. (1996) developed a method that models the elements of the logarithm of the covariance

matrix as a linear function of the covariates. Hoff and Niu (2012) proposed a regression model where the covariance matrix is expressed as a quadratic function of the explanatory variables. Zou et al. (2017) linked the matrix outcome to a linear combination of similarity matrices derived from the covariates and examined the asymptotic properties of different estimators under this framework. Zhao et al. (2021) introduced the Covariate Assisted Principal (CAP) regression model for multiple covariance matrix outcomes, focusing on identifying linear projections of the covariance matrices associated with the covariates. Kong et al. (2020) and Hu et al. (2021) developed linear and nonparametric regression methods for high-dimensional, low-rank matrices using nuclear norm regularization. However, these approaches often either impose specific structural assumptions on the covariance matrices or require the estimation of a large number of parameters.

In this paper, we focus on general regression analysis where the responses are symmetric positive-definite (SPD) matrices and the predictors are Euclidean within the framework of the Fréchet regression model (Petersen and Müller, 2019). This model defines a global regression function that links response data in an arbitrary metric space with Euclidean predictors. A key consideration in this context is the choice of an appropriate metric for SPD matrices. Various metrics are available, including the Log-Euclidean metric (Arsigny et al., 2007), the Cholesky metric (Dryden et al., 2009) and the Log-Cholesky metric (Lin, 2019). Each of these metrics has distinct properties and implications for regression analysis. However, the Bures-Wasserstein metric W , introduced by Bures (1969), stands out for its desirable mathematical and geometric properties. It is defined for any pair of SPD matrices $A, B \in \mathcal{S}_d^{++}$ as

$$W(A, B) = \left[\text{tr } A + \text{tr } B - 2 \text{tr} \left(A^{1/2} B A^{1/2} \right)^{1/2} \right]^{1/2}. \quad (1)$$

This distance possesses desirable properties under both scaling and rotation: for positive scalar $c > 0$ and orthogonal matrix $O \in \mathcal{O}_d$, it satisfies

$$\begin{aligned} W(cA, cB) &= c^{1/2} W(A, B) \\ W(OAO^\top, OBO^\top) &= W(A, B) \end{aligned}$$

Known as the Bures distance in quantum information theory (Bures, 1969), this metric is equivalent to the Wasserstein distance between two centered Gaussian distributions with specified covariance matrices. The Wasserstein distance, a fundamental concept in optimal transport (OT) theory (Villani, 2003, 2009), which blends optimization, analysis, and geometry, has become highly valuable in various statistical and machine learning applications (Abadie and Imbens, 2006; Deb and Sen, 2023; Arjovsky et al., 2017; Redko et al., 2017; Hallin et al., 2021). This equivalence provides the Bures-Wasserstein distance with a clear and meaningful probabilistic interpretation as the minimal "effort" required to transport one Gaussian distribution, defined by its covariance matrix, to another. In single-cell genomics, most measurement technologies are destructive, meaning that the same cell cannot be observed multiple times or profiled over time. Consequently, measurements at each time point are often modeled as distributions, making optimal transport techniques well-suited for analyzing the associated dynamics (Schiebinger et al., 2019; Bunne et al., 2022; Somnath et al., 2023; Bunne et al., 2023a,b). Given its connection to optimal transport theory, the

Bures-Wasserstein metric is a natural choice for studying gene expression covariance matrices. When equipped with this metric, the space \mathcal{S}_d^{++} becomes a Riemannian manifold, known as the Bures-Wasserstein manifold (Bhatia et al., 2019).

In contrast, the Log-Euclidean metric, the Cholesky metric and the Log-Cholesky metric lack clear practical interpretations and do not naturally align with the structure of distributions or physical models. In addition, the Log-Euclidean metric d_{LE} is insensitive to scaling; specifically, $d_{LE}(cA, cB) = d_{LE}(A, B)$ for any $c > 0$, making it less suitable for applications where the magnitude of matrices is a critical factor. The Log-Cholesky metric d_{LC} treats the diagonal and off-diagonal components of the Cholesky factor L_A differently, resulting in $d_{LC}(cA, cB)$ not being proportional to $d_{LC}(A, B)$ for $c > 0$, which further limits its applicability. Additionally, both the Cholesky metric d_C and the Log-Cholesky metric d_{LC} lack rotational invariance; specifically, for $O \in \mathcal{O}_d$,

$$\begin{aligned} d_C(OAO^\top, OBO^\top) &\neq d_C(A, B), \\ d_{LC}(OAO^\top, OBO^\top) &\neq d_{LC}(A, B). \end{aligned}$$

These limitations reduce the practical utility of these metrics for applications involving SPD matrices where both scaling and rotation are essential, such as in aligning data from different sources (Ma et al., 2024).

Using the Bures-Wasserstein distance as the metric for SPD matrices in the Fréchet regression model, we consider independent and identically distributed (i.i.d.) pairs of predictor and response variables $(X_1, Q_1), \dots, (X_n, Q_n) \in \mathbb{R}^p \times \mathcal{S}_d^{++}$. The objective is to conduct inference, particularly to test the effect of the covariate X on the response variable Q . Previous research on the Fréchet regression model has primarily focused on consistency in the asymptotic regime (Petersen and Müller, 2019; Chen and Müller, 2022), with inference being considered only in the specific case of one-dimensional (1D) density curves as response variables under the Wasserstein metric (Petersen et al., 2021). Notably, for any pair of 1D distributions $\mu, \nu \in \mathcal{P}(\mathbb{R})$ with distribution functions F_μ, F_ν , the 2-Wasserstein distance between them is given by $\left[\int_0^1 |F_\mu^{-1}(t) - F_\nu^{-1}(t)|^2 dt \right]^{1/2}$. Therefore, the 1D Wasserstein space is unique in that it has zero sectional curvature (Ambrosio et al., 2005), and can be *isometrically* embedded into a Hilbert space. This flat geometry allows for closed-form expressions in 1D Fréchet inference and provides well-understood solutions to various problems (Panaretos and Zemel, 2016; Chen et al., 2023; Bigot et al., 2017). However, most metric spaces of interest are nonlinear and exhibit nonzero curvature. For example, the Wasserstein space of distributions in d dimension is positively curved for $d > 1$ (Ambrosio et al., 2005), with the Bures-Wasserstein manifold being a special case. Consequently, *isometric* embeddings of such curved metric spaces into a Hilbert space cannot be assumed, making it difficult to derive distributional results for Fréchet regression in these spaces. This complexity poses additional challenges for statistical inference, particularly in providing guarantees on significance levels and power.

1.1 Main contribution

We focus on Fréchet regression on the Bures-Wasserstein manifold, and our main contributions are threefold.

First, we establish a non-asymptotic \sqrt{n} -rate of convergence (up to $\log n$ factors) for the regression estimate $\hat{Q}(x)$ uniformly over the region $\|x\| \lesssim \sqrt{\log n}$. To the best of our knowledge, this is the first non-asymptotic uniform convergence result for Fréchet regression over a potentially diverging region. Beyond the standard assumptions of light tails for X and Q , we only require well-separation and a local curvature lower bound, which are mild conditions and are verified in a simple case. These results are crucial for later deriving the asymptotic null distribution and power of our proposed test for the association between covariance matrices and covariates.

Second, we derive a central limit theorem for the point estimate $\hat{Q}(x)$, leading to the construction of a pointwise confidence region. The covariance operator of the limiting Gaussian distribution is shown to have contributions from two sources: the intrinsic variability in Q and the imperfect information regarding X .

Third, we carefully construct a test statistic with a tractable asymptotic null distribution, which is represented as a weighted sum of χ_p^2 distributions. The weights are determined by the covariance of the tangent vector, which can be interpreted as a generalization of classical noise variance. The proposed test is also shown to be powerful against a sequence of contiguous alternatives. To the best of our knowledge, this is the first test developed for Fréchet regression on a space with nonzero sectional curvature.

Finally, we validate our theoretical results through numerical simulations and real applications.

1.2 Related works

Statistical OT In addition to advances in computational optimal transport (OT) (Cuturi, 2013; Peyré and Cuturi, 2019; Altschuler et al., 2017), there has been a growing interest in the statistical aspects of OT, where the stability of estimated densities (Weed and Berthet, 2019), Wasserstein distances (Barrio and Loubes, 2019; Mena and Niles-Weed, 2019; del Barrio et al., 2023; Altschuler et al., 2022), transport maps (Hütter and Rigollet, 2021; Pooladian and Niles-Weed, 2022; Manole et al., 2022; Gonzalez-Sanz et al., 2022; Pooladian et al., 2023; Manole et al., 2023), and Fréchet means (Agueh and Carlier, 2011; Kim and Pass, 2017; Le Gouic and Loubes, 2017; Le Gouic et al., 2022; Altschuler et al., 2021) are investigated in the presence of sampling noise.

For the Wasserstein Fréchet mean, Le Gouic et al. (2022) established a parametric rate of convergence for the empirical Fréchet mean in the more general Alexandrov spaces, which include the 2-Wasserstein space as a special case, by introducing a bi-extendibility condition that translates into regularity conditions on the Kantorovich potentials. This condition was later relaxed by Chewi et al. (2020) when establishing the linear rate of convergence for gradient descent algorithms over the Wasserstein space.

Taking this further, Panaretos and Zemel (2016) and Agueh and Carlier (2017) established central limit theorems for the empirical Fréchet mean of 1D distributions. Subsequently, a central limit theorem for multivariate Gaussians was established by exploiting the first-order differentiability of optimal transport maps (Kroshnin et al., 2021).

Fréchet mean The Fréchet mean is a natural generalization of the concept of an average to abstract metric spaces. For its properties in general curved metric spaces, see Ohta (2012); Yokota (2016); Le Gouic et al. (2022) and references therein. The existence and

uniqueness of the Fréchet mean in the context of Riemannian manifolds and Wasserstein spaces are established in Agueh and Carlier (2011); Kim and Pass (2017); Le Gouic and Loubes (2017). The asymptotic properties of the empirical Fréchet mean on a Riemannian manifold are addressed in Bhattacharya and Patrangenaru (2003, 2005); Le Gouic et al. (2022).

Fréchet regression The Fréchet regression model can be viewed as an extension of the Fréchet mean by incorporating a weighted average and was first introduced by Petersen and Müller (2019). Petersen et al. (2021) proposed an F-test specifically for the case of 1D density responses. In their approach, uniform convergence was not necessary for inference because the problem could be embedded into a Hilbert space, allowing for explicit expressions. However, in our case, demonstrating uniform convergence is crucial to ensure that the contributions from the remainder term in the Taylor expansion are negligible.

It is worth noting that while the uniform convergence of Fréchet regression is also addressed in Petersen and Müller (2019) and Chen and Müller (2022), their results are asymptotic and only uniform over a fixed compact set of x . In contrast, our result is non-asymptotic, with uniformity achieved within a compact set that expands in diameter to accommodate the potential unboundedness of $\text{supp } X$.

Regression models on manifolds It is also important to note that the Fréchet regression model is defined purely in terms of distance, making it applicable to any abstract metric space. Meanwhile, a separate line of research focuses on regression on manifolds (Yuan et al., 2012; Cornea et al., 2017; Lin et al., 2023; Chen et al., 2023). These regression models are grounded in the concept of tangent spaces from differential geometry. Because tangent spaces are linear, they allow regression on manifolds to essentially reduce to classical linear regression within these tangent spaces.

1.3 Organization

The remainder of the paper is organized as follows. In Section 2, we provide the necessary background on optimal transport. Section 3.1 formulates the Fréchet regression model, followed by the assumptions outlined in Section 3.2. The main results are presented in Section 4, where we demonstrate the uniform convergence of our estimator in Section 4.1 and propose our test along with theoretical guarantees in Section 4.2. Then, Section 5 presents a Riemannian gradient descent algorithm and numerical simulations to validate our theory. Finally, Section 6 applies our test to a single-cell gene expression dataset, revealing age-related changes in gene co-expression networks. The proofs of our theorems and technical lemmas are provided in the Appendix. We conclude with a discussion in Section 7.

1.4 Notation

We denote by \mathbb{Z} and \mathbb{R}_+ the set of integers and the set of non-negative real numbers. For any $a, b \in \mathbb{R}$, we write $a \vee b = \max(a, b)$ and $a \wedge b = \min(a, b)$. For any $x > 0$, we write $\log^+(x) := \log(x) \vee 1$. For any integer $K \geq 1$, $[K] = \{1, \dots, K\}$. Given $z_i \in \mathbb{R}$ for $i \in [n]$, the set $\{z_1, \dots, z_n\}$ is denoted by z_1^n . The Euclidean norm on \mathbb{R}^p is denoted $\|\cdot\|$. For any $x \in \mathbb{R}^p$ and $L > 0$, let $B_x(L) = B(x, L) = \{y \in \mathbb{R}^d : \|x - y\| \leq L\}$. We denote by $\mathcal{S}_d, \mathcal{S}_d^+, \mathcal{S}_d^{++}$ the set of all $d \times d$ symmetric, positive semi-definite and positive definite matrices. For any real

$a < b$, we define $\mathcal{S}_d(a, b) := \{A \in \mathcal{S}_d : aI_d \prec A \prec bI_d\}$. The subscript d is omitted when it's clear from context. Given any $A \in \mathcal{S}_d$, denote the largest and smallest eigenvalue of A by $\lambda_{\max}(A)$ and $\lambda_{\min}(A)$. The Frobenius norm and operator norm of a matrix A are denoted by $\|A\|_F$ and $\|A\|_{\text{op}}$. Given any matrix $A \in \mathbb{R}^{m,n}$, let $\text{vec } A \in \mathbb{R}^{mn}$ denote the vectorized A obtained by stacking columns of A . Given a random variable X and $\alpha > 0$, the ψ_α -norm of X , denoted by $\|X\|_{\psi_\alpha}$, is defined in Appendix B.2. The support of a probability distribution is denoted by $\text{supp}(\cdot)$.

Given normed spaces Y and Z , let $L(Y; Z)$ denote the space of all bounded linear operators from Y to Z . Given a function $\phi : Y \rightarrow Z$ and integer $k \geq 0$, the k -th differential $d^k\phi$, its operator norm $\|d^k\phi\|$ and symmetric norm $\|d^k\phi\|$ are defined in Section 2.

Finally, the quantities C and c will refer to constants whose values may change from line to line. Given sequences $(a_n)_{n=1}^\infty$ and $(b_n)_{n=1}^\infty$, we write $a_n \lesssim b_n$ if there exists $C > 0$ such that $a_n \leq Cb_n$, and we also write $a_n \asymp b_n$ if $b_n \lesssim a_n \lesssim b_n$. The constant C is always permitted to depend on d and other problem parameters when they are clear from context.

2. Preliminaries

We provide a concise overview of fundamental concepts in optimal transport, along with associated differential properties, specifically focusing on the case of centered Gaussian distributions.

Given a Polish space (E, d) , let $\mathcal{P}_2(E)$ denote the collection of all (Borel) probability measures μ on E such that $\mathbb{E}_{X \sim \mu} d(X, y)^2 < \infty$ for some $y \in E$. One can show that the definition of $\mathcal{P}_2(E)$ is independent of the choice of y . We specialize to the case when $E = \mathbb{R}^d$ with Euclidean distance. For any pair of measures $\mu, \nu \in \mathcal{P}_2(\mathbb{R}^d)$, let $\Pi(\mu, \nu)$ be the set of couplings of between μ and ν , that is, the collection of probability measures π on $\mathbb{R}^d \times \mathbb{R}^d$ such that if $(X, Y) \sim \pi$, then $X \sim \mu$ and $Y \sim \nu$. The 2-Wasserstein distance between μ and ν is defined as

$$W(\mu, \nu) := \left[\inf_{\pi \in \Pi(\mu, \nu)} \mathbb{E}_{(X, Y) \sim \pi} \|X - Y\|^2 \right]^{1/2} \quad (2)$$

Here, with a slight abuse of notation, we use W to denote both the Wasserstein distance between two distributions and the Bures-Wasserstein distance between two PSD matrices. This notation is justified by the fact that both distances coincide and have a closed-form expression (1) when we identify centered Gaussian distributions with their covariance matrices.

Let $\mathcal{P}_{2, \text{ac}}(\mathbb{R}^d)$ denote the subset of measures in $\mathcal{P}_2(\mathbb{R}^d)$ that are absolutely continuous with respect to the Lebesgue measure. Given $\mu_0, \mu_1 \in \mathcal{P}_{2, \text{ac}}(\mathbb{R}^d)$, Brenier's theorem guarantees the existence of a unique optimal coupling $\pi^* \in \Pi(\mu_0, \mu_1)$ that achieves the minimum in (2) and that it is induced by the optimal transport map $T_{\mu_0}^{\mu_1} : \mathbb{R}^d \rightarrow \mathbb{R}^d$ in the sense that $T_{\mu_0}^{\mu_1}(X) \sim \mu_1$ whenever $X \sim \mu_0$. Specifically, when μ_0, μ_1 are centered Gaussian distributions with covariance matrices Q, S , the optimal transport map is given by the linear map

$$T_Q^S = S^{1/2} \left(S^{1/2} Q S^{1/2} \right)^{-1/2} S^{1/2} = Q^{-1/2} \left(Q^{1/2} S Q^{1/2} \right)^{1/2} Q^{-1/2} \quad (3)$$

For completeness, additional background on the geometry of optimal transport is provided in Appendix A.1.

Kroshnin et al. (2021) showed that for any fixed $S \in \mathcal{S}_d^{++}$, T_Q^S is (Fréchet) differentiable with respect to Q , with the differential at Q denoted by dT_Q^S . They also showed that for fixed $S \in \mathcal{S}_d^{++}$, the squared Wasserstein distance $W^2(\cdot, S) : \mathcal{S}_d^{++} \rightarrow \mathbb{R}$ is twice differentiable, with the corresponding 1st and 2nd differential $dW^2(Q, S), d^2W^2(Q, S)$ satisfying

$$\begin{aligned} dW^2(Q, S)(X) &= \langle I - T_Q^S, X \rangle \\ d^2W^2(Q, S)(X, Y) &= -\langle X, dT_Q^S(Y) \rangle \end{aligned} \quad (4)$$

In this paper, higher-order differentials $d^k T_Q^S$ are essential for the development of the theory. Hence, additional background on functional calculus is provided in Appendix A.2 for self-containedness.

3. Problem Formulation

In this section, we formulate the Fréchet regression model in Section 3.1 and outline the assumptions in Section 3.2.

3.1 Fréchet regression on Bures-Wasserstein manifold

Given a metric space (\mathcal{Y}, d) , let \mathbb{P} be a probability distribution over $\mathbb{R}^p \times \mathcal{Y}$ that generates random pairs (X, Y) . The primary challenge in formulating a regression model between Y and X lies in the fact that the metric-space-valued response Y does not lend itself to linear operations. The Fréchet regression model (Petersen and Müller, 2019) seeks to generalize classical linear regression to a general metric space, building upon the concept of the conditional Fréchet mean, which we introduce first.

The unconditional Fréchet mean $\mathbb{E}_{\text{Fréchet}} Y$ generalizes the notion of an average in a general metric space (\mathcal{Y}, d) and is defined as:

$$\mathbb{E}_{\text{Fréchet}} Y := \operatorname{argmin}_{y \in \mathcal{Y}} \mathbb{E}_Y d^2(y, Y) \quad (5)$$

The above definition is motivated by the fact that when \mathcal{Y} is a Euclidean space, $\mathbb{E}_{\text{Fréchet}} Y$ coincides with the classical expectation $\mathbb{E}Y$. Similarly, the conditional Fréchet mean, denoted by $\mathbb{E}_{\text{Fréchet}} [Y|X = x]$, is defined as

$$\mathbb{E}_{\text{Fréchet}} [Y|X = x] := \operatorname{argmin}_{y \in \mathcal{Y}} \mathbb{E}_{Y|X=x} [d^2(y, Y)|X = x] \quad (6)$$

The existence and uniqueness of both the conditional and unconditional Fréchet mean in metric spaces can be ensured under additional assumptions; see Petersen and Müller (2019) and references therein for further details.

With the notion of the conditional Fréchet mean established, the Fréchet regression model proposed by Petersen and Müller (2019) satisfies:

$$\mathbb{E}_{\text{Fréchet}} [Y|X = x] = \operatorname{argmin}_{y \in \mathcal{Y}} \mathbb{E}_{(X, Y) \sim \mathbb{P}} [w(x, X) d^2(y, Y)] \quad (7)$$

where the weight function w is defined as

$$w(x, X) = 1 + (x - \mu)^\top \Sigma^{-1} (X - \mu), \quad \mu = \mathbb{E}X, \Sigma = \text{Var}(X) \quad (8)$$

The objective function $\mathbb{E} [w(x, X)d^2(y, Y)]$ on the right-hand side of (7) generalizes the Fréchet mean (5) by introducing a weighted expectation, similar in spirit to kernel estimators in non-parametric statistics (Wasserman, 2006). The specific choice of the weight $w(x, X)$ in (8) is motivated by the requirement that under the classical linear regression model, the minimizer corresponds to the desired conditional expectation of Y at x . Specifically, when $\mathcal{Y} = \mathbb{R}$ and $\mathbb{E}[Y|X = x] = a^\top(x - \mu) + b$, one can verify that

$$a^\top(x - \mu) + b = \operatorname{argmin}_{y \in \mathbb{R}} \mathbb{E}_{(X, Y)} [w(x, X)(y - Y)^2] \quad (9)$$

When specialized to the Bures-Wasserstein manifold $(\mathcal{Y}, d) = (\mathcal{S}_d^{++}, W)$, we use $Q \in \mathcal{S}_d^{++}$ to denote the response. The existence and uniqueness of both the conditional and unconditional Fréchet mean are guaranteed under mild moment conditions. The following lemma follows from Theorem 2.1 in Kroshnin et al. (2021).

Lemma 1 *On the Bures-Wasserstein manifold (\mathcal{S}_d^{++}, W) , if $\mathbb{E} \operatorname{tr} Q < +\infty$, then the Fréchet mean $\mathbb{E}_{\text{Fréchet}} Q$ exists uniquely in \mathcal{S}_d^{++} . If $\mathbb{E}[\operatorname{tr} Q|X = x] < +\infty$, then the conditional Fréchet mean $\mathbb{E}_{\text{Fréchet}} [Q|X = x]$ exists uniquely in \mathcal{S}_d^{++} .*

We denote $Q^*(x) := \mathbb{E}_{\text{Fréchet}} [Q|X = x]$ as the conditional Fréchet mean of Q given $X = x$, and our Fréchet regression model assumes

$$Q^*(x) = \operatorname{argmin}_{S \in \mathcal{S}_d^{++}} F(x, S), \quad \text{where } F(x, S) := \mathbb{E}_{(X, Q)} [w(x, X)W^2(S, Q)] \quad (10)$$

In general, there is no closed-form expression for $Q^*(x)$, similar to the situation with Bures-Wasserstein barycenters (Agueh and Carlier, 2011). To gain intuition about the types of functional relationships between Q and X that the Fréchet regression model (10) captures, consider a special case where Q is concentrated on matrices that commute with each other. In this setting, for any commuting matrices $S, Q \in \mathcal{S}_d^{++}$, the Bures-Wasserstein distance (1) between Q and S simplifies to $W(Q, S) = \|Q^{1/2} - S^{1/2}\|_F$, which forms an isometric Hilbert embedding of $\operatorname{supp} Q$. Under this simplification, the Fréchet regression model (10) essentially reduces to a linear regression model on the square roots of the covariance matrices in a similar flavor as (9):

$$Q^*(x) = \operatorname{argmin}_{S \in \mathcal{S}_d^{++}} \mathbb{E}_{(X, Q)} \left[w(x, X) \|S^{1/2} - Q^{1/2}\|_F^2 \right]$$

A specific example of this is provided in Example 1 in Section 5.

3.2 Assumptions

To establish rigorous theoretical guarantees for estimation and hypothesis testing, we impose the following model assumptions, which we believe are both theoretically minimal and sufficiently general.

We start with conditions on the marginal distribution of the covariate X and the conditional distribution of Q given X , as outlined in Assumption 1 and 2.

Assumption 1 X is sub-Gaussian with $\|X\|_{\psi_2} \leq C_{\psi_2}$ and $\lambda_{\min}(\Sigma) \geq c_{\Sigma}$ for some constants $C_{\psi_2}, c_{\Sigma} > 0$.

Assumption 2 Given $X = x \in \text{supp } X$, the eigenvalues of Q are bounded away from 0 and infinity in the sense that

$$\mathbb{P}(Q \in \mathcal{S}_d(\gamma_{\Lambda}(\|x - \mu\|)^{-1}, \gamma_{\Lambda}(\|x - \mu\|)) | X = x) = 1 \quad (11)$$

where $\gamma_{\Lambda} : \mathbb{R}^+ \rightarrow \mathbb{R}^+$ is defined as

$$\gamma_{\Lambda}(t) := c_{\Lambda}(t \vee 1)^{C_{\Lambda}} \quad (12)$$

for some constants $c_{\Lambda} \geq 1$ and $C_{\Lambda} \geq 0$.

Assumption 2 implies upper and lower bounds on both the conditional expectation $Q^*(x)$ (see Lemma 33) and noise. These bounds become progressively weaker as $x \rightarrow \infty$. In the literature on covariance matrix estimation, an upper bound on the population covariance matrix is commonly assumed (Cai et al., 2010). The lower bound on $\lambda_{\min}(Q^*(x))$ here is analogous to the uniform upper bound on the conditional densities in Fréchet regression for 1D density response curves (Petersen et al., 2021, Assumption T4), since density is inversely proportional to the standard deviation in a 1D location-scale family. Assumptions on both upper and lower bounds are natural in the context of optimal transport. For example, Hütter and Rigollet (2021), Manole et al. (2022), and Pooladian and Niles-Weed (2022) assumed smoothness and strong convexity of the Brenier potential for optimal transport map estimation, which translates to upper and lower bounds on the eigenvalues of the covariance matrix in the Gaussian case. Similarly, Altschuler et al. (2021) assume both upper and lower bounds on the eigenvalues to ensure a variance inequality proposed in Chewi et al. (2020), which is critical for proving the \sqrt{n} -convergence of the empirical barycenter and the linear convergence of a gradient descent algorithm on the Bures-Wasserstein manifold.

Remark 2 The bounds presented here depend on $\gamma_{\Lambda}(\|x - \mu\|)$, which diverges as $x \rightarrow \infty$. This is motivated by the behavior of the conditional mean $\mathbb{E}(Y|X = x)$ in classical linear regression, which also diverges as $x \rightarrow \infty$. Specifically, when $\mu = 0$ and $\mathbb{E}[Y|X] = a^{\top}X + b$, it can be shown that $|\mathbb{E}[Y|X]| \leq \gamma_{\Lambda}(X)$ with $c_{\Lambda} = \|a\| + |b|$ and $C_{\Lambda} = 1$. In our context, we assume a polynomial growth rate $\gamma_{\Lambda}(t) \lesssim t^{C_{\Lambda}}$, a condition permitted by (12) and commonly satisfied in practical applications.

Remark 3 The bounded noise assumption can be relaxed by assuming that Q has a light tail conditional on $Q^*(X)$. Specifically, we can assume that

$$\mathbb{P}\left\{\frac{\lambda_{\min}(Q^*(x))}{\lambda_{\min}(Q)} \vee \frac{\lambda_{\max}(Q)}{\lambda_{\max}(Q^*(x))} > t \middle| X = x\right\} \lesssim \exp(-ct^{\alpha}), \quad \forall t > 0$$

for some constant $\alpha > 0$. The proof presented in our paper remains valid by incorporating additional concentration arguments.

Assumption 1 and 2, together with Lemma 1, imply that the conditional Fréchet mean $Q^*(x)$ is well-defined. Based on this, we consider the Fréchet regression model given by (10).

Assumption 3 *The Fréchet regression model holds. In other words, for any $x \in \text{supp } X$, equation (10) is satisfied.*

To ensure estimation consistency, some regularity conditions on $F(x, \cdot)$ are needed. Specifically, Assumption 4 below concerns the global behavior of $F(x, \cdot)$ outside local neighborhoods around its minimizer $Q^*(x)$, while Assumption 5 focuses locally on the lower bound of the eigenvalue of the second differential of $F(x, \cdot)$ at $Q^*(x)$.

Assumption 4 *For any $x \in \text{supp } X$, $F(x, \cdot) : \mathcal{S}_d^{++} \rightarrow \mathbb{R}$ has a unique minimizer. Moreover, there exist constants $\alpha_F \geq 1$ and $\delta_F > 0$ such that for any $x \in \text{supp } X$ and any (δ, Δ) that satisfies $0 \leq \delta \leq \delta_F \leq \Delta$, the following*

$$\inf \{F(x, S) - F(x, Q^*(x)) : \delta \leq \|S - Q^*(x)\|_F \leq \Delta\} \geq \frac{\delta^{\alpha_F}}{\gamma_F(\|x - \mu\|, \Delta)} \quad (13)$$

holds where $\gamma_F : \mathbb{R}^+ \times \mathbb{R}^+ \rightarrow \mathbb{R}^+$ is defined as

$$\gamma_F(t_1, t_2) = c_F (t_1 \vee 1)^{C_F} (t_2 \vee 1)^{C_F}$$

for constants $c_F \geq 1, C_F \geq 0$.

Remark 4 *Assumption 3 states that $Q^*(x)$ is a minimizer of $F(x, \cdot)$, while Assumption 4 further assumes that it is the unique minimizer to guarantee the consistency of our M-estimation-type estimator (15). Additionally, condition (13) is motivated by the well-separated-maximizer assumption in M-Estimation (Van Der Vaart and Wellner, 1996, Lemma 3.2.1), which is also assumed in (Petersen and Müller, 2019). In the special case when X and Q are independent, Lemma 34 establishes that Assumption 4 holds with $\delta_F = 1$, $\alpha_F = 2$, $C_F = 1$ and some constant c_F large enough. Here we allow for a polynomial dependence on $\|x - \mu\|$ in the definition of γ_λ so that (13) is still expected to hold in the general case when X and Q are not independent.*

Assumption 5 *For any $x \in \text{supp } X$, consider the symmetric linear operator*

$$\mathbb{E} \left(-w(x, X) dT_{Q^*(x)}^Q \right) : \mathcal{S}_d \rightarrow \mathcal{S}_d,$$

which is the second differential of $F(x, \cdot)$ as $Q^*(x)$. This operator has a lower bound for its minimum eigenvalue given by:

$$\lambda_{\min} \left(-\mathbb{E} w(x, X) dT_{Q^*(x)}^Q \right) \geq \frac{1}{\gamma_\lambda(\|x - \mu\|)} \quad (14)$$

where $\gamma_\lambda : \mathbb{R}^+ \times \mathbb{R}^+ \rightarrow \mathbb{R}^+$ is defined as

$$\gamma_\lambda(t) := c_\lambda (t \vee 1)^{C_\lambda}$$

for some constants $c_\lambda \geq 1$ and $C_\lambda \geq 0$.

Remark 5 We provide some clarification on the term $\mathbb{E} \left(-w(x, X) dT_{Q^*(x)}^Q \right)$ here. It is equal to the second differential of $F(x, \cdot) : \mathcal{S}_d^{++} \rightarrow \mathbb{R}$ at $Q^*(x)$. Note that under the Fréchet regression model (10), $Q^*(x)$ is assumed to be the minimizer of the objective function $F(x, \cdot) : \mathcal{S}_d^{++} \rightarrow \mathbb{R}$. As a result, the second differential of $F(x, \cdot)$ at $Q^*(x)$ is positive semi-definite. Similar to the non-singular Fisher information assumption in maximum likelihood estimation (van der Vaart, 1998), we further assume that the second differential $\mathbb{E} \left(-w(x, X) dT_{Q^*(x)}^Q \right)$ has positive eigenvalues that are lower-bounded as in (14). For relevant concepts in functional calculus, see Appendix A.2, and for the justification that $\mathbb{E} \left(-w(x, X) dT_{Q^*(x)}^Q \right)$ is equal to the second differential, see Appendix B.1.

Remark 6 In the special case when X and Q are independent, which is a consequence of Assumption 6 and the null hypothesis of no effect (21) below, one can show that (Lemma 34 in Appendix B.3) Assumption 5 holds for $C_\lambda = 0$ and c_λ large enough. Again, dependence on $\|x - \mu\|$ is allowed in order to account for the possible unboundedness of x when X and Q are independent.

Finally, we assume conditional independence between X and Q given $Q^*(X)$ for hypothesis testing. However, this assumption is not necessary for the results on uniform convergence (Theorem 7) or the central limit theorem (Theorem 11).

Assumption 6 X and Q are independent conditionally on $Q^*(X)$.

4. Statistical Inference for Fréchet Regression on Bures-Wasserstein manifold

Building on the assumptions outlined in Section 3, we now focus on hypothesis testing within the Fréchet regression model. We begin by establishing the uniform convergence of the Fréchet regression estimator and proving a central limit theorem in Section 4.1. The uniform convergence is not only theoretically important but also crucial for determining the asymptotic size and power of our proposed test. In Section 4.2.1, we introduce the test statistic, followed by an analysis of its asymptotic null distribution and asymptotic power in Section 4.2.2.

4.1 Estimation under the Fréchet regression model

We define the Fréchet regression estimator as follows. For $\rho = n^{-1}$, let the empirical mean and the regularized covariance estimator be given by

$$\bar{X} = n^{-1} \sum_{i=1}^n X_i, \quad \hat{\Sigma}_\rho = n^{-1} \sum_{i=1}^n (X_i - \bar{X}) (X_i - \bar{X})^\top + \rho I_p.$$

The estimator is then defined as

$$\hat{Q}_\rho(x) := \operatorname{argmin}_{S \in \mathcal{S}_d^{++}} F_{n,\rho}(x, S), \quad F_{n,\rho}(x, S) := \frac{1}{n} \sum_{i=1}^n [w_{n,\rho}(x, X_i) W^2(S, Q_i)] \quad (15)$$

where $w_{n,\rho}(x, X_i) = 1 + (x - \bar{X})\widehat{\Sigma}_\rho^{-1}(X_i - \bar{X})$.

The estimator was initially studied by Petersen and Müller (2019) with $\rho = 0$, where its asymptotic properties in a general metric space were analyzed under additional assumptions on the covering number, based on the theory of M-estimation. Here, we introduce an additional regularization term, ρI_p , to the empirical covariance matrix. This modification offers several benefits: First, similar to ridge regression, the regularization term enhances numerical stability due to the involvement of the estimated precision matrix in $w_{n,\rho}$. Next, as shown in Theorem 7, beyond providing a high-probability bound for the estimation error in (16), it enables us to establish in (17) a uniform non-asymptotic \sqrt{n} -rate of convergence for the expected estimation error defined as:

$$\mathbb{E} \sup_{\|x - \mu\| \lesssim \sqrt{\log n}} \left\| \widehat{Q}_\rho(x) - Q^*(x) \right\|_{\text{F}}.$$

For $\rho = 0$, as studied by Petersen and Müller (2019), it is challenging even to prove the existence of such a moment. However, by setting $\rho = n^{-1}$, we can show that this moment exists and converges uniformly at a parametric rate. Similar results have been reported in the literature on high-dimensional precision matrix estimation (Tony Cai and Luo, 2011; Cai et al., 2016). Furthermore, the inclusion of the regularization term $n^{-1}I_p$ does not affect the asymptotic central limit theorem (Theorem 11) or the asymptotic level and power of our test (Theorem 16, 18). This is because it only introduces a bias on the order of n^{-1} which is negligible compared to the typical stochastic variation of order $n^{-1/2}$. In fact, ρ can be set to any n^{-c} with a constant $c > 1/2$. For simplicity, we fix $\rho = n^{-1}$ unless otherwise specified.

Our results on the non-asymptotic uniform convergence of the estimator are summarized in Theorem 7 below.

Theorem 7 *Let $1 \leq L_n \leq C_L \sqrt{\log n}$ for some constant $C_L > 0$. Suppose Assumption 1-5 hold. Then*

1. *for $\rho \in \{0, n^{-1}\}$ and any arbitrarily large fixed constant $\tau > 0$, with probability at least $1 - C_3 n^{-\tau}$, we have*

$$\sup_{\|x - \mu\| \leq L_n} \left\| \widehat{Q}_\rho(x) - Q^*(x) \right\|_{\text{F}} \leq C_1 \frac{\log^{C_2} n}{\sqrt{n}} \quad (16)$$

for some constants $C_1, C_2, C_3 > 0$ independent of n .

2. *for $\rho = n^{-1}$, we have*

$$\mathbb{E} \sup_{\|x - \mu\| \leq L_n} \left\| \widehat{Q}_\rho(x) - Q^*(x) \right\|_{\text{F}} \leq C_4 \frac{\log^{C_5} n}{\sqrt{n}} \quad (17)$$

for some constants $C_4, C_5 > 0$ independent of n .

Remark 8 *The factor $\sqrt{\log n}$ in the upper bound of diameter L_n can be refined to $\log^{\xi_L} n$ for an arbitrarily large fixed constant $\xi_L > 0$. One implication of Theorem 7 is that for $\rho \in \{0, n^{-1}\}$, we have*

$$\max_{i \in [n]} \left\| \widehat{Q}_\rho(X_i) - Q^*(X_i) \right\|_{\text{F}} \lesssim \frac{\text{polylog}(n)}{\sqrt{n}} \quad (18)$$

with probability at least $1 - O(n^{-\tau})$ for any fixed $\tau > 0$. This result is crucial for deriving the asymptotic size and power of our proposed test. Petersen and Müller (2019) investigated asymptotic uniform convergence over a fixed compact set. In our context, their results state that for a fixed $L > 0$,

$$\sup_{\|x-\mu\| \leq L} W\left(\widehat{Q}_0(x), Q^*(x)\right) = O_p\left(n^{-\frac{1}{2(\alpha'-1)}}\right)$$

for any $\alpha' > 2$. However, their results do not guarantee (18) unless the norm of the covariate X is bounded by L , and therefore are insufficient for deriving the size and power of our test when X has unbounded support. Consequently, we employ non-asymptotic bounds and derive (18) from (16).

Remark 9 We use the Frobenius norm to quantify error here instead of the Wasserstein distance for the following reasons:

- *Differential properties of $W^2(\cdot, \cdot)$:* The Wasserstein distance is used in the objective function (10) due to its favorable properties under scaling and rotation, as well as its meaningful probabilistic interpretation, as discussed in Section 1. After adopting the Bures-Wasserstein metric in our formulation of the Fréchet regression model, we need to leverage the differential properties of this metric to establish the statistical properties of the estimator. These differential properties are more naturally expressed using the Frobenius norm. For instance, for any $Q, \tilde{Q} \in \mathcal{S}_d^{++}$ and any $A \in \mathcal{S}_d$, we have

$$d_Q W^2(Q, \tilde{Q})(A) = \left\langle I_d - T_Q^{\tilde{Q}}, A \right\rangle$$

where the Frobenius inner product is used, and $T_Q^{\tilde{Q}}$ is the optimal transport map from the normal distribution $N(0, Q)$ to $N(0, \tilde{Q})$. Similarly, existing Taylor expansions for the Bures-Wasserstein metric are also formulated in terms of the Frobenius norm. Hence, it is more straightforward to express the estimation error using the Frobenius norm. For further details on the differential properties of W , see Appendix B.1.

- *Relationship Between Frobenius Norm and Wasserstein Distance:* The Frobenius norm and the Wasserstein distance are closely related. Specifically, Lemma 21 in Appendix B.1 shows that for any $Q, \tilde{Q} \in \mathcal{S}_d^{++}$, the Wasserstein distance $W(Q, \tilde{Q})$ can be upper-bounded in terms of their Euclidean distance. Consequently, we can derive analogous results for the Wasserstein distance, as in Corollary 10 below. In fact, the two rates are equivalent up to a $\text{polylog}(n)$ factor in our setting.

Corollary 10 Let $1 \leq L_n \leq C_L \sqrt{\log n}$ for some constant $C_L > 0$. Suppose Assumption 1-5 hold. Then

1. for $\rho \in \{0, n^{-1}\}$ and any arbitrarily large fixed constant $\tau > 0$, with probability at least $1 - C_{W3}n^{-\tau}$, we have

$$\sup_{\|x-\mu\| \leq L_n} W\left(\widehat{Q}_\rho(x), Q^*(x)\right) \leq C_{W1} \frac{\log^{C_{W2}} n}{\sqrt{n}}$$

for some constants $C_{W1}, C_{W2}, C_{W3} > 0$ independent of n .

2. for $\rho = n^{-1}$, we have

$$\mathbb{E} \sup_{\|x - \mu\| \leq L_n} W \left(\widehat{Q}_\rho(x), Q^*(x) \right) \leq C_{W4} \frac{\log^{C_{W5}} n}{\sqrt{n}}$$

for some constants $C_{W4}, C_{W5} > 0$ independent of n .

To the best of our knowledge, Theorem 7 and Corollary 10 provide the first non-asymptotic uniform convergence results for Fréchet regression over a potentially diverging region. The proof of Theorem 7 is intricate, and we provide an outline here. First, we use an approach similar to Agueh and Carlier (2011) to establish that the largest eigenvalues of the estimates $\widehat{Q}_\rho(x)$ are uniformly bounded from above. Next, we show that $\widehat{Q}_\rho(x)$ converges uniformly at a slow rate by employing the chaining method. Special attention is required here due to the Hölder continuity of the squared distance W^2 ; see Lemma 38 for details. Finally, we achieve a uniform fast rate of convergence by solving a quadratic inequality related to the convergence rate. A detailed proof of Theorem 7 is provided in Appendix C.

Theorem 7 implies the pointwise consistency of the Fréchet regression estimator (15). Moving one step further, we establish a pointwise central limit theorem that is elusive in general metric spaces with nonzero curvature (Petersen and Müller, 2019). In order to present the theorem, we pause to introduce several notations. First, for any $x \in \mathbb{R}^p$, define $\vec{x} := (1 \ x^\top)^\top$. For any random vector $X \in \mathbb{R}^p$ with covariance matrix Σ , denote $\vec{\Sigma} = \mathbb{E} \vec{X} \vec{X}^\top \in \mathbb{R}^{(p+1) \times (p+1)}$. Next, let $[\mathbb{E}(-w(x, X) dT_{Q^*(x)}^Q)]^{-1}$ denote the inverse of $\mathbb{E}(-w(x, X) dT_{Q^*(x)}^Q)$ which is a linear operator in $L(\mathcal{S}_d; \mathcal{S}_d)$ (Appendix B.1). Last, given elements x_1, x_2 of a Hilbert space \mathcal{H} , the tensor product operator $x_1 \otimes x_2 : \mathcal{H} \rightarrow \mathcal{H}$ is defined by $(x_1 \otimes x_2)y = \langle x_1, y \rangle x_2$ for any $y \in \mathcal{H}$; see Hsing and Eubank (2015) for properties of the tensor product operator and its role in the central limit theorem for random elements of a Hilbert space.

With these notations in place, the central limit theorem for the Fréchet regression estimator $\widehat{Q}_\rho(x)$ is stated as Theorem 11.

Theorem 11 *Let $\rho \in \{0, n^{-1}\}$. Suppose Assumption 1-5 hold. Then for any fixed $x \in \text{supp } X$, the following central limit theorem*

$$\sqrt{n} \left[\widehat{Q}_\rho(x) - Q^*(x) \right] \xrightarrow{w} \left[\mathbb{E}_{(X, Q)} \left(-w(x, X) dT_{Q^*(x)}^Q \right) \right]^{-1} Z_x, \quad (19)$$

holds. Here $Z_x \sim \mathcal{N}(0, \Xi_x)$ is a Gaussian random element of $\mathbb{R}^{d \times d}$ with covariance operator Ξ_x equal to $\mathbb{E} V_x \otimes V_x$ where

$$\begin{aligned} V_x &= V_{x,1} + V_{x,2} \\ V_{x,1} &= w(x, X) \left(T_{Q^*(x)}^Q - I_d \right) \\ V_{x,2} &= - \left(\vec{x}^\top \vec{\Sigma}^{-1} (\vec{X} \vec{X}^\top - \vec{\Sigma}) \vec{\Sigma}^{-1} \otimes I_d \right) \cdot \left(\mathbb{E} \vec{X} \otimes (T_{Q^*(x)}^Q - I_d) \right) \end{aligned}$$

Remark 12 Since a linear transformation of a multivariate normal distribution is still normal, Theorem 11 implies that asymptotically the entries of $\sqrt{n}(\widehat{Q}_\rho(x) - Q^*(x))$ jointly follow a multivariate normal distribution in \mathbb{R}^{d^2} when vectorized. More specifically, we have

$$\sqrt{n} \left(\text{vec } \widehat{Q}_\rho(x) - \text{vec } Q^*(x) \right) \xrightarrow{w} \mathcal{N}(0, \Omega_x) \quad (20)$$

Here $\Omega_x = H_x^{-1} \mathbb{E}(\text{vec } V_x)(\text{vec } V_x)^\top H_x^{-1}$ where $H_x \in \mathbb{R}^{d^2 \times d^2}$ is the matrix representing the invertible linear operator $\mathbb{E}(-w(x, X) dT_{Q^*(x)}^Q)$; see Appendix B.1 for the expression of H_x .

Remark 13 Theorem 11 can be utilized to obtain confidence regions for entries of $Q^*(x)$. The asymptotic variance can be estimated using the plug-in method as follows:

$$\widehat{\Xi}_x = \frac{1}{n} \sum_{i=1}^n V_{x,i} \otimes V_{x,i}$$

where

$$\begin{aligned} V_{x,i} &= V_{x,1,i} + V_{x,2,i} \\ V_{x,1,i} &= w_{n,\rho}(x, X_i) \left(T_{\widehat{Q}_\rho(x)}^{Q_i} - I_d \right) \\ V_{x,2,i} &= - \left(\vec{x}^\top \widehat{\Sigma}_\rho^{-1} (\vec{X}_i \vec{X}_i^\top - \widehat{\Sigma}) \widehat{\Sigma}_\rho^{-1} \otimes I_d \right) \cdot \left(\frac{1}{n} \sum_{j=1}^n \vec{X}_j \otimes (T_{\widehat{Q}_\rho(x)}^{Q_j} - I_d) \right) \end{aligned}$$

Similarly, $\left[\mathbb{E}_{(X,Q)} \left(-w(x, X) dT_{Q^*(x)}^Q \right) \right]^{-1}$ can be estimated by

$$\left[\frac{1}{n} \sum_{i=1}^n \left(-w_{n,\rho}(x, X_i) dT_{\widehat{Q}_\rho(x)}^{Q_i} \right) \right]^{-1}$$

In Figure 1 in Section 5.2, the asymptotic variance of each entry of $Q^*(x)$ is extracted using (20) and appropriate indexing.

Theorem 11 shows that the covariance operator Ξ_x has contributions from both $V_{x,1}$ and $V_{x,2}$. If the expectation μ and the covariance matrix Σ of the predictor X are known, and we obtain an estimate $\widehat{Q}_0^*(x)$ of $Q^*(x)$ by directly optimizing an 'oracle' objective function $F_n^*(x, S)$ defined as $F_n^*(x, S) = \frac{1}{n} \sum_{i=1}^n w(x, X_i) W^2(S, Q_i)$, then $\widehat{Q}_0^*(x)$ would follow a central limit theorem with covariance operator exactly equal to $\mathbb{E} V_{1,x} \otimes V_{1,x}$. When μ and Σ are unknown and empirical estimates $\widehat{\mu}$ and $\widehat{\Sigma}$ are plugged in as in (15), there is an additional contribution to Ξ_x from $V_{x,2}$. Theorem 11 is supported by numerical experiments in Section 5, and the proof is given in Appendix D. To the best of our knowledge, the expression for Ξ_x cannot be further simplified in general. However, under the hypothesis that X and Q are independent, which can be a consequence of the null hypothesis in Section 4.2 and Assumption 6, the term $V_{x,2}$ vanishes, leaving Ξ_x with contributions only from $V_{x,1}$. This result is summarized in Corollary 14 below; see Appendix E for the proof.

Corollary 14 *Instate the assumptions in Theorem 11. If X and Q are independent, then*

$$\sqrt{n} \left[\widehat{Q}_\rho(x) - Q^*(x) \right] \xrightarrow{w} \left(\mathbb{E}_{(X,Q)} \left(-w(x, X) dT_{Q^*(x)}^Q \right) \right)^{-1} Z'_x,$$

holds where $Z'_x \sim \mathcal{N}(0, \Xi'_x)$ and

$$\Xi'_x = \mathbb{E} \left[w(x, X) \left(T_{Q^*}^Q - I_d \right) \otimes \left(T_{Q^*}^Q - I_d \right) w(x, X) \right]$$

4.2 Hypothesis testing

In this section, we focus on testing the global null hypothesis of no effects within the Fréchet regression model on the Bures-Wasserstein manifold, defined as follows:

$$\mathcal{H}_0 : Q^*(x) \equiv Q^* \quad \text{for some unknown } Q^*, \quad (21)$$

where $Q^*(x)$ is constant across all x .

We begin by providing the motivation and the formulation of the test statistic in Section 4.2.1. Subsequently, we explore its asymptotic null distribution in Section 4.2.2, followed by an analysis of the asymptotic size and power of the proposed test.

4.2.1 TEST STATISTIC

Note that a key distinction between the Fréchet regression model (10) and classical linear or generalized linear regression models is that the conditional expectation in the Fréchet model is determined by an optimization problem, rather than through a link function. Consequently, there is no parameter analogous to the slope β found in linear models. Therefore, the null hypothesis can only be tested by directly aggregating comparisons between the estimated predictions $\widehat{Q}_\rho(X_i)$ and the Fréchet mean Q^* .

Given the uniform consistency of the Fréchet regression estimator $\widehat{Q}_\rho(x)$ in Theorem 7, we propose the following test statistic for testing the null hypothesis in (21).

$$\widehat{\mathcal{T}}_\rho = \sum_{i=1}^n \left\| \widehat{H}_\rho \cdot \left(\widehat{Q}_\rho(X_i) - \widehat{Q}_\rho(\bar{X}) \right) \right\|_{\mathbb{F}}^2, \quad \text{where} \quad \widehat{H}_\rho = -\frac{1}{n} \sum_{i=1}^n dT_{\widehat{Q}_\rho(\bar{X})}^{Q_i} \quad (22)$$

To develop some intuition about $\widehat{\mathcal{T}}_\rho$, consider that under the null hypothesis (21), the "difference" between $\widehat{Q}_\rho(X_i)$ and Q^* should be small. Since Q^* is unknown, we estimate it by $\widehat{Q}_\rho(\bar{X})$. Note that $\widehat{Q}_\rho(\bar{X})$ coincides with the Fréchet mean of Q_1, \dots, Q_n . Specifically, by definition (15), we have $w_{n,\rho}(\bar{X}, X_i) = 1$, which implies:

$$\widehat{Q}_\rho(\bar{X}) = \operatorname{argmin}_{S \in \mathcal{S}_d^{++}} \frac{1}{n} \sum_{i=1}^n W^2(S, Q_i)$$

This is precisely the Fréchet mean of Q_1, \dots, Q_n . Therefore, a sensible test statistic would be of the form

$$\mathcal{T}_f = \sum_{i=1}^n f \left(\widehat{Q}_\rho(X_i), \widehat{Q}_\rho(\bar{X}) \right)$$

where $f : \mathbb{R}^{d \times d} \times \mathbb{R}^{d \times d} \rightarrow \mathbb{R}$ is some function that measures the "difference" between $\hat{Q}_\rho(X_i)$ and $\hat{Q}_\rho(\bar{X})$. The function f is expected to satisfy $f(\cdot, \cdot) \geq 0$, $f(Q, Q) = 0$ and $f(Q, S) = f(S, Q)$. Assuming tightness of the second-order Taylor approximation, the uniform consistency of $\hat{Q}_\rho(x)$ (Theorem 7) then implies

$$\tau_f \approx \sum_{i=1}^n \left\langle \frac{1}{2} H_f \left(\hat{Q}_\rho(X_i) - \hat{Q}_\rho(\bar{X}) \right), \hat{Q}_\rho(X_i) - \hat{Q}_\rho(\bar{X}) \right\rangle \quad (23)$$

where H_f is the Hessian of f at $\hat{Q}_\rho(\bar{X})$. This justifies the quadratic form of (22).

Furthermore, the specification of H_f should depend on the detailed distribution of $\hat{Q}_\rho(X_i) - \hat{Q}_\rho(\bar{X})$ such that the quadratic form in (23) has a well-defined asymptotic distribution. To derive the asymptotic distribution of $\hat{Q}_\rho(X_i) - \hat{Q}_\rho(\bar{X})$, we rely on the optimality condition $\sum_j w_{n,\rho}(x, X_j)(T_{\hat{Q}_\rho(x)}^{Q_j} - I_d) = 0$. Under the null hypothesis (21), and assuming tightness of the first order approximation of the optimality conditions at X_i and \bar{X} we have

$$\begin{aligned} 0 &\approx \sum_j w_{n,\rho}(X_i, X_j)(T_{Q^*}^{Q_j} - I_d) + \sum_j w_{n,\rho}(X_i, X_j) dT_{Q^*}^{Q_j} (\hat{Q}_\rho(X_i) - Q^*) \\ 0 &\approx \sum_j w_{n,\rho}(\bar{X}, X_j)(T_{Q^*}^{Q_j} - I_d) + \sum_j w_{n,\rho}(\bar{X}, X_j) dT_{Q^*}^{Q_j} (\hat{Q}_\rho(\bar{X}) - Q^*) \end{aligned}$$

Take the difference and rearrange, we arrive at

$$\begin{aligned} \left(-\frac{1}{n} \sum_j dT_{Q^*}^{Q_j} \right) \cdot \sqrt{n}(\hat{Q}_\rho(X_i) - Q^*) &\approx \underbrace{\frac{1}{\sqrt{n}} \sum_j (w_{n,\rho}(X_i, X_j) - 1)(T_{Q^*}^{Q_j} - I_d)}_{a_1(X_i)} \\ &+ \underbrace{\frac{1}{\sqrt{n}} \sum_j (w_{n,\rho}(X_i, X_j) - 1) dT_{Q^*}^{Q_j} (\hat{Q}_\rho(X_i) - Q^*)}_{a_2(X_i)} \end{aligned}$$

It can be shown that $a_2(X_i)$ is negligible compared to $a_1(X_i)$, leading to

$$\left(-\frac{1}{n} \sum_j dT_{Q^*}^{Q_j} \right) \cdot \sqrt{n}(\hat{Q}_\rho(X_i) - Q^*) \approx a_1(X_i).$$

Specifically, the intuition behind the negligibility is as follows. The null hypothesis (21) and Assumption 6 imply that X and Q are independent. As a result, $\frac{1}{\sqrt{n}} \sum_j (w(x, X_j) - 1) dT_{Q^*}^{Q_j}$ has zero expectation and is of order $O_p(1)$ by the central limit theorem, suggesting that

$$\frac{1}{\sqrt{n}} \sum_j (w_{n,\rho}(X_i, X_j) - 1) dT_{Q^*}^{Q_j} = O_p(1)$$

by an approximation argument. Then the consistency of $\hat{Q}_\rho(x)$ implies that $a_2(X_i)$ is of order $o_p(1)$, making it negligible compared to $a_1(X_i)$.

Further calculations confirm a tractable asymptotic distribution for $\sum \|a_1(X_i)\|^2$, which suggests setting $H_f = (-\frac{1}{n} \sum_j dT_{Q^*}^{Q_j}) \otimes (-\frac{1}{n} \sum_j dT_{Q^*}^{Q_j})$. Since Q^* is unknown, H_f is approximated by $\hat{H}_\rho \otimes \hat{H}_\rho$, leading directly to our test statistic \hat{T}_ρ as defined in (22).

Remark 15 *Petersen et al. (2021) proposed the following test statistic for the case where responses are 1D densities, which has a simpler form compared to ours:*

$$\hat{T}_{\rho,1D} = \sum_{i=1}^n W^2(\hat{Q}_\rho(X_i), \hat{Q}_\rho(\bar{X}))$$

with $\rho = 0$. However, their results rely heavily on the isometric Hilbert embedding of the 1D Wasserstein space and do not generalize to higher dimensions. In contrast, our test statistic (22) is applicable in any dimension and is motivated by the Wald statistic used in generalized linear models. Moreover, one can show that \hat{T}_ρ and $\hat{T}_{n,1D}$ are equivalent in the special case of 1D Gaussian distributions. Since our test statistic (22) can also be viewed as a generalization of the numerator of the global F-test in multiple linear regression, we refer to \hat{T}_ρ in (22) as the Wasserstein F-statistic following Petersen et al. (2021).

4.2.2 THEORETICAL PROPERTIES

We now discuss the theoretical guarantees of our test statistic \hat{T}_ρ and the corresponding test. To begin, Theorem 16 provides the asymptotic null distribution of \hat{T}_ρ .

Theorem 16 *Let $\rho \in \{0, n^{-1}\}$. Suppose Assumption 1-3 and 6 hold. Then under the null (21), the test statistic \hat{T}_ρ satisfies*

$$\hat{T}_\rho \xrightarrow{w} \sum_i \lambda_i w_i \quad (24)$$

where w_i are i.i.d. χ_p^2 random variables, and λ_i are the eigenvalues of the following operator:

$$\mathbb{E} \left[\left(T_{Q^*(\mu)}^Q - I_d \right) \otimes \left(T_{Q^*(\mu)}^Q - I_d \right) \right]$$

Remark 17 *The proof of Theorem 16 relies on the uniform consistency established in Theorem 7, since \hat{T}_ρ involves estimated predictions $\hat{Q}_\rho(\cdot)$ at random covariates $\{X_i\}_{i \in [n]}$. In contrast, the uniform consistency is not needed in Petersen et al. (2021) because the Wasserstein space $W_2(\mathbb{R})$ is essentially flat (has zero sectional curvature), which allows for a closed-form expression for $\hat{T}_{n,1D}$. However, the Bures-Wasserstein manifold (S_d^{++}, W) is positively curved when $d > 1$ (Ambrosio et al., 2005), and no closed-form expression is available for \hat{T}_ρ . Thus, we rely on uniform consistency to ensure the tightness of the Taylor approximation. For the proof, see Appendix F.*

Theorem 16 asserts that \hat{T}_ρ converges weakly to a weighted sum of χ_p^2 s with weights determined by the eigenvalues of the covariance operator $\mathbb{E}(T_{Q^*}^Q - I_d) \otimes (T_{Q^*}^Q - I_d)$. To get a corresponding test, note that the asymptotic null distribution in Theorem 16 depends on unknown parameters, namely the eigenvalues λ_i , which must be approximated to formulate

a rejection region. A natural approach would be to estimate the eigenvalues $\hat{\lambda}_i$ of the sample average $\frac{1}{n} \sum_{i=1}^n T_{\hat{Q}_\rho(\bar{X})}^{Q_i} \otimes T_{\hat{Q}_\rho(\bar{X})}^{Q_i}$ and let $\hat{q}_{1-\alpha}$ be the $1 - \alpha$ quantile of $\sum_{i=1} \hat{\lambda}_i w_i$. Then we define our test Φ_α for any $\alpha \in (0, 1)$ by

$$\Phi_{\rho,\alpha} = \mathbf{1} \left(\hat{\mathcal{T}}_\rho > \hat{q}_{1-\alpha} \right) \quad (25)$$

With Theorem 16 established, Proposition 1 below shows that $\Phi_{\rho,\alpha}$ has an asymptotic size of α under the null hypothesis. See Appendix G for the proof.

Proposition 1 *Suppose Assumption 1-3 and 6 hold. Then under the null (21),*

$$\mathbb{P} \left(\hat{\mathcal{T}}_\rho > \hat{q}_{1-\alpha} \right) \rightarrow \alpha$$

as $n \rightarrow \infty$.

Finally, we turn to an analysis of the power of the test $\Phi_{\rho,\alpha}$ under a sequence of contiguous alternatives. To this end, we denote by \mathfrak{P} the set of distributions of (X, Q) that satisfy Assumption 1 - 6,

$$\mathfrak{P} := \{ \mathbb{P} \in \mathcal{P}_2(\mathbb{R}^p \times \mathcal{S}_d^+) : \mathbb{P} \text{ satisfies Assumption 1 - 6} \}$$

For any $\mathbb{P} \in \mathfrak{P}$, we measure the deviation of $Q^*(x)$ from being a constant function of x by $\mathbb{E} \text{dist}^2(Q^*(X), Q^*(\mu))$ where dist is chosen to be either the Wasserstein distance or the Frobenius distance. We define the corresponding alternatives under each distance as follows:

$$\begin{aligned} H_{1,n} : \mathbb{P} \in \mathfrak{P}_F(a_n) &:= \left\{ \tilde{\mathbb{P}} \in \mathfrak{P} : \mathbb{E}_{(X,Q) \sim \tilde{\mathbb{P}}} \|Q^*(X) - Q^*(\mu)\|_F^2 \geq a_n^2 \right\}, \\ \tilde{H}_{1,n} : \mathbb{P} \in \mathfrak{P}_W(a_n) &:= \left\{ \tilde{\mathbb{P}} \in \mathfrak{P} : \mathbb{E}_{(X,Q) \sim \tilde{\mathbb{P}}} W^2(Q^*(X), Q^*(\mu)) \geq a_n^2 \right\}. \end{aligned}$$

Theorem 18 shows that $\Phi_{\rho,\alpha}$ is powerful against both $H_{1,n}$ and $\tilde{H}_{1,n}$ whenever $a_n \gtrsim n^{-(1/2-\alpha_2)}$ for some constant $\alpha_2 > 0$. The proof is provided in Appendix H.

Theorem 18 *Let $\rho \in \{0, n^{-1}\}$. Consider a sequence of alternative hypotheses $H_{1,n}$ with a_n being a sequence such that $a_n \gtrsim \frac{1}{n^{1/2-\alpha_2}}$ for some constant $\alpha_2 > 0$. Then the worst case power converges uniformly to 1, that is*

$$\inf_{\mathbb{P} \in \mathfrak{P}_F(a_n)} \mathbb{P} \left(\hat{\mathcal{T}}_\rho > \hat{q}_{1-\alpha} \right) \rightarrow 1$$

as $n \rightarrow \infty$. The same result also holds for alternative hypotheses $\tilde{H}_{1,n}$ defined by the Wasserstein distance, that is

$$\inf_{\mathbb{P} \in \mathfrak{P}_W(a_n)} \mathbb{P} \left(\hat{\mathcal{T}}_\rho > \hat{q}_{1-\alpha} \right) \rightarrow 1$$

as $n \rightarrow \infty$.

5. Algorithm and Numerical Experiments

In this section, we propose a Riemannian gradient descent algorithm for optimizing (15) in Section 5.1. We then present a series of numerical experiments in Section 5.2 to validate our theoretical results on the central limit theorem (Theorem 11), asymptotic null distribution (Theorem 16) and power (Theorem 18). Additionally, in Section 5.3, we conduct simulations under a setting where Q is not directly observed but is estimated from data, to assess the deviation from the setting with perfect observations.

5.1 Riemannian gradient descent algorithm

Motivated by the Bures-Wasserstein gradient descent algorithm (Chewi et al., 2020; Altschuler et al., 2021) for the vanilla Bures-Wasserstein barycenter, we propose a gradient descent algorithm to compute $\hat{Q}_\rho(x)$ in (15), which is given as Algorithm 1.

Algorithm 1 GD for Fréchet regression

- 1: **Input:** covariates $\{X_i\}_{i=1}^n$, responses $\{Q_i\}_{i=1}^n$, $\rho \in \{0, n^{-1}\}$, covariate x , learning rate η , initialization S_0 , maximum number of iterations T , threshold eps .
- 2: Initialize $S \leftarrow S_0$.
- 3: **for** $t = 1, \dots, T$ **do**
- 4: Set

$$G \leftarrow I_d + \eta \cdot \frac{1}{n} \sum_{i=1}^n w_{n,\rho}(x, X_i)(T_S^{Q_i} - I_d) \quad (26)$$

- 5: Set

$$S \leftarrow \text{GSG} \quad (27)$$

- 6: **if** $\|G\|_F < \text{eps}$ **then**
 - 7: **break**
 - 8: **Output:** S .
-

Algorithm 1 can be viewed as a Riemannian gradient descent algorithm (see Appendix A and Panaretos and Zemel (2016); Chewi et al. (2020); Altschuler et al. (2021)). Intuitively, $-\frac{1}{n} \sum_{i=1}^n w_{n,\rho}(x, X_i)(T_S^{Q_i} - I_d)$ is the derivative of the objective function $F_{n,\rho}(x, S)$ in (15) in the tangent space (Ambrosio et al., 2005, Corollary 10.2.7) and (26) corresponds to one gradient step in the tangent space with step size η . Then (27) is mapping the gradient step in the tangent space back to \mathcal{S}_d^{++} through the exponential map (Appendix A)

The algorithm terminates if the Frobenius norm of the gradient G falls below the threshold eps , indicating that the relative change in the gradient update is less than eps . For the initialization S_0 , optimization over the Euclidean space typically starts near the origin (Chen et al., 2019; Ye and Du, 2021). However, since the space of symmetric positive definite (SPD) matrices, \mathcal{S}_d^{++} , is nonlinear, the natural counterpart of the origin in this space is the identity matrix I_d . Therefore, we initialize at $S_0 = I_d$. In Appendix I, we use numerical simulations to compare this initialization with random initialization and initialization at the mean, and observe that it consistently performs at least as well as these alternatives. For the step size η , Altschuler et al. (2021) observed through numerical simulations that while the convergence rate of Euclidean gradient descent is highly sensitive to its step size, Riemannian gradient descent requires no tuning and works effectively with $\eta = 1$ when computing the Bures-Wasserstein barycenter. In our simulations, we also find that $\eta = 1$ performs at least as well as (and often better than) smaller step sizes. See Appendix I for details.

5.2 Simulation setup and results

To validate our theory and demonstrate the practical applicability of our inferential procedures, we conduct a series of numerical experiments with $\rho = n^{-1}$. We start with two illustrative examples that follow the Fréchet regression model. In Example 1, the covariance matrices share a common eigenspace and commute, which effectively reduces the Fréchet regression model to a linear regression model on the square roots. In contrast, Example 2 considers the case where the covariance matrices do not commute.

Example 1 Let $\{X_i = (X_{i1} \ \cdots \ X_{ip}), i \in [n]\}$ be i.i.d. random covariates in \mathbb{R}^p with $X_i \sim \text{Uniform}[-1, 1]^p$. The response matrices $Q_1, \dots, Q_n \in \mathbb{R}^{d \times d}$ are generated as:

$$Q_i = UV_i f(X_i; \delta)^2 V_i U^\top$$

where U and $\{X_i, V_i\}_{i \in [n]}$ are independent, and:

- $f(\cdot; \delta) : \mathbb{R}^p \rightarrow \mathbb{R}^{d \times d}$ is a mapping from \mathbb{R}^p to diagonal matrices defined by:

$$x = (x_1 \ \cdots \ x_p) \mapsto f(x; \delta) = (f(x; \delta)_{kl})_{k,l \in [d]},$$

where

$$f(x; \delta)_{kk} = 1.5 + \frac{k}{2} + \delta \cdot \sum_{j=1}^p x_j,$$

with $\delta \in (-2p^{-1}, 2p^{-1})$ being a parameter that quantifies the deviation of the model from the null hypothesis (21).

- $U \in \mathcal{O}_d$ is a random orthogonal matrix following the Haar measure.
- $V_1, \dots, V_n \in \mathbb{R}^d$ are random diagonal matrices with i.i.d. diagonal entries $V_{i,kk} \sim \text{Uniform}[-0.1, 0.1]$.

It can be verified that the pair (X, Q) satisfies the Fréchet regression model with the conditional expectation $Q^*(\cdot)$ satisfying

$$Q^*(X_i) = U f(X_i)^2 U^\top$$

Example 2 Let $X = (X_1 \ \cdots \ X_p)$ be a random covariate in \mathbb{R}^p with $X \sim \text{Uniform}[-1, 1]^p$. The response matrix $Q \in \mathbb{R}^{d \times d}$, where d is an even number, is generated as:

$$Q = UVg(X; \delta)^2 VU^\top,$$

where U, V and X are independent, and:

- $g(\cdot; \delta) : \mathbb{R}^p \rightarrow \mathbb{R}^{d \times d}$ is a mapping from \mathbb{R}^p to diagonal matrices defined by:

$$x = (x_1 \ \cdots \ x_p) \mapsto g(x; \delta) = (g(x; \delta)_{kl})_{k,l \in [d]},$$

where

$$g(x; \delta)_{kk} = 1.5 + 0.5 \cdot \lceil k/2 \rceil + \delta \cdot \sum_{j=1}^p x_j,$$

with $\lceil \cdot \rceil$ denoting the ceiling function, and $\delta \in (-2p^{-1}, 2p^{-1})$ being a parameter that quantifies the deviation of the model from the null hypothesis (21).

- $U \in \mathcal{O}_d$ is a random orthogonal matrix with a block-diagonal structure

$$U = \text{diag}(U^{(1)}, \dots, U^{(\lfloor d/2 \rfloor)})$$

where $U^{(1)}, \dots, U^{(\lfloor d/2 \rfloor)}$ are i.i.d. random 2×2 orthogonal matrices following the Haar measure.

- $V \in \mathbb{R}^d$ is a diagonal matrix with i.i.d. diagonal entries $V_{ii} \sim \text{Uniform}[-0.1, 0.1]$.

It can be verified that the pair (X, Q) satisfies the Fréchet regression model with

$$Q^*(x) = g(x)^2$$

With Example 1 and 2 in hand, we proceed to check the validity of the central limit theorem for $\hat{Q}_\rho(x)$ as stated in Theorem 11. To this end, we generate random predictor-response pairs (X, Q) based on Example 1 with parameters $d = 5$, $p = 5$ and $\delta = 0$; and Example 2 with parameters $d = 6$, $p = 5$ and $\delta = 0$. For each trial, we generate $n = 200$ samples of (X_i, Q_i) and obtain the Fréchet regression estimate $\hat{Q}_\rho(x)$ using Algorithm 1 with $\rho = n^{-1}$. We then compute normalized error

$$\left[\tilde{Q}(x) \right]_{ij} = \frac{\sqrt{n} \left[\hat{Q}_\rho(x) - Q^*(x) \right]_{ij}}{\sqrt{\hat{v}_{x,ij}}}$$

where $\hat{v}_{x,ij}$ denotes the plug-in estimate for the asymptotic variance of the (i, j) -entry of $\sqrt{n}(\hat{Q}_\rho(x) - Q^*(x))$, following Theorem 11 and (20). Specifically, the covariance operator Ξ_x in Theorem 11 is estimated as follows:

$$\hat{\Xi}_x = \frac{1}{n} \sum_{i=1}^n V_{x,i} \otimes V_{x,i}$$

where

$$\begin{aligned} V_{x,i} &= V_{x,1,i} + V_{x,2,i} \\ V_{x,1,i} &= w_{n,\rho}(x, X_i) \left(T_{\hat{Q}_\rho(x)}^{Q_i} - I_d \right) \\ V_{x,2,i} &= - \left(\vec{x}^\top \hat{\Sigma}_\rho^{-1} (\vec{X}_i \vec{X}_i^\top - \hat{\Sigma}) \hat{\Sigma}_\rho^{-1} \otimes I_d \right) \cdot \left(\frac{1}{n} \sum_{j=1}^n \vec{X}_j \otimes (T_{\hat{Q}_\rho(x)}^{Q_j} - I_d) \right) \end{aligned}$$

The estimated covariance $\hat{v}_{x,ij}$ can be extracted from $\hat{\Xi}_x$ by appropriately indexing its elements.

Figure 1 presents the Q-Q (quantile-quantile) plots of $\tilde{Q}_{ij}(x)$ against the standard normal distribution, based on 200 Monte Carlo trials at $x = 0$. According to Theorem 11, $\tilde{Q}_{ij}(x)$ should asymptotically follow a standard normal distribution, which would be indicated by a Q-Q plot that exhibits a linear relationship with a slope of one and an intercept of zero. As shown in Figure 1, the empirical quantiles of $\tilde{Q}_{ij}(0)$ align closely with the theoretical quantiles of $\mathcal{N}(0, 1)$, providing strong empirical support for the validity of Theorem 11.

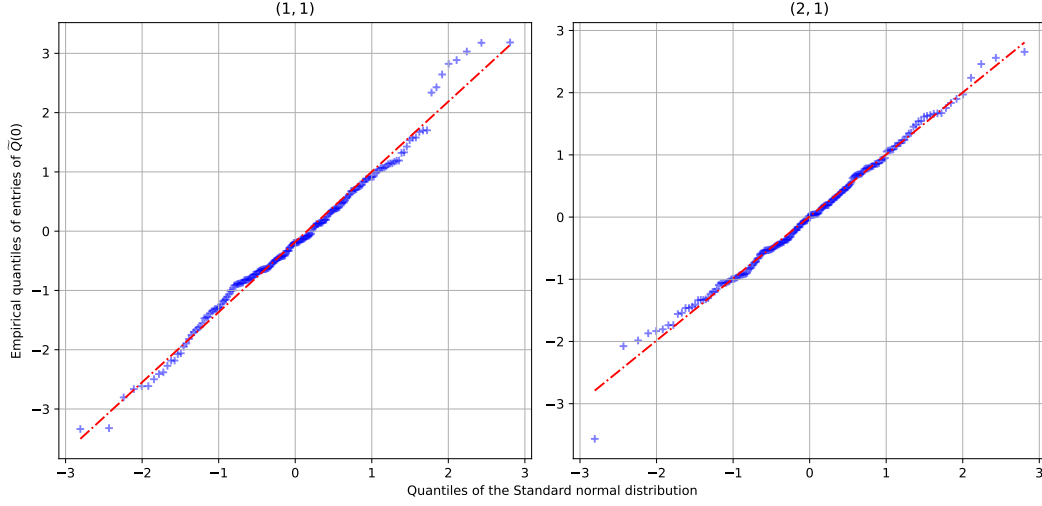
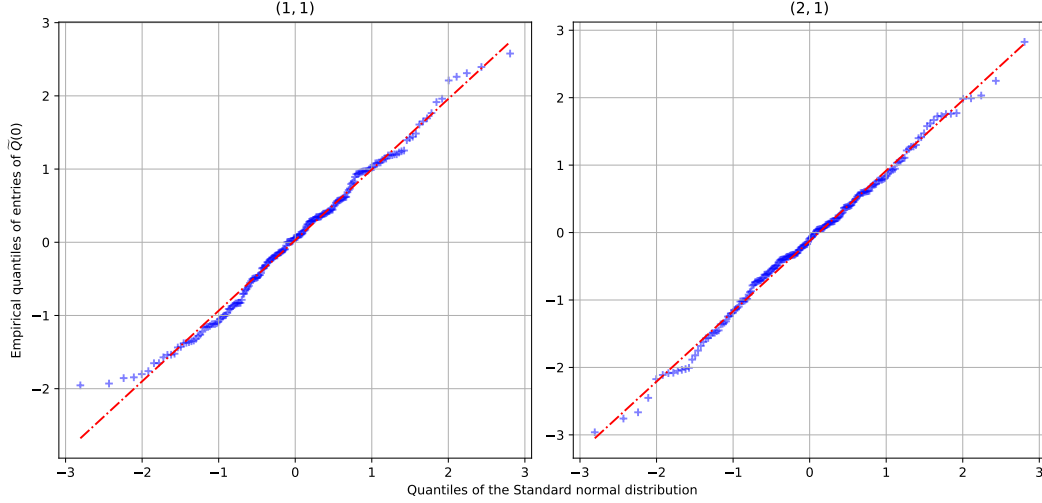

 (a) Example 1, $d = 5$

 (b) Example 2, $d = 6$

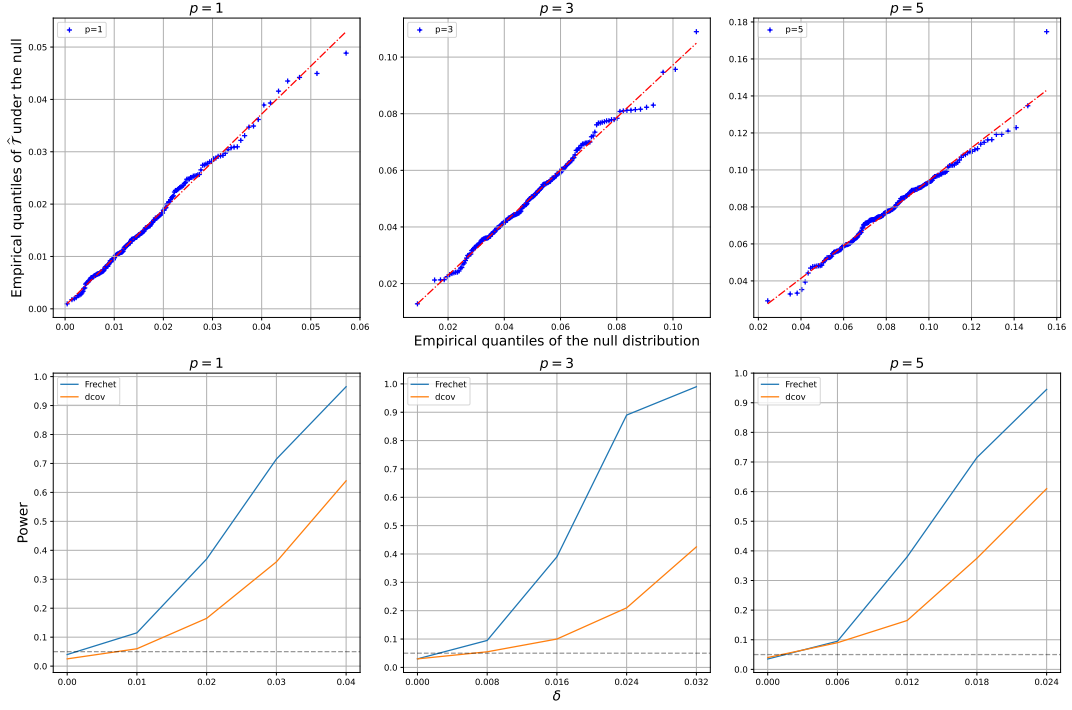
Figure 1: Q-Q plots of $\tilde{Q}_{11}(0)$ and $\tilde{Q}_{21}(0)$ with parameters $p = 5, \delta = 0, n = 200$. (a) Example 1, $d = 5$; (b) Example 2, $d = 6$.

We then turn our attention to Theorem 16, which characterizes the asymptotic null distribution of the test statistic $\hat{\mathcal{T}}_\rho$, and Theorem 18, which focuses on the power of the test. Figure 2 illustrates the results for Example 1 with parameters $n = 200, p \in \{1, 3, 5\}, d \in \{5, 10\}$; Figure 3 presents results for Example 2 with parameters $n = 200, p \in \{1, 3, 5\}, d \in \{6, 10\}$. The Q-Q plots demonstrate a linear fit with slope 1 and intercept 0. Additionally, the figures indicate that as p and d increase, the test statistic also increases. The power of the test quickly approaches 1 as the effect size δ grows. We compare the power of

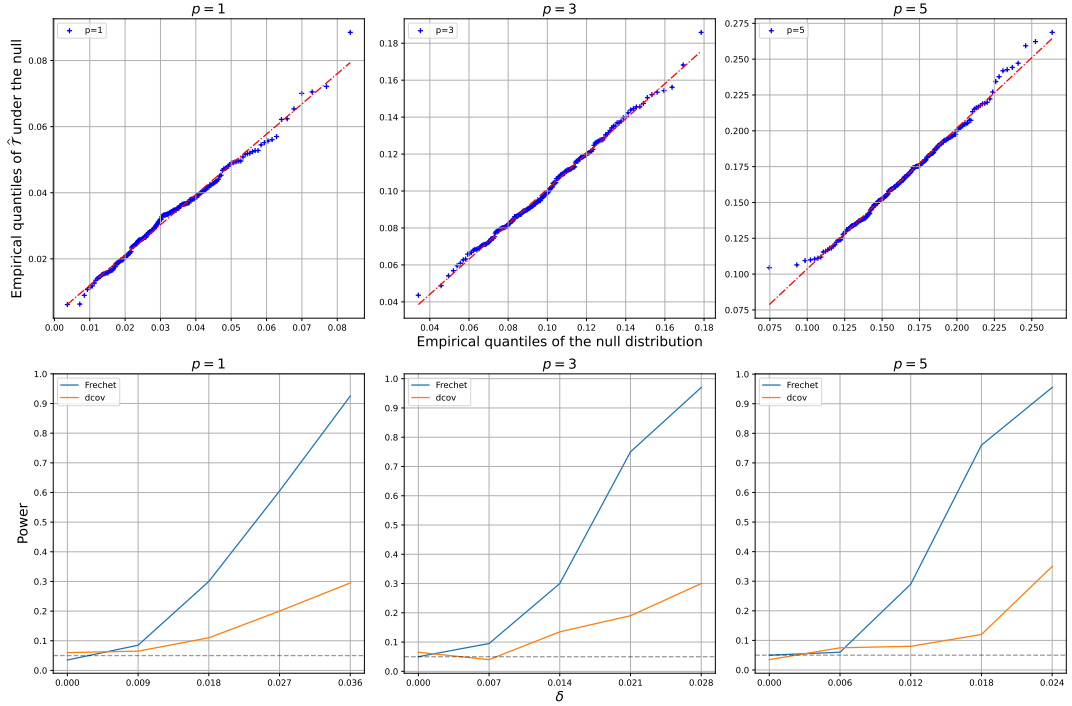
our test with that of the distance covariance test (dcov) (Székely et al., 2007) for testing independence, using the Python dcor package (Ramos-Carreño and Torrecilla, 2023). Given that the distance covariance test is nonparametric, our test is expected to outperform dcov in this context.

In Example 1 and 2 above, the covariate X is supported on the fixed compact domain $[-1, 1]^p$. To illustrate the uniform convergence over a diverging domain $\|x\| \lesssim \sqrt{\log n}$ as stated in Theorem 7, we perform additional experiments with $X \sim \text{Uniform}[-\sqrt{\log n}, \sqrt{\log n}]^p$ for $n \in \{100, 200, 400, 800, 1600\}$. The maximum in-sample estimation error is computed and is observed to scale approximately as $n^{-0.46}$, providing empirical support for the parametric rate of convergence up to logarithmic terms. For more details, see Appendix I.3.

TEST ON THE BURES-WASSERSTEIN MANIFOLD



(a) $d = 5$



(b) $d = 10$

Figure 2: Q-Q plots of the test statistic $\hat{\tau}_\rho$ against its asymptotic null distribution and power curves compared to dcov as a function of the effect size δ for Example 1, with parameters $n = 200$, $p \in \{1, 3, 5\}$. (a) $d = 5$; (b) $d = 10$.₂₅

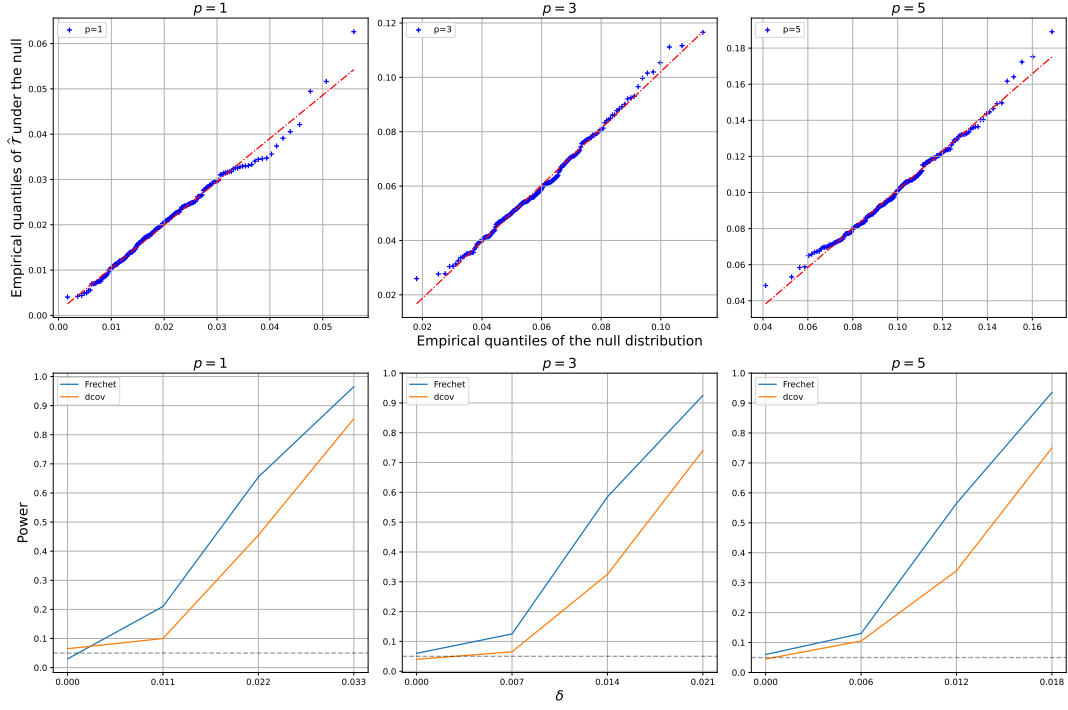
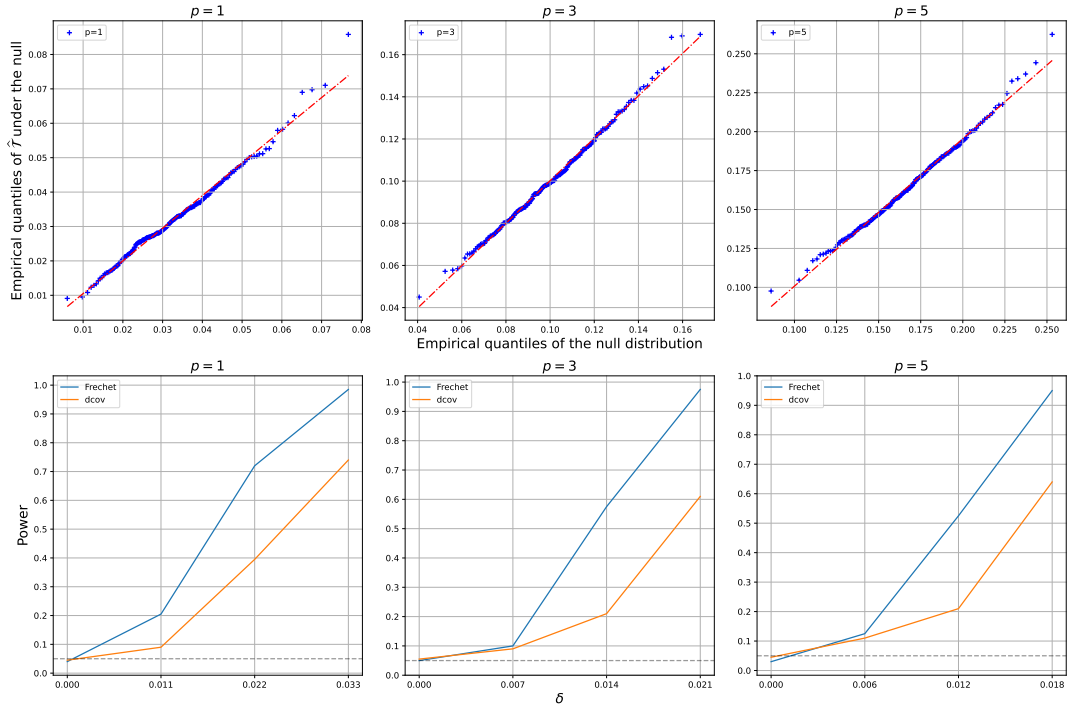

 (a) $d = 6$

 (b) $d = 10$

Figure 3: Q-Q plots of the test statistic \hat{T}_p against its asymptotic null distribution and power curves compared to dcov as a function of the effect size δ for Example 2, with parameters $n = 200$, $p \in \{1, 3, 5\}$. (a) $d = 6$; (b) $d = 10$.₂₆

5.3 Robustness of the results when covariance matrices are estimated

In this section, we examine the numerical performance of our results in a setting where Q is not directly observed but is instead estimated from the data. Specifically, consider the case where we have access only to $\{(X_i; Z_{i1}, Z_{i2}, \dots, Z_{i\tilde{n}})\}_{i \in [n]}$ with $Z_{i1}, Z_{i2}, \dots, Z_{i\tilde{n}} \stackrel{i.i.d.}{\sim} \mathcal{N}(0, Q_i)$. In this scenario, a natural plug-in approach for Fréchet estimation and testing is to estimate Q_i using the sample covariance, defined as $\bar{Q}_i := \tilde{n}^{-1} \sum_{j=1}^{\tilde{n}} Z_{i1} Z_{i1}^\top$. This estimate \bar{Q}_i is then substituted for Q_i in downstream estimators. Specifically, we define the estimator $\hat{Q}_{\rho, \tilde{n}}(x)$ and test statistic $\hat{\mathcal{T}}_{\rho, \tilde{n}}$ with $\rho = n^{-1}$ as follows.

$$\begin{aligned} \hat{Q}_{\rho, \tilde{n}}(x) &= \operatorname{argmin}_{S \in \mathcal{S}_d^{++}} \frac{1}{n} \sum_{i=1}^n w_{n, \rho}(x, X_i) W^2(S, \bar{Q}_i) \\ \hat{\mathcal{T}}_{\rho, \tilde{n}} &= \sum_{i=1}^n \left\| \hat{H}_{\rho, \tilde{n}} \cdot \left(\hat{Q}_{\rho, \tilde{n}}(X_i) - \hat{Q}_{\rho, \tilde{n}}(\bar{X}) \right) \right\|_F^2, \quad \text{where} \quad \hat{H}_{\rho, \tilde{n}} = -\frac{1}{n} \sum_{i=1}^n dT_{\hat{Q}_{\rho, \tilde{n}}(\bar{X})}^{\bar{Q}_i} \end{aligned}$$

The asymptotic null distribution of $\hat{\mathcal{T}}_{\rho, \tilde{n}}$ is computed as $\sum_j \hat{\lambda}_{\tilde{n}, j} w_i$ where w_i are i.i.d. χ_p^2 random variables and $\hat{\lambda}_{\tilde{n}, j}$ are the eigenvalues of the following estimated operator:

$$\frac{1}{n} \sum_{i=1}^n \left(T_{\hat{Q}_{\rho, \tilde{n}}(\bar{X})}^{\bar{Q}_i} - I_d \right) \otimes \left(T_{\hat{Q}_{\rho, \tilde{n}}(\bar{X})}^{\bar{Q}_i} - I_d \right)$$

We denote by $\hat{q}_{1-\alpha, \tilde{n}}$ the $1-\alpha$ quantile of $\sum_j \hat{\lambda}_{\tilde{n}, j} w_i$. This quantile is then used to construct a test:

$$\Phi_{\rho, \tilde{n}, \alpha} = \mathbf{1} \left(\hat{\mathcal{T}}_{\rho, \tilde{n}} > \hat{q}_{1-\alpha, \tilde{n}} \right)$$

Similarly, the covariance operator $\hat{\Xi}_{x, \tilde{n}}$ is computed as

$$\hat{\Xi}_{x, \tilde{n}} = \frac{1}{n} \sum_{i=1}^n V_{x, \tilde{n}, i} \otimes V_{x, \tilde{n}, i} \quad (28)$$

where

$$\begin{aligned} V_{x, i} &= V_{x, \tilde{n}, 1, i} + V_{x, \tilde{n}, 2, i} \\ V_{x, \tilde{n}, 1, i} &= w_{n, \rho}(x, X_i) \left(T_{\hat{Q}_{\rho, \tilde{n}}(x)}^{\bar{Q}_i} - I_d \right) \\ V_{x, \tilde{n}, 2, i} &= - \left(\bar{x}^\top \hat{\Sigma}_\rho^{-1} (\bar{X}_i \bar{X}_i^\top - \hat{\Sigma}) \hat{\Sigma}_\rho^{-1} \otimes I_d \right) \cdot \left(\frac{1}{n} \sum_{j=1}^n \bar{X}_j \otimes (T_{\hat{Q}_{\rho, \tilde{n}}(x)}^{\bar{Q}_j} - I_d) \right) \end{aligned}$$

First, we examine the central limit theorem (Theorem 11) for $(X_i, Q_i)_{i \in [n]}$ satisfying the conditions in Example 1 with parameters $\tilde{n} \in \{50, 100, +\infty\}$, $n = 200$, $p = 5$, $d = 5$ and $\delta = 0$. Here $\tilde{n} = +\infty$ indicates Q_i is directly observed. Using the estimated covariance operator $\hat{\Xi}_{x, \tilde{n}}$ defined in (28), we define the normalized error as:

$$\left[\tilde{Q}_{\rho, \tilde{n}}(x) \right]_{i, j} := \frac{\sqrt{n} \left[\hat{Q}_{\rho, \tilde{n}}(x) - Q^*(x) \right]_{ij}}{\sqrt{\hat{v}_{x, \tilde{n}; ij}}}$$

where $\widehat{v}_{x,\tilde{n};ij}$ is extracted from $\widehat{\Xi}_{x,\tilde{n}}$ with appropriate indexing. Figure 4 shows the Q-Q plots of $[\widehat{Q}_{\rho,\tilde{n}}(x)]_{ij}$ against the standard normal distribution. Generally, as $\tilde{n} \rightarrow \infty$, the Q-Q plots for finite \tilde{n} approach those of $\tilde{n} = \infty$, and can still be approximated by a straight line. This observation suggests that the central limit theorem for the entries still holds. Interestingly, the detailed behavior of the bias differs between diagonal and off-diagonal entries. For off-diagonal entries, Figure 4b indicates that $[\widehat{Q}_{\rho,\tilde{n}}(0)]_{ij}$ is unbiased. In contrast, for diagonal entries, Figure 4a reveals that the Q-Q plots for finite \tilde{n} are negatively shifted compared to that of $\tilde{n} = \infty$ (blue), suggesting that $[\widehat{Q}_{\rho,\tilde{n}}(0)]_{ii}$ is negatively biased. In this paper, we focus on the theoretical properties of the Fréchet regression estimator (15) in the setting where covariance matrices are directly observed. We leave a detailed theoretical investigation of its refined behavior in the setting where covariance matrices are estimated for future work.

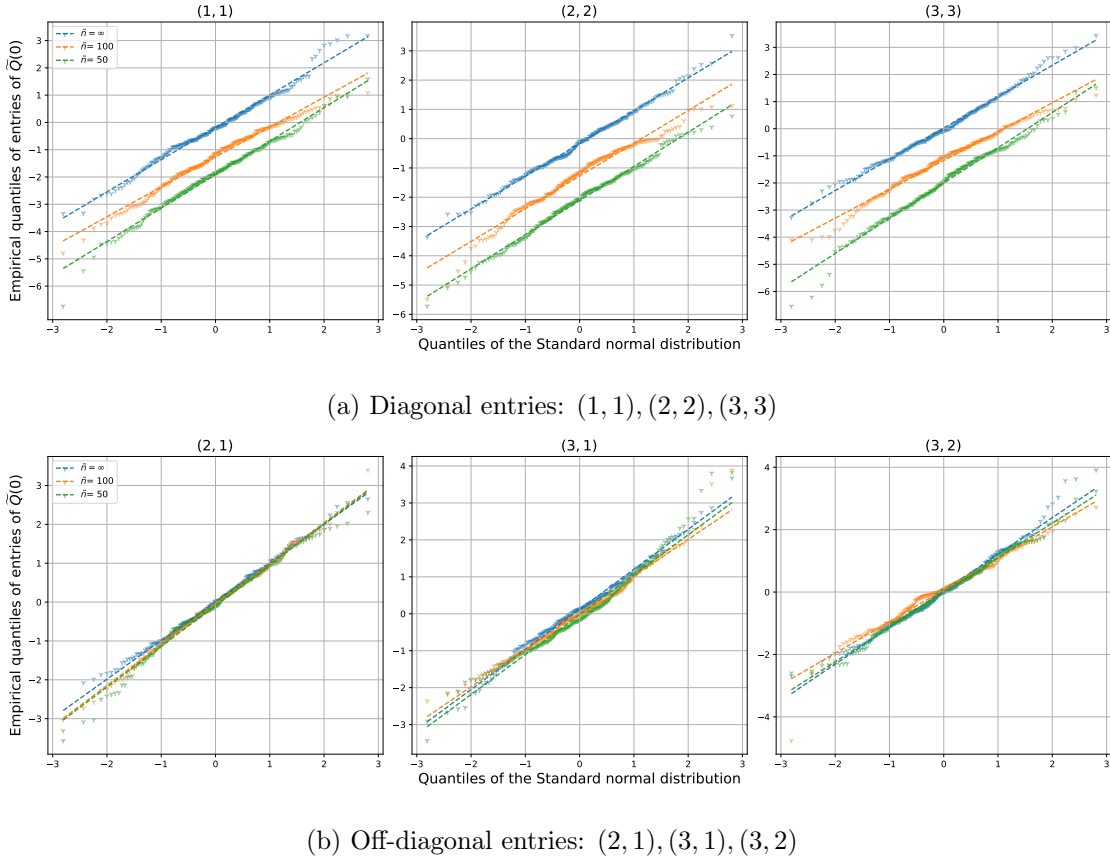


Figure 4: Q-Q plots of diagonal and off-diagonal entries of $\widehat{Q}_{\rho,\tilde{n}}(0)$ with parameters $d = 5, n = 200$ and varying $\tilde{n} \in \{50, 100, \infty\}$.

Figure 5 shows the Q-Q plot of $\widehat{\mathcal{T}}_{\rho,\tilde{n}}$ against $\sum_j \widehat{\lambda}_{\tilde{n},j} w_i$ and the power of the test $\Phi_{\rho,\tilde{n},\alpha}$ for (X_i, Q_i) satisfying the conditions of Example 1 with parameters $\tilde{n} \in \{50, 100, +\infty\}$, $n = 200$, $p = 5$ and $d = 5$. Notably, the Q-Q plots exhibit a linear relationship with slope 1 and intercept 0, indicating that $\widehat{\mathcal{T}}_{\rho,\tilde{n}}$ follows the asymptotic null distribution $\sum_j \lambda_{\tilde{n},j} w_i$,

where w_i are i.i.d. χ_p^2 random variables and $\lambda_{\tilde{n},j}$ are the eigenvalues of the following operator:

$$\mathbb{E} \left[\left(T_{Q^*(\mu)}^{\bar{Q}} - I_d \right) \otimes \left(T_{Q^*(\mu)}^{\bar{Q}} - I_d \right) \right]$$

where $\bar{Q} = \text{Cov } Z$. This observation suggests that the plug-in method remains valid in this context. The power plot indicates that while the power of the test decreases when Q_i is not directly observed, our test still demonstrates higher power compared to the distance covariance test (dcov) in this setting.

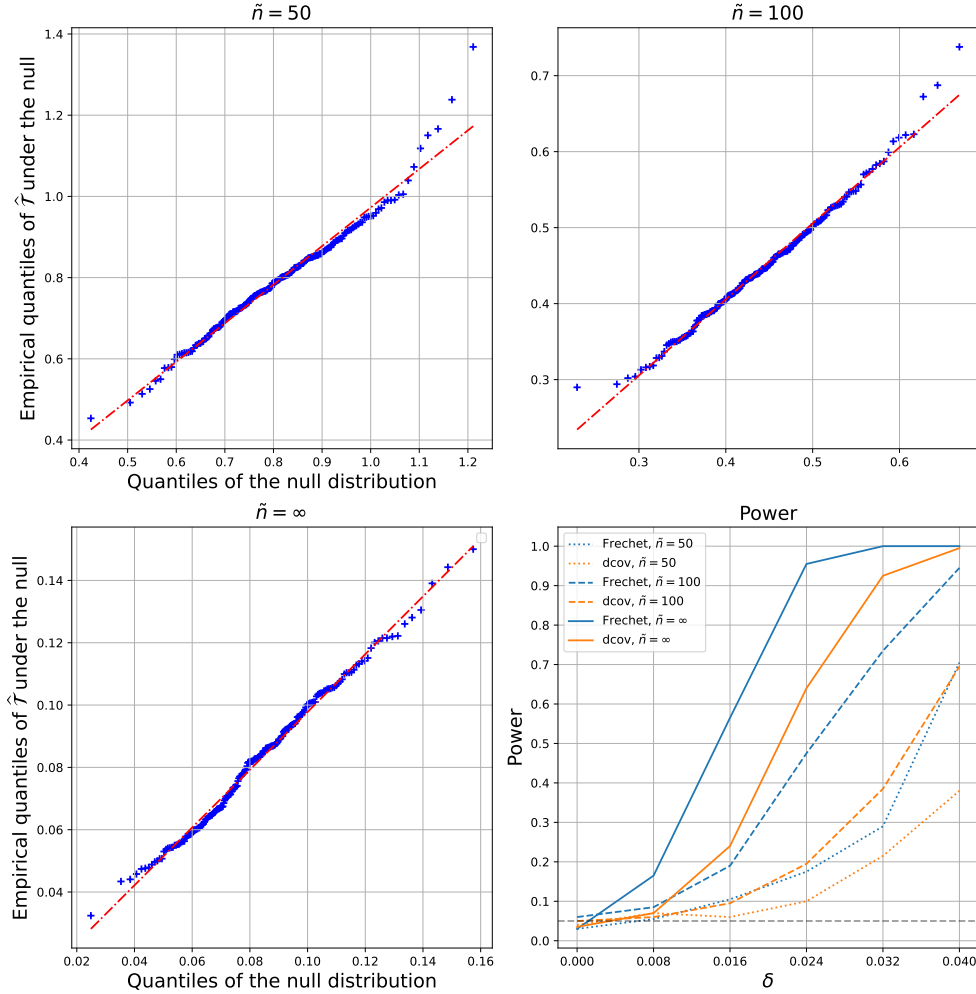


Figure 5: Q-Q plots of the test statistic $\hat{T}_{\rho,\tilde{n}}$ against the estimated plug-in null distribution $\sum_j \hat{\lambda}_{\tilde{n},j} w_i$ and power curves of $\Phi_{\rho,\tilde{n},\alpha}$ compared to dcov as a function of the effect size δ .

6. Application to Single-cell Gene Co-expression Networks

Aging is a complex process of accumulation of molecular, cellular, and organ damage, leading to loss of function and increased vulnerability to disease and death. Nutrient-sensing pathways, namely insulin/insulin-like growth factor signaling and target-of-rapamycin can substantially increase healthy life span of laboratory model organisms (Davinelli et al., 2012; de Lucia et al., 2020). These nutrient signaling pathways are conserved in various organisms. We are interested in understanding the co-expression structure of 61 genes in this KEGG nutrient-sensing pathways based on the recently published population scale single cell RNA-seq data of human peripheral blood mononuclear cells (PBMCs) from blood samples of over 982 healthy individuals with ages ranging from 20 to 90 (Yazar et al., 2022).

We focus our analysis on CD4+ naive and central memory T (CD4NC) cells, which is the most common cell type observed in the data. Age-associated changes in CD4 T-cell functionality have been linked to chronic inflammation and decreased immunity (Elyahu et al., 2019). There are a total of 51 genes that are expressed in this cell type. Even though the Fréchet regression still makes sense when the covariance matrix is potentially degenerate (see the remarks after Example 2), our theory relies on the strict positive definiteness. Hence, we retain only the genes that have nonzero variances at any age, resulting in a total of 37 genes, see Figure 6 for a concise overview of these covariance matrices for individuals at different ages, showing difference across different ages.

In genetics research, such covariance matrices represent individual-specific gene co-expression networks. We are interested in testing whether such networks are associated with ages by testing whether there is an age effect on the gene expression covariance matrices, i.e. $\mathcal{H}_0 : Q^*(t) \equiv Q^*$ for some Q^* . The test we propose yields a p-value of 0.00019. When the analysis is performed separately for males and females, we obtain a p-value of 0.00034 for the male group and 0.00012 for the female group, suggesting a strong age effect on the gene expression covariance matrices.

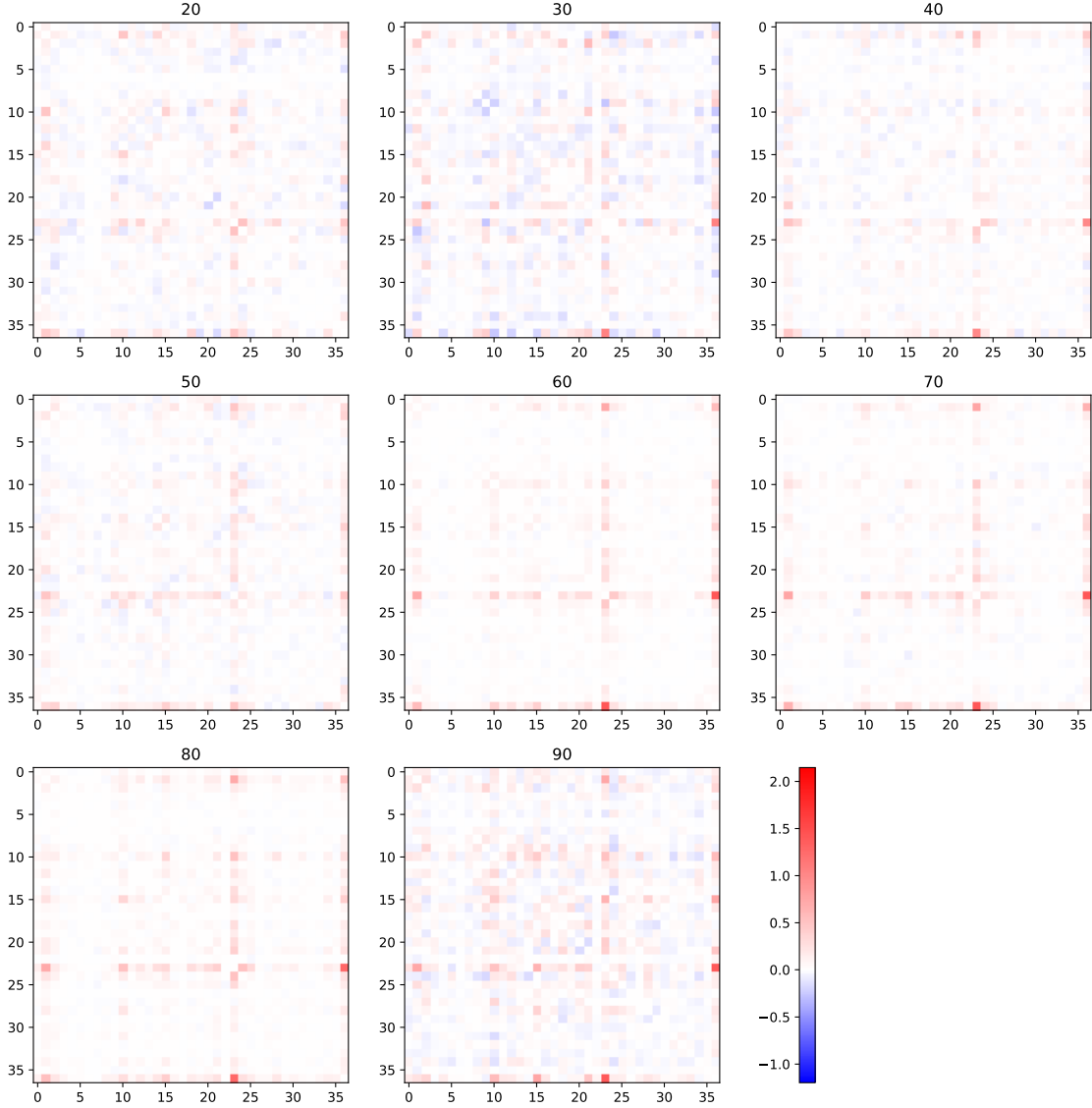


Figure 6: Heatmap of the gene expression covariance matrices with diagonal elements omitted for individual at age 20, 30, \dots , 90, respectively.

7. Discussion

The present paper provides a framework for statistical inference in the context of Fréchet regression on the Bures-Wasserstein manifold, where covariance matrices serve as outcomes. Our contributions include establishing uniform convergence rates for estimating the conditional Fréchet mean and deriving the asymptotic distribution of the estimator. Leveraging these results, we further develop a hypothesis test to rigorously assess associations between Euclidean covariates and covariance matrix-valued outcomes. The finite-sample performance of our methods is validated through numerical simulations, and their practical

utility is demonstrated in an application to gene co-expression analysis. Specifically, we test the association between age and co-expression patterns in the nutrient sensing pathway, revealing how these patterns evolve with age.

Beyond genomics, our framework has broad applicability in neuroimaging, where covariance matrices (or correlation matrices after standardization) of multiple brain regions are used to summarize as functional connectivity matrices. The proposed methods can be used to identify the factors that are associated with such functional connectivity matrices.

In this work, we focus on the theoretical analysis of the Fréchet regression estimator (15) under the assumption that outcome covariance matrices are directly observed. This setting aligns with prior frameworks in Petersen and Müller (2019) and Petersen et al. (2021). While our empirical results suggest the proposed hypothesis test remains valid when covariance matrices must first be estimated from data, formal theoretical guarantees for this setting remain an open problem. Additionally, relaxing the eigenvalue lower-bound assumption to accommodate singular covariance matrices would broaden the applicability of our framework, particularly in high-dimensional or sparse-data regimes.

Appendix A. Background on optimal transport and functional calculus

In this section, we collect relevant background about optimal transport and functional calculus to make the paper more self-contained.

A.1 Geometry of optimal transport

We begin with the geometry of optimal transport, and then specialize the general concepts to the Bures–Wasserstein manifold. For introductory expositions of optimal transport, we refer to Villani (2003); Santambrogio (2015); Panaretos and Zemel (2020). For a more comprehensive treatment, we refer to Ambrosio et al. (2005); Villani (2009).

Given $\mu_0, \mu_1 \in \mathcal{P}_{2,\text{ac}}(\mathbb{R}^d)$, the constant-speed geodesic $(\mu_t)_{t \in [0,1]}$ connecting μ_0 to μ_1 is characterized by

$$\mu_t = [\text{id} + t(T_{\mu_0}^{\mu_1} - \text{id})]_{\#} \mu_0, \quad t \in [0, 1]$$

Here, $\#$ denotes the pushforward operation defined by $T_{\#} \mu(E) = \mu(T^{-1}(E))$ for any Borel set $E \subset \mathbb{R}^d$. Then define the tangent vector of the geodesic $(\mu_t)_{t \in [0,1]}$ at $t = 0$ to be the mapping $T_{\mu_0}^{\mu_1} - \text{id}$. The tangent space $T_{\mu_0} \mathcal{P}_{2,\text{ac}}(\mathbb{R}^d)$ to $\mathcal{P}_{2,\text{ac}}(\mathbb{R}^d)$ at μ_0 is defined in Ambrosio et al., 2005, Thm 8.5.1 as

$$T_{\mu_0} \mathcal{P}_{2,\text{ac}}(\mathbb{R}^d) := \overline{\{\lambda(T_{\mu_0}^{\nu} - \text{id}) : \lambda > 0, \nu \in \mathcal{P}_{2,\text{ac}}(\mathbb{R}^d)\}}^{L^2(\mu_0)}$$

Here the overline denotes closure with respect to the $L^2(\mu_0)$ measure.

Given two covariance matrices $Q, S \in \mathcal{S}_d^{++}$, the constant-speed geodesic connecting the corresponding centered Gaussians is given by

$$(I_d + t(T_Q^S - I_d)) Q (I_d + t(T_Q^S - I_d)), \quad t \in [0, 1]$$

The tangent space $T_Q \mathcal{S}_d^{++}$ can be identified with the space \mathcal{S}_d of symmetric $d \times d$ matrices. For any $\tilde{S} \in T_Q \mathcal{S}_d^{++}$, its norm in the tangent space is given by

$$\|\tilde{S}\|_Q := \langle \tilde{S}, Q\tilde{S} \rangle^{1/2}$$

A.2 Functional calculus

To consider higher order differentials of T_Q^S on the Bures–Wasserstein manifold, we give a brief review of some key concepts in functional calculus that are essential for the development, and direct readers to Dudley and Norvaiša (2011) for further details.

Let Y and Z be normed spaces with the norm on each denoted by $\|\cdot\|$, and let U be an open subset of Y . Let $L(Y, Z)$ denote the space of all bounded linear operators from Y into Z . A function $\phi : U \rightarrow Z$ is called (Fréchet) differentiable at $u \in U$ if there exists an $L(u) \in L(Y, Z)$ such that for each $y \in Y$ with $u + Y \in U$,

$$\lim_{y \rightarrow 0} \frac{\|\phi(u + y) - \phi(u) - L(u)y\|}{\|y\|} = 0$$

The linear operator $L(u)$ is unique, is called the derivative of ϕ at u , and is denoted by $D\phi(u)$. If the function ϕ is differentiable at each $u \in U$, then we say that ϕ is differentiable on U .

To consider higher order derivatives of ϕ , let $L^1(Y, Z) := L(Y, Z)$ with the usual operator norm, and let $L^{k+1}(Y, Z) := L(Y, L^k(Y, Z))$ with the operator norm recursively for $k = 1, 2, \dots$. For $k \geq 2$, we say that ϕ is (Fréchet) differentiable of order k at u if ϕ has a $(k-1)$ st derivative $D^{k-1}\phi(y)$ at each point y of some neighborhood of u , and the mapping $D^{k-1}\phi$ is differentiable at u . Then $D^k\phi(u)$, the k th derivative of ϕ at u , is defined as the derivative of $D^{k-1}\phi$ at u . If ϕ is differentiable of order k at u for each $u \in U$ then we say that ϕ is differentiable of order k on U . A mapping ϕ is called a C^k function on $U \subset Y$ if the derivatives of ϕ through order k all exist on U and are continuous. A mapping ϕ is called a C^∞ function on $U \subset X$ if it is a C^k function for each k .

It would be convenient to introduce the notion of multilinear mappings in order study the properties of higher order derivatives $D^k\phi$. A function $A : Y^k \rightarrow Z$ is called k -linear if for each $j \in [k]$, $A(y_1, \dots, y_k)$ is linear in y_j for any fixed values of y_i , $i \neq j$. The function A is called bounded if

$$\|A\| := \sup \{\|A(y_1, \dots, y_k)\| : \|y_j\| \leq 1, j \in [k]\} < \infty \quad (29)$$

Let $\mathcal{M}_k(Y, Z)$ be the set of all bounded k -linear maps from Y^k to Z with norm define by (29).

The space $L^k(Y, Z)$ can be identified with $\mathcal{M}_k(Y, Z)$ through the natural isomorphism $\Phi_{(k)} : L^k(Y, Z) \rightarrow \mathcal{M}_k(Y, Z)$ defined by

$$\begin{aligned} \Phi_{(k)}(A)(y_1, \dots, y_k) &:= A(y_1)(y_2) \cdots (y_k) \\ &:= [\cdots [A(y_1)](y_2) \cdots](y_k) \end{aligned}$$

Then the k th differential at u is defined by $d^k\phi(u) := \Phi_{(k)}(D^k\phi(u))$.

The k th differential $d^k\phi(u)$ is symmetric in the sense that

$$d^k\phi(u)(y_{\sigma(1)}, \dots, y_{\sigma(k)}) = d^k\phi(u)(y_1, \dots, y_k)$$

for any permutation σ of $[k]$. For simplicity, we also denote $d^k\phi(u)(y, \dots, y)$ by $d^k\phi(u) \cdot y^{\otimes k}$.

Let $\mathcal{M}_{k,s}(Y, Z)$ denote the subspace of all symmetric elements of $\mathcal{M}_k(Y, Z)$. Then aside from the operator norm inherited from $\mathcal{M}_k(Y, Z)$, one can define another norm $\|\cdot\|$ on $\mathcal{M}_{k,s}(Y, Z)$ by

$$\|P\| := \sup \{\|P(y, y, \dots, y)\| : \|y\| \leq 1\}$$

Properties of the high order differentials $d^k T_Q^S$ are investigated in Appendix B.1.

Appendix B. Technical lemmas

In this section, we collect technical lemmas and relevant notations for the proof. First, properties of differentials of T_Q^S are collected in Appendix B.1. Then various concentration results are given in Appendix B.2. Finally, properties of $F(\cdot, \cdot)$ and $Q^*(\cdot)$ are summarized in Appendix B.3.

B.1 Differentials of optimal transport maps

Recall that the optimal transport map between two centered Gaussian distributions $\mathcal{N}(0, Q)$ and $\mathcal{N}(0, S)$ has the closed-form expression $T_Q^S = S^{1/2}(S^{1/2}QS^{1/2})^{-1/2}S^{1/2}$. In all relevant

analysis, we need to consider differentials of T_Q^S when viewed as a function of Q for fixed S , which we denote as $d^k T_Q^S$ for $k \geq 1$. It is shown in Kroshnin et al. (2021) that dT_Q^S can be defined as follows. For any $H \in \mathcal{S}_d$,

$$dT_Q^S(H) := -S^{1/2}U^\top \Lambda^{-1/2} \delta \Lambda^{-1/2} U S^{1/2}$$

where $U^\top \Lambda U$ is an eigenvalue decomposition of $S^{1/2} Q S^{1/2}$ with $U U^\top = U^\top U = I$ and $\delta = (\delta_{ij})_{i,j=1}^d$ with

$$\delta_{ij} = \begin{cases} \frac{\Delta_{ij}}{\sqrt{\lambda_i} + \sqrt{\lambda_j}} & i, j \leq \text{rank}(S) \\ 0 & \text{otherwise} \end{cases}, \quad \Delta = U S^{1/2} H S^{1/2} U^\top$$

To consider higher order differentials of T_Q^S , we start with the map $Q \mapsto Q^{1/2}$ in Lemma 19, next move on to $Q \mapsto Q^{-1/2}$ in Lemma 20 which then leads to T_Q^S in Lemma 21. Connections between T_Q^S and $W^2(Q, S)$ are also investigated in Lemma 21.

Lemma 19 *The square root functional $\phi : Q \in \mathcal{S}_d^{++} \rightarrow \phi(Q) = Q^{1/2} \in \mathcal{S}_d^{++}$ is Fréchet differentiable at any order on \mathcal{S}_d^{++} . Moreover, for any $Q \in \mathcal{S}_d^{++}$ and any $n \geq 0$, we have the estimates*

$$\| \| d^{n+1} \phi(Q) \| \| \leq C_{d,n} \lambda_{\min}(Q)^{-(n+1/2)}$$

Here $C_{d,n} = d^{n/2} \cdot n! \binom{2n}{n} \cdot 2^{-(2n+1)}$.

Proof See Del Moral and Niclas (2018) Theorem 1.1. ■

Lemma 20 *The inverse square root functional $\varphi : Q \in \mathcal{S}_d^{++} \mapsto \varphi(Q) = Q^{-1/2} \in \mathcal{S}_d^{++}$ is Fréchet differentiable at any order on \mathcal{S}_d^{++} . Moreover, for any $A \in \mathcal{S}_d^{++}, H \in \mathbb{S}^d$, the following holds.*

$$\begin{aligned} d\varphi(Q) \cdot H &= -Q^{-1/2} (d\phi(Q) \cdot H) Q^{-1/2} \\ d^2 \varphi(Q) \cdot H^{\otimes 2} &= -Q^{-1/2} (d^2 \phi(Q) \cdot H^{\otimes 2}) Q^{-1/2} + 2Q^{-1/2} (d\phi(Q) \cdot H) Q^{-1/2} (d\phi(Q) \cdot H) Q^{-1/2} \\ d^3 \varphi(Q) \cdot H^{\otimes 3} &= -Q^{-1/2} (d^3 \phi(Q) \cdot H^{\otimes 3}) Q^{-1/2} + 3Q^{-1/2} (d^2 \phi(Q) \cdot H^{\otimes 2}) Q^{-1/2} (d\phi(Q) \cdot H) Q^{-1/2} \\ &\quad + 3Q^{-1/2} (d\phi(Q) \cdot H) Q^{-1/2} (d^2 \phi(Q) \cdot H^{\otimes 2}) Q^{-1/2} \\ &\quad - 6Q^{-1/2} (d\phi(Q) \cdot H) Q^{-1/2} (d\phi(Q) \cdot H) Q^{-1/2} (d\phi(Q) \cdot H) Q^{-1/2} \end{aligned}$$

with

$$\begin{aligned} \| d\varphi(Q) \| &\leq C_{d,0} \lambda_{\min}(Q)^{-3/2} \\ \| d^2 \varphi(Q) \| &\leq (C_{d,1} + 2C_{d,0}^2) \cdot (\lambda_{\min}(Q))^{-5/2} \\ \| d^3 \varphi(Q) \| &\leq (C_{d,2} + 6C_{d,1}C_{d,0} + 6C_{d,0}^3) \lambda_{\min}(Q)^{-7/2} \end{aligned}$$

where $C_{d,n}$ is defined in Lemma 19.

Proof By Lemma 19 we have for infinitesimal $H \in \mathcal{S}_d$ that

$$(Q + H)^{1/2} = Q^{1/2} + \underbrace{d\phi(Q) \cdot H}_{Z_1} + \underbrace{\frac{1}{2}d^2\phi(Q) \cdot H^{\otimes 2}}_{Z_2} + \underbrace{\frac{1}{6}d^3\phi(Q) \cdot H^{\otimes 3}}_{Z_3} + \dots$$

Let $Z = Z_1 + Z_2 + Z_3 + Z_4$, and we obtain for infinitesimal $H \in \mathcal{S}_d$ that

$$\begin{aligned} (Q + H)^{-1/2} &= \left(Q^{1/4}(I_d + Q^{-1/4}ZQ^{-1/4})Q^{1/4} \right)^{-1} \\ &= Q^{-1/4}(I_d + Q^{-1/4}ZQ^{-1/4})^{-1}Q^{-1/4} \\ &\stackrel{(i)}{=} Q^{-1/4} \left(I_d - Q^{-1/4}ZQ^{-1/4} + Q^{-1/4}ZQ^{-1/2}ZQ^{-1/4} - Q^{-1/4}ZQ^{-1/2}ZQ^{-1/2}ZQ^{-1/4} + \dots \right) Q^{-1/4} \\ &= Q^{-1/2} - Q^{-1/2}ZQ^{-1/2} + Q^{-1/2}ZQ^{-1/2}ZQ^{-1/2} - Q^{-1/2}ZQ^{-1/2}ZQ^{-1/2}ZQ^{-1/2} + \dots \\ &\stackrel{(ii)}{=} Q^{-1/2} - \underbrace{Q^{-1/2}Z_1Q^{-1/2}}_{\text{1st order in } H} - \underbrace{Q^{-1/2}Z_2Q^{-1/2} + Q^{-1/2}Z_1Q^{-1/2}Z_1Q^{-1/2}}_{\text{2nd order in } H} \\ &\quad - \underbrace{Q^{-1/2}Z_3Q^{-1/2} + Q^{-1/2}Z_2Q^{-1/2}Z_1Q^{-1/2} + Q^{-1/2}Z_1Q^{-1/2}Z_2Q^{-1/2} - Q^{-1/2}Z_1Q^{-1/2}Z_1Q^{-1/2}Z_1Q^{-1/2}}_{\text{3rd order in } H} \\ &\quad + \text{higher order terms in } H \end{aligned}$$

Here (i) follows from the von Neumann series expansion, and (ii) is obtained by arranging terms according to their orders in H . Then (20) follows, and we have the following estimate

$$\begin{aligned} \|d^2\varphi(Q) \cdot H^{\otimes 2}\|_F &\leq \left\| Q^{-1/2} (d^2\phi(Q) \cdot H^{\otimes 2}) Q^{-1/2} \right\|_F + 2 \left\| Q^{-1/2} (d\phi(Q) \cdot H) Q^{-1/2} (d\phi(Q) \cdot H) Q^{-1/2} \right\|_F \\ &\leq \lambda_{\min}(Q)^{-1} \cdot \|d^2\phi(Q)\| \cdot \|H\|_F^2 + 2\lambda_{\min}(Q)^{-3/2} \|d\phi(Q)\|^2 \cdot \|H\|_F^2 \\ &\leq (C_{d,1} + 2C_{d,0}^2)\lambda_{\min}(Q)^{-5/2} \cdot \|H\|_F^2 \end{aligned}$$

and similarly

$$\begin{aligned} \|d^3\varphi(Q) \cdot H^{\otimes 3}\|_F &\leq \lambda_{\min}(Q)^{-1} \|d^3\varphi(Q)\| \cdot \|H\|_F^3 + 6\lambda_{\min}(Q)^{-3/2} \|d^2\varphi(Q)\| \cdot \|d\varphi(Q)\| \cdot \|H\|_F^3 \\ &\quad + 6\lambda_{\min}(Q)^{-2} \|d\varphi(Q)\|^3 \cdot \|H\|_F^3 \\ &\leq (C_{d,2} + 6C_{d,1}C_{d,0} + 6C_{d,0}^3)\lambda_{\min}(Q)^{-7/2} \cdot \|H\|_F^3 \end{aligned}$$

Here the last inequality follows by applying Lemma 19. And we arrive at the desired result. \blacksquare

Lemma 21 *The following properties hold for the 2-Wasserstein distance $W(Q, S)$ and the optimal transport map T_Q^S . For any $Q, Q_1, Q_2 \in \mathcal{S}_d^{++}$, $S \in \mathcal{S}_d^+$ and $X, Y \in \mathcal{S}_d$,*

(i) $W^2(Q, S)$ is upper bounded by

$$W^2(Q, S) \leq 2d(\lambda_{\max}(Q) + \lambda_{\max}(S))$$

(ii) $W^2(Q, S)$ is twice differentiable with

$$\begin{aligned} d_Q W^2(Q, S)(X) &= \langle I - T_Q^S, X \rangle \\ d_Q^2 W^2(Q, S)(X, Y) &= -\langle X, dT_Q^S(Y) \rangle \end{aligned}$$

Moreover, the following quadratic approximation holds:

$$\begin{aligned} & \frac{2}{\left(1 + \lambda_{\max}^{1/2}(Q')\right)^2} \langle -dT_{Q_0}^S(Q_1 - Q_0), Q_1 - Q_0 \rangle \\ & \leq W^2(Q_1, S) - W^2(Q_0, S) + \langle T_{Q_0}^S - I, Q_1 - Q_0 \rangle \\ & \leq \frac{2}{\left(1 + \lambda_{\min}^{1/2}(Q')\right)^2} \langle -dT_{Q_0}^S(Q_1 - Q_0), Q_1 - Q_0 \rangle \end{aligned}$$

with $Q' := Q_0^{-1/2} Q_1 Q_0^{-1/2}$.

(iii) dT_Q^S is self-adjoint, negative semi-definite and enjoys the following two-sided bound.

$$\frac{\lambda_{\min}^{1/2}(S^{1/2} Q S^{1/2})}{2} \|Q^{-1/2} X Q^{-1/2}\|_F^2 \leq \langle -dT_Q^S(X), X \rangle \leq \frac{\lambda_{\max}^{1/2}(S^{1/2} Q S^{1/2})}{2} \|Q^{-1/2} X Q^{-1/2}\|_F^2$$

(iv) $W^2(Q_0, Q_1)$ can be upper and lower bounded by the Frobenius norm as follows

$$\frac{1}{2} \frac{\lambda_{\max}(Q_0) \lambda_{\min}^{-2}(Q_0)}{1 + \lambda_{\min}^{-1}(Q_0) \lambda_{\max}(Q_1)} \|Q_1 - Q_0\|_F^2 \leq W^2(Q_0, Q_1) \leq \frac{\lambda_{\max}(Q_0) \lambda_{\min}^{-2}(Q_0)}{1 + \lambda_{\max}^{-1}(Q_0) \lambda_{\min}(Q_1)} \|Q_1 - Q_0\|_F^2$$

(v) $\|d^k T_Q^S\|$ can be upper bounded by

$$\begin{aligned} \|dT_Q^S\| &\leq \frac{\lambda_{\min}(Q)^{-2}}{2} \cdot \left(\lambda_{\max}(S^{1/2} Q S^{1/2})\right)^{1/2} \\ \|d^2 T_Q^S\| &\leq \lambda_{\max}(S)^3 (C_{d,1} + 2C_{d,0}^2) \cdot \left(\lambda_{\min}(S^{1/2} Q S^{1/2})\right)^{-5/2} \\ \|d^3 T_Q^S\| &\leq \lambda_{\max}(S)^4 (C_{d,2} + 6C_{d,1}C_{d,0} + 6C_{d,0}^3) \left(\lambda_{\min}(S^{1/2} Q S^{1/2})\right)^{-7/2} \end{aligned}$$

Moreover, if $S, Q \in \mathcal{S}_d(M^{-1}, M)$, then we have

$$\begin{aligned} \|dT_Q^S\| &\leq \frac{1}{2} M^3 \\ \|d^2 T_Q^S\| &\leq (C_{d,1} + 2C_{d,0}^2) M^8 \\ \|d^3 T_Q^S\| &\leq (C_{d,2} + 6C_{d,1}C_{d,0} + 6C_{d,0}^3) M^{11} \end{aligned}$$

Proof

Proof of (i): By the closed form expression for $W^2(Q, S)$, one has

$$\begin{aligned} W^2(Q, S) &= \text{tr} \left[Q + S - 2(S^{1/2}QS^{1/2})^{1/2} \right] \\ &\leq \text{tr} \left[Q + S + 2(S^{1/2}QS^{1/2})^{1/2} \right] \\ &\leq d \left(\lambda_{\max}(Q) + \lambda_{\max}(S) + 2\lambda_{\max}(Q)^{1/2}\lambda_{\max}(S)^{1/2} \right) \\ &\leq 2d \left(\lambda_{\max}(Q) + \lambda_{\max}(S) \right) \end{aligned}$$

Proof of (ii), (iii): see Kroshnin et al. (2021, Lemma A.2, A.3, A.4, A.6).

Proof of (iv): Set $S = Q_0$ in ((ii)), one can obtain

$$\begin{aligned} \frac{2}{\left(1 + \lambda_{\max}^{1/2}(Q')\right)^2} \left\langle -dT_{Q_0}^{Q_0}(Q_1 - Q_0), Q_1 - Q_0 \right\rangle &\leq W^2(Q_1, Q_0) \\ &\leq \frac{2}{\left(1 + \lambda_{\min}^{1/2}(Q')\right)^2} \left\langle -dT_{Q_0}^{Q_0}(Q_1 - Q_0), Q_1 - Q_0 \right\rangle \end{aligned} \quad (30)$$

where $Q' = Q_0^{-1/2}Q_1Q_0^{-1/2}$. Next, apply (iii) to get that for any $X \in \mathcal{S}_d$, one has

$$\frac{\lambda_{\min}(Q_0)}{2} \lambda_{\max}^{-2}(Q_0) \|X\|_F^2 \leq \left\langle -dT_{Q_0}^{Q_0}(X), X \right\rangle \leq \frac{\lambda_{\max}(Q_0)}{2} \lambda_{\min}^{-2}(Q_0) \|X\|_F^2 \quad (31)$$

Combine (30) and (31) to get that

$$\begin{aligned} W^2(Q_1, Q_0) &\leq \frac{\lambda_{\max}(Q_0) \lambda_{\min}^{-2}(Q_0)}{\left(1 + \lambda_{\min}^{1/2}(Q')\right)^2} \cdot \|Q_1 - Q_0\|_F^2 \\ &\leq \frac{\lambda_{\max}(Q_0) \lambda_{\min}^{-2}(Q_0)}{1 + \lambda_{\min}(Q')} \cdot \|Q_1 - Q_0\|_F^2 \\ &\leq \frac{\lambda_{\max}(Q_0) \lambda_{\min}^{-2}(Q_0)}{1 + \lambda_{\max}^{-1}(Q_0) \lambda_{\min}(Q_1)} \cdot \|Q_1 - Q_0\|_F^2 \end{aligned}$$

and similarly

$$\begin{aligned} W^2(Q_1, Q_0) &\geq \frac{\lambda_{\min}(Q_0) \lambda_{\max}^{-2}(Q_0)}{\left(1 + \lambda_{\max}^{1/2}(Q')\right)^2} \cdot \|Q_1 - Q_0\|_F^2 \\ &\geq \frac{1}{2} \frac{\lambda_{\max}(Q_0) \lambda_{\min}^{-2}(Q_0)}{1 + \lambda_{\max}(Q')} \cdot \|Q_1 - Q_0\|_F^2 \\ &\geq \frac{1}{2} \frac{\lambda_{\max}(Q_0) \lambda_{\min}^{-2}(Q_0)}{1 + \lambda_{\min}^{-1}(Q_0) \lambda_{\max}(Q_1)} \cdot \|Q_1 - Q_0\|_F^2 \end{aligned}$$

Proof of (v): First, note that results for dT_Q^S follows directly from (iii).

Next since $T_Q^S = S^{1/2}(S^{1/2}QS^{1/2})^{-1/2}S^{1/2}$, applying Lemma 20 gives that for H small enough, one has

$$\begin{aligned} d^2 T_Q^S \cdot H^{\otimes 2} &= S^{1/2} \left[d^2 \varphi(S^{1/2}QS^{1/2}) \cdot (S^{1/2}HS^{1/2})^{\otimes 2} \right] S^{1/2} \\ d^3 T_Q^S \cdot H^{\otimes 3} &= S^{1/2} \left[d^3 \varphi(S^{1/2}QS^{1/2}) \cdot (S^{1/2}HS^{1/2})^{\otimes 3} \right] S^{1/2} \end{aligned}$$

which implies

$$\begin{aligned} \|d^2 T_Q^S\| &\leq \lambda_{\max}(S)^3 \cdot \|d^2 \varphi(S^{1/2}QS^{1/2})\| \\ &\leq \lambda_{\max}(S)^3 (C_{d,1} + 2C_{d,0}^2) \cdot \left(\lambda_{\min}(S^{1/2}QS^{1/2}) \right)^{-5/2} \\ \|d^3 T_Q^S\| &\leq \lambda_{\max}(S)^4 \|d^3 \varphi(S^{1/2}QS^{1/2})\| \\ &\leq \lambda_{\max}(S)^4 (C_{d,2} + 6C_{d,1}C_{d,0} + 6C_{d,0}^3) \lambda_{\min}(S^{1/2}QS^{1/2})^{-7/2} \end{aligned}$$

■

Lemma 22 *Let Y and Z be normed spaces with the norm on each denoted by $\|\cdot\|$. For any $y \in Y$, $P \in \mathcal{M}_{k,s}(Y, Z)$, $k \geq 2$, let Py denote a $(k-1)$ -linear function defined by*

$$Py(y_1, \dots, y_{k-1}) = P(y, y_1, \dots, y_{k-1})$$

Then $Py \in \mathcal{M}_{k-1,s}(Y, Z)$ and

$$\|Py\| \leq \|P\| \|y\| \leq \frac{k^k}{k!} \|P\| \|y\| \quad (32)$$

Proof $Py \in \mathcal{M}_{k-1,s}(Y, Z)$ can be proved by definition. To prove (32), note that for any $\|\tilde{y}\| \leq 1$, one has

$$\begin{aligned} \|Py(\tilde{y}, \dots, \tilde{y})\| &\leq \|P\| \cdot \|y\| \cdot \|\tilde{y}\|^{k-1} \\ &\leq \|P\| \cdot \|y\| \\ &\stackrel{(i)}{\leq} \frac{k^k}{k!} \|P\| \cdot \|y\| \end{aligned}$$

Here (i) follows from Dudley and Norvaiša (2011, Theorem 5.7).

■

B.2 Concentration inequalities and uniform convergence

First, let us introduce some additional notation. Given a random variable X we denote $\|X\|_{\psi_\alpha}$ for $\alpha > 0$ as follows.

$$\|X\|_{\psi_\alpha} := \inf \{ \eta > 0 : \mathbb{E} \psi_\alpha(|X/\eta|) \leq 1 \}, \quad \text{where } \psi_\alpha(x) := \exp(x^\alpha) - 1 \quad \text{for } x \geq 0 \quad (33)$$

Note that by the definition and Markov's inequality, if $\|X\|_{\psi_\alpha} < \infty$, then

$$\mathbb{P}(|X| \geq t) \leq 2 \exp \left(- \frac{t^\alpha}{\|X\|_{\psi_\alpha}^\alpha} \right) \quad (34)$$

For $\alpha \geq 1$, $\|\cdot\|_{\psi_\alpha}$ is a norm (Vershynin, 2018), and for $\alpha < 1$, it is equivalent to a norm (Talagrand, 1989).

Lemma 23 *Let X be a random variable and $\alpha \geq 1$. Then the following properties are equivalent; the parameters $K_i > 0$ appearing in these properties differ from each other by at most an absolute constant factor depending on α .*

1. *The ψ_α -norm of X satisfies $\|X\|_{\psi_\alpha} \leq K_1$.*
2. *The tails of X satisfy*

$$\mathbb{P}\{|X| \geq t\} \leq 2 \exp(-t^\alpha/K_2^\alpha) \quad \forall t \geq 0.$$

3. *The moments of X satisfy*

$$\|X\|_p \leq K_3 p^{1/\alpha}$$

Proof See Vershynin (2018, Exercise 2.7.3). ■

For a random vector $X \in \mathbb{R}^p$, denote

$$\|X\|_{\psi_\alpha} := \sup_{u \in \mathbb{R}^p, \|u\|=1} \|u^\top X\|_{\psi_\alpha}$$

Lemma 24 *Let $C_1, c_1, \alpha, \beta, B > 0$. Then*

$$\min \left\{ 1, B^\beta \cdot C_1 \exp(-c_1 t^\alpha) \right\} \leq C_1 \exp \left(-c_1 \frac{t^\alpha}{1 + \frac{\beta \log(1 \vee B)}{\log(1 \vee C_1)}} \right)$$

Proof Without loss of generality, assume $C_1, B \geq 1$. Denote

$$\begin{aligned} g(t) &:= C_1 \exp(-c_1 t^\alpha) \\ f(t) &:= C_1 \exp \left(-c_1 \frac{t^\alpha}{1 + \frac{\beta \log B}{\log C_1}} \right) \end{aligned}$$

$f, g : (0, +\infty) \rightarrow \mathbb{R}$ satisfy the following properties

- g, f are both positive and decreasing for $t \geq 0$.
- denote $t_0 = \left[\frac{\beta \log B + \log C_1}{c_1} \right]^{1/\alpha}$, then

$$1 \wedge g(t_0) = 1 \leq f(t_0)$$

- derivatives satisfy

$$\frac{d}{dt} [\log(g(t))] \leq \frac{d}{dt} [\log(f(t))], \quad \forall t \geq t_0$$

Therefore,

$$1 \wedge g(t) \leq f(t), \quad \forall t > 0$$

■

Lemma 25 *Let $\alpha \geq 1$. Let X, Q be random elements in $\mathbb{R}^d, \mathbb{R}^{d \times d}$ and let Φ_k be a random symmetric operator in $\mathcal{L}((\mathbb{R}^{d \times d})^{\times k}; \mathbb{R}^{d \times d})$ for some $k \in \mathbb{N}_+$. Then*

- I. $\|X\|_{\psi_\alpha} \leq \| \|X\|_2 \|_{\psi_\alpha} \leq c_0(2d+1)^{1/\alpha} \|X\|_{\psi_\alpha}$ for any $\alpha \geq 1$ where $c_0 > 0$ is some absolute constant independent of d and α .
- II. If $Q \in \mathbb{R}^{d \times d}$ is symmetric, then for any $\|v\| = 1$,

$$\left\| v^\top Q v \right\|_{\psi_\alpha} \leq \left\| \|Q\|_{\text{op}} \right\|_{\psi_\alpha} \leq c_1 \left(2d \cdot \frac{\log 3}{\log 2} + 1 \right)^{1/\alpha} \sup_{\|v\|=1} \left\| v^\top Q v \right\|_{\psi_\alpha}$$

where $c_1 > 0$ is some absolute constant independent of d and α .

- III. For any $U \in \mathbb{R}^{d \times d}$ with unit Frobenius norm

$$\|\langle U, Q \rangle\|_{\psi_\alpha} \leq \| \|Q\|_{\text{F}} \|_{\psi_\alpha} \lesssim \sup_{U: \|U\|_{\text{F}}=1} \|\langle U, Q \rangle\|_{\psi_\alpha}$$

- IV. For any $U \in \mathbb{R}^{d \times d}$ with unit Frobenius norm

$$\left\| \left\| \Phi_k \cdot U^{\otimes k} \right\|_{\text{F}} \right\|_{\psi_\alpha} \leq \| \Phi_k \|_{\psi_\alpha} \lesssim \sup_{U: \|U\|_{\text{F}}=1} \left\| \left\| \Phi_k \cdot U^{\otimes k} \right\|_{\text{F}} \right\|_{\psi_\alpha}$$

The constants behind \lesssim only depend on dimension d , k and possible on α .

Proof

I: The first inequality $\|X\|_{\psi_\alpha} \leq \| \|X\|_2 \|_{\psi_\alpha}$ follows since $\langle v, X \rangle \leq \|X\|$ for any $\|v\| = 1$.

For the second inequality, by Lemma 23, we can assume without loss of generality that $\|X\|_{\psi_\alpha}$ satisfies $K_2 = 1$, i.e.

$$\sup_{\|v\|=1} \mathbb{P} \{ |\langle v, X \rangle| \geq t \} \leq 2 \exp(-t^\alpha)$$

Following the 1/2-net argument in Lemma 1 in Jin et al. (2019), one can obtain

$$\begin{aligned}\mathbb{P}\{\|X\| \geq t\} &\leq 4^d \sup_{\|v\|=1} \mathbb{P}(\langle v, X \rangle \geq t/2) \\ &\leq \underbrace{2 \cdot 4^d \exp(-t^\alpha/2^\alpha)}_{=:g(t)}\end{aligned}$$

Lemma 24 implies that for any $t \geq 0$,

$$\mathbb{P}\{\|X\| \geq t\} \leq 2 \exp\left(-\frac{t^\alpha}{(2d+1) \cdot 2^\alpha}\right) =: f(t)$$

Therefore, Lemma 23 implies that

$$\| \|X\|_2 \|_{\psi_\alpha} \leq c_0 (2d+1)^{1/\alpha} \|X\|_{\psi_\alpha}$$

where $c_0 > 0$ is some absolute constant independent of d and α .

II: We prove the second inequality here. By Lemma 23, we can assume without loss of generality that $\sup_{\|v\|=1} \|v^\top Qv\|_{\psi_\alpha}$ satisfies $K_2 = 1$, i.e.

$$\sup_{\|v\|=1} \mathbb{P}\left\{|v^\top Qv| \geq t\right\} \leq 2 \exp(-t^\alpha)$$

Following the 1/4-net argument in Vershynin (2018, Exercise 4.4.3, Theorem 4.4.5.), one can obtain that for any $t \geq 0$,

$$\mathbb{R}\left\{\| \|Q\|_{\text{op}} \|_{\psi_\alpha} \geq t\right\} \leq 9^d \cdot 2 \exp(-t^\alpha/2^\alpha)$$

Lemma 24 then implies that

$$\mathbb{R}\left\{\| \|Q\|_{\text{op}} \|_{\psi_\alpha} \geq t\right\} \leq 2 \exp\left(-\frac{t^\alpha}{(2d \cdot \frac{\log 3}{\log 2} + 1) \cdot 2^\alpha}\right), \quad \forall t \geq 0$$

Hence Lemma 23 implies that

$$\| \|Q\|_{\text{op}} \|_{\psi_\alpha} \leq c_1 \left(2d \cdot \frac{\log 3}{\log 2} + 1\right)^{1/\alpha} \sup_{\|v\|=1} \|v^\top Qv\|_{\psi_\alpha}$$

where $c_1 > 0$ is some absolute constant independent of d and α .

The proofs for statements **III** and **IV** follow a similar approach. ■

Next, we give some properties of the weights $w(x, X)$ and $w_{n,\rho}(x, X_i)$ that will be crucial to prove uniform concentration as well as a central limit theorem for the estimate $\widehat{Q}_\rho(x)$. Recall that by definition, $w(x, X) = 1 + (x - \mu)^\top \Sigma^{-1}(X - \mu)$ and $w_{n,\rho}(x) = 1 + (x - \bar{X})^\top \widehat{\Sigma}_\rho^{-1}(X_i - \bar{X})$. For any vector $z \in \mathbb{R}^p$, denote $\vec{z} = (1, z^\top)^\top$ and

$$\vec{\Sigma} := \mathbb{E} \vec{X} \vec{X}^\top = \begin{pmatrix} 1 & \mu^\top \\ \mu & \Sigma + \mu \mu^\top \end{pmatrix}$$

$$\begin{aligned}\vec{\Sigma}_\rho &:= \begin{pmatrix} 1 & \mu^\top \\ \mu & \Sigma_\rho + \mu\mu^\top \end{pmatrix}, \quad \text{where } \Sigma_\rho = \Sigma + \rho I_p \\ \hat{\vec{\Sigma}} &:= n^{-1} \sum_{i=1}^n \vec{X}_i \vec{X}_i^\top = \begin{pmatrix} 1 & \hat{\mu}^\top \\ \hat{\mu} & \hat{\Sigma} + \hat{\mu}\hat{\mu}^\top \end{pmatrix} \\ \hat{\vec{\Sigma}}_\rho &:= \begin{pmatrix} 1 & \hat{\mu}^\top \\ \hat{\mu} & \hat{\Sigma}_\rho + \hat{\mu}\hat{\mu}^\top \end{pmatrix}\end{aligned}$$

Lemma 26 Suppose $X, X_1, \dots, X_n \stackrel{\text{i.i.d.}}{\sim} \mathbb{P} \in \mathcal{P}_2(\mathbb{R}^p)$. Let $\hat{\vec{\Sigma}} = n^{-1} \sum_{i=1}^n \vec{X}_i \vec{X}_i^\top$. Then for any $x \in \mathbb{R}^p$,

$$w(x, X) = \vec{x}^\top \vec{\Sigma}^{-1} \vec{X} \quad (35)$$

$$w_{n,\rho}(x, X_i) = \vec{x}^\top \hat{\vec{\Sigma}}_\rho^{-1} \vec{X} \quad (36)$$

Proof For (35), by definition one has

$$\vec{\Sigma} = \begin{pmatrix} 1 & \mu^\top \\ \mu & \Sigma + \mu\mu^\top \end{pmatrix}$$

Computing the inverse of the block matrix $\vec{\Sigma}$ then gives

$$\vec{\Sigma}^{-1} = \begin{pmatrix} 1 + \mu^\top \Sigma^{-1} \mu & -\mu^\top \Sigma^{-1} \\ -\Sigma^{-1} \mu & \Sigma^{-1} \end{pmatrix}$$

Finally we arrive at (35) by computing $\vec{x}^\top \Lambda^{-1} \vec{X}$. (36) follows from similar arguments. \blacksquare

Lemma 27 Suppose Assumption 1 holds. Set

$$\begin{aligned}W_{0n} &:= W_{0n}(x) = -\bar{X}^\top \hat{\Sigma}_\rho^{-1}(x - \bar{X}) + \mu^\top \hat{\Sigma}_\rho^{-1}(x - \mu) \\ W_{1n} &:= W_{1n}(x) = \Sigma^{-1}(x - \mu) - \hat{\Sigma}_\rho^{-1}(x - \bar{X})\end{aligned}$$

one has

$$w_{n,\rho}(x, X_i) - w(x, X_i) = W_{0n} + W_{1n}^\top X_i \quad (37)$$

Moreover, for any $\rho \in \{0, n^{-1}\}$, $L \geq 1$ and $\tau \geq 0$, there exists a constant $C > 0$ independent of n such that

$$\mathbb{P} \left\{ \sup_{x \in B_\mu(L)} |W_{0n}(x)| > \frac{L}{\sqrt{n}} C(1 + \tau) \log n \right\} \lesssim n^{-(1+\tau)} \quad (38)$$

$$\mathbb{P} \left\{ \sup_{x \in B_\mu(L)} \|W_{1n}(x)\| > \frac{L}{\sqrt{n}} C(1 + \tau) \log n \right\} \lesssim n^{-(1+\tau)} \quad (39)$$

Proof Under Assumption 1, we can assume without loss of generality that $\mu = 0$ and $\Sigma = I_p$. Then

$$\begin{aligned} w_{n,\rho}(x, X_i) &= 1 + (x - \bar{X})^\top \hat{\Sigma}_\rho^{-1} (X_i - \bar{X}) \\ w(x, X) &= 1 + x^\top X_i \end{aligned}$$

Proof of (37): Direct computation gives

$$\begin{aligned} w_{n,\rho}(x, X_i) - w(x, X_i) &= (x - \bar{X})^\top \hat{\Sigma}_\rho^{-1} (X_i - \bar{X}) - x^\top X_i \\ &= \underbrace{-x^\top \hat{\Sigma}_\rho^{-1} \bar{X} + \bar{X}^\top \hat{\Sigma}_\rho^{-1} \bar{X}}_{=W_{0n}(x)} + \underbrace{\left((\hat{\Sigma}_\rho^{-1} - I_p)x + \hat{\Sigma}_\rho^{-1} \bar{X} \right)^\top}_{=W_{1n}(x)} X_i \end{aligned}$$

Proof of (38), (39): For $W_{0n}(x)$:

$$\begin{aligned} \mathbb{P} \left\{ \sup_{x \in B_\mu(L)} |W_{0n}(x)| > t \right\} &\leq \mathbb{P} \left(\left\{ \left\| \hat{\Sigma}_\rho^{-1} \right\|_{\text{op}} \leq 2 \right\} \cap \left\{ L \|\bar{X}\| + \|\bar{X}\|^2 \geq t/2 \right\} \right) \\ &\quad + \mathbb{P} \left(\left\{ 2 < \left\| \hat{\Sigma}_\rho^{-1} \right\|_{\text{op}} \leq n \right\} \cap \left\{ L \|\bar{X}\| + \|\bar{X}\|^2 \geq t/n \right\} \right) \\ &\quad + \mathbb{P} \left\{ \left\| \hat{\Sigma}_\rho^{-1} \right\|_{\text{op}} > n \right\} \\ &\leq \underbrace{\mathbb{P} \left(\left\{ L \|\bar{X}\| + \|\bar{X}\|^2 \geq t/2 \right\} \right)}_{\text{(I)}} \\ &\quad + \underbrace{\mathbb{P} \left(\left\{ 2 < \left\| \hat{\Sigma}_\rho^{-1} \right\| \leq n \right\} \right) \wedge \mathbb{P} \left(\left\{ L \|\bar{X}\| + \|\bar{X}\|^2 \geq t/n \right\} \right)}_{\text{(II)}} \\ &\quad + \underbrace{\mathbb{P} \left\{ \left\| \hat{\Sigma}_\rho^{-1} \right\| > n \right\}}_{\text{(III)}} \end{aligned}$$

Note that for any $s \geq 0$, one has

$$\begin{aligned} \mathbb{P} \left\{ L \|\bar{X}\| + \|\bar{X}\|^2 \geq s \right\} &\leq \mathbb{P} \left\{ L \|\bar{X}\| \geq \frac{\sqrt{n}L}{\sqrt{n}L+1} s \right\} + \mathbb{P} \left\{ \|\bar{X}\|^2 \geq \frac{1}{\sqrt{n}L+1} s \right\} \\ &= \mathbb{P} \left\{ \|\bar{X}\| \geq \frac{\sqrt{n}}{\sqrt{n}L+1} s \right\} + \mathbb{P} \left\{ \|\bar{X}\|^2 \geq \frac{1}{\sqrt{n}L+1} s \right\} \\ &\lesssim \exp(-cs^2/L^2) + \exp(-c\sqrt{n}s/L) \end{aligned}$$

- If $\rho = 0$, then

$$\begin{aligned} \text{(II)} + \text{(III)} &\leq \mathbb{P} \left(\left\| \hat{\Sigma}_\rho^{-1} \right\|_{\text{op}} > 2 \right) \\ &\leq \mathbb{P} \left(\left\| \hat{\Sigma}_\rho - I_p \right\|_{\text{op}} > 1/2 \right) \end{aligned}$$

$$\lesssim \exp(-cn)$$

Hence

$$\begin{aligned} \mathbb{P} \left\{ \sup_{x \in B_\mu(L)} |W_{0n}(x)| > t \right\} &\lesssim \exp(-cnt^2/L^2) + \exp(-c\sqrt{nt}/L) + \exp(-cn) \\ &\lesssim \begin{cases} \exp(-c\sqrt{nt}/L) + \exp(-cn), & t \geq L/\sqrt{n} \\ \exp(-cnt^2/L^2) + \exp(-cn), & t < L/\sqrt{n} \end{cases} \end{aligned}$$

or equivalently for any $s > 0$,

$$\mathbb{P} \left\{ \sup_{x \in B_\mu(L)} |W_{0n}(x)| > \frac{L}{\sqrt{n}} s \right\} \lesssim \begin{cases} \exp(-cs^2) & s < 1 \\ \exp(-cs), & 1 \leq s \leq n \\ \exp(-cn), & s > n \end{cases}$$

- If $\rho = 1/n$, then **III** = 0. As a result, for any $s \geq 1$,

$$\begin{aligned} &\mathbb{P} \left\{ \sup_{x \in B_\mu(L)} |W_{0n}(x)| > s \frac{L}{\sqrt{n}} \right\} \\ &\lesssim \exp(-cs) + \exp(-cn) \wedge \begin{cases} \exp(-cs^2/n^2), & 1 \leq s \leq n \\ \exp(-cs/n), & s > n \end{cases} \\ &= \exp(-cs) + \begin{cases} \exp(-cn), & 1 \leq s \leq n^2 \\ \exp(-cs/n), & s > n^2 \end{cases} \\ &\lesssim \begin{cases} \exp(-cs), & 1 \leq s \leq n \\ \exp(-cn), & n < s \leq n^2 \\ \exp(-cs/n), & s > n^2 \end{cases} \end{aligned}$$

For $W_{1n}(x)$: note that $\widehat{\Sigma}_\rho^{-1} - I_p = -\widehat{\Sigma}_\rho^{-1} (\widehat{\Sigma}_\rho - I_p)$, then one has

$$\begin{aligned} &\mathbb{P} \left\{ \sup_{x \in B_\mu(L)} \|W_{1n}(x)\| > t \right\} \\ &\leq \mathbb{P} \left\{ \left\| \widehat{\Sigma}_\rho^{-1} \right\|_{\text{op}} \left\| \widehat{\Sigma}_\rho - I_p \right\|_{\text{op}} L + \left\| \widehat{\Sigma}_\rho^{-1} \right\|_{\text{op}} \|\bar{X}\| \geq t \right\} \\ &\leq \mathbb{P} \left\{ \left\| \widehat{\Sigma}_\rho^{-1} \right\|_{\text{op}} \left\| \widehat{\Sigma}_\rho - I_p \right\|_{\text{op}} L \geq \frac{L}{L+1} t \right\} + \mathbb{P} \left\{ \left\| \widehat{\Sigma}_\rho^{-1} \right\|_{\text{op}} \|\bar{X}\| \geq \frac{1}{L+1} t \right\} \end{aligned}$$

Hence for $L \geq 1$, one has

$$\begin{aligned} &\mathbb{P} \left\{ \sup_{x \in B_\mu(L)} \|W_{1n}(x)\| > t \right\} \\ &\leq \mathbb{P} \left\{ \left\| \widehat{\Sigma}_\rho^{-1} \right\|_{\text{op}} \left\| \widehat{\Sigma}_\rho - I_p \right\|_{\text{op}} \geq \frac{t}{2L} \right\} + \mathbb{P} \left\{ \left\| \widehat{\Sigma}_\rho^{-1} \right\|_{\text{op}} \|\bar{X}\| \geq \frac{t}{2L} \right\} \end{aligned}$$

$$\begin{aligned}
 &\leq \mathbb{P} \left(\left\{ \left\| \widehat{\Sigma}_\rho^{-1} \right\|_{\text{op}} \leq 2 \right\} \cap \left\{ \left\| \widehat{\Sigma}_\rho - I_p \right\|_{\text{op}} \geq t/(4L) \right\} \right) \\
 &+ \mathbb{P} \left(\left\{ 2 < \left\| \widehat{\Sigma}_\rho^{-1} \right\|_{\text{op}} \leq n \right\} \cap \left\{ \left\| \widehat{\Sigma}_\rho - I_p \right\|_{\text{op}} \geq t/(2nL) \right\} \right) \\
 &+ \mathbb{P} \left(\left\{ \left\| \widehat{\Sigma}_\rho^{-1} \right\|_{\text{op}} > n \right\} \right) \\
 &+ \mathbb{P} \left(\left\{ \left\| \widehat{\Sigma}_\rho^{-1} \right\|_{\text{op}} \leq 2 \right\} \cap \{ \|\bar{X}\| \geq t/(4L) \} \right) \\
 &+ \mathbb{P} \left(\left\{ 2 < \left\| \widehat{\Sigma}_\rho^{-1} \right\|_{\text{op}} \leq n \right\} \cap \{ \|\bar{X}\| \geq t/(2nL) \} \right) \\
 &+ \mathbb{P} \left(\left\{ \left\| \widehat{\Sigma}_\rho^{-1} \right\|_{\text{op}} > n \right\} \right) \\
 &\lesssim \underbrace{\mathbb{P} \left\{ \left\| \widehat{\Sigma}_\rho - I_p \right\|_{\text{op}} \geq \frac{t}{4L} \right\} + \mathbb{P} \left\{ \|\bar{X}\| \geq \frac{t}{4L} \right\}}_{\text{(i)}} \\
 &+ \underbrace{\mathbb{P} \left\{ 2 < \left\| \widehat{\Sigma}_\rho^{-1} \right\|_{\text{op}} \leq n \right\} \wedge \left[\mathbb{P} \left\{ \left\| \widehat{\Sigma}_\rho - I_p \right\|_{\text{op}} \geq t/(2nL) \right\} + \mathbb{P} \{ \|\bar{X}\| \geq t/(2nL) \} \right]}_{\text{(ii)}} \\
 &+ \underbrace{\mathbb{P} \left(\left\{ \left\| \widehat{\Sigma}_\rho^{-1} \right\|_{\text{op}} > n \right\} \right)}_{\text{(iii)}}
 \end{aligned}$$

- If $\rho = 0$, then

$$\text{(ii)} + \text{(iii)} \leq \mathbb{P} \left\{ \left\| \widehat{\Sigma}_\rho^{-1} \right\|_{\text{op}} > 2 \right\} \lesssim \exp(-cn)$$

As a result, for any $L \geq 1$,

$$\begin{aligned}
 \mathbb{P} \left\{ \sup_{x \in B_\mu(L)} \|W_{1n}(x)\| > t \right\} &\lesssim \exp(-cnt/L) \vee \exp(-cnt^2/L^2) + \exp(-cnt^2/L^2) + \exp(-cn) \\
 &= \begin{cases} \exp(-cnt^2/L^2), & t < L \\ \exp(-cn), & t \geq L \end{cases}
 \end{aligned}$$

Therefore, for any $s > 0$, $L \geq 1$,

$$\mathbb{P} \left\{ \sup_{x \in B_\mu(L)} \|W_{1n}(x)\| > \frac{L}{\sqrt{n}} s \right\} \lesssim \begin{cases} \exp(-cs^2), & s < \sqrt{n} \\ \exp(-cn), & s > \sqrt{n} \end{cases}$$

- If $\rho = n^{-1}$, then $\text{(iii)} = 0$. As a result, for any $t = sL/\sqrt{n}$ and $L \geq 1$,

$$\mathbb{P} \left\{ \sup_{x \in B_\mu(L)} \|W_{1n}(x)\| > \frac{L}{\sqrt{n}} s \right\}$$

$$\lesssim \begin{cases} \exp(-cs^2), & 0 < s < \sqrt{n} \\ \exp(-cn), & \sqrt{n} \leq s < n^{3/2} \\ \exp(-cs/\sqrt{n}), & s \geq n^{3/2} \end{cases}$$

Combining results for $W_{0n}(x)$ and $W_{1n}(x)$, one can obtain that for any $\rho \in \{0, n^{-1}\}$, $L \geq 1$, $\tau \geq 0$ and $1 \leq s \leq \sqrt{n}$,

$$\mathbb{P} \left\{ \sup_{x \in B_\mu(L)} |W_{0n}(x)| > \frac{L}{\sqrt{n}} s \right\} \vee \mathbb{P} \left\{ \sup_{x \in B_\mu(L)} \|W_{1n}(x)\| > \frac{L}{\sqrt{n}} s \right\} \lesssim \exp(-cs)$$

By taking $s = O((1 + \tau) \log n)$, the proof is then complete. \blacksquare

Lemma 28 *Let $\{(X_i, Q_i)\}_{i=1}^n$ be i.i.d. samples satisfying Assumption 1-2. Let $(\mathcal{V}, \|\cdot\|)$ be a normed vector space and $\psi : \mathcal{S}_d^{++} \times \Theta \rightarrow \mathcal{V}$ is a mapping parametrized by $\theta \in \Theta$. Suppose there exists an event \mathcal{E}_0 under which*

- $\|X_i - \mu\| \leq L_n$ for $i = 1, \dots, n$.
- $\|\psi(Q_i, \theta)\| \leq K_n$ for $i = 1, \dots, n$ uniformly for $\theta \in \Theta$.

Denote

$$G(L, \Theta; n) := \sup_{x \in B_\mu(L)} \sup_{\theta \in \Theta} \left\| \frac{1}{n} \sum_{i=1}^n [w_{n,\rho}(x, X_i) - w(x, X_i)] \psi(Q_i; \theta) \right\|$$

Then for any $\rho \in \{0, n^{-1}\}$, $L \geq 1$ and any $\tau \geq 0$, there exists constant C independent of n such that

$$\mathbb{P} \left\{ G(L, \Theta; n) \geq CK_n(L_n + 1) \frac{L}{\sqrt{n}} (1 + \tau) \log n \right\} \leq O(n^{-(1+\tau)}) + \mathbb{P}(\mathcal{E}_0^c)$$

Proof Given $s \geq 1$, define event $\mathcal{E}_1(s)$

$$\mathcal{E}_1(s) := \sup_{x \in B_\mu(L)} \sup_{\theta \in \Theta} \left\| \frac{1}{n} \sum_{i=1}^n [w_{n,\rho}(x, X_i) - w(x, X_i)] \psi(Q_i; \theta) \right\| \geq 2K_n(L_n + 1) \cdot \frac{L}{\sqrt{n}} s$$

Then

$$\mathbb{P}(\mathcal{E}_1(s)) \leq \mathbb{P}(\mathcal{E}_1(s) \cap \mathcal{E}_0) + \mathbb{P}(\mathcal{E}_0^c)$$

By (37) in Lemma 27, one can obtain for any x and $\theta \in \Theta$,

$$\begin{aligned} & \left\| \frac{1}{n} \sum_{i=1}^n w_{n,\rho}(x, X_i) \psi(Q_i; \theta) - \frac{1}{n} \sum_{i=1}^n w(x, X_i) \psi(Q_i; \theta) \right\| \\ &= \left\| W_{0n}(x) \frac{1}{n} \sum_{i=1}^n \psi(Q_i; \theta) + \frac{1}{n} \sum_{i=1}^n W_{1n}(x)^\top X_i \psi(Q_i; \theta) \right\| \end{aligned}$$

$$\leq |W_{0n}(x)| \cdot \frac{1}{n} \sum_{i=1}^n \|\psi(Q_i; \theta)\| + \|W_{1n}(x)\|_2 \frac{1}{n} \sum_{i=1}^n \|X_i\|_2 \|\psi(Q_i; \theta)\|$$

Hence

$$\begin{aligned} \mathbb{P}(\mathcal{E}_1(s) \cap \mathcal{E}_0) &\leq \mathbb{P} \left\{ \sup_{x \in B_\mu(L)} [|W_{0n}(x)| \cdot K_n + \|W_{1n}(x)\|_2 \cdot L_n K_n] \geq 2K_n(L_n + 1) \cdot \frac{L}{\sqrt{n}} s \right\} \\ &= \mathbb{P} \left\{ \sup_{x \in B_\mu(L)} [|W_{0n}(x)| + \|W_{1n}(x)\|_2] \geq 2 \cdot \frac{L}{\sqrt{n}} s \right\} \\ &\leq \mathbb{P} \left\{ \sup_{x \in B_\mu(L)} |W_{0n}(x)| \geq \frac{L}{\sqrt{n}} s \right\} + \mathbb{P} \left\{ \sup_{x \in B_\mu(L)} \|W_{1n}(x)\|_2 \geq \frac{L}{\sqrt{n}} s \right\} \end{aligned}$$

Applying (38) and (38) in Lemma 27 gives the desired result. \blacksquare

We will need uniform upper bounds for various quantities of the form

$$\sup_{\theta \in \Theta} f(Z_1^n; \theta)$$

where Z_1^n denotes (Z_1, \dots, Z_n) . To this end, we decompose the above quantity as follows.

$$\underbrace{\sup_{\theta \in \Theta} f(Z_1^n; \theta) - \mathbb{E} \sup_{\theta \in \Theta} f(Z_1^n; \theta)}_{\text{perturbation}} + \underbrace{\mathbb{E} \sup_{\theta \in \Theta} f(Z_1^n; \theta)}_{\text{expectation}}$$

The expectation term is bounded in Lemma 29 below with a chaining argument, while the perturbation term is shown to concentrate in Lemma 31 by exploiting its bounded difference property. The proof of Lemma 29 is deferred to Appendix B.2.1.

Lemma 29 *Given a function class*

$$\mathcal{F}(\Theta) = \{f(\cdot; \theta) : \mathbb{R}^p \rightarrow \mathbb{R} : \theta \in \Theta\}$$

where f is continuous jointly in (z, θ) and (Θ, d) is a separable metric space with finite diameter $D := \sup_{\theta_1, \theta_2 \in \Theta} d(\theta_1, \theta_2) \gtrsim 1$. Suppose a random vector $Z \in \mathbb{R}^p$ satisfies the following inequality

$$\|f(Z; \theta_1) - f(Z; \theta_2)\|_{\psi_2} \leq \tau(d(\theta_1, \theta_2)), \quad \forall \theta_1, \theta_2 \in \Theta \quad (40)$$

for some increasing function $\tau : \mathbb{R}^+ \rightarrow \mathbb{R}^+$ that satisfies $\tau(0) = 0, \tau(+\infty) = +\infty$. Then

- the following inequality holds

$$\mathbb{E} \sup_{\theta \in \Theta} |f(Z; \theta)| \leq C \int_0^{\tau(D)} \sqrt{\log N(\tau^{-1}(t); \Theta)} \cdot dt + \sup_{\theta \in \Theta} \mathbb{E} |f(Z; \theta)| \quad (41)$$

where $C > 0$ is a fixed absolute constant and $N(\epsilon; \Theta)$ is the ϵ -covering number of Θ .

- Specifically, if τ has the form

$$\tau(\epsilon) = \frac{K}{\sqrt{n}} (\epsilon \vee \epsilon^{\alpha_0})$$

for some constant $\alpha_0 > 0$, then for any $D \gtrsim 1$, one has

$$\int_0^{\tau(D)} \sqrt{\log N(\tau^{-1}(t); \Theta)} \cdot dt \lesssim \frac{KD^{1 \vee \alpha_0}}{\sqrt{n}} \sqrt{\log^+ D} \quad (42)$$

- Specifically, if $Z_1, \dots, Z_n \in \mathbb{R}^p$ are i.i.d. and $f : (\mathbb{R}^p)^k \times \Theta \rightarrow \mathbb{R}$ is equal to

$$f(Z_1^n; \theta) = \left\| \frac{1}{n} \sum_{i=1}^n \psi(Z_i; \theta) - \mathbb{E} \frac{1}{n} \sum_{i=1}^n \psi(Z_i; \theta) \right\|, \quad Z_1^n := (Z_1, \dots, Z_n)$$

where $\psi(z; \theta) \in \mathcal{M}_{k,s}(\mathbb{R}^m; \mathbb{R}^m)$ is a symmetric k -linear operator for any $z \in \mathbb{R}^p, \theta \in \Theta$. Then the following inequality

$$\left\| f(Z_1^n; \theta) - f(Z_1^n; \tilde{\theta}) \right\|_{\psi_2} \lesssim \frac{1}{\sqrt{n}} \left\| \left\| \psi(Z_1; \theta) - \psi(Z_1; \tilde{\theta}) \right\| \right\|_{\psi_2} \quad (43)$$

holds.

Remark 30 (43) will be applied in three ways as follows.

- $m = 1, k = 1$: $\psi(z; \theta) \in \mathbb{R}$ and $\|\psi(z; \theta)\|$ reduces to the absolute value of $\psi(z; \theta)$. Examples include $\psi(X, Q; x, S) = w(x, X)W^2(Q, S)$ in Lemma 38
- $m = d, k = 1$: $\psi(z; \theta) \in \mathcal{L}(\mathbb{R}^m; \mathbb{R}^m)$ can be viewed as a matrix. Moreover, $\|\psi(z; \theta)\|$ reduces to the matrix operator norm. Examples include $\psi(X, Q; x) = w(x, X)(T_{Q^*(x)}^Q - I_d)$ in Lemma 40.
- $m = d \times d, k = 1$: this is a special case of the previous one by identifying $d \times d$ matrices as a vector in \mathbb{R}^{d^2} . Examples include $\psi(X, Q; x) = -w(x, X)dT_{Q^*(x)}^Q$ in Lemma 40.
- $m = d \times d, k = 2$: Examples include $\psi(X, Q; x, S) = w(x, X)d^2T_S^Q$ in Lemma 40.

Lemma 31 Let $\psi : \mathbb{R}^p \times \Theta \rightarrow \mathbb{R}^q$ be a class of functions indexed by $\theta \in \Theta$. Suppose ψ is uniformly bounded in the sense that there exists a finite constant K such that $\|\psi(x; \theta)\|_2 \leq K$ for any $x \in \mathbb{R}^p$ and $\theta \in \Theta$. For any fixed $y_0 \in \mathcal{W}$, define $f : (\mathbb{R}^p)^n \times \Theta \rightarrow \mathbb{R}^q$ as

$$f(x_1^n) := \sup_{\theta \in \Theta} \left\| \frac{1}{n} \sum_{i=1}^n \psi(x_i; \theta) - y_0 \right\|_2$$

Then f satisfies the bounded difference property with parameter $2K/n$, i.e. for any $x_1^n, \tilde{x}_1^n \in (\mathbb{R}^p)^n$ such that $\sum_{i=1}^n \mathbb{I}(x_i \neq \tilde{x}_i) \leq 1$,

$$\|f(x_1^n) - f(\tilde{x}_1^n)\|_2 \leq \frac{2K}{n}, \quad (44)$$

Moreover, suppose X_1, \dots, X_n are i.i.d. random element in \mathbb{R}^p , then

$$\|f(X_1^n) - \mathbb{E}f(X_1^n)\|_{\psi_2} \lesssim \frac{2K}{\sqrt{n}} \quad (45)$$

Proof Without loss of generality, assume $x_1 \neq \tilde{x}_1$ and $x_j = \tilde{x}_j$ for $j = 2, \dots, n$. Then one has

$$\begin{aligned}
 & \left\| \frac{1}{n} \sum_{i=1}^n \psi(x_i; \theta) - y_0 \right\|_2 - \sup_{\tilde{\theta} \in \Theta} \left\| \frac{1}{n} \sum_{i=1}^n \psi(\tilde{x}_i; \tilde{\theta}) - y_0 \right\|_2 \\
 & \leq \left\| \frac{1}{n} \sum_{i=1}^n \psi(x_i; \theta) - y_0 \right\|_2 - \left\| \frac{1}{n} \sum_{i=1}^n \psi(\tilde{x}_i; \theta) - y_0 \right\|_2 \\
 & \leq \left\| \frac{1}{n} \sum_{i=1}^n \psi(x_i; \theta) - y_0 - \frac{1}{n} \sum_{i=1}^n \psi(\tilde{x}_i; \theta) + y_0 \right\|_2 \\
 & = \left\| \frac{1}{n} \psi(x_1; \theta) - \psi(\tilde{x}_1; \theta) \right\|_2 \\
 & \leq \frac{2K}{n}
 \end{aligned}$$

Taking supremum over θ then gives

$$f(x_1^n) - f(\tilde{x}_1^n) \leq \frac{2K}{n}$$

The other direction can be obtained with the role x_1^n and \tilde{x}_1^n reversed. Therefore we get (44). By Corollary 2.21 in Wainwright (2019), we get (45). The proof is then complete. \blacksquare

Finally, even though Theorem 7 shows fast convergence $\hat{Q}_\rho(x) \rightarrow Q^*(x)$, one needs $\hat{Q}_\rho(\bar{X}) \rightarrow Q^*(\mu)$ when considering power in Theorem 18, and this is stated in Lemma 32 below. A crucial observation here is that $\hat{Q}_\rho(\bar{X})$ and $Q^*(\mu)$ are the empirical and population Fréchet mean respectively. The proof is built upon Le Gouic et al. (2022) and Altschuler et al. (2021).

Lemma 32 *Assume Assumption 1 and 2 hold. Then there exists an event \tilde{E} with probability at least $1 - n^{-100}$ under which the following holds*

$$W^2 \left(\hat{Q}_\rho(\bar{X}), Q^*(\mu) \right) \lesssim \frac{\text{polylog}(n)}{\sqrt{n}} \quad (46)$$

As a result, under \tilde{E} , one has

$$\left\| \hat{Q}_\rho(\bar{X}) - Q^*(\mu) \right\|_F \lesssim \frac{\text{polylog}(n)}{\sqrt{n}} \quad (47)$$

Proof Note that $\hat{Q}_\rho(\bar{X})$ and $Q^*(\mu)$ are equal to the empirical and population barycenter respectively. For simplicity, we write Q^* for $Q^*(\mu)$.

Proof of (46): From (3.8)-(3.9) in Le Gouic et al. (2022), the variance inequality in Chewi et al. (2020, Theorem 6) and results in Altschuler et al. (2021), one can obtain the following

$$W^2\left(\widehat{Q}_\rho(\bar{X}), Q^*\right) \leq C_b \left\| \frac{1}{n} \sum_{i=1}^n \left(T_{Q^*}^{Q_i} - I_d\right) \right\|_{Q^*} \quad (48)$$

where $C_b > 0$ is a constant independent of n and $\|A\|_{Q^*(\mu)} := \langle A, Q^* A \rangle$. Moreover, Lemma 33 implies that $c_1^{-1} I_d \preceq Q^* \preceq c_1 I_d$; see Assumption 2 for the definition of c_1 . As a result, one can obtain

$$\begin{aligned} \left\| \frac{1}{n} \sum_{i=1}^n \left(T_{Q^*}^{Q_i} - I_d\right) \right\|_{Q^*} &\leq \lambda_{\max}(Q^*) \cdot \left\| \frac{1}{n} \sum_{i=1}^n \left(T_{Q^*}^{Q_i} - I_d\right) \right\|_{\text{F}} \\ &\leq c_\Lambda \left\| \frac{1}{n} \sum_{i=1}^n \left(T_{Q^*}^{Q_i} - I_d\right) \right\|_{\text{F}} \end{aligned} \quad (49)$$

Note that by the optimality condition for barycenter, one has $\mathbb{E}\left(T_{Q^*}^Q - I_d\right) = 0$. Then one can apply Lemma 25 to get that

$$\begin{aligned} \left\| \left\| \frac{1}{n} \sum_{i=1}^n \left(T_{Q^*}^{Q_i} - I_d\right) \right\|_{\text{F}} \right\|_{\psi_2} &\lesssim \sup_{\|U\|_{\text{F}} \leq 1} \left\| \left\langle U, \frac{1}{n} \sum_{i=1}^n \left(T_{Q^*}^{Q_i} - I_d\right) \right\rangle \right\|_{\psi_2} \\ &\lesssim \sup_{\|U\|_{\text{F}} \leq 1} \frac{1}{\sqrt{n}} \left\| \left\langle U, T_{Q^*}^Q - I_d \right\rangle \right\|_{\psi_2} \\ &\leq \frac{1}{\sqrt{n}} \left\| \left\| T_{Q^*}^Q - I_d \right\|_{\text{F}} \right\|_{\psi_2} \end{aligned} \quad (50)$$

Recall that $T_S^Q = S^{-1/2} (S^{1/2} Q S^{1/2})^{1/2} S^{-1/2}$, then one has

$$\begin{aligned} \left\| \left\| T_{Q^*}^Q - I_d \right\|_{\text{F}} \right\|_{\psi_2} &\stackrel{(i)}{\lesssim} \left\| \left\| T_{Q^*}^Q - I_d \right\|_{\text{op}} \right\|_{\psi_2} \\ &\leq \left\| 1 + c_\Lambda^{3/2} \lambda_{\max}(Q)^{1/2} \right\|_{\psi_2} \\ &\stackrel{(ii)}{\leq} \left\| 1 + c_\Lambda^{3/2} c_\Lambda^{1/2} \|X - \mu\|^{1/2} \right\|_{\psi_2} \\ &\leq \left\| 1 + c_\Lambda^{3/2} c_\Lambda^{1/2} (1 \vee \|X - \mu\|) \right\|_{\psi_2} \\ &\leq \left\| 1 + c_\Lambda^2 (1 + \|X - \mu\|) \right\|_{\psi_2} \\ &\lesssim 1 \end{aligned} \quad (51)$$

Here (i) follows since dimension d is fixed and absorbed into the constant factor independent of n , (ii) is a result of Assumption 2.

Finally, combining (48) (49) (50) and (51) gives

$$\left\| W^2\left(\widehat{Q}_\rho(\bar{X}), Q^*(\mu)\right) \right\|_{\psi_2} \lesssim \frac{1}{\sqrt{n}} \quad (52)$$

which implies (46).

Proof of (47): Apply ((iv)) in Lemma 21 to get that under the event \tilde{E} ,

$$\begin{aligned} \left\| \hat{Q}_\rho(\bar{X}) - Q^* \right\|_F &\leq W^2 \left(\hat{Q}_\rho(\bar{X}), Q^* \right) \cdot \frac{1 + \lambda_{\min}^{-1}(Q^*) \lambda_{\max} \left(\hat{Q}_\rho(\bar{X}) \right)}{\lambda_{\max}(Q^*) \cdot \lambda_{\min}^{-2}(Q^*)} \\ &\lesssim W^2 \left(\hat{Q}_\rho(\bar{X}), Q^* \right) \cdot \left(1 + \lambda_{\max} \left(\hat{Q}_\rho(\bar{X}) \right) \right) \\ &\leq C \frac{\sqrt{\log n}}{\sqrt{n}} \cdot \left(1 + \lambda_{\max} \left(\hat{Q}_\rho(\bar{X}) \right) \right) \end{aligned} \quad (53)$$

for constant $C > 0$ large enough. Note that this implies

$$\begin{aligned} \lambda_{\max} \left(\hat{Q}_\rho(\bar{X}) \right) &\leq \lambda_{\max}(Q^*) + \left\| \hat{Q}_\rho(\bar{X}) - Q^* \right\|_F \\ &\leq c_\Lambda + C \frac{\sqrt{\log n}}{\sqrt{n}} \cdot \left(1 + \lambda_{\max} \left(\hat{Q}_\rho(\bar{X}) \right) \right) \end{aligned}$$

Solving for $\lambda_{\max} \left(\hat{Q}_\rho(\bar{X}) \right)$ gives

$$\begin{aligned} \lambda_{\max} \left(\hat{Q}_\rho(\bar{X}) \right) &\leq \frac{c_\Lambda + C \frac{\sqrt{\log n}}{\sqrt{n}}}{1 - C \frac{\sqrt{\log n}}{\sqrt{n}}} \\ &\lesssim 1 \quad \text{for } n \text{ large enough} \end{aligned} \quad (54)$$

Finally, plugging (54) back into (53) gives (47). ■

B.2.1 PROOF OF LEMMA 29

Proof of (41): Since $f(z, \theta)$ is continuous in θ and Θ is separable, we have for any countable, dense subset $\mathbb{U} \subset \Theta$, the following equality

$$\mathbb{E} \sup_{\theta \in \Theta} |f(Z, \theta)| = \mathbb{E} \sup_{S \in \mathbb{U}} |f(Z, \theta)|$$

holds. By the monotone convergence theorem, it suffices to assume that \mathbb{U} is finite and get an upper bound that is independent of the cardinality of \mathbb{U} .

For each $k \in \mathbb{Z}$, let $\mathbb{U}_k \subset \mathbb{U}$ be a minimal ϵ_k -covering set of \mathbb{U} where ϵ_k is defined by

$$\epsilon_k = \tau^{-1}(2^{-k}) \quad (55)$$

Let $N(\epsilon_k, \mathbb{U})$ be the cardinality of the minimal ϵ_k -covering set \mathbb{U}_k of \mathbb{U} . Since \mathbb{U} is a subset of Θ , $N(\epsilon_k, \mathbb{U})$ can be upper bounded by

$$\log N(\epsilon_k; \mathbb{U}) = \log |\mathbb{U}_k| \leq \log N(\epsilon_k; \Theta)$$

Since \mathbb{U} is finite, there is a largest $\eta \in \mathbb{Z}$ and a smallest integer $H \in \mathbb{Z}$ such that

$$\mathbb{U}_\eta = \{\theta_0\} \text{ for some } \theta_0 \in \mathbb{U}, \quad \mathbb{U}_H = \mathbb{U}$$

For each $k \in \mathbb{Z}$, define the mapping $\pi_k : \Theta \rightarrow \mathbb{U}_k$ via

$$\pi_k(\theta) = \underset{\tilde{\theta} \in \mathbb{U}_k}{\operatorname{argmin}} d(\tilde{\theta}, \theta)$$

so that $\pi_k(S)$ is the best approximation of $\theta \in \Theta$ from the set \mathbb{U}_k .

For any $\theta \in \mathbb{U}$, apply the triangle inequality to see that

$$\begin{aligned} \mathbb{E} \left| \max_{\theta \in \mathbb{U}} f(Z, \theta) \right| &\leq \mathbb{E} \left| \max_{\theta \in \mathbb{U}} (f(Z, \theta) - f(Z, \theta_0)) \right| + \mathbb{E} |f(Z; \theta_0)| \\ &\leq \mathbb{E} \max_{\theta \in \mathbb{U}} |f(Z, \theta) - f(Z, \theta_0)| + \mathbb{E} |f(Z; \theta_0)| \\ &\stackrel{(i)}{=} \mathbb{E} \max_{\theta \in \mathbb{U}} \left| \sum_{k=\eta+1}^H (f(Z, \pi_k(\theta)) - f(Z, \pi_{k-1}(\theta))) \right| + \mathbb{E} |f(Z; \theta_0)| \\ &\leq \mathbb{E} \max_{\theta \in \mathbb{U}} \sum_{k=\eta+1}^H |f(Z, \pi_k(\theta)) - f(Z, \pi_{k-1}(\theta))| + \mathbb{E} |f(Z; \theta_0)| \\ &\leq \sum_{k=\eta+1}^H \mathbb{E} \max_{\theta \in \mathbb{U}} |f(Z, \pi_k(\theta)) - f(Z, \pi_{k-1}(\theta))| + \mathbb{E} |f(Z; \theta_0)| \end{aligned} \tag{56}$$

Here (i) is a consequence of decomposing $f(Z, \theta) - f(Z, \theta_0)$ as a telescoping sum

$$f(Z, \theta) - f(Z, \theta_0) = \sum_{k=\eta+1}^H (f(Z, \pi_k(\theta)) - f(Z, \pi_{k-1}(\theta)))$$

For any fixed $\theta \in \Theta$, one has

$$\begin{aligned} \|f(\cdot, \pi_k(\theta)) - f(\cdot, \pi_{k-1}(\theta))\|_{\psi_2} &\leq \|f(\cdot, \pi_k(\theta)) - f(\cdot, \theta)\|_{\psi_2} + \|f(\cdot, \theta) - f(\cdot, \pi_{k-1}(\theta))\|_{\psi_2} \\ &\stackrel{(I)}{\leq} \tau(\epsilon_k) + \tau(\epsilon_{k-1}) \\ &\leq 2\tau(\epsilon_{k-1}) \\ &\stackrel{(II)}{=} 2 \cdot 2^{-(k-1)} \end{aligned}$$

Here (I) is a result of (40), and (II) follows from (55). Then, one can obtain

$$\begin{aligned} \mathbb{E} \max_{\theta \in \mathbb{U}} |f(Z, \pi_k(\theta)) - f(Z, \pi_{k-1}(\theta))| &\stackrel{(i)}{\lesssim} \sqrt{|\mathbb{U}_{k-1}| \cdot |\mathbb{U}_k|} \cdot 2 \cdot 2^{-(k-1)} \\ &\lesssim \sqrt{\log N(\epsilon_k; \Theta)} \cdot 2^{-(k-1)} \end{aligned} \tag{57}$$

Here (i) follows from the properties of the maximum of finitely many sub-Gaussian random variables and the fact that the maximum is taken over at most $|\mathbb{U}_{k-1}| \cdot |\mathbb{U}_k| \leq N(\epsilon_k; \Theta)^2$

random variables. Therefore,

$$\begin{aligned}
 \sum_{k=\eta+1}^H \mathbb{E} \max_{\theta \in \mathbb{U}} |f(Z, \pi_k(\theta)) - f(Z, \pi_{k-1}(\theta))| &\lesssim \sum_{k=\eta+1}^H \sqrt{\log N(\epsilon_k; \Theta)} \cdot 2^{-(k-1)} \\
 &= \sum_{k=\eta+1}^H \sqrt{\log N(\tau^{-1}(2^{-k}); \Theta)} \cdot 2^{-(k-1)} \\
 &\lesssim \int_0^\infty \sqrt{\log N(\tau^{-1}(t); \Theta)} \cdot dt \\
 &\stackrel{(I)}{=} \int_0^{\tau(D)} \sqrt{\log N(\tau^{-1}(t); \Theta)} \cdot dt
 \end{aligned} \tag{58}$$

Here (I) follows by noticing that $\log N(\tau^{-1}(t); \Theta) = 0$ for any $t \geq \tau(D)$.

Combine (56) and (58) to see that

$$\mathbb{E} \sup_{\theta \in \Theta} |f(Z; \theta)| \leq C \int_0^{\tau(D)} \sqrt{\log N(\tau^{-1}(t); \Theta)} \cdot dt + \sup_{\theta \in \Theta} \mathbb{E} |f(Z; \theta)|$$

for some constant $C > 0$.

Proof of (42): Consider two cases $\alpha_0 \in (0, 1]$ and $\alpha_0 > 1$.

Case I $\alpha_0 \in (0, 1]$: Note that since $\alpha_0 \in (0, 1]$, one has

$$\begin{aligned}
 \tau^{-1}(t) &= \left(\frac{\sqrt{nt}}{K} \right) \wedge \left(\frac{\sqrt{nt}}{K} \right)^{1/\alpha_0} \\
 &= \begin{cases} \frac{\sqrt{nt}}{K} & \sqrt{nt} \geq K \\ \left(\frac{\sqrt{nt}}{K} \right)^{1/\alpha_0} & \sqrt{nt} \leq K \end{cases}
 \end{aligned} \tag{59}$$

As a result, one has

$$\begin{aligned}
 &\int_0^{\tau(D)} \sqrt{\log N(\tau^{-1}(t); \Theta)} \cdot dt \\
 &\stackrel{(i)}{=} \frac{1}{\sqrt{n}} \int_0^{\sqrt{n}\tau(D)} \sqrt{\log N\left(\tau^{-1}\left(\frac{t}{\sqrt{n}}\right); \Theta\right)} \cdot dt \\
 &\stackrel{(ii)}{=} \frac{1}{\sqrt{n}} \int_0^K \sqrt{\log N\left((t/K)^{1/\alpha_0}; \Theta\right)} \cdot dt + \frac{1}{\sqrt{n}} \int_K^{\sqrt{n}\tau(D)} \sqrt{\log N(t/K; \Theta)} \cdot dt \\
 &\lesssim \frac{1}{\sqrt{n}} \int_0^K \sqrt{\log^+\left(\frac{D}{(t/K)^{1/\alpha_0}}\right)} \cdot dt + \frac{1}{\sqrt{n}} \int_K^{\sqrt{n}\tau(D)} \sqrt{\log^+\left(\frac{D}{(t/K)}\right)} \cdot dt \\
 &\stackrel{(iii)}{=} \underbrace{\frac{K}{\sqrt{n}} \int_0^1 \sqrt{\log^+\left(\frac{D}{t^{1/\alpha_0}}\right)} \cdot dt}_{a_1} + \underbrace{\frac{K}{\sqrt{n}} \int_1^{\sqrt{n}\tau(D)/K} \sqrt{\log^+\left(\frac{D}{t}\right)} \cdot dt}_{a_2}
 \end{aligned} \tag{60}$$

Here (i), (iii) follows from a change of variables, and (ii) follows from (59).

- a_1 : one has

$$\begin{aligned}
 a_1 &\stackrel{(i)}{\leq} \int_0^1 \sqrt{\log^+ \left(\frac{1}{t^{1/\alpha_0}} \right)} dt + \int_0^1 \sqrt{\log^+ D} dt \\
 &\stackrel{(ii)}{\lesssim} \sqrt{\log^+ D}
 \end{aligned} \tag{61}$$

Here (i) follows from the inequality $\sqrt{s+t} \leq \sqrt{s} + \sqrt{t}$ for $s, t \geq 0$, and (ii) follows from the assumption that $D \gtrsim 1$.

- a_2 : for $D \gtrsim 1$,

$$\begin{aligned}
 a_2 &\stackrel{(i)}{\leq} \int_1^{\sqrt{n}\tau(D)/K} \sqrt{\log^+ D} dt \\
 &\stackrel{(ii)}{\leq} \frac{\sqrt{n}\tau(D)}{K} \cdot \sqrt{\log^+ D} \\
 &= (D \vee D^{\alpha_0}) \sqrt{\log^+ D} \\
 &\stackrel{(iii)}{\lesssim} D \sqrt{\log^+ D}
 \end{aligned} \tag{62}$$

Here (i) a consequence of the fact that $D/t \leq D$ for $t \geq 1$, (ii) results from substituting the definition of $\tau(\cdot)$ and (iii) follows from the assumption $\alpha_0 \in (0, 1]$.

Combine (60), (61) and (62) to see that

$$\int_0^{\tau(D)} \sqrt{\log N(\tau^{-1}(t); \Theta)} \cdot dt \lesssim \frac{K}{\sqrt{n}} D \sqrt{\log^+ D}$$

which proves (42) for $\alpha_0 \in (0, 1]$.

Case II $\alpha_0 > 1$: for $\alpha_0 > 1$, one has

$$\begin{aligned}
 \tau^{-1}(t) &= \left(\frac{\sqrt{nt}}{K} \right) \wedge \left(\frac{\sqrt{nt}}{K} \right)^{1/\alpha_0} \\
 &= \begin{cases} \frac{\sqrt{nt}}{K} & \sqrt{nt} \leq K \\ \left(\frac{\sqrt{nt}}{K} \right)^{1/\alpha_0} & \sqrt{nt} > K \end{cases}
 \end{aligned} \tag{63}$$

As a result, one can obtain

$$\begin{aligned}
 & \int_0^{\tau(D)} \sqrt{\log N(\tau^{-1}(t); \Theta)} \cdot dt \\
 & \stackrel{(i)}{=} \frac{1}{\sqrt{n}} \int_0^{\sqrt{n}\tau(D)} \sqrt{\log N\left(\tau^{-1}\left(\frac{t}{\sqrt{n}}\right); \Theta\right)} \cdot dt \\
 & \stackrel{(ii)}{=} \frac{1}{\sqrt{n}} \int_0^K \sqrt{\log N(t/K; \Theta)} \cdot dt + \frac{1}{\sqrt{n}} \int_K^{\sqrt{n}\tau(D)} \sqrt{\log N((t/K)^{1/\alpha_0}; \Theta)} \cdot dt \\
 & \lesssim \frac{1}{\sqrt{n}} \int_0^K \sqrt{\log^+\left(\frac{D}{t/K}\right)} \cdot dt + \frac{1}{\sqrt{n}} \int_K^{\sqrt{n}\tau(D)} \sqrt{\log^+\left(\frac{D}{(t/K)^{1/\alpha_0}}\right)} \cdot dt \\
 & \stackrel{(iii)}{=} \underbrace{\frac{K}{\sqrt{n}} \int_0^1 \sqrt{\log^+\left(\frac{D}{t}\right)} \cdot dt}_{a_3} + \underbrace{\frac{K}{\sqrt{n}} \int_1^{\sqrt{n}\tau(D)/K} \sqrt{\log^+\left(\frac{D}{t^{1/\alpha_0}}\right)} \cdot dt}_{a_4}
 \end{aligned} \tag{64}$$

Here (i), (iii) follows from a change of variables, and (ii) follows from (63).

- a_3 : one has

$$\begin{aligned}
 a_3 & \stackrel{(i)}{\leq} \int_0^1 \sqrt{\log^+\left(\frac{1}{t}\right)} dt + \int_0^1 \sqrt{\log^+ D} dt \\
 & \stackrel{(ii)}{\lesssim} \sqrt{\log^+ D}
 \end{aligned} \tag{65}$$

Here (i) follows from the inequality $\sqrt{s+t} \leq \sqrt{s} + \sqrt{t}$ for $s, t \geq 0$, and (ii) follows from the assumption that $D \gtrsim 1$.

- a_4 : for $D \gtrsim 1$,

$$\begin{aligned}
 a_4 & \stackrel{(i)}{\leq} \int_1^{\sqrt{n}\tau(D)/K} \sqrt{\log^+ D} dt \\
 & \stackrel{(ii)}{\leq} \frac{\sqrt{n}\tau(D)}{K} \cdot \sqrt{\log^+ D} \\
 & = (D \vee D^{\alpha_0}) \sqrt{\log^+ D} \\
 & \stackrel{(iii)}{\lesssim} D^{\alpha_0} \sqrt{\log^+ D}
 \end{aligned} \tag{66}$$

Here (i) a consequence of the fact that $D/(t^{1/\alpha_0}) \leq D$ for $t \geq 1$, (ii) results from substituting the definition of $\tau(\cdot)$ and (iii) follows from the assumption $\alpha_0 > 1$.

Combine (64), (65) and (66) to see that

$$\int_0^{\tau(D)} \sqrt{\log N(\tau^{-1}(t); \Theta)} \cdot dt \lesssim \frac{K}{\sqrt{n}} D^{\alpha_0} \sqrt{\log^+ D}$$

which proves (42) for $\alpha_0 > 1$.

Proof of (43): Let $\Psi_n(\theta) := \frac{1}{n} \sum_{i=1}^n \psi(Z_i; \theta)$, then one has

$$\begin{aligned}
 & \left\| f(Z_1^n; \theta) - f(Z_1^n; \tilde{\theta}) \right\|_{\psi_2} \\
 &= \left\| \left\| \Psi_n(\theta) - \mathbb{E} \Psi_n(\theta) \right\| - \left\| \Psi_n(\tilde{\theta}) - \mathbb{E} \Psi_n(\tilde{\theta}) \right\| \right\|_{\psi_2} \\
 &\leq \left\| \left\| \Psi_n(\theta) - \mathbb{E} \Psi_n(\theta) - \Psi_n(\tilde{\theta}) + \mathbb{E} \Psi_n(\tilde{\theta}) \right\| \right\|_{\psi_2} \\
 &\stackrel{(i)}{\lesssim} \sup_{u, v \in \mathbb{S}^{m-1}} \left\| \left\langle u, \left[\tilde{\Psi}_n(\theta) - \mathbb{E} \tilde{\Psi}_n(\theta) \right] - \left[\tilde{\Psi}_n(\tilde{\theta}) + \tilde{\Psi}_n(\tilde{\theta}) \right] \cdot v^{\otimes k} \right\rangle \right\|_{\psi_2} \\
 &\stackrel{(ii)}{\lesssim} \frac{1}{\sqrt{n}} \sup_{u, v \in \mathbb{S}^{m-1}} \left\| \left\langle u, \left[\psi(Z; \theta) - \psi(Z; \tilde{\theta}) \right] - \mathbb{E} \left[\psi(Z; \theta) - \psi(Z; \tilde{\theta}) \right] \cdot v^{\otimes k} \right\rangle \right\|_{\psi_2} \\
 &\stackrel{(iii)}{\lesssim} \frac{1}{\sqrt{n}} \sup_{u, v \in \mathbb{S}^{m-1}} \left\| \left\langle u, \left[\psi(Z; \theta) - \psi(Z; \tilde{\theta}) \right] \cdot v^{\otimes k} \right\rangle \right\|_{\psi_2} \\
 &\leq \frac{1}{\sqrt{n}} \sup_{v \in \mathbb{S}^{m-1}} \left\| \left\| \left[\psi(Z; \theta) - \psi(Z; \tilde{\theta}) \right] \cdot v^{\otimes k} \right\|_{\mathbb{F}} \right\|_{\psi_2} \\
 &\stackrel{(iv)}{\leq} \frac{1}{\sqrt{n}} \sup_{v \in \mathbb{S}^{m-1}} \left\| \left\| \psi(Z; \theta) - \psi(Z; \tilde{\theta}) \right\| \cdot \|v\|_{\mathbb{F}}^k \right\|_{\psi_2} \\
 &\leq \frac{1}{\sqrt{n}} \left\| \left\| \psi(Z; \theta) - \psi(Z; \tilde{\theta}) \right\| \right\|_{\psi_2}
 \end{aligned}$$

Here (i) is a result of Lemma 25, (ii) arises due to independence, (iii) follows from Lemma 2.6.8 in Vershynin (2018), and (iv) is derived from the definition of $\|\cdot\|$. The proof is then complete.

B.3 Properties of F and Q^*

This section collects properties of $F(\cdot, \cdot)$ and $Q^*(\cdot)$ that are needed later in the proof.

Lemma 33 *Assume Assumption 1 and 2 holds. Then the following inequalities*

$$\gamma_\Lambda(\|x - \mu\|)^{-1} I_d \preceq Q^*(x) \preceq \gamma_\Lambda(\|x - \mu\|) I_d, \quad \forall x \in \text{supp } X \quad (67)$$

Moreover, there exist constant $C_0^*, C_1^* > 0$ which only depend on c_Σ, C_{ψ_2} from Assumption 1 and c_Λ, C_1 (from Assumption 2) such that

$$Q^*(x) \preceq [C_0^* + C_1^* \|x - \mu\|]^2 I_d \quad (68)$$

Proof Without loss of generality, assume X have expectation $\mu = 0$ and covariance $\Sigma = I_p$. Denote $\lambda_1(x) = \lambda_{\max}(Q^*(x))$ and $\lambda_d(x) = \lambda_{\min}(Q^*(x))$

Proof of (67): By the optimality condition of $Q^*(x) = \text{argmin}_S \mathbb{E} [W^2(S, Q) | X = x]$. By the differential properties of W^2 in Lemma 21, one can obtain

$$I_d = \mathbb{E} \left[T_{Q^*(x)}^Q | X = x \right]$$

Recall that $T_{Q^*(x)}^Q = Q^*(x)^{-1/2} (Q^*(x)^{1/2} Q Q^*(x)^{1/2})^{1/2} Q^*(x)^{-1/2}$, multiplying both sides of the equation with $Q^*(x)^{1/2}$ then gives

$$Q^*(x) = \mathbb{E} \left[\left(Q^*(x)^{1/2} Q Q^*(x)^{1/2} \right)^{1/2} \mid X = x \right]$$

Then one can obtain

$$\begin{aligned} \lambda_1(x) &= \lambda_{\max} \left(\mathbb{E} \left[\left(Q^*(x)^{1/2} Q Q^*(x)^{1/2} \right)^{1/2} \mid X = x \right] \right) \\ &\leq \mathbb{E} \left[\lambda_{\max} \left(\left(Q^*(x)^{1/2} Q Q^*(x)^{1/2} \right)^{1/2} \mid X = x \right) \right] \\ &\leq \gamma_{\Lambda}(\|x - \mu\|)^{1/2} \lambda_1(x)^{1/2} \end{aligned}$$

Here the second line follows from the convexity of the largest eigenvalue function $\lambda_{\max}(\cdot)$ over PSD matrices. The last inequality follows from Assumption 2. Therefore, one can readily obtain $\lambda_1(x) \leq \gamma_{\Lambda}(\|x - \mu\|)$. The lower bound $\lambda_d(x) \geq \gamma_{\Lambda}(\|x - \mu\|)^{-1}$ can be derived similarly by the concavity of the function $\lambda_{\min}(\cdot)$.

Proof of (68): By the optimality condition of $Q^*(x) = \arg\min_S \mathbb{E} w(x, X) W^2(S, Q)$, one has

$$0 = \mathbb{E} w(x, X) \left(T_{Q^*(x)}^Q - I_d \right)$$

Note that $\mathbb{E} w(x, X) = 1$ for any x , hence one has

$$I_d = \mathbb{E} w(x, X) T_{Q^*(x)}^Q$$

Recall that $T_{Q^*(x)}^Q = Q^*(x)^{-1/2} (Q^*(x)^{1/2} Q Q^*(x)^{1/2})^{1/2} Q^*(x)^{-1/2}$, multiplying both sides of the equation with $Q^*(x)^{1/2}$ then gives

$$Q^*(x) = \mathbb{E} w(x, X) \left(Q^*(x)^{1/2} Q Q^*(x)^{1/2} \right)^{1/2}$$

Taking the largest eigenvalue on both sides, one has

$$\begin{aligned} \lambda_1(x) &\leq \mathbb{E} |w(x, X)| \cdot \left\| \left(Q^*(x)^{1/2} Q Q^*(x)^{1/2} \right)^{1/2} \right\|_{\text{op}} \\ &\leq \mathbb{E} |w(x, X)| \cdot \lambda_{\max}(Q)^{1/2} \lambda_1^{1/2}(x) \end{aligned}$$

which implies

$$\lambda_1(x) \leq \left[\mathbb{E} (1 + \|x\| \cdot \|X\|) \lambda_{\max}(Q)^{1/2} \right]^2$$

Under Assumption 1 and Assumption 2, we have $\lambda_{\max}(Q_i) \leq \gamma_{\Lambda}(\|X_i\|)$, which implies that

$$\begin{aligned} \lambda_1(x) &\leq \left[\mathbb{E} \gamma_{\Lambda}(\|X\|)^{1/2} + \|x\| \mathbb{E} \|X\| \gamma_{\Lambda}(\|X\|)^{1/2} \right]^2 \\ &\leq [C_0^* + C_1^* \|x\|]^2 \end{aligned}$$

where $C_0^*, C_1^* > 0$ are finite constants only depend on c_{Σ}, C_{ψ_2} from Assumption 1 and c_{Λ}, C_{Λ} (from Assumption 2). \blacksquare

Lemma 34 *Assume Assumption 1 and 2 hold. If X and Q are independent, then the following holds.*

- $Q^*(x) \equiv Q^*(\mu)$
- Assumption 5 holds for $C_\lambda = 0$ and c_λ large enough.
- Assumption 4 holds for $\delta_F = 1$, $\alpha_F = 2$, $C_F = 1$ and some constant c_F large enough.

Proof Independence between X and Q implies that

$$\begin{aligned} F(x, S) &= \mathbb{E} w(x, X) W^2(S, Q) \\ &= \mathbb{E} w(x, X) \cdot \mathbb{E} W^2(S, Q) \\ &= \mathbb{E} W^2(S, Q) \end{aligned}$$

The existence and uniqueness of the minimizer of $F(x, \cdot)$ follow from Panaretos and Zemel (2020, Proposition 3.2.3, Proposition 3.2.7). Moreover, since $\mathbb{E} W^2(S, Q)$ does not depend on x , we have $Q^*(x) \equiv Q^*(\mu)$. We denote $Q^* := Q^*(\mu)$ from now on. Note that Lemma 33 implies that $Q^* \in \mathcal{S}_d(c_\Lambda^{-1}, c_\Lambda)$

Verification of Assumption 5: Again by independence, one has

$$\mathbb{E} \left[-w(x, X) dT_{Q^*}^Q \right] = \mathbb{E} \left[-dT_{Q^*}^Q \right]$$

Therefore, one can obtain

$$\begin{aligned} \lambda_{\min} \left(\mathbb{E} \left[-w(x, X) dT_{Q^*}^Q \right] \right) &= \lambda_{\min} \left(\mathbb{E} \left[-dT_{Q^*}^Q \right] \right) \\ &\stackrel{(i)}{\geq} \mathbb{E} \frac{\lambda_{\min}^{1/2}(Q^{1/2} Q^* Q^{1/2})}{2} \cdot \lambda_{\min}^2(Q^{*-1}) \\ &\gtrsim \mathbb{E} \lambda_{\min}^{1/2}(Q) \\ &\stackrel{(ii)}{\gtrsim} \mathbb{E} [\gamma_\Lambda(\|X - \mu\|)]^{-1/2} \\ &\gtrsim \mathbb{E} \frac{1}{1 \vee \|X - \mu\|^{C_\Lambda/2}} \\ &\stackrel{(iii)}{\geq} c_{\text{eig}} \end{aligned}$$

for some constant c_{eig} small enough. Here (i) follows from Lemma 21, (ii) is a result of Assumption 2 and (iii) is due to Assumption 1. The proof is then complete.

Verification of Assumption 4: In order to verify (13). Let us denote $Q' = Q^{*-1/2}SQ^{*-1/2}$. Then ((ii)) and ((iii)) in Lemma 21 imply the following lower bound.

$$\begin{aligned}
 & F(x, S) - F(x, Q^*) \\
 &= \mathbb{E} [W^2(S, Q) - W^2(Q^*, Q)] \\
 &\geq \mathbb{E} \left\langle I_d - T_{Q^*}^Q, S - Q^* \right\rangle + \mathbb{E} \frac{2}{\left(1 + \lambda_{\max}^{1/2}(Q'(x))\right)^2} \left\langle -dT_{Q^*}^Q(S - Q^*), S - Q^* \right\rangle \\
 &\stackrel{(i)}{\geq} \mathbb{E} \frac{2}{\left(1 + \lambda_{\max}^{1/2}(Q'(x))\right)^2} \cdot \frac{\lambda_{\min}^{1/2}(Q^{1/2}Q^*Q^{1/2})}{2} \cdot \left\| Q^{*-1/2}(S - Q^*)Q^{*-1/2} \right\|_F^2
 \end{aligned} \tag{69}$$

Here (i) follows from the fact that $\mathbb{E}(T_{Q^*}^Q - I_d) = 0$ which is a consequence of the optimality condition and independence.

Triangle inequality implies that

$$\{S \in \mathcal{S}_d^{++} : \delta \leq \|S - Q^*\|_F \leq \Delta\} \subset \{S \in \mathcal{S}_d^{++} : \|S\|_{\text{op}} \leq c_\Lambda + \Delta\}$$

which combined with (69) implies for any $\Delta \geq \delta$,

$$\begin{aligned}
 & \inf_{\delta \leq \|S - Q^*\|_F \leq \Delta} F(x, S) - F(x, Q^*) \\
 &\geq \inf_{\delta \leq \|S - Q^*\|_F \leq \Delta} \mathbb{E} \frac{2}{\left(1 + \lambda_{\max}^{1/2}(Q')\right)^2} \cdot \frac{\lambda_{\min}^{1/2}(Q^{1/2}Q^*Q^{1/2})}{2} \cdot \left\| Q^{*-1/2}(S - Q^*)Q^{*-1/2} \right\|_F^2 \\
 &\stackrel{(i)}{\gtrsim} \inf_{\delta \leq \|S - Q^*\|_F \leq \Delta} \frac{1}{\left(1 + \Delta^{1/2}\right)^2} \cdot \|S - Q^*\|_F^2 \\
 &\stackrel{(ii)}{\gtrsim} \inf_{\delta \leq \|S - Q^*\|_F \leq \Delta} \frac{1}{1 + \Delta} \cdot \|S - Q^*\|_{\text{op}}^2 \\
 &\geq c_{\text{sep}} \frac{1}{\Delta} \delta^2
 \end{aligned}$$

where $c_{\text{sep}} > 0$ is a constant independent of n that is small enough. Here (i) follows from the inequalities that $\lambda_{\max}(Q') \lesssim \lambda_{\max}(S) \lesssim \Delta$, $\lambda_{\min}(Q^{1/2}Q^*Q^{1/2}) \gtrsim \lambda_{\min}(Q)$ and $\mathbb{E}\lambda_{\min}(Q)^{1/2} \gtrsim 1$; (ii) holds by choosing $\delta_F = 1$ so that $\Delta \geq \delta_F = 1$.

Combining results above, Assumption 4 holds with $\delta_F = 1$, $\alpha_F = 2$, $C_F = 1$ and $c_F = 1/c_{\text{sep}}$. The proof is then complete. ■

Recall the definition of $\gamma_\Lambda(\cdot)$ from Assumption 2 and the definition of α_F, δ_F, C_F from Assumption 4. Denote $M_L := \gamma_\Lambda(L)$ for any $L > 0$. With these notations in place, we now present the continuity theorem for $Q^*(\cdot)$ as follows.

Lemma 35 (Hölder continuity of $Q^*(\cdot)$) Suppose Assumption 1-5 hold. Then for any $L \geq e$ that satisfies $M_L > \delta_F$, any $x, \tilde{x} \in B_\mu(L)$,

$$\|Q^*(x) - Q^*(\tilde{x})\| \leq C_H \max \left\{ \left(L^{C_F} M_L^{C_F+1} \right)^{1/\alpha_F} \|x - \tilde{x}\|^{1/\alpha_F}, L^{C_F} M_L^{C_F+1} \|x - \tilde{x}\| \right\} \quad (70)$$

holds as long as the constant C_H is large enough. Here C_H is independent of n but depends on d .

Proof Note that $M_L \geq 1$ by the properties of γ_Λ in Assumption 2. The proof is divided into 3 steps.

1. First, we should that $F(x, S)$ is Lipschitz in x for any fixed $S \in \mathcal{S}_d(M_L^{-1}, M_L)$.
2. Next, the Lipschitzness in step 1 and Assumption 4 imply the local Hölder continuity of $Q^*(x)$ with respect to x .
3. Then, an argument based on linear interpolation between x and \tilde{x} implies Lipschitz continuity for x, \tilde{x} separated far apart.

Combining the 3 step above finally gives (70).

To start with, note that Lemma 33 implies

$$Q^*(x) \in \mathcal{S}_d(M_L^{-1}, M_L), \quad \text{for any } x \in B(\mu, L)$$

Recall that $F(x, S) = \mathbb{E} [w(x, X)W^2(S, Q)]$.

- 1. F is Lipschitz in x :** for any $x, \tilde{x} \in B_\mu(L)$, one can obtain

$$\begin{aligned} |F(x, S) - F(\tilde{x}, S)| &= |\mathbb{E}(x - \tilde{x})\Sigma^{-1}(X - \mu)W^2(S, Q)| \\ &\lesssim \|x - \tilde{x}\| \cdot \mathbb{E} \|(X - \mu)W^2(S, Q)\| \\ &\stackrel{(i)}{\lesssim} \|x - \tilde{x}\| \cdot \mathbb{E} \|X - \mu\| (\lambda_{\max}(S) + \lambda_{\max}(Q)) \\ &= \|x - \tilde{x}\| \cdot [\mathbb{E} \|X - \mu\| \lambda_{\max}(S) + \mathbb{E} \|X - \mu\| \lambda_{\max}(Q)] \\ &\stackrel{(ii)}{\lesssim} \|x - \tilde{x}\| (\lambda_{\max}(S) + 1) \end{aligned} \quad (71)$$

Here (i) follows from ((i)) in Lemma 21, and (ii) follows from the concentration of X (Assumption 1) and boundedness of Q given X (Assumption 2).

- 2. Local Hölder continuity:** for any $x, \tilde{x} \in B_\mu(L)$, one has

$$\begin{aligned} 0 &\leq F(x, Q^*(\tilde{x})) - F(x, Q^*(x)) \\ &= (F(x, Q^*(\tilde{x})) - F(\tilde{x}, Q^*(\tilde{x}))) + \underbrace{(F(\tilde{x}, Q^*(\tilde{x})) - F(\tilde{x}, Q^*(x)))}_{\leq 0} + (F(\tilde{x}, Q^*(x)) - F(x, Q^*(x))) \\ &\leq |F(x, Q^*(\tilde{x})) - F(\tilde{x}, Q^*(\tilde{x}))| + |F(\tilde{x}, Q^*(x)) - F(x, Q^*(x))| \\ &\stackrel{(i)}{\lesssim} \|x - \tilde{x}\| \cdot (\lambda_{\max}(Q^*(x)) + \lambda_{\max}(Q^*(\tilde{x})) + 2) \\ &\stackrel{(ii)}{\leq} CM_L \|x - \tilde{x}\| \end{aligned} \quad (72)$$

provided that constant $C > 0$ is large enough. Here (i) follows from (71) and (ii) follows since $Q^*(x), Q^*(\tilde{x}) \in \mathcal{S}_d(M_L^{-1}, M_L)$ for $x, \tilde{x} \in B_\mu(L)$ by Lemma 33.

Note also that for any $x, \tilde{x} \in B_\mu(L)$, one has $\|Q^*(x) - Q^*(\tilde{x})\|_F \leq \sqrt{d}M_L$. Recall the definition of $\delta_F, \alpha_F, \gamma_2(\cdot, \cdot)$ from Assumption 4. Then for any $x, \tilde{x} \in B_\mu(L)$ that satisfy

$$\|x - \tilde{x}\| \leq \frac{\delta_F^{\alpha_F}}{CM_L \gamma_2(L, \sqrt{d}M_L)} =: t_0 \quad (73)$$

one can obtain

$$\|Q^*(x) - Q^*(\tilde{x})\|_F \leq \left(CM_L \gamma_2(L, \sqrt{d}M_L)\right)^{1/\alpha_F} \cdot \|x - \tilde{x}\|^{1/\alpha_F} \quad (74)$$

To see this, note that first we have $\|Q^*(x) - Q^*(\tilde{x})\|_F < \delta_F$ since otherwise Assumption 4 then implies that

$$F(x, Q^*(\tilde{x})) - F(x, Q^*(x)) > \frac{\delta_F^{\alpha_F}}{\gamma_2(L, \sqrt{d}M_L)} \stackrel{(i)}{\geq} CM_L \|x - \tilde{x}\|$$

Here (i) results from (73). This is a contradiction with (72). Next applying Assumption 4 with $\delta = \|Q^*(x) - Q^*(\tilde{x})\|_F$ and $\Delta = \sqrt{d}M_L$ gives

$$F(x, Q^*(\tilde{x})) - F(x, Q^*(x)) \geq \frac{\|Q^*(x) - Q^*(\tilde{x})\|_F^{\alpha_F}}{\gamma_2(L, \sqrt{d}M_L)}$$

which combined with (71) implies (74).

3. x, \tilde{x} far apart: for any $x, \tilde{x} \in B_\mu(L)$ such that $\|x - \tilde{x}\| > t_0$, let $K := \lceil \|x - \tilde{x}\| / t_0 \rceil$ and

$$x_k := \left(1 - \frac{k}{K}\right)x + \frac{k}{K}\tilde{x}, \quad k = 0, \dots, K$$

Then $\|x_{k+1} - x_k\| \leq t_0$, and one can apply (74) to see that

$$\begin{aligned} \|Q^*(x) - Q^*(\tilde{x})\| &\leq \sum_{k=0}^{K-1} \|Q^*(x_{k+1}) - Q^*(x_k)\| \\ &\leq \sum_{k=0}^{K-1} \left(CM_L \gamma_2(L, \sqrt{d}M_L)\right)^{1/\alpha_F} \cdot \|x_{k+1} - x_k\|^{1/\alpha_F} \\ &\leq \sum_{k=0}^{K-1} \left(CM_L \gamma_2(L, \sqrt{d}M_L)\right)^{1/\alpha_F} \cdot t_0^{1/\alpha_F} \\ &= K \delta_F \\ &\leq 2 \frac{\|x - \tilde{x}\|}{t_0} \delta_F \\ &\lesssim M_L \gamma_2(L, \sqrt{d}M_L) \cdot \|x - \tilde{x}\| \end{aligned} \quad (75)$$

Here the last line follows since δ_F is a constant defined in Assumption 4.

Finally, combine (74) and (75) to see that for any $x, \tilde{x} \in B_\mu(L)$, one has

$$\|Q^*(x) - Q^*(\tilde{x})\| \leq C_H \cdot \max \left\{ (M_L \gamma_2(L, M_L))^{1/\alpha_F} \|x - \tilde{x}\|^{1/\alpha_F}, M_L \gamma_2(L, M_L) \|x - \tilde{x}\| \right\}$$

for some constant C_H independent of n (but depends on d). \blacksquare

Appendix C. Proof of Theorem 7 and Corollary 10

C.1 Proof outline for uniform convergence

We start by introducing some notations. Define

$$\begin{aligned} \hat{F}_\rho(x, S) &= \frac{1}{n} \sum_{i=1}^n w_{n,\rho}(x, X_i) W^2(S, Q_i) \\ \hat{A}_\rho &:= \frac{1}{n} \sum_{i=1}^n w_{n,\rho}(x, X_i) \left(T_{Q^*(x)}^{Q_i} - I_d \right) \\ \hat{\Phi}_\rho(x) &:= \frac{1}{n} \sum_{i=1}^n w_{n,\rho}(x, X_i) dT_{Q^*(x)}^{Q_i} \\ \hat{\Psi}_\rho(x, S) &:= \frac{1}{n} \sum_{i=1}^n w_{n,\rho}(x, X_i) d^2 T_S^{Q_i} \end{aligned} \tag{76}$$

and

$$\begin{aligned} \tilde{F}_n(x, S) &:= \frac{1}{n} \sum_{i=1}^n w(x, X_i) W^2(S, Q_i) \\ \tilde{A}_n(x) &:= \frac{1}{n} \sum_{i=1}^n w(x, X_i) \left(T_{Q^*(x)}^{Q_i} - I_d \right) \\ \tilde{\Phi}_n(x) &:= \frac{1}{n} \sum_{i=1}^n w(x, X_i) dT_{Q^*(x)}^{Q_i} \\ \tilde{\Psi}_n(x, S) &:= \frac{1}{n} \sum_{i=1}^n w(x, X_i) d^2 T_S^{Q_i} \end{aligned} \tag{77}$$

For any $t \geq 1$, define the event $E_0(t)$ as

$$E_0(t) := \{ \|X_i - \mu\| \leq t, Q_i \in \mathcal{S}_d(M_t^{-1}, M_t), i \in [n] \}$$

where $M_t := \gamma_\Lambda(t)$ and $\gamma_\Lambda(\cdot)$ is defined in Assumption 2. Specifically, let $L_\tau = \sqrt{(1+\tau) \log n}$ for $\tau \geq 0$ and define

$$E_0 = E_0(C_{\psi_2} L_\tau), \tag{78}$$

Under Assumption 1 and 2, we have $\mathbb{P}(E_0^c) \leq 2n^{-(1+\tau)}$.

We now give an outline of the proof strategy for our uniform convergence theory, namely, Theorems 7 and Corollary 10.

- Lemma 36 establishes a tail probability bound for the largest eigenvalue of $\widehat{Q}_\rho(x)$.
- Lemma 37 provides uniform concentration guarantees for \widehat{F}_ρ .
- Lemma 38 demonstrates uniform consistency of $\widehat{Q}_\rho(x)$, albeit with a potentially slow rate;
- Lemma 40-42 derive upper bounds for quantities associated with $\widehat{A}_\rho(x)$, $\widehat{\Phi}_\rho(x)$ and $\widehat{\Psi}_\rho(x; S)$.
- Lemma 43 arrives at the non-asymptotic \sqrt{n} -rate uniform convergence in the Frobenius norm with high probability.
- Lemma 44 combines Lemma 36 and 43 to obtain a \sqrt{n} -rate moment bound for the uniform Frobenius error.
- Lemma 45 generalizes the preceding moment result to the Wasserstein distance.

Theorem 7 and Corollary 10 then follow immediately by combining Lemmas 36-45.

Lemma 36 *Suppose Assumption 1-2 hold. If $n \geq 3$ and $1 \leq L \leq \sqrt{n}$, then there exist constants $C_b, c_b, c_e > 0$ (independent of n) such that for $t_0 = C_b L$, the following hold:*

- **Case $\rho = 0$:** for all $t \geq t_0$, the tail probability satisfies

$$\mathbb{P} \left(\sup_{x \in B_\mu(L)} \widehat{\lambda}_1(x) \geq t^2 \right) \lesssim h_0(t) + \exp(-c_e n)$$

- **Case $\rho = n^{-1}$:** for all $t \geq t_0$, the tail probability satisfies

$$\mathbb{P} \left(\sup_{x \in B_\mu(L)} \widehat{\lambda}_1(x) \geq t^2 \right) \lesssim \begin{cases} h_0(t) + \exp(-c_e n), & t \in [t_0, nt_0] \\ h_0(t) + \exp(-c_e n) \wedge h_0(t/n), & t \in (nt_0, +\infty) \end{cases}$$

where in the above two cases, $h_0 : \mathbb{R}_+ \rightarrow \mathbb{R}_+$ is defined as

$$h_0(t) := \exp \left(-\sqrt{\frac{t}{c_b L}} \right) + n \cdot \exp \left[-\left(\frac{t}{c_b L} \right)^{2/(C_\Lambda + 2)} \right]$$

As a result, denote

$$\widetilde{E}_0 := \left\{ \sup_{x \in B_\mu(L)} \lambda_{\max} \left(\widehat{Q}_\rho(x) \right) < \widetilde{M}_L \right\} \quad (79)$$

$$\text{where } \widetilde{M}_L := (c_b \vee C_b)^2 L^2 L_\tau^{4+(4\vee 2C_1)}$$

then

$$\mathbb{P}(\widetilde{E}_0^c) \leq C_{\lambda, \tau} n^{-\tau} \quad (80)$$

for some constant $C_{\lambda, \tau}$ independent of n .

Proof See Appendix C.2. ■

Denote $F(x, S) = \mathbb{E}\tilde{F}_n(x, S)$.

Lemma 37 *Instate the notations and assumptions in Theorem 7 and Lemma 36. Then for $\rho \in \{0, n^{-1}\}$ and any $\tau > 0$,*

$$\sup_{\substack{x \in B_\mu(L) \\ S \in \mathcal{S}_d(0, \tilde{M}_L)}} \left| \hat{F}_\rho(x, S) - F(x, S) \right| \leq C_{F,\tau} \frac{L\tilde{M}_L^3}{\sqrt{n}}$$

with probability greater than $1 - c_{F,\tau}n^{-\tau}$ for constants $C_{F,\tau}, c_{F,\tau}$ independent of n .

Proof See Appendix C.3. ■

Lemma 38 *Instate the notations and assumptions in Theorem 7 and Lemma 37. Let $1 \leq L = O(\sqrt{\log n})$. Then for $\rho \in \{0, n^{-1}\}$ and any $\tau > 0$, there exists an event $\tilde{E}_1 \subset \tilde{E}_0$ with probability greater than $1 - c_{\delta,\tau}n^{-\tau}$ under which*

$$\sup_{x \in B_\mu(L)} \left\| \hat{Q}_\rho(x) - Q^*(x) \right\|_F \leq C_{\delta,\tau} \frac{\text{polylog}(n)}{n^{1/(2\alpha_F)}} \quad (81)$$

for constants $C_{\delta,\tau}, c_{\delta,\tau} > 0$ independent of n . As a result, under \tilde{E}_1 ,

$$\hat{Q}_\rho(x) \in \mathcal{S}_d((C_{\text{slow}}M_L)^{-1}, C_{\text{slow}}M_L), \quad \forall x \in B_\mu(L) \quad (82)$$

for some constant $C_{\text{slow}} > 0$ independent of n .

Proof See Appendix C.4. ■

Remark 39 *We remind readers that in Lemma 38, $\alpha_F \geq 1$ is defined in Assumption 4. Therefore, Lemma 38 might only lead to a uniform convergence rate slower than \sqrt{n} . For example, it is shown in Lemma 34 that $\alpha_F = 2$ when X and Q are independent, which results by (81) in a uniform convergence rate of $n^{1/4}$. Such a slow rate is not enough to derive the asymptotic null distribution of our test statistic in Theorem 16. Therefore, we apply a finer analysis in the following part of the proof to further boost the uniform convergence rate to $n^{1/2}$.*

Lemma 40 *Instate the notations and assumptions in Theorem 7 and Lemma 38. Then there exists an event $E_{2,1} \subset \tilde{E}_1$ with probability greater than $1 - c_{\tau,1}n^{-\tau}$ under which*

$$\sup_{x \in B_\mu(L)} \left\| \tilde{A}_n(x) \right\|_F \leq C_{\tau,1} \frac{\text{polylog}(n)}{\sqrt{n}} \quad (83)$$

$$\sup_{x \in B_\mu(L)} \left\| \tilde{\Phi}_n(x) - \mathbb{E}\tilde{\Phi}_n(x) \right\| \leq C_{\tau,1} \frac{\text{polylog}(n)}{\sqrt{n}} \quad (84)$$

$$\sup_{\substack{x \in B_\mu(L) \\ S \in \mathcal{S}_d((C_{\text{slow}} M_L)^{-1}, C_{\text{slow}} M_L)}} \left\| \tilde{\Psi}_n(x; S) - \mathbb{E} \tilde{\Psi}_n(x; S) \right\| \leq C_{\tau,1} \frac{\text{polylog}(n)}{\sqrt{n}} \quad (85)$$

$$\sup_{\substack{x \in B_\mu(L) \\ S \in \mathcal{S}_d((C_{\text{slow}} M_L)^{-1}, C_{\text{slow}} M_L)}} \left\| \mathbb{E} \tilde{\Psi}_n(x; S) \right\| \leq C_{\tau,1} \text{polylog}(n) \quad (86)$$

for constants $C_{\tau,1}, c_{\tau,1}$ independent of n .

Proof See Appendix C.5. ■

Lemma 41 *Instate the notations and assumptions in Theorem 7 and Lemma 38. Then there exists event $E_{2,2} \subset \tilde{E}_1$ with probability greater than $1 - c_{\tau,2} n^{-\tau}$ under which*

$$\sup_{x \in B_\mu(L)} \left\| \hat{A}_\rho(x) - \tilde{A}_n(x) \right\|_{\text{F}} \leq C_{\tau,2} \frac{\text{polylog}(n)}{\sqrt{n}} \quad (87)$$

$$\sup_{x \in B_\mu(L)} \left\| \hat{\Phi}_\rho(x) - \tilde{\Phi}_n(x) \right\| \leq C_{\tau,2} \frac{\text{polylog}(n)}{\sqrt{n}} \quad (88)$$

$$\sup_{\substack{x \in B_\mu(L) \\ S \in \mathcal{S}_d((C_{\text{slow}} M_L)^{-1}, C_{\text{slow}} M_L)}} \left\| \hat{\Psi}_\rho(x; S) - \tilde{\Psi}_n(x; S) \right\| \leq C_{\tau,2} \frac{\text{polylog}(n)}{\sqrt{n}} \quad (89)$$

for constants $C_{\tau,2}, c_{\tau,2}$ independent of n .

Proof See Appendix C.6. ■

With Lemma 40 and 41 in hand, we define event $\tilde{E}_2 := E_{2,1} \cap E_{2,2}$ under which $\hat{A}_\rho(x)$, $\hat{\Psi}_\rho(x; S)$ and $\hat{\Phi}_\rho^{-1}(x)$ are uniformly bounded. This is summarized in Lemma 42.

Lemma 42 *Instate the notations and assumptions in Theorem 7, Lemma 38, 40 and 41. Then $\mathbb{P}(\tilde{E}_2^c) \leq \tilde{c}_{\tau,2} n^{-\tau}$ and under \tilde{E}_2 the following hold:*

$$\sup_{x \in B_\mu(L)} \left\| \hat{A}_\rho(x) \right\|_{\text{F}} \leq \tilde{C}_{\tau,2} \frac{\text{polylog}(n)}{\sqrt{n}} \quad (90)$$

$$\sup_{\substack{x \in B_\mu(L) \\ S \in \mathcal{S}_d((C_{\text{slow}} M_L)^{-1}, C_{\text{slow}} M_L)}} \left\| \hat{\Psi}_\rho(x; S) \right\| \leq \tilde{C}_{\tau,2} \text{polylog}(n) \quad (91)$$

$$\inf_{x \in B_\mu(L)} \lambda_{\min} \left(-\hat{\Phi}_\rho(x) \right) \geq \frac{1}{\tilde{C}_{\tau,2} \text{polylog}(n)} \quad (92)$$

As a result, under \tilde{E}_2 the operator $-\hat{\Phi}_\rho$ is invertible and

$$\sup_{x \in B_\mu(L)} \left\| -\hat{\Phi}_\rho^{-1}(x) \right\| \leq \tilde{C}_{\tau,2} \text{polylog}(n) \quad (93)$$

Proof See Appendix C.7. ■

Lemma 43 *Instate the notations and assumptions in Theorem 7, Lemma 36 - 41. then under \tilde{E}_2 ,*

$$\sup_{x \in B_\mu(L)} \left\| \hat{Q}_\rho(x) - Q^*(x) \right\|_F \leq C_{\text{fast}, \tau} \frac{\text{polylog}(n)}{\sqrt{n}}$$

for constant $C_{\text{fast}, \tau}$ independent of n .

Proof See Appendix C.8. ■

Lemma 44 *Let $\rho = 1/n$. Instate the notations and assumptions in Theorem 7, Lemma 36 - 43. Then*

$$\mathbb{E} \sup_{x \in B_\mu(L)} \left\| \hat{Q}_\rho(x) - Q^*(x) \right\|_F \leq C_{E, F} \frac{\text{polylog}(n)}{\sqrt{n}}$$

for some constant $C_{E, F}$ independent of n .

Proof See Appendix C.9. ■

Lemma 45 *Let $\rho = 1/n$. Instate the notations and assumptions in Theorem 7, Lemma 36 - 43. Then*

$$\mathbb{E} \sup_{x \in B_\mu(L)} W \left(\hat{Q}_\rho(x), Q^*(x) \right) \leq C_{E, W} \frac{\text{polylog}(n)}{\sqrt{n}}$$

for some constant $C_{E, W}$ independent of n .

Proof See Appendix C.10. ■

C.2 Proof of Lemma 36

Without loss of generality, assume $\mu = 0$, $\Sigma = I_p$, $C_{\psi_2} = 1$ and $c_\Lambda = 1$ (defined in Assumption 2). Denote $\hat{\lambda}_1(x) = \lambda_{\max}(\hat{Q}_\rho(x))$.

By the optimality condition for $\hat{Q}_\rho(x)$, one has

$$0 = \frac{1}{n} \sum_{i=1}^n w_{n, \rho}(x, X_i) \left(T_{\hat{Q}_\rho(x)}^{Q_i} - I_d \right)$$

Note that $n^{-1} \sum_{i=1}^n w_{n, \rho}(x, X_i) = 1$ for any x , hence one has

$$I_d = \frac{1}{n} \sum_{i=1}^n w_{n, \rho}(x, X_i) T_{\hat{Q}_\rho(x)}^{Q_i}$$

Recall that $T_{\hat{Q}_\rho(x)}^Q = \hat{Q}_\rho(x)^{-1/2} \left(\hat{Q}_\rho(x)^{1/2} Q \hat{Q}_\rho(x)^{1/2} \right)^{1/2} \hat{Q}_\rho(x)^{-1/2}$, multiplying both sides of the equation with $\hat{Q}_\rho(x)$ then gives

$$\hat{Q}_\rho(x) = \frac{1}{n} \sum_{i=1}^n w_{n,\rho}(x, X_i) \left(\hat{Q}_\rho(x)^{1/2} Q_i \hat{Q}_\rho(x)^{1/2} \right)^{1/2}$$

Taking the largest eigenvalue on both sides, one has

$$\begin{aligned} \hat{\lambda}_1(x) &\leq \frac{1}{n} \sum_{i=1}^n |w_{n,\rho}(x, X_i)| \cdot \left\| \left(\hat{Q}_\rho(x)^{1/2} Q_i \hat{Q}_\rho(x)^{1/2} \right)^{1/2} \right\|_{\text{op}} \\ &\leq \frac{1}{n} \sum_{i=1}^n |w_{n,\rho}(x, X_i)| \cdot \lambda_{\max}(Q_i)^{1/2} \hat{\lambda}_1^{1/2}(x) \end{aligned}$$

which implies

$$\begin{aligned} \hat{\lambda}_1(x) &\leq \left[\frac{1}{n} \sum_{i=1}^n \left| 1 + (x - \bar{X})^\top \hat{\Sigma}_\rho^{-1}(X_i - \bar{X}) \right| \cdot \lambda_{\max}(Q_i)^{1/2} \right]^2 \\ &\leq \left[\frac{1}{n} \sum_{i=1}^n \left(1 + \|x - \bar{X}\| \cdot \left\| \hat{\Sigma}_\rho^{-1} \right\|_{\text{op}} \cdot \|X_i - \bar{X}\| \right) \lambda_{\max}(Q_i)^{1/2} \right]^2 \end{aligned}$$

Hence

$$\begin{aligned} &\mathbb{P} \left(\sup_{x \in B_\mu(L)} \hat{\lambda}_1(x) \geq t^2 \right) \\ &\leq \mathbb{P} \left(\left\{ \left\| \hat{\Sigma}_\rho^{-1} \right\|_{\text{op}} \leq 2 \right\} \cap \left\{ \sup_{x \in B_\mu(L)} \frac{1}{n} \sum_{i=1}^n (1 + \|x - \bar{X}\| \cdot \|X_i - \bar{X}\|) \lambda_{\max}(Q_i)^{1/2} \geq t/2 \right\} \right) \\ &\quad + \mathbb{P} \left(\left\{ 2 < \left\| \hat{\Sigma}_\rho^{-1} \right\|_{\text{op}} \leq n \right\} \cap \left\{ \sup_{x \in B_\mu(L)} \frac{1}{n} \sum_{i=1}^n (1 + \|x - \bar{X}\| \cdot \|X_i - \bar{X}\|) \lambda_{\max}(Q_i)^{1/2} \geq t/n \right\} \right) \\ &\quad + \mathbb{P} \left(\left\| \hat{\Sigma}_\rho^{-1} \right\|_{\text{op}} > n \right) \\ &\leq \underbrace{\mathbb{P} \left(\sup_{x \in B_\mu(L)} \frac{1}{n} \sum_{i=1}^n (1 + \|x - \bar{X}\| \cdot \|X_i - \bar{X}\|) \lambda_{\max}(Q_i)^{1/2} \geq t/2 \right)}_{\text{(I)}} \\ &\quad + \underbrace{\mathbb{P} \left(2 < \left\| \hat{\Sigma}_\rho^{-1} \right\|_{\text{op}} \leq n \right) \wedge \mathbb{P} \left(\sup_{x \in B_\mu(L)} \frac{1}{n} \sum_{i=1}^n (1 + \|x - \bar{X}\| \cdot \|X_i - \bar{X}\|) \lambda_{\max}(Q_i)^{1/2} \geq t/n \right)}_{\text{(II)}} \\ &\quad + \underbrace{\mathbb{P} \left(\left\| \hat{\Sigma}_\rho^{-1} \right\|_{\text{op}} > n \right)}_{\text{(III)}} \end{aligned}$$

Claim 1 Suppose Assumption 1-2 hold with $\mu = 0$, $\Sigma = I_p$, $C_{\psi_2} = 1$ and $c_\Lambda = 1$. If $n \geq 3$, $1 \leq L \leq \sqrt{n}$. then for $\rho \in \{0, n^{-1}\}$ and any $t \geq 6L$,

$$\begin{aligned} \mathbb{P} \left\{ \left\| \widehat{\Sigma}_\rho^{-1} \right\|_{\text{op}} > 2 \right\} &\lesssim \exp(-c_e n) \\ \mathbb{P} \left\{ \sup_{x \in B_\mu(L)} \frac{1}{n} \sum_{i=1}^n (1 + \|x - \bar{X}\| \cdot \|X_i - \bar{X}\|) \lambda_{\max}(Q_i)^{1/2} \geq t \right\} &\lesssim h_0(t) \end{aligned}$$

where

$$h_0(t) = \exp \left(-\sqrt{\frac{t}{c_b L}} \right) + n \cdot \exp \left[-\left(\frac{t}{c_b L} \right)^{2/(C_\Lambda+2)} \right]$$

for some constants $c_b, c_e > 0$ independent of n .

See Appendix C.2.1 for the proof.

Combining the above results, one can obtain that

- If $\rho = 0$, then

$$\text{(II)} + \text{(III)} \leq \mathbb{P} \left(\left\| \widehat{\Sigma}_\rho^{-1} \right\|_{\text{op}} > 2 \right)$$

As a result, for any $t \geq 6L$, the tail probability satisfies

$$\mathbb{P} \left(\sup_{x \in B_\mu(L)} \widehat{\lambda}_1(x) \geq t^2 \right) \lesssim h_0(t) + \exp(-c_e n)$$

- If $\rho = n^{-1}$, then $\text{(III)} = 0$. As a results, the tail probability satisfies

$$\mathbb{P} \left(\sup_{x \in B_\mu(L)} \widehat{\lambda}_1(x) \geq t^2 \right) \lesssim \begin{cases} h_0(t) + \exp(-c_e n), & t \in [t_0, nt_0] \\ h_0(t) + \exp(-c_e n) \wedge h_0(t/n), & t \in (nt_0, +\infty) \end{cases}$$

with $t_0 = 6L$.

Finally, set

$$\tilde{t}_L = (C_b \vee c_b) L_\tau^{2+(2 \vee 2C_1)} L$$

Note that $\tilde{t}_L \geq t_0$, hence

$$\begin{aligned} \mathbb{P} \left(\sup_{x \in B_\mu(L)} \widehat{\lambda}_1(x) \geq \tilde{t}_L^2 \right) &\lesssim h_0(\tilde{t}_L) + \exp(-c_e n) \\ &\lesssim n^{-(1+\tau)} + \exp(-c_e n) \\ &\leq C_\tau n^{-(1+\tau)} \end{aligned}$$

for some constant C_τ independent of n (but depends on τ).

This finishes the proof of Lemma 36.

C.2.1 PROOF OF CLAIM 1

The first inequality follows by noticing that $\widehat{\Sigma}_\rho = n^{-1} \sum_{i=1}^n X_i X_i^\top - \bar{X} \bar{X}^\top + \rho I_p$ with $\rho \in \{0, 1/n\}$. Applying the triangle inequality, one can obtain that for $n \geq 3$,

$$\begin{aligned} \mathbb{P} \left\{ \left\| \widehat{\Sigma}_\rho^{-1} \right\|_{\text{op}} > 2 \right\} &\leq \mathbb{P} \left\{ \left\| \widehat{\Sigma}_\rho - I_p \right\|_{\text{op}} > 1/2 \right\} \\ &\leq \mathbb{P} \left(\left\| n^{-1} \sum_{i=1}^n X_i X_i^\top - I_p \right\|_{\text{op}} > 1/12 \right) + \mathbb{P} \left(\left\| \bar{X} \bar{X}^\top \right\|_{\text{op}} > 1/12 \right) \end{aligned}$$

Note that the last line holds for both $\rho \in \{0, n^{-1}\}$. Under the sub-Gaussian assumption in Assumption 1, one can obtain that

$$\begin{aligned} \mathbb{P} \left\{ \left\| n^{-1} \sum_{i=1}^n X_i X_i^\top - I_p \right\|_{\text{op}} > 1/12 \right\} &\lesssim \exp(-cn) \\ \mathbb{P} \left\{ \left\| \bar{X} \bar{X}^\top \right\|_{\text{op}} > 1/12 \right\} &\lesssim \exp(-cn) \end{aligned}$$

Now we move on to the second inequality. Note that

$$\begin{aligned} &\frac{1}{n} \sum_{i=1}^n (1 + \|x - \bar{X}\| \cdot \|X_i - \bar{X}\|) \lambda_{\max}(Q_i)^{1/2} \\ &\leq \frac{1}{n} \sum_{i=1}^n (1 + \|x\| \cdot \|X_i\|) \lambda_{\max}(Q_i)^{1/2} + \frac{1}{n} \sum_{i=1}^n \|\bar{X}\| \cdot \|X_i\| \lambda_{\max}(Q_i)^{1/2} \\ &\quad + \frac{1}{n} \sum_{i=1}^n \|x\| \cdot \|\bar{X}\| \lambda_{\max}(Q_i)^{1/2} + \frac{1}{n} \sum_{i=1}^n \|\bar{X}\|^2 \lambda_{\max}(Q_i)^{1/2} \\ &= \left(1 + \|x\| \|\bar{X}\| + \|\bar{X}\|^2\right) \cdot \frac{1}{n} \sum_{i=1}^n \lambda_{\max}(Q_i)^{1/2} + (\|\bar{X}\| + \|x\|) \cdot \frac{1}{n} \sum_{i=1}^n \|X_i\| \lambda_{\max}(Q_i)^{1/2} \end{aligned}$$

Therefore, for any $L \geq 1$ and $t > 0$, one has

$$\begin{aligned} &\mathbb{P} \left\{ \sup_{x \in B_\mu(L)} \frac{1}{n} \sum_{i=1}^n (1 + \|x - \bar{X}\| \cdot \|X_i - \bar{X}\|) \lambda_{\max}(Q_i)^{1/2} \geq t \right\} \\ &\leq \mathbb{P} \left\{ \left(1 + L \|\bar{X}\| + \|\bar{X}\|^2\right) \cdot \frac{1}{n} \sum_{i=1}^n \lambda_{\max}(Q_i)^{1/2} \geq \frac{1}{1+L} t \right\} \\ &\quad + \mathbb{P} \left\{ (\|\bar{X}\| + \|x\|) \cdot \frac{1}{n} \sum_{i=1}^n \|X_i\| \lambda_{\max}(Q_i)^{1/2} \geq \frac{L}{1+L} t \right\} \\ &\leq \underbrace{\mathbb{P} \left\{ \left(1 + L \|\bar{X}\| + \|\bar{X}\|^2\right) \cdot \frac{1}{n} \sum_{i=1}^n \lambda_{\max}(Q_i)^{1/2} \geq \frac{t}{2L} \right\}}_{(\mathbf{A})} \end{aligned}$$

$$+ \underbrace{\mathbb{P} \left\{ (\|\bar{X}\| + L) \cdot \frac{1}{n} \sum_{i=1}^n \|X_i\| \lambda_{\max}(Q_i)^{1/2} \geq \frac{t}{2} \right\}}_{(\mathbf{B})}$$

Analysis for (A): for any $1 \leq L \leq \sqrt{n}$, one can obtain that

$$\begin{aligned} (\mathbf{A}) &\leq \mathbb{P} \left\{ \frac{1}{n} \sum_{i=1}^n \lambda_{\max}(Q_i)^{1/2} \geq \frac{1}{1 + L/\sqrt{n} + 1/n} \cdot \frac{t}{2L} \right\} \\ &\quad + \mathbb{P} \left\{ L \|\bar{X}\| \frac{1}{n} \sum_{i=1}^n \lambda_{\max}(Q_i)^{1/2} \geq \frac{L/\sqrt{n}}{1 + L/\sqrt{n} + 1/n} \cdot \frac{t}{2L} \right\} \\ &\quad + \mathbb{P} \left\{ \|\bar{X}\|^2 \frac{1}{n} \sum_{i=1}^n \lambda_{\max}(Q_i)^{1/2} \geq \frac{1/n}{1 + L/\sqrt{n} + 1/n} \cdot \frac{t}{2L} \right\} \\ &\leq \underbrace{\mathbb{P} \left\{ \frac{1}{n} \sum_{i=1}^n \lambda_{\max}(Q_i)^{1/2} \geq \frac{1}{3} \cdot \frac{t}{2L} \right\}}_{\mathbf{A}_1} + \underbrace{\mathbb{P} \left\{ \|\bar{X}\| \frac{1}{n} \sum_{i=1}^n \lambda_{\max}(Q_i)^{1/2} \geq \frac{1/\sqrt{n}}{3} \cdot \frac{t}{2L} \right\}}_{\mathbf{A}_2} \\ &\quad + \underbrace{\mathbb{P} \left\{ \|\bar{X}\|^2 \frac{1}{n} \sum_{i=1}^n \lambda_{\max}(Q_i)^{1/2} \geq \frac{1/n}{3} \cdot \frac{t}{2L} \right\}}_{\mathbf{A}_2} \end{aligned}$$

• \mathbf{A}_1 :

$$\mathbf{A}_1 \leq \mathbb{P} \left\{ \|\bar{X}\| \geq \frac{1}{\sqrt{n}} \cdot \sqrt{\frac{t}{6L}} \right\} + \mathbb{P} \left\{ \frac{1}{n} \sum_{i=1}^n \lambda_{\max}(Q_i)^{1/2} \geq \sqrt{\frac{t}{6L}} \right\}$$

• \mathbf{A}_2 :

$$\mathbf{A}_2 \leq \mathbb{P} \left\{ \|\bar{X}\|^2 \geq \frac{1}{n} \cdot \sqrt{\frac{t}{6L}} \right\} + \mathbb{P} \left\{ \frac{1}{n} \sum_{i=1}^n \lambda_{\max}(Q_i)^{1/2} \geq \sqrt{\frac{t}{6L}} \right\}$$

Hence for $t \geq 6L$,

$$\begin{aligned} (\mathbf{A}) &\lesssim \mathbb{P} \left\{ \frac{1}{n} \sum_{i=1}^n \lambda_{\max}(Q_i)^{1/2} \geq \sqrt{\frac{t}{6L}} \right\} + \mathbb{P} \left\{ \|\bar{X}\|^2 \geq \frac{1}{n} \cdot \sqrt{\frac{t}{6L}} \right\} \\ &\lesssim \mathbb{P} \left\{ \frac{1}{n} \sum_{i=1}^n \lambda_{\max}(Q_i)^{1/2} \geq \sqrt{\frac{t}{6L}} \right\} + \exp \left(-c \sqrt{\frac{t}{6L}} \right) \end{aligned}$$

Analysis for (B):

$$(\mathbf{B}) \leq \mathbb{P} \left\{ L \cdot \frac{1}{n} \sum_{i=1}^n \|X_i\| \lambda_{\max}(Q_i)^{1/2} \geq \frac{\sqrt{n}L}{\sqrt{n}L + 1} \frac{t}{2} \right\}$$

$$\begin{aligned}
 & + \mathbb{P} \left\{ \|\bar{X}\| \cdot \frac{1}{n} \sum_{i=1}^n \|X_i\| \lambda_{\max}(Q_i)^{1/2} \geq \frac{1}{\sqrt{nL}+1} \frac{t}{2} \right\} \\
 & \leq \mathbb{P} \left\{ \frac{1}{n} \sum_{i=1}^n \|X_i\| \lambda_{\max}(Q_i)^{1/2} \geq \frac{t}{4L} \right\} + \mathbb{P} \left\{ \|\bar{X}\| \cdot \frac{1}{n} \sum_{i=1}^n \|X_i\| \lambda_{\max}(Q_i)^{1/2} \geq \frac{t}{4\sqrt{nL}} \right\} \\
 & \leq \mathbb{P} \left\{ \frac{1}{n} \sum_{i=1}^n \|X_i\| \lambda_{\max}(Q_i)^{1/2} \geq \frac{t}{4L} \right\} \\
 & + \mathbb{P} \left\{ \|\bar{X}\| \geq \sqrt{\frac{t}{4nL}} \right\} + \mathbb{P} \left\{ \frac{1}{n} \sum_{i=1}^n \|X_i\| \lambda_{\max}(Q_i)^{1/2} \geq \sqrt{\frac{t}{4L}} \right\}
 \end{aligned}$$

Hence for $t \geq 4L$,

$$\begin{aligned}
 (\mathbf{B}) & \lesssim \mathbb{P} \left\{ \|\bar{X}\| \geq \sqrt{\frac{t}{4nL}} \right\} + \mathbb{P} \left\{ \frac{1}{n} \sum_{i=1}^n \|X_i\| \lambda_{\max}(Q_i)^{1/2} \geq \sqrt{\frac{t}{4L}} \right\} \\
 & \lesssim \exp(-ct/L) + \mathbb{P} \left\{ \frac{1}{n} \sum_{i=1}^n \|X_i\| \lambda_{\max}(Q_i)^{1/2} \geq \sqrt{\frac{t}{4L}} \right\}
 \end{aligned}$$

Combining results for **(A)** and **(B)**, one can obtain that for $t \geq 6L$,

$$\begin{aligned}
 & (\mathbf{A}) + (\mathbf{B}) \\
 & \lesssim \exp \left(-c\sqrt{\frac{t}{6L}} \right) + \exp(-ct/L) + \mathbb{P} \left\{ \frac{1}{n} \sum_{i=1}^n \lambda_{\max}(Q_i)^{1/2} \geq \sqrt{\frac{t}{6L}} \right\} \\
 & + \mathbb{P} \left\{ \frac{1}{n} \sum_{i=1}^n \|X_i\| \lambda_{\max}(Q_i)^{1/2} \geq \sqrt{\frac{t}{4L}} \right\} \\
 & \lesssim \exp \left(-c\sqrt{\frac{t}{6L}} \right) + \mathbb{P} \left\{ \frac{1}{n} \sum_{i=1}^n \lambda_{\max}^{1/2}(Q_i) \geq \sqrt{\frac{t}{6L}} \right\} + \mathbb{P} \left\{ \frac{1}{n} \sum_{i=1}^n \|X_i\| \lambda_{\max}^{1/2}(Q_i) \geq \sqrt{\frac{t}{4L}} \right\}
 \end{aligned}$$

Note that for any $s \geq 1$,

$$\begin{aligned}
 \mathbb{P} \left\{ \frac{1}{n} \sum_{i=1}^n \lambda_{\max}(Q_i)^{1/2} \geq s \right\} & \leq n \cdot \mathbb{P} \left\{ \lambda_{\max}(Q_i)^{1/2} \geq s \right\} \\
 & \leq n \cdot \mathbb{P} \left\{ (1 \vee \|X_i\|)^{C_\Lambda/2} \geq s \right\} \\
 & = n \cdot \mathbb{P} \left\{ \|X_i\| \geq s^{2/C_\Lambda} \right\} \\
 & \lesssim n \cdot \exp \left(-cs^{4/C_\Lambda} \right)
 \end{aligned}$$

and

$$\mathbb{P} \left\{ \frac{1}{n} \sum_{i=1}^n \|X_i\| \lambda_{\max}(Q_i)^{1/2} \geq s \right\} \leq n \cdot \mathbb{P} \left\{ \|X_i\| \lambda_{\max}(Q_i)^{1/2} \geq s \right\}$$

$$\begin{aligned}
 &\leq n \cdot \mathbb{P} \left\{ \|X_i\| (1 \vee \|X_i\|)^{C_\Lambda/2} \geq s \right\} \\
 &= n \cdot \mathbb{P} \left\{ \|X_i\|^{1+C_\Lambda/2} \geq s \right\} \\
 &\lesssim n \cdot \exp \left(-cs^{4/(C_\Lambda+2)} \right)
 \end{aligned}$$

Therefore, for any $t \geq 6L$,

$$\begin{aligned}
 (\mathbf{A}) + (\mathbf{B}) &\lesssim \exp \left(-c\sqrt{\frac{t}{L}} \right) + n \cdot \exp \left[-c(t/L)^{2/C_\Lambda} \right] + n \cdot \exp \left[-c(t/L)^{2/(C_\Lambda+2)} \right] \\
 &\lesssim \exp \left(-c\sqrt{\frac{t}{L}} \right) + n \cdot \exp \left[-c(t/L)^{2/(C_\Lambda+2)} \right]
 \end{aligned}$$

The proof of Claim 1 is complete by adjusting constant c here.

C.3 Proof of Lemma 37

Without loss of generality, assume $\mu = 0$, $\Sigma = I_p$, $C_{\psi_2} = 1$. Recall notation $\mathcal{S}_d(a, b) := \{A \in \mathcal{S}_d : aI_d \prec A \prec bI_d\}$. Denote parameter $\theta = (x, S)$ and parameter space $\Theta = B_\mu(L) \times \mathcal{S}_d(0, \widetilde{M}_L)$. Let $F(\theta)$ denote $F(x, S)$ and similarly for other functions of (x, S) . By triangle inequality, one can obtain that

$$\sup_{\theta \in \Theta} \left| \widehat{F}_\rho(\theta) - F(\theta) \right| \leq \underbrace{\sup_{\theta \in \Theta} \left| \widehat{F}_\rho(\theta) - \widetilde{F}_n(\theta) \right|}_{\zeta_1} + \underbrace{\sup_{\theta \in \Theta} \left| \widetilde{F}_n(\theta) - F(\theta) \right|}_{\zeta_2} \quad (94)$$

For any $t_1, t_2 > 0$, define $C_{\text{sup}}(t_1, t_2) := \sup \{W^2(Q, S) : Q \in \mathcal{S}_d(0, t_1), S \in \mathcal{S}_d(0, t_2)\}$. By Lemma 21, one has

$$C_{\text{sup}}(t_1, t_2) \lesssim t_1 + t_2 \quad (95)$$

Here dimension d is viewed as a fixed constant and absorbed into \lesssim .

- ζ_1 : under event E_0 (defined in (78)), one has

$$\begin{aligned}
 &- \|X_i - \mu\| \leq L_\tau = \sqrt{(1 + \tau) \log n} \text{ for } i = 1, \dots, n. \\
 &- W^2(Q_i, S) \lesssim M_{L_\tau} \vee \widetilde{M}_L \text{ for any } S \in \mathcal{S}_d(0, \widetilde{M}_L) \text{ and } i \in [n].
 \end{aligned}$$

Lemma 28 then implies that

$$\zeta_1 \lesssim \left(M_{L_\tau} \vee \widetilde{M}_L \right) \frac{LL_\tau^3}{\sqrt{n}} \quad (96)$$

with probability greater than $1 - O(n^{-(1+\tau)})$.

- ζ_2 can be bounded by truncation and the uniform concentration in Lemma 29. Let $\widetilde{L} \geq 1$ be a parameter to be specified later, and let $M_{\widetilde{L}} := \gamma_\Lambda(\widetilde{L})$. Then Assumption 2 implies that a.s.

$$X \in B_\mu(\widetilde{L}) \implies Q \in \mathcal{S}_d \left(M_{\widetilde{L}}^{-1}, M_{\widetilde{L}} \right) \quad (97)$$

Let \tilde{X} be a truncated form of X defined as

$$\tilde{X} := \tilde{X}(\tilde{L}) = X \mathbf{1}(\|X\| \leq \tilde{L}) \quad (98)$$

Let $\tilde{X}_{>} := X - \tilde{X}$. By definition, one has $X = \tilde{X} + \tilde{X}_{>}$, $\|\tilde{X}\| \leq \tilde{L}$ and

$$\tilde{X}_{>} = \begin{cases} 0, & \|X\| \leq \tilde{L} \\ X, & \|X\| > \tilde{L} \end{cases} \quad (99)$$

As a result, one can obtain the following decomposition.

$$\begin{aligned} & w(x, X)W^2(S, Q) \\ &= \left(1 + x^\top X\right) W^2(S, Q) \\ &= \left(1 + x^\top \tilde{X}\right) W^2(S, Q) + x^\top \tilde{X}_{>} W^2(S, Q) \\ &= \left(1 + x^\top \tilde{X}\right) W^2(S, Q) \mathbf{1}\left(Q \in \mathcal{S}\left(M_{\tilde{L}}^{-1}, M_{\tilde{L}}\right)\right) \\ &\quad + \left(1 + x^\top \tilde{X}\right) W^2(S, Q) \mathbf{1}\left(Q \notin \mathcal{S}\left(M_{\tilde{L}}^{-1}, M_{\tilde{L}}\right)\right) + x^\top \tilde{X}_{>} W^2(S, Q) \\ &\stackrel{(i)}{=} \left(1 + x^\top \tilde{X}\right) W^2(S, Q) \mathbf{1}\left(Q \in \mathcal{S}\left(M_{\tilde{L}}^{-1}, M_{\tilde{L}}\right)\right) \\ &\quad + W^2(S, Q) \mathbf{1}\left(Q \notin \mathcal{S}\left(M_{\tilde{L}}^{-1}, M_{\tilde{L}}\right)\right) + x^\top \tilde{X}_{>} W^2(S, Q) \end{aligned}$$

Here (i) follows from (97). Indeed, if $Q \notin \mathcal{S}\left(M_{\tilde{L}}^{-1}, M_{\tilde{L}}\right)$, then as a result of (97), one has $\|X\| > \tilde{L}$ which then implies $\tilde{X} = 0$.

As a consequence, one can then obtain following decomposition of $\tilde{F}_n(\theta) - F(\theta)$.

$$\begin{aligned} \tilde{F}_n(\theta) - F(\theta) &= \alpha_1(\theta) + \alpha_2(\theta) + \alpha_3(\theta), \quad \text{where} \\ \alpha_1(\theta) &:= \frac{1}{n} \sum_{i=1}^n \left(1 + x^\top \tilde{X}_i\right) W^2(S, Q_i) \mathbf{1}\left(Q_i \in \mathcal{S}\left(M_{\tilde{L}}^{-1}, M_{\tilde{L}}\right)\right) \\ &\quad - \mathbb{E} \left(1 + x^\top \tilde{X}\right) W^2(S, Q) \mathbf{1}\left(Q \in \mathcal{S}\left(M_{\tilde{L}}^{-1}, M_{\tilde{L}}\right)\right) \\ \alpha_2(\theta) &:= \frac{1}{n} \sum_{i=1}^n x^\top \tilde{X}_{i,>} W^2(S, Q_i) - \mathbb{E} x^\top \tilde{X}_{>} W^2(S, Q) \\ \alpha_3(\theta) &:= \frac{1}{n} \sum_{i=1}^n W^2(S, Q_i) \mathbf{1}\left(Q_i \notin \mathcal{S}\left(M_{\tilde{L}}^{-1}, M_{\tilde{L}}\right)\right) \\ &\quad - \mathbb{E} W^2(S, Q) \mathbf{1}\left(Q \notin \mathcal{S}\left(M_{\tilde{L}}^{-1}, M_{\tilde{L}}\right)\right) \end{aligned}$$

Then triangle inequality gives

$$\zeta_2 \leq \sup_{\theta \in \Theta} |\alpha_1(\theta)| + \sup_{\theta \in \Theta} |\alpha_2(\theta)| + \sup_{\theta \in \Theta} |\alpha_3(\theta)| \quad (100)$$

- $\alpha_2(\theta), \alpha_3(\theta)$: $\sup_{\theta \in \Theta} |\alpha_2(\theta)|$ and $\sup_{\theta \in \Theta} |\alpha_3(\theta)|$ can be upper bounded as in Claim 2. The proof is deferred to Appendix C.3.1.

Claim 2 *There exists constant \tilde{C} independent of n such that by taking $\tilde{L} = \tilde{C}C_{\psi_2}\sqrt{(1+\tau)\log n}$, the following inequality*

$$\sup_{\theta \in \Theta} |\alpha_2(\theta)| \vee \sup_{\theta \in \Theta} |\alpha_3(\theta)| = O\left(\frac{\tilde{M}_L \tilde{L} + \tilde{L}^{1+C_\Lambda}}{n^{1+\tau}} \cdot L\right)$$

holds with probability greater than $1 - O(n^{-\tau})$.

- $\alpha_1(\theta)$: Apply triangle inequality to see that

$$\sup_{\theta \in \Theta} |\alpha_1(\theta)| = \underbrace{\sup_{\theta \in \Theta} |\alpha_1(\theta)| - \mathbb{E} \sup_{\theta \in \Theta} |\alpha_1(\theta)|}_{b_1} + \underbrace{\mathbb{E} \sup_{\theta \in \Theta} |\alpha_1(\theta)|}_{b_2} \quad (101)$$

After truncation, $\alpha_1(\theta)$ is uniformly bounded. Hence b_1 can be upper bounded by exploiting the bounded difference property (Lemma 31). For b_2 , we first show that $W^2(Q, S)$ is only Hölder continuous in S which then implies Hölder (rather than Lipschitz) continuity of the corresponding sub-Gaussian norm. By generalizing Dudley's integral inequality via chaining in Lemma 29, we arrive at an upper bound for b_2 . The results are summarized below in Claim 3 whose proof is deferred to Appendix C.3.2.

Claim 3 *Instate the notations and assumptions in Lemma 38 and Claim 2.*

$$\sup_{\theta \in \Theta} |\alpha_1(\theta)| = O\left(\frac{L\tilde{M}_L^2}{\sqrt{n}} \sqrt{\log \tilde{M}_L}\right)$$

with probability at least $1 - O(n^{-(1+\tau)})$.

- combine Claim 2, Claim 3 and (100) to see that

$$\zeta_2 = O\left(\frac{L\tilde{M}_L^3}{\sqrt{n}}\right) \quad (102)$$

with probability at least $1 - O(n^{-\tau})$.

- Combining results for ζ_1 and ζ_2 , one has

$$\begin{aligned} \sup_{\theta \in \Theta} \left| \hat{F}_\rho(\theta) - F(\theta) \right| &\leq \zeta_1 + \zeta_2 \\ &= O\left(\frac{L\tilde{M}_L^3}{\sqrt{n}}\right) \end{aligned}$$

with probability greater than $1 - O(n^{-\tau})$.

The proof of Lemma 37 is now complete.

C.3.1 PROOF OF CLAIM 2

Without loss of generality, assume $\|X\|_{\psi_2} \leq \| \|X\| \|_{\psi_2} \leq 1$.

Proof for $\alpha_2(\theta)$: a crude upper bound suffices here. By triangle inequality, one has

$$\begin{aligned} & \sup_{\theta \in \Theta} |\alpha_2(\theta)| \\ & \leq \sup_{\theta \in \Theta} \left| \mathbb{E} x^\top \tilde{X}_> W^2(S, Q) \right| + \sup_{\theta \in \Theta} \left| \frac{1}{n} \sum_{i=1}^n x^\top \tilde{X}_{i,>} W^2(S, Q_i) \right| \\ & \leq \sup_{\theta \in \Theta} \|x\| \cdot \mathbb{E} \left\| \tilde{X}_> \right\| \cdot W^2(S, Q) + \sup_{\theta \in \Theta} \frac{1}{n} \sum_{i=1}^n \|x\| \cdot \left\| \tilde{X}_{i,>} \right\| \cdot W^2(S, Q_i) \end{aligned}$$

Therefore, for any $t > 0$, one has

$$\begin{aligned} \mathbb{P} \left\{ \sup_{\theta \in \Theta} |\alpha_2(\theta)| \geq t \right\} & \leq \underbrace{\mathbb{P} \left\{ \sup_{\theta \in \Theta} \|x\| \cdot \mathbb{E} \left\| \tilde{X}_> \right\| \cdot W^2(S, Q) \geq t \right\}}_{=:\mathbf{I}(t)} \\ & \quad + \underbrace{\mathbb{P} \left\{ \sup_{\theta \in \Theta} \frac{1}{n} \sum_{i=1}^n \|x\| \cdot \left\| \tilde{X}_{i,>} \right\| \cdot W^2(S, Q_i) > 0 \right\}}_{=:\mathbf{II}(t)} \end{aligned}$$

I(t): Lemma 21 and Assumption 2 imply that

$$\sup_{\theta \in \Theta} W^2(S, Q) \lesssim \tilde{M}_L + \gamma_\Lambda(\|X\|)$$

which further gives

$$\begin{aligned} (\mathbf{I}) & \leq \mathbb{P} \left\{ \mathbb{E} \left\| \tilde{X}_> \right\| \left(\tilde{M}_L + \|X\|^{C_\Lambda} \right) \geq t/(Lc) \right\} \\ & \leq \mathbb{P} \left\{ \mathbb{E} \left\| \tilde{X}_> \right\| \geq \frac{t}{2cL\tilde{M}_L} \right\} + \mathbb{P} \left\{ \mathbb{E} \left\| \tilde{X}_> \right\| \|X\|^{C_\Lambda} \geq t/(2Lc) \right\} \end{aligned}$$

By the definition of $\tilde{X}_>$, one can obtain that $\left\| \tilde{X}_> \right\| \|X\|^{C_\Lambda} = \left\| \tilde{X}_> \right\|^{1+C_\Lambda}$ and for any $s \geq 0$,

$$\mathbb{P} \left\{ \left\| \tilde{X}_> \right\| \geq \tilde{L} + s \right\} \leq 2 \exp(-\tilde{L}^2) \exp(-s^2)$$

Therefore, one can obtain that for any $\tilde{L} \geq 1$, one has

$$\begin{aligned} \mathbb{E} \left\| \tilde{X}_> \right\| & \leq 2 \exp(-\tilde{L}^2) \int_0^\infty (\tilde{L} + s) \exp(-s^2) ds \\ & \lesssim \exp(-\tilde{L}^2) \tilde{L} \end{aligned}$$

and

$$\mathbb{E} \left\| \tilde{X}_> \right\|^{1+C_\Lambda} \leq 2 \exp(-\tilde{L}^2) \int_0^\infty (\tilde{L} + s)^{1+C_\Lambda} \exp(-s^2) ds$$

$$\lesssim \exp(-\tilde{L}^2) \tilde{L}^{1+C_\Lambda}$$

Therefore, by taking $\tilde{L} = O(L_\tau)$, there exists constant $C_I > 0$ such that

$$\mathbf{I}(\mathbf{t}) = 0, \quad \forall t \geq C_I n^{-(1+\tau)} \cdot (\tilde{M}_L + M_{\tilde{L}}) L \tilde{L}$$

II: Note that $\|\tilde{X}_{i,>}\| > 0$ if and only if $\|X_i\| > \tilde{L}$. Hence

$$\begin{aligned} (\mathbf{II}) &\leq n \cdot \mathbb{P} \left\{ \|\|X_i\| > \tilde{L}\| \right\} \\ &\lesssim n \cdot \exp(-\tilde{L}^2) \\ &\lesssim n^{-\tau} \end{aligned}$$

Combining (I), (II), there exists a constant $\tilde{C}_2 > 0$ such that for any $s \geq 0$,

$$\mathbb{P} \left\{ \sup_{\theta \in \Theta} |\alpha_2(\theta)| \geq n^{-(1+\tau)} \tilde{C}_2 L \tilde{L} (\tilde{M}_L + \tilde{L}^{C_\Lambda}) \right\} \lesssim n^{-\tau}$$

Proof for $\alpha_3(\theta)$:

$$\begin{aligned} \alpha_3(\theta) &:= \frac{1}{n} \sum_{i=1}^n W^2(S, Q_i) \mathbf{1} \left(Q_i \notin \mathcal{S} \left(M_{\tilde{L}}^{-1}, M_{\tilde{L}} \right) \right) \\ &\quad - \mathbb{E} W^2(S, Q) \mathbf{1} \left(Q \notin \mathcal{S} \left(M_{\tilde{L}}^{-1}, M_{\tilde{L}} \right) \right) \end{aligned}$$

Following a similar argument to the one above, there exists constant $\tilde{C}_3 > 0$ such that

$$\mathbb{P} \left\{ \sup_{\theta \in \Theta} |\alpha_3(\theta)| \geq n^{-(1+\tau)} \tilde{C}_3 L (\tilde{M}_L + \tilde{L}^{C_\Lambda}) \right\} \lesssim n^{-\tau}$$

Finally, combining above results on $\alpha_2(\theta)$ and $\alpha_3(\theta)$ proves Claim 2.

C.3.2 PROOF OF CLAIM 3

Recall that we assume without loss of generality that $C_{\psi_2} = 1$, $\mu = 0$, $\Sigma = I_p$, $\tilde{L} = L_\tau$, $M_{\tilde{L}} = \gamma_\Lambda(\tilde{L})$ and \tilde{M}_L is defined in Lemma 36 as

$$\tilde{M}_L \asymp L^2 L_\tau^{4+(4 \vee 2C_1)} \gtrsim c_\Lambda (1 \vee \tilde{L})^{C_\Lambda} = M_{\tilde{L}}$$

Denote $Z = (X, Q)$ and define \tilde{X} as in (98). Define

$$f(Z; \theta) := \left(1 + (x - \mu)^\top \Sigma^{-1} (\tilde{X} - \mu) \right) W^2(S, Q) \mathbf{1} \left(Q \in \mathcal{S} \left(M_{\tilde{L}}^{-1}, M_{\tilde{L}} \right) \right)$$

Then one has $\alpha_1(\theta) = n^{-1} \sum_{i=1}^n f(Z_i; \theta) - \mathbb{E} f(Z; \theta)$.

Analysis of b_1 : due to truncation, f is uniformly bounded. To see this, note that recall the definition of Θ ,

$$\theta \in \Theta \iff x \in B_\mu(L), S \in \mathcal{S}_d(0, \widetilde{M}_L)$$

Then for any $\theta \in \Theta$, one has

$$\begin{aligned} \sup_{\theta \in \Theta} \sup_Z |f(Z; \theta)| &= \sup_{\theta \in \Theta} \sup_Z \left| (1 + x^\top \widetilde{X}) W^2(Q, S) \right| \mathbb{1} \left(Q \in \mathcal{S}_d \left(M_{\widetilde{L}}^{-1}, M_{\widetilde{L}} \right) \right) \\ &\leq \sup_{\theta \in \Theta} \sup_Z \left(1 + \|x - \mu\| \cdot \left\| \widetilde{X} - \mu \right\| \right) \cdot |W^2(Q, S)| \mathbb{1} \left(Q \in \mathcal{S}_d \left(M_{\widetilde{L}}^{-1}, M_{\widetilde{L}} \right) \right) \\ &\lesssim (1 + L\widetilde{L}) \cdot (M_{\widetilde{L}} + \widetilde{M}_L) \\ &\lesssim L\widetilde{L}\widetilde{M}_L \end{aligned} \tag{103}$$

By Lemma 31, one has $\|b_1(x)\|_{\psi_2} \lesssim \frac{L\widetilde{L}\widetilde{M}_L}{\sqrt{n}}$ which implies

$$b_1 = O \left(\frac{L\widetilde{L}\widetilde{M}_L L_\tau}{\sqrt{n}} \right) \tag{104}$$

with probability at least $1 - O(n^{-(1+\tau)})$ for some absolute constant $C > 0$.

Analysis of b_2 : to apply Lemma 29, we follow the steps below.

- first we define a metric d on Θ by

$$d(\theta, \tilde{\theta}) = \max \left\{ \|x - \tilde{x}\|_2, \|S - \tilde{S}\|_{\text{F}} \right\}$$

for any $\theta = (x, S)$ and $\tilde{\theta} = (\tilde{x}, \tilde{S})$. Since $\mathcal{S}_d(0, \widetilde{M}_L)$ is a subset of $\{S \in \mathbb{R}^{d \times d} : |S_{ij}| \leq \widetilde{M}_L\}$ (by noticing that for any $S \in \mathcal{S}_{\widetilde{M}_L}^+$, one has $0 \leq S_{ii} \leq \widetilde{M}_L$ and $|S_{ij}| \leq \sqrt{S_{ii}S_{jj}} \leq \widetilde{M}_L$), the diameter D and metric entropy of Θ can be upper bounded by

$$D \leq \max \left\{ \text{diam}(B_\mu(\widetilde{L})), \text{diam}(\mathcal{S}_d(\widetilde{M}_L^{-1}, \widetilde{M}_L)) \right\} \lesssim L \vee \widetilde{M}_L \stackrel{(i)}{\lesssim} \widetilde{M}_L \tag{105}$$

and

$$\begin{aligned} \log N(\epsilon; \Theta) &\leq \log \left(N(\epsilon; B_\mu(\widetilde{L})) \cdot N(\epsilon; \mathcal{S}_{\widetilde{M}_L}^+) \right) \\ &\lesssim \log^+ \left(\frac{L \vee \widetilde{M}_L}{\epsilon} \right) \\ &\stackrel{(ii)}{\lesssim} \log^+ \left(\frac{\widetilde{M}_L}{\epsilon} \right) \end{aligned} \tag{106}$$

Here both (i) and (ii) arise due to $\widetilde{M}_L \gtrsim L$.

- Next, let us consider the sub-Gaussian norm of $\alpha_1(\theta) - \alpha_1(\tilde{\theta})$ as (40) in Lemma 29. By (43) in Lemma 29, one has

$$\begin{aligned}
 \left\| \alpha_1(\theta) - \alpha_1(\tilde{\theta}) \right\|_{\psi_2} &\lesssim \frac{1}{\sqrt{n}} \left\| f(Z; \theta) - f(Z; \tilde{\theta}) \right\|_{\psi_2} \\
 &= \frac{1}{\sqrt{n}} \left\| f(Z; x, S) - f(Z; \tilde{x}, \tilde{S}) \right\|_{\psi_2} \\
 &\leq \frac{1}{\sqrt{n}} \underbrace{\left\| f(Z; x, S) - f(Z; x, \tilde{S}) \right\|_{\psi_2}}_{\beta_1} + \frac{1}{\sqrt{n}} \underbrace{\left\| f(Z; x, \tilde{S}) - f(Z; \tilde{x}, \tilde{S}) \right\|_{\psi_2}}_{\beta_2}
 \end{aligned} \tag{107}$$

- β_1 : Let $g(Q; S) := (Q^{1/2} S Q^{1/2})^{1/2}$. Then for any $Q \in \mathcal{S}_d(M_{\tilde{L}}^{-1}, M_{\tilde{L}})$,

$$\begin{aligned}
 \left\| g(Q; S) - g(Q; \tilde{S}) \right\|_F &= \left\| (Q^{1/2} S Q^{1/2})^{1/2} - (Q^{1/2} \tilde{S} Q^{1/2})^{1/2} \right\|_F \\
 &\lesssim \left\| Q^{1/2} S Q^{1/2} - Q^{1/2} \tilde{S} Q^{1/2} \right\|_F^{1/2} \\
 &\leq \left\| S - \tilde{S} \right\|_F^{1/2} \|Q\|_{\text{op}}^{1/2} \\
 &\leq M_{\tilde{L}}^{1/2} \left\| S - \tilde{S} \right\|_F^{1/2}
 \end{aligned}$$

Here in the second inequality, we exploit the following the Hölder continuity of matrix square root in Wihler (2009); Carlsson (2018).

$$\left\| A^{1/2} - B^{1/2} \right\|_F \leq C_d \|A - B\|_F^{1/2}$$

where C_d is a constant only depending on dimension d . Hence for any $\theta, \tilde{\theta} \in \Theta$,

$$\begin{aligned}
 \beta_1 &= \left\| f(X, Q; x, S) - f(X, Q; x, \tilde{S}) \right\|_{\psi_2} \\
 &\leq \left\| 1 + x^\top \tilde{X} \right\|_{\psi_2} \cdot \left\| \left(g(Q; S) - g(Q; \tilde{S}) \right) \cdot \mathbb{1} \left(Q \in \mathcal{S}_d \left(M_{\tilde{L}}^{-1}, M_{\tilde{L}} \right) \right) \right\|_F \\
 &\leq \left(1 + \|x\| \cdot \left\| \tilde{X} \right\|_2 \right) \cdot M_{\tilde{L}}^{1/2} \left\| S - \tilde{S} \right\|_F^{1/2} \\
 &\lesssim L \cdot M_{\tilde{L}}^{1/2} \left\| S - \tilde{S} \right\|_F^{1/2}
 \end{aligned} \tag{108}$$

- β_2 : for any $\theta, \tilde{\theta} \in \Theta$, one has

$$\begin{aligned}
 \beta_2 &= \left\| (x - \tilde{x})^\top \tilde{X} W^2(Q, S) \mathbb{1} \left(Q \in \mathcal{S}_d \left(M_{\tilde{L}}^{-1}, M_{\tilde{L}} \right) \right) \right\|_{\psi_2} \\
 &\stackrel{(i)}{\lesssim} \|x - \tilde{x}\| \left\| \tilde{X} \right\|_{\psi_2} \cdot (M_{\tilde{L}} + \tilde{M}_L) \\
 &\lesssim \tilde{M}_L \|x - \tilde{x}\|_2
 \end{aligned} \tag{109}$$

Here (i) follows from (95).

– As a result of (108) and (109), one can obtain

$$\begin{aligned}
 \left\| f(Z; \theta) - f(Z; \tilde{\theta}) \right\|_{\psi_2} &\leq \beta_1 + \beta_2 \\
 &\lesssim LM_{\tilde{L}}^{1/2} d(\theta, \tilde{\theta})^{1/2} + \tilde{M}_L d(\theta, \tilde{\theta}) \\
 &\lesssim L\tilde{M}_L \left(d(\theta, \tilde{\theta})^{1/2} \vee d(\theta, \tilde{\theta}) \right)
 \end{aligned} \tag{110}$$

Combining (107) and (110) then gives

$$\left\| \alpha_1(\theta) - \alpha_1(\tilde{\theta}) \right\|_{\psi_2} \lesssim \frac{L\tilde{M}_L}{\sqrt{n}} \left(d(\theta, \tilde{\theta})^{1/2} \vee d(\theta, \tilde{\theta}) \right)$$

Hence $\tau(\epsilon)$ in Lemma 29 can be chosen as

$$\tau(\epsilon) = \frac{K}{\sqrt{n}} \left(\epsilon^{1/2} \vee \epsilon \right), \quad K = C_f L\tilde{M}_L \tag{111}$$

• by Lemma 29, one has

$$b_2 := \mathbb{E} \sup_{\theta \in \Theta} |\alpha_1(Z; \theta)| \lesssim \underbrace{\frac{KD}{\sqrt{n}} \sqrt{\log^+ D}}_{\zeta_1} + \underbrace{\sup_{\theta \in \Theta} \mathbb{E} |\alpha_1(Z; \theta)|}_{\zeta_2}$$

– ζ_1 : by (42) in Lemma 29 and (105), one can obtain

$$\zeta_1 \lesssim \frac{L\tilde{M}_L^2}{\sqrt{n}} \sqrt{\log \tilde{M}_L} \tag{112}$$

– ζ_2 : For any $\theta \in \Theta$, we have

$$\begin{aligned}
 \|f(Z; \theta)\|_{\psi_2} &\leq \|f(Z; \theta)\|_{\psi_2} \\
 &\leq \left\| 1 + (x - \mu)^\top \Sigma^{-1} (\tilde{X} - \mu) \right\|_{\psi_2} \\
 &\quad \cdot \sup_{S \in \mathcal{S}_d(0, \tilde{M}_L)} \left[W^2(Q, S) \mathbb{1} \left(Q \in \mathcal{S}_d \left(M_{\tilde{L}}^{-1}, M_{\tilde{L}} \right) \right) \right] \\
 &\lesssim \left(1 + \|x - \mu\| \left\| \tilde{X} - \mu \right\|_{\psi_2} \right) \cdot (M_{\tilde{L}} + \tilde{M}_L) \\
 &\lesssim L\tilde{M}_L
 \end{aligned}$$

Therefore for any $\theta \in \Theta$, one has $\|\alpha_1(\theta)\|_{\psi_2} \lesssim L\tilde{M}_L/\sqrt{n}$ and

$$\mathbb{E} |\alpha_1(Z; \theta)| \lesssim \|\alpha_1(\theta)\|_{\psi_2} \lesssim \frac{L\tilde{M}_L}{\sqrt{n}}$$

Take supremum over $\theta \in \Theta$ to see that

$$\zeta_2 \lesssim \frac{L\tilde{M}_L}{\sqrt{n}} \tag{113}$$

– As a result of (112) and (113), one can obtain

$$b_2 \lesssim \frac{L\widetilde{M}_L^2}{\sqrt{n}} \sqrt{\log \widetilde{M}_L} \quad (114)$$

Finally, combine (104) and (114) to see that

$$\sup_{\theta \in \Theta} |\alpha_1(\theta)| \leq b_1 + b_2 \lesssim C \frac{L\widetilde{M}_L^2}{\sqrt{n}} \sqrt{\log \widetilde{M}_L} \quad (115)$$

with probability at least $1 - O(n^{-(1+\tau)})$. The proof of Claim 3 is then complete.

C.4 Proof of Lemma 38

Without loss of generality, assume $\mu = 0$, $\Sigma = I_p$ and $C_{\psi_2} = 1$ in Assumption 1 and Assumption 2. To prove the uniform convergence of \widehat{Q}_ρ in (81), observe that for any $\delta_n, \epsilon_n > 0$, one has

$$\begin{aligned} & \left\{ \sup_{x \in B_\mu(L)} \left\| \widehat{Q}_\rho(x) - Q^*(x) \right\|_F \leq \delta_n \right\} \\ & \stackrel{(i)}{\supset} \left\{ \inf_{S: \|S - Q^*(x)\|_F \leq \delta_n} \widehat{F}_\rho(x, S) < \inf_{S: \|S - Q^*(x)\|_F \geq \delta_n} \widehat{F}_\rho(x, S), \forall x \in B_\mu(L) \right\} \\ & \stackrel{(ii)}{\supset} \left\{ \widehat{F}_\rho(x, Q^*(x)) < \inf_{S: \|S - Q^*(x)\|_F \geq \delta_n} \widehat{F}_\rho(x, S), \forall x \in B_\mu(L) \right\} \cap \left\{ \widehat{Q}_\rho(x) \preceq \widetilde{M}_L I_d, \forall x \in B_\mu(L) \right\} \\ & \stackrel{(iii)}{\supset} \left\{ \widehat{F}_\rho(x, Q^*(x)) < \inf_{\substack{S: \|S - Q^*(x)\|_F \geq \delta_n \\ S \preceq \widetilde{M}_L I_d}} \widehat{F}_\rho(x, S), \forall x \in B_\mu(L) \right\} \cap \left\{ \widehat{Q}_\rho(x) \preceq \widetilde{M}_L I_d, \forall x \in B_\mu(L) \right\} \\ & \supset \underbrace{\left\{ \sup_{\substack{x \in B_\mu(L) \\ S \preceq \widetilde{M}_L I_d}} \left| \widehat{F}_\rho(x, S) - F(x, S) \right| \leq \epsilon_n \right\}}_{=: \mathcal{E}_1} \\ & \cap \underbrace{\left\{ F(x, Q^*(x)) \leq \inf_{\substack{S: \|S - Q^*(x)\|_F \geq \delta_n \\ S \preceq \widetilde{M}_L I_d}} F(x, S) - 3\epsilon_n, \forall x \in B_\mu(L) \right\}}_{=: \mathcal{E}_2} \\ & \cap \underbrace{\left\{ \widehat{Q}_\rho(x) \preceq \widetilde{M}_L I_d, \forall x \in B_\mu(L) \right\}}_{=: \widetilde{E}_0 \text{ in Lemma 36}} \end{aligned} \quad (116)$$

Here (i) and (iii) follow since $\widehat{Q}_\rho(x)$ is the minimizer of $F_{n,\rho}(x, \cdot)$; (ii) is obtained by setting $S = Q^*(x)$.

With (116) in place, it suffices to choose appropriate δ_n, ϵ_n such that $\mathcal{E}_1 \cap \mathcal{E}_2$ occurs with high probability and define $E_2 = \mathcal{E}_1 \cap \mathcal{E}_2 \cap \widetilde{E}_0$. To this end, we first find the uniform

convergence rate ϵ_n so that $\mathbb{P}(\mathcal{E}_1^c) \lesssim n^{-\tau}$. Next, δ_n is set based on ϵ_n and properties of F so that $\mathbb{P}(\mathcal{E}_2^c) \lesssim n^{-\tau}$. For \tilde{E}_0 , we already showed in (80) that $\mathbb{P}(\tilde{E}_0^c) \lesssim n^{-\tau}$. Combining results on $\mathcal{E}_1, \mathcal{E}_2$ and \tilde{E}_0 , the proof is then complete.

Analysis of \mathcal{E}_1 : Let $\epsilon_0 = \delta_F^{\alpha_F} / \gamma_F(L, \sqrt{d}\tilde{M}_L)$ (δ_F and $\gamma_F(\cdot, \cdot)$ are defined in Assumption 4) and set

$$\epsilon_n := C_{F,\tau} \frac{L\tilde{M}_L^3}{\sqrt{n}} \wedge \frac{\epsilon_0}{3}$$

Note that for $L = O(\sqrt{\log n})$, $C_{F,\tau} \frac{L\tilde{M}_L^3}{\sqrt{n}} \leq \epsilon_0/3$ for large n . Therefore, Lemma 37 implies that for any $\tau \geq 0$,

$$\mathbb{P}\{\mathcal{E}_1\} \geq 1 - c_{\delta,\tau} n^{-\tau}$$

for some constant $c_{\delta,\tau} > 0$ independent of n .

Analysis of \mathcal{E}_2 : Note that for any $x \in B_\mu(L)$ and $S \preceq \tilde{M}_L I_d$, one has

$$\|S - Q^*(x)\|_F \leq \sqrt{d} \max\{\|S\|_{\text{op}}, \|Q^*(x)\|_{\text{op}}\} \leq \sqrt{d}\tilde{M}_L$$

Hence one has

$$\inf_{\substack{S: \|S - Q^*(x)\|_F \geq \delta_n \\ S \preceq \tilde{M}_L I_d}} F(x, S) - F(x, Q^*(x)) \geq \inf_{S: \delta_n \leq \|S - Q^*(x)\|_F \leq \sqrt{d}\tilde{M}_L} F(x, S) - F(x, Q^*(x))$$

Set

$$\delta_n := \left[3\epsilon_n \gamma_F(L, \sqrt{d}\tilde{M}_L) \right]^{1/\alpha_F}$$

By definition, $\delta_n \leq \delta_F$. Therefore, Assumption 4 implies that $\mathbb{P}(\mathcal{E}_2) = 1$ for n large enough.

Combining the above results on $\mathcal{E}_1, \mathcal{E}_2$ and \tilde{E}_0 , one can obtain

$$\mathbb{P} \left(\sup_{x \in B_\mu(L)} \left\| \hat{Q}_\rho(x) - Q^*(x) \right\|_F \leq C_{\delta,\tau} \left[\frac{L\tilde{M}_L^3 \gamma_F(L, \sqrt{d}\tilde{M}_L)}{\sqrt{n}} \right]^{1/\alpha_F} \right) \geq 1 - c_{\delta,\tau} n^{-\tau}$$

for constants $C_{\delta,\tau}, c_{\delta,\tau}$ independent of n . Note that by definition, $L, \tilde{M}_L, \gamma_F(L, \sqrt{d}\tilde{M}_L) \lesssim \text{polylog}(n)$, one can then arrive at (81).

Finally, Eq. (82) follows by noticing that $\delta_F \wedge \delta_n = o(M_L)$.

C.5 Proof of Lemma 40

Without loss of generality, assume $\mu = 0$, $\Sigma = I_p$, and $C_{\psi_2} = 1$ in Assumption 1-Assumption 2. If X is unbounded, then one can apply the truncation as (98) and follow similar arguments as in Lemma 38. Hence we can assume without loss of generality that $\|X - \mu\| \leq \tilde{L}$ almost surely with $\tilde{L} := C_{\psi_2} L_\tau = C_{\psi_2} \sqrt{(1 + \tau) \log n}$. Then we have a.s.

$$\begin{aligned} \|X - \mu\| &\leq \tilde{L} \\ Q, Q^*(X) &\in \mathcal{S}_d(M_{\tilde{L}}^{-1}, M_{\tilde{L}}) \\ \mathbb{1}(E_0) &= 1 \end{aligned} \tag{117}$$

where we remind readers that $E_0 := \left\{ \|X_i - \mu\| \leq \tilde{L}, Q_i \in \mathcal{S}_d(M_{\tilde{L}}^{-1}, M_{\tilde{L}}), i \in [n] \right\}$ is defined in (78).

Proof of (83): one can obtain the following decomposition.

$$\sup_{x \in B_\mu(L)} \left\| \tilde{A}_n(x) \right\|_F \leq \underbrace{\sup_{x \in B_\mu(L)} \left\| \tilde{A}_n(x) \right\|_F - \mathbb{E} \sup_{x \in B_\mu(L)} \left\| \tilde{A}_n(x) \right\|_F}_{a_1} + \underbrace{\mathbb{E} \sup_{x \in B_\mu(L)} \left\| \tilde{A}_n(x) \right\|_F}_{a_2} \quad (118)$$

Let $\varphi(Z; x) = w(x, X) \left(T_{Q^*(x)}^Q - I_d \right)$.

Analysis of a_1 : By (117), one has almost surely that for any $x \in B_\mu(L)$,

$$\begin{aligned} \|\varphi(Z; x)\|_F &\lesssim (1 + \|x - \mu\| \cdot \|X - \mu\|) \cdot \sup_{\substack{Q \in \mathcal{S}_d(M_{\tilde{L}}^{-1}, M_{\tilde{L}}) \\ S \in \mathcal{S}_d((M_L)^{-1}, M_L)}} \left\| T_S^Q - I_d \right\|_F \\ &\stackrel{(i)}{\lesssim} L \tilde{L} M_{\tilde{L}}^{1/2} M_L^{3/2} \end{aligned}$$

Here (i) follows by noticing that

$$\sup_{\substack{Q \in \mathcal{S}_d(M_{\tilde{L}}^{-1}, M_{\tilde{L}}) \\ S \in \mathcal{S}_d((M_L)^{-1}, M_L)}} \left\| T_S^Q \right\|_{\text{op}} = \sup_{\substack{Q \in \mathcal{S}_d(M_{\tilde{L}}^{-1}, M_{\tilde{L}}) \\ S \in \mathcal{S}_d((M_L)^{-1}, M_L)}} \left\| S^{-1/2} (S^{1/2} Q S^{1/2})^{1/2} S^{-1/2} \right\|_{\text{op}} \leq M_{\tilde{L}}^{1/2} M_L^{3/2}$$

Lemma 31 then implies that $\|a_1\|_{\psi_2} \lesssim \frac{L \tilde{L} M_{\tilde{L}}^{1/2} M_L^{3/2}}{\sqrt{n}}$. Therefore, one can obtain

$$a_1 \lesssim \frac{L \tilde{L} M_{\tilde{L}}^{1/2} M_L^{3/2} L_\tau}{\sqrt{n}} \quad (119)$$

with probability at least $1 - O(n^{-(1+\tau)})$.

Analysis of a_2 : to apply Lemma 29, we follow the steps below.

- First, we consider the sub-Gaussian norm of $\left\| \tilde{A}_n(x) \right\|_F - \left\| \tilde{A}_n(\tilde{x}) \right\|_F$ as in (40) in Lemma 29. For any $x, \tilde{x} \in B_\mu(L)$, one has

$$\left\| \left\| \tilde{A}_n(x) \right\|_F - \left\| \tilde{A}_n(\tilde{x}) \right\|_F \right\|_{\psi_2} \leq \left\| \left\| \tilde{A}_n(x) - \tilde{A}_n(\tilde{x}) \right\|_F \right\|_{\psi_2} \quad (120)$$

Note that

$$\tilde{A}_n(x) - \tilde{A}_n(\tilde{x}) = \frac{1}{n} \sum_{i=1}^n \left[w(x, X_i) \left(T_{Q^*(x)}^{Q_i} - I_d \right) - w(\tilde{x}, X_i) \left(T_{Q^*(\tilde{x})}^{Q_i} - I_d \right) \right]$$

Since $\mathbb{E}\tilde{A}_n(x) = 0$ by the optimality condition (147), one has

$$\begin{aligned}
 \left\| \tilde{A}_n(x) - \tilde{A}_n(\tilde{x}) \right\|_{\mathbb{F}} \Big\|_{\psi_2} &\stackrel{(i)}{\lesssim} \frac{1}{\sqrt{n}} \left\| \left\| w(x, X) \left(T_{Q^*(x)}^Q - I_d \right) - w(\tilde{x}, X) \left(T_{Q^*(\tilde{x})}^Q - I_d \right) \right\|_{\mathbb{F}} \right\|_{\psi_2} \\
 &\leq \frac{1}{\sqrt{n}} \underbrace{\left\| \left\| w(x, X) \left(T_{Q^*(x)}^Q - I_d \right) - w(x, X) \left(T_{Q^*(\tilde{x})}^Q - I_d \right) \right\|_{\mathbb{F}} \right\|_{\psi_2}}_{b_1} \\
 &\quad + \frac{1}{\sqrt{n}} \underbrace{\left\| \left\| w(x, X) \left(T_{Q^*(\tilde{x})}^Q - I_d \right) - w(\tilde{x}, X) \left(T_{Q^*(\tilde{x})}^Q - I_d \right) \right\|_{\mathbb{F}} \right\|_{\psi_2}}_{b_2}
 \end{aligned} \tag{121}$$

Here (i) follows from (43) in Lemma 29.

– b_1 : For any $x, \tilde{x} \in B_\mu(L)$, one has $Q^*(x), Q^*(\tilde{x}) \in \mathcal{S}_d(M_L^{-1}, M_L)$. Then one can obtain

$$\begin{aligned}
 b_1 &= \left\| \left\| w(x, X) \left(T_{Q^*(x)}^Q - T_{Q^*(\tilde{x})}^Q \right) \right\|_{\mathbb{F}} \right\|_{\psi_2} \\
 &\lesssim \left\| 1 + \|x - \mu\| \cdot \|X - \mu\| \cdot \left\| T_{Q^*(x)}^Q - T_{Q^*(\tilde{x})}^Q \right\|_{\mathbb{F}} \right\|_{\psi_2} \\
 &\lesssim \|1 + \|x - \mu\| \cdot \|X - \mu\|\|_{\psi_2} \cdot \sup_{Q \in \mathcal{S}_d(M_L^{-1}, M_L)} \left\| T_{Q^*(x)}^Q - T_{Q^*(\tilde{x})}^Q \right\|_{\mathbb{F}} \\
 &\lesssim L \cdot \sup_{\substack{Q \in \mathcal{S}_d(M_L^{-1}, M_L) \\ x, \tilde{x} \in B_\mu(L)}} \left\| T_{Q^*(x)}^Q - T_{Q^*(\tilde{x})}^Q \right\|_{\mathbb{F}} \\
 &\stackrel{(i)}{=} L \cdot \sup_{\substack{Q \in \mathcal{S}_d(M_L^{-1}, M_L) \\ Q' \in \mathcal{S}_d(M_L^{-1}, M_L)}} \left\| dT_{Q'}^Q \cdot (Q^*(x) - Q^*(\tilde{x})) \right\|_{\mathbb{F}} \\
 &\stackrel{(ii)}{\leq} L \cdot \|Q^*(x) - Q^*(\tilde{x})\|_{\mathbb{F}} \cdot \sup_{\substack{Q \in \mathcal{S}_d(M_L^{-1}, M_L) \\ S \in \mathcal{S}_d(M_L^{-1}, M_L)}} \left\| dT_S^Q \right\| \\
 &\stackrel{(iii)}{\lesssim} L \cdot \text{poly}(M_L, M_{\tilde{L}}) \cdot \|Q^*(x) - Q^*(\tilde{x})\|_{\mathbb{F}}
 \end{aligned} \tag{122}$$

Here (i) is a consequence of the mean value theorem (Dudley and Norvaiša, 2011, Theorem 5.3) for some Q' that lies between $Q^*(x)$ and $Q^*(\tilde{x})$, (ii) arises from Lemma 22, and (iii) follows from Lemma 21.

– b_2 : one can obtain that

$$\begin{aligned}
 b_2 &\leq \left\| (x - \tilde{x}) \Sigma^{-1} (X - \mu) \right\|_{\psi_2} \sup_{\substack{Q \in \mathcal{S}_d(M_L^{-1}, M_L) \\ S \in \mathcal{S}_d(M_L^{-1}, M_L)}} \left\| T_S^Q - I_d \right\|_{\mathbb{F}} \\
 &\stackrel{(i)}{\lesssim} \|x - \tilde{x}\| \left\| X - \mu \right\|_{\psi_2} \text{poly}(M_L, M_{\tilde{L}}) \\
 &\lesssim \|x - \tilde{x}\| \text{poly}(M_L, M_{\tilde{L}})
 \end{aligned} \tag{123}$$

almost surely. Here (i) arises from Lemma 25 and the bounds on T_S^Q in Lemma 21.

Combining (120), (121), (122) and (123) gives that for any $x, \tilde{x} \in B_\mu(L)$,

$$\begin{aligned}
 & \left\| \left\| \tilde{A}_n(x) \right\|_F - \left\| \tilde{A}_n(\tilde{x}) \right\|_F \right\|_{\psi_2} \\
 & \lesssim \frac{1}{\sqrt{n}} \cdot L \text{poly}(M_L, M_{\tilde{L}}) (\|x - \tilde{x}\| + \|Q^*(x) - Q^*(\tilde{x})\|_F) \\
 & \stackrel{(i)}{\lesssim} \frac{1}{\sqrt{n}} \cdot L \text{poly}(M_L, M_{\tilde{L}}) \left(\|x - \tilde{x}\| + M_L \gamma_F(L, M_L) \left(\|x - \tilde{x}\|_2 \vee \|x - \tilde{x}\|^{1/\alpha_F} \right) \right) \\
 & \stackrel{(ii)}{\leq} \frac{C}{\sqrt{n}} \cdot L^{1+C_F} M_L^{4+C_F} \left(\|x - \tilde{x}\|_2 \vee \|x - \tilde{x}\|^{1/\alpha_F} \right)
 \end{aligned} \tag{124}$$

Here (i) follows from Lemma 35 and the fact that $M \gamma_F(L, M) \geq 1$.

- With the Hölder continuity (124) in place, we can apply Lemma 29 with $\tau(\epsilon)$ chosen as

$$\tau(\epsilon) = \frac{K}{\sqrt{n}} \left(\epsilon^{1/\alpha_F} \vee \epsilon \right), \quad K = CL^{1+C_F} M_L^{4+C_F}$$

which gives

$$\begin{aligned}
 a_2 & \lesssim \frac{KL^{1 \vee \alpha_F^{-1}}}{\sqrt{n}} \sqrt{\log^+ L} + \sup_{x \in B_\mu(L)} \mathbb{E} \left\| \tilde{A}_n(x) \right\|_F \\
 & \stackrel{(i)}{=} \underbrace{\frac{KL}{\sqrt{n}} \sqrt{\log^+ L}}_{\zeta_1} + \underbrace{\sup_{x \in B_\mu(L)} \mathbb{E} \left\| \tilde{A}_n(x) \right\|_F}_{\zeta_2}
 \end{aligned}$$

Here (i) follows since $\alpha_F \geq 1$ as defined in Assumption 4.

– ζ_1 : direct calculation gives that

$$\zeta_1 \lesssim \frac{L^{2+C_F} M_L^{4+C_F}}{\sqrt{n}} \sqrt{\log^+ L} \tag{125}$$

– ζ_2 : for any $U \in \mathbb{R}^{d \times d}$ with unit Frobenius norm, one has

$$\begin{aligned}
 \sup_{x \in B_\mu(L)} \left\| \left\| \tilde{A}_n(x) \right\|_F \right\|_{\psi_2} &\stackrel{(i)}{\lesssim} \sup_{x \in B_\mu(L)} \sup_{\|U\|_F=1} \left\| \left\langle U, \tilde{A}_n(x) \right\rangle \right\|_{\psi_2} \\
 &\stackrel{(ii)}{\lesssim} \sup_{x \in B_\mu(L)} \sup_{\|U\|_F=1} \left\| \left\langle U, \frac{1}{n} \sum_{i=1}^n \varphi(Z_i; x) - \mathbb{E} \varphi(Z; x) \right\rangle \right\|_{\psi_2} \\
 &\lesssim \sup_{x \in B_\mu(L)} \sup_{\|U\|_F=1} \frac{1}{\sqrt{n}} \left\| \langle U, \varphi(Z; x) \rangle \right\|_{\psi_2} \\
 &\leq \sup_{x \in B_\mu(L)} \frac{1}{\sqrt{n}} \left\| \left\| \varphi(Z; x) \right\|_F \right\|_{\psi_2} \\
 &\leq \frac{1}{\sqrt{n}} \sup_{x \in B_\mu(L)} \left\| \left\| w(x, X) \right\| \right\|_{\psi_2} \cdot \sup_{\substack{Q \in \mathcal{S}_d(M_{\tilde{L}}^{-1}, M_{\tilde{L}}) \\ S \in \mathcal{S}_d(M_{\tilde{L}}^{-1}, M_L)}} \left\| T_S^Q - I_d \right\|_F \\
 &\lesssim \frac{1}{\sqrt{n}} L \cdot \text{poly}(M_L, M_{\tilde{L}})
 \end{aligned}$$

Here (i) results from Lemma 25, (ii) follows from independence. Therefore,

$$\zeta_2 \lesssim \frac{L \cdot \text{poly}(M_L, M_{\tilde{L}})}{\sqrt{n}} \quad (126)$$

• As a result of (125) and (126), one can obtain

$$a_2 \lesssim \frac{\text{poly}(L, M_L, M_{\tilde{L}})}{\sqrt{n}} \sqrt{\log^+ L} \quad (127)$$

Finally, combining (118), (119) and (127) gives (83).

Proof of (84): one can obtain the following decomposition.

$$\begin{aligned}
 \sup_{x \in B_\mu(L)} \left\| \left\| \tilde{\Phi}_n(x) - \mathbb{E} \tilde{\Phi}_n(x) \right\| \right\| &\leq \underbrace{\sup_{x \in B_\mu(L)} \left\| \left\| \tilde{\Phi}_n(x) - \mathbb{E} \tilde{\Phi}_n(x) \right\| - \mathbb{E} \sup_{x \in B_\mu(L)} \left\| \left\| \tilde{\Phi}_n(x) - \mathbb{E} \tilde{\Phi}_n(x) \right\| \right\|}_{a_3} \\
 &\quad + \underbrace{\mathbb{E} \sup_{x \in B_\mu(L)} \left\| \left\| \tilde{\Phi}_n(x) - \mathbb{E} \tilde{\Phi}_n(x) \right\| \right\|}_{a_4}
 \end{aligned} \quad (128)$$

Let $Z = (X, Q)$. Define $\phi(Z; x) = w(x, X) dT_{Q^*(x)}^Q$.

Analysis of a_3 : By (117), one has almost surely that for any $x \in B_\mu(L)$,

$$\begin{aligned}
 \left\| \phi(Z; x) \right\| &\lesssim L \tilde{L} \cdot \sup_{\substack{Q \in \mathcal{S}_d(M_{\tilde{L}}^{-1}, M_{\tilde{L}}) \\ S \in \mathcal{S}_d(M_{\tilde{L}}^{-1}, M_L)}} \left\| dT_S^Q \right\| \\
 &\lesssim L \tilde{L} \text{poly}(M_L, M_{\tilde{L}})
 \end{aligned}$$

By Lemma 31, one can then obtain

$$\|a_3\|_{\psi_2} \lesssim \frac{L\tilde{L}\text{poly}(M_L, M_{\tilde{L}})}{\sqrt{n}} \quad (129)$$

Analysis of a_4 : to apply Lemma 29, we follow the steps below.

- First, let us consider the sub-Gaussian norm of $\left\| \tilde{\Phi}_n(x) - \mathbb{E}\tilde{\Phi}_n(x) \right\| - \left\| \tilde{\Phi}_n(\tilde{x}) - \mathbb{E}\tilde{\Phi}_n(\tilde{x}) \right\|$. For any $x, \tilde{x} \in B_\mu(L)$, by (43) in Lemma 29, one can obtain

$$\left\| \left\| \tilde{\Phi}_n(x) - \mathbb{E}\tilde{\Phi}_n(x) \right\| - \left\| \tilde{\Phi}_n(\tilde{x}) - \mathbb{E}\tilde{\Phi}_n(\tilde{x}) \right\| \right\|_{\psi_2} \lesssim \frac{1}{\sqrt{n}} \left\| \phi(Z; x) - \phi(Z; \tilde{x}) \right\|_{\psi_2} \quad (130)$$

Moreover,

$$\begin{aligned} & \left\| \left\| \phi(Z; x) - \phi(Z; \tilde{x}) \right\| \right\|_{\psi_2} \\ &= \left\| \left\| w(x, X)dT_{Q^*(x)}^Q - w(\tilde{x}, X)dT_{Q^*(\tilde{x})}^Q \right\| \right\|_{\psi_2} \\ &\leq \underbrace{\left\| \left\| w(x, X)dT_{Q^*(x)}^Q - w(x, X)dT_{Q^*(\tilde{x})}^Q \right\| \right\|_{\psi_2}}_{b_3} \\ &\quad + \underbrace{\left\| \left\| w(x, X)dT_{Q^*(\tilde{x})}^Q - w(\tilde{x}, X)dT_{Q^*(\tilde{x})}^Q \right\| \right\|_{\psi_2}}_{b_4} \end{aligned} \quad (131)$$

– b_3 : we have for any $x, \tilde{x} \in B_\mu(L)$,

$$\begin{aligned} b_3 &\stackrel{(i)}{=} \left\| \left\| w(x, X)d^2T_{Q'}^Q \cdot (Q^*(x) - Q^*(\tilde{x})) \right\| \right\|_{\psi_2} \\ &\lesssim L \|X - \mu\|_{\psi_2} \cdot \|Q^*(x) - Q^*(\tilde{x})\|_F \cdot \sup_{\substack{Q \in \mathcal{S}_d(M_L^{-1}, M_{\tilde{L}}) \\ S \in \mathcal{S}_d(M_L^{-1}, M_L)}} \left\| d^2T_S^Q \right\| \\ &\stackrel{(ii)}{\lesssim} L \cdot \text{poly}(M_L, M_{\tilde{L}}) \|Q^*(x) - Q^*(\tilde{x})\|_F \end{aligned} \quad (132)$$

Here (i) is a consequence of the mean value theorem (Dudley and Norvaiša, 2011, Theorem 5.3) for some Q' that lies between $Q^*(x)$ and $Q^*(\tilde{x})$ and (ii) follows from Lemma 21.

– b_4 : for any $x, \tilde{x} \in B_\mu(L)$,

$$\begin{aligned} b_4 &\leq \left\| \left\| w(x, X) - w(\tilde{x}, X) \right\| \right\|_{\psi_2} \sup_{\substack{Q \in \mathcal{S}_d(M_L^{-1}, M_{\tilde{L}}) \\ S \in \mathcal{S}_d(M_L^{-1}, M_L)}} \left\| dT_S^Q \right\| \\ &\lesssim \text{poly}(M_L, M_{\tilde{L}}) \|x - \tilde{x}\|_2 \end{aligned} \quad (133)$$

Combining (130), (131), (132) and (133), one can obtain that

$$\begin{aligned}
 & \left\| \left\| \tilde{\Phi}_n(x) - \mathbb{E} \tilde{\Phi}_n(x) \right\| - \left\| \tilde{\Phi}_n(\tilde{x}) - \mathbb{E} \tilde{\Phi}_n(\tilde{x}) \right\| \right\|_{\psi_2} \\
 & \lesssim \frac{1}{\sqrt{n}} \cdot L \text{poly}(M_L, M_{\tilde{L}}) (\|x - \tilde{x}\| + \|Q^*(x) - Q^*(\tilde{x})\|_F) \\
 & \lesssim \frac{1}{\sqrt{n}} \cdot L \text{poly}(M_L, M_{\tilde{L}}) \left(\|x - \tilde{x}\| \vee \|x - \tilde{x}\|^{1/\alpha_F} \right)
 \end{aligned}$$

- With the Hölder continuity above, we can apply Lemma 29 with $\tau(\epsilon)$ chosen as

$$\tau(\epsilon) = \frac{K}{\sqrt{n}} \left(\epsilon^{1/\alpha_F} \vee \epsilon \right), \quad K = \text{poly}(L, M_L)$$

which gives

$$a_4 \lesssim \underbrace{\frac{KL}{\sqrt{n}} \sqrt{\log^+ L}}_{\zeta_3} + \underbrace{\sup_{x \in B_\mu(L)} \mathbb{E} \left\| \tilde{\Phi}_n(x) \right\|}_{\zeta_4}$$

- ζ_3 : direct calculation gives

$$\zeta_3 \lesssim \frac{\text{poly}(L, M_L)}{\sqrt{n}} \tag{134}$$

- ζ_4 : for any $U, V \in \mathbb{R}^{d \times d}$ with unit Frobenius norm, one has

$$\begin{aligned}
 \sup_{x \in B_\mu(L)} \left\| \left\| \tilde{\Phi}_n(x) \right\| \right\|_{\psi_2} & \stackrel{(i)}{\lesssim} \sup_{x \in B_\mu(L)} \sup_{\|U\|_F = \|V\|_F = 1} \left\| \left\langle V, \tilde{\Phi}_n(x) \cdot U \right\rangle \right\|_{\psi_2} \\
 & \stackrel{(ii)}{\lesssim} \sup_{x \in B_\mu(L)} \sup_{\|U\|_F = \|V\|_F = 1} \frac{1}{\sqrt{n}} \left\| \left\langle V, [\phi(Z; x) - E\phi(Z; x)] \cdot U \right\rangle \right\|_{\psi_2} \\
 & \lesssim \sup_{x \in B_\mu(L)} \sup_{\|U\|_F = \|V\|_F = 1} \frac{1}{\sqrt{n}} \left\| \left\langle V, \phi(Z; x) \cdot U \right\rangle \right\|_{\psi_2} \\
 & \leq \sup_{x \in B_\mu(L)} \frac{1}{\sqrt{n}} \left\| \left\| \phi(Z; x) \right\| \right\|_{\psi_2} \\
 & \leq \frac{1}{\sqrt{n}} \sup_{x \in B_\mu(L)} \|w(x, X)\|_{\psi_2} \cdot \sup_{\substack{Q \in S_d(M_{\tilde{L}}^{-1}, M_{\tilde{L}}) \\ S \in S_d(M_{\tilde{L}}^{-1}, M_L)}} \left\| dT_S^Q \right\| \\
 & \stackrel{(iii)}{\lesssim} \frac{1}{\sqrt{n}} L \text{poly}(M_L, M_{\tilde{L}})
 \end{aligned}$$

Here (i) results from Lemma 25, (ii) follows from independence and (iii) is due to Lemma 21. Hence

$$\zeta_4 \lesssim \frac{L \text{poly}(M_L, M_{\tilde{L}})}{\sqrt{n}} \tag{135}$$

- As a result of (134) and (135), one can obtain

$$a_4 \lesssim \frac{\text{poly}(L, M_L, M_{\tilde{L}})}{\sqrt{n}} \tag{136}$$

Finally, combining (128), (129) and (136) gives (84).

Proof of (85): Denote $\Theta = B_\mu(L) \times \mathcal{S}_d((C_{\text{slow}}M_{\tilde{L}})^{-1}, C_{\text{slow}}M_{\tilde{L}})$ and $\theta = (x, S) \in \Theta$.

$$\begin{aligned} \sup_{\theta \in \Theta} \left\| \tilde{\Psi}_n(\theta) - \mathbb{E} \tilde{\Psi}_n(\theta) \right\| &\leq \underbrace{\sup_{\theta \in \Theta} \left\| \tilde{\Psi}_n(\theta) - \mathbb{E} \tilde{\Psi}_n(\theta) \right\| - \mathbb{E} \sup_{\theta \in \Theta} \left\| \tilde{\Psi}_n(\theta) - \mathbb{E} \tilde{\Psi}_n(\theta) \right\|}_{a_5} \\ &\quad + \underbrace{\mathbb{E} \sup_{\theta \in \Theta} \left\| \tilde{\Psi}_n(\theta) - \mathbb{E} \tilde{\Psi}_n(\theta) \right\|}_{a_6} \end{aligned} \quad (137)$$

Let $Z = (X, Q)$. Define $\psi(Z; \theta) = w(x, X)d^2T_S^Q$ and $\bar{\psi}(Z; \theta) = \psi(Z; \theta) - \mathbb{E}\psi(Z; \theta)$
Analysis of a_5 : By (117), one has almost surely that

$$\begin{aligned} \sup_{Z, \theta} \left\| \psi(Z; \theta) \right\| &\leq \sup_{\substack{x \in B_\mu(L) \\ X \in B_\mu(\tilde{L})}} |w(x, X)| \cdot \sup_{\substack{Q \in \mathcal{S}_d(M_{\tilde{L}}^{-1}, M_{\tilde{L}}) \\ S \in \mathcal{S}_d((C_{\text{slow}}M_L)^{-1}, C_{\text{slow}}M_L)}} \left\| d^2T_S^Q \right\| \\ &\lesssim L\tilde{L} \cdot \text{poly}(M_L, M_{\tilde{L}}) \end{aligned}$$

Lemma 31 then implies that , $\|a_5\|_{\psi_2} \lesssim \frac{L\tilde{L} \cdot \text{poly}(M_{\tilde{L}}M_L)}{\sqrt{n}}$. Therefore, one can obtain

$$a_5 \lesssim L\tilde{L} \cdot \text{poly}(M_L M_{\tilde{L}}) \frac{L_\tau}{\sqrt{n}} \quad (138)$$

with probability at least $1 - O(n^{-(1+\tau)})$.

Analysis of a_6 : to apply Lemma 29, we follow the steps below.

- First, let us consider the sub-Gaussian norm of $\left\| \tilde{\Psi}_n(\theta) - \mathbb{E} \tilde{\Psi}_n(\theta) \right\| - \left\| \tilde{\Psi}_n(\tilde{\theta}) - \mathbb{E} \tilde{\Psi}_n(\tilde{\theta}) \right\|$.
 For any $\theta, \tilde{\theta} \in \Theta$, one has

$$\left\| \left\| \tilde{\Psi}_n(\theta) - \mathbb{E} \tilde{\Psi}_n(\theta) \right\| - \left\| \tilde{\Psi}_n(\tilde{\theta}) - \mathbb{E} \tilde{\Psi}_n(\tilde{\theta}) \right\| \right\|_{\psi_2} \stackrel{(i)}{\lesssim} \frac{1}{\sqrt{n}} \left\| \left\| \psi(Z; \theta) - \psi(Z; \tilde{\theta}) \right\| \right\|_{\psi_2} \quad (139)$$

Here (i) follows from Lemma 29. Moreover, one has

$$\begin{aligned} &\left\| \left\| \psi(Z; \theta) - \psi(Z; \tilde{\theta}) \right\| \right\|_{\psi_2} \\ &= \left\| \left\| w(x, X)d^2T_S^Q - w(\tilde{x}, X)d^2T_{\tilde{S}}^Q \right\| \right\|_{\psi_2} \\ &\leq \underbrace{\left\| \left\| w(x, X)d^2T_S^Q - w(x, X)d^2T_{\tilde{S}}^Q \right\| \right\|_{\psi_2}}_{b_5} + \underbrace{\left\| \left\| w(x, X)d^2T_S^Q - w(\tilde{x}, X)d^2T_{\tilde{S}}^Q \right\| \right\|_{\psi_2}}_{b_6} \end{aligned} \quad (140)$$

– b_5 : for any $\theta, \tilde{\theta} \in \Theta$, one can obtain

$$b_5 \stackrel{(i)}{=} \left\| \left\| w(x, X)d^3T_{Q'}^Q \cdot (S - \tilde{S}) \right\| \right\|_{\psi_2}$$

$$\begin{aligned}
 &\leq \|w(x, X)\|_{\psi_2} \cdot \|S - \tilde{S}\|_F \cdot \sup_{\substack{Q \in \mathcal{S}_d(M_{\tilde{L}}^{-1}, M_{\tilde{L}}) \\ S \in \mathcal{S}_d((C_{\text{slow}} M_L)^{-1}, C_{\text{slow}} M_L)}} \|d^3 T_S^Q\| \\
 &\stackrel{(ii)}{\lesssim} L \cdot \text{poly}(M_L, M_{\tilde{L}}) \|S - \tilde{S}\|_F
 \end{aligned} \tag{141}$$

Here (i) is a consequence of the mean value theorem (Dudley and Norvaiša, 2011, Theorem 5.3) for some Q' that lies between $Q^*(x)$ and $Q^*(\tilde{x})$ and (ii) follows from Lemma 21.

– b_6 : for any $\theta, \tilde{\theta} \in \Theta$, one has

$$\begin{aligned}
 b_6 &\leq \|w(x, X) - w(\tilde{x}, X)\|_{\psi_2} \sup_{\substack{Q \in \mathcal{S}_d(M_{\tilde{L}}^{-1}, M_{\tilde{L}}) \\ S \in \mathcal{S}_d((C_{\text{slow}} M_L)^{-1}, C_{\text{slow}} M_L)}} \|d^2 T_S^Q\| \\
 &\lesssim \text{poly}(M_L, M_{\tilde{L}}) \|x - \tilde{x}\|
 \end{aligned} \tag{142}$$

Combining (130), (140), (141) and (142), one can obtain

$$\left\| \left\| \tilde{\Psi}_n(\theta) - \mathbb{E} \tilde{\Psi}_n(\theta) \right\| - \left\| \tilde{\Psi}_n(\tilde{\theta}) - \mathbb{E} \tilde{\Psi}_n(\tilde{\theta}) \right\| \right\|_{\psi_2} \lesssim \frac{L \text{poly}(M_L, M_{\tilde{L}})}{\sqrt{n}} \cdot d(\theta, \tilde{\theta})$$

- With the Lipschitz continuity above, we can apply Lemma 29 with $\tau(\epsilon)$ chosen as

$$\tau(\epsilon) = \frac{K}{\sqrt{n}} \epsilon, \quad K = C \cdot L \text{poly}(M_L, M_{\tilde{L}})$$

which gives

$$a_6 \lesssim \underbrace{\frac{K M_{\tilde{L}}}{\sqrt{n}} \sqrt{\log^+ M_{\tilde{L}}}}_{\zeta_5} + \underbrace{\sup_{\theta \in \Theta} \mathbb{E} \left\| \tilde{\Psi}_n(\theta) \right\|}_{\zeta_6}$$

– ζ_5 : direct calculation gives

$$\zeta_5 \lesssim \frac{L \text{poly}(M_L M_{\tilde{L}})}{\sqrt{n}} \tag{143}$$

– ζ_6 : for any $U, V \in \mathbb{R}^{d \times d}$ with unit Frobenius norm, one has

$$\begin{aligned}
 \left\| \left\langle V, \tilde{\Psi}_n(\theta) \cdot U \right\rangle \right\|_{\psi_2} &\lesssim \frac{1}{\sqrt{n}} \|V, [\psi(Z; \theta) - \mathbb{E} \psi(Z; \theta)] \cdot U\|_{\psi_2} \\
 &\lesssim \frac{1}{\sqrt{n}} \|V, \psi(Z; \theta) \cdot U\|_{\psi_2} \\
 &\leq \frac{1}{\sqrt{n}} \|\psi(Z; \theta)\|_{\psi_2} \\
 &\leq \frac{1}{\sqrt{n}} \|w(x, X)\|_{\psi_2} \cdot \sup_{\substack{Q \in \mathcal{S}_d(M_{\tilde{L}}^{-1}, M_{\tilde{L}}) \\ S \in \mathcal{S}_d((C_{\text{slow}} M_L)^{-1}, C_{\text{slow}} M_L)}} \|d^2 T_S^Q\| \\
 &\lesssim \frac{1}{\sqrt{n}} L \cdot \text{poly}(M_L M_{\tilde{L}})
 \end{aligned}$$

which combined with Lemma 25 implies

$$\zeta_6 \lesssim \frac{L \text{poly}(M_L M_{\tilde{L}})}{\sqrt{n}} \quad (144)$$

- As a result of (143) and (144), one can obtain

$$a_6 \lesssim \frac{L \text{poly}(M_L M_{\tilde{L}})}{\sqrt{n}} \quad (145)$$

Finally, combining (137), (138) and (145) gives (85).

Proof of (86): by definition, one has

$$\mathbb{E} \tilde{\Psi}_n(x; S) = \mathbb{E}_{(X, Q)} \left[w(x, X) d^2 T_S^Q \right]$$

Therefore, by the truncation assumption (117), one can obtain

$$\begin{aligned} & \sup_{\substack{x \in B_\mu(L) \\ S \in \mathcal{S}_d((C_{\text{slow}} M_{\tilde{L}})^{-1}, C_{\text{slow}} M_{\tilde{L}})}} \left\| \mathbb{E} \tilde{\Psi}_n(x; S) \right\| \\ & \leq \left[\sup_{x \in B_\mu(L)} \mathbb{E} |w(x, X)| \right] \cdot \left[\sup_{\substack{Q \in \mathcal{S}_d(M_{\tilde{L}}^{-1}, M_{\tilde{L}}) \\ S \in \mathcal{S}_d((C_{\text{slow}} M_L)^{-1}, C_{\text{slow}} M_L)}} \left\| d^2 T_S^Q \right\| \right] \\ & \stackrel{(i)}{\lesssim} L \cdot \text{poly}(M_L M_{\tilde{L}}) \end{aligned}$$

Here (i) follows from Lemma 21.

Finally, note that by definition, $\tilde{L} \asymp \sqrt{\log n}$ and $M_{\tilde{L}} = \text{poly}(\tilde{L})$, one can obtain (86).

C.6 Proof of Lemma 41

As argued at the beginning of Appendix C.5, we can assume without loss of generality that $\|X - \mu\| \leq \tilde{L}$ almost surely with \tilde{L} defined in (78). As a result, we have almost surely that

$$\begin{aligned} & \|X - \mu\| \leq \tilde{L} \\ & Q, Q^*(X) \in \mathcal{S}_d(M_{\tilde{L}}^{-1}, M_{\tilde{L}}) \\ & \mathbb{1}(E_0) = 1 \end{aligned} \quad (146)$$

where $E_0 := \left\{ \|X_i - \mu\| \leq L, Q_i \in \mathcal{S}_d(M_{\tilde{L}}^{-1}, M_{\tilde{L}}), i \in [n] \right\}$.

With the boundedness condition (146) in place, Lemma 21 implies the following upper bounds

$$\begin{aligned}
 \sup_{\substack{Q \in \mathcal{S}_d(M_{\tilde{L}}^{-1}, M_{\tilde{L}}) \\ S \in \mathcal{S}_d((C_{\text{slow}} M_L)^{-1}, C_{\text{slow}} M_L)}} \|T_S^Q - I_d\|_{\text{F}} &\leq \text{poly}(M_L, M_{\tilde{L}}) \\
 \sup_{\substack{Q \in \mathcal{S}_d(M_{\tilde{L}}^{-1}, M_{\tilde{L}}) \\ S \in \mathcal{S}_d((C_{\text{slow}} M_L)^{-1}, C_{\text{slow}} M_L)}} \|dT_S^Q\| &\leq \text{poly}(M_L, M_{\tilde{L}}) \\
 \sup_{\substack{Q \in \mathcal{S}_d(M_{\tilde{L}}^{-1}, M_{\tilde{L}}) \\ S \in \mathcal{S}_d((C_{\text{slow}} M_L)^{-1}, C_{\text{slow}} M_L)}} \|d^2 T_S^Q\| &\leq \text{poly}(M_L, M_{\tilde{L}})
 \end{aligned}$$

almost surely. Applying Lemma 28 then gives the desired results.

C.7 Proof of Lemma 42

Proof of (90): apply the triangle inequality to see that

$$\begin{aligned}
 \sup_{x \in B_\mu(L)} \|\hat{A}_\rho(x)\|_{\text{F}} &\leq \sup_{x \in B_\mu(L)} \|\hat{A}_\rho(x) - \tilde{A}_n(x)\|_{\text{F}} + \sup_{x \in B_\mu(L)} \|\tilde{A}_n(x)\|_{\text{F}} \\
 &\stackrel{(i)}{\leq} \frac{\text{polylog}(n)}{\sqrt{n}}
 \end{aligned}$$

Here (i) follows from Lemma 40 and Lemma 41

Proof of (91): apply the triangle inequality to see that

$$\begin{aligned}
 &\sup_{\substack{x \in B_\mu(L) \\ S \in \mathcal{S}_d((2M_L)^{-1}, 2M_L)}} \|\hat{\Psi}_\rho(x, S)\| \\
 &\leq \sup_{\substack{x \in B_\mu(L) \\ S \in \mathcal{S}_d((2M_L)^{-1}, 2M_L)}} \|\hat{\Psi}_\rho(x, S) - \tilde{\Psi}_n(x, S)\| + \sup_{\substack{x \in B_\mu(L) \\ S \in \mathcal{S}_d((2M_L)^{-1}, 2M_L)}} \|\tilde{\Psi}_n(x, S) - \mathbb{E}\tilde{\Psi}_n(x, S)\| \\
 &\quad + \sup_{\substack{x \in B_\mu(L) \\ S \in \mathcal{S}_d((2M_L)^{-1}, 2M_L)}} \|\mathbb{E}\tilde{\Psi}_n(x, S)\| \\
 &\stackrel{(i)}{\leq} \frac{\text{polylog}(n)}{\sqrt{n}} + \frac{\text{polylog}(n)}{\sqrt{n}} + \text{polylog}(n) \\
 &\leq \text{polylog}(n)
 \end{aligned}$$

Here (i) follows from Lemma 40 and Lemma 41.

Proof of (92), (93): By the remark after Assumption 5 and the eigenvalue stability inequality, one can obtain

$$\begin{aligned}
 & \inf_{x \in B_{\mu(L)}} \lambda_{\min} \left(-\widehat{\Phi}_\rho(x) \right) \\
 & \geq \inf_{x \in B_{\mu(L)}} \lambda_{\min} \left(-\mathbb{E} \widetilde{\Phi}_n(x) \right) - \sup_{x \in B_{\mu(L)}} \left\| \widetilde{\Phi}_n(x) - \mathbb{E} \widetilde{\Phi}_n(x) \right\| - \sup_{x \in B_{\mu(L)}} \left\| \widetilde{\Phi}_n(x) - \widehat{\Phi}_\rho(x) \right\| \\
 & \stackrel{(i)}{\geq} \frac{1}{\text{polylog}(n)} - \frac{\text{polylog}(n)}{\sqrt{n}} - \frac{\text{polylog}(n)}{\sqrt{n}} \\
 & \geq \frac{1}{\text{polylog}(n)}
 \end{aligned}$$

Here (i) follows by noticing that $\mathbb{E}(-\widetilde{\Phi}_n(x)) = \mathbb{E} \left(-w(x, X) dT_{Q^*(x)}^Q \right)$, Assumption 5 and applying Lemma 40, Lemma 41.

By the remark after Assumption 5 and (92), one then has

$$\left\| -\widehat{\Phi}_\rho^{-1}(x) \right\| = \lambda_{\max} \left(-\widehat{\Phi}_\rho^{-1}(x) \right) \leq \frac{1}{\lambda_{\min} \left(-\widehat{\Phi}_\rho(x) \right)}$$

Then (93) follows from (92).

C.8 Proof of Lemma 43

Recall definitions of $Q^*(x)$ and $\widehat{Q}_\rho(x)$ that

$$\begin{aligned}
 Q^*(x) &= \operatorname{argmin}_{S \in \mathcal{S}_d^{++}} \mathbb{E} [w(x, X) W^2(S, Q)] \\
 \widehat{Q}_\rho(x) &= \operatorname{argmin}_{S \in \mathcal{S}_d^{++}} \frac{1}{n} \sum_{i=1}^n w_{n,\rho}(x, X_i) W^2(S, Q_i)
 \end{aligned}$$

First, the differential properties of W^2 (Lemma 21) imply the following optimality conditions for $Q^*(x)$ and $\widehat{Q}_\rho(x)$.

$$\mathbb{E} w(x, X) \left(T_{Q^*(x)}^Q - I_d \right) = 0 \tag{147}$$

$$\frac{1}{n} \sum_{i=1}^n w_{n,\rho}(x, X_i) \left(T_{\widehat{Q}_\rho(x)}^{Q_i} - I_d \right) = 0 \tag{148}$$

Next, one can apply Lemma 21 again to get the 2nd order Taylor expansion of (148) around $Q^*(x)$ as follows.

$$\begin{aligned}
 0 &= \frac{1}{n} \sum_{i=1}^n w_{n,\rho}(x, X_i) \left(T_{Q^*(x)}^{Q_i} - I_d \right) + \frac{1}{n} \sum_{i=1}^n w_{n,\rho}(x, X_i) dT_{Q^*(x)}^{Q_i} \cdot \left(\widehat{Q}_\rho(x) - Q^*(x) \right) \\
 &\quad + \frac{1}{2n} \sum_{i=1}^n w_{n,\rho}(x, X_i) d^2 T_{Q^*(x)}^{Q_i} \cdot \left(\widehat{Q}_\rho(x) - Q^*(x) \right)^{\otimes 2}
 \end{aligned}$$

where $\tilde{Q}_n(x)$ lies between $Q^*(x)$ and $\hat{Q}_\rho(x)$. Rearranging then gives

$$\begin{aligned} \hat{Q}_\rho(x) - Q^*(x) &= \left(-\frac{1}{n} \sum_{i=1}^n w_{n,\rho}(x, X_i) dT_{Q^*(x)}^{Q_i} \right)^{-1} \cdot \underbrace{\frac{1}{n} \sum_{i=1}^n w_{n,\rho}(x, X_i) (T_{Q^*(x)}^{Q_i} - I_d)}_{=\hat{A}_\rho(x)} \\ &+ \left(-\frac{1}{n} \sum_{i=1}^n w_{n,\rho}(x, X_i) dT_{Q^*(x)}^{Q_i} \right)^{-1} \cdot \underbrace{\frac{1}{2} \cdot \frac{1}{n} \sum_{i=1}^n w_{n,\rho}(x, X_i) d^2 T_{\tilde{Q}_n(x)}^{Q_i}}_{=\hat{\Psi}_\rho(x; \tilde{Q}_n(x))} \cdot (\hat{Q}_\rho(x) - Q^*(x))^{\otimes 2} \end{aligned} \quad (149)$$

Taking Frobenius norm on both sides of (149) gives the following quadratic inequality for $\|\hat{Q}_\rho(x) - Q^*(x)\|_F$.

$$\begin{aligned} &\|\hat{Q}_\rho(x) - Q^*(x)\|_F \\ &\leq \left\| -\hat{\Phi}_\rho(x)^{-1} \right\| \cdot \|\hat{A}_\rho(x)\|_F + \frac{1}{2} \left\| -\hat{\Phi}_\rho(x)^{-1} \right\| \cdot \left\| \hat{\Psi}_\rho(x, \tilde{Q}_n(x)) \right\| \cdot \|\hat{Q}_\rho(x) - Q^*(x)\|_F^2 \end{aligned} \quad (150)$$

With Lemma 42 in place, one can then derive from (150) that the following quadratic inequality holds uniformly for any $x \in B_\mu(L)$ under \tilde{E}_2 .

$$\|\hat{Q}_\rho(x) - Q^*(x)\|_F \leq a_0 + a_2 \|\hat{Q}_\rho(x) - Q^*(x)\|_F^2$$

Here $a_0, a_2 > 0$ are uniform for any $x \in B_\mu(L)$ and satisfies

$$\begin{aligned} a_0 &= \tilde{C}'_{\tau,1} \frac{\text{polylog}(n)}{\sqrt{n}} \\ a_2 &= \tilde{C}'_{\tau,2} \text{polylog}(n) \text{polylog}(n) \end{aligned}$$

for constants $\tilde{C}'_{\tau,1}, \tilde{C}'_{\tau,2} > 0$ independent of n .

Therefore, taking the supremum over x gives

$$\sup_{x \in B_\mu(L)} \|\hat{Q}_\rho(x) - Q^*(x)\|_F \leq a_0 + a_2 \sup_{x \in B_\mu(L)} \|\hat{Q}_\rho(x) - Q^*(x)\|_F^2$$

Solving the above quadratic inequality for $\sup_{x \in B_\mu(L)} \|\hat{Q}_\rho(x) - Q^*(x)\|_F$ then gives

$$\sup_{x \in B_\mu(L)} \|\hat{Q}_\rho(x) - Q^*(x)\|_F \in \left[0, \frac{1 - \sqrt{1 - 4a_0a_2}}{2a_2} \right] \cup \left[\frac{1 + \sqrt{1 - 4a_0a_2}}{2a_2}, +\infty \right)$$

Since $\tilde{E}_2 \subset \tilde{E}_1$ and the slow rate of convergence (Lemma 38) implies that only the smaller branch should be retained. Therefore, one has

$$\begin{aligned} \sup_{x \in B_\mu(L)} \left\| \hat{Q}_\rho(x) - Q^*(x) \right\|_F &\leq \frac{1 - \sqrt{1 - 4a_0a_2}}{2a_2} \\ &= \frac{4a_0a_2}{2a_2(1 + \sqrt{1 - 4a_0a_2})} \\ &\leq C_{\text{fast},\tau} \frac{\text{polylog}(n)}{\sqrt{n}} \end{aligned}$$

for some constant $C_{\text{fast},\tau}$ independent of n .

The proof of Lemma 43 is then complete.

C.9 Proof of Lemma 44

Fix $\tau = 10$. Denote

$$\begin{aligned} \Delta_{n,\rho} &= \sup_{x \in B_\mu(L)} \left\| \hat{Q}_{n,\rho}(x) - Q^*(x) \right\|_F \\ \tilde{E}_\rho &= \left\{ \sup_{x \in B_\mu(L)} \left\| \hat{Q}_{n,\rho}(x) \right\|_{\text{op}} \leq n^2 \tilde{M}_L \right\}, \end{aligned}$$

where $\tilde{M}_L = (C_B \vee c_b)^2 L^2 L_\tau^{4+(4\vee 2C_1)} \geq t_0^2$ is defined in Lemma 36. By definition, one has $\tilde{E}_2 \subset \tilde{E}_0 \subset \tilde{E}_\rho$. Hence one can obtain the following decomposition

$$\begin{aligned} \mathbb{E} \Delta_{n,\rho} &= \mathbb{E} \Delta_{n,\rho} \mathbf{1} \left\{ \tilde{E}_2 \right\} + \mathbb{E} \Delta_{n,\rho} \mathbf{1} \left\{ \tilde{E}_2^c \cap \tilde{E}_\rho \right\} + \mathbb{E} \Delta_{n,\rho} \mathbf{1} \left\{ \tilde{E}_\rho^c \right\} \\ &\leq C_{\text{fast},\tau} \frac{\text{polylog}(n)}{\sqrt{n}} + \mathbb{E} \Delta_{n,\rho} \mathbf{1} \left\{ \tilde{E}_2^c \cap \tilde{E}_\rho \right\} + \mathbb{E} \Delta_{n,\rho} \mathbf{1} \left\{ \tilde{E}_\rho^c \right\} \end{aligned}$$

- Under event $\tilde{E}_2^c \cap \tilde{E}_\rho$, one has

$$\begin{aligned} \Delta_{n,\rho} &\lesssim \sup_{x \in B_\mu(L)} \left\| \hat{Q}_\rho(x) - Q^*(x) \right\|_{\text{op}} \\ &\leq \sup_{x \in B_\mu(L)} \left[\left\| \hat{Q}_\rho(x) \right\|_{\text{op}} \vee \|Q^*(x)\|_{\text{op}} \right] \\ &\stackrel{(i)}{\lesssim} n^2 \tilde{M}_L \end{aligned}$$

Here (i) follows from Lemma 33, 36 and the definition of \tilde{E}_ρ . Hence

$$\begin{aligned} \mathbb{E} \Delta_{n,\rho} \mathbf{1} \left\{ \tilde{E}_2^c \cap \tilde{E}_\rho \right\} &\lesssim n^2 \tilde{M}_L \mathbb{P}(\tilde{E}_2^c) \\ &\lesssim \frac{\text{polylog}(n)}{n^{\tau-2}} \end{aligned}$$

- Under event \tilde{E}_ρ^c , one has

$$\begin{aligned}\Delta_{n,\rho} &\lesssim \sup_{x \in B_\mu(L)} \left[\left\| \hat{Q}_\rho(x) \right\|_{\text{op}} + \|Q^*(x)\|_{\text{op}} \right] \\ &\lesssim \sup_{x \in B_\mu(L)} \left\| \hat{Q}_\rho(x) \right\|_{\text{op}}\end{aligned}$$

Lemma 36 implies that under \tilde{E}_ρ^c , for any $t \geq n^2 \widetilde{M}_L$,

$$\begin{aligned}\mathbb{P} \left\{ \sup_{x \in B_\mu(L)} \left\| \hat{Q}_\rho(x) \right\|_{\text{op}} \geq t \right\} &\lesssim h_0(\sqrt{t}/n) \\ &\lesssim n \cdot \exp \left[- \left(\frac{\sqrt{t}}{c_b L n} \right)^{2/(C_\Lambda \vee 2 + 2)} \right] \\ &= n \cdot \exp \left[- \left(\frac{t}{c_b^2 L^2 n^2} \right)^{1/(C_\Lambda \vee 2 + 2)} \right]\end{aligned}$$

Hence

$$\begin{aligned}\mathbb{E} \Delta_{n,\rho} \mathbf{1} \left\{ \tilde{E}_\rho^c \right\} &\lesssim \mathbb{E} \sup_{x \in B_\mu(L)} \left\| \hat{Q}_\rho(x) \right\|_{\text{op}} \mathbf{1} \left\{ \sup_{x \in B_\mu(L)} \left\| \hat{Q}_\rho(x) \right\|_{\text{op}} > n^2 \widetilde{M}_L \right\} \\ &\lesssim \int_{n^2 \widetilde{M}_L}^\infty n t \exp \left[- \left(\frac{t}{c_b^2 L^2 n^2} \right)^{1/(C_\Lambda \vee 2 + 2)} \right] dt \\ &= \int_{\widetilde{M}_L}^\infty n^5 t \exp \left[- \left(\frac{t}{c_b^2 L^2} \right)^{1/(C_\Lambda \vee 2 + 2)} \right] dt \\ &\stackrel{(i)}{\lesssim} \int_{\widetilde{M}_L}^\infty n^5 t \left\{ \exp \left[- \left(\frac{\widetilde{M}_L}{c_b^2 L^2} \right)^{1/(C_\Lambda \vee 2 + 2)} \right] - \left(\frac{t}{c_b^2 L^2} \right)^{1/(C_\Lambda \vee 2 + 2)} \right\} dt \\ &\lesssim \int_{\widetilde{M}_L}^\infty n^5 t \left\{ n^{-(1+\tau)} \exp \left[- \left(\frac{t}{c_b^2 L^2} \right)^{1/(C_\Lambda \vee 2 + 2)} \right] \right\} dt \\ &\lesssim n^{4-\tau} L^4 \int_{\widetilde{M}_L/(c_b L^2)}^\infty t \exp \left[- t^{1/(C_\Lambda \vee 2 + 2)} \right] dt \\ &\lesssim n^{4-\tau} L^4\end{aligned}$$

Here (i) follows from the inequality

$$(a+b)^\alpha \geq (a^\alpha + b^\alpha)/2^{1-\alpha}, \quad a, b \geq 0, 0 < \alpha \leq 1$$

Finally, it follows by combining the results above that for $\tau \geq 9/2$,

$$\mathbb{E} \Delta_{n,\rho} \leq \frac{\text{polylog}(n)}{\sqrt{n}}$$

This finishes the proof of Lemma 44.

C.10 Proof of Lemma 45

Let $\tau > 9/2$. Lemma 21 (iv) implies that under \tilde{E}_2 , one also has

$$\sup_{x \in B_\mu(L)} W\left(\hat{Q}_\rho(x), Q^*(x)\right) \leq \frac{\text{polylog}(n)}{\sqrt{n}}$$

The proof then follows similar steps to those in Lemma 44. Specifically, denote

$$\tilde{\Delta}_{n,\rho} = \sup_{x \in B_\mu(L)} W\left(\hat{Q}_\rho(x), Q^*(x)\right)$$

Then one can obtain

$$\begin{aligned} \mathbb{E}\tilde{\Delta}_{n,\rho} &= \mathbb{E}\tilde{\Delta}_{n,\rho} \mathbf{1}\{\tilde{E}_2\} + \mathbb{E}\tilde{\Delta}_{n,\rho} \mathbf{1}\{\tilde{E}_2^c \cap \tilde{E}_\rho\} + \mathbb{E}\tilde{\Delta}_{n,\rho} \mathbf{1}\{\tilde{E}_\rho^c\} \\ &\leq \frac{\text{polylog}(n)}{\sqrt{n}} + \mathbb{E}\tilde{\Delta}_{n,\rho} \mathbf{1}\{\tilde{E}_2^c \cap \tilde{E}_\rho\} + \mathbb{E}\tilde{\Delta}_{n,\rho} \mathbf{1}\{\tilde{E}_\rho^c\} \end{aligned}$$

- Under event $\tilde{E}_2^c \cap \tilde{E}_\rho$, Lemma 21 (i) implies that

$$\begin{aligned} \tilde{\Delta}_{n,\rho} &\lesssim \sup_{x \in B_\mu(L)} \left[\left\| \hat{Q}_\rho(x) \right\|_{\text{op}} \vee \|Q^*(x)\|_{\text{op}} \right] \\ &\stackrel{(i)}{\lesssim} n^2 \tilde{M}_L \end{aligned}$$

Here (i) follows from Lemma 33, 36 and the definition of \tilde{E}_ρ . Hence

$$\begin{aligned} \mathbb{E}\tilde{\Delta}_{n,\rho} \mathbf{1}\{\tilde{E}_2^c \cap \tilde{E}_\rho\} &\leq n^2 \tilde{M}_L \mathbb{P}(\tilde{E}_2^c) \\ &\leq \frac{\text{polylog}(n)}{n^{\tau-2}} \end{aligned}$$

- Under event \tilde{E}_ρ^c , one can similarly obtain that

$$\tilde{\Delta}_{n,\rho} \lesssim \sup_{x \in B_\mu(L)} \left\| \hat{Q}_\rho(x) \right\|_{\text{op}}$$

and

$$\begin{aligned} \mathbb{E}\tilde{\Delta}_{n,\rho} \mathbf{1}\{\tilde{E}_\rho^c\} &\lesssim \mathbb{E} \sup_{x \in B_\mu(L)} \left\| \hat{Q}_\rho(x) \right\|_{\text{op}} \mathbf{1}\left\{ \sup_{x \in B_\mu(L)} \left\| \hat{Q}_\rho(x) \right\|_{\text{op}} > n^2 \tilde{M}_L \right\} \\ &\lesssim n^{-(\tau-4)} \end{aligned}$$

Finally, one can arrive at the conclusion that for $\tau \geq 9/2$,

$$\mathbb{E}\tilde{\Delta}_{n,\rho} \leq \frac{\text{polylog}(n)}{\sqrt{n}}$$

This finishes the proof of Lemma 45.

Appendix D. Proof of Theorem 11

To simplify notation, we fix x and write Q^* for $Q^*(x)$, $\hat{Q}_{n,\rho}$ for $\hat{Q}_{n,\rho}(x)$ when there is no ambiguity.

Following the same argument as (149) in Appendix C, one can obtain

$$\mathbb{E}_{(X,Q) \sim \mathbb{P}} w(x, X) \left(T_{Q^*}^Q - I_d \right) = 0 \quad (151)$$

$$\begin{aligned} \sqrt{n} \left(\hat{Q}_{n,\rho}(x) - Q^*(x) \right) &= \underbrace{\left(-\frac{1}{n} \sum_{i=1}^n w_{n,\rho}(x, X_i) dT_{Q^*(x)}^{Q_i} \right)^{-1}}_{=\hat{\Phi}_\rho(x)} \cdot \underbrace{\frac{1}{\sqrt{n}} \sum_{i=1}^n w_{n,\rho}(x, X_i) \left(T_{Q^*(x)}^{Q_i} - I_d \right)}_{=:a_n(x)} \\ &+ \underbrace{\left(-\frac{1}{n} \sum_{i=1}^n w_{n,\rho}(x, X_i) dT_{Q^*(x)}^{Q_i} \right)^{-1} \cdot \frac{\sqrt{n}}{2} \cdot \frac{1}{n} \sum_{i=1}^n w_{n,\rho}(x, X_i) d^2 T_{\hat{Q}_{n,\rho}(x)}^{Q_i} \cdot \left(\hat{Q}_{n,\rho}(x) - Q^*(x) \right)^{\otimes 2}}_{=:b_n(x)} \end{aligned} \quad (152)$$

Analysis of $\hat{\Phi}_\rho(x)$: Lemma 40 and 41 together imply that for any fixed x ,

$$\hat{\Phi}_\rho(x) \xrightarrow{p} \mathbb{E} \left[-w(x, X) dT_{Q^*(x)}^Q \right]$$

Then by Assumption 5, one has

$$\left[\hat{\Phi}_\rho(x) \right]^{-1} \xrightarrow{p} \left[\mathbb{E} \left(-w(x, X) dT_{Q^*(x)}^Q \right) \right]^{-1} \quad (153)$$

Analysis of $b_n(x)$: Theorem 7, Lemma 40 and 41 together imply that for any fixed x , with probability at least $1 - O(n^{-100})$,

$$\begin{aligned} |b_n(x)| &\leq \frac{\sqrt{n}}{2} \cdot \left\| \frac{1}{n} \sum_{i=1}^n w_{n,\rho}(x, X_i) d^2 T_{\hat{Q}_{n,\rho}(x)}^{Q_i} \right\| \cdot \left\| \hat{Q}_{n,\rho}(x) - Q^*(x) \right\|_F^2 \\ &\leq \frac{\sqrt{n}}{2} \cdot \text{polylog}(n) \cdot \frac{\text{polylog}(n)}{n} \\ &= \frac{\text{polylog}(n)}{\sqrt{n}} \end{aligned} \quad (154)$$

Analysis of $a_n(x)$: by Lemma 26, one has $w(x, X) = \vec{x}^\top \vec{\Sigma}^{-1} \vec{X}$ and $w_{n,\rho}(x, X) = \vec{x}^\top \widehat{\vec{\Sigma}}_\rho^{-1} \vec{X}$. Hence one can obtain

$$\begin{aligned}
 a_n(x) &= \frac{1}{\sqrt{n}} \sum_{i=1}^n \vec{x}^\top \widehat{\vec{\Sigma}}_\rho^{-1} \vec{X}_i \left(T_{Q^*}^Q - I_d \right) \\
 &= \left(\vec{x}^\top \widehat{\vec{\Sigma}}_\rho^{-1} \otimes I_d \right) \cdot \frac{1}{\sqrt{n}} \sum_{i=1}^n \vec{X}_i \otimes \left(T_{Q^*}^Q - I_d \right) \\
 &= \underbrace{\left(\vec{x}^\top \vec{\Sigma}^{-1} \otimes I_d \right) \cdot \frac{1}{\sqrt{n}} \sum_{i=1}^n \vec{X}_i \otimes \left(T_{Q^*}^Q - I_d \right)}_{=: a_{n,1}(x)} \\
 &\quad + \underbrace{\sqrt{n} \left(\vec{x}^\top (\widehat{\vec{\Sigma}}_\rho^{-1} - \vec{\Sigma}^{-1}) \otimes I_d \right) \cdot \mathbb{E} \vec{X} \otimes \left(T_{Q^*}^Q - I_d \right)}_{=: a_{n,2}(x)} \\
 &\quad + \underbrace{\sqrt{n} \left(\vec{x}^\top (\widehat{\vec{\Sigma}}_\rho^{-1} - \vec{\Sigma}^{-1}) \otimes I_d \right) \cdot \left[\frac{1}{n} \sum_{i=1}^n \vec{X}_i \otimes \left(T_{Q^*}^Q - I_d \right) - \mathbb{E} \vec{X} \otimes \left(T_{Q^*}^Q - I_d \right) \right]}_{=: a_{n,3}(x)}
 \end{aligned} \tag{155}$$

- $a_{n,1}(x)$: by (151) and Lemma 26, one can obtain

$$\mathbb{E} a_{n,1}(x) = 0 \tag{156}$$

To see this, it suffices to show $\mathbb{E} \left(\vec{x}^\top \vec{\Sigma}^{-1} \otimes I_d \right) \cdot \left(\vec{X} \otimes (T_{Q^*}^Q - I_d) \right) = 0$. Direct computation shows that the LHS of the equality is equal to

$$\mathbb{E} \vec{x}^\top \vec{\Sigma}^{-1} \vec{X} \left(T_{Q^*}^Q - I_d \right) = \mathbb{E} w(x, X) (T_{Q^*}^Q - I_d) = 0$$

Here the equality follows from the optimality condition (151).

- $a_{n,2}(x)$: using the formula $A^{-1} - B^{-1} = A^{-1}(B - A)B^{-1}$, one can obtain

$$\begin{aligned}
 a_{n,2}(x) &= \sqrt{n} \left(\vec{x}^\top \widehat{\vec{\Sigma}}_\rho^{-1} (\vec{\Sigma} - \widehat{\vec{\Sigma}}_\rho^{-1}) \vec{\Sigma}^{-1} \otimes I_d \right) \cdot \mathbb{E} \vec{X} \otimes \left(T_{Q^*}^Q - I_d \right) \\
 &= \sqrt{n} \left(\vec{x}^\top \vec{\Sigma}^{-1} (\vec{\Sigma} - \widehat{\vec{\Sigma}}_\rho) \vec{\Sigma}^{-1} \otimes I_d \right) \cdot \mathbb{E} \vec{X} \otimes \left(T_{Q^*}^Q - I_d \right) \\
 &\quad - \sqrt{n} \left(\vec{x}^\top (\vec{\Sigma}^{-1} - \widehat{\vec{\Sigma}}_\rho^{-1}) (\vec{\Sigma} - \widehat{\vec{\Sigma}}_\rho^{-1}) \vec{\Sigma}^{-1} \otimes I_d \right) \cdot \mathbb{E} \vec{X} \otimes \left(T_{Q^*}^Q - I_d \right) \\
 &= -\frac{1}{\sqrt{n}} \left(\sum_{i=1}^n \vec{x}^\top \vec{\Sigma}^{-1} (\vec{X}_i \vec{X}_i^\top - \vec{\Sigma}) \vec{\Sigma}^{-1} \otimes I_d \right) \cdot \left(\mathbb{E} \vec{X} \otimes (T_{Q^*}^Q - I_d) \right) + O_p \left(\frac{1}{\sqrt{n}} \right)
 \end{aligned} \tag{157}$$

Here the last line follows from the fact that $\widehat{\vec{\Sigma}}_\rho - \vec{\Sigma} = O_p(n^{-1/2})$ which is itself due to the sub-Gaussianity of X .

- $a_{n,3}(x)$: note that $\left[\frac{1}{n} \sum_{i=1}^n \vec{X}_i \otimes (T_{Q^*}^Q - I_d) - \mathbb{E} \vec{X} \otimes (T_{Q^*}^Q - I_d)\right] = O_p(n^{-1/2})$ again by the sub-Gaussianity of X and Assumption 2, one can arrive at

$$a_{n,3}(x) = O_p(n^{-1/2}) \quad (158)$$

Combining (155), (156), (157), (158) and the functional central limit theorem (Hsing and Eubank, 2015, Theorem 7.7.6), one can obtain

$$\begin{aligned} a_n(x) &= \frac{1}{\sqrt{n}} \sum_{i=1}^n \left[\left(\vec{x}^\top \vec{\Sigma}^{-1} \otimes I_d \right) \cdot \left(\vec{X}_i \otimes (T_{Q^*}^Q - I_d) \right) - \right. \\ &\quad \left. \left(\vec{x}^\top \vec{\Sigma}^{-1} (\vec{X}_i \vec{X}_i^\top - \vec{\Sigma}) \otimes I_d \right) \cdot \left(\mathbb{E} \vec{X} \otimes (T_{Q^*}^Q - I_d) \right) \right] + o_p(1) \\ &\xrightarrow{w} Z_x \end{aligned} \quad (159)$$

where $Z_x \sim \mathcal{N}(0, \Xi_x)$ is a Gaussian random matrix with mean 0 and covariance Ξ_x . Here $\Xi_x := \mathbb{E} V_x \otimes V_x$ with

$$V_x = w(x, X) \left(T_{Q^*}^Q - I_d \right) - \left(\vec{x}^\top \vec{\Sigma}^{-1} (\vec{X} \vec{X}^\top - \vec{\Sigma}) \otimes I_d \right) \cdot \left(\mathbb{E} \vec{X} \otimes (T_{Q^*}^Q - I_d) \right)$$

Finally, (152), (153), (154) and (159) together imply

$$\sqrt{n} \left(\hat{Q}_{n,\rho} - Q^* \right) \xrightarrow{w} \left(-\mathbb{E} w(x, X) dT_{Q^*}^Q \right)^{-1} \cdot Z_x$$

which completes the proof.

Appendix E. Proof of Corollary 14

The optimality condition for $Q^*(x)$ implies that

$$\mathbb{E} w(x, X) \left(T_{Q^*(x)}^Q - I_d \right) = 0$$

Note that independence between X and Q implies that

$$\begin{aligned} \mathbb{E} w(x, X) \left(T_{Q^*(x)}^Q - I_d \right) &= \mathbb{E} w(x, X) \mathbb{E} \left(T_{Q^*(x)}^Q - I_d \right) \\ &\stackrel{(i)}{=} \mathbb{E} \left(T_{Q^*(x)}^Q - I_d \right) \end{aligned}$$

Here (i) follows from the fact that $\mathbb{E} w(x, X) = 1$. Therefore, one has

$$\mathbb{E} \left(T_{Q^*(x)}^Q - I_d \right) = 0$$

By independence, the above equality then implies

$$\mathbb{E} \vec{X} \otimes (T_{Q^*}^Q - I_d) = 0$$

Hence one has $V_{x,2} = 0$, the proof is then complete.

Appendix F. Proof of Theorem 16

First, we demonstrate in Lemma 46 below uniform fast convergence under Assumption 1, 2, 5, 6 and the null hypothesis. In order to apply Theorem 7, it suffices to verify Assumption 4 and Assumption 5. Note that Assumption 6 and the null hypothesis implies the independence between X and Q , then both Assumption 4 and 5 are consequences of Lemma 34.

Lemma 46 *Instate the assumptions in Theorem 16. Then with probability at least $1 - O(n^{-100})$, one has*

$$\begin{aligned} \sup_{x \in B_\mu(L_n)} \left\| \widehat{Q}_\rho(x) - Q^*(x) \right\|_F &\leq \frac{\text{polylog}(n)}{\sqrt{n}} \\ \sup_{1 \leq i \leq n} \left\| \widehat{Q}_\rho(X_i) - Q^*(X_i) \right\|_F &\leq \frac{\text{polylog}(n)}{\sqrt{n}} \end{aligned}$$

With Lemma 46 in place, we consider the Taylor expansion of the optimality condition for $\widehat{Q}_\rho(x)$. Recall that the optimality condition for $\widehat{Q}_\rho(x)$ gives

$$\frac{1}{n} \sum_{i=1}^n w_{n,\rho}(x, X_i) \left(T_{\widehat{Q}_\rho(x)}^{Q_i} - I_d \right) = 0$$

Then one can apply Lemma 21 to get the following 2nd order Taylor expansion around $Q^*(x)$.

$$0 = \frac{1}{n} \sum_{i=1}^n w_{n,\rho}(x, X_i) \left(T_{Q^*(x)}^{Q_i} - I_d \right) + \frac{1}{n} \sum_{i=1}^n w_{n,\rho}(x, X_i) dT_{Q^*(x)}^{Q_i} \cdot \left(\widehat{Q}_\rho(x) - Q^*(x) \right) + R_2(x) \quad (160)$$

where $R_2(x)$ is the 2nd order remainder term with some $\widetilde{Q}_n(x)$ lying between $Q^*(x)$ and $\widehat{Q}_\rho(x)$ defined as follows.

$$R_2(x) := \frac{1}{2n} \sum_{i=1}^n w_{n,\rho}(x, X_i) d^2 T_{\widetilde{Q}_n(x)}^{Q_i} \cdot \left(\widehat{Q}_\rho(x) - Q^*(x) \right)^{\otimes 2}$$

Under the null hypothesis (21), one has $Q^*(x) \equiv Q^*$ and (160) reduces to

$$0 = \frac{1}{n} \sum_{i=1}^n w_{n,\rho}(x, X_i) \left(T_{Q^*}^{Q_i} - I_d \right) + \frac{1}{n} \sum_{i=1}^n w_{n,\rho}(x, X_i) dT_{Q^*}^{Q_i} \cdot \left(\widehat{Q}_\rho(x) - Q^*(x) \right) + R_2(x) \quad (161)$$

Setting $x = \bar{X}$ in (161) then gives

$$0 = \frac{1}{n} \sum_{i=1}^n \left(T_{Q^*}^{Q_i} - I_d \right) + \frac{1}{n} \sum_{i=1}^n dT_{Q^*}^{Q_i} \cdot \left(\widehat{Q}_\rho(x) - Q^*(x) \right) + R_2(x) \quad (162)$$

Take difference between (161) and (162) and rearrange, one can obtain

$$\begin{aligned}
 & \underbrace{\left(-\frac{1}{n} \sum_{i=1}^n dT_{Q^*}^{Q_i}\right) \cdot (\hat{Q}_\rho(x) - \hat{Q}_\rho(\bar{X}))}_{=:\tau(x)} \\
 &= \underbrace{\frac{1}{n} \sum_{i=1}^n (w_{n,\rho}(x, X_i) - 1)(T_{Q^*}^{Q_i} - I_d)}_{=:\alpha_0(x)} + \underbrace{\frac{1}{n} \sum_{i=1}^n (w_{n,\rho}(x, X_i) - 1) dT_{Q^*}^{Q_i} \cdot (\hat{Q}_\rho(x) - \hat{Q}_\rho(\bar{X}))}_{=:\alpha_1(x)} \\
 &+ \underbrace{R_2(x) - R_2(\bar{X})}_{=:\alpha_2(x)}
 \end{aligned} \tag{163}$$

Before delving further into the proof, we pause to introduce more notations. Define

$$\hat{\tau}(x) = \left(-\frac{1}{n} \sum_{i=1}^n dT_{\hat{Q}_\rho(\bar{X})}^{Q_i}\right) \cdot (\hat{Q}_\rho(x) - \hat{Q}_\rho(\bar{X}))$$

Then by definition, the test statistic $\hat{\mathcal{T}}_\rho = \sum_{i=1}^n \|\hat{\tau}(X_i)\|_F^2$, and we have the following decomposition of $\tau(x)$.

$$\begin{aligned}
 \hat{\tau}(x) &= \tau(x) + \hat{\tau}(x) - \tau(x) \\
 &= \tau(x) + \underbrace{\left(\frac{1}{n} \sum_{i=1}^n dT_{Q^*}^{Q_i} - \frac{1}{n} \sum_{i=1}^n dT_{\hat{Q}_\rho(\bar{X})}^{Q_i}\right) \cdot (\hat{Q}_\rho(x) - \hat{Q}_\rho(\bar{X}))}_{=:\alpha_3(x)} \\
 &= \alpha_0(x) + \underbrace{\alpha_1(x) + \alpha_2(x) + \alpha_3(x)}_{=:R(x)}
 \end{aligned}$$

We also define counterparts of $\alpha_0, \dots, \alpha_2$ and R_2 by replacing $w_{n,\rho}(\cdot, \cdot)$ with $w(\cdot, \cdot)$ as follows.

$$\begin{aligned}
 \tilde{\alpha}_0(x) &:= \frac{1}{n} \sum_{i=1}^n (w(x, X_i) - 1)(T_{Q^*}^{Q_i} - I_d) \\
 \tilde{\alpha}_1(x) &:= \frac{1}{n} \sum_{i=1}^n (w(x, X_i) - 1) dT_{Q^*}^{Q_i} \cdot (\hat{Q}_\rho(x) - \hat{Q}_\rho(\bar{X})) \\
 \tilde{\alpha}_2(x) &:= \tilde{R}_2(x) - \tilde{R}_2(\bar{X})
 \end{aligned}$$

where

$$\tilde{R}_2(x) = \frac{1}{2n} \sum_{i=1}^n w(x, X_i) d^2 T_{\hat{Q}_n(x)}^{Q_i} \cdot (\hat{Q}_\rho(x) - Q^*)^{\otimes 2}$$

With the above notation in place, one can obtain

$$\hat{\mathcal{T}}_\rho = \sum_{k=1}^n \|\alpha_0(X_k)\|_F^2 + \underbrace{2 \sum_{k=1}^n \langle \alpha_0(X_k), R(X_k) \rangle + \sum_{k=1}^n \|R(X_k)\|_F^2}_{\text{Rem}_n} \tag{164}$$

The proof is then divided into three steps.

- First, we give upper bounds for $\tilde{\alpha}_i$ as well as $\tilde{\alpha}_i - \alpha_i$ and their consequences.
- Next, we show that the remainder term Rem_n is negligible.
- Then, we demonstrate that $\sum_i \|\alpha_0(X_i)\|_F^2$ converges weakly to the desired asymptotic null distribution (24).

Analysis of $\tilde{\alpha}$ and $\tilde{\alpha} - \alpha$: We give uniform upper bounds for $\tilde{\alpha}_i(x) - \alpha_i(x)$ for $i = 0, 1, 2$ in Lemma 47 as uniform upper bounds for $\tilde{\alpha}_i(x)$ and $\alpha_3(x)$ in Lemma 48; see Appendix F.1, F.2 for the proof.

Lemma 47 *Instate the notations and assumptions in Theorem 16.*

$$\sup_{x \in B_\mu(L)} \|\alpha_0(x) - \tilde{\alpha}_0(x)\|_F \leq \frac{\text{polylog}(n)}{\sqrt{n}} \quad (165)$$

$$\sup_{x \in B_\mu(L)} \|\alpha_1(x) - \tilde{\alpha}_1(x)\|_F \leq \frac{\text{polylog}(n)}{n} \quad (166)$$

$$\sup_{x \in B_\mu(L)} \|\alpha_2(x) - \tilde{\alpha}_2(x)\|_F \leq \frac{\text{polylog}(n)}{n^{3/2}} \quad (167)$$

with probability at least $1 - O(n^{-99})$.

Lemma 48 *Instate the notations and assumptions in Theorem 16.*

$$\sup_{x \in B_\mu(L)} \|\tilde{\alpha}_0(x)\|_F \leq \frac{\text{polylog}(n)}{\sqrt{n}} \quad (168)$$

$$\sup_{x \in B_\mu(L)} \|\tilde{\alpha}_1(x)\|_F \leq \frac{\text{polylog}(n)}{n} \quad (169)$$

$$\sup_{x \in B_\mu(L)} \|\tilde{\alpha}_2(x)\|_F \leq \frac{\text{polylog}(n)}{n} \quad (170)$$

$$\sup_{x \in B_\mu(L)} \|\alpha_3(x)\|_F \leq \frac{\text{polylog}(n)}{n} \quad (171)$$

with probability at least $1 - O(n^{-99})$.

With the above lemmas in place, one can readily obtain that with probability at least $1 - O(n^{-99})$,

$$\sup_{i \in [n]} \|\alpha_0(X_i)\|_F \leq \sup_{i \in [n]} \|\tilde{\alpha}_0(X_i) - \alpha_0(X_i)\|_F + \sup_{i \in [n]} \|\tilde{\alpha}_0(X_i)\|_F \leq \frac{\text{polylog}(n)}{\sqrt{n}} \quad (172)$$

$$\sup_{i \in [n]} \|\alpha_1(X_i)\|_F \leq \sup_{i \in [n]} \|\tilde{\alpha}_1(X_i) - \alpha_1(X_i)\|_F + \sup_{i \in [n]} \|\tilde{\alpha}_1(X_i)\|_F \leq \frac{\text{polylog}(n)}{n} \quad (173)$$

$$\sup_{i \in [n]} \|\alpha_2(X_i)\|_F \leq \sup_{i \in [n]} \|\tilde{\alpha}_2(X_i) - \alpha_2(X_i)\|_F + \sup_{i \in [n]} \|\tilde{\alpha}_2(X_i)\|_F \leq \frac{\text{polylog}(n)}{n} \quad (174)$$

which then implies

$$\sup_{i \in [n]} \|R(X_i)\|_F \leq \sum_{k=1}^3 \sup_{i \in [n]} \|\alpha_k(X_i)\|_F \leq \frac{\text{polylog}(n)}{n} \quad (175)$$

Negligibility of Rem_n : We consider two terms $\sum_{k=1}^n \langle \alpha_0(X_k), R(X_k) \rangle$ and $\sum_{k=1}^n \|R(X_k)\|_F^2$ separately.

Analysis of $\sum_{k=1}^n \langle \alpha_0(X_k), R(X_k) \rangle$: By (172) and (175), one has with probability at least $1 - O(n^{-99})$,

$$\begin{aligned} \sum_{k=1}^n \langle \alpha_0(X_k), R(X_k) \rangle &\leq \sum_{k=1}^n \|\alpha_0(X_k)\|_F \cdot \|R(X_k)\|_F \\ &\leq \sum_{k=1}^n \frac{\text{polylog}(n)}{n^{3/2}} \\ &= \frac{\text{polylog}(n)}{\sqrt{n}} \end{aligned}$$

Analysis of $\sum_{k=1}^n \|R(X_k)\|_F^2$: Similarly, by (175), one has with probability at least $1 - O(n^{-99})$,

$$\sum_{k=1}^n \|R(X_k)\|_F^2 \leq \frac{\text{polylog}(n)}{n}$$

Therefore, the above results imply that with probability at least $1 - O(n^{-99})$,

$$\text{Rem}_n \leq \frac{\text{polylog}(n)}{\sqrt{n}} \quad (176)$$

Analysis of α_0 : To consider the main term $\sum_{i=1}^n \|\alpha_0(X_i)\|_F^2$, one has

$$\begin{aligned} &\sum_{k=1}^n \|\alpha_0(X_k)\|_F^2 \\ &= \frac{1}{n^2} \sum_{k=1}^n \left\langle \sum_{i=1}^n (X_k - \bar{X})^\top \hat{\Sigma}_\rho^{-1} (X_i - \bar{X}) (T_{Q^*}^{Q_i} - I_d), \sum_{j=1}^n (X_k - \bar{X})^\top \hat{\Sigma}_\rho^{-1} (X_j - \bar{X}) (T_{Q^*}^{Q_j} - I_d) \right\rangle \\ &= \frac{1}{n^2} \sum_{i,j=1}^n \sum_{k=1}^n (X_i - \bar{X})^\top \hat{\Sigma}_\rho^{-1} (X_k - \bar{X}) (X_k - \bar{X})^\top \hat{\Sigma}_\rho^{-1} (X_j - \bar{X}) \langle T_{Q^*}^{Q_i} - I_d, T_{Q^*}^{Q_j} - I_d \rangle \\ &= \frac{1}{n} \sum_{i,j=1}^n (X_i - \bar{X})^\top \hat{\Sigma}_\rho^{-1} (X_j - \bar{X}) \langle T_{Q^*}^{Q_i} - I_d, T_{Q^*}^{Q_j} - I_d \rangle \\ &= \left\| \frac{1}{\sqrt{n}} \sum_{i=1}^n \hat{\Sigma}_\rho^{-1/2} (X_i - \bar{X}) \otimes (T_{Q^*}^{Q_i} - I_d) \right\|_F^2 \end{aligned}$$

Note that

$$\begin{aligned}
 & \frac{1}{\sqrt{n}} \sum_{i=1}^n \widehat{\Sigma}_\rho^{-1/2} (X_i - \bar{X}) \otimes (T_{Q^*}^{Q_i} - I_d) \\
 &= \left[(\widehat{\Sigma}_\rho^{-1/2} \Sigma^{1/2}) \otimes I_d \right] \cdot \left[\frac{1}{\sqrt{n}} \sum_{i=1}^n \Sigma^{-1/2} (X_i - \bar{X}) \otimes (T_{Q^*}^{Q_i} - I_d) \right] \\
 &= (1 + o_p(1)) \frac{1}{\sqrt{n}} \sum_{i=1}^n \Sigma^{-1/2} (X_i - \bar{X}) \otimes (T_{Q^*}^{Q_i} - I_d)
 \end{aligned} \tag{177}$$

Also, one can obtain

$$\begin{aligned}
 & \left(\frac{1}{\sqrt{n}} \sum_{i=1}^n \Sigma^{-1/2} (X_i - \bar{X}) \otimes (T_{Q^*}^{Q_i} - I_d) \right) - \left(\frac{1}{\sqrt{n}} \sum_{i=1}^n \Sigma^{-1/2} (X_i - \mu) \otimes (T_{Q^*}^{Q_i} - I_d) \right) \\
 &= \frac{1}{\sqrt{n}} \sum_{i=1}^n \Sigma^{-1/2} (\mu - \bar{X}) \otimes (T_{Q^*}^{Q_i} - I_d) \\
 &= \Sigma^{-1/2} (\mu - \bar{X}) \otimes \left[\frac{1}{\sqrt{n}} \sum_{i=1}^n (T_{Q^*}^{Q_i} - I_d) \right] \\
 &= o_p(1)
 \end{aligned} \tag{178}$$

Here the last line follows since $\bar{X} - \mu = o_p(1)$ and $\frac{1}{\sqrt{n}} \sum_{i=1}^n (T_{Q^*}^{Q_i} - I_d)$ is asymptotically normal with zero mean. The zero mean is justified in the following claim whose proof is deferred to Appendix F.3.

Claim 4 *Under the null hypothesis (21) and Assumption 4, 6, X and Q are independent and one has*

$$\mathbb{E}(X - \mu) \otimes (T_{Q^*}^Q - I_d) = 0 \tag{179}$$

$$\mathbb{E}(T_{Q^*}^Q - I_d) = 0 \tag{180}$$

Claim 4 also implies that $\mathbb{E} \Sigma^{-1/2} (X - \mu) \otimes (T_{Q^*}^Q - I_d) = 0$. Therefore, by the functional central limit theorem (Hsing and Eubank, 2015, Theorem 7.7.6), one can obtain

$$\frac{1}{\sqrt{n}} \sum_{i=1}^n \Sigma^{-1/2} (X_i - \mu) \otimes (T_{Q^*}^{Q_i} - I_d) \xrightarrow{w} \mathcal{N} \left(0, I_p \otimes \mathbb{E} \left[(T_{Q^*}^Q - I_d) \otimes (T_{Q^*}^Q - I_d) \right] \right) \tag{181}$$

Combining (177), (178) and (181), we arrive at

$$\frac{1}{\sqrt{n}} \sum_{i=1}^n \widehat{\Sigma}_\rho^{-1/2} (X_i - \bar{X}) \otimes (T_{Q^*}^{Q_i} - I_d) \xrightarrow{w} \mathcal{N} \left(0, I_p \otimes \mathbb{E} \left[(T_{Q^*}^Q - I_d) \otimes (T_{Q^*}^Q - I_d) \right] \right)$$

which then implies that

$$\sum_{k=1}^n \|\alpha_0(X_k)\|^2 \xrightarrow{w} \sum_i \lambda_i w_i \tag{182}$$

where w_i are i.i.d. χ_p^2 random variables and λ_i are the eigenvalues of $\mathbb{E} \left[(T_{Q_*}^Q - I_d) \otimes (T_{Q_*}^Q - I_d) \right]$.

Finally, taking (164) (176) and (182) collectively yields

$$\widehat{\mathcal{T}}_\rho \xrightarrow{w} \sum_i \lambda_i w_i$$

F.1 Proof of Lemma 47

Apply Lemma 46 and triangle inequality to see that under the null (21), with probability at least $1 - O(n^{-100})$, one has

$$\begin{aligned} \sup_{x \in B_\mu(L)} \left\| \widehat{Q}_\rho(x) - Q^* \right\|_F &\leq \frac{\text{polylog}(n)}{\sqrt{n}} \\ \sup_{x \in B_\mu(L)} \left\| \widehat{Q}_\rho(x) - \widehat{Q}_\rho(\bar{X}) \right\|_F &\leq \sup_{x \in B_\mu(L)} \left\| \widehat{Q}_\rho(x) - Q^* \right\|_F + \sup_{x \in B_\mu(L)} \left\| \widehat{Q}_\rho(\bar{X}) - Q^* \right\|_F \leq \frac{\text{polylog}(n)}{\sqrt{n}} \end{aligned} \quad (183)$$

Proof of (165): note that

$$\begin{aligned} \alpha_0(x) - \widetilde{\alpha}_0(x) &= \frac{1}{n} \sum_{i=1}^n (w_{n,\rho}(x, X_i) - w(x, X_i)) (T_{Q_*}^{Q_i} - I_d) \\ &= A_n(x) - \widetilde{A}_n(x) \end{aligned}$$

Hence (165) follows from Lemma 41.

Proof of (166): one can obtain

$$\begin{aligned} \|\alpha_1(x) - \widetilde{\alpha}_1(x)\|_F &= \left\| \frac{1}{n} \sum_{i=1}^n (w_{n,\rho}(x, X_i) - w(x, X_i)) dT_{Q_*}^{Q_i} \cdot (\widehat{Q}_\rho(x) - \widehat{Q}_\rho(\bar{X})) \right\|_F \\ &\leq \left\| \frac{1}{n} \sum_{i=1}^n (w_{n,\rho}(x, X_i) - w(x, X_i)) dT_{Q_*}^{Q_i} \right\| \cdot \left\| \widehat{Q}_\rho(x) - \widehat{Q}_\rho(\bar{X}) \right\|_F \\ &= \left\| \Phi_n(x) - \widetilde{\Phi}_n(x) \right\| \cdot \left\| \widehat{Q}_\rho(x) - \widehat{Q}_\rho(\bar{X}) \right\|_F \end{aligned}$$

Hence (166) follows from Lemma 41 and (183).

Proof of (167): one can obtain

$$\begin{aligned} \|\alpha_2(x) - \widetilde{\alpha}_2(x)\|_F &= \left\| \frac{1}{n} \sum_{i=1}^n (w_{n,\rho}(x, X_i) - w(x, X_i)) d^2 T_{\widetilde{Q}_n(x)}^{Q_i} \cdot (\widehat{Q}_\rho(x) - Q^*)^{\otimes 2} \right\|_F \\ &\leq \left\| \frac{1}{n} \sum_{i=1}^n (w_{n,\rho}(x, X_i) - w(x, X_i)) d^2 T_{\widetilde{Q}_n(x)}^{Q_i} \right\| \cdot \left\| \widehat{Q}_\rho(x) - \widehat{Q}_\rho(\bar{X}) \right\|_F^2 \\ &= \left\| \Psi_n(x, \widetilde{Q}_n(x)) - \widetilde{\Psi}_n(x, \widetilde{Q}_n(x)) \right\| \cdot \left\| \widehat{Q}_\rho(x) - \widehat{Q}_\rho(\bar{X}) \right\|_F^2 \end{aligned}$$

Hence (167) follows from Lemma 41 and (183).

F.2 Proof of Lemma 48

Apply Lemma 46 and triangle inequality to see that under the null (21), with probability at least $1 - O(n^{-100})$, one has

$$\begin{aligned} \sup_{x \in B_\mu(L)} \left\| \widehat{Q}_\rho(x) - Q^* \right\|_F &\leq \frac{\text{polylog}(n)}{\sqrt{n}} \\ \sup_{x \in B_\mu(L)} \left\| \widehat{Q}_\rho(x) - \widehat{Q}_\rho(\bar{X}) \right\|_F &\leq \sup_{x \in B_\mu(L)} \left\| \widehat{Q}_\rho(x) - Q^* \right\|_F + \sup_{x \in B_\mu(L)} \left\| \widehat{Q}_\rho(\bar{X}) - Q^* \right\|_F \leq \frac{\text{polylog}(n)}{\sqrt{n}} \end{aligned} \quad (184)$$

Proof of (168): Recall the definition of $\widetilde{A}_n(x)$ in (77), one can see that (168) follows from (83) in Lemma 40 with a slight and straightforward modification to accommodate the $w(x, X) - 1$ term here. For brevity, we omit the proof.

Proof of (169):

$$\|\widetilde{\alpha}_1\|_F \leq \left\| \frac{1}{n} \sum_{i=1}^n (w(x, X_i) - 1) dT_{Q^*}^{Q_i} \right\| \cdot \left\| \widehat{Q}_\rho(x) - \widehat{Q}_\rho(\bar{X}) \right\|_F \quad (185)$$

Recall the definition of $\widetilde{\Phi}_n(x)$ in (77), with a slight modification of the proof of (84) in Lemma 40, one can obtain

$$\sup_{x \in B_\mu(L)} \left\| \frac{1}{n} \sum_{i=1}^n (w(x, X_i) - 1) dT_{Q^*}^{Q_i} \right\| \leq \frac{\text{polylog}(n)}{\sqrt{n}} \quad (186)$$

For brevity, the proof is omitted.

Combining (184), (185) and (186) gives (169).

Proof of (170): Apply triangle inequality to see that

$$\begin{aligned} \sup_{x \in B_\mu(L)} \|\widetilde{\alpha}_2(x)\|_F &\leq \left\| \widetilde{R}_2(\bar{X}) \right\|_F + \sup_{x \in B_\mu(L)} \left\| \widetilde{R}_2(x) \right\|_F \\ &\leq 2 \sup_{x \in B_\mu(L)} \left\| \widetilde{R}_2(x) \right\|_F \end{aligned}$$

Moreover, one can obtain

$$\begin{aligned} \sup_{x \in B_\mu(L)} \left\| \widetilde{R}_2(x) \right\|_F &\lesssim \sup_{x \in B_\mu(L)} \left\| \frac{1}{n} \sum_{i=1}^n w(x, X_i) d^2 T_{\widehat{Q}_n(x)}^{Q_i} \right\| \cdot \left\| \widehat{Q}_\rho(x) - Q^* \right\|_F^2 \\ &\leq \sup_{\substack{x \in B_\mu(L) \\ S \in \mathcal{S}_d((2M_L)^{-1}, 2M_L)}} \left\| \widetilde{\Psi}_n(x, S) \right\| \cdot \sup_{x \in B_\mu(L)} \left\| \widehat{Q}_\rho(x) - Q^* \right\|_F^2 \\ &\stackrel{(i)}{\leq} \frac{\text{polylog}(n)}{n} \end{aligned}$$

Here (i) follows from (85), (86) in Lemma 40 as well as (184).

Proof of (171): one can obtain

$$\begin{aligned} \sup_{x \in B_\mu(L)} \|\alpha_3(x)\|_F &= \sup_{x \in B_\mu(L)} \left\| \left(\frac{1}{n} \sum_{i=1}^n dT_{Q^*}^{Q_i} - \frac{1}{n} \sum_{i=1}^n dT_{\hat{Q}_\rho(\bar{X})}^{Q_i} \right) \cdot (\hat{Q}_\rho(x) - \hat{Q}_\rho(\bar{X})) \right\|_F \\ &\leq \left\| \frac{1}{n} \sum_{i=1}^n dT_{Q^*}^{Q_i} - \frac{1}{n} \sum_{i=1}^n dT_{\hat{Q}_\rho(\bar{X})}^{Q_i} \right\| \cdot \sup_{x \in B_\mu(L)} \|\hat{Q}_\rho(x) - \hat{Q}_\rho(\bar{X})\|_F \quad (187) \end{aligned}$$

Moreover, one has

$$\begin{aligned} \left\| \frac{1}{n} \sum_{i=1}^n dT_{Q^*}^{Q_i} - \frac{1}{n} \sum_{i=1}^n dT_{\hat{Q}_\rho(\bar{X})}^{Q_i} \right\| &\stackrel{(i)}{\leq} \frac{1}{n} \sum_{i=1}^n \left\| dT_{Q^*}^{Q_i} - dT_{\hat{Q}_\rho(\bar{X})}^{Q_i} \right\| \\ &\stackrel{(ii)}{\leq} \frac{1}{n} \sum_{i=1}^n \left\| d^2 T_{Q'_{i,n}}^{Q_i} \cdot (\hat{Q}_\rho(\bar{X}) - Q^*) \right\| \\ &\stackrel{(iii)}{\leq} \frac{1}{n} \sum_{i=1}^n \left\| d^2 T_{Q'_{i,n}}^{Q_i} \right\| \cdot \|\hat{Q}_\rho(\bar{X}) - Q^*\|_F \quad (188) \end{aligned}$$

Here (i) follows from triangle inequality, (ii) from the mean value theorem (Dudley and Norvaiša, 2011, Theorem 5.3) for some $Q'_{i,n}$ that lies on the segment between Q^* and $\hat{Q}_\rho(\bar{X})$, and (iii) arises from Lemma 22. Therefore, with a slight modification of the proof of (86) in Lemma 40, one can obtain with probability at least $1 - O(n^{-100})$,

$$\frac{1}{n} \sum_{i=1}^n \left\| d^2 T_{Q'_{i,n}}^{Q_i} \right\| \leq \text{polylog}(n) \quad (189)$$

For brevity, the proof is omitted.

Finally, combining (187), (188) and (189) gives

$$\begin{aligned} \sup_{x \in B_\mu(L)} \|\alpha_3(x)\|_F &\lesssim \left[\frac{1}{n} \sum_{i=1}^n \left\| d^2 T_{Q'_{i,n}}^{Q_i} \right\| \right] \cdot \left[\sup_{x \in B_\mu(L)} \|\hat{Q}_\rho(x) - Q^*(x)\|_F \right]^2 \\ &\leq \frac{\text{polylog}(n)}{n} \end{aligned}$$

which finishes the proof for (171).

F.3 Proof of Claim 4

The independence follows directly from Assumption 6 and the null hypothesis (21).

Proof of (179): by independence, we have

$$\begin{aligned} \mathbb{E}(X - \mu) \otimes (T_{Q^*}^Q - I_d) &= [\mathbb{E}(X - \mu)] \otimes [\mathbb{E}(T_{Q^*}^Q - I_d)] \\ &= 0 \end{aligned}$$

Proof of (180): The optimality condition of $Q^*(x)$ gives that

$$\mathbb{E}w(x, X) \left(T_{Q^*(x)}^Q - I_d \right) = 0$$

By independence, one then has

$$\begin{aligned} \mathbb{E} \left(T_{Q^*(x)}^Q - I_d \right) &\stackrel{(i)}{=} [\mathbb{E}w(x, X)] \cdot \left[\mathbb{E}w(x, X) \left(T_{Q^*(x)}^Q - I_d \right) \right] \\ &= \mathbb{E}w(x, X) \left(T_{Q^*(x)}^Q - I_d \right) \\ &= 0 \end{aligned}$$

Here (i) follows from the fact that $\mathbb{E}w(x, X) \equiv 1$.

Appendix G. Proof of Proposition 1

Note that under the null (21), it holds that $Q^*(\bar{X}) = Q^*$. Then Theorem 7 implies that with probability at least $1 - O(n^{-100})$,

$$\left\| \hat{Q}_\rho(\bar{X}) - Q^* \right\|_F \leq \frac{\text{polylog}(n)}{\sqrt{n}}$$

Therefore, one has with probability at least $1 - O(n^{-100})$,

$$\hat{Q}_\rho(\bar{X}) \in \mathcal{S}_d((2c_1)^{-1}, 2c_1) \quad (190)$$

for n large enough. Here we recall that $c_1 \geq 1$ is a constant defined in Assumption 2.

One can apply the triangle inequality to see that

$$\begin{aligned} &\left\| \frac{1}{n} \sum_{i=1}^n \left(T_{\hat{Q}_\rho(\bar{X})}^{Q_i} - I_d \right) \otimes \left(T_{\hat{Q}_n(\bar{X})}^{Q_i} - I_d \right) - \mathbb{E} \left(T_{Q^*}^Q - I_d \right) \otimes \left(T_{Q^*}^Q - I_d \right) \right\|_F \\ &\leq \left\| \frac{1}{n} \sum_{i=1}^n \left(T_{\hat{Q}_\rho(\bar{X})}^{Q_i} - I_d \right) \otimes \left(T_{\hat{Q}_n(\bar{X})}^{Q_i} - I_d \right) - \mathbb{E} \left(T_{\hat{Q}_\rho(\bar{X})}^Q - I_d \right) \otimes \left(T_{\hat{Q}_n(\bar{X})}^Q - I_d \right) \right\|_F \\ &\quad + \left\| \mathbb{E} \left(T_{\hat{Q}_\rho(\bar{X})}^Q - I_d \right) \otimes \left(T_{\hat{Q}_n(\bar{X})}^Q - I_d \right) - \mathbb{E} \left(T_{Q^*}^Q - I_d \right) \otimes \left(T_{Q^*}^Q - I_d \right) \right\|_F \\ &\stackrel{(i)}{\leq} \underbrace{\sup_{S \in \mathcal{S}_d((2c_1)^{-1}, 2c_1)} \left\| \frac{1}{n} \sum_{i=1}^n \left(T_S^{Q_i} - I_d \right) \otimes \left(T_S^{Q_i} - I_d \right) - \mathbb{E} \left(T_S^Q - I_d \right) \otimes \left(T_S^Q - I_d \right) \right\|_F}_{\zeta_1} \\ &\quad + \underbrace{\sup_{S: \|S - Q^*\|_F \leq \text{polylog}(n)/\sqrt{n}} \left\| \mathbb{E} \left(T_S^Q - I_d \right) \otimes \left(T_S^Q - I_d \right) - \mathbb{E} \left(T_{Q^*}^Q - I_d \right) \otimes \left(T_{Q^*}^Q - I_d \right) \right\|_F}_{\zeta_2} \end{aligned} \quad (191)$$

Here (i) follows from (190).

Upper bounds for ζ_1, ζ_2 in (191) are summarized in the lemma below whose proof is deferred to Appendix G.1. Note that Lemma 49 do not assume the null hypothesis so that it can be reused later for the proof of the power (Theorem 18).

Lemma 49 *Suppose Assumption 1-6 hold. Then with probability at least $1 - O(n^{-100})$,*

$$\zeta_1 \leq \frac{\text{polylog}(n)}{\sqrt{n}} \quad (192)$$

$$\zeta_2 \leq \frac{\text{polylog}(n)}{\sqrt{n}} \quad (193)$$

Lemma 49 combined with (191) then implies

$$\left\| \frac{1}{n} \sum_{i=1}^n \left(T_{\hat{Q}_\rho(\bar{X})}^{Q_i} - I_d \right) \otimes \left(T_{\hat{Q}_n(\bar{X})}^{Q_i} - I_d \right) - \mathbb{E} \left(T_{Q^*}^Q - I_d \right) \otimes \left(T_{Q^*}^Q - I_d \right) \right\|_{\text{F}} \leq \frac{\text{polylog}(n)}{\sqrt{n}} \quad (194)$$

As a result, if $(\lambda_i)_{i \in [d^2]}$ are sorted in order, then one has $\hat{\lambda}_i \rightarrow \lambda_i$ uniformly for $i \in [d^2]$ in probability which further implies that

$$\sum_{i=1}^{d^2} \hat{\lambda}_i w_i \xrightarrow{p} \sum_{i=1}^{d^2} \lambda_i w_i$$

Then the continuous mapping theorem implies that $\hat{q}_{1-\alpha} \rightarrow q_{1-\alpha}$ in probability. Then one can obtain

$$\mathbb{P} \left(\hat{\mathcal{T}}_\rho > \hat{q}_{1-\alpha} \right) \leq \mathbb{P} \left(\hat{\mathcal{T}}_\rho > q_{1-\alpha} - \epsilon \right) + \mathbb{P} \left(|\hat{q}_{1-\alpha} - q_{1-\alpha}| > \epsilon \right)$$

Taking the limit as $n \rightarrow \infty$, followed by letting $\epsilon \rightarrow 0$ to get that

$$\limsup_{n \rightarrow \infty} \mathbb{P} \left(\hat{\mathcal{T}}_\rho > \hat{q}_{1-\alpha} \right) \leq \alpha \quad (195)$$

A similar lower bound shows

$$\liminf_{n \rightarrow \infty} \mathbb{P} \left(\hat{\mathcal{T}}_\rho > \hat{q}_\alpha \right) \geq \alpha \quad (196)$$

Finally, combining (195) and (196) completes the proof.

G.1 Proof of Lemma 49

Proof of (192): the proof is similar to Lemma 40, and is hence omitted for brevity.

Proof of (193): First, recall differential properties (Lemma 21) that

$$\begin{aligned} \|T_S^Q\|_{\text{op}} &\leq \lambda_{\max}(Q) \cdot \lambda_{\min}(Q)^{-1/2} \cdot \lambda_{\min}(S)^{-1/2} \\ \|dT_S^Q\| &\leq \frac{1}{2} \lambda_{\max}(Q)^{1/2} \cdot \lambda_{\min}(S)^{-2} \cdot \lambda_{\max}(S)^{1/2} \end{aligned} \quad (197)$$

Also, denote $\phi(Q, S) := \left(T_S^Q - I_d \right) \otimes \left(T_S^Q - I_d \right)$.

With these results in place, one can obtain

$$\begin{aligned}
 \|\phi(Q, S) - \phi(Q, Q^*)\|_F &\leq \left\| (T_S^Q - I_d) \otimes (T_S^Q - T_{Q^*}^Q) \right\|_F + \left\| (T_S^Q - T_{Q^*}^Q) \otimes (T_{Q^*}^Q - I_d) \right\|_F \\
 &\leq \|T_S^Q - I_d\|_F \cdot \|T_S^Q - T_{Q^*}^Q\|_F + \|T_S^Q - T_{Q^*}^Q\|_F \cdot \|T_{Q^*}^Q - I_d\|_F \\
 &\leq \|T_S^Q - I_d\|_F \cdot \|dT_{S'}^Q\| \cdot \|S - Q^*\|_F + \|T_{Q^*}^Q - I_d\|_F \cdot \|dT_{S'}^Q\| \cdot \|S - Q^*\|_F
 \end{aligned}$$

where S' lies between Q^* and S . Note that Assumption 2 and the condition $\|S - Q^*\|_F \leq \text{polylog}(n)/\sqrt{n}$ implies that

$$S, Q^* \in \mathcal{S}_d((2c_1)^{-1}, 2c_1)$$

Then for any $S \in \mathcal{S}_d((2c_1)^{-1}, 2c_1)$, one has

$$\begin{aligned}
 \mathbb{E} \|T_S^Q - I_d\|_F \cdot \|dT_{S'}^Q\| &\stackrel{(i)}{\lesssim} \mathbb{E} \left(\|X - \mu\|^{3C_1/2} + 1 \right) \cdot \|X - \mu\|^{C_1/2} \\
 &\stackrel{(ii)}{\lesssim} 1
 \end{aligned}$$

Here C_1 is defined in Assumption 2, (i) follows from (197) and (ii) is a result of the sub-Gaussianity of X . Similarly, one has

$$\mathbb{E} \|T_{Q^*}^Q - I_d\|_F \cdot \|dT_{S'}^Q\| \lesssim 1$$

Combining results above, one can obtain

$$\begin{aligned}
 \zeta_2 &\leq \sup_{S: \|S - Q^*\|_F \leq \text{polylog}(n)/\sqrt{n}} \mathbb{E}_Q \|\phi(Q, S) - \phi(Q, Q^*)\|_F \\
 &\lesssim \sup_{S: \|S - Q^*\|_F \leq \text{polylog}(n)/\sqrt{n}} 1 \cdot \|S - Q^*\|_F \\
 &\leq \frac{\text{polylog}(n)}{\sqrt{n}}
 \end{aligned}$$

The proof is then complete.

Appendix H. Proof of Theorem 18

As argued at the beginning of Appendix C.5, one can assume without loss of generality that $\|X - \mu\| \leq L$ almost surely with $L = C_L \sqrt{\log n}$ for some constant C_L large enough as in (78). Recall the notation that $M_L := \gamma_\Lambda(L)$.

Power under Frobenius norm: First we demonstrate concentration of various quantities of interest and derive their consequences in Lemma 50 below. The proof is in Appendix H.1.

Lemma 50 *Instate the notations and assumptions in Theorem 18. Assume in addition that $\|X - \mu\| \leq L$ almost surely. Then there exists an event E_n that satisfies $\mathbb{P}(E_n) \geq 1 - O(n^{-100})$ for any $\mathbb{P} \in \mathfrak{P}$, under which the following inequalities*

$$\|X_i - \mu\| \leq L, \quad Q_i \in \mathcal{S}_d(M_L^{-1}, M_L) \quad \text{for } i \in [n] \quad (198)$$

$$\sup_{x \in B_\mu(L)} \left\| \widehat{Q}_\rho(x) - Q^*(x) \right\|_F \leq \frac{\text{polylog}(n)}{\sqrt{n}} \quad (199)$$

$$\left\{ \widehat{Q}_\rho(x) : x \in B_\mu(L) \right\} \subset \mathcal{S}_d \left(\left(2^{1/6} M_L \right)^{-1}, 2^{1/6} M_L \right) \quad (200)$$

$$\sum_{i=1}^n \|Q^*(X_i) - Q^*(\mu)\|_F^2 \geq \frac{n a_n^2}{2} \quad (201)$$

$$\left\| \widehat{Q}_\rho(\bar{X}) - Q^*(\mu) \right\|_F \leq \frac{\text{polylog}(n)}{\sqrt{n}} \quad (202)$$

$$|\widehat{\lambda}_i| \leq 2\lambda_1, \quad i \in [d^2] \quad (203)$$

hold for n large enough.

Next, note that ((iii)) in Lemma 21 implies that for any $Q, S \in \mathcal{S}_d \left((2^{1/6} M_L)^{-1}, 2^{1/6} M_L \right)$ (here $2^{1/6}$ is chosen only for technical computation), one has

$$\frac{1}{2\sqrt{2}M_L^3} \leq \lambda_{\min} \left(-dT_S^Q \right) \leq \lambda_{\max} \left(-dT_S^Q \right) \leq \frac{\sqrt{2}}{2} M_L^3$$

which then implies that under E_n , the following holds.

$$\lambda_{\min} \left(\widehat{H} \right) \geq \frac{1}{n} \sum_{i=1}^n \lambda_{\min} \left(-dT_{\widehat{Q}_\rho(\bar{X})}^{Q_i} \right) \geq \frac{1}{2\sqrt{2}M_L^3}$$

Therefore, under E_n , one has

$$\widehat{\mathcal{T}}_\rho \geq \frac{1}{8M_L^6} \sum_{i=1}^n \left\| \widehat{Q}_\rho(X_i) - \widehat{Q}_\rho(\bar{X}) \right\|_F^2 \quad (204)$$

Then from the following decomposition

$$\widehat{Q}_\rho(X_i) - \widehat{Q}_\rho(\bar{X}) = Q^*(X_i) - Q^*(\mu) + \underbrace{\widehat{Q}_\rho(X_i) - Q^*(X_i) + Q^*(\mu) - \widehat{Q}_\rho(\bar{X})}_{\Delta_i}$$

one can obtain

$$\begin{aligned} & \sum_{i=1}^n \left\| \widehat{Q}_\rho(X_i) - \widehat{Q}_\rho(\bar{X}) \right\|_F^2 \\ & \stackrel{(i)}{=} \sum_{i=1}^n \|Q^*(X_i) - Q^*(\mu)\|_F^2 + 2 \sum_{i=1}^n \langle Q^*(X_i) - Q^*(\mu), \Delta_i \rangle + \sum_{i=1}^n \|\Delta_i\|_F^2 \\ & \stackrel{(ii)}{\geq} \sum_{i=1}^n \|Q^*(X_i) - Q^*(\mu)\|_F^2 - 2 \left(\sum_{i=1}^n \|Q^*(X_i) - Q^*(\mu)\|_F^2 \right)^{1/2} \cdot \left(\sum_{i=1}^n \|\Delta_i\|_F^2 \right)^{1/2} + \sum_{i=1}^n \|\Delta_i\|_F^2 \\ & \geq \left(\sum_{i=1}^n \|Q^*(X_i) - Q^*(\mu)\|_F^2 \right) \cdot \left(1 - 2 \left(\frac{\sum_{i=1}^n \|\Delta_i\|_F^2}{\sum_{i=1}^n \|Q^*(X_i) - Q^*(\mu)\|_F^2} \right)^{1/2} \right) \end{aligned} \quad (205)$$

Here (i) follows by developing the square, (ii) is a consequence of the Cauchy-Schwarz inequality.

Lemma 50 implies that under E_n , one has

$$\begin{aligned} \|\Delta_i\|_F &\leq \left\| \widehat{Q}_\rho(X_i) - Q^*(X_i) \right\|_F + \left\| Q^*(\mu) - \widehat{Q}_\rho(\bar{X}) \right\|_F \\ &\lesssim \frac{\text{polylog}(n)}{\sqrt{n}} \end{aligned} \quad (206)$$

Therefore, (205) and (206) together imply the following inequality for n large enough.

$$\sum_{i=1}^n \left\| \widehat{Q}_\rho(X_i) - \widehat{Q}_\rho(\bar{X}) \right\|_F^2 \stackrel{(I)}{\geq} \frac{na_n^2}{2} \cdot \frac{3}{4} \quad (207)$$

Here (I) arises due to Lemma 50, (206) as well as the fact that $\text{polylog}(n) = o(na_n^2)$.

Combining (204) and (207) then gives that under E_n , one has

$$\begin{aligned} \widehat{\mathcal{T}}_\rho &\geq \frac{3na_n^2}{64M_L^6} \\ &\geq \sqrt{na_n^2} \end{aligned} \quad (208)$$

for n large enough.

Denote $\widetilde{q}_{1-\alpha}$ the $1 - \alpha$ quantile of $\sum_{i=1}^{d^2} 2\lambda_1 w_i$, which is a fixed constant. Then as a consequence of Lemma 50, $\sum_{i=1}^{d^2} \widehat{\lambda}_i w_i$ is stochastically dominated by $\sum_{i=1}^{d^2} 2\lambda_1 w_i$ under E_n , which then implies that

$$\widehat{q}_{1-\alpha} < \widetilde{q}_{1-\alpha}, \quad \text{under } E_n$$

Therefore, for any $\mathbb{P} \in \mathfrak{P}$, one can obtain

$$\begin{aligned} \mathbb{P}(\widehat{\mathcal{T}}_\rho > \widehat{q}_{1-\alpha}) &\geq \mathbb{P}\left(\left\{\widehat{\mathcal{T}}_\rho > \widetilde{q}_{1-\alpha}\right\} \cap \left\{\widetilde{q}_{1-\alpha} > \widehat{q}_{1-\alpha}\right\}\right) \\ &\geq \mathbb{P}\left(\left\{\widehat{\mathcal{T}}_\rho > \widetilde{q}_{1-\alpha}\right\} \cap E_n\right) \\ &\stackrel{(i)}{\geq} \mathbb{P}\left(\left\{\sqrt{na_n^2} > \widetilde{q}_{1-\alpha}\right\} \cap E_n\right) \end{aligned}$$

for n large enough. Here (i) follows from (208).

Finally, one can take $n \rightarrow \infty$ to see that

$$\begin{aligned} \liminf_{n \rightarrow \infty} \inf_{\mathbb{P} \in \mathfrak{P}} \mathbb{P}(\widehat{\mathcal{T}}_\rho > \widehat{q}_{1-\alpha}) &\geq \liminf_{n \rightarrow \infty} \inf_{\mathbb{P} \in \mathfrak{P}} \mathbb{P}\left(\left\{\sqrt{na_n^2} > \widetilde{q}_{1-\alpha}\right\} \cap E_n\right) \\ &\geq 1 \end{aligned}$$

which implies

$$\lim_{n \rightarrow \infty} \inf_{\mathbb{P} \in \mathfrak{P}} \mathbb{P}(\widehat{\mathcal{T}}_\rho > \widehat{q}_{1-\alpha}) = 1$$

The proof is then complete.

Power under Wasserstein distance: by Lemma 21 and the boundedness assumption, one has

$$\mathbb{E} \|Q^*(X) - Q^*(\mu)\|_F^2 \geq \frac{1}{\text{polylog}(n)} \mathbb{E} W^2(Q^*(X), Q^*(\mu))$$

Therefore, the 1st part of the proof (Frobenius norm) can be applied.

H.1 Proof of Lemma 50

(198)-(200) are shown in the proof of Theorem 7 and (202) is due to Lemma 32.

Proof of (201): With these in place, we have almost surely that

$$\begin{aligned} \|Q^*(X) - Q^*(\mu)\|_F^2 &\leq d \|Q^*(X) - Q^*(\mu)\|_{\text{op}}^2 \\ &\leq d M_L^2 =: M_F \asymp \text{polylog}(n) \end{aligned} \quad (209)$$

Define random variable Y_i as follows.

$$Y_i := \|Q^*(X_i) - Q^*(\mu)\|_F^2 - \mathbb{E} \|Q^*(X) - Q^*(\mu)\|_F^2$$

With (209) in place, one then has $Y_i \in [-M_F, M_F]$, $\mathbb{E} Y_i = 0$ and

$$\begin{aligned} \text{Var } Y_i &= \mathbb{E} Y_i^2 \\ &\leq \mathbb{E} |Y_i| M_F \\ &\leq 2 \mathbb{E} \|Q^*(X_i) - Q^*(\mu)\|_F^2 \cdot M_F \\ &= 2 a_n^2 \cdot M_F \end{aligned}$$

To prove (201), it suffices to show that ,

$$\left| \sum_i^n Y_i \right| \leq \frac{n a_n^2}{2} \quad \text{with probability at least } 1 - O(n^{-100}) \quad (210)$$

To this end, note that by Bernstein's inequality (Vershynin, 2018, Theorem 2.8.4), one has with probability at least $1 - O(n^{-100})$,

$$\begin{aligned} \left| \sum_i^n Y_i \right| &\lesssim M_F \log n + \sqrt{n \text{Var } Y_i} \cdot \sqrt{\log n} \\ &\lesssim M_F \log n + \sqrt{n a_n^2 M_F} \cdot \sqrt{\log n} \\ &\stackrel{(i)}{\leq} \text{polylog}(n) \cdot \left(1 + \sqrt{n a_n^2}\right) \end{aligned} \quad (211)$$

Then (211) implies (210) for n large enough due to the assumption that $n a_n^2 \gtrsim n^{2\alpha_2}$ for some constant $\alpha_2 > 0$. The proof is then complete.

Proof of (203): Note that the proof of Lemma 49 does not assume the null hypothesis. Therefore, with (202) in place, one can exactly follow (190)-(194) as in the proof of Lemma 49 to get

$$\left\| \frac{1}{n} \sum_{i=1}^n \left(T_{\hat{Q}_\rho(\bar{X})}^{Q_i} - I_d \right) \otimes \left(T_{\hat{Q}_\rho(\bar{X})}^{Q_i} - I_d \right) - \mathbb{E} \left(T_{Q^*}^Q - I_d \right) \otimes \left(T_{Q^*}^Q - I_d \right) \right\|_F \leq \frac{\text{polylog}(n)}{\sqrt{n}}$$

which implies (203).

Appendix I. Additional simulations

We conduct simulations to compare different initialization methods and step sizes, as described in Section 5.1.

I.1 Initialization

To assess the performance of different initialization methods, we conduct simulations based on Example 2 with parameters $n = 200$, $p = 5$ and $d = 6$, setting $\eta = 1$, $T = 30$ and $\text{eps} = 10^{-6}$ in Algorithm 1. We consider three different initialization methods:

1. Identity Initialization : $S_0 = I_d$.
2. Random Initialization: $S_0 = U \exp(M) U^\top$ where U and M are independent; $U \in \mathcal{O}_d$ is a random orthogonal matrix following the Haar measure, and M is a random diagonal matrix with i.i.d. diagonal entries $M_{kk} \sim \mathcal{N}(0, 1/\sqrt{d})$, $k \in [d]$.
3. Mean Initialization: $S_0 = \sum_{i=1}^n Q_i$.

Table 1 reports the number of steps before termination and the relative error with respect to the identity initialization, averaged over 100 simulations. In each simulation, after generating $n = 200$ pairs of (X_i, Q_i) , we randomly generate $\tilde{X} \sim \text{Uniform}[-1, 1]^p$ and compute $\hat{Q}_\rho(\tilde{X})$ using Algorithm 1 with the three initialization methods described above. It can be observed

Table 1: Comparison of different initialization methods.

δ	Initialization method	Steps	Relative error ($\times \text{eps}$)
0	I_d	3	0
	Random	3.4	5.3e-4
	Mean	3	4.8e-6
0.2	I_d	3	0
	Random	3.4	5.9e-4
	Mean	3	4.8e-6
0.4	I_d	3	0
	Random	3.3	6.0e-4
	Mean	3	5.9e-6

that all three methods converge to the same point within nearly the same number of steps, as the relative error is significantly smaller than the threshold eps .

I.2 Step size

To examine the performance of different step sizes, we conduct simulations based on Example 2 with parameters $n = 200$, $p = 5$ and $d = 6$, setting $S_0 = I_d$, $T = 30$ and $\text{eps} = 10^{-6}$ in Algorithm 1. Table 2 reports the number of steps before termination and the relative error with respect to step size 1, averaged over 100 simulations. In each simulation, after generating $n = 200$ pairs of (X_i, Q_i) , we randomly generate $\tilde{X} \sim \text{Uniform}[-1, 1]^p$ and compute $\hat{Q}_\rho(\tilde{X})$ using Algorithm 1 with step size $\eta \in \{1, 0.9, 0.8, 0.7, 0.6\}$

Table 2: Comparison of different step sizes.

δ	η	Steps	Relative error ($\times \text{eps}$)
0	1	3	0
	0.9	8	0.013
	0.8	10	0.13
	0.7	13	0.20
	0.6	17	0.22
0.2	1	3	0
	0.9	8	0.013
	0.8	10	0.13
	0.7	13	0.20
	0.6	17	0.22
0.4	1	3	0
	0.9	7.94	0.017
	0.8	10	0.13
	0.7	13	0.20
	0.6	15.97	0.22

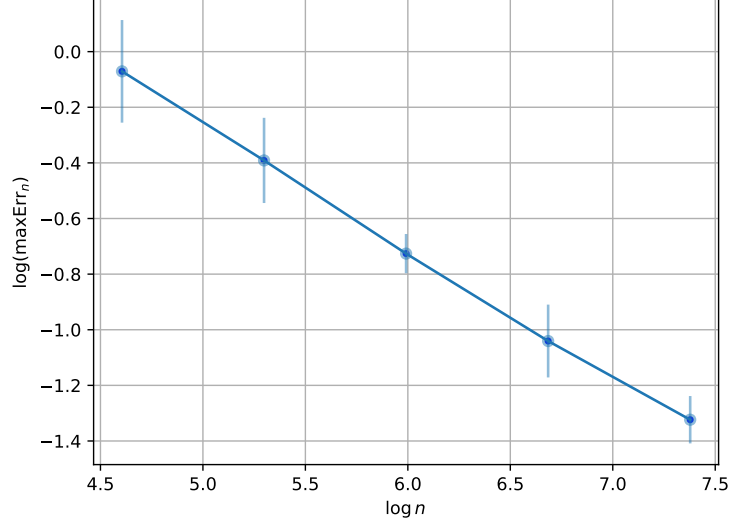
It can be observed that all three methods converge to the same point, as the relative error is smaller than the threshold eps . Additionally, setting $\eta = 1$ results in the fastest convergence.

I.3 Uniform convergence over a diverging domain

To investigate the convergence over a diverging domain, we slightly modify Example 2 by considering $X \sim \text{Uniform}[-\sqrt{\log n}, \sqrt{\log n}]^p$ with parameters $p = 5$, $d = 6$, $\delta = 0.1$ and $n \in \{100, 200, 400, 800, 1600\}$. For each n , we compute the maximum in-sample error

$$\text{maxErr}_n := \max_{i \in [n]} \left\| \hat{Q}_\rho(X_i) - Q^*(X_i) \right\|_F, \quad \rho = 1/n$$

Figure 7 plots $\log \overline{\text{maxErr}_n}$ against $\log n$ where $\overline{\text{maxErr}_n}$ represents the average of maxErr_n over 5 independent trials. The plot shows a linear trend with a slope -0.46 , providing empirical support for the parametric rate of convergence up to logarithmic factors.


 Figure 7: $\log \overline{\max \text{Err}_n}$ vs. $\log n$

References

- A. Abadie and G. W. Imbens. Large Sample Properties of Matching Estimators for Average Treatment Effects. *Econometrica*, 74(1):235–267, 2006.
- M. Agueh and G. Carlier. Barycenters in the Wasserstein Space. *SIAM Journal on Mathematical Analysis*, 43(2):904–924, 2011.
- M. Agueh and G. Carlier. Vers un théorème de la limite centrale dans l’espace de Wasserstein? *Comptes Rendus. Mathématique*, 355(7):812–818, 2017.
- J. Altschuler, J. Niles-Weed, and P. Rigollet. Near-linear time approximation algorithms for optimal transport via Sinkhorn iteration. *Advances in neural information processing systems*, 30, 2017.
- J. Altschuler, S. Chewi, P. R. Gerber, and A. J. Stromme. Averaging on the Bures-Wasserstein manifold: dimension-free convergence of gradient descent. *Advances in neural information processing systems*, 34:22132–22145, 2021.
- J. M. Altschuler, J. Niles-Weed, and A. J. Stromme. Asymptotics for Semidiscrete Entropic Optimal Transport. *SIAM Journal on Mathematical Analysis*, 54(2):1718–1741, 2022.
- L. Ambrosio, N. Gigli, and G. Savaré. *Gradient Flows In Metric Spaces and in the Space of Probability Measures*. Birkhäuser-Verlag, 2005.
- M. Arjovsky, S. Chintala, and L. Bottou. Wasserstein Generative Adversarial Networks. In *Proceedings of the 34th International Conference on Machine Learning*, pages 214–223. PMLR, 2017.

- M. Arnaudon, F. Barbaresco, and L. Yang. Riemannian medians and means with applications to radar signal processing. *IEEE Journal of Selected Topics in Signal Processing*, 7(4):595–604, 2013.
- V. Arsigny, P. Fillard, X. Pennec, and N. Ayache. Geometric means in a novel vector space structure on symmetric positive-definite matrices. *SIAM Journal on Matrix Analysis and Applications*, 29(1):328–347, 2007.
- E. d. Barrio and J.-M. Loubes. Central limit theorems for empirical transportation cost in general dimension. *The Annals of Probability*, 47(2):926–951, 2019.
- R. Bhatia, T. Jain, and Y. Lim. On the Bures–Wasserstein distance between positive definite matrices. *Expositiones Mathematicae*, 37(2):165–191, 2019.
- R. Bhattacharya and V. Patrangenaru. Large sample theory of intrinsic and extrinsic sample means on manifolds. *The Annals of Statistics*, 31(1):1–29, 2003.
- R. Bhattacharya and V. Patrangenaru. Large sample theory of intrinsic and extrinsic sample means on manifolds—II. *The Annals of Statistics*, 33(3):1225–1259, 2005.
- J. Bigot, R. Gouet, T. Klein, and A. López. Geodesic PCA in the Wasserstein space by convex PCA. *Annales de l’Institut Henri Poincaré, Probabilités et Statistiques*, 53(1):1 – 26, 2017.
- C. Bunne, L. Papaxanthos, A. Krause, and M. Cuturi. Proximal Optimal Transport Modeling of Population Dynamics. In *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics*, pages 6511–6528. PMLR, 2022.
- C. Bunne, Y.-P. Hsieh, M. Cuturi, and A. Krause. The Schrödinger Bridge between Gaussian Measures has a Closed Form. In *Proceedings of The 26th International Conference on Artificial Intelligence and Statistics*, pages 5802–5833. PMLR, 2023a.
- C. Bunne, S. G. Stark, G. Gut, J. S. del Castillo, M. Levesque, K.-V. Lehmann, L. Pelkmans, A. Krause, and G. Rätsch. Learning single-cell perturbation responses using neural optimal transport. *Nature Methods*, 20(11):1759–1768, 2023b.
- D. Bures. An extension of Kakutani’s theorem on infinite product measures to the tensor product of semifinite w^* -algebras. *Transactions of the American Mathematical Society*, 135(0):199–212, 1969.
- T. T. Cai, C.-H. Zhang, and H. H. Zhou. Optimal rates of convergence for covariance matrix estimation. *The Annals of Statistics*, 38(4):2118–2144, 2010.
- T. T. Cai, W. Liu, and H. H. Zhou. Estimating sparse precision matrix: Optimal rates of convergence and adaptive estimation. *The Annals of Statistics*, 44(2):455 – 488, 2016.
- M. Carlsson. Perturbation theory for the matrix square root and matrix modulus, 2018. arXiv:1810.01464.

- R. Caseiro, P. Martins, J. F. Henriques, and J. Batista. A nonparametric Riemannian framework on tensor field with application to foreground segmentation. *Pattern Recognition*, 45(11):3997–4017, 2012.
- Y. Chen and H.-G. Müller. Uniform convergence of local Fréchet regression with applications to locating extrema and time warping for metric space valued trajectories. *The Annals of Statistics*, 50(3):1573–1592, 2022.
- Y. Chen, Y. Chi, J. Fan, and C. Ma. Gradient descent with random initialization: fast global convergence for nonconvex phase retrieval. *Mathematical Programming*, 176(1):5–37, July 2019. ISSN 1436-4646.
- Y. Chen, Z. Lin, and H.-G. Müller. Wasserstein Regression. *Journal of the American Statistical Association*, 118(542):869–882, 2023.
- S. Chewi, T. Maunu, P. Rigollet, and A. J. Stromme. Gradient descent algorithms for Bures-Wasserstein barycenters. In *Proceedings of Thirty Third Conference on Learning Theory*, pages 1276–1304. PMLR, 2020.
- T. Y. M. Chiu, T. Leonard, and K.-W. Tsui. The Matrix-Logarithmic Covariance Model. *Journal of the American Statistical Association*, 91(433):198–210, 1996.
- E. Cornea, H. Zhu, P. Kim, and J. G. Ibrahim. Regression models on Riemannian symmetric spaces. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 79(2):463–482, 2017.
- M. Cuturi. Sinkhorn Distances: Lightspeed Computation of Optimal Transport. In C. J. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 26, 2013.
- S. Davinelli, D. Willcox, and G. Scapagnini. Extending healthy ageing: nutrient sensitive pathway and centenarian population. *Immunity & Ageing*, 9:9, 2012.
- C. de Lucia, T. Murphy, C. Steves, and et al. Lifestyle mediates the role of nutrient-sensing pathways in cognitive aging: cellular and epidemiological evidence. *Communications Biology*, 3:157, 2020.
- N. Deb and B. Sen. Multivariate Rank-Based Distribution-Free Nonparametric Testing Using Measure Transportation. *Journal of the American Statistical Association*, 118(541):192–207, 2023.
- E. del Barrio, A. G. Sanz, J.-M. Loubes, and J. Niles-Weed. An Improved Central Limit Theorem and Fast Convergence Rates for Entropic Transportation Costs. *SIAM Journal on Mathematics of Data Science*, 5(3):639–669, 2023.
- P. Del Moral and A. Niclas. A Taylor expansion of the square root matrix function. *Journal of Mathematical Analysis and Applications*, 465(1):259–266, 2018.
- I. L. Dryden, A. Koloydenko, and D. Zhou. Non-Euclidean statistics for covariance matrices, with applications to diffusion tensor imaging. *The Annals of Applied Statistics*, 3(3):1102–1123, 2009.

- R. M. Dudley and R. Norvaiša. *Concrete Functional Calculus*. Springer, New York, NY, 2011.
- Y. Elyahu, I. Hekselman, I. Eizenberg-Magar, O. Berner, I. Strominger, M. Schiller, K. Mittal, A. Nemirovsky, E. Eremenko, A. Vital, E. Simonovsky, V. Chalifa-Caspi, N. Friedman, E. Yeger-Lotem, and A. Monsonego. Aging promotes reorganization of the CD4 T cell landscape toward extreme regulatory and effector phenotypes. *Science Advances*, 5(8): eaaw8330, 2019.
- P. Fillard, V. Arsigny, X. Pennec, K. M. Hayashi, P. M. Thompson, and N. Ayache. Measuring brain variability by extrapolating sparse tensor fields measured on sulcal lines. *Neuroimage*, 34(2):639–650, 2007.
- K. J. Friston. Functional and effective connectivity: a review. *Brain Connectivity*, 1(1): 13–36, 2011.
- A. Gonzalez-Sanz, J.-M. Loubes, and J. Niles-Weed. Weak limits of entropy regularized Optimal Transport; potentials, plans and divergences, 2022. arXiv:2207.07427.
- M. Hallin, G. Mordant, and J. Segers. Multivariate goodness-of-fit tests based on Wasserstein distance. *Electronic Journal of Statistics*, 15(1):1328–1371, 2021.
- P. D. Hoff and X. Niu. A covariance regression model. *Statistica Sinica*, pages 729–753, 2012.
- T. Hsing and R. Eubank. *Theoretical foundations of functional data analysis, with an introduction to linear operators*, volume 997. John Wiley & Sons, 2015.
- W. Hu, T. Pan, D. Kong, and W. Shen. Nonparametric matrix response regression with application to brain imaging data analysis. *Biometrics*, 77(4):1227–1240, 2021.
- J.-C. Hütter and P. Rigollet. Minimax estimation of smooth optimal transport maps. *The Annals of Statistics*, 49(2):1166–1194, 2021.
- C. Jin, P. Netrapalli, R. Ge, S. M. Kakade, and M. I. Jordan. A Short Note on Concentration Inequalities for Random Vectors with SubGaussian Norm, 2019. arXiv:1902.03736.
- Y.-H. Kim and B. Pass. Wasserstein barycenters over Riemannian manifolds. *Advances in Mathematics*, 307:640–683, 2017.
- D. Kong, B. An, J. Zhang, and H. Zhu. L2rm: Low-rank linear regression models for high-dimensional matrix responses. *Journal of the American Statistical Association*, 115(529): 403–424, 2020.
- A. Kroshnin, V. Spokoiny, and A. Suvorikova. Statistical inference for Bures–Wasserstein barycenters. *The Annals of Applied Probability*, 31(3):1264–1298, 2021.
- T. Le Gouic and J.-M. Loubes. Existence and consistency of Wasserstein barycenters. *Probability Theory and Related Fields*, 168(3-4):901–917, 2017.

- T. Le Gouic, Q. Paris, P. Rigollet, and A. J. Stromme. Fast convergence of empirical barycenters in Alexandrov spaces and the Wasserstein space. *Journal of the European Mathematical Society*, 25(6):2229–2250, 2022.
- Z. Lin. Riemannian Geometry of Symmetric Positive Definite Matrices via Cholesky Decomposition. *SIAM Journal on Matrix Analysis and Applications*, 40(4):1353–1370, 2019.
- Z. Lin, H.-G. Müller, and B. U. Park. Additive models for symmetric positive-definite matrices and Lie groups. *Biometrika*, 110(2):361–379, 2023.
- R. Ma, E. D. Sun, D. Donoho, and J. Zou. Principled and interpretable alignability testing and integration of single-cell data. *Proceedings of the National Academy of Sciences*, 121(10):e2313719121, 2024.
- T. Manole, S. Balakrishnan, J. Niles-Weed, and L. Wasserman. Plugin Estimation of Smooth Optimal Transport Maps, 2022. arXiv:2107.12364.
- T. Manole, S. Balakrishnan, J. Niles-Weed, and L. Wasserman. Central Limit Theorems for Smooth Optimal Transport Maps, 2023. arXiv:2312.12407.
- G. Mena and J. Niles-Weed. Statistical bounds for entropic optimal transport: sample complexity and the central limit theorem. *Advances in neural information processing systems*, 32, 2019.
- S.-I. Ohta. Barycenters in Alexandrov spaces of curvature bounded below. *Advances in Geometry*, 12(4):571–587, 2012.
- V. M. Panaretos and Y. Zemel. Amplitude and phase variation of point processes. *The Annals of Statistics*, 44(2):771–812, 2016.
- V. M. Panaretos and Y. Zemel. *An Invitation to Statistics in Wasserstein Space*. Springer Nature, 2020.
- A. Petersen and H.-G. Müller. Fréchet regression for random objects with Euclidean predictors. *The Annals of Statistics*, 47(2):691–719, 2019.
- A. Petersen, X. Liu, and A. A. Divani. Wasserstein F-tests and confidence bands for the Fréchet regression of density response curves. *The Annals of Statistics*, 49(1):590–611, 2021.
- G. Peyré and M. Cuturi. Computational Optimal Transport: With Applications to Data Science. *Foundations and Trends® in Machine Learning*, 11(5-6):355–607, 2019.
- A.-A. Pooladian and J. Niles-Weed. Entropic estimation of optimal transport maps, May 2022. arXiv:2109.12004.
- A.-A. Pooladian, V. Divol, and J. Niles-Weed. Minimax estimation of discontinuous optimal transport maps: The semi-discrete case. In *International Conference on Machine Learning*, pages 28128–28150. PMLR, 2023.

- C. Ramos-Carreño and J. L. Torrecilla. dcor: Distance correlation and energy statistics in Python. *SoftwareX*, 22, 2 2023. doi: 10.1016/j.softx.2023.101326.
- I. Redko, A. Habrard, and M. Sebban. Theoretical Analysis of Domain Adaptation with Optimal Transport. In M. Ceci, J. Hollmén, L. Todorovski, C. Vens, and S. Džeroski, editors, *Machine Learning and Knowledge Discovery in Databases*, pages 737–753, 2017.
- F. Santambrogio. *Optimal Transport for Applied Mathematicians: Calculus of Variations, PDEs, and Modeling*, volume 87. Springer, NY, 2015.
- G. Schiebinger, J. Shu, M. Tabaka, B. Cleary, V. Subramanian, A. Solomon, J. Gould, S. Liu, S. Lin, P. Berube, L. Lee, J. Chen, J. Brumbaugh, P. Rigollet, K. Hochedlinger, R. Jaenisch, A. Regev, and E. S. Lander. Optimal-Transport Analysis of Single-Cell Gene Expression Identifies Developmental Trajectories in Reprogramming. *Cell*, 176(4): 928–943, 2019.
- V. R. Somnath, M. Pariset, Y.-P. Hsieh, M. R. Martinez, A. Krause, and C. Bunne. Aligned Diffusion Schrödinger Bridges. In *Proceedings of the Thirty-Ninth Conference on Uncertainty in Artificial Intelligence*, volume 216, pages 1985–1995. PMLR, 2023.
- G. J. Székely, M. L. Rizzo, and N. K. Bakirov. Measuring and testing dependence by correlation of distances. *The Annals of Statistics*, 35(6):2769 – 2794, 2007.
- M. Talagrand. Isoperimetry and Integrability of the Sum of Independent Banach-Space Valued Random Variables. *The Annals of Probability*, 17(4):1546–1570, 1989.
- W. L. Tony Cai and X. Luo. A constrained l1 minimization approach to sparse precision matrix estimation. *Journal of the American Statistical Association*, 106(494):594–607, 2011.
- A. W. van der Vaart. *Asymptotic Statistics*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 1998.
- A. W. Van Der Vaart and J. A. Wellner. *Weak Convergence and Empirical Processes*. Springer Series in Statistics. Springer, New York, NY, 1996.
- R. Vershynin. *High-Dimensional Probability: An Introduction with Applications in Data Science*. Cambridge University Press, Cambridge, 2018.
- C. Villani. *Topics in Optimal Transportation*. American Mathematical Society, 2003.
- C. Villani. *Optimal Transport: Old and New*, volume 338. Springer, Berlin, Heidelberg, 2009.
- M. J. Wainwright. *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*. Cambridge University Press, 2019.
- L. Wasserman. *All of Nonparametric Statistics*. Springer Texts in Statistics. Springer, New York, NY, 2006. ISBN 978-0-387-25145-5. doi: 10.1007/0-387-30623-4.

- J. Weed and Q. Berthet. Estimation of smooth densities in Wasserstein distance. In *conference on Learning Theory*, pages 3118–3119. PMLR, 2019.
- T. P. Wihler. On the Hölder continuity of matrix functions for normal matrices. *Journal of inequalities in pure and applied mathematics*, 10(10), 2009.
- S. Yazar, J. Alquicira-Hernandez, K. Wing, A. Senabouth, M. G. Gordon, S. Andersen, Q. Lu, A. Rowson, T. R. P. Taylor, L. Clarke, K. Maccora, C. Chen, A. L. Cook, C. J. Ye, K. A. Fairfax, A. W. Hewitt, and J. E. Powell. Single-cell eqtl mapping identifies cell type-specific genetic control of autoimmune disease. *Science*, 376(6589):eabf3041, 2022.
- T. Ye and S. S. Du. Global convergence of gradient descent for asymmetric low-rank matrix factorization. In *Advances in Neural Information Processing Systems*, 2021.
- T. Yokota. Convex functions and barycenter on CAT (1)-spaces of small radii. *Journal of the Mathematical Society of Japan*, 68(3):1297–1323, 2016.
- Y. Yuan, H. Zhu, W. Lin, and J. S. Marron. Local polynomial regression for symmetric positive definite matrices. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 74(4):697–719, 2012.
- Y. Zhao, B. Wang, S. H. Mostofsky, B. S. Caffo, and X. Luo. Covariate assisted principal regression for covariance matrix outcomes. *Biostatistics*, 22(3):629–645, 2021.
- T. Zou, W. Lan, H. Wang, and C.-L. Tsai. Covariance Regression Analysis. *Journal of the American Statistical Association*, 112(517):266–281, 2017.