London School of Economics and Political Science

Applications of Optimal Transport in Multivariate Statistics

Xuzhi Yang

A thesis submitted to the Department of Statistics of the London School of Economics and Political Science for the degree of Doctor of Philosophy

May, 2025

To Zhong & Junyan, for their unwavering support
To Yueran, for her love beyond measure

Declaration

I certify that the thesis I have presented for examination for the PhD degree of the London School of Economics and Political Science is solely my own work other than where I have clearly indicated that it is the work of others (in which case the extent of any work carried out jointly by me and any other person is clearly identified in it).

I confirm that Chapter 2 is co-authored with Professor Tengyao Wang, and has been published in *The Thirty Seventh Annual Conference on Learning Theory*. Chapter 3 is a joint work with Doctor Mona Azadikia and Professor Tengyao Wang. We plan to submit Chapter 3 for publication soon.

Abstract

Finding a desirable generalisation of rank-based statistical methods to multivariate case has been a timeless statistical endeavor. While various concepts of multivariate rank/quantile have been proposed in the past decades, most of them do not maintain crucial properties enjoyed by the traditional univariate rank/quantile, such as distribution-freeness and strong consistency. A novel multivariate rank/quantile from the perspective of optimal transport (OT) was proposed by Chernozhukov et al. [Che+17] and Hallin et al. [Hal+21]. This OT-based concept extends most of desirable properties of traditional rank/quanilte on real line to multidimensional space, thus has drawn an increasing attention in recent years.

In this essay, we apply the OT-based multivariate rank/quantile on two statistical domains: multiple-output quantile regression and nonparametric independence testing between random vectors. On the first direction, we introduce a robust estimation of multiple-output linear model coefficient by extending the traditional univariate composite quantile regression to the case of multivariate response variable through OT-based techniques. Both the consistency and the convergence rate of the proposed estimator is established under multivariate heavy-tailed random error case. For the second direction, we proposed an geometrically intuitive correlation coefficient for random vectors utilising the OT-based multivariate rank. The proposed coefficient enjoys an entirely distribution-free asymptotic theory under the independent assumption, thus avoiding any permutation-based p-value calculations. Moreover, unlike many existing measurement, the proposed coefficient is capable of detecting not only functional dependency but also spurious correlation via confounders.

Acknowledgements

My deepest gratitude to my primary supervisor, Tengyao Wang, who is the most patient, supportive, and considerate mentor and friend. I am grateful for his tremendous encouragement and support, especially during these times of great uncertainty. What I have learned from him will serve me throughout the rest of my life. I would also like to express my appreciation to Yining Chen, for his constantly guidance during this journey. I extend my sincere appreciation to all participants of Tengyao's reading group, where I gained invaluable knowledge beyond my own research.

To my dear friends—Tao, Yutong, Xinyi, Sixing, Yudong, August, Pingfan, Zetai, Pouya, Trevor, Kaixin, Shakeel—thank you for the laughter, the late-night talks, and the countless ways you've made this journey brighter. I couldn't have done it without you.

I also thank many professors and professional staffs in the department for fostering an environment of equality where everyone is willing to pause and listen to a PhD student's voice. Especially, I thank Zoltán Zsábo, for his selfless dedication to our discussions on various topics, even sometimes at the cost of his dinner.

No words can express my gratitude to my parents, Zhong and Junyan, and my fiancée Yueran– they are the real heroes.

Contents

List of Figures

Chapter 1

Introduction

Modern scientific research routinely confronts datasets contaminated by systematic errors and outliers. This contamination poses significant challenges for traditional statistical methods that rely on distributional assumptions or are sensitive to extreme values. The concept of rank-based statistics emerges as a powerful alternative, offering robustness against data contamination while remaining desirable property without distribution assumptions on the data. Rather than working with raw observations, rank-based methods operate on the relative ordering of data points, effectively mitigating the impact of outliers and reducing the influence of measurement errors. This approach, pioneered by early statisticians like Wilcoxon [Wil92], Mann-Whitney [MW47] and Siegel [Sid57], has gained renewed relevance in applications ranging from high-throughput genomics to financial data analysis [DK14; VG20; PA24]. However, modern treatment of this topic focuses more on handling multivariate data. For instance, in the area of robust mean estimation, Diakonikolas, Kane, and Pensia [DKP20], Lugosi and Mendelson [LM21], Depersin and Lecué [DL22], and Minasyan and Zhivotovskiy [MZ23] have proposed various extensions of univariate robust mean procedures such as the trimmed mean estimator [TM63] and median of means estimator [NY83; JVV86; AMS96] to the multivariate setting. We witness a similar surge in research interest in the area of robust covariance estimation [MZ20; AZ22; MZ23].

However, the extension of rank-based methods to multivariate settings presents fundamental challenges, as there is no natural ordering of points in higher dimensions that preserves the desirable theoretical properties of ranks on the real line. Various concepts have been considered, e.g. depth-based ranks [Tuk75; LS93; ZS00], spatial ranks [MO95; Cha96; Kol97], componentwise ranks [Hod55; Bic65], Mahalanobis ranks [HP02b; HP02a], but none of them enjoy the *distribution-freeness* and *essential maximal ancillarity* while the traditional rank notion on real line does.

Monge-Kantorovich rank, a concept of multivariate rank proposed in Chernozhukov et al. [Che+17], Hallin [Hal17], and Hallin et al. [Hal+21] provides a new insight of traditional ranks from the perspective of optimal transport (OT) map. It is robust to the outliers in natural, and more importantly, it enjoys desirable properties that make the success of univariate rank. It has been applied successfully in a variety of multivariate statistical problems, including two-sample testing [DS21; Shi+21; SDH22a; HS23], multiple-output regression [CCG16; HHH23; BSH24; YW24], dependency measurement [Shi+22; DS23], semiparametric estimation [HLL22], etc.

This thesis contributes two applications of the Monge-Kantorovich rank in multivariate statistics: robust estimation of multiple-output linear model coefficient and dependency measurement

between multivariate random vectors. In Chapter 2, we propose a multivariate extension of the traditional univariate composite quantile regression developed by Zou and Yuan [ZY08]. We establish both consistency and convergence rate results for this extension. Our work addresses the gap in existing literature, as previous studies on multiple-output quantile regression using OT have primarily focused on quantile contour estimation rather than linear model coefficient estimation. In Chapter 3, we introduce a novel rank-based correlation coefficient with the following features: 1) it leverages the Monge-Kantorovich rank to measure dependency between random vectors; 2) it enjoys a distribution-free asymptotic theory; 3) it is capable to detect not only functional correlation, but also implicit functional correlation.

In the rest of this chapter, we establish foundations of our work by detailed presenting fundamental concepts and reviewing existing literature in rank-based statistical methods. The background material serves to provide readers with a comprehensive introduction to the field while contextualising our proposed method within the current literature. We begin with traditional univariate rank-based methods and their desirable properties in Section 1.1. We then review existing approaches for generalising the concept of rank from the real line to the multivariate case in Section 1.2. Finally, we introduce the concept of Monge-Kantorovich rank and quantile, along with their applications in regression and nonparametric statistics in Section 1.3.

Notations. For any integer $d \geq 1$, we write $\mathcal{B}$ as the Borel σ-algebra of $\mathbb{R}^d$. Write $\mathcal{P}(\mathbb{R}^d)$ as the set of Borel probability measures defined on $(\mathbb{R}^d, \mathcal{B})$. For any random variable X, we denote P^X to be its induced Borel probability measure and $\sigma(X)$ to be the Borel sigma-algebra generated by X. We write $\mathcal{N}(\mu, \Sigma)$ as the Gaussian distribution with mean vector $\mu \in \mathbb{R}^p$ and covariance matrix $\Sigma \in \mathbb{R}^p \times \mathbb{R}^p$. Given $x = (x_1, \ldots, x_d) \in \mathbb{R}^d$, we denote $\|x\| = \left(\sum_{i=1}^n x_i^2\right)^{1/2}$ to be the Euclidean norm of x. Let $\xrightarrow{\text{d}}$ denote convergence in distribution and $\xrightarrow{\text{a.s.}}$ denote almost sure convergence.

1.1 Rank-based statistical methods

For the univariate data, rank-based methods have become foundational in robust estimation and nonparametric statistics. Particularly, *quantile regression* [KB78] and nonparametric independence testing exemplify the remarkable efficiency of rank transformations. In this section, we examine the problem formulation and existing approaches within these two areas.

1.1.1 Quantile regression

We consider a covariate-response pair (X, Y) with joint distribution $P^{(X,Y)} \in \mathcal{P}(\mathbb{R}^p \times \mathbb{R})$ is generated from linear regression model

$$Y = \beta_*^\top X + \varepsilon, \tag{1.1}$$

where $\beta_* \in \mathbb{R}^p$ is the regression coefficient and random noise ε is independent of X. Given i.i.d. covariate-response pairs $(X_1, Y_1), \ldots, (X_n, Y_n)$ drawn from $P^{(X,Y)}$, then the corresponding error terms are $\varepsilon_i = Y_i - \beta_*^\top X_i$, $i = 1, \ldots, n$. The goal is to estimate β_* when the random error term possibly follows a heavy-tailed distribution.

The *M-estimator* represents a broad class of robust estimation methods. It is defined as the empirical risk minimiser of a loss function $\ell : \mathbb{R} \to \mathbb{R}_{\geq 0}$:

$$\widehat{\beta} \in \operatorname*{arg\,min}_{\beta \in \mathbb{R}^p} \frac{1}{n} \sum_{i=1}^{n} \ell(Y_i - \beta^\top X_i).$$

In particular, by taking $\ell(t) = t^2$, we obtain the *ordinary least squared* (OLS) estimator of β_*:

$$\widehat{\beta}^{\mathrm{OLS}} = \operatorname*{arg\,min}_{\beta \in \mathbb{R}^p} \frac{1}{n} \sum_{i=1}^{n} (Y_i - \beta^\top X_i)^2. \tag{1.2}$$

It is known that under the Gauss-Markov assumptions: 1) $\mathbb{E}\,\varepsilon_i = 0$; 2) $\mathrm{Var}(\varepsilon_1) = \mathrm{Var}(\varepsilon_2) = \ldots = \mathrm{Var}(\varepsilon_n) = \sigma^2 < +\infty$; 3) $\mathrm{Cov}(\varepsilon_i, \varepsilon_j) = 0$ for all $i \neq j$, the OLS estimator $\widehat{\beta}^{\mathrm{OLS}}$ attains the minimal possible variance among all the linear unbiased estimator of β_*. Moreover, under the same set of assumptions plus that $D_0 := \frac{1}{n}\sum_{i=1}^{n} X_i X_i^\top$ is non-degenerate, we have asymptotic normality

$$\sqrt{n}(\widehat{\beta}^{\mathrm{OLS}} - \beta_*) \xrightarrow{\mathrm{d}} \mathcal{N}\Big(0, \sigma^2 D_0^{-1}\Big), \quad \text{as } n \to \infty. \tag{1.3}$$

However, one can even extend the optimality to non-linear estimator by the *maximal likelihood estimator* (MLE) provided the density function of the random error is known.

If $\varepsilon_1, \ldots, \varepsilon_n$ have a known absolutely continuous density function p_0 such that the *Fisher information* $i_{\varepsilon_1} = \int (\frac{\partial}{\partial x}\ell(x))^2 p_0(x)\,\mathrm{d}x$ is finite, then we can obtain the MLE $\widehat{\beta}^{\mathrm{MLE}}$ by letting $\ell = -\log p_0$. In particular, when $\varepsilon_1, \ldots, \varepsilon_n$ follows a standard Gaussian distribution, the $\widehat{\beta}^{\mathrm{MLE}}$ coincides with the OLS estimator. Moreover, under some regularity conditions, we have asymptotic normality [Van00, Theorem 5.39]

$$\sqrt{n}(\widehat{\beta}^{\mathrm{MLE}} - \beta_*) \xrightarrow{\mathrm{d}} \mathcal{N}\Big(0, \frac{\{\mathbb{E}(X_1 X_1^\top)\}^{-1}}{i_{\varepsilon_1}}\Big). \tag{1.4}$$

Indeed, the asymptotic variance in (1.4) attains the Cramer-Rao lower bound thus the MLE estimator exhibits the optimality in the sense that it is *asymptotically* uniformly minimal-variance unbiased estimator [see e.g. Van00, Section 5.5].

However, in practice, since p_0 is typically unknown, the MLE estimator is not directly available, and the OLS estimator may not be optimal because its variance is proportional to σ^2 (see (1.3)) and can blow up when the random error follows a heavy-tailed distribution. On popular way to tackle the heavy-tailed error is based on the *quantile regression* (QR) [KB78]. The quantile regression approach determines the estimator of β_* by solving the following optimisation problem:

$$\hat{\beta}^{QR} = \operatorname*{arg\,min}_{\beta \in \mathbb{R}^p} \frac{1}{n} \sum_{i=1}^{n} \rho_\tau(Y_i - \beta^\top X_i), \tag{1.5}$$

where $\tau \in (0,1)$ and $\rho_\tau : \mathbb{R} \to \mathbb{R}_{\geq 0}$ is the so-called *check function* defined by $\rho_\tau(x) = \max\{x, 0\} + (\tau - 1)x$, for any $x \in \mathbb{R}$ (see Fig. 1.1). Unlike the OLS estimator, which estimates the conditional mean function, the quantile regression estimator aims to estimate the conditional quantile function of Y given X. Specifically, note that the population counterpart of optimisation (1.2) is

$\min \mathbb{E}(Y - \beta^\top X)^2$. Since the conditional mean $E(Y \mid X)$ is the best projection of Y onto the L^2-space consisting of all $\sigma(X)$-measurable functions, OLS is effectively estimating $E(Y \mid X)$ through a linear model. On the other hand, the population counterpart of (1.5) is $\min \mathbb{E}\, \rho_\tau(Y - \beta^\top X)$, and by calculating the saddle point of the objective function one can show that the minimiser of such an optimisation problem is the conditional quantile function $q_{Y|X}(\tau) := \inf\{y \in \mathbb{R} : \mathbb{P}(Y \leq y|X) \geq \tau\}$ for any $\tau \in (0,1)$, which implies that quantile regression aims to estimate the conditional quantile function of Y with a linear model. Since quantile as a functional is more robust compared to the expectation functional, quantile regression is a more robust estimation method. We guide the readers to [EL11] for analysis of mean and quantile as functionals from the perspective of influence function.

Moreover, under some continuity conditions on p_0, the estimator $\hat{\beta}^{QR}$ has asymptotic normality as follows:

$$\sqrt{n}(\hat{\beta}^{QR} - \beta_*) \xrightarrow{\mathrm{d}} \mathcal{N}(0, \omega^2 D_0^{-1}), \tag{1.6}$$

where $\omega^2 = \tau(1-\tau)/f_i^2(q_i(\tau))$ with f_i and q_i are the density function and quantile function of ε_i. Compared to the OLS estimator, the quantile regression estimator achieves $\sqrt{n}$-consistency and asymptotic normality without requiring finite variance; however, its relative efficiency can be arbitrarily small because the term $f_i^2(q_i(\tau))$ in (1.6) can be close to zero. Zou and Yuan [ZY08] proposed a solution to this issue through the *composite quantile regression* (CQR) method, whose loss function aggregates multiple quantile regression loss functions. Specifically, for any $K \in \mathbb{N}$, the CQR estimator $\widehat{\beta}^{\mathrm{CQR}}$ is obtained by the following optimisation problem

$$(\hat{q}_1, \ldots, \hat{q}_K, \widehat{\beta}^{\mathrm{CQR}}) = \underset{q_1,\ldots,q_K \in \mathbb{R},\ \beta \in \mathbb{R}^p}{\arg\min} \sum_{i=1}^{n} \sum_{k=1}^{K} \rho_{\tau_k}(Y_i - \beta^\top X_i - q_k),$$

for $\tau_k = k/(K+1)$. Zou and Yuan [ZY08] showed that the CQR estimator can achieve at least 70% relative efficiency compared to the OLS estimator even for Gaussian noise, and for the case of non-Gaussian random error, it typically enjoys a much smaller variance.

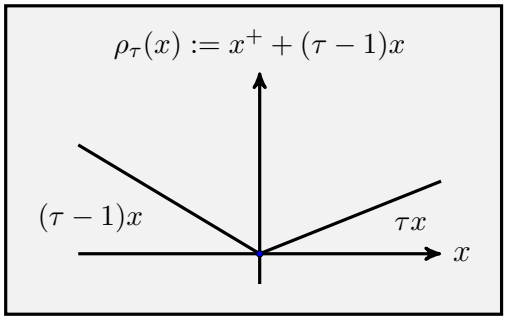
However, when Y_i is a multivariate response variable, neither the QR estimator (1.5) nor the CQR estimator in (2.3) has a natural extension, due to the lack of well-defined multivariate ranks/quantiles and corresponding multivariate check functions. While there is extensive literature on multiple-output quantile regression (please see Section 1.2.1 and 1.3.2 for details), the extension of the CQR estimator remains unexplored. To bridge this gap, we propose a multiple-output CQR estimator based on OT theory in Chapter 2.

1.1.2 Nonparametric independence testing

Another strength of rank-based methods lies in the problem of mutual independence testing. Specifically, given P^X and P^Y are probability measures on $\mathbb{R}^{d_1}$ and $\mathbb{R}^{d_2}$, respectively. Suppose $P^{(X,Y)} \in \mathcal{P}(\mathbb{R}^d)$ is the joint distribution of P^X and P^Y, where $d = d_1 + d_2$. In this section, we focus on the case of univariate marginal distribution, i.e. $d_1 = d_2 = 1$, and consider the following hypothesis

$$\mathcal{H}_0 : P^{(X,Y)} = P^X \otimes P^Y \quad \text{v.s.} \quad \mathcal{H}_1 : P^{(X,Y)} \neq P^X \otimes P^Y \tag{1.7}$$

Figure 1.1: Illustration of check function.



Traditional approaches to the two-sample independence testing problem often rely on specific distributional assumptions. The Pearson correlation test [Pea20], which remains widely used in practice, assumes bivariate normality and focuses exclusively on linear relationships. The likelihood ratio test [Wil38], requires explicit parametric specifications of the underlying distributions, making its power heavily dependent on the validity of these assumptions. These distributional dependencies have motivated the development of rank-based nonparametric testing procedures, such as Spearman's rank correlation and Kendall's τ coefficient [Ken38; Spe04].

There have been various of non-parametric methods proposed to try to mitigate such distributional assumptions. One of the earliest line of work are based on joint cumulative distribution functions and ranks. Hoeffding [Hoe94] pioneered this approach by introducing a test statistic based on the discrepancy between the empirical joint distribution function and the product of marginal distribution functions, which was later extended in Blum, Kiefer, and Rosenblatt [BKR61] and Yanagimoto [Yan70]. Mosteller [Mos46] introduced a dependency measure called the *quadrant count ratio*, further developed in [Blo50]. Subsequently, Rosenblatt [Ros75] proposed a test procedure based on density estimation. Bergsma and Dassios [BD14] introduced a modified version of Kendall's τ that ensures consistency under mild conditions on the joint distribution F, with further theoretical analysis provided in Nandy, Weihs, and Drton [NWD16] and Weihs, Drton, and Leung [WDL16]. Another line of work is based on kernel method. The Hilbert-Schmidt Independence Criterion (HSIC), introduced by Gretton et al. [Gre+05a; Gre+07], has proven particularly versatile. This approach was subsequently adapted by Sen and Sen [SS14] for independence testing in linear models, while Ramdas et al. [Ram+15] established its theoretical properties in high-dimensional settings. The framework was further generalized by Pfister et al. [Pfi+18] to handle K-sample ($K \geq 2$) independence testing. Other proposals include coefficient based on copulas [Skl59; SW81; DSS13; LHS13; Fuc24; GJT22; SDS24]; correlation coefficient based on pairwise distance [SR09; SRB07; HHG13], OT-based method [NSM21; MS22; Wie22; MS20]. We refer the readers to [DK01; JH16; TOS22; Cha24] and the references therein for extensive reviewing on this area.

Despite some of these correlation coefficients being consistent under fixed alternatives, there are several common problems. First, most of the correlation coefficients are designed for independence testing not for measuring the strength of the relationship between the variables. Secondly, although some of the coefficient enjoys consistency under a fixed alternative, most of them are lack of a distribution-free null asymptotic theory. In the absence of such result, one need to resort to permutation-based method to obtain a p-value, which is quite computationally expensive, and the issue becomes even more prohibitive under the case of multiple testing.

Chatterjee [Cha21] proposed a simple rank-based coefficient of correlation solves the issues mentioned above. Specifically, given n independent copies $(X_1, Y_1), \ldots, (X_n, Y_n) \overset{\text{iid}}{\sim} P^{(X,Y)} \in \mathcal{P}(\mathbb{R}^2)$, and assuming no ties in $(X_i)_{1\le i\le n}$ and $(Y_i)_{1\le i\le n}$. Then we can rearrange the data as $(X_{(1)}, Y_{(1)}), \ldots, (X_{(n)}, Y_{(n)})$ such that $X_{(1)} < \cdots < X_{(n)}$. Chatterjee's correlation is then defined as

$$\xi_n^{(X,Y)} := 1 - \frac{\sum_{i=1}^{n-1} |r_{i+1} - r_i|}{(n^2-1)/3}, \tag{1.8}$$

where $r_i := \#\{j : Y_{(j)} \le Y_{(i)}\}$ is the rank of $Y_{(i)}$. It is shown to enjoy the following desirable properties:

(I) If Y is not almost surely a constant, then as $n \to \infty$, we have

$$\xi_n^{(X,Y)} \overset{\text{a.s.}}{\to} \xi^{(X,Y)} := \frac{\int \operatorname{Var}(\mathbb{E}(\mathbb{1}_{Y\ge t} \mid X))\, \mathrm{d}\mu(t)}{\int \operatorname{Var}(\mathbb{1}_{Y\ge t})\, \mathrm{d}\mu(t)}. \tag{1.9}$$

Moreover, $\xi^{(X,Y)} = 0$ if and only if X and Y are independent, and $\xi^{(X,Y)} = 1$ if and only if there exists a measurable function $f : \mathbb{R} \to \mathbb{R}$ such that $Y = f(X)$ almost surely.

(II) When X and Y are independent, as $n \to \infty$, ξ_n enjoys the following asymptotic normality:

$$\sqrt{n}\xi_n^{(X,Y)} \overset{\text{d}}{\to} \mathcal{N}(0, 2/5).$$

(III) The coefficient can be computed in time $O(n \log n)$.

Given its simplicity and nonparametric nature, Chatterjee's correlation has been applied across various practical fields [e.g. Sad22; Suo+24], despite being a relatively recent development. With these desirable properties, the Chatterjee's coefficient has attracted much attention recently, such as the power analysis of the coefficient [SDH22b; ADN21; LH23; Bic22], asymptotic theory under the alternative [LH22; Kro24], multivariate extension [AC21; DGS20; Han21; AF24], measuring conditional independence[AC21; HDS22; SDH24; HH24].

As one may note, Chatterjee's correlation coefficient (1.8) is intentionally asymmetric, i.e., $\xi_n^{(X,Y)} \neq \xi_n^{(Y,X)}$. This design aims to detect the functional relationship between variables, specifically, whether Y is a function of X or vice versa. However, this means that Chatterjee's correlation may not be powerful in detecting dependence between X and Y mediated through their respective functional dependence on some hidden variable H. Moreover, although there have been some existing work on the multivariate extension of $\xi_n^{(X,Y)}$ [e.g. DGS20; AF24], none of them are exact distribution-free.

In Chapter 3, we introduce a new correlation coefficient motivated by a geometric interpretation of Chatterjee's correlation. We demonstrate that the proposed coefficient satisfies both consistency and asymptotic normality under only absolutely continuous condition on the marginal distributions, while maintaining computational efficiency with an $O(n \log n)$ algorithm in univariate case. Furthermore, by leveraging the concept of Monge-Kantorovich rank, our proposed coefficient provides a natural multivariate framework for measuring dependence between random vectors X and Y.

1.2 Statistical methods based on multivariate rank/quantile

Building upon the univariate rank-based methods discussed in the previous section, we now turn our attention to their multivariate extensions. While univariate rank transformations provide powerful tools for robust estimation and nonparametric testing, extending these concepts to multivariate case introduces significant theoretical and computational challenges. The absence of a natural ordering in multivariate spaces requires more sophisticated approaches to define ranks and quantiles. In this section, we examine several key frameworks that generalise rank-based statistical methods to multivariate settings, focusing particularly on the theoretical foundations and statistical properties that emerge in multivariate applications.

1.2.1 Multiple-output quantile regression

We consider similar coefficient estimation problem as in Section 1.1.1, but under multiple-output linear model. Suppose we have covariate-response pair (X, Y) with joint distribution $P^{(X,Y)} \in \mathcal{P}(\mathbb{R}^p \times \mathbb{R}^d)$, where $d, p \geq 1$, that is generated from

$$Y = b_* X + \varepsilon, \tag{1.10}$$

with regression coefficient $b_* \in \mathbb{R}^{d\times p}$ and a noise vector ε taking values in $\mathbb{R}^d$ independent of X. Given $(X_1, Y_1), \ldots, (X_n, Y_n)$ are i.i.d. samples from $P^{(X,Y)}$, we aim to develop a robust estimation for the regression coefficient b_* in the case of a multivariate heavy-tailed random error term.

The traditional OLS estimator (1.2) can be adopted for multiple-output linear model (1.10) by letting $\ell(t) = \|t\|^2$, where $t \in \mathbb{R}^d$. However, since the asymptotic variance of such OLS estimator depends on the covariance matrix of ε, it is not a reliable estimation when the random error term follows some multivariate heavy-tailed distribution, e.g. the multivariate t-distribution [Rot12]. This motivates us to consider a multivariate extension of the quantile regression estimator.

The generalisation of the idea of quantile regression to the multivariate case has been a rather long history in statistics. Many concepts have been considered in the literature, including coordinatewise quantiles, projection method, spatial quantile, halfspace depth method. The idea of coordinatewise rank/quantile dates back to 1950s and 1960s by Hodges [Hod55], Bickel [Bic65], Puri and Sen [PS66], Sen and Puri [SP67], and Puri and Sen [PS67; PS71]. In the context of $\mathbb{R}^d$, the coordinatewise rank is defined to be a tuple of marginal orderings on the real line. However, this approach lacks rotation invariance and fails to achieve distribution-freeness, limiting its applicability. As a result, our investigation focuses on the latter three methodologies mentioned above: projection-based methods, spatial quantiles, and halfspace depth methods.

Given a d-dimensional random variable Y, the key idea of the projection method is that for each unit vector $u \in \mathbb{R}^d$ and $\tau \in (0, 1)$ the τ-quantile of $u^\top Y$ is well-defined. By utilising this idea, Kong and Mizera [KM12b] defines the *directional quantile* of order τ for vector u as $q_{Y;\mathrm{KM}}(\tau u) := q_{u^\top Y}(\tau) \cdot u$, where

$$q_{u^\top Y}(\tau) = \inf\{y : \mathbb{P}(u^\top Y \leq y) \geq \tau\},$$

which can induce a quantile halfspace $H^+_{\mathrm{KM},\tau u} := \{z \in \mathbb{R}^d : u^\top(z - q_{Y;\mathrm{KM}}(\tau u)) \geq 0\}$. Therefore, for any fixed quantile level $\tau \in (0, 1/2)$, the τ-quantile contour can be constructed by $D_{\mathrm{KM}}(\tau) := \bigcap_{u\in\mathcal{S}^{d-1}} H^+_{\mathrm{KM},\tau u}$. However, in practice, one needs to sample a large number of unit

vector on $\mathcal{S}^{d-1}$ to obtain an estimator of D_{KM}, which can be computationally inefficient. Moreover, this definition depends on the choice of origin and also lack of affine equivariance, which are undesirable.

Another important step on this direction is proposed by Hallin, Paindaveine, and Šiman [HPŠ10]. Instead of considering the quantile of $u^\top Y$ directly, [HPŠ10] projects Y onto u and its orthogonal space, and then they construct the quantile hyperplane by running a regular quantile regression. Specifically, for any fixed $u \in \mathcal{S}^{d-1}$, let $\Gamma_u \in \mathbb{R}^{d\times(d-1)}$ such that $(u, \Gamma_u) \in \mathbb{R}^{d\times d}$ be a unit orthonomal basis. Let $Z_u := u^\top Z$ and $Z_u^\perp := \Gamma_u^\top Z$. Then the HPŠ's τu-quantile hyperplane is obtained by regressing Z_u on $Z_u^\perp$ and an intercept term under the check function. In detail, the HPŠ's τu-quantile hyperplane is

$$H_{\mathrm{HP\check{S}},\tau u} := \{z \in \mathbb{R}^d : u^\top z = b_\tau^\top \Gamma_u^\top z + a_\tau\}, \tag{1.11}$$

where (b_τ, a_τ) is obtained by

$$(b_\tau, a_\tau) \in \operatorname*{arg\,min}_{(a,b)\in\mathbb{R}^d} \mathbb{E}\,\rho_\tau(Z_u - b^\top Z_u^\perp - a). \tag{1.12}$$

Thus the corresponding quantile halfspace follows as $H^+_{\mathrm{HP\check{S}},\tau u} := \{z \in \mathbb{R}^d : u^\top z \geq b_\tau^\top \Gamma_u^\top z + a_\tau\}$, and thus the τ-quantile contour is $D_{\mathrm{HP\check{S}}}(\tau) = \bigcap_{u\in\mathcal{S}^{d-1}} \cap H^+_{\mathrm{HP\check{S}},\,\tau u}$, where $\cap H^+_{\mathrm{HP\check{S}},\,\tau u}$ is intersection of all hyperplanes that satisfy (1.11). It is shown in [HPŠ10] that this concept of multivariate quantile enjoys many desirable properties including, affine-invariance, strong consistency, asymptotic normality and Bahadur-type representation. Moreover, in term of the empirical estimation, the method also enjoy an efficient algorithm due to its close relationship with Tukey's depth; see also [PŠ12; PŠ12]. This concept of multivariate quantile can also be immediately adopted into the problem of multiple-output nonparametric quantile regression (see [HPŠ10, Section 6]), however, as pointed out in [Hal+15], the resulting quantile contours carries little information on the population conditional quantile of Y given $X = x$, but some average version of the latter. We refer the readers to [Hal+15] for more discussions on this direction.

Spatial/geometric quantile is another concept of multivariate quantile pineered by Chaudhuri [Cha96], and later developed in [MO95; MOT97; Kol97; CM97; BMG14; CC96; CC17; CC14; KP23] and many others. The definition starts from the form of traditional check function $\rho_\tau(\cdot)$. Note for any $z \in \mathbb{R}$ the traditional check function can be rewritten in the following way

$$\rho_\tau(z) = \frac{1}{2}(|z| + (2\tau - 1)z) = \frac{1}{2}(|z| + vz),$$

where $v = 2\tau - 1$. Thus a natural extension of the check function to the multi-dimensional case is

$$\Phi_v(z) := \frac{1}{2}(\|z\| + v^\top z),$$

where $v = \tau u$, and $u \in \mathcal{S}^{d-1}$. Therefore, given $\tau \in (0, 1)$ and $u \in \mathcal{S}^{d-1}$, for a multivariate random variable Y, we may define its τu-quantile as

$$q_{Y;\mathrm{SP}}(\tau u) := \operatorname*{arg\,min}_{y\in\mathbb{R}^d} \mathbb{E}\big[\Phi_{\tau u}(Y - y) - \Phi_{\tau u}(Y)\big],$$

where the additional $\Phi_{\tau u}(Y)$ term is to make sure the optimisation problem is well-defined even for Y does not have finite first moment. In fact, by differentiating with respect to y, the solution of the above is equivalent to the following equation

$$\mathbb{E}\Big[\frac{Y - q_{Y;\mathrm{SP}}(\tau u)}{\|Y - q_{Y;\mathrm{SP}}(\tau u)\|}\Big] = -\tau u.$$

Intuitively speaking, this indicates that $q_{Y;\mathrm{SP}}(\tau u)$ defines a point in $\mathbb{R}^d$ such that the average unit vector from it to other random samples is τu. The generalisation to quantile regression setting is quite straightforward by applying the spatial quantile definition to $Y - bX$. However, similar to the concept proposed in Kong and Mizera [KM12b], given a quantile level, in order to come up with a quantile contour via the definition above one need to sample a large number of unit vector on $\mathcal{S}^{d-1}$, which makes it computationally inefficient.

The idea of depth-based methods also lies at the core of multivariate quantile/rank theory. Unlike spatial methods or projection-based approaches, which attempt to generalise the analytic formulation of univariate quantiles, statistical depth describes multivariate data from a geometric perspective. Tukey [Tuk75] first introduced the concept of halfspace depth in 1970s, and was popularised by Donoho and Gasko [DG92]. Later, Liu [Liu90] proposed simplicial depth, Liu [Liu92], Zuo and Serfling [ZS00], and Zuo [Zuo03] considered projection depth, Vardi and Zhang [VZ00] proposed spatial depth. For the regression setting, pioneer regression depth was proposed in Rousseeuw and Hubert [RH99], and we refer the readers to Zuo [Zuo21] for a comprehensive survey. Although depth-based methods typically offer appealing geometric intuition and affine-invariance properties, their implementation frequently results in computationally inefficient algorithms. Furthermore, most depth formulations can only characterise convex support regions, creating limitations when analysing distributions supported on non-convex domains. Chernozhukov et al. [Che+17] and Hallin et al. [Hal+21] addressed this limitation by introducing the Monge-Kantorovich depth, founded on measure transportation theory, which effectively captures non-convex support structures. We will investigate this innovative approach in detail in Section 1.3.

Among the various concepts of multivariate quantile mentioned above and their applications in regression settings, two significant problems persist: 1) the constructed quantile contours cannot capture potential non-convexity in the distribution of interest; and 2) most concepts are suitable only for non-parametric quantile regression, with few can be adopted for the problem of linear model coefficient estimation. In Chapter 2, we propose a regression coefficient estimation method that leverages the concept of Monge-Kantorovich rank/quantile, effectively addressing these two problems.

1.2.2 Nonparametric independence testing for random vectors.

In Section 1.1.2, we introduced the problem of nonparametric independence testing under univariate marginal distributions. Many approaches have been proposed for nonparametric independence testing under the multivariate setting. Specifically, we consider the same hypothesis as in (1.7), but extend the framework for $\max\{d_1, d_2\} \geq 2$.

The problem turns out to be much more challenging, and the solutions have not been discovered until recent years. The first significant line of work was initiated by Székely, Rizzo, and Bakirov [SRB07], who proposed a new dependence measure termed *distance covariance* for any

two random vectors with finite first moments. Tests based on distance covariance offer many appealing properties, such as computational efficiency and consistency against all alternatives with finite means; see [MS19] for further discussions. Subsequently, Székely and Rizzo [SR09] explored a generalisation to stochastic processes, while Lyons [Lyo13] and Jakobsen [Jak17] developed generalisations to general metric spaces. Another distance-based method propose recently is [KBW20], where the authors proposed a projection test statistics based on Cramér-von Mises divergence. The second important track of work is kernel-based methods, pioneered by Gretton et al. [Gre+05b; Gre+07], where they proposed the HSIC, and Gretton et al. [Gre+12] introduced a class of distance between probability measures, call *maxiaml mean discrepancy (MMD)*. Interestingly, in [Sej+13], the authors showed an equivalence between the distance-based method and the kernel-based method in general metric spaces. Other constructions include pariwise distance based methods [HGH12], Wasserstein distance-based method [Wie22; MS20; MS22; Oza+19; XW19]. Rank-based method is another active area in recent years. The key challenge is to generalise the concept of rank on the real line to higher dimensional space so that one can adopt those traditional rank-based methods mentioned in Section 1.1.2 to the case of multivariate marginal distributions. In the rest of this section, we will concentrate on this specific direction since it is more relevant to later chapters of this essasy.

Rank-based method is mainly classifed into two categories: graph-based method and OT-based method. The correlation coefficient proposed by Chatterjee [Cha21] and Azadkia and Chatterjee [AC21] initiate a line of work based on k-nearest neighbor graph. To clarify, let's revisit the limit of Chatterjee's correlation coefficient defined in (1.9) and examine its numerator:

$$\int \mathrm{Var}(\mathbb{E}(\mathbb{1}_{\{Y\geq t\}} \mid X))\, \mathrm{d}\mu(t) = \int \mathbb{E}\big(\mathbb{P}^2(Y \geq t \mid X)\big)\, \mathrm{d}\mu(t) - \int \mathbb{P}^2(Y \geq t)\, \mathrm{d}\mu(t).$$

A key step used in [Cha21] to approximate the conditional probability $\mathbb{P}^2(Y \geq t \mid X)$ on the right-hand side of the above is the following approximation:

$$\mathbb{P}(Y_1 \geq t) \approx \mathbb{P}(Y_{N(1)} \geq t),$$

where $N(1)$ is the index $i \in [n] \setminus \{1\}$ such that X_i is the nearest neighbor of X_1. Heuristically, the distribution of Y_1 should be close to the conditional distribution of $Y_{N(1)}$ due to the closeness of X_1 and $X_{N(1)}$ ([See Cha21, Corollary A.9.]). By leveraging this idea, Deb, Ghosal, and Sen [DGS20] and Huang, Deb, and Sen [HDS22] constructed new dependency measurement and the corresponding empirical estimation for independence and conditional independence test on general metric spaces, and Deb, Ghosal, and Sen [DGS20] also generalise Chatterjee's correlation to the case of multivariate marginal distributions. The graph-based methods have many desirable properties including distribution-freeness, strong consistency, computational efficiency, etc. However, as pointed out in [SDH22b; SDH24; ADN21; CB20], graph-based methods can lack power in certain scenarios where traditional coefficients, such as Kendall's τ, achieve optimal rates. Several modifications have been proposed to enhance the power performance of graph-based methods [LH23].

Another class of rank-based methods relies on the concept of multivariate rank. This direction has flourished in recent years, particularly following Chernozhukov et al. [Che+17] and Hallin et al. [Hal+21]'s introduction of OT-based multivariate rank/quantile. We will explore this approach in Section 1.3.3, after briefly introducing the OT-based ranks/quantiles.

1.3 OT-based rank/quantile and its applications

In the last section, we reviewed traditional statistical methods for addressing multivariate multiple-output quantile regression and nonparametric independence testing with multivariate marginal distributions. For the regression problem, existing frameworks yield only convex quantile level sets, which becomes problematic when the underlying random error follows a distribution with non-convex support. For the independence testing problem, most previous approaches fail to achieve exact or even asymptotic distribution-free properties, requiring permutation-based methods to compute p-values, leading to high computational costs.

Chernozhukov et al. [Che+17] and Hallin et al. [Hal+21] introduced a new multivariate rank/quantile framework based on OT theory. This approach preserves many of the desirable properties that have made univariate rank/quantile methods successful. In this section, we first introduce this new concept of rank/quantile in Section 1.3.1, then dive into its applications in quantile regression and independence testing in Section 1.3.2 and 1.3.3.

1.3.1 OT-based rank/quantile

We first give a very brief overview of OT theory. Given $\mu \in \mathcal{P}(\mathbb{R}^d)$ is a probability measure on $\mathbb{R}^d$. For any map $T : \mathbb{R}^d \to \mathbb{R}^d$, we define $\nu := T\#\mu$ to be another probability measure on $\mathbb{R}^d$ such that $(T\#\mu)(A) = \mu(T^{-1}(A))$ for all $A \subset \mathcal{B}$. We call ν is a *push-forward measure* of μ under T. Thus, on the space of random variable, if $X \sim \mu$, then $T(X) \sim \nu$.

The OT problem was first formulated in 1781 by Monge [Mon81], i.e. the Monge's problem. Specifically, we consider the following optimisation problem:

$$\inf_T \int_{\mathbb{R}^d} \|x - T(x)\|^2 \, \mathrm{d}\mu(x) \quad \text{such that } T\#\mu = \nu.$$

We call the solution of this problem as the *optimal transport map*. Then a natural question concerns the existence and uniqueness of solutions to this optimisation problem. Unfortunately, solutions do not always exist, and the problem itself can be even be ill-posed. For instance, one can never find a pushfoward map from a finite discrete distribution to a continuous distribution. However, the following phenomenal theorem guarantees the existence when μ and ν are absolutely continuous.

Theorem 1.1 (Brenier's Theorem). *Let μ, ν be absolutely continuous probability measures on $\mathbb{R}^d$, with finite second moments. Then there exists a convex function φ and its Legendre conjugate $\varphi^*(v) = \sup_{u \in \mathbb{R}^d}\{\langle v, u\rangle - \varphi(u)\}$ for $v \in \mathbb{R}^d$, such that the maps $R := \nabla\varphi$ and $Q := \nabla\varphi^*$ satisfy: 1) R is the optimal transport map from μ to ν, Q is the optimal transport map from ν to μ; 2) R and Q are almost everywhere unique; 3) $R \circ Q(x) = x$ and $Q \circ R(y) = y$, for $x, y \in \mathbb{R}^d$ almost everywhere.*

Note that, when $d = 1$, the cumulative distribution function of μ is the optimal transport map between μ to the uniform distribution $\mathrm{Unif}([0,1])$, and the quantile function is the optimal transport map from $\mathrm{Unif}([0,1])$ to μ. Based on this instructive observation and Theorem 1.1, define a new notion of multivariate rank/quantile in the population level can be defined as follows.

Definition 1.1 ([Che+17; Hal+21]). Given ν to be an absolutely continuous reference distribution. Then for any absolutely continuous distribution μ, the optimal transport map from μ to ν is defined to be the population rank map and the optimal transport map from ν to μ is defined to be the population quantile map.

Note the definition allows multiple choices for the reference distribution ν, including uniform sphere distribution, $\mathrm{Unif}([0,1]^d)$, and even normal distribution. There is no optimal choice for the reference distribution, as it depends on the specific problem to be solved. For example, Deb and Sen [DS23] uses $\mathrm{Unif}([0,1]^d)$ for theoretical analysis convenience, while Hallin et al. [Hal+21] employs a spherical uniform distribution to construct a center-outward rank for maximal ancillarity. Moreover, the population distribution μ is typically not available in practice, instead, we only have access to a sample of data $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \mu$. Therefore, one need to come up with an empirical estimation of the rank/quantile map. There are two types of approaches to define the empirical multivariate ranks and quantiles using the notion in Definition 1.1: discrete-discrete type [DS23; Hal+21], i.e. discrete dataset with discrete samples from the reference distribution, discrete-continuous type [GS22], i.e. discrete dataset with continuous reference distribution. We summarise these two different notions here.

For the discrete-discrete type of rank, estimating the population rank of $X_1, \ldots, X_n$ requires a set of d-dimensional discrete points $\mathcal{U} = \{u_1, \ldots, u_n\}$. The empirical population rank is then obtained as the optimal transport map between two empirical distributions: $\frac{1}{n}\sum_{i=1}^n \delta_{X_i}$ and $\frac{1}{n}\sum_{i=1}^n \delta_{u_i}$, which can be formulated as a *optimal assignment problem*. Specifically, let $\mathcal{S}_n$ be the set of all permutations of $[n]$, and

$$\pi^* := \operatorname*{arg\,min}_{\pi \in \mathcal{S}_n} \sum_{i=1}^n \|X_i - u_{\pi(i)}\|^2. \tag{1.13}$$

Then the empirical rank is

$$\widehat{R}^X(X_i) = u_{\pi^*(i)}, \quad \text{for } i = 1, \ldots, n,$$

and the empirical quantile map is its inverse. There are several possible choices for $\mathcal{U}$. One approach is to select $\mathcal{U}$ as a deterministic sequence that approximates the population reference distribution, such as $\mathrm{Unif}([0,1]^d)$. For example, Deb and Sen [DS23] uses the Halton sequence to approximate the uniform distribution on $[0,1]^d$, while Hallin et al. [Hal+21] employs a fixed regular grid to approximate a spherical uniform distribution. In this case, one may usually prefer to choose a $\mathcal{U}$ that has a low approximation error to the population reference distribution and computationally efficient. The readers may find more discussion from [DS23, Section D.3]. Another strategy is that instead of using deterministic sequence, one can simply draw a set of random samples from the reference distribution [Che+17]. While empirical ranks constructed this way offer theoretical advantages, they lose deterministic properties due to their random nature. Moreover, this approach requires recalculation of all samples whenever the sample size n increases by one. While these approaches differ in their theory and implementation, the resulting empirical ranks/quantiles both share the elegant property of distribution-freeness.

For the discrete-continuous rank type, we calculate the empirical rank map directly by finding the optimal transport map between the empirical measure $\frac{1}{n}\sum_{i=1}^n \delta_{X_i}$ and an absolutely continuous reference distribution ν. This approach eliminates the need to approximate the reference

distribution using a set of data points $\mathcal{U}$. By Theorem 1.1, the quantile function, i.e. the optimal transport map from ν to $\frac{1}{n}\sum_{i=1}^n \delta_{X_i}$ can be a.e.-uniquely obtained by

$$\widehat{Q}^X := \arg\min_T \int_{\mathbb{R}^d} \|x - T(x)\|^2 \, \mathrm{d}\nu(x) \quad \text{such that } T\#\nu = \frac{1}{n}\sum_{i=1}^n \delta_{X_i}. \tag{1.14}$$

Moreover, by Theorem 1.1, except a zero mass set, $\widehat{Q}^X$ can be expressed as $\widehat{Q}^X = \nabla\hat{\varphi}$ for some convex function $\hat{\varphi} : \mathbb{R}^d \to \mathbb{R}\cup\{+\infty\}$; when $\hat{\varphi}$ is not differentiable at some point $u \in \mathbb{R}^d$, $\widehat{Q}^X(u)$ is defined to be any point within the subdifferential of $\hat{\varphi}$, i.e. $\partial\hat{\varphi}(u)$. Therefore, $\widehat{Q}^X$ defines a partition of $\mathrm{Supp}(\nu) = \cup_{i=1}^n P_i$, where $P_i = \{u \in \mathrm{Supp}(\nu) : \widehat{Q}(u) = X_i\}$, and $\nu(P_i) = 1/n$. Then suppose $\hat{\varphi}^* : \mathbb{R}^d \to \mathbb{R}$ is the Legendre conjugate of $\hat{\varphi}$, i.e. $\hat{\varphi}^*(v) = \sup_{u\in\mathbb{R}^d}\{\langle u, v\rangle - \hat{\varphi}(u)\}$, the empirical rank transformation $\widehat{R}^X : \mathbb{R}^d \to \mathbb{R}^d$ is defined to be

$$\widehat{R}^X := \nabla\hat{\varphi}^*.$$

However, due to the non-continuity of the empirical measure, the rank transformation might not be uniquely defined on $X_1, \ldots, X_n$. In fact, $\widehat{R}^X(X_i)$ can be any point of set P_i. In [GS22], the authors define $\widehat{R}^X(X_i)$ to be a random point drawn from the conditional distribution of ν restricted on P_i, i.e.

$$\widehat{R}^X(X_i) \mid X_1, \ldots, X_n \sim \nu_i,$$

where $\nu_i(B) = n\nu(P_i \cap B)$ for any $B \in \mathcal{B}$. In particular, if ν is a uniform distribution, then $\widehat{R}^X(X_i)$ is drawn uniformly from P_i. Although this specific choice of $\widehat{R}^X(X_i)$ introduce additional randomness, it is shown in [GS22, Lemma 3.4] that the resulting marginal distribution of the multivariate rank is indeed distribution-free.

Comparing two types of empirical OT-based rank/quantile maps above, the discrete-discrete type of method immediately generates the rank map and quantile map, and maintains distribution-freeness. However, it does not lead to a smooth quantile function/contour, some smoothing interpolation techniques are required [Hal+21]. In contrast, the second method naturally yields to a notion of quantile function, but does not automatically lead to an empirical rank map. From a computational perspective, the optimisation problem (1.13) can be solved by classical Hungairan algorithm with a time complexity $O(n^3)$ (faster approximation solution is possible [see e.g. SDH22a, Section 5]), while the semi-continuous OT problem involves the construction of *power diagram* thus result in an algorithm with worst complexity of $O(n^{\lfloor d/2\rfloor})$ for $d > 2$; see [Aur87] for details.

Compared with other notion of multivariate ranks/quantile mentioned in Section 1.2.1, the new concept based on OT enjoys many desirable properties including exact distribution-freeness (see [Hal+21, Proposition 2.5], [DS23, Proposition 2.2] and [GS22, Lemma 3.4]) and strong consistency (see [Hal+21, Proposition 2.4], [DS23, Theorem 2.1], [GS22, Theorem 4.1]). Moreover, unlike other multivariate quantile notion, the OT-based quantile contour is not constructed via intersection of halfspace or averaging directions, thus it can capture the non-convexity support of the interest distribution [Che+17; BSH24], which makes the induced statistical methods more robust and flexible in handling complex data structures.

In the next two sections, we review some applications of this new notion of multivariate ranks/quantiles in modern multivariate statistics, especially in the context of quantile regression and independence testing.

1.3.2 Multiple-output quantile regression via OT-based quantile function

A pioneer work attempts to generalise quantile regression to the case of multiple-output linear model via OT theory was conducted by Carlier, Chernozhukov, and Galichon [CCG16]. The authors consider a heteroskedastic linear model

$$q_{Y|X}(U) = \beta_*(U)^\top X, \tag{1.15}$$

where $Y, U \in \mathbb{R}^d$ and $X \in \mathbb{R}^p$ is independent with U, $q_{Y|X} : [0,1] \times \mathbb{R}^p \to \mathbb{R}$ is the conditional quantile function of Y defined as $q_{Y|X=x}(\tau) = \inf\{y : \mathbb{P}(Y \le y | X = x) \ge \tau\}$ for any $\tau \in [0,1]$, $x \in \mathbb{R}^p$, and $\beta_* : \mathbb{R}^d \to \mathbb{R}^{p\times d}$ is the regression coefficient functional. The authors observed that if equation (1.15) holds, then U is a solution to the optimisation problem:

$$\max\{\mathbb{E}(V^\top Y) : V \sim F_U,\ \mathbb{E}(X|V) = \mathbb{E}\, X\}. \tag{1.16}$$

Furthermore, solving the dual problem of the above yields the true coefficient functional β_* from (1.15) (see [CCG16, Theorem 3.2]). Although the authors didn't mention the concept of multivariate quantile, the optimisation problem (1.16) implicitly solves for the optimal transport map between Y and U.

Building on the optimal transport (OT)-based quantile framework introduced in [Hal+21], Barrio, Sanz, and Hallin [BSH24] developed a non-parametric approach for multiple-output quantile regression. Their method constructs a smooth interpolation based on the empirical quantile map to generate quantile tubes in multi-dimensional spaces. This approach was later extended to manifold settings by Hallin and Liu [HL24]. However, both of them focus on non-parametric quantile regression and concentrate on estimating the quantile contours rather than focusing on the robust estimation of the regression coefficients defined in (1.10).

In Chapter 2, we propose a multiple-output composite quantile regression estimator based on optimal transport theory. To the best of our knowledge, this is the first work applies the idea of OT-based quantiles to robust coefficient estimation in multiple-output linear models.

1.3.3 Multivariate nonparametric independence testing via OT-based rank

The emergence of OT-based multivariate ranks has inspired many new distribution-free approaches on independence testing between two random vectors. Given $(X_1, Y_1), \ldots, (X_n, Y_n) \overset{\text{iid}}{\sim} P^{(X,Y)}$, we denote $Z_i = (X_i, Y_i)$ for $i = 1, \ldots, n$, consider the hypothesis (1.7) with $\max\{d_1, d_2\} > 1$.

Ghosal and Sen [GS22] proposed a test statistic, which measures the L_2-distance between the OT-based rank of Z_i and the product of OT-based ranks of X_i and Y_i, and similar test statistic can be constructed for independence testing of more than two samples. However, the asymptotic theory under the null is not established, thus a permutation-based method is required to obtain the p-value. A more systematic investigate is carried out by Deb and Sen [DS23], where both two-sample goodness-of-fit test and two-sample independence test are considered. The authors revisit the distance covariance by [SRB07]. As mentioned in Section 1.1.2, a notable drawback of distance covariance method is that its null distribution depend on the marginal distribution P^X and P^Y. As a consequence, the test are no longer distribution-free and permutation analysis has to be conducted in order to implement them. To overcome this, Deb and Sen [DS23] developed an

asymptotically distribution-free test statistic by replacing the original observations X_i and Y_i in the empirical distance covariance with their respective OT-based multivariate ranks $\widehat{R}^X(X_i)$ and $\widehat{R}^Y(Y_i)$, enabling explicit p-value computation. Independently, Shi, Drton, and Han [SDH22a] developed a similar test statistic, though they employed a different concept of OT-based multivariate rank with a different reference distribution than that used in Deb and Sen [DS23]. Building on this line of research, Shi et al. [Shi+22] established a general framework for designing consistent independence tests using the center-outward OT-based rank proposed by Hallin et al. [Hal+21]. Another work by Shi et al. [Shi+21], take up the same concept of multivariate rank to develop the multivariate analogues of the sign quadrant statistic, Kedall's tau and Spearman's rho. Their work not only proves the asymptotic distribution-free property of these tests but also establishes a multivariate Chernoff-Savage type lower bound to demonstrate the relative efficiency of the proposed test.

However most of the coefficient in the above can only characterise the potential functional correlation between X and Y, for instance, see Property (I) for Chatterjee's coefficient. In the case of spurious correlation through some confounders, the effectiveness of these methods are unkonwn. We bridge this gap by proposing a new coefficient of correlation for random vectors through OT-based rank in Chapter 3. Unlike many existing coefficients, the proposed coefficient can not only detect functional correlations, but also spurious correlations. Moreover, thanks to the distribution-freeness of OT-based rank, the proposed coefficient enjoys a simple asymptotic theory without any assumptions on the marginal distributions, which avoiding any permutation techniques to obtain p-values.

1.4 Overview

The rest of this essay is organised as follows. Chapter 2 presents a robust method for estimating coefficients in multiple-output linear models, which extends univariate CQR to multivariate response variables by employing OT-based techniques. In Chapter 3, by leveraging the OT-based multivariate ranks, we propose a new correlation coefficient for measuring dependence between two random vectors which enjoys a simple form and a distribution-free asymptotic theory under the null.

Chapter 2

Multiple-output quantile regression via optimal transport

2.1 Introduction

The area of robust statistics has seen a revival of interest in recent years, both in Statistics and Computer Science. This is partly due to the fact that the massive surge in data volumes brings about a significant demand for efficient and precise analysis of heavy-tailed or partially corrupted data [ENK16; WPL15; Sze+14]. Compared to earlier works in this area pioneered by Tukey and McLaughlin [TM63] and Huber [Hub64; Hub65], modern treatment of this topic focuses more on handling multivariate data. For instance, in the area of robust mean estimation, Diakonikolas, Kane, and Pensia [DKP20], Lugosi and Mendelson [LM21], Depersin and Lecué [DL22], and Minasyan and Zhivotovskiy [MZ23] have proposed various extensions of univariate robust mean procedures such as the trimmed mean estimator [TM63] and median of means estimator [NY83; JVV86; AMS96] to the multivariate setting. We witness a similar surge in research interest in the area of robust covariance estimation [MZ20; AZ22; MZ23].

In this work, we focus on the topic of robust linear regression with potentially multivariate response variable, where a covariate-response pair $(X, Y) \in \mathbb{R}^p \times \mathbb{R}^d$ with joint distribution $P^{(X,Y)}$ is generated from

$$Y = b^* X + \varepsilon, \tag{2.1}$$

with regression coefficients $b^* \in \mathbb{R}^{d\times p}$, a zero-mean covariate vector $X \in \mathbb{R}^p$ and a noise vector ε taking values in $\mathbb{R}^d$ independent of X. Given independent and identically distributed (i.i.d.) covariate-response pairs $(X_1, Y_1), \dots (X_n, Y_n)$ drawn from $P^{(X,Y)}$, our goal is to estimate b^*. The contamination of a linear model is mainly captured by two different mechanisms: heavy-tailed noise [Cat12; LM19] and outlier contamination [Sze+14; Hub04]. When $d = 1$, both directions have thrived in recent years [NT13; FLW17; SZF20; SF20; PJL20; Ado+23]. However, in the context of multiple-output linear regression, where $d > 1$, the literature is notably scant. In this work, we go beyond the case of the univariate response variable to the case of the multiple-output linear model under possibly heavy-tailed noise.

One popular way to tackle the heavy-tailed error is based on the quantile regression [KB78; WLJ07; LZ08; ZY08; WL09; BC11]. In the case of univariate linear regression, although the ordinary least square (OLS) estimator is widely recognized as the best unbiased estimator when

the random error follows a Gaussian distribution since it attains the Cramer–Rao lower bound, it may not perform well when the random error is heavy-tailed, as the mean squared error of the OLS estimator is proportional to the second moment of the random error term. This issue can be addressed by using the quantile regression estimator [KB78]. Unlike the OLS estimator, which estimates the conditional mean function, the quantile regression estimator aims to estimate the conditional quantile function of Y given X. Thanks to the robustness of quantiles, the quantile regression estimator is less affected by outliers or heavy-tailed distributions. However, the relative efficiency of the quantile regression estimator compared to the OLS estimator, i.e. the asymptotic variance of OLS estimator to that of the CQR estimator, can be arbitrarily small based on their respective asymptotic variances. Zou and Yuan [ZY08] proposed a solution to this issue through the composite quantile regression (CQR) method, whose loss function aggregates multiple quantile regression loss functions. Specifically, for $d = 1$ and any $K \in \mathbb{N}$, the CQR estimator $\tilde{b}$ is obtained by the following optimization problem

$$(\hat{q}_1, \dots, \hat{q}_K, \tilde{b}) = \underset{q_1,\dots,q_K \in \mathbb{R},\ b \in \mathbb{R}^{d\times p}}{\arg\min} \sum_{i=1}^{n} \sum_{k=1}^{K} \rho_{\tau_k}(Y_i - bX_i - q_k), \tag{2.2}$$

where $\rho_\tau(t)$ is the so-called check function defined as $\rho_\tau(t) = \max\{t, 0\} + (\tau - 1)t$ for any $t \in \mathbb{R}$, and $\tau_k = k/(K+1)$. Zou and Yuan [ZY08] showed that the CQR estimator can achieve at least 70% relative efficiency compared to the OLS estimator even for Gaussian noise. However, when $d \geq 2$, the CQR estimator $\tilde{b}$ does not have a natural extension due to the lack of a proper definition for multivariate rank/quantile and the corresponding multivariate check function.

One of the key contributions of this study is the development of a multiple-output composite quantile regression (MCQR) estimator. The definition of our proposed estimator is closely related to the concept of the Monge–Kantorovich (MK) ranks/quantiles, which are multivariate generalization of ranks and quantiles from the view of optimal transport developed by Chernozhukov et al. [Che+17] and Hallin et al. [Hal+21]. Intuitively, the univariate cumulative distribution function (CDF) and the quantile function of any probability distribution P^X can be viewed as optimal transport maps between P^X and a reference distribution, e.g. the uniform distribution $U[0, 1]$. This perspective allows for a natural extension of ranks and quantiles to multivariate distributions. Compared to many previous extensions based on Tukey's depth [Tuk75], MK-ranks/quantiles have several advantages, including the ability to capture more complex and possibly non-convex quantile contours and allowing for distribution-free inference in multivariate settings. Please refer to Hallin [Hal22] for a comprehensive introduction to the MK-ranks/quantiles.

A crucial observation in constructing our MCQR estimation is that the univariate CQR loss function can be equivalently described as the *Wasserstein product* between the empirical distribution of the residuals $(Y_i - bX_i : i = 1, \dots, n)$ and the uniform distribution $U[0, 1]$. Here, the 'Wasserstein product' between two distributions P and Q is the maximum of $\mathbb{E}(XY)$ over all couplings (X, Y) with marginal distributions $X \sim P$ and $Y \sim Q$. When Q is viewed as a reference distribution, this optimal coupling is exactly the same as in MK-quantiles. See (2.4) for a formal definition and more detailed discussion. This alternative viewpoint allows us to circumvent the need of defining individual multivariate check functions and instead formulate the MCQR loss in terms of the MK-quantiles. It is worthwhile to note that while various previous studies in the literature have attempted to extend the concept of quantile regression to the multiple-output setting [HPŠ10; KM12a; Hal+15; CCG16; BSH24], the majority have concentrated on estimating

the quantile contours rather than focusing on the robust estimation of the regression coefficients. See Section 2.2 for a more detailed discussion of our proposed method.

Then in Section 2.3 we investigate the theoretical guarantees of the MCQR estimator. We first prove the consistency result when the random noise is only assumed to have finite ℓ-th moment for some $\ell > 2$ (see Theorem 2.1). Then a faster convergence rate is established when we assume a noise distribution with a sub-Weibull tail (see Theorem 2.2). We highlight that the MCQR procedure represents an M-estimation problem incorporating the Wasserstein distance within its loss function, for which the empirical process theory tools used in traditional M-estimators are not directly applicable. To the best of our knowledge, Theorem 2.1 and Theorem 2.2 are the first results that establish the consistency and convergence rate of an M-estimation where the loss function involves the 2-Wasserstein distance. New theoretical tools were developed along the way, which we believe may be of independent interest in future research. Please refer to Section 2.3 for detailed descriptions of the Theorems and proof sketches.

2.1.1 Related works

Various definitions of multiple-output quantile regression have been proposed in the past, including the depth-based directional method [HPŠ10; KM12a; Hal+15], the M-quantile [Kol97], the spatial quantile [Cha96; CC14], among others. As remarked above, unlike our work, all these approaches focus on estimating the quantile contours of the response variable. In addition, these definition of multivariate quantiles do not preserve the quintessential attributes of the univariate quantile, notably distribution-freeness and the Glivenko-Cantelli property [Hal+21]. Furthermore, their quantile contours are constrained to be convex, which hinders performance when data distribution exhibits non-convex level sets.

In contrast, Chernozhukov et al. [Che+17] and Hallin et al. [Hal+21] introduced a novel multivariate quantile/rank framework based on optimal transport. This framework adeptly captures level set non-convexities while retaining the distribution-freeness and the Glivenko-Cantelli property, hallmarks of the univariate rank/quantile [Che+17; Hal+21]. Several applications in multivariate statistics have been established successfully [DS21; BSH24; HHH23; Shi+24]. We refer to a comprehensive survey [Hal22] and references therein. Building upon this groundwork, Carlier, Chernozhukov, and Galichon [CCG16] and Barrio, Sanz, and Hallin [BSH24] proposed two notions of multiple-output quantile regression, though concentrating primarily on the estimation of conditional quantile functions rather than the regression coefficients themselves.

2.1.2 Notation

For $n \in \mathbb{N}$, write $[n] := \{1, \dots, n\}$. For any vector $v \in \mathbb{R}^d$, we write $\|v\| := (\sum_{j\in[d]} v_j^2)^{1/2}$. For any matrix $M \in \mathbb{R}^{p\times d}$, we define $\|M\|_{\mathrm{F}} := \big(\mathrm{Tr}\big(M^\top M\big)\big)^{1/2}$. We denote $\mathcal{S}^{d-1}$ to be the unit sphere in $\mathbb{R}^d$. For any measurable function $f : X \to \mathbb{R}$, we denote $f^+(x) := \max\{f(x), 0\}$ as its positive part, and $f^-(x) := \max\{-f(x), 0\}$ as its negative part. We write $\mathcal{B}$ as the Borel σ-algebra of $\mathbb{R}^d$. Write $\mathcal{P}_\ell(\mathbb{R}^d)$ as the set of Borel probability measures defined on $(\mathbb{R}^d, \mathcal{B})$ with finite ℓ-th order moments for $\ell \in \mathbb{N}$ and $\mathcal{P}_{ac}(\mathbb{R}^d)$ be the set of probability measures on the same space that are absolutely continuous with respect to the Lebesgue measure. For any random variable X on $\mathbb{R}^d$, write P^X for the associated probability measure and $P_n^X := \frac{1}{n}\sum_{i=1}^n \delta_{X_i}$ for the associated

empirical distribution where $X_1, \dots, X_n$ are n independent copies of X and δ_x denote the Dirac measure on x.

2.2 MCQR

In this section, we present a generalization of the traditional CQR when the dimension of the response variable d is greater than 1. We start by revisiting the univariate CQR estimator, and showing that at the population level, it can be seen as the minimizer of the Wasserstein product between P^{Y-bX} and the uniform reference distribution $U[0,1]$, which allows a multivariate generalization. Moreover, we justify that the choice of the reference distribution does not affect the population minimizer in this problem, thus allowing us to select more natural reference distributions in multivariate settings.

2.2.1 Univariate CQR revisited

Since $q_1, \dots, q_K$ in (2.2) have the interpretation of quantiles associated with $\tau_1, \dots, \tau_K$, it is natural to further constrain the optimization by assuming $q_1 \le \cdots \le q_K$. Let $\mathcal{M}$ denote the set of all increasing functions on $\mathbb{R}$, then (2.2) with this additional constraint can be viewed as the empirical version of the following optimization problem

$$\operatorname*{arg\,min}_{q\in\mathcal{M},\, b\in\mathbb{R}^{1\times p}} \mathbb{E}\Big\{\rho_T\big(Y - bX - q(T)\big)\Big\} = \operatorname*{arg\,min}_{q\in\mathcal{M},\, b\in\mathbb{R}^{1\times p}} \mathbb{E}\Big\{\int_0^1 \rho_\tau\big(Y - bX - q(\tau)\big)\mathrm{d}\tau\Big\}, \tag{2.3}$$

where $(X, Y) \sim P^{(X,Y)}$ and $T \sim U[0,1]$. The following lemma indicates that, when $d = 1$, the true regression coefficient b^* in (2.1) and the quantile function $q^*_\varepsilon : \tau \mapsto \inf\{y \in \mathbb{R} : P^\varepsilon(-\infty, y] \ge \tau\}$ of ε form a solution of (2.3). As we will see from Lemma 2.2 and Proposition 2.1, this is actually the unique solution to the problem.

Lemma 2.1. *Under the linear model (2.1), we have*

$$(b^*, q^*_\varepsilon) \in \operatorname*{arg\,min}_{b\in\mathbb{R}^{1\times p},\, q\in\mathcal{M}} \mathbb{E}\int_0^1 \rho_\tau(Y - bX - q(\tau))\,\mathrm{d}\tau.$$

In fact, an inspection of the proof (see Section 2.5.3) of the above lemma reveals that if $\tau_1, \dots, \tau_K$ converges to a distribution P^Z with support $\mathcal{Z}$ rather than to $U[0,1]$, then a similar result to Lemma 2.1 holds provided that we modify the convex check functions $\rho_\tau : \mathbb{R} \to \mathbb{R}^+$ for $\tau \in \mathcal{Z}$ so that they satisfy $F_W^{-1} \circ F_Z(\tau) \in \arg\min_\theta \mathbb{E}\,\rho_\tau(W - \theta)$ for all random variables W with absolutely continuous distributions. However, generalizing the check functions beyond the univariate setting is difficult. While some attempts have been made [Cha96; Kol97], the resulting multivariate quantiles, defined through the minimizer of these generalized check functions, lack key properties of their univariate counterparts (see our discussion in Section 2.1.1, as well as empirical comparisons in Section 2.4). Instead, our work takes a different approach and generalizes the CQR population loss function as a whole rather than individual check functions. A key observation that allows us to achieve this is the following reformulation of the loss function of (2.3) in

Lemma 2.2 below. To state the lemma, we define the *Wasserstein product* between $P, Q \in \mathcal{P}_2(\mathbb{R}^d)$ as

$$\langle\!\langle P, Q \rangle\!\rangle_{\mathcal{W}_2} := \sup_{\gamma \in \mathcal{C}(P,Q)} \int \langle x, y \rangle \, \mathrm{d}\gamma(x, y), \tag{2.4}$$

where $\mathcal{C}(P, Q)$ denotes the set of all couplings between P and Q, i.e. for any $\gamma \in \mathcal{C}(P, Q)$, and measureable subsets $A, B \subset \mathbb{R}^d$, we have $\gamma(A \times \mathbb{R}^d) = P(A)$ and $\gamma(\mathbb{R}^d \times B) = Q(B)$. The name 'Wasserstein product' stems from its intrinsic link with the 2-Wasserstein distance: $\frac{1}{2}\mathcal{W}_2^2(P, Q) = \frac{1}{2}\int \|x\|^2 \, \mathrm{d}P(x) + \frac{1}{2}\int \|y\|^2 \, \mathrm{d}Q(y) - \langle\!\langle P, Q \rangle\!\rangle_{\mathcal{W}_2}$. We will often slightly abuse notation to write $\langle\!\langle X, Y \rangle\!\rangle_{\mathcal{W}_2}$ instead of $\langle\!\langle P^X, P^Y \rangle\!\rangle_{\mathcal{W}_2}$.

Lemma 2.2. *Suppose that $X \sim P^X$ is mean-zero with finite second moment. For $U \sim U[0, 1]$, and a fixed $b \in \mathbb{R}^{1\times p}$, we have*

$$\inf_{q \in \mathcal{M}} \mathbb{E}\Big\{ \int_0^1 \rho_\tau\big(Y - bX - q(\tau)\big) \, \mathrm{d}\tau \Big\} + \frac{1}{2}\mathbb{E}\, Y = \langle\!\langle Y - bX, U \rangle\!\rangle_{\mathcal{W}_2}.$$

The proof is deferred to Section 2.5.4. Writing $\mathcal{L}(b; U) := \langle\!\langle Y - bX, U \rangle\!\rangle_{\mathcal{W}_2}$, Lemma 2.2 and Equation (2.3) imply that, the optimizer in b for the population CQR loss function in (2.3) is equal to $\arg\min_{b \in \mathbb{R}^{d\times p}} \mathcal{L}(b; U)$ when $d = 1$.

2.2.2 Multiple-output CQR via optimal transport

With the help of Lemma 2.2, we may regard $\mathcal{L}(b; U)$ as a generalized population CQR loss function for the multiple-output case ($d \geq 2$) for suitably chosen reference random vector U. The following proposition (see Section 2.5.5 for proof) verifies that under a mild condition this loss has a unique minimizer and that is independent of the specific choice of U (see Section 2.7 for an intuitive illustration).

Proposition 2.1. *If $P^\varepsilon, P^U \in \mathcal{P}_2(\mathbb{R}^d) \cap \mathcal{P}_{ac}(\mathbb{R}^d)$ and P^X is not a point mass, then b^* is the unique minimizer of $\mathcal{L}(b; U)$.*

There are various choices of the reference distribution of U, including the uniform distribution on the unit cube [Che+17; DS21] and the spherical uniform distribution [Hal+21; BSH24]. In this paper, we opt for the standard multivariate normal distribution as the reference distribution, primarily motivated by its advantageous theoretical characteristics. Moreover, we will also omit the specification of the reference distribution in the loss function and simply write it as $\mathcal{L}(b)$ throughout the rest of the paper.

Proposition 2.1 motivates the following natural estimator of b^* based on the Wasserstein product of the empirical distributions.

Definition 2.1. Given i.i.d. covariate-response pairs $(X_1, Y_1), \ldots, (X_n, Y_n)$ generated as in (2.1) and a reference distribution $P^U \in \mathcal{P}_2(\mathbb{R}^d) \cap \mathcal{P}_{ac}(\mathbb{R}^d)$ and $U_1, \ldots, U_m \overset{i.i.d.}{\sim} P^U$, the *MCQR estimator* for b^* is defined as

$$\hat{b} \in \arg\min_{b \in \mathbb{R}^{d\times p}} \mathcal{L}_{n,m}(b), \text{ where } \mathcal{L}_{n,m}(b) := \langle\!\langle P_n^{Y-bX}, P_m^U \rangle\!\rangle_{\mathcal{W}_2}. \tag{2.5}$$

The optimization procedure above is an M-estimation problem. However, unlike classical M-estimation problems, the empirical loss function cannot be viewed as an empirical process of the population loss (in fact, $\mathbb{E}\langle\!\langle P_n^{Y-bX}, P_m^U \rangle\!\rangle_{\mathcal{W}_2} \neq \langle\!\langle P^{Y-bX}, P^U \rangle\!\rangle_{\mathcal{W}_2}$), which prevents us from applying traditional empirical process theory techniques to obtain the convergence rate results directly. Instead, a collection of new theoretical results is developed to better understand both the population and empirical version of the Wasserstein product loss. Please refer to Section 2.3 for more details. Secondly, it is worth noting that the empirical reference distribution P_m^U is distinct from the distribution of τ_k's in (2.2) when $d = 1$. Instead, we employ it as the reference distribution to redefine the distribution function and the quantile function (refer to Section 2.7 for an example). Thus, even when $d = 1$ with a uniform reference distribution, the plug-in estimator in (2.5) does not reduce to the univariate CQR estimator (2.2). This can also be seen from the proof of Lemma 2.2. Therefore, our proposed MCQR estimator (2.5) is different from the univariate CQR estimator that is studied in [ZY08] but shares the same loss function at the population level. See also Figure 2.3a and Figure 2.3b for an interesting difference in their robustness to contamination in one dimension.

2.2.3 Solving MCQR via linear programming

We describe here how the optimization problem can be solved in practice. Given $\{(X_i, Y_i)\}_{i=1}^n \subset \mathbb{R}^p \times \mathbb{R}^d$ and $\{U_i\}_{i=1}^m$, we define $X = (X_1, \ldots, X_n)^\top \in \mathbb{R}^{n\times p}$ and $Y = (Y_1, \ldots, Y_n)^\top \in \mathbb{R}^{n\times d}$ and $U = (U_1, \ldots, U_m)^\top \in \mathbb{R}^{m\times d}$. Define

$$\mathcal{C}_{n,m} = \{A \in \mathbb{R}_+^{m\times n} : A\mathbf{1}_n = \mathbf{1}_m/m \text{ and } A^\top \mathbf{1}_m = \mathbf{1}_n/n\}.$$

Every $\pi \in \mathcal{C}_{n,m}$ represents a coupling of $P_n^{(X,Y)}$ and P_m^U in the sense that $\pi_{i,j}$ denotes the mass to be transported from (X_i, Y_i) to U_j. Then by the definition of $\langle\!\langle \cdot, \cdot \rangle\!\rangle_{\mathcal{W}_2}$, the optimization problem in (2.5) can be written as

$$\begin{aligned}\min_{b\in\mathbb{R}^{d\times p}} \max_{\pi\in\mathcal{C}_{n,m}} \operatorname{Tr}\big(U^\top \pi (Y - Xb^\top)\big) &= \max_{\pi\in\mathcal{C}_{n,m}} \min_{b\in\mathbb{R}^{d\times p}} \operatorname{Tr}\big(U^\top \pi (Y - Xb^\top)\big) \\ &= \max_{\pi\in\mathcal{C}_{n,m}} \min_{b\in\mathbb{R}^{d\times p}} \big\{\operatorname{Tr}\big(U^\top \pi Y\big) - \operatorname{Tr}\big(U^\top \pi X b^\top\big)\big\},\end{aligned}$$

where the exchange of the minimum and maximum is allowed as the objective is linear [Neu28]. The dual formulation on the right-hand side is easier to handle since its inner minimum is equal to $-\infty$ unless $U^\top \pi X = 0$. Hence, the dual problem of (2.5) is

$$\begin{aligned}\max_{\pi\in\mathcal{C}_{n,m}} \quad & \operatorname{Tr}\big(U^\top \pi Y\big) \\ \text{s.t.} \quad & U^\top \pi X = 0,\end{aligned}$$

which can be solved by standard linear programming solvers. After obtaining the dual optimizer $\hat{\pi}$, the MCQR estimator $\hat{b}$ is obtained via complementary slackness.

2.3 Theory

In this section, we investigate the theoretical performance of the proposed estimator when adopting a standard Gaussian reference distribution $U \sim \mathcal{N}(0, I_d)$. In Theorem 2.1, we provide a non-asymptotic bound for the estimation error when only assuming a finite $2 + \delta$ moment condition

on the random noise term. Furthermore, we demonstrate in Theorem 2.2 that in cases where the distributions of both the covariates and the noise exhibit a sub-Weibull tail, the MCQR estimator enjoys a faster rate of convergence to the truth.

Given a positive definite matrix $\Sigma \in \mathbb{R}^{p\times p}$ and any matrix $A \in \mathbb{R}^{d\times p}$, we define the *matrix Mahalanobis norm* of A with respect to Σ as $\|A\|_\Sigma := \mathrm{Tr}^{1/2}(A\Sigma A^\top) = \|A\Sigma^{1/2}\|_{\mathrm{F}}$. We will assume throughout this section that $\mathbb{E}(XX^\top) = \Sigma$.

Assumption 1 *X follows an elliptical distribution, i.e., there exists independent random variable R on $\mathbb{R}_+$ and random vector $Q \sim U(\mathcal{S}^{d-1})$ such that $X = \Sigma^{1/2}QR$, and P^ε is absolutely continuous.*

Under this assumption, we first consider the case when the random noise ε is only assumed to satisfy a finite moment condition.

Theorem 2.1. *Suppose $(X, Y), (X_1, Y_1), \ldots, (X_n, Y_n)$ are i.i.d. pairs generated according to (2.1), $U_1, \ldots, U_m \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, I_d)$. Assume $m \geq n > 1$ and that Assumption 1 holds. If $P^X, P^\varepsilon \in \mathcal{P}_\ell(\mathbb{R}^d)$ for $\ell > 2$ then there exists $C > 0$ depending only on ℓ, d and p such that with probability at least $1 - 4(\log n)^{-1}$, the MCQR estimator defined in (2.5) satisfies*

$$\|\hat{b} - b^*\|_\Sigma^2 \wedge 1 \leq C\big(n^{-\frac{1}{4}} + n^{-\frac{1}{d\vee p}} + n^{-\frac{\ell-2}{2\ell}}\big)\log m.$$

An immediate consequence of Theorem 2.1 is that if taking n and m to be large enough such that

$$C\big(n^{-\frac{1}{4}} + n^{-\frac{1}{d\vee p}} + n^{-\frac{\ell-2}{2\ell}}\big)\log m < 1, \tag{2.6}$$

then we have

$$\|\hat{b} - b^*\|_\Sigma^2 \leq C\big(n^{-\frac{1}{4}} + n^{-\frac{1}{d\vee p}} + n^{-\frac{\ell-2}{2\ell}}\big)\log m \tag{2.7}$$

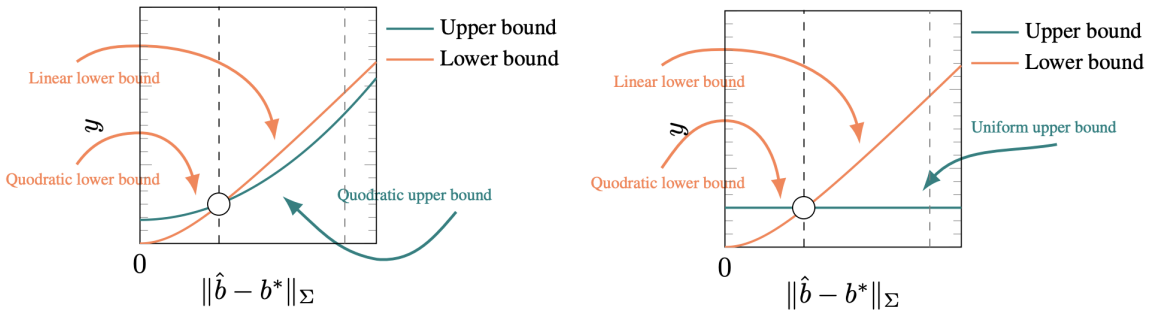
holds with probability at least $1 - 4(\log n)^{-1}$. We make a few remarks here. Firstly, to the best of our knowledge, this is the first consistency result for an M-estimator whose loss function involves a multivariate 2-Wasserstein distance term. Bernton et al. [Ber+19] studied the convergence rate and asymptotic distribution of a minimum Wasserstein estimator, but their result is restricted to 1-Wasserstein distance in the univariate setting, for which explicit characterization of the optimal transport is available. In our setting, the traditional M-estimator/Z-estimator argument [VW96, Chapter 3.2-3.3] that derives consistency and rate of convergence of an M-estimator by analyzing the curvature of the loss function is infeasible. Instead, our proof relies on several new lemmas that reveal important properties of the Wasserstein product.

To briefly sketch the proof of Theorem 2.1, we first introduce the following lemmas.

Lemma 2.3. *Let Z and ε be independent random vectors in $\mathbb{R}^d$ and $U \sim \mathcal{N}(0, I_d)$. If P^ε and P^Z are absolutely continuous with finite second moments, then*

$$\langle\!\langle Z + \varepsilon, U\rangle\!\rangle_{\mathcal{W}_2}^2 \geq \langle\!\langle Z, U\rangle\!\rangle_{\mathcal{W}_2}^2 + \langle\!\langle \varepsilon, U\rangle\!\rangle_{\mathcal{W}_2}^2.$$

This lemma is proved by constructing a sequence of couplings of the triple (Z, ε, U) via the Slepian smart path interpolation [see e.g. Ver18, Chapter 7.2.1]. The best induced coupling of $(Z + \varepsilon, U)$ provides the desired lower bound of $\langle\!\langle Z + \varepsilon, U\rangle\!\rangle_{\mathcal{W}_2}$. See Section 2.5.7 for the proof. We remark that the lower bound in Lemma 2.3 is sharp, as can be seen from Lemma 2.13.



(a) The upper and lower bound constructed in the Proof of Theorem 2.1.

(b) The upper and lower bound constructed in the proof of Theorem 2.2.

Figure 2.1: Illustration of proofs.

Lemma 2.4. *Let X_1, X_2, Y_1, Y_2 be random elements taking values in a normed space $(\mathcal{X}, \|\cdot\|)$. Then we have*

$$\left|\langle\!\langle X_1, X_2\rangle\!\rangle_{\mathcal{W}_2} - \langle\!\langle Y_1, Y_2\rangle\!\rangle_{\mathcal{W}_2}\right| \le \left(\mathbb{E}\,\|Y_2\|^2\right)^{1/2}\mathcal{W}_2(P^{X_1}, P^{Y_1}) + \left(\mathbb{E}\,\|X_1\|^2\right)^{1/2}\mathcal{W}_2(P^{X_2}, P^{Y_2}).$$

This lemma links $\mathcal{W}_2(P^{X_1}, P^{X_2})$, $\mathcal{W}_2(P^{Y_1}, P^{Y_2})$ with $\mathcal{W}_2(P^{X_1}, P^{Y_1})$, $\mathcal{W}_2(P^{X_2}, P^{Y_2})$. This is useful when transforming a two-sample problem into two one-sample problems. Please refer to Section 2.5.8 for the proof.

Proof sketch of Theorem 2.1. We start with the basic inequality:

$$\mathcal{L}(\hat{b}) - \mathcal{L}(b^*) \le \mathcal{L}(\hat{b}) - \mathcal{L}_{n,m}(\hat{b}) + \mathcal{L}_{n,m}(b^*) - \mathcal{L}(b^*). \tag{2.8}$$

The proof strategy involves establishing a lower bound for the left-hand side of (2.8) with respect to $\|\hat{b} - b^*\|_\Sigma$ and an upper bound for the right-hand side of (2.8) in terms of $\|\hat{b} - b^*\|_\Sigma$. Then by solving the resulting inequality, we can derive an expression bounding $\|\hat{b} - b^*\|_\Sigma$.

For a lower bound of the left-hand side of (2.8), since for any $b \in \mathbb{R}^{d\times p}$, we have $\mathcal{L}(b) - \mathcal{L}(b^*) = \langle\!\langle (b^* - b)X + \varepsilon, U\rangle\!\rangle_{\mathcal{W}_2} - \langle\!\langle \varepsilon, U\rangle\!\rangle_{\mathcal{W}_2}$, by applying Lemma 2.3 and the explicit form for $\langle\!\langle (b^* - b)X, U\rangle\!\rangle_{\mathcal{W}_2}$ we can show that

$$\mathcal{L}(b) - \mathcal{L}(b^*) \ge \sqrt{r^2 + \|b^* - b\|_\Sigma^2} - r, \tag{2.9}$$

where $r := \langle\!\langle \varepsilon, U\rangle\!\rangle_{\mathcal{W}_2}$ is a constant. This lower bound grows quadratically in $\|\hat{b} - b^*\|_\Sigma$ when $\|\hat{b} - b^*\|_\Sigma$ is close to zero and linearly when $\|\hat{b} - b^*\|_\Sigma$ is large (see Figure 2.1a for an illustration).

To upper bound the right-hand side of (2.8), by applying Lemma 2.4 we have for each $b \in \mathbb{R}^{d\times p}$,

$$|\mathcal{L}(b) - \mathcal{L}_{n,m}(b)| \le \left(\frac{1}{m}\sum_{i=1}^m \|U_i\|^2\right)^{1/2} \mathcal{W}_2(P^{Y-bX}, P_n^{Y-bX}) + (\mathbb{E}\,\|Y - bX\|^2)^{1/2}\mathcal{W}_2(P^U, P_m^U). \tag{2.10}$$

Here $\mathcal{W}_2(P^{Y-bX}, P_n^{Y-bX})$ and $\mathcal{W}_2(P^U, P_m^U)$ are one-sample empirical Wasserstein distance, and the state-of-art convergence rate can be applied [see e.g. FG15] (the actual proof is more involved

in the sense that we need to establish the same result uniformly over b). Then a direct calculation on the right-hand side of (2.10) leads to a quadratic upper bound in terms of $\|b^* - b\|_\Sigma$. The result follows by combining the upper bound with the lower bound (2.9). See Section 2.5.6 for the proof. □

Before we state a faster convergence rate result, we first introduce the following assumptions.

Assumption 2 *For some $\sigma_1, \sigma_2 > 0$ and $\alpha, \beta \in (0, 2]$, it holds that the distribution of $\Sigma^{-1/2}X$ is (σ_1, α)-*sub-Weibull *and P^ε is (σ_2, β)-*sub-Weibull*, in the sense that*

$$\mathbb{E}\exp\left\{\frac{1}{2}(\|\Sigma^{-1/2}X\|/\sigma_1)^\alpha\right\} \le 2 \quad \textit{and} \quad \mathbb{E}\exp\left\{\frac{1}{2}(\|\varepsilon\|/\sigma_2)^\beta\right\} \le 2 \tag{2.11}$$

Assumption 3 *For some $\gamma_1, \gamma_2 > 0$, the density function of ε, write as f_ε, satisfies the following anti-concentration property*

$$f_\varepsilon(e) \ge \gamma_1 \exp\{(-\gamma_2\|e\|^2)\}, \quad \textit{for } \|e\| \ge 1. \tag{2.12}$$

On the one hand, Assumption 3 immediately implies the following anti-concentration bound

$$\mathbb{P}(\|\varepsilon\| \ge r) \ge \frac{\pi^{d/2}\big((r+1)^d - r^d\big)}{\Gamma(\frac{d}{2}+1)}\gamma_1 \exp\big(-2\gamma_2 r^2 - 2\gamma_2\big), \quad \text{for } r \ge 1.$$

This indicates that the random noise ε possesses a heavier tail than the sub-gaussian tail outside the unit ball. On the other hand, by proposition 2.5(i), the sub-Weibull assumption implies that $\mathbb{P}(\|\varepsilon\| \ge r) \le 2e^{-\frac{1}{2}(r/\sigma_2)^\beta}$. The anti-concentration condition in (2.12) is a relaxation of the so-called (γ_1, γ_2)-regularity defined in [PW16]. The merit of employing this relaxation becomes apparent when examining Lemma 2.15, where it is demonstrated that the convolution of two independent probability densities adhering to (2.12) continues to satisfy the anti-concentration inequality. In contrast, the convolution of two independent regular densities may not be regular.

Equipped with these assumptions, we are ready to state an improved convergence rate.

Theorem 2.2. *Under the same setup of Theorem 2.1 and suppose that Assumptions 2 and 3 are satisfied. For m, n large enough such that (2.6) is satisfied, there exists some constant $M > 0$ depending only on $d, \alpha, \beta, \sigma_1, \sigma_2, \gamma_1, \gamma_2$ such that with probability at least $1 - 33(\log n)^{-1}$, we have*

$$\|b^* - \hat{b}\|_\Sigma^2 \le M\big((p/n)^{1/2} + n^{-2/d}\big)(\log m)^{\frac{8}{2\wedge\alpha\wedge\beta}}. \tag{2.13}$$

When $d > 4$, up to a factor of the logarithm, the empirical Wasserstein distance estimation error $n^{-2/d}$ is the dominant term. This is derived from a uniform empirical Wasserstein distance control (see (2.14) and Proposition 2.4), and its minimax optimality has been established in [SP18]. Compared to (2.7), this improved bound in (2.13) removes the dependence on p in the exponent. Moreover, unlike the convergence rate result established for the projected Wasserstein distance in Wang, Gao, and Xie [WGX21; WGX22], our argument does not require the distribution of ε to have compact support. When $d \le 4$, the parametric rate $(p/n)^{1/2}$ dominates the estimation error. However, this does not translate into a the root-n consistency even when $d = 1$. We conjecture

that this is likely due to an artifact of our proof. Specifically, due to a lack of effective tools to analyze the curvation of the loss function that incorporates the Wasserstein distance, we were unable to obtain concentration results for $\frac{\partial}{\partial b}(\mathcal{L}(b) - \mathcal{L}_{n,m}(b))$ uniformly over b in a similar way that we have done for $\mathcal{L}(b) - \mathcal{L}_{n,m}(b)$. Exploration along this direction remains an area for future work. We briefly sketch the proof below. See Section 2.5.9 for a complete proof.

Proof sketch of Theorem 2.2. Assume the setting of Theomem 2.1, error bound (2.7) implies that on a high probability event, $\hat{b}$ will lie in a bounded ball centered at b^*, denoted by $\mathcal{B}$. Thus the basic inequality (2.8) indicates the following uniform bound

$$\begin{aligned}\mathcal{L}(\hat{b}) - \mathcal{L}(b^*) &\le 2 \sup_{b\in\mathcal{B}} |\mathcal{L}(b) - \mathcal{L}_{n,m}(b)| \\ &\le \left|\frac{1}{m}\sum_{i=1}^{m} \|U_i\|^2 - \mathbb{E}\,\|U\|^2\right| + \sup_{b\in\mathcal{B}}\left|\frac{1}{n}\sum_{i=1}^{n} \|Y_i - bX_i\|^2 - \mathbb{E}\,\|Y - bX\|^2\right| \\ &\quad + \sup_{b\in\mathcal{B}}\left|\mathcal{W}_2^2(P^{Y-bX}, P^U) - \mathcal{W}_2^2(P_n^{Y-bX}, P_m^U)\right|. \end{aligned} \tag{2.14}$$

Utilizing the same lower bound for the left-hand side as in (2.9), it remains to derive an upper bound for the right-hand side of the above inequality. While the initial two terms of (2.14) can be effectively controlled through the application of statistical concentration arguments, as elucidated in Lemma 2.11, achieving control over the last term demands much more effort. Motivated by the duality argument presented in Manole and Niles-Weed [MN24, Theorem 13], we establish a non-asymptotic *uniform* error bound for the empirical 2-Wasserstein distance (Proposition 2.4; see also Figure 2.1b for an illustration), which forms the key ingredient of the proof. □

2.4 Simulations

In this section, we compare the empirical performance of MCQR with other robust regression estimators. The MCQR estimator is obtained by solving the linear programming problem in Section 2.2.3. The competitors used in the simulation studies include the ordinary least squares estimator (LS), the spatial quantile regression (SpQR) with zero quantile level [Cha96], and coordinate-wise CQR (CoorCQR), i.e. independently applying CQR to each component of the response variable. We refer readers to Section 2.8 for more details about SpQR.

In each experiment, we draw i.i.d. data $(X_1, Y_1), \ldots, (X_n, Y_n)$ according to model (2.1), where the regression coefficients $b^* \in \mathbb{R}^{d\times p}$ has independent $\mathcal{N}(5, 5)$ entries and is kept fixed for all repetitions. Covariates $X_i \in \mathbb{R}^p$, $i = 1, \ldots, n$, are drawn from $N(0, \Sigma)$ with a Toeplitz covariance matrix $\Sigma = (2^{-|i-j|})_{i,j} \in \mathbb{R}^{p\times p}$. The noise ε is generated from one of the following distributions:

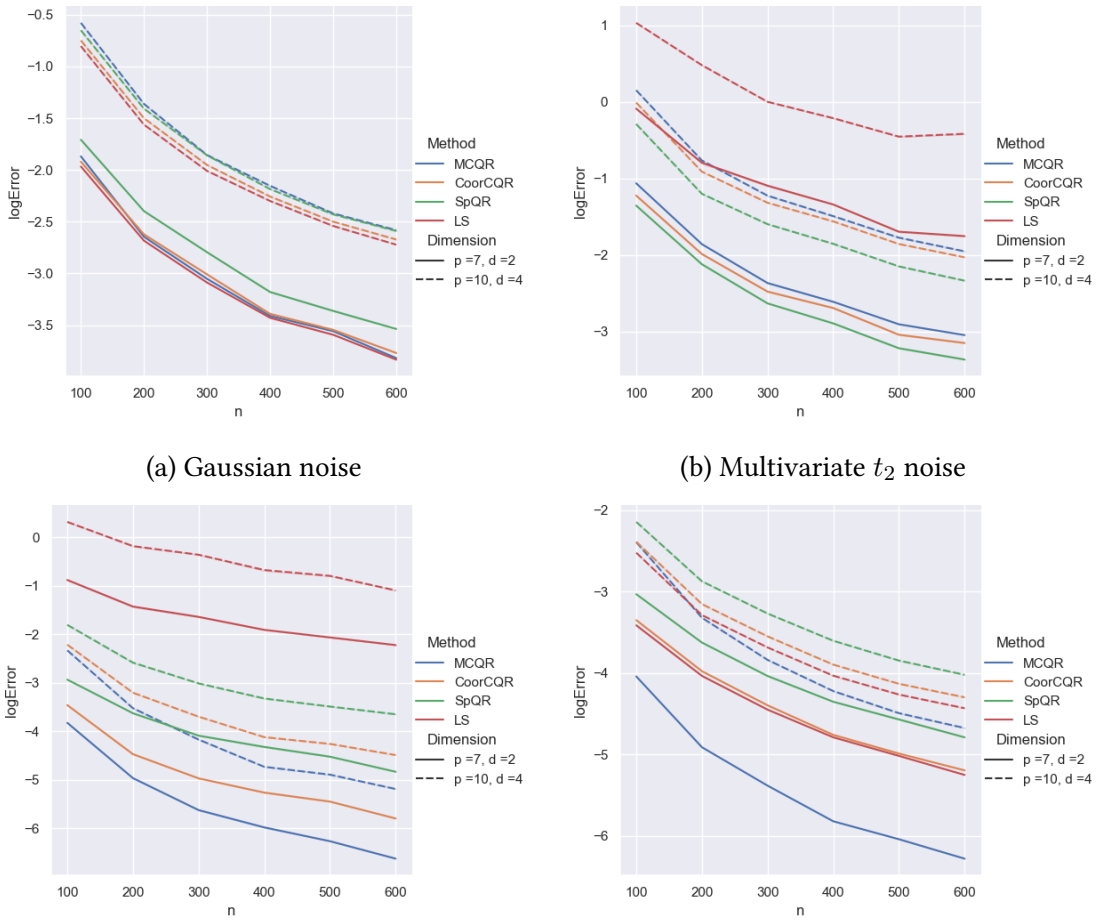
(1a) $\varepsilon \sim \mathcal{N}(0, I_d)$

(1b) $\varepsilon \sim t_2(0, I_d)$ follows a multivariate t_2 distribution

(1c) ε has each marginal distributed with $\mathrm{Pareto}(-2, 2, 1)$ [1] and the same copula as $\mathcal{N}(0, \Sigma')$, where $\Sigma' = (0.9^{|i-j|})_{i,j} \in \mathbb{R}^{d\times d}$

(1d) ε follows a centered Banana-shaped distribution, i.e. $\varepsilon_i \overset{d}{=} (B_{d-1}, \|B_{d-1}\|^2 - \frac{2}{d+2}) + 0.3B_d$, where B_d is uniformly distributed in the unit ball in $\mathbb{R}^d$

[1] the Pareto distribution $\mathrm{Pareto}(k, \alpha, s)$ has density function $f(x) \propto \frac{\alpha s^{\alpha+1}}{(x-k)^{\alpha+1}}$ for all $x \ge 1 + k$, with shape parameter $\alpha > 0$, location parameter $k \in \mathbb{R}$ and scale parameter $s > 0$. Here $\mathrm{Pareto}(-2, 2, 1)$ has mean 0.



(a) Gaussian noise

(b) Multivariate t_2 noise

(c) Pareto copula noise

(d) Banana-shaped noise

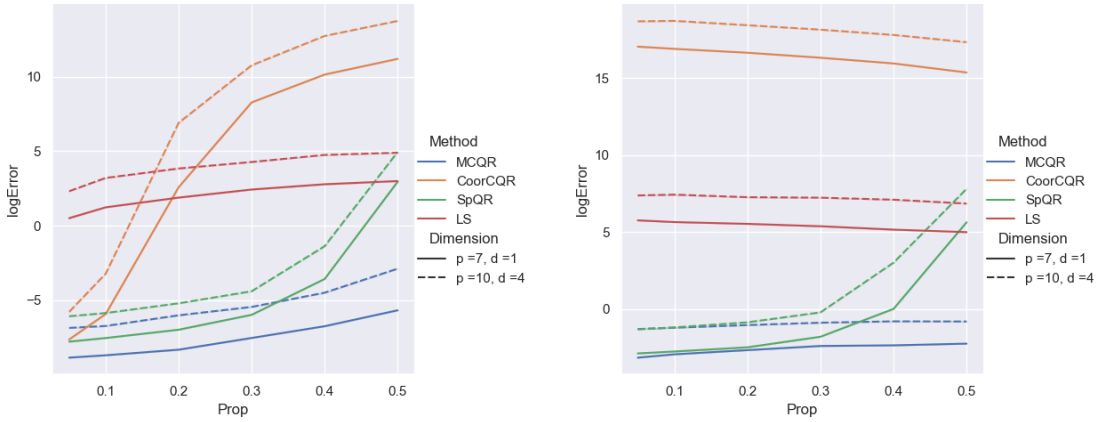
Figure 2.2: Logarithmic average loss, measured in matrix Mahalanobis norm, of the regression coefficient estimated by MCQR, CoorCQR, SpQR and LS for data generated according to the mechanism described in Section 2.4 for various sample size n, covariate dimension p and response dimension d and four different noise distributions (panels (a) to (d)).

Figure 2.2 reports the average matrix Mahalanobis norm error (estimated over 100 Monte Carlo repetitions) of MCQR, LS, SpQR and CoorCQR over the four noise distributions mentioned above for $n \in \{100, 200, \ldots, 600\}$ and $(d, p) \in \{(2, 7), (4, 10)\}$. We see that MCQR has done well over all settings considered here. In contrast, LS estimator performs the best under Gaussian noise but has poor performance under heavy-tailed noise or noise with non-convex support. CoorCQR and SpQR have relatively good performance in panels (a) and (b) when the noise is spherically symmetric but their performance deteriorated when the noise exhibits strong cross-sectional dependence in panels (c) and (d).

While our theoretical results have mostly concerned with heavy-tailed noise, we also investigate the empirical performance of MCQR in the presence of outlier contamination. Here, we consider two cases of ϵ-contaminated noise, for some $\epsilon \in (0, 1)$:

(2a) $\varepsilon \sim (1-\epsilon)P_1 + \epsilon P_2$; here P_1 is a Pareto copula with $\mathrm{Pareto}(-\frac{10}{9}, 10, 1)$ marginals and copula generated by $\mathcal{N}(0, \Sigma')$ as in case (1c) and P_2 is a heavier-tailed location-shifted Pareto copula with marginals distributed as $\mathrm{Pareto}(10, 2, 10)$.

(2b) $\varepsilon \sim (1-\epsilon)\mathcal{N}(0, I_d) + \epsilon\mathcal{N}(100, I_d)$



(a) Pareto contamination (b) Gaussian contamination

Figure 2.3: Logarithmic average estimation loss, measured in matrix Mahalanobis norm, of the regression coefficient estimated by MCQR, CoorCQR, SpQR and LS for data generated according to the mechanism described in Section 2.4 for various outlier contamination proportion (from 0.05 to 0.5), covariate dimension p and response dimension d and two different noise contamination models. We fix $n = 200$.

Figure 2.3 shows the performance of the four procedures for increasing levels of contamination proportion ϵ. We observe that MCQR is generally more robust than other competitors when we add additional outliers to the random error. Interestingly, we see that in the case where $d = 1$, the CoorCQR, which reduces to the univariate CQR, shows a lack of robustness against the outlier contamination, while the 1-dimensional version of MCQR maintains its robustness even with a high proportion of contamination.

2.5 Proofs

We first record here some notations and several classical results on optimal transport theory that will be used throughout our theoretical analysis.

2.5.1 Preliminaries on optimal transport theory

Define the rescaled squared ℓ_2-distance as $L_2(x,y) := \frac{1}{2}\|x-y\|^2$ for any $x, y \in \mathbb{R}^d$. In this notation, for two distributions P and Q on $\mathbb{R}^d$, we have

$$\frac{1}{2}\mathcal{W}_2^2(P,Q) = \inf_{\gamma\in\mathcal{C}(P,Q)} \int L_2(x,y)d\gamma(x,y) =: I_2(P,Q). \tag{2.15}$$

Our proof depends on the following Kantorovich duality [see e.g., Vil21, Theorem 1.3]

$$I_2(P,Q) = \sup_{\varphi,\psi\in\Phi_2} J_{P,Q}(\varphi,\psi), \tag{2.16}$$

where $\Phi_2 := \{(\varphi,\psi) \in L^1(P)\times L^1(Q) : \varphi(x)+\psi(y) \le L_2(x,y)\}$ and

$$J_{P,Q}(\varphi,\psi) := \int \varphi(x)dP(x) + \int \psi(y)dQ(y).$$

By taking advantage of the particular form of L_2, we also have for $\tilde{\Phi} := \{(\varphi,\psi) \in L^1(P)\times L^1(Q) : \varphi(x)+\psi(y) \ge x^Ty\}$ that

$$\int \frac{\|x\|^2}{2}dP(x) + \int \frac{\|y\|^2}{2}dQ(y) - \sup_{\varphi,\psi\in\Phi_2} J_{P,Q}(\varphi,\psi) = \inf_{\varphi,\psi\in\tilde{\Phi}} J_{P,Q}(\varphi,\psi) := \tilde{I}_2(P,Q). \tag{2.17}$$

Thus solve the problem of (2.16) degenerates to solve the problem of $\tilde{I}_2(P,Q)$.

For any $\varphi \in L^1(P)$, define its *Legendre transform* as $\varphi^*(y) := \sup_{x\in\mathbb{R}^d}(x^Ty - \varphi(x))$. Then it can be shown that φ^* is a *convex lower semi-continuous (l.s.c.)* function. This definition immediately implies that for any $(\varphi,\psi) \in \tilde{\Phi}$, $\psi(y) \ge \varphi^*(y)$, $\forall y \in \mathbb{R}^d$. Thus we have $J_{P,Q}(\varphi,\psi) \ge J_{P,Q}(\varphi,\varphi^*)$. Similarily, we have $\varphi(x) \ge \sup_{y\in\mathbb{R}^d}\big(x^Ty - \varphi^*(y)\big) = \varphi^{**}(x)$, $\forall x \in \mathbb{R}^d$, which further implies that $J_{P,Q}(\varphi,\varphi^*) \ge J_{P,Q}(\varphi^{**},\varphi^*)$. In the end, we deduced that

$$\inf_{\varphi,\psi\in\tilde{\Phi}} J_{P,Q}(\varphi,\psi) \ge \inf_{\varphi\in L^1(P)} J_{P,Q}(\varphi^{**},\varphi^*) \ge \inf_{\varphi \text{ is convex l.s.c.}} J_{P,Q}(\varphi^*,\varphi).$$

In fact, it can be shown [see e.g. Vil21, Theorem 2.9] that the equality above holds, i.e. there exists a convex l.s.c. function φ_0 such that the *conjugate pair* (φ_0,φ_0^*) is the optimal solution to $\tilde{I}_2(P,Q)$. Now we are ready to state a fundemental theorem for the optimal transport theory with L_2 loss function.

Theorem 2.3. *[Vil21, Theorem 2.12 and Remark 2.13(iii)] Let P and Q be probability measures on $\mathbb{R}^d$, with finite second moment. We consider the Kantorovich dual problem associated with the rescaled squared ℓ_2-distance L_2. Then $\gamma \in \mathcal{C}(P,Q)$ is optimal if and only if there exists a convex l.s.c. function φ_0 such that*

$$\mathrm{Supp}(\gamma) \subset \partial\varphi_0,$$

or equivalently, for γ-almost all (x, y),

$$y \in \partial\varphi_0(x).$$

Moreover, there exists a conjugate pair (φ_0, φ_0^*) that is a minimizer of $\tilde{I}_2(P, Q)$. Thus $(\|\cdot\|^2/2 - \varphi_0, \|\cdot\|^2/2 - \varphi_0^*)$ solves the Kantorovich dual problem $I_2(P, Q)$.

The 1-Wasserstein distance satisfies the following Kantorovich–Rubinstein duality.

Theorem 2.4 (Kantorovich–Rubinstein theorem)**.** *Suppose $\mathcal{X}$ is a subset of $\mathbb{R}^d$, define the diameter of $\mathcal{X}$ as $\operatorname{diam}(\mathcal{X}) := \sup_{x,y\in\mathcal{X}} \|x - y\|$. Let $\operatorname{Lip}(\mathcal{X})$ denote the space of all Lipschitz function on $\mathcal{X}$ and for any f within this space define*

$$\|f\|_{\operatorname{Lip}(\mathcal{X})} := \max\Big\{ \sup_{\substack{x,y\in\mathcal{X} \\ x\neq y}} \frac{|f(x) - f(y)|}{\|x - y\|},\ \frac{\|f\|_\infty}{\operatorname{diam}(\mathcal{X})} \Big\}.$$

Then

$$\mathcal{W}_1(P, Q) = \sup\left\{ \int f(x)dP(x) - \int f(y)dQ(y) : f \in \mathrm{L}^1(|P - Q|),\ f \in \operatorname{Lip}_1(\mathcal{X}) \right\}, \quad (2.18)$$

where $\operatorname{Lip}_1(\mathcal{X}) := \{f : \|f\|_{\operatorname{Lip}(\mathcal{X})} \leq 1\}$.

In particular, the 1-Wasserstein distance can be seen as a special case of a integral probability metric (defined below) with respect to the Lip_1 function class.

Definition 2.2 (Integral Probability Metrics)**.** Given probability measures P and Q as before, the integral probability metrics (IPMs) with respect to function class $\mathcal{F}$ is defined as

$$\operatorname{IPM}(P, Q; \mathcal{F}) = \sup_{f\in\mathcal{F}} \Big\{ \int f(x)dP(x) - \int f(y)dQ(y) \Big\}. \quad (2.19)$$

2.5.2 Additional notation

Suppose T is a map from a measurable space X, equipped with a measure μ, to an arbitrary space Y, we denote by $T\#\mu$ as the push-forward of μ by T. Specifically, $(T\#\mu)(A) = \mu(T^{-1}(A))$ for any measurable set A.

Suppose $X_1, \ldots, X_n$ are random samples from some probability distribution P. Then given any function class $\mathcal{F}$, define the Rademacher complexity of $\mathcal{F}$ as

$$\mathcal{R}_n(\mathcal{F}, P) := \mathbb{E}\Big(\sup_{f\in\mathcal{F}} \frac{1}{n} \sum_{i=1}^n \xi_i f(X_i) \Big), \quad (2.20)$$

where ξ_i's are independent Rademacher random variables, independent from $X_1, \ldots, X_n$. The p-dimensional closed ball in centered at $x \in \mathbb{R}^p$ with radius $r > 0$ is denoted by $\mathcal{B}_{x,r}^p := \{y \in \mathbb{R}^p : \|y\| \leq r\}$ and we omit r when $r = 1$: $\mathcal{B}_{x,1}^p := \mathcal{B}_x^p$. The matrix operator norm is denoted by $\|\cdot\|_{\mathrm{op}}$, so that $\|A\|_{\mathrm{op}} := \sup_{x:\|x\|=1} \|Ax\|$.

2.5.3 Proof for Lemma 2.1

Proof. For any fixed $\tau \in (0,1)$, by the definition of check function ρ_τ we have

$$q_Y(\tau) \in \arg\min_\theta \mathbb{E}\,\rho_\tau(Y-\theta),$$

where $q_Y(\cdot)$ is the quantile function of Y. Thus under the linear model (2.1) we have for any $x \in \mathbb{R}^p$,

$$(b^*, q_\varepsilon^*(\tau)) \in \arg\min_{b\in\mathbb{R}^{1\times d}, q\in\mathbb{R}} \mathbb{E}[\rho_\tau(Y-bX-q) \mid X=x]. \tag{2.21}$$

For any $b \in \mathbb{R}^{1\times p}$ and $q \in \mathbb{R}$, define $g(x;b,q) := \mathbb{E}[\rho_\tau(Y-bX-q) \mid X=x]$, then (2.21) implies that

$$g(x; b^*, q_\varepsilon^*(\tau)) \le g(x;b,q),$$

thus

$$\int_{\mathbb{R}^p} g(x; b^*, q_\varepsilon^*(\tau))\,dx \le \int_{\mathbb{R}^p} g(x;b,q)\,dx.$$

Then by the Fubini Theorem and the Law of iterated expectation, we have

$$\mathbb{E}[\rho_\tau(Y-b^*X-q_\varepsilon^*(\tau))] \le \mathbb{E}[\rho_\tau(Y-bX-q)]. \tag{2.22}$$

Because the quantile function $q_\varepsilon^* \in \mathcal{M}$, thus (2.22) implies that for any $q(\cdot) \in \mathcal{M}$

$$\int_0^1 \mathbb{E}[\rho_\tau(Y-b^*X-q_\varepsilon^*(\tau))]\,d\tau \le \int_0^1 \mathbb{E}[\rho_\tau(Y-bX-q(\tau))]d\tau.$$

Therefore the result follows by applying the Fubini Theorem once again. □

2.5.4 Proof for Lemma 2.2

Proof. Let $\mathcal{C}$ denote the class of convex functions on $[0,1]$. By the definition of the check function ρ_τ and the fact that X is mean-zero, we have

$$\begin{aligned}
&\inf_{q\in\mathcal{M}} \mathbb{E}\Big\{\int_0^1 \rho_\tau\big(Y-bX-q(\tau)\big)d\tau\Big\} + \frac{1}{2}\mathbb{E}\,Y \\
&\qquad = \inf_{q\in\mathcal{M}}\Big\{\mathbb{E}\int_0^1 (Y-q(\tau)-bX)^+ d\tau + \int_0^1 (1-\tau)q(\tau)d\tau\Big\} \\
&\qquad = \inf_{q\in\mathcal{M}}\Big\{\mathbb{E}\max_{t\in[0,1]}\int_0^t (Y-q(\tau)-bX)\,d\tau + \int_0^1\int_\tau^1 q(\tau)\,du\,d\tau\Big\} \\
&\qquad = \inf_{\phi\in\mathcal{C}}\Big\{\mathbb{E}\max_{t\in[0,1]}(t(Y-bX)-\phi(t)) + \mathbb{E}\phi(U)\Big\} \\
&\qquad = \inf_{\phi\in\mathcal{C}}\mathbb{E}\big\{\phi^*(Y-bX)+\mathbb{E}\phi(U)\big\},
\end{aligned} \tag{2.23}$$

where $\phi^*(t) := \max_{t\in[0,1]}\{ut-\phi(u)\}$ is the Legendre conjugate of $\phi:[0,1]\to\mathbb{R}$ and we used Fubini's theorem and a change of variable $q \mapsto \phi \in \mathcal{C}$ defined by $\phi(t) = \int_0^t q(\tau)\,d\tau$ in the penultimate step.

Let ϕ_0 be the optimizer of (2.23) and ϕ_0^* its Legendre conjugate, then by Villani [Vil21, Theorem 2.9], we have

$$\begin{aligned}\mathbb{E}\phi_0^*(Y-bX)+\mathbb{E}\phi_0(U) &= \inf_{\phi\in\mathcal{C}}\big\{\mathbb{E}\phi^*(Y-bX)+\mathbb{E}\phi(U)\big\}\\ &= \inf_{\phi,\psi\in\mathcal{C}:\phi(x)+\psi(y)\geq xy}\big\{\mathbb{E}\psi(Y-bX)+\mathbb{E}\phi(U)\big\}.\end{aligned}$$

Then by the arguments in Villani [Vil21, Sec 2.1.2], the pair $(\tilde{\phi}_0,\tilde{\psi}_0)$ defined by $\tilde{\phi}_0(u)=u^2/2-\phi_0(u)$ and $\tilde{\psi}_0(y)=y^2/2-\phi_0^*(y)$ is the optimizer of the Kantorovich dual formulation of the optimal transport problem between P^{Y-bX} and P^U, i.e.

$$\mathbb{E}\tilde{\psi}_0(Y-bX)+\mathbb{E}\tilde{\phi}_0(U)=\sup_{\substack{\tilde{\phi},\tilde{\psi}\in L^1(\mathbb{R})\\ \tilde{\phi}(x)+\tilde{\psi}(y)\leq(x-y)^2/2}}\mathbb{E}\tilde{\psi}(Y-bX)+\mathbb{E}\tilde{\phi}(U). \tag{2.24}$$

By the strong duality theorem [Vil21, Theorem 1.3], we have

$$\begin{aligned}\frac{1}{2}\mathcal{W}_2^2\big(P^{Y-bX},P^U\big) &= \mathbb{E}\tilde{\psi}_0(Y-bX)+\mathbb{E}\tilde{\phi}_0(U)\\ &= \mathbb{E}\bigg\{\frac{1}{2}(Y-bX)^2-\phi_0^*(Y-bX)\bigg\}+\mathbb{E}\bigg\{\frac{1}{2}U^2-\phi_0(U)\bigg\},\end{aligned} \tag{2.25}$$

which together with the definition of $\langle\!\langle\cdot,\cdot\rangle\!\rangle_{\mathcal{W}_2}$ implies that

$$\langle\!\langle P^{Y-bX},P^U\rangle\!\rangle_{\mathcal{W}_2}=\mathbb{E}\phi_0^*(Y-bX)+\mathbb{E}\phi_0(U).$$

The result follows by combining the above identity with the optimality of ϕ_0 in (2.23). □

2.5.5 Proof for Proposition 2.1

Proof. By Brenier's Theorem, there is a unique (invertible) optimal transport map $\phi:\mathbb{R}^d\to\mathbb{R}^d$ from P^U to P^ε, which induces a coupling $P^{(U,\varepsilon)}:=(\phi\otimes\mathrm{Id})\#P^U\in\mathcal{C}(P^U,P^\varepsilon)$. Then $P^{(U,\varepsilon)}\otimes P^{(b^*-b)X}$ is a joint distribution of $(U,\varepsilon,(b^*-b)X)$, which induces a joint distribution $P^{(U,Y-bX)}\in\mathcal{C}(P^U,P^{Y-bX})$ through the map $(u,e,z)\mapsto(u,e+z)$. Observe that the squared L_2 transport cost associated with $P^{(U,Y-bX)}$ is

$$\begin{aligned}\int\|u-v\|_2^2\,dP^{(U,Y-bX)}(u,v) &= \int\|u-(e+z)\|_2^2\,d(P^{(U,\varepsilon)}\otimes P^{(b^*-b)X})(u,e,z)\\ &= \int\|\phi(u)-u\|_2^2\,dP^U(u)+\int\|z\|_2^2\,dP^{(b^*-b)X}(z)\\ &= \mathcal{W}_2^2(P^U,P^\varepsilon)+\mathbb{E}\,\|(b^*-b)X\|_2^2.\end{aligned} \tag{2.26}$$

Therefore, we have

$$\begin{aligned}\mathcal{L}(b;U)-\mathcal{L}(b^*;U) &= -\mathcal{W}_2^2(P^U,P^{Y-bX})+\mathcal{W}_2^2(P^U,P^\varepsilon)+\mathbb{E}\,\|(b^*-b)X\|_2^2\\ &= \int\|u-v\|_2^2\,dP^{(U,Y-bX)}(u,v)-\inf_{Q\in\mathcal{C}(P^U,P^{Y-bX})}\int\|u-v\|_2^2\,dQ(u,v)\geq 0.\end{aligned} \tag{2.27}$$

This implies that $b^* \in \arg\min \mathcal{L}(b; U)$. To prove the uniqueness, by Brenier's Theorem, since $P^U \in \mathcal{P}_{\mathrm{ac}}(\mathbb{R}^d)$, the optimal transport map from P^U to P^{Y-bX} is unique, thus the equality can only be achieved in (2.27) if $P^{(U,Y-bX)}$ is the optimal coupling. In such a case, by the Knott-Smith optimality criterion [Vil21, Theorem 2.12(i)], there exists a unique convex lower semi-continuous function $h : \mathbb{R}^d \to \mathbb{R}^d$ such that $\mathrm{Supp}(P^{(U,Y-bX)}) \subset \mathrm{Graph}(\nabla h)$ in the sense that, for any $(u, v) \in \mathrm{Supp}(P^{(U,Y-bX)})$, we have $v = \nabla h(u)$. Define an event $A = \{\nabla h(\phi^{-1}(\varepsilon))) = \varepsilon + (b^* - b)X\}$. Then

$$\begin{aligned}\mathbb{P}(A) &= \mathbb{P}\big((\phi^{-1}(\varepsilon), \varepsilon + (b^* - b)X) \in \{(u, v) : \nabla h(u) = v\}\big) \\ &= P^{(U,Y-bX)}\{(u, v) : \nabla h(u) = v\} = 1.\end{aligned}$$

This implies that $\varepsilon + (b^* - b)X = \nabla h(\phi^{-1}(\varepsilon))$ almost surely. Because X is independent of ε, and is not a point mass, the only way to make this equality hold is when $b = b^*$ as desired. □

2.5.6 Proof for Theorem 2.1

For notation simplicity, write $S := \Sigma^{-1/2}X$ and $S_i := \Sigma^{-1/2}X_i$ for $i \in [n]$ throughout the rest of the paper.

Proof. By the definition of $\hat{b}$ in (2.5), we have the following basic inequality:

$$\mathcal{L}(\hat{b}) - \mathcal{L}(b^*) \le \mathcal{L}(\hat{b}) - \mathcal{L}_{n,m}(\hat{b}) + \mathcal{L}_{n,m}(b^*) - \mathcal{L}(b^*). \tag{2.28}$$

By the explicit formula for the 2-Wasserstein distance between two elliptical distributions [see Gel90, Theorem 2.1], we have

$$\begin{aligned}\langle\!\langle P^{(b^*-b)X}, P^U \rangle\!\rangle_{\mathcal{W}_2} &= \frac{1}{2}\Big\{\mathbb{E}\,\|(b^* - b)X\|^2 + \mathbb{E}\,\|U\|^2 - \mathcal{W}_2^2(P^{(b^*-b)X}, P^U)\Big\} \\ &= \frac{1}{2}\Big\{\mathbb{E}\,\|(b^* - b)X\|^2 + \mathbb{E}\,\|U\|^2 - \big\|\big((b^* - b)\Sigma(b^* - b)^T\big)^{1/2} - I_d\big\|_{\mathrm{F}}^2\Big\} \\ &= \mathrm{Tr}\big\{\big((b^* - b)\Sigma(b^* - b)^T\big)^{1/2}\big\} \qquad (2.29) \\ &\ge \mathrm{Tr}^{1/2}\big\{(b^* - b)\Sigma(b^* - b)^T\big\} = \|b^* - b\|_\Sigma. \qquad (2.30)\end{aligned}$$

Hence, writing $r := \langle\!\langle P^\varepsilon, P^U \rangle\!\rangle_{\mathcal{W}_2}$, we have by Lemma 2.3 that for any $b \in \mathbb{R}^{d\times p}$,

$$\begin{aligned}\mathcal{L}(b) - \mathcal{L}(b^*) &= \langle\!\langle P^{(b^*-b)X+\varepsilon}, P^U \rangle\!\rangle_{\mathcal{W}_2} - \langle\!\langle P^\varepsilon, P^U \rangle\!\rangle_{\mathcal{W}_2} \\ &\ge \sqrt{r^2 + \langle\!\langle P^{(b^*-b)X}, P^U \rangle\!\rangle_{\mathcal{W}_2}^2} - r \ge \sqrt{r^2 + \|b^* - b\|_\Sigma^2} - r. \qquad (2.31)\end{aligned}$$

On the other hand, by Lemma 2.4, we have

$$\begin{aligned}|\mathcal{L}(b) - \mathcal{L}_{n,m}(b)| &= \Big|\langle\!\langle P^{Y-bX}, P^U \rangle\!\rangle_{\mathcal{W}_2} - \langle\!\langle P_n^{Y-bX}, P_m^U \rangle\!\rangle_{\mathcal{W}_2}\Big| \\ &\le \alpha_m \mathcal{W}_2(P^{Y-bX}, P_n^{Y-bX}) + (\mathbb{E}\,\|Y - bX\|^2)^{1/2}\mathcal{W}_2(P^U, P_m^U), \qquad (2.32)\end{aligned}$$

where $\alpha_m := \big(\frac{1}{m}\sum_{i=1}^m \|U_i\|^2\big)^{1/2}$. We control the two terms on the right-hand side of (2.32) separately. For the first term, suppose P_1 is the optimal coupling between P^S and P_n^S, and P_2 is

the optimal coupling between P^{ε} and P_n^{ε}. Since $P_1 \otimes P_2$ induces a coupling between P^{Y-bX} and P_n^{Y-bX} through the relation $Y - bX = (b^* - b)\Sigma^{1/2}S + \varepsilon$, we have

$$\begin{aligned}
&\mathcal{W}_2^2(P^{Y-bX}, P_n^{Y-bX}) \\
&\qquad \le \int \|(b^* - b)\Sigma^{1/2}s_1 + e_1 - (b^* - b)\Sigma^{1/2}s_2 - e_2\|^2 d(P_1 \otimes P_2)(s_1, s_2, e_1, e_2) \\
&\qquad \le \int \|b^* - b\|_\Sigma^2 \|s_1 - s_2\|^2 dP_1(s_1, s_2) + \int \|e_1 - e_2\|^2 dP_2(e_1, e_2) \\
&\qquad = \|b^* - b\|_\Sigma^2 \mathcal{W}_2^2(P^S, P_n^S) + \mathcal{W}_2^2(P^{\varepsilon}, P_n^{\varepsilon}).
\end{aligned}$$

Thus,

$$\mathcal{W}_2(P^{Y-bX}, P_n^{Y-bX}) \le \|b^* - b\|_\Sigma \mathcal{W}_2(P^S, P_n^S) + \mathcal{W}_2(P^{\varepsilon}, P_n^{\varepsilon}) =: I_n(\|b^* - b\|_\Sigma). \tag{2.33}$$

For the second term on the right-hand side of (2.32), define $s^2 := \mathbb{E}\,\|\varepsilon\|^2$, we have

$$\begin{aligned}
(\mathbb{E}\,\|Y - bX\|^2)^{1/2} &= (\mathbb{E}\,\|(b^* - b)X + \varepsilon\|^2)^{1/2} \le \{2\mathbb{E}\|(b^* - b)X\|^2 + 2\mathbb{E}\|\varepsilon\|^2\}^{1/2} \\
&= \left\{2\|b^* - b\|_\Sigma^2 + 2s^2\right\}^{1/2}.
\end{aligned} \tag{2.34}$$

Combining (2.32), (2.33) and (2.34), we obtain that

$$|\mathcal{L}(b) - \mathcal{L}_{n,m}(b)| \le \alpha_m I_n(\|b^* - b\|_\Sigma) + \left\{2\|b^* - b\|_\Sigma^2 + 2s^2\right\}^{1/2} \mathcal{W}_2(P^U, P_m^U). \tag{2.35}$$

Since (2.31) and (2.35) holds for arbitrary $b \in \mathbb{R}^{d\times p}$, we have by (2.28) that

$$\begin{aligned}
\left\{r^2 + \|b^* - \hat{b}\|_\Sigma^2\right\}^{1/2} - r \le \alpha_m I_n(\|b^* - \hat{b}\|_\Sigma) + \left\{2\|b^* - \hat{b}\|_\Sigma^2 + 2s^2\right\}^{1/2} \mathcal{W}_2(P^U, P_m^U) \\
+ \alpha_m I_n(0) + s\sqrt{2}\mathcal{W}_2(P^U, P_m^U).
\end{aligned}$$

We apply Lemma 2.12 to the left-hand side of the above and combine with the fact that $r^2 \le s^2 d$, we deduce that for some constant $C > 0$ only depending on d, the following inequality holds:

$$\begin{aligned}
&\frac{(2\|b^* - \hat{b}\|_\Sigma - 1) \wedge \|b^* - \hat{b}\|_\Sigma^2}{(\|b^* - \hat{b}\|_\Sigma \vee 1)} \\
&\qquad \le C(2 + 2s)\big(\alpha_m \mathcal{W}_2(P^S, P_n^S) + (\sqrt{2} + 2s\sqrt{2})\mathcal{W}_2(P^U, P_m^U) + 2\alpha_m \mathcal{W}_2(P^{\varepsilon}, P_n^{\varepsilon})\big).
\end{aligned} \tag{2.36}$$

Thus we only need to control the right-hand side of the above.

Note by Markov's inequality, $E_0^{(m)} := \{\alpha_m \le \sqrt{d \log m}\}$ holds with probability at least $1 - (\log m)^{-1}$. Similarily, by the convergence rate of empirical 2-Wasserstein distance in Theorem 2.5 implies that there exists constants $C_1 > 0$ depending only on p and ℓ and $C_2, C_3 > 0$ depending only on d, ℓ such that for all $m, n > 1$, events $E_1^{(n)} := \{\mathcal{W}_2(P^S, P_n^S) \le C_1 \tau_n^{1/2}(p, \ell) \log^{1/2} n\}$, $E_2^{(n)} := \{\mathcal{W}_2(P^{\varepsilon}, P_n^{\varepsilon}) \le C_2 \tau_n^{1/2}(d, \ell) \log^{1/2} n\}$ and $E_3^{(m)} := \{\mathcal{W}_2(P^U, P_m^U) \le C_3 \tau_m^{1/2}(d, \ell) \log^{1/2} m\}$ hold with probability at least $1 - (\log n)^{-1}$, $1 - (\log n)^{-1}$, $1 - (\log m)^{-1}$, respectively. Therefore, for all $n > 1$ and $m > n$, let $E^{(n,m)} := E_0^{(m)} \bigcap E_1^{(n)} \bigcap E_2^{(n)} \bigcap E_3^{(m)}$, we have $\mathbb{P}(E^{(n,m)}) \ge 1 - 4(\log n)^{-1}$.

Note

$$\frac{(2\|b^* - \hat{b}\|_\Sigma - 1) \wedge \|b^* - \hat{b}\|_\Sigma^2}{(\|b^* - \hat{b}\|_\Sigma \vee 1)} \geq \|b^* - \hat{b}\|_\Sigma^2 \wedge 1.$$

Then combining this with (2.36), and working on the event $E^{(n,m)}$, there exists some constant $\tilde{M} > 0$ depending only on d, ℓ, p such that

$$\begin{aligned}\|b^* - \hat{b}\|_\Sigma^2 \wedge 1 &\leq \tilde{M}(1+s)\big(\tau_n^{1/2}(p,\ell) + s\tau_n^{1/2}(d,\ell)\big)\log^{1/2} m \\ &\leq \tilde{M}\big(n^{-1/4} + n^{-\frac{1}{d\vee p}} + n^{\frac{2-\ell}{2\ell}}\big)\log m,\end{aligned}$$

where a positive constant depending on d is absorbed in $\tilde{M}$ in the final inequality, while we stick with notation $\tilde{M}$ for simplicity. □

2.5.7 Proof for Lemma 2.3

Proof. By the Brenier's Theorem [Vil09, Theorem 2.12 (ii)], there exists optimal transport maps $\phi, \psi : \mathbb{R}^d \to \mathbb{R}^d$ such that $\phi \# P^\varepsilon = P^U$ and $\psi \# P^Z = P^U$. Now, for any fixed $t \in [0,1]$, we define $M_t(z,e) := \sqrt{1-t}\psi(z) + \sqrt{t}\phi(e)$, for all $z, e \in \mathbb{R}^d$. Since $M_t(Z,\varepsilon) \overset{\mathrm{d}}{=} U$, there exists a coupling $P^{(Z,\varepsilon,U)} \in \mathcal{C}(P^Z \otimes P^\varepsilon, P^U)$ whose associated transport map is M_t (more specifically, $P^{(Z,\varepsilon,U)} = (\mathrm{Id} \otimes M_t)\#(P^Z \otimes P^\varepsilon)$). Thus, we have

$$\begin{aligned}\langle\!\langle Z+\varepsilon, U\rangle\!\rangle_{\mathcal{W}_2} &\geq \int \langle z+e, u\rangle dP^{(Z,\varepsilon,U)}(z,e,u) \\ &= \int \big\langle z+e, \sqrt{1-t}\psi(z) + \sqrt{t}\phi(e)\big\rangle d(P^Z \otimes P^\varepsilon)(z,e) \\ &= \sqrt{1-t}\int \langle z, \psi(z)\rangle dP^Z(z) + \sqrt{t}\int \langle e, \phi(e)\rangle dP^\varepsilon(e) \\ &= \sqrt{1-t}\langle\!\langle Z, U\rangle\!\rangle_{\mathcal{W}_2} + \sqrt{t}\langle\!\langle \varepsilon, U\rangle\!\rangle_{\mathcal{W}_2},\end{aligned}$$

where in the penultimate step we used the fact that ε is independent from Z. Now, taking $t = \frac{\langle\!\langle \varepsilon, U\rangle\!\rangle_{\mathcal{W}_2}^2}{\langle\!\langle \varepsilon, U\rangle\!\rangle_{\mathcal{W}_2}^2 + \langle\!\langle Z, U\rangle\!\rangle_{\mathcal{W}_2}^2}$, we have

$$\langle\!\langle Z+\varepsilon, U\rangle\!\rangle_{\mathcal{W}_2}^2 \geq \langle\!\langle Z, U\rangle\!\rangle_{\mathcal{W}_2}^2 + \langle\!\langle \varepsilon, U\rangle\!\rangle_{\mathcal{W}_2}^2$$

as desired. □

2.5.8 Proof for Lemma 2.4

Proof. Let $\mathcal{X}_1, \mathcal{X}_2, \mathcal{Y}_1, \mathcal{Y}_2$ denote four copies of $\mathcal{X}$. By Lemma 2.5, there exists a distribution η on $\mathcal{X}_1 \times \mathcal{X}_2 \times \mathcal{Y}_1 \times \mathcal{Y}_2$ with marginals P^{X_1}, P^{X_2}, P^{Y_1}, P^{Y_2}, such that $\eta|_{\mathcal{X}_1\times\mathcal{X}_2}$, $\eta|_{\mathcal{X}_2\times\mathcal{Y}_2}$, $\eta|_{\mathcal{X}_1\times\mathcal{Y}_1}$ are

optimal couplings between X_1 and X_2, X_2 and Y_2, and X_1 and Y_1 respectively. Then we have

$$\begin{aligned}
&\langle\!\langle X_1, X_2 \rangle\!\rangle_{\mathcal{W}_2} - \langle\!\langle Y_1, Y_2 \rangle\!\rangle_{\mathcal{W}_2} \\
&\qquad = \sup_{\mu \in \mathcal{C}(P^{X_1}, P^{X_2})} \int \langle x_1, x_2 \rangle \, d\mu(x_1, x_2) - \sup_{\nu \in \mathcal{C}(P^{Y_1}, P^{Y_2})} \int \langle y_1, y_2 \rangle \, d\nu(y_1, y_2) \\
&\qquad \leq \int \langle x_1, x_2 \rangle \, d\eta|_{\mathcal{X}_1 \times \mathcal{X}_2}(x_1, x_2) - \int \langle y_1, y_2 \rangle \, d\eta|_{\mathcal{Y}_1 \times \mathcal{Y}_2}(y_1, y_2) \\
&\qquad \leq \int \langle x_1, x_2 - y_2 \rangle - \langle y_1 - x_1, y_2 \rangle \, d\eta(x_1, x_2, y_1, y_2) \\
&\qquad \leq \Big(\int \|x_2 - y_2\|^2 d\eta|_{\mathcal{X}_2 \times \mathcal{Y}_2}(x_2, y_2) \Big)^{1/2} \Big(\int \|x_1\|^2 d\eta|_{\mathcal{X}_1}(x_1) \Big)^{1/2} \\
&\qquad\qquad + \Big(\int \|x_1 - y_1\|^2 d\eta|_{\mathcal{X}_1 \times \mathcal{Y}_1}(x_1, y_1) \Big)^{1/2} \Big(\int \|y_2\|^2 d\eta|_{\mathcal{Y}_2}(y_2) \Big)^{1/2} \\
&\qquad = \mathcal{W}_2(P^{X_2}, P^{Y_2}) \cdot \big(\mathbb{E} \, \|X_1\|^2 \big)^{1/2} + \mathcal{W}_2(P^{X_1}, P^{Y_1}) \cdot \big(\mathbb{E} \, \|Y_2\|^2 \big)^{1/2},
\end{aligned}$$

where we used the Cauchy–Schwarz inequality in the final inequality. Similarly, we can find $\tilde{\eta}$ such that $\tilde{\eta}|_{\mathcal{Y}_1 \times \mathcal{Y}_2}$, $\tilde{\eta}|_{\mathcal{X}_2 \times \mathcal{Y}_2}$, $\tilde{\eta}|_{\mathcal{X}_1 \times \mathcal{Y}_1}$ are the corresponding optimal couplings between Y_1 and Y_2, X_2 and Y_2, and X_1 and Y_1 respectively. Then,

$$\begin{aligned}
\langle\!\langle Y_1, Y_2 \rangle\!\rangle_{\mathcal{W}_2} - \langle\!\langle X_1, X_2 \rangle\!\rangle_{\mathcal{W}_2} &\leq \int \langle y_1, y_2 \rangle \, d\tilde{\eta}|_{\mathcal{Y}_1 \times \mathcal{Y}_2}(y_1, y_2) - \int \langle x_1, x_2 \rangle \, d\tilde{\eta}|_{\mathcal{X}_1 \times \mathcal{X}_2}(x_1, x_2) \\
&\leq \int \langle y_1 - x_1, y_2 \rangle - \langle x_1, x_2 - y_2 \rangle \, d\tilde{\eta}(x_1, x_2, y_1, y_2) \\
&\leq \mathcal{W}_2(P^{X_1}, P^{Y_1}) \cdot \big(\mathbb{E} \, \|Y_2\|^2 \big)^{1/2} + \mathcal{W}_2(P^{X_2}, P^{Y_2}) \cdot \big(\mathbb{E} \, \|X_1\|^2 \big)^{1/2}.
\end{aligned}$$

Combining the above two bounds, we get the desried results. □

Lemma 2.5. *For $L \in \mathbb{N}$, write $V = \{1, \ldots, L\}$. Let $(\mathcal{X}_i, \Omega_i, \nu_i)$, $i \in V$ be L probability spaces. Suppose that for some $E \subseteq V \times V$, and for each $(i, j) \in E$, we have a pre-specified joint probability measure $\xi_{i,j}$ on $(\mathcal{X}_i \times \mathcal{X}_j, \Omega_i \otimes \Omega_j)$ such that $\xi_{i,j}|_{\mathcal{X}_i} = \nu_i$ and $\xi_{i,j}|_{\mathcal{X}_j} = \nu_j$. If the simple undirected graph $G = (V, E)$ is acyclic, then there exists a joint probability measure ρ on $\big(\prod_{i=1}^L \mathcal{X}_i, \bigotimes_{i=1}^L \Omega_i\big)$ such that $\rho|_{\mathcal{X}_i} = \nu_i$ for all $i \in V$ and $\rho|_{\mathcal{X}_i \times \mathcal{X}_j} = \xi_{i,j}$ for all $(i, j) \in E$.*

Proof. We assume first that G is connected. Then, there exists a traversal of all the vertices in G such that apart from the first vertex in the traversal, each vertex has exactly one edge connected to an earlier vertex. This can be done by using e.g. depth-first search or breadth first search, after arbitrarily assigning a root node, and each node is connected only to its parent node when first visited. Hence, without loss of generality, we may relabel the nodes so that this traversal is given by the ordering $1, 2, \ldots, L$. We now prove by induction that for any $\ell \in \{1, \ldots, L\}$, there exists a measure $\rho_{1,\ldots,\ell}$ on $\mathcal{X}_1 \times \cdots \times \mathcal{X}_\ell$ such that $\rho_{1,\ldots,\ell}|_{\mathcal{X}_i} = \nu_i$ for all $i \in \{1, \ldots, \ell\}$ and $\rho_{1,\ldots,\ell}|_{\mathcal{X}_i \times \mathcal{X}_j} = \xi_{i,j}$ for all $(i, j) \in E \cap \{1, \ldots, \ell\}^2$.

The base case of the induction is trivially true as we can take $\rho_1 = \nu_1$. Now assume that we have successfully constructed $\rho_{1,\ldots,\ell-1}$ for some $\ell \in \{2, \ldots, L\}$. Let ℓ' be the only neighbour of ℓ in $\{1, \ldots, \ell - 1\}$ (the existence and uniqueness of ℓ' is guaranteed by the traversal ordering of the vertices in the previous paragraph). By the Disintegration Theorem [see e.g. GM89], there

exists a probability measure $\xi_{\ell|\ell'}(\cdot \mid x_{\ell'})$ on $\mathcal{X}_\ell$ such that $d\xi_{\ell|\ell'}(x_\ell \mid x_{\ell'})d\nu_{\ell'}(x_{\ell'}) = d\xi_{\ell',\ell}(x_{\ell'}, x_\ell)$. Now, we define

$$d\rho_{1,\dots,\ell}(x_1,\dots,x_\ell) = d\rho_{1,\dots,\ell-1}(x_1,\dots,x_{\ell-1})d\xi_{\ell|\ell'}(x_\ell \mid x_{\ell'}).$$

To see that $\rho_{1,\dots,\ell}$ satisfies the required conditions, we check that for any $B \in \Omega_i$, $\rho_{1,\dots,\ell}|_{\mathcal{X}_i}(B) = \rho_{1,\dots,\ell-1}|_{\mathcal{X}_i}(B) = \nu_i(B)$ if $i \le \ell - 1$ and

$$\begin{aligned}\rho_{1,\dots,\ell}|_{\mathcal{X}_\ell}(B) &= \rho_{1,\dots,\ell}(\mathcal{X}_1 \times \cdots \times \mathcal{X}_{\ell-1} \times B) = \int_{\mathcal{X}_{\ell'}} \int_B d\xi_{\ell|\ell'}(x_\ell \mid x_{\ell'}) d\rho_{1,\dots,\ell-1}|_{\mathcal{X}_{\ell'}}(x_{\ell'}) \\ &= \int_{\mathcal{X}_{\ell'}} \int_B d\xi_{\ell|\ell'}(x_\ell \mid x_{\ell'}) d\nu_{\ell'}(x_{\ell'}) = \xi_{\ell',\ell}(\mathcal{X}_{\ell'} \times B) = \nu_\ell(B),\end{aligned}$$

if $i = \ell$. Moreover, if $(i,j) \in E \cap \{1,\dots,\ell\}^2$, then for $A \in \Omega_i$ and $B \in \Omega_j$, we either have $(i,j) \in E \cap \{1,\dots,\ell-1\}^2$, in which case $\rho_{1,\dots,\ell}|_{\mathcal{X}_i \times \mathcal{X}_j}(A \times B) = \rho_{1,\dots,\ell-1}|_{\mathcal{X}_i \times \mathcal{X}_j}(A \times B) = \xi_{i,j}(A \times B)$, or $(i,j) = (\ell', \ell)$ (or (ℓ, ℓ') which can be handled symmetrically), in which case,

$$\begin{aligned}\rho_{1,\dots,\ell}|_{\mathcal{X}_{\ell'} \times \mathcal{X}_\ell}(A \times B) &= \int_A \int_B d\xi_{\ell|\ell'}(x_\ell \mid x_{\ell'}) d\rho_{1,\dots,\ell-1}|_{\mathcal{X}_{\ell'}}(x_{\ell'}) \\ &= \int_A \int_B d\xi_{\ell|\ell'}(x_\ell \mid x_{\ell'}) d\nu_{\ell'}(x_{\ell'}) = \xi_{\ell',\ell}(A \times B).\end{aligned}$$

This completes the induction. In particular, $\rho_{1,\dots,L}$ satisfies the desired properties of ρ in the lemma. □

2.5.9 Proof for Theorem 2.2

Define event $\Theta := \{\|\hat{b} - b^*\|_\Sigma < 1\}$, then in the regime of (2.6) we have $\mathbb{P}(\Theta) \ge 1 - 4(\log n)^{-1}$. We henceforth work on the event Θ throughout the proof. Write the linear transformation $A(b) = (b^* - b)X + \varepsilon$ for any $b \in \mathbb{R}^{d \times p}$.

Our proof strategy for Theorem 2.2 is to use the fact that b^* maximizes $\mathcal{L}$ and $\hat{b}$ maximizes $\mathcal{L}_n$ to bound $\mathcal{L}(\hat{b}) - \mathcal{L}(b^*)$ by $|\mathcal{L}(b^*) - \mathcal{L}_n(b^*)| + |\mathcal{L}(\hat{b}) - \mathcal{L}_n(\hat{b})|$. Write $\mathcal{B} := \{b \in \mathbb{R}^{d \times p} : \|b - b^*\|_\Sigma < 1\}$. Then on the event Θ, the key to control the latter is to establish a bound on

$$\sup_{b \in \mathcal{B}} \left| \mathcal{W}_2^2(P^{A(b)}, P^U) - \mathcal{W}_2^2(P_n^{A(b)}, P_m^U) \right|$$

in Proposition 2.4. The proof of Proposition 2.4 relies on rewriting the Wasserstein distances using the Kantorovich dual formulation. Specifically, writing $\tilde{\Phi}_b := \{(f,g) \in L^1(P_n^{A(b)}) \times L^1(P_m^U) : v^T u \le f(v) + g(u),\ \forall (v,u) \in \mathrm{Supp}(P_n^{A(b)}) \times \mathrm{Supp}(P_m^U)\}$, then for any fixed $b \in \mathcal{B}$, by Theorem 2.3 and Lemma 2.16, there exists a conjugate pair $(\tilde{\varphi}_{b;n,m}, \tilde{\varphi}^*_{b;n,m})$ such that

$$\begin{aligned}(\tilde{\varphi}^*_{b;n,m}, \tilde{\varphi}_{b;n,m}) &= \operatorname*{arg\,min}_{(f,g) \in \tilde{\Phi}_b} \int f \, dP_n^{A(b)} + \int g \, dP_m^U, \qquad (2.37) \\ \frac{1}{2}\mathcal{W}_2^2(P_n^{A(b)}, P_m^U) &= \int \|v\|^2/2 - \tilde{\varphi}^*_{b;n,m}(v) \, dP_n^{A(b)}(v) + \int \|u\|^2/2 - \tilde{\varphi}_{b;n,m}(u) \, dP_m^U(u),\end{aligned}$$

and

$$\|u\|^2/2 \le \tilde{\varphi}_{b;n,m}(u) \le \|u\|^2/2 + L_{b;n,m}, \quad \|v\|^2/2 - L_{b;n,m} \le \tilde{\varphi}^*_{b;n,m}(v) \le \|v\|^2/2, \tag{2.38}$$

where $L_{b;n,m} := \max\{L_2(A(b)_i, U_j) : 1 \le i \le n,\ 1 \le j \le m\}$.

Before stating Proposition 2.4, we first establish two results on extensions of $\tilde{\varphi}_{b;n,m}$ and $\tilde{\varphi}^*_{b;n,m}$ onto the entire $\mathbb{R}^d$, which will form the core of the argument in the proof of Proposition 2.4.

Proposition 2.2. *Let $\tilde{\varphi}$ and $\tilde{\varphi}^*$ be defined as in (2.37) and set $L_{b;n,m} := \max_{i\in[n],j\in[m]} L_2(A(b)_i, U_j)$. Let $\zeta_{b;n,m}$, $\varphi_{b;n,m}$ and $\varphi^*_{b;n,m}$ be defined such that for all $v \in \mathbb{R}^d$,*

$$\begin{aligned}
\zeta_{b;n,m}(v) &:= \sup_{u\in \mathrm{Supp}(P_n^{A(b)})} \big\{v^T u - \tilde{\varphi}_{b;n,m}(u)\big\} \vee \Big(\frac{\|v\|^2}{2} - L_{b;n,m}\Big),\\
\varphi_{b;n,m}(v) &:= \sup_{u\in\mathbb{R}^d} \big\{v^T u - \zeta_{b;n,m}(u)\big\},\\
\varphi^*_{b;n,m}(v) &:= \sup_{u\in\mathbb{R}^d} \big\{v^T u - \varphi_{b;n,m}(u)\big\}.
\end{aligned}$$

Then we have

(i) for any $(u, v) \in \mathbb{R}^d \times \mathbb{R}^d$, $v^T u \le \varphi_{b;n,m}(u) + \varphi^*_{b;n,m}(v)$;

(ii) $\varphi_{b;n,m}(u) = \tilde{\varphi}_{b;n,m}(u)$ for $u \in \mathrm{Supp}(P_n^{A(b)})$ and $\varphi^*_{b;n,m}(v) = \tilde{\varphi}^*_{b;n,m}(v)$ for $v \in \mathrm{Supp}(P_m^U)$;

(iii) for $u, v \in \mathbb{R}^d$, $-L_{b;n,m} \le \frac{\|u\|^2}{2} - \varphi_{b;n,m}(u) \le 0$ and $0 \le \frac{\|v\|^2}{2} - \varphi^*_{b;n,m}(v) \le L_{b;n,m}$;

(iv) Let $\pi_{b;n,m} \in \mathcal{C}(P_n^{A(b)}, P_m^U)$ be the optimal coupling between $P_n^{A(b)}$ and P_m^U. Then for any $(u, v) \in \mathrm{Supp}(\pi_{b;n,m})$, we have $v \in \partial\varphi_{b;n,m}(u)$ and $u \in \partial\varphi^*_{b;n,m}(v)$.

Proof. Note (i) is immediately followed by the definition of $\varphi_{b;n,m}$ and $\varphi^*_{b;n,m}$. For part (ii), note for any $u \in \mathrm{Supp}(P_m^U)$

$$\varphi_{b;n,m}(u) \le \sup_{v\in\mathbb{R}^d} \big\{v^T u - v^T u + \tilde{\varphi}_{b;n,m}(u)\big\} = \tilde{\varphi}_{b;n,m}(u). \tag{2.39}$$

For any $v \in \mathrm{Supp}(P_n^{A(b)})$,

$$\begin{aligned}
\varphi^*_{b;n,m}(v) &\le \sup_{u\in\mathbb{R}^d} \big\{v^T u - v^T u + \zeta_{b;n,m}(v)\big\}\\
&= \zeta_{b;n,m}(v) \le \tilde{\varphi}^*_{b;n,m}(v) \vee \Big(\frac{\|v\|^2}{2} - \|c\|_\infty\Big) \le \tilde{\varphi}^*_{b;n,m}(v).
\end{aligned} \tag{2.40}$$

Assume any of (2.39) or (2.40) holds strictly, then because $P_n^{A(b)}$ and P_m^U are finitely support it follows that

$$\int \varphi_{b;n,m}(u) dP_m^U(u) + \int \varphi^*_{b;n,m}(v) dP_n^{A(b)}(v) < \int \tilde{\varphi}_{b;n,m}(u) dP_m^U(u) + \int \tilde{\varphi}^*_{b;n,m}(v) dP_n^{A(b)}(v),$$

which contradicts to the optimality of $(\tilde{\varphi}_{b;n,m}, \tilde{\varphi}^*_{b;n,m})$. This completes the proof for (ii).

For part (iii), by the bounded property (2.38) and preceding constructions we have for $u \in \mathbb{R}^d$

$$\|u\|^2/2 - \varphi_{b;n,m}(u) \geq \inf_{v\in\mathbb{R}^d}\{L_2(u,v) - L_{b;n,m}\} = -L_{b;n,m}. \tag{2.41}$$

Moreover, we have

$$\begin{aligned}\|u\|^2/2 - \varphi_{b;n,m}(u) &\leq -\big(\|u\|^2/2 - \zeta_{b;n,m}(u)\big)\\ &= -\inf_{u'\in\mathrm{Supp}(P_n^{A(b)})}\big(L(u,u') - (\|u'\|^2/2 - \tilde{\varphi}_{b;n,m}(u'))\big) \wedge L_{b;n,m} \leq 0,\end{aligned} \tag{2.42}$$

where the last step follows by the fact that $\|u'\|^2/2 - \tilde{\varphi}_{b;n,m}(u') \leq 0$, for all $u' \in \mathrm{Supp}(P_n^{A(b)})$. Here, we proved that $-L_{b;n,m} \leq \frac{\|u\|^2}{2} - \varphi_{b;n,m}(u) \leq 0$ and the result holds. For any $v \in \mathbb{R}^d$, by (2.42) we have

$$\|v\|^2/2 - \varphi^*_{b;n,m}(v) = \inf_{u\in\mathbb{R}^d}\Big(L_2(u,v) - (\|u\|^2/2 - \varphi_{b;n,m}(u))\Big) \geq 0. \tag{2.43}$$

Moreover, by (2.41) it follows that

$$\|v\|^2/2 - \varphi^*_{b;n,m}(v) \leq -(\|v\|^2/2 - \varphi_{b;n,m}(v)) \leq L_{b;n,m}. \tag{2.44}$$

Thus we have $0 \leq \frac{\|v\|^2}{2} - \varphi^*_{b;n,m}(v) \leq L_{b;n,m}$ as desired.

To prove (iv), note (ii) implies that

$$\int (\varphi_{b;n,m}(u) + \varphi^*_{b;n,m}(v) - v^T u)\, d\pi_{b;n,m}(u,v) = 0.$$

Furthermore, part (i) implies that the integrand of the above is nonnegative. Thus it follows that

$$\varphi_{b;n,m}(u) + \varphi^*_{b;n,m}(v) = v^T u, \quad \forall (u,v) \in \mathrm{Supp}(\pi_{b;n,m}).$$

Then the conclusion follows by [Vil21, Proposition 2.4]. □

Now we argue that for all $b \in \mathcal{B}$, $\varphi^*_{b;n,m}$ (and similarly, $\varphi_{b;n,m}$) is a piecewise Lipschitz function on a high probability event that does not depend on b. The following lemma plays a key role in the argument. It implies that the local Lipschitz constant of $\varphi^*_{b;n,m}$ is largely driven by the magnitude of the subdifferential of $\varphi^*_{b;n,m}$. The proof is analogous to Manole and Niles-Weed [MN24, Lemma 10], but for the sake of completeness, we provide it here.

Lemma 2.6. *Suppose P and Q are two distributions on $\mathbb{R}^d$. Let (φ_0, φ_0^*) be the conjugate pair that solves $\tilde{I}_2(P,Q)$ (see (2.17)). Then for any $r \geq 1$, $\varphi_0 : \mathcal{B}^d_{0,r} \to \mathbb{R}$ and $\varphi_0^* : \mathcal{B}^d_{0,r} \to \mathbb{R}$ are Lipschitz continuous with parameters L_0 and L_0^* respectively, where*

$$L_0 := \sup\big\{\|y\| : y \in \partial\varphi_0\big(\mathcal{B}^d_{0,r}\big)\big\} \quad \text{, and} \quad L_0^* := \sup\big\{\|z\| : z \in \partial\varphi_0^*\big(\mathcal{B}^d_{0,r}\big)\big\}$$

Proof. We focus on φ_0 and the same argument can be used for φ_0^*. Firstly, by Villani [Vil21, Proposition 2.4], for any $v \in \mathcal{B}^d_{0,r}$, φ_0 admits the following representation

$$\varphi_0(v) = \sup_{u\in\partial\varphi_0(v)}\big\{u^T v - \varphi_0^*(u)\big\}.$$

Thus, there exists a sequence of $u_k \in \partial\varphi_0(v)$ such that

$$\varphi_0(v) \le u_k^T v - \varphi_0^*(u_k) + \frac{1}{k}, \quad \text{for } k = 1, 2, \dots.$$

Then for any $v' \in \mathcal{B}_{0,r}^d$, we have

$$\begin{aligned}\varphi_0(v) - \varphi_0(v') &\le u_k^T v - \varphi_0^*(u_k) + \frac{1}{k} - u_k^T v' + \varphi_0^*(u_k) \\ &= u_k^T(v - v') + \frac{1}{k} \le L_0\|v - v'\| + \frac{1}{k},\end{aligned}$$

and the Lipschitz property follows by letting $k \to +\infty$. □

For all $j \ge 0$, define $L_j := [-3^j, 3^j]^d$ and let $P_j := L_j \setminus L_{j-1}$. We note that each P_j can be further partitioned into $N := 3^d - 1$ cubes, say $\{P_{j,k}\}_{k=1,\dots,N}$, that are each congruent to L_{j-1}. We note that all elements of P_j has norm bounded by $\ell_j := \sup_{z\in P_j}\|z\| = 3^j\sqrt{d}$.

For any $I \subset \mathbb{R}^d$, we write $\mathcal{C}(I)$ for the set of all the convex function on I. We define $\mathcal{C}_{m,u}(I) := \{f \in \mathcal{C}(I) : \exists m, u > 0, \text{ s.t. } |f(x) - f(y)| \le m\|x - y\|,\ |f(x)| \le u,\ \forall x, y \in I\}$ to be the class of m-Lipschitz convex functions on I bounded in value by u. Given a sequence M and U, define

$$\mathcal{C}_{M,U} := \big\{f : \mathbb{R}^p \times \mathbb{R}^d \to \mathbb{R} : f|_{P_{j,k}} \in \mathcal{C}_{M_j,U_j}(P_{j,k}),\ j \ge 0,\ 1 \le k \le N\big\}.$$

We now prove that for suitable choices of M, U and R, T, $\varphi_{b;n,m}^* - \varphi_{b;n,m}^*(0) \in \mathcal{C}_{M,U}$ and $\varphi_{b;n,m} - \varphi_{b;n,m}(0) \in \mathcal{C}_{R,T}$ on a high probability event that does not depend on b. Recalling that we write $S = \Sigma^{-1/2}X$ and $S_i = \Sigma^{-1/2}X_i$ for $i \in [n]$.

Let's first discuss the concentration property of P^U and $P^{A(b)}$ and their empirical counterparts P_m^U and $P_n^{A(b)}$. In fact, due to the Gaussian assumption, P^U is a $(\sqrt{2d}, 2)$-sub-Weibull distribution. Moreover, by the sub-Weibull assumptions on S and ε, there exists a constant $\sigma > 0$ depends on σ_1, σ_2 such that $\|S\| + \|\varepsilon\| \sim (\sigma, \alpha \wedge \beta)$-sub-Weibull. Thus by noting that $\|A(b)\| \le \|S\| + \|\varepsilon\|$ for all $b \in \mathcal{B}$, $P^{A(b)}$ is a $(\sigma, \alpha \wedge \beta)$-sub-Weibull random vector as well. However, the concentration of the corresponding empirical measures introduces extra randomness on the sub-Weibull parameters, as defined here

$$E_{1,m} = \int \exp\Big(\frac{\|u\|^2}{4d}\Big)\, dP_m^U, \quad \text{and} \quad E_{b;2,n} = \int \exp\Big(\frac{\|v\|^{\alpha\wedge\beta}}{4\sigma^{\alpha\wedge\beta}}\Big)\, dP_n^{A(b)}.$$

The following lemma constructs the sub-Weibull properties of P_m^U and $P_n^{A(b)}$.

Lemma 2.7. *Define $E_{2,n} := \sup_{b\in\mathcal{B}} E_{b;2,n}$. Then for any fixed $n, m \ge 1$ we have that P_m^U is $((2dE_{1,m})^{1/2}, 2)$-sub-Weibull and $P_n^{A(b)}$ is $(\sigma(2E_{2,n})^{1/(\alpha\wedge\beta)}, \alpha \wedge \beta)$-sub-Weibull, where $E_{1,m} \le 2 + \sqrt{\frac{\log m}{m}}$ with probability at least $1 - 2(\log m)^{-1}$ and $E_{2,n} \le 2 + \sqrt{\frac{\log n}{n}}$ with probability at least $1 - 2(\log n)^{-1}$.*

Proof. We only need to note that $E_{1,m} \ge 1$, and Jensen's inequality yields that

$$\int \exp\Big(\frac{\|u\|^2}{4dE_{1,m}}\Big)\, dP_m^U \le E_{1,m}^{\frac{1}{E_{1,m}}} \le 2.$$

One the other hand, for each fixed $b \in \mathcal{B}$, a similar calculation can be applied to $P_n^{A(b)}$ and obtain that $P_n^{A(b)} \sim (\sigma(2E_{b;2,n})^{1/(\alpha\wedge\beta)}, \alpha \wedge \beta)$-sub-Weibull. Thus by noting that

$$\int \exp\Big(\frac{\|v\|^{\alpha\wedge\beta}}{4E_{b;2,n}\sigma^{\alpha\wedge\beta}}\Big)\, dP_n^{A(b)}(v) \geq \int \exp\Big(\frac{\|v\|^{\alpha\wedge\beta}}{4E_{2,n}\sigma^{\alpha\wedge\beta}}\Big)\, dP_n^{A(b)}(v)$$

we have $P_n^{A(b)} \sim (\sigma(2E_{2,n})^{1/(\alpha\wedge\beta)}, \alpha \wedge \beta)$-sub-Weibull

Now we control the sub-Weibull parameters. Define $\Gamma_1 := \big\{E_{1,m} \leq 2 + \sqrt{\frac{\log m}{m}}\big\}$, then by the Chebyshev's inequality we have

$$\mathbb{P}(\Gamma_1^c) \leq \mathbb{P}\Big(|E_{1,m} - \mathbb{E}\, E_{1,m}| \geq \sqrt{\frac{\log m}{m}}\Big) \leq \frac{m\mathrm{Var}(E_{1,m})}{\log m} \leq \frac{2}{\log m}.$$

To control $E_{2,n}$, we first note

$$\mathbb{E}\, E_{2,n} = \mathbb{E} \sup_{b\in\mathcal{B}} \exp\Big(\frac{1}{4}\big(\frac{\|A(b)\|}{\sigma}\big)^{\alpha\wedge\beta}\Big) \leq \mathbb{E} \exp\Big(\frac{1}{4}\big(\frac{\|S\| + \|\varepsilon\|}{\sigma}\big)^{\alpha\wedge\beta}\Big) \leq 2.$$

Then define $\Gamma_2 := \Big\{E_{2,n} \leq 2 + \sqrt{\frac{\log n}{n}}\Big\}$, then we have

$$\begin{aligned}
\mathbb{P}(\Gamma_2^c) &\leq \mathbb{P}\Big(E_{2,n} - \mathbb{E}\exp\Big(\frac{1}{4}\big(\frac{\|S\| + \|\varepsilon\|}{\sigma}\big)^{\alpha\wedge\beta}\Big) \geq \sqrt{\frac{\log n}{n}}\Big) \\
&\leq \mathbb{P}\Big(\frac{1}{n}\sum_{i=1}^n \exp\Big(\frac{1}{4}\big(\frac{\|S_i\| + \|\varepsilon_i\|}{\sigma}\big)^{\alpha\wedge\beta}\Big) - \mathbb{E}\exp\Big(\frac{1}{4}\big(\frac{\|S\| + \|\varepsilon\|}{\sigma}\big)^{\alpha\wedge\beta}\Big) \geq \sqrt{\frac{\log n}{n}}\Big) \\
&\leq \frac{2}{\log n},
\end{aligned}$$

where the final inequality is obtained by Chebyshev's inequality. □

Proposition 2.3. *Let* $J_n = \Big\lfloor \frac{1}{2}\log_3\Big(\frac{\log n}{16\gamma_2 d}\Big)\Big\rfloor$, $I_m = \Big\lfloor \frac{1}{2}\log_3\Big(\frac{\log m}{8d}\Big)\Big\rfloor$ *and* $l_{n,m} = (\log m) \vee (\log n)^{2/(\alpha\wedge\beta)}$. *Then there exist an event* Υ *with probability at least* $1 - 12(\log n)^{-1}$ *and constants* $C_i', C_i', \tilde{C}_i, \tilde{C}_i > 0$ *depends on* $d, \gamma_1, \gamma_2, \sigma_1, \sigma_2, \alpha, \beta$ *such that on* Υ, *for all* $b \in \mathcal{B}$, *we have* $\varphi^*_{b;n,m} - \varphi^*_{b;n,m}(0) \in \mathcal{C}_{M,U}$ *and* $\varphi_{b;n,m} - \varphi_{b;n,m}(0) \in \mathcal{C}_{R,T}$ *where* M *and* U *are chosen as*

$$M_j = \begin{cases} C_0'\ell_j, & 0 \leq j \leq J_n \\ C_1' l_{n,m}\ell_j, & j > J_n \end{cases}, \quad U_j = \begin{cases} \tilde{C}_0\ell_j^3, & 0 \leq j \leq J_n \\ \tilde{C}_1 l_{n,m}\ell_j^3, & j > J_n \end{cases}, \tag{2.45}$$

and R *and* T *are chosen as*

$$R_i = \begin{cases} C_2'\ell_i, & 0 \leq i \leq I_m \\ C_3' l_{n,m}\ell_i, & i > I_m \end{cases}, \quad T_i = \begin{cases} \tilde{C}_2\ell_i^3, & 0 \leq i \leq I_m \\ \tilde{C}_3 l_{n,m}\ell_i^3, & i > I_m \end{cases}. \tag{2.46}$$

Proof. Note Lemma 2.6 implies that in order to quantify the Lipschitz constant of $\varphi^*_{b;n,m}$ on $P_{j,k}$, we only need to bound the magnitude of $\sup\{\|y\| : y \in \partial\varphi^*_{b;n,m}(P_{j,k})\}$. To this end, we first note that $\partial\varphi^*_{b;n,m}(v) = \partial^c(\|\cdot\|^2/2 - \varphi^*_{b;n,m})(v)$ and $\|\cdot\|^2/2 - \varphi^*_{b;n,m}$ is obviously a c-concave function.

Thus by Lemma 2.2(iv) and Lemma 2.7, we can apply Manole and Niles-Weed [MN24, Theorem 11] to obtain[2] that there exists a constant $C_0 > 0$ depends on d such that for any $v \in P_{j,k}$ and $y \in \partial\varphi^*_{b;n,m}(v)$, we have

$$\|y\| \le C_0(2dE_{1,m})^{1/2}\Big\{(\|v\|+1) \vee \sup_{w:\|v-w\|\le 2}\Big[\log\Big(\frac{1}{P_n^{A(b)}(\mathcal{B}^d_{w,3})}\Big)\Big]^{1/2}\Big\}. \tag{2.47}$$

Thus to upper bound the magnitude of $\partial\varphi^*_{b;n,m}(v)$ we only need to prove an anticoncentration bound for $P_n^{A(b)}$.

We first note that from (2.12), for any $0 \le j \le J_n$, $v \in P_j$ and w such that $\|w - v\| \le 2$, we have

$$\begin{aligned} P^\varepsilon(\mathcal{B}^d_{w,2}) \ge \int_{\mathcal{B}^d_{w,2}\setminus\mathcal{B}^d_0} \gamma_1 \exp\big(-\gamma_2\|e\|^2\big)\,de &\ge \frac{\pi^{d/2}(2^d-1)}{\Gamma(d/2+1)}\gamma_1 \exp\big(-2\gamma_2\|z\|^2 - 50\gamma_2\big) \\ &\ge 2K_1\exp\big(-2\gamma_2\ell_j^2\big), \end{aligned} \tag{2.48}$$

where $K_1 \in (0,1)$ is a constant depending on d, γ_1, γ_2. Observe that the right-hand side does not depend on z or w, hence, we may take infimum over $v \in P_j$ and w such that $\|w - v\| \le 2$ and have the same lower bound. Hence, we have

$$P^\varepsilon \otimes P^S(\mathcal{B}^d_{w,2} \times \mathcal{B}^p_0) = P^\varepsilon(\mathcal{B}^d_{w,2})P^S(\mathcal{B}^p_0) \ge 2K_1'\exp\big(-2\gamma_2\ell_j^2\big),$$

for some $K_1' \in (0,1)$ depends on $d, \gamma_1, \gamma_2, \sigma_1$ and α, where the sub-Weibull assumption on S has been exploited in the final inequality.

On the other hand, let $\mathcal{B}^d := \{\mathcal{B}^d_{a,r} : a \in \mathbb{R}^d, r > 0\}$ be the set of all balls in $\mathbb{R}^d$. Let $\tilde{u} = \sqrt{\frac{160d\log n}{n}}$ and define

$$\Upsilon_1 := \Big\{\sup_{B\in\mathcal{B}^d}|P_n^\varepsilon \otimes P_n^S(B \times \mathcal{B}^p_{0,}) - P^\varepsilon \otimes P^S(B \times \mathcal{B}^p_0)| < \tilde{u}\Big\}.$$

Thus, since $\tilde{u} \lesssim K_1' n^{-1/8} \le K_1' e^{-2\gamma_2\ell_j^2}$ for $0 \le j \le J_n$, working on Υ_1 we have $P_n^\varepsilon \otimes P_n^S(\mathcal{B}^d_{w,2} \times \mathcal{B}^d_0) \ge K_1'\exp\big(-2\gamma_2\ell_j^2\big)$. Thus consider the event

$$\Upsilon_2 := \bigcap_{j=0}^{J_n}\Big\{\inf_{v\in P_j}\inf_{w:\|v-w\|\le 2} P_n^\varepsilon \otimes P_n^S(\mathcal{B}^d_{w,2} \times \mathcal{B}^p_0) \ge K_1'\exp\big(-2\gamma_2\ell_j^2\big)\Big\},$$

we have $\Upsilon_1 \subset \Upsilon_2$. Note the Vapnik–Chervonenkis (VC) dimension of $\mathcal{B}^d$ is no more than $d+2$ [See e.g. DGL13, Corollary 13.2], by the VC-inequality [see VC15, Theorem 2] we have

$$\mathbb{P}(\Upsilon_1^c) \lesssim n^{d+2}\exp\big(-n\tilde{u}^2/32\big) \le n^{2-4d} \le n^{-2}, \tag{2.49}$$

[2]We remark that the bound given below uses the probability mass on $\mathcal{B}^d_{w,3}$ whereas the original formulation in Manole and Niles-Weed [MN24, Theorem 11] has $\mathcal{B}^d_{w,1}$ instead. We have used a slightly different radius here for the convenience of the subsequent argument. The exact radius is unimportant in the argument used in that theorem and the same proof will work verbatim with radius changed to 3.

whence $\mathbb{P}(\Upsilon_2) \geq 1 - n^{-2}$. Thus working on $\Upsilon_2 \cap \Gamma_1$, by $\|A(b)_i\| \leq \|S_i\| + \|\varepsilon_i\|$ for all $b \in \mathcal{B}$ and $i \in [n]$, we have $P_n^{A(b)}(\mathcal{B}_{w,3}^d) \geq K_1' \exp\big(-2\gamma_2 \ell_j^2\big)$, and combining this with (2.47), we conclude that for any $1 \leq k \leq N$ and $0 \leq j \leq J_n$, there exits some sufficiently large constant $C_0' > 0$ depends on $d, \gamma_1, \gamma_2, \sigma_1, \alpha$ such that

$$\begin{aligned}\sup_{y \in \partial\varphi^*_{b;n,m}(P_{j,k})} \|y\| &\leq C_0 (2dE_{1,m})^{1/2} \Big(\ell_j + 1 + \sqrt{2}\ell_j \gamma_2^{1/2} + \sqrt{\log(1/K_1')}\Big) \\ &\leq C_0' E_{1,m}^{1/2} \ell_j \leq C_0' \Big(2 + \sqrt{\frac{\log m}{m}}\Big)^{1/2} \ell_j \lesssim C_0' \ell_j := M_j. \end{aligned} \tag{2.50}$$

When $j > J_n$, by Lemma 2.2(iii) and Manole and Niles-Weed [MN24, Proposition 16], we only need to bound $L_{b;n,m}$. Note $L_{b;n,m} \leq L_{n,m} := 2\max_{i\in[n]} \|\Sigma^{-1/2} X_i\|^2 + 2\max_{i\in[n]} \|\varepsilon_i\|^2 + 2\max_{j\in[m]} \|U_j\|^2$. Define $r_{n,m} := 2\sigma_1^2 (4\log n)^{2/\alpha} + 2\sigma_2^2 (4\log n)^{2/\beta} + 8d \log m$ and consider the event $\Upsilon_3 := \{L_{n,m} < r_{n,m}\}$. By part(i) of Proposition 2.5 and union bound, it follows that

$$\begin{aligned}\mathbb{P}\big(\Upsilon_3^c\big) &\leq \mathbb{P}\big(\max_{i\in[n]} \|\Sigma^{-1/2} X_i\|^2 \geq \sigma_1^2 (4\log n)^{2/\alpha}\big) + \mathbb{P}\big(\max_{i\in[n]} \|\varepsilon_i\|^2 \geq \sigma_2^2 (4\log n)^{2/\beta}\big) \\ &\quad + \mathbb{P}\big(\max_{j\in[m]} \|U_j\|^2 \geq 8d\log m\big) \leq 4n^{-1} + 2m^{-1}. \end{aligned} \tag{2.51}$$

Therefore on the event Υ_3, by Manole and Niles-Weed [MN24, Proposition 16] we have that there exists a universal constant $C_1 > 0$ and a sufficiently large $C_1' > 0$ depends on σ_1, σ_2 such that for any $1 \leq k \leq N$,

$$\sup_{y \in \partial\varphi^*_{b;n,m}(P_{j,k})} \|y\| \leq C_1(\ell_j + r_{n,m}) \leq C_1' l_{n,m} \ell_j =: M_j, \quad \text{for all } j > J_n. \tag{2.52}$$

Putting (2.50) and (2.52) together, for some constants $\tilde{C}_0, \tilde{C}_1 > 0$ depend on $d, \gamma_1, \gamma_2, \sigma_1, \sigma_2, \alpha$, we have $\varphi^*_{b;n,m} - \varphi^*_{b;n,m}(0) \in \mathcal{C}_{M,U}$ on the event $\Upsilon' := \Upsilon_2 \cap \Gamma_1 \cap \Upsilon_3$, where $M = (M_j)_{j\geq 0}$ and $U = (U_j)_{j\geq 0}$ are chosen as

$$M_j = \begin{cases} C_0' \ell_j, & 0 \leq j \leq J_n \\ C_1' l_{n,m} \ell_j, & j > J_n \end{cases}, \quad U_j = \begin{cases} \tilde{C}_0 \ell_j^3, & 0 \leq j \leq J_n \\ \tilde{C}_1 l_{n,m} \ell_j^3, & j > J_n \end{cases},$$

as desired.

A similar argument can be applied to study the Lipschitz property of $\varphi_{b;n,m}$. Since $U \sim \mathcal{N}(0, I_d)$, for all $i \leq I_m$, $u \in P_i$ and all w such that $\|w - u\| \leq 2$, we have

$$P^U(\mathcal{B}_w^d) = \int_{\mathcal{B}_w^d} (2\pi)^{-d/2} \exp\big(-\|y\|^2/2\big)\, dy \geq 2K_2 e^{-\ell_i^2},$$

where $K_2 \in (0,1)$ is a constant depends only on d. Let $\tilde{v} = \sqrt{\frac{160 d \log m}{m}}$, and define

$$\Upsilon_4 := \Big\{ \sup_{B\in\mathcal{B}^d} |P_m^U(B) - P^U(B)| < \tilde{v} \Big\}, \text{ and } \Upsilon_5 := \bigcap_{i=0}^{I_m} \Big\{ \inf_{u\in P_i} \inf_{w:\|u-w\|\leq 2} P_n^U(\mathcal{B}_w^d) \geq K_2 e^{-\ell_i^2} \Big\}.$$

Then since $\tilde{v} \leq m^{-1/8} \leq K_2 e^{-\ell_{I_m}^2}$ we have $\Upsilon_4 \subset \Upsilon_5$. Furthermore, by leveraging the VC-inequality again, we can deduce that $\mathbb{P}(\Upsilon_4^c) \leq m^{-2}$, which implies that $\mathbb{P}(\Upsilon_5) \geq 1 - m^{-2}$. On

the event $\Upsilon_5 \cap \Gamma_2$, by applying Manole and Niles-Weed [MN24, Theorem 11] and Lemma 2.6 again we obtain that for $0 \le i \le I_m$, there exists constants $C_2 > 0$ depends on d, α, β and $C_2' > 0$ depends on d, σ, α, β such that

$$\begin{aligned}\sup_{z \in \partial\varphi_{b;n,m}(u)} \|z\| &\le C_2 \sigma (2E_{2,n})^{1/(\alpha\wedge\beta)} (2\ell_i + 1 + \sqrt{\log(1/K_2)}) \\ &\le C_2' E_{2,n}^{1/(\alpha\wedge\beta)} \ell_i \le C_2' \Big(2 + \sqrt{\frac{\log n}{n}}\Big)^{1/(\alpha\wedge\beta)} \ell_i \lesssim C_2' \ell_i := R_i. \end{aligned} \tag{2.53}$$

When $i > I_m$, since we still have $|\|u\|^2/2 - \varphi_{b;n,m}| \le L_{b;n,m} \le L_{n,m}$ by Lemma 2.2(iii), working on the event Υ_3, there exists an absolute constant $C_3 > 0$, and $C_3' > 0$ depends on σ_1, σ_2 such that for $1 \le k \le N$.

$$\sup_{z \in \partial\varphi_{b;n,m}(P_{i,k})} \|z\| \le C_3(\ell_i + r_{n,m}) \le C_3' l_{n,m} \ell_i := R_i \quad \text{for } i > I_m. \tag{2.54}$$

Thus combine (2.53) and (2.54) we can deduce that there exists constants $\tilde{C}_2, \tilde{C}_3 > 0$ depend on $d, \alpha, \beta, \sigma_1, \sigma_2$ such that $\varphi_{b;n,m} - \varphi_{b;n,m}(0) \in \mathcal{C}_{R,T}$ on the event $\Upsilon = \Upsilon' \cap \Upsilon_5 \cap \Gamma_2$, where

$$R_i = \begin{cases} C_2' \ell_i, & 0 \le i \le I_m \\ C_3' l_{n,m} \ell_i, & i > I_m \end{cases}, \quad T_i = \begin{cases} \tilde{C}_2 \ell_i^3, & 0 \le i \le I_m \\ \tilde{C}_3 l_{n,m} \ell_i^3, & i > I_m \end{cases}.$$

Finally, combining the controls on the probability of $\Upsilon_2, \Upsilon_3, \Upsilon_5, \Gamma_1$ and Γ_2, we arrive at the the upper bound $\mathbb{P}(\Upsilon) \ge 1 - n^{-2} - 4n^{-1} - 2m^{-1} - m^{-2} - 2(\log m)^{-1} - 2(\log n)^{-1} \ge 1 - 12(\log n)^{-1}$ when $m \ge n$, which completes the proof. □

Now we are ready to introduce the core proposition in the proof.

Proposition 2.4. *There exists a constant $C > 0$ depending on $d, \alpha, \beta, \sigma_1, \sigma_2, \gamma_1, \gamma_2$ such that for any fixed $n, m \ge 1$, the following inequality holds*

$$\sup_{b \in \mathcal{B}} \big| \mathcal{W}_2^2(P^{A(b)}, P^U) - \mathcal{W}_2^2(P_n^{A(b)}, P_m^U) \big| \le C (\log m)^{\frac{8}{2\wedge\alpha\wedge\beta}} \left(\sqrt{\frac{p}{n}} + \frac{1}{n^{2/d}} \right) \tag{2.55}$$

with probability at least $1 - 29(\log n)^{-1}$.

Proof. Note part (i) and part (iii) of Lemma 2.2 implies that $(\|\cdot\|^2 - \varphi_{b;n,m}^*, \|\cdot\|^2 - \varphi_{b;n,m})$ is a feasible pair to the duality of the Kantorovich problem between $P^{A(b)}$ and P^U. This yields that

$$\begin{aligned}\frac{1}{2}\mathcal{W}_2^2(P^{A(b)}, P^U) &\ge \int \Big(\frac{\|v\|^2}{2} - \varphi_{b;n,m}^*(v) \Big) dP^{A(b)}(v) + \int \Big(\frac{\|u\|^2}{2} - \varphi_{b;n,m}(u) \Big) dP^U(u) \\ &= \int \frac{\|v\|^2}{2} dP_n^{A(b)}(v) + \int \frac{\|u\|^2}{2} dP_m^U(u) \\ &\quad + \int \frac{\|v\|^2}{2} (dP^{(A(b)} - dP_n^{A(b)})(v) + \int \frac{\|u\|^2}{2} (dP^U - dP_m^U)(u) \\ &\quad - \Big\{ \int \varphi_{b;n,m}^*(v) dP_n^{A(b)}(v) + \varphi_{b;n,m}(u) dP_m^U(u) \Big\} \\ &\quad - \Big\{ \int \varphi_{b;n,m}^*(v) d\big(P^{A(b)} - P_n^{A(b)}\big)(v) + \int \varphi_{b;n,m}(u) d\big(P^U - P_m^U\big)(u) \Big\}. \end{aligned}$$

By the definition of $(\varphi_{b;n,m}, \varphi^*_{b;n,m})$, we have $\mathcal{W}_2^2(P_n^{A(b)}, P_m^U) = \int \big(\frac{\|v\|^2}{2} - \varphi^*_{b;n,m}(v)\big) dP_n^{A(b)}(v) + \int \big(\frac{\|u\|^2}{2} - \varphi_{b;n,m}(u)\big) dP_m^U(u)$. Consequently, from the above display, we deduce that

$$\begin{aligned}
&\frac{1}{2}\mathcal{W}_2^2(P_n^{A(b)}, P_m^U) - \frac{1}{2}\mathcal{W}_2^2(P^{A(b)}, P^U) \\
&\qquad \le \underbrace{\int \varphi^*_{b;n,m}(v) d\big(P^{A(b)} - P_n^{A(b)}\big)(v) + \int \varphi_{b;n,m}(u) d\big(P^U - P_m^U\big)(u)}_{=:E_{b;n,m}} \\
&\qquad\quad + \underbrace{\int \frac{\|v\|^2}{2} d\big(P_n^{A(b)} - P^{A(b)}\big)(v) + \int \frac{\|u\|^2}{2} d\big(P_m^U - P^U\big)(u)}_{=:F_{b;n,m}}. \qquad (2.56)
\end{aligned}$$

On the other hand, define $\Psi_b := \{(f,g) \in L^1(P^{A(b)}) \times L^1(P^U) : v^T u \le f(v) + g(u),\ \forall (v,u) \in \mathrm{Supp}(P^{A(b)}) \times \mathrm{Supp}(P^U)\}$, then Theorem 2.3 implies that for any $b \in \mathcal{B}$ there exists a conjugate pair (ψ_b^*, ψ_b) such that

$$\begin{aligned}
(\psi_b^*, \psi_b) &= \operatorname*{arg\,min}_{f,g \in \Psi_b} \int f \, dP^{A(b)} + \int g \, dP^U, \\
\frac{1}{2}\mathcal{W}_2^2(P^{A(b)}, P^U) &= \int \|v\|^2/2 - \psi_b^*(v) \, dP^{A(b)}(v) + \int \|u\|^2/2 - \psi_b(u) \, dP^U(u).
\end{aligned}$$

Since $\Psi_b \subseteq \tilde{\Phi}_b$, $(\|v\|^2/2 - \psi_b^*(v), \|u\|^2/2 - \psi_b(u))$ is a feasible solution for the duality between $P_n^{A(b)}$ and P_m^U. Therefore, we can rerun the previous derivation and obtain that

$$\begin{aligned}
&\frac{1}{2}\mathcal{W}_2^2(P_n^{A(b)}, P_m^U) - \frac{1}{2}\mathcal{W}_2^2(P^{A(b)}, P^U) \\
&\qquad \ge \underbrace{\int \psi_b^*(v) d\big(P^{A(b)} - P_n^{A(b)}\big)(v) + \int \psi_b(u) d\big(P^U - P_m^U\big)(u)}_{=:G_{b;n,m}} \\
&\qquad\quad + \int \frac{\|v\|^2}{2} d\big(P_n^{A(b)} - P^{A(b)}\big)(v) + \int \frac{\|u\|^2}{2} d\big(P_m^U - P^U\big)(u). \qquad (2.57)
\end{aligned}$$

Write the first two terms and the last two terms of (2.56) as $E_{n,m}$ and $F_{n,m}$ respectively, and write the first two terms of (2.57) as $G_{n,m}$. Then combining (2.56) and (2.57), for ϑ_k defined in (2.68), we have

$$\begin{aligned}
\sup_{b \in \mathcal{B}} \Big| \frac{1}{2}\mathcal{W}_2^2(P_n^{A(b)}, P_m^U) - \frac{1}{2}\mathcal{W}_2^2(P^{A(b)}, P^U) \Big| &\le \sup_{b \in \mathcal{B}} |E_{b;n,m}| + 2 \sup_{b \in \mathcal{B}} |F_{b;n,m}| + \sup_{b \in \mathcal{B}} |G_{b;n,m}| \\
&\lesssim (\log m)^{\frac{6}{2 \wedge \alpha \wedge \beta}} \big(\vartheta_n + \sqrt{\frac{p}{n}} + \sqrt{\frac{\log n}{n}} + \vartheta_m + \sqrt{\frac{\log m}{m}} \big), \\
&\le (\log m)^{\frac{8}{2 \wedge \alpha \wedge \beta}} \left(\sqrt{\frac{p}{n}} + \frac{1}{n^{2/d}} \right). \qquad (2.58)
\end{aligned}$$

with probability at least $1 - 29(\log n)^{-1}$, where we have used Lemmas 2.8, 2.11 and 2.10 to bound each of the three terms in the penultimate inequality. □

Lemma 2.8. *There exists $C > 0$, depending only on $d, \alpha, \beta, \gamma_2, \sigma_1, \sigma_2$, and an event Ω with probability at least $1 - 18(\log n)^{-1}$, such that on Ω, for any $b \in \mathcal{B}$, we have*

$$\left| \int \varphi^*_{b;n,m}(v)\, d\big(P^{A(b)} - P_n^{A(b)}\big)(v) \right| \le C(\log m)^{\frac{6}{2\wedge\alpha\wedge\beta}} \left(\vartheta_n + \sqrt{\frac{p}{n}} + \sqrt{\frac{2\log n}{n}} \right)$$
$$\left| \int \varphi_{b;n,m}(u)\, d\big(P^{U} - P_m^{U}\big)(u) \right| \le C(\log m)^{\frac{6}{2\wedge\alpha\wedge\beta}} \left(\vartheta_m + \sqrt{\frac{2\log m}{m}} \right),$$

where ϑ_n is defined as (2.68).

Proof. We note that the value of the integrals on the left-hand side of both inequalities will not change if we add any constant to the functions $\phi^*_{b;n,m}$ and $\phi_{b;n,m}$. Hence, we may assume without loss of generality throughout this proof that $\phi^*_{b;n,m}(0) = \phi_{b;n,m}(0) = 0$.

Note that due to the sub-Weibull assumptions on ε and S, and combining with Proposition 2.6(ii), we have $(\varepsilon, S) \sim (\rho, \alpha\wedge\beta)$-sub-Weibull for $\rho > 0$ depending only on σ_1 and σ_2. Then let $\kappa = \rho(4\log n)^{1/(\alpha\wedge\beta)}$ and $\Omega_1 := \{\max_{1\le i\le n} \|(\varepsilon_i, S_i)\| \le \kappa\}$, and by Proposition 2.5(i), we have

$$\mathbb{P}(\Omega_1^c) \le n\,\mathbb{P}(\|(\varepsilon, S)\| \ge \kappa) \le 2n \exp\left\{ -\frac{1}{2}(\kappa/\rho)^{\alpha\wedge\beta} \right\} \le \frac{2}{n}.$$

For any $b \in \mathcal{B}$, define the linear projection $T_b : \mathbb{R}^p \times \mathbb{R}^d \to \mathbb{R}^d$ such that

$$T_b(s, e) := (b^* - b)\Sigma^{1/2} s + e. \tag{2.59}$$

Write $E_b = \{T_b(s,e) \in \mathbb{R}^d : (s,e) \in \mathcal{B}_{0,\kappa}^{d+p}\}$. Working on the event Ω_1 and observing that $\|T_b\|_{\mathrm{op}} \le 1$ for any $b \in \mathcal{B}$, we have

$$\begin{aligned}
\int_{\mathbb{R}^d\setminus E_b} \varphi^*_{b;n,m}(v)\, d\big(P^{A(b)} - P_n^{A(b)}\big)(v) &= \int_{\mathbb{R}^{d+p}\setminus\mathcal{B}_{0,\kappa}^{d+p}} \varphi^*_{b;n,m} \circ T_b(e,s)\, dP^\varepsilon \otimes P^S(e,s) \\
&\overset{(a)}{\le} \int_{\mathbb{R}^{d+p}\setminus\mathcal{B}_{0,\kappa}^{d+p}} \left(\frac{\|T_b(e,s)\|^2}{2} + r_{n,m} \right) d(P^\varepsilon \otimes P^S)(e,s) \\
&\le \int_{\mathbb{R}^{d+p}\setminus\mathcal{B}_{0,\kappa}^{d+p}} \left(\frac{\|(e,s)\|^2}{2} + r_{n,m} \right) d(P^\varepsilon \otimes P^S)(e,s) \\
&\overset{(b)}{\le} C_4 e^{-\frac{1}{4}(\frac{\kappa}{\rho})^{\alpha\wedge\beta}} + \frac{2r_{n,m}}{n^2} \lesssim \frac{C_4}{n},
\end{aligned} \tag{2.60}$$

where we use part (iii) of Proposition 2.2 to obtain (a) and Lemma 2.14 to obtain (b) and $C_4 > 0$ is a constant only depending on $d, \sigma_1, \sigma_2, \alpha, \beta$.

On the other hand, for $\mathcal{X} \subseteq \mathbb{R}^d$, we define $\mathrm{Lip}_{1,1}(\mathcal{X}) := \{f \in \mathrm{Lip}_1(\mathcal{X}) : \sup_{x\in\mathcal{X}} |f(x)| \le 1\}$ to be the class of 1-Lipschitz functions on $\mathcal{X}$ uniformly bounded by 1. Consider the following function class

$$\mathcal{F} := \left\{ (s,e) \mapsto (\varphi \circ T_b)(s,e)\, \mathbb{1}_{\mathcal{B}_0^{d+p}}(s,e) : b \in \mathcal{B},\ \varphi \in \mathrm{Lip}_{1,1}(\mathcal{B}_0^d) \right\}. \tag{2.61}$$

Let $j_n = (J_n + 1) + \lceil \log_3(\rho(4\log n)^{1/(\alpha\wedge\beta)}/d^{1/2}) \rceil$. Then we have $3^{j_n}\sqrt{d} \ge \kappa$, which implies that $\mathcal{B}_{0,\kappa}^d \subseteq \bigcup_{j=0}^{j_n} \bigcup_{k=1}^{N} P_{j,k}$ for $P_{j,k}$ defined before Lemma 2.7. Let Υ be the event with

probability $1 - 12(\log n)^{-1}$ on which Proposition 2.3 holds. Then, from Proposition 2.3, we have $\varphi^*_{b;n,m}|_{\mathcal{B}^d_{0,\kappa}}$ is Lipschitz continuous with parameter M_{j_n} and upper bound U_{j_n}, for M_j and U_j are defined in (2.45). Specifically, since $j_n > J_n$, from (2.45), there exists $C_5 > 0$, depending only on $d, \alpha, \beta, \sigma_1, \sigma_2, \gamma_2$, such that $M_{j_n} \vee U_{j_n} \leq C_5(\log m)^{\frac{5}{2\wedge\alpha\wedge\beta}}$. Whence, observing that

$$\frac{\varphi^*_{b;n,m}\big(\kappa T_b(\cdot,\cdot)\big)\,\mathbb{1}_{\mathcal{B}_0^{d+p}}(\cdot,\cdot)}{C_5(\log m)^{\frac{5}{2\wedge\alpha\wedge\beta}}} \in \mathcal{F},$$

we deduce that

$$\begin{aligned}
&\int_{E_b} \frac{\varphi^*_{b;n,m}(v)}{C_5\kappa(\log m)^{\frac{5}{2\wedge\alpha\wedge\beta}}}\, d\big(P^{A(b)} - P_n^{A(b)}\big)(v) \\
&\qquad = \int_{\mathcal{B}^{d+p}_{0,\kappa}} \frac{\varphi^*_{b;n,m}(T_b(s,e))}{C_5\kappa(\log m)^{\frac{5}{2\wedge\alpha\wedge\beta}}}\, d(P^\varepsilon\otimes P^S - P_n^\varepsilon\otimes P_n^S)(e,s) \\
&\qquad = \int_{\mathcal{B}^{d+p}_{0,1}} \frac{\varphi^*_{b;n,m}(\kappa T_b(s,e))}{C_5(\log m)^{\frac{5}{2\wedge\alpha\wedge\beta}}}\, d(P^\varepsilon\otimes P^S - P_n^\varepsilon\otimes P_n^S)(e,s) \\
&\qquad \leq \sup_{f\in\mathcal{F}}\Big\{\int f(s,e)\, d(P^\varepsilon\otimes P^S - P_n^\varepsilon\otimes P_n^S)(e,s)\Big\}. \qquad (2.62)
\end{aligned}$$

By Lemma 2.9 and Wainwright [Wai19, Theorem 4.10], there exists an event Ω_2 with probability at least $1 - n^{-1}$, on which for some constant $C' > 0$, depending only on d, we have

$$\sup_{f\in\mathcal{F}}\Big|\int f\, d(P^\varepsilon\otimes P^S - P_n^\varepsilon\otimes P_n^S)\Big| \leq 2C'\Big(\vartheta_n + \sqrt{\frac{p}{n}}\Big) + \sqrt{\frac{2\log n}{n}}. \qquad (2.63)$$

Combining (2.60), (2.62) and (2.63), we have on event $\Upsilon\cap\Omega_1\cap\Omega_2$ that

$$\begin{aligned}
\int \varphi^*_{b;n,m}(v)\, d\big(P^{A(b)} - P_n^{A(b)}\big)(v) &\leq C_5\kappa(\log m)^{\frac{5}{2\wedge\alpha\wedge\beta}}\bigg(2C'\Big(\vartheta_n + \sqrt{\frac{p}{n}}\Big) + \sqrt{\frac{2\log n}{n}}\bigg) + \frac{C_4}{n} \\
&\leq C_5'(\log m)^{\frac{6}{2\wedge\alpha\wedge\beta}}\bigg(\vartheta_n + \sqrt{\frac{p}{n}} + \sqrt{\frac{2\log n}{n}}\bigg),
\end{aligned}$$

for some $C_5' > 0$ depending only on $d, \alpha, \beta, \sigma_1, \sigma_2, \gamma_2$. A symmetric argument shows that on $\Upsilon\cap\Omega_1\cap\Omega_2$, $\int -\varphi^*_{b;n,m} d(P^{A(b)} - P_n^{A(b)})$ can be controlled by the same upper bound. This establishes the first claim of the lemma.

A similar argument is applied to obtain the bound for the empirical process of $\varphi_{b;n,m}$. Let $\gamma = 2\sqrt{2d\log m}$, and define $\Omega_3 := \{\max_{1\leq i\leq m}\|U_i\| \leq \gamma\}$. Then by a union bound we have $\mathbb{P}(\Omega_3^c) \leq m\,\mathbb{P}\big(\|U_1\| \geq \gamma\big) \leq 2m\exp\Big(-\frac{1}{2}\frac{\gamma^2}{2d}\Big) \leq \frac{2}{m}$. Working on Ω_3 we deduce that for some absolute constant $C_6 > 0$,

$$\begin{aligned}
\int_{\mathbb{R}^d\setminus\mathcal{B}^d_{0,\gamma}} \varphi_{b;n,m}(u)\, d(P^U - P_m^U)(u) &= \int_{\mathbb{R}^d\setminus\mathcal{B}^d_{0,\gamma}} \varphi_{b;n,m}(u)\, dP^U(u) \\
&\overset{(c)}{\leq} \int_{\mathbb{R}^d\setminus\mathcal{B}^d_{0,\gamma}} \frac{\|u\|^2}{2}\, dP^U(u) \overset{(d)}{\leq} \frac{C_6}{m^2}. \qquad (2.64)
\end{aligned}$$

In the above, we use part (iii) in the Proposition 2.2 to obtain (c) and Lemma 2.14 in inequality (d).

Define

$$\mathcal{H} = \{g\,\mathbb{1}_{\mathcal{B}_0^d} : g \in \mathrm{Lip}_{1,1}(\mathcal{B}_0^d)\}. \tag{2.65}$$

Let $i_m := (I_m + 1) + \lceil \frac{1}{2}\log_3(8d\log m)\rceil$. Observe that $3^{i_m}\sqrt{d} \geq \gamma$ thus we have $\mathcal{B}_{0,\gamma}^d \subset \bigcup_{i=0}^{i_m}\bigcup_{k=1}^{N} P_{i,k}$. Since $\varphi_{b;n,m} \in \mathcal{C}_{R,T}$ on Υ according to Proposition 2.3, we have that $\varphi_{b;n,m}|_{\mathcal{B}_{0,\gamma}^d}$ is bounded and Lipschitz continuous with upper bound T_{i_m} and Lipshictz constant R_{i_m} as defined in (2.46). Moreover, by the explicit display of (2.46), there exists a constant C_7 depends on $d, \sigma_1, \sigma_2, \alpha, \beta, \gamma_2$ such that $R_{i_m} \vee T_{i_m} \leq C_7(\log m)^{\frac{5}{2\wedge\alpha\wedge\beta}}$. Therefore, on Υ, we have

$$\frac{\varphi_{b;n,m}(\langle\gamma,\cdot\rangle)}{C_7(\log m)^{\frac{5}{2\wedge\alpha\wedge\beta}}}\,\mathbb{1}_{\mathcal{B}_{0,1}^d}(\cdot) \in \mathcal{H},$$

and consequently,

$$\frac{1}{C_7\gamma(\log m)^{\frac{5}{2\wedge\alpha\wedge\beta}}}\int_{\mathcal{B}_{0,\gamma}^d}\varphi_{b;n,m}(u)\,d(P^U - P_m^U)(u) \leq \sup_{h\in\mathcal{H}}\Big\{\int h(u)\,d(P^U - P_m^U)(u)\Big\}. \tag{2.66}$$

Then applying Lemma 2.9 and Wainwright [Wai19, Theorem 4.10], we derive that there exists an event Ω_4 with probability at least $1 - m^{-1}$ such that on this event we have

$$\sup_{h\in\mathcal{H}}\left|\int h\,d(P^U - P_m^U)\right| \leq 2\vartheta_m + \sqrt{\frac{2\log m}{m}}. \tag{2.67}$$

Consequently, combining (2.64), (2.66) and (2.67), and working on the event $\Upsilon \cap \Omega_3 \cap \Omega_4$, we obtain

$$\begin{aligned}\int \varphi_{b;n,m}(u)\,d(P^U - P_m^U) &\leq C_7(\log m)^{\frac{5}{2\wedge\alpha\wedge\beta}}\gamma\Big(2\vartheta_m + \sqrt{\frac{2\log m}{m}}\Big) + \frac{C_6}{m},\\ &\leq C_7'(\log m)^{\frac{6}{2\wedge\alpha\wedge\beta}}\Big(\vartheta_m + \sqrt{\frac{2\log m}{m}}\Big).\end{aligned}$$

for some C_7' depends on $d, \sigma_1, \sigma_2, \alpha, \beta, \gamma_2$. A symmetric argument can be applied to establish the upper bound for $\int -\varphi_{b;n,m}(u)\,d(P^U - P_m^U)$ and the second claim follows. Finally, the proof is complete by observing that $\mathbb{P}(\Upsilon \cap \Omega_1 \cap \Omega_2 \cap \Omega_3 \cap \Omega_4) \geq 1 - 18(\log n)^{-1}$. □

Lemma 2.9. *Suppose that $\mathcal{T}$ be a subset of linear maps from $\mathbb{R}^p$ to $\mathbb{R}^d$ whose operator norms are bounded by 1 and let $\mathcal{L}$ be a subset of $\{g : g \in \mathrm{Lip}_{1,1}(\mathcal{B}_0^d), g(0) = 0$ and g is convex$\}$. Define $\mathcal{F} := \{(g\circ h)\mathbb{1}_{\mathcal{B}_0^p} : h \in \mathcal{T}, g \in \mathcal{L}\}$. Let $P \in \mathcal{P}_2(\mathbb{R}^p)$. Then exists $C > 0$, depending only on d, such that*

$$\mathcal{R}_n(\mathcal{F}, P) \leq C\left(\vartheta_n + \sqrt{\frac{p}{n}}\right),$$

where

$$\vartheta_k := \begin{cases} k^{-2/d}, & \text{if } d \geq 5,\\ k^{-1/2}\log k, & \text{if } d = 4,\\ k^{-1/2}, & \text{if } d \leq 3.\end{cases} \tag{2.68}$$

for $k \in \mathbb{N}$.

Proof. For any fixed $\delta \in (0,1)$, let $\mathcal{G}$ be a δ-covering set of $\mathcal{L}$ with respect to $\|\cdot\|_{L^\infty(\mathcal{B}_0^d)}$. By Bronshtein [Bro76, Remark 1 and Theorem 6], we have $N_0 := |\mathcal{G}| \le e^{C_8(4/\delta)^{d/2}}$ for some $C_8 > 0$, depending only on d. Similarly, let $\mathcal{H}$ be a δ-covering set of $\mathcal{T}$ with respect to $\|\cdot\|_{\mathrm{op}}$. By Wainwright [Wai19, Lemma 5.7], we have $N_1 := |\mathcal{H}| \le (1+2/\delta)^{dp}$. Now, given any $f = g \circ h \in \mathcal{F}$, we can find $g' \in \mathcal{G}$ and $h' \in \mathcal{H}$ such that $\|g'-g\|_{L^\infty(\mathcal{B}_0^d)} \le \delta$ and $\|h'-h\|_{\mathrm{op}} \le \delta$. Consequently, for $X \sim P$, we have

$$\begin{aligned}\|(g\circ h - g'\circ h')\,\mathbb{1}_{\mathcal{B}_0^p}\|_{L^2(P)} &\le \|(g\circ h - g'\circ h)\,\mathbb{1}_{\mathcal{B}_0^p}\|_{L^2(P)} + \|(g'\circ h - g'\circ h')\,\mathbb{1}_{\mathcal{B}_0^p}\|_{L^2(P)} \\ &= \left\{\mathbb{E}\left|(g-g')\circ h(X)\,\mathbb{1}_{\{\|X\|\le 1\}}\right|^2\right\}^{1/2} + \left\{\mathbb{E}\left|g'\circ(h-h')(X)\,\mathbb{1}_{\{\|X\|\le 1\}}\right|^2\right\}^{1/2} \le 2\delta.\end{aligned}$$

which implies

$$\log N(2\delta, \mathcal{F}, \|\cdot\|_{L^2(P)}) \le \log(N_0 N_1) \le C_8\left(\frac{4}{\delta}\right)^{d/2} + dp\log\left(1+\frac{2}{\delta}\right) \le C_8\left(\frac{4}{\delta}\right)^{d/2} + \frac{2dp}{\delta}. \tag{2.69}$$

Since all functions in $\mathcal{F}$ are uniformly bounded by 1, the $L_2(P)$-diameter of $\mathcal{F}$ is bounded by 2. Thus, by Dudley's chaining [see e.g. Wai19, Theorem 5.22], for any $\epsilon \in [0,1]$, we have

$$\begin{aligned}\mathcal{R}_n(\mathcal{F}, P) &\le 2\epsilon + \frac{32}{\sqrt{n}}\,\mathbb{E}\int_{\epsilon/4}^2 \log^{1/2} N(\delta, \mathcal{F}, \|\cdot\|_{L^2(P)})\,d\delta \\ &\le 2\epsilon + \frac{2^{5+3d/4} C_8^{1/2}}{n^{1/2}}\int_{\epsilon/4}^2 \frac{1}{\delta^{d/4}}\,d\delta + \frac{64(dp)^{1/2}}{n^{1/2}}\int_{\epsilon/4}^2 \frac{1}{\delta^{1/2}}\,d\delta.\end{aligned}$$

By choosing $\epsilon \asymp n^{-2/d}$ if $d \ge 4$ and $\epsilon = 0$ otherwise, we deduce from the previous inequality that there exists $C > 0$ depending only on d such that

$$\mathcal{R}_n(\mathcal{F}, P) \le C\begin{cases} n^{-2/d} + (p/n)^{1/2}, & \text{if } d \ge 5 \\ n^{-1/2}(\log n + p^{1/2}), & \text{if } d = 4 \\ (p/n)^{1/2}, & \text{if } d \le 3, \end{cases},$$

completing the proof. □

Lemma 2.10. *There exists $C > 0$ depending only on $d, \alpha, \beta, \sigma_1, \sigma_2, \gamma_1, \gamma_2$, such that with probability at least $1 - 6/n$, both of the following inequalities hold:*

$$\begin{aligned}\sup_{b\in\mathcal{B}}\left|\int \psi_b^*(v)\,d(P^{A(b)} - P_n^{A(b)})(v)\right| &\le C(\log m)^{\frac{2}{2\wedge\alpha\wedge\beta}} n^{-1/2} \\ \sup_{b\in\mathcal{B}}\left|\int \psi_b(u)\,d(P^U - P_m^U)(u)\right| &\le C(\log m)^{\frac{2}{2\wedge\alpha\wedge\beta}} m^{-1/2}.\end{aligned}$$

Proof. Since adding a constant to ψ_b^* or ψ_b will not change the value of $\int \psi_b^*(v)\,d(P^{A(b)} - P_n^{A(b)})(v)$ or $\int \psi_b(u)\,d(P^U - P_m^U)(u)$, we assume $\psi_b^*(0) = \psi_b(0) = 0$ with out loss of generality. We first

note that $\mathbb{E}\,\|(b^*-b)\Sigma^{1/2}S\|^2 = \|b^*-b\|_\Sigma^2 \le 1$, for any $b \in \mathcal{B}$. By Lemma 2.15 and the anti-concentration inequality of ε given in (2.12), there exists a constant $M_1 > 0$ depends on γ_1 and γ_2 such that the density function of $A(b)$, write as $f_{A(b)}$, have the anti-concentration inequality

$$f_{A(b)}(v) \ge M_1 \exp\bigl(-2\gamma_2\|v\|^2\bigr), \quad \text{for all } \|v\| \ge 2.$$

Then by recalling that $P^U \sim (\sqrt{2d}, 2)$-sub-Weibull, we apply Manole and Niles-Weed [MN24, Theorem 11][3] to obtain that $\|\nabla\psi_b^*(v)\| \le C(\|v\|+1)$ for all $v \in \mathbb{R}^d$, where $C > 0$ is a constant depending on d, γ_1, γ_2. Therefore, applying mean value theorem, we have $|\psi_b^*(v)| \le C(\|v\|+1)^2$ for all $v \in \mathbb{R}^d$.

Define $\Omega_1 := \{\max_{1\le i\le n}\|(\varepsilon_i, S_i)\| \le \kappa\}$ and $E_b := \{T_b(s,e) : (s,e) \in \mathcal{B}_{0,\kappa}^{d+p}\}$ for each fixed $b \in \mathcal{B}$. From the proof of Lemma 2.8, we have $\mathbb{P}(\Omega_1) \ge 1 - 2/n$. Then on Ω_1 we can obtain that

$$\begin{aligned}
\sup_{b\in\mathcal{B}}\Bigl|\int_{\mathbb{R}^d\setminus E_b}\psi_b^*(v)\,d(P^{A(b)}-P_n^{A(b)})(v)\Bigr| &= \sup_{b\in\mathcal{B}}\Bigl|\int_{\mathbb{R}^{d+p}\setminus\mathcal{B}_{0,\kappa}^{d+p}}\psi_b^*\circ T_b(s,e)\,d(P^S\otimes P^\varepsilon)(s,e)\Bigr| \\
&\le C\sup_{b\in\mathcal{B}}\int_{\mathbb{R}^{d+p}\setminus\mathcal{B}_{0,\kappa}^{d+p}}(1+\|T_b(s,e)\|)^2\,d(P^S\otimes P^\varepsilon)(s,e) \\
&\le C\sup_{b\in\mathcal{B}}\int_{\mathbb{R}^{d+p}\setminus\mathcal{B}_{0,\kappa}^{d+p}}(1+\|(s,e)\|)^2\,d(P^S\otimes P^\varepsilon)(s,e) \\
&\le \frac{C'}{n},
\end{aligned}$$

for some constant $C' > 0$ depending on $d, \alpha, \beta, \sigma_1, \sigma_2, \gamma_1, \gamma_2$, where we used the fact that $P^{(S,\varepsilon)} \sim (\rho, \alpha\wedge\beta)$-sub-Weibull and Lemma 2.14 in the final inequality. It therefore remains to control

$$G := \sup_{b\in\mathcal{B}}\Bigl|\int_{E_b}\psi_b^*(v)\,d(P^{A(b)}-P_n^{A(b)})(v)\Bigr| = \sup_{b\in\mathcal{B}}\Bigl|\int_{\mathcal{B}_{0,\kappa}^{d+p}}\psi_b^*\circ T_b(s,e)\,d(P^S\otimes P^\varepsilon - P_n^S\otimes P_n^\varepsilon)(s,e)\Bigr|.$$

To simplify the notation, define the centered function

$$\bar\psi_b^*(s,e) := \psi_b^*\circ T_b(s,e)\,\mathbb{1}\{\|(s,e)\| \le \kappa\} - \mathbb{E}[\psi_b^*\circ T_b(S,\varepsilon)\,\mathbb{1}\{\|(S,\varepsilon)\| \le \kappa\}],$$

then it follows that $\|\bar\psi_b^*\|_\infty \le 2C(\kappa+1)^2 \le C(\log m)^{\frac{1}{2\wedge\alpha\wedge\beta}}$. In this notation, we have $G = \sup_{b\in\mathcal{B}}|n^{-1}\sum_{i\in[n]}\bar\psi_b^*(S_i,\varepsilon_i)|$. By Markov's inequality, we then have

$$\mathbb{E}(G) = \int_0^{+\infty}\mathbb{P}(G\ge t)\,dt \le n^{-1/2} + C\int_{n^{-1/2}}^{+\infty}\frac{(\log m)^{\frac{2}{2\wedge\alpha\wedge\beta}}}{nt^2}dt \lesssim \frac{(\log m)^{\frac{2}{2\wedge\alpha\wedge\beta}}}{\sqrt{n}}. \tag{2.70}$$

We now claim that G, when viewed as a function of $(s_1,e_1),\dots,(s_n,e_n)$, satisfies the bounded difference property [see e.g. Wai19, (2.32)]. By symmetry, it suffices to consider a perturbation on (s_1,e_1). Define $v = (v_i)_{i=1}^n, v' = (v_i')_{i=1}^n$ where each $v_i = (s_i,e_i), v_i' = (s_i',e_i') \in \mathbb{R}^{d+p}$, such

[3] In the original Theorem 11 of [MN24], a regular condition is required on the density function of the source probability measure. Nevertheless, it is indeed sufficient to reestablish the result by merely assuming an anti-concentration inequality on the density function of the source probability measure, as we have proven for $f_{A(b)}$ here.

that $v_i = v_i'$ for any $i \neq 1$. We have

$$\sup_{b\in\mathcal{B}}\left|\frac{1}{n}\sum_{i=1}^n \bar{\psi}_b^*(v_i)\right| - \sup_{b\in\mathcal{B}}\left|\frac{1}{n}\sum_{i=1}^n \bar{\psi}_b^*(v_i')\right| \leq \sup_{b\in\mathcal{B}}\left\{\left|\frac{1}{n}\sum_{i=1}^n \bar{\psi}_b^*(v_i)\right| - \left|\frac{1}{n}\sum_{i=1}^n \bar{\psi}_b^*(v_i')\right|\right\}$$
$$\leq \frac{1}{n}\sup_{b\in\mathcal{B}}\left|\bar{\psi}_b^*(v_1) - \bar{\psi}_b^*(v_1')\right| \leq \frac{2C(\log m)^{\frac{1}{2\wedge\alpha\wedge\beta}}}{n},$$

estbalishing, the bounded difference property for G. Thus by McDiarmid's inequality [see e.g. Wai19, Corollary 2.21], we obtain that the event

$$\Lambda_1 := \left\{G \leq \mathbb{E}G + \frac{\sqrt{2}C(\log m)^{\frac{2}{2\wedge\alpha\wedge\beta}}}{\sqrt{n}}\right\},$$

occurs with probability at least $1 - 1/m$.

Thus, working on the event $\Omega_1 \cap \Lambda_1$, we deduce from (2.70) that

$$\sup_{b\in\mathcal{B}}\left|\int \psi_b^*(v)d(P^{A(b)} - P_n^{A(b)})(v)\right| \leq \frac{C(\log m)^{\frac{2}{2\wedge\alpha\wedge\beta}}}{\sqrt{n}} + \frac{C'}{n} \leq \frac{C(\log m)^{\frac{2}{2\wedge\alpha\wedge\beta}}}{\sqrt{n}},$$

for some constant $C > 1$ depends on $d, \alpha, \beta, \gamma_1, \gamma_2, \sigma_1, \sigma_2$, which completes the first claim of the lemma.

For the second claim, in order to bound $\int \psi_b(u)\, d(P^U - P_m^U)(u)$, we notice that the anti-concentration property of P^U holds due to the Gaussian assumption. Thus Manole and Niles-Weed [MN24, Theorem 11] implies that $\|\nabla\psi_b(u)\| \leq \tilde{C}(1 + \|u\|)^{\frac{2}{\alpha\wedge\beta}}$ for some $\tilde{C} > 0$ depending on $d, \sigma_1, \sigma_2, \alpha, \beta$, and it follows that $|\psi_b(u)| \leq \tilde{C}(1 + \|u\|)^{\frac{2}{\alpha\wedge\beta}+1}$.

Define $\Omega_2 := \{\max_{1\leq i\leq m}\|U_i\| \leq \gamma\}$. From the proof of Lemma 2.8 again, we have $\mathbb{P}(\Omega_2) \geq 1 - 2/n$. Working on Ω_2, we have

$$\sup_{b\in\mathcal{B}}\left|\int_{\mathbb{R}^d\setminus\mathcal{B}_{0,\gamma}} \psi_b(u)d(P^U - P_m^U)(u)\right| = \sup_{b\in\mathcal{B}}\left|\int_{\mathbb{R}^d\setminus\mathcal{B}_{0,\gamma}} \psi_b(u)dP^U(u)\right|$$
$$\leq \tilde{C}\int_{\mathbb{R}^d\setminus\mathcal{B}_{0,\gamma}} (1 + \|u\|)^{\frac{2}{\alpha\wedge\beta}+1}dP^U(u) \leq \frac{\tilde{C}}{m},$$

for some constant $\tilde{C} > 0$ depending on $d, \alpha, \beta, \sigma_1, \sigma_2$. Now, defining $\tilde{G} := \sup_{b\in\mathcal{B}}\left|\int_{\mathcal{B}_{0,\gamma}} \psi_b d(P^U - P_m^U)\right|$, by the same argument as in the proof of the first part of this lemma, there is an event Λ_2 with probability at least $1 - m^{-1}$, such that on $\Omega_2 \cap \Lambda_2$, we have

$$\sup_{b\in\mathcal{B}}\left|\int \psi_b(u)d(P^U - P_m^U)(u)\right| \leq \tilde{G} + \frac{\tilde{C}'}{m} \leq \mathbb{E}\tilde{G} + \frac{\bar{C}(\log m)^{\frac{2}{2\wedge\alpha\wedge\beta}}}{\sqrt{m}} + \frac{\tilde{C}}{m} \leq \frac{\bar{C}(\log m)^{\frac{2}{2\wedge\alpha\wedge\beta}}}{\sqrt{m}}, \tag{2.71}$$

for $\bar{C} > 0$ depending only on $d, \alpha, \beta, \sigma_1, \sigma_2, \gamma_2$. □

Lemma 2.11. *There exists $C > 0$ depending only on $d, \alpha, \beta, \sigma_1, \sigma_2, \gamma_1, \gamma_2$, such that with probability at least $1 - 5/n$, we have*

$$\sup_{b\in\mathcal{B}}\left|\int \|v\|^2 \, d\big(P_n^{A(b)} - P^{A(b)}\big)(v)\right| \le C(\log m)^{\frac{2}{2\wedge\alpha\wedge\beta}} n^{-1/2},$$
$$\left|\int \|u\|^2 \, d\big(P^U - P_m^U\big)(u)\right| \le C\sqrt{\frac{\log m}{m}}.$$

Proof. Observe that the only property of ψ_b^* that we used in the first part of the proof of Lemma 2.10 is that $\|\nabla\psi_b^*(v)\| \le C(\|v\| + 1)$ for all $b \in \mathcal{B}$ and $v \in \mathbb{R}^d$. The same property is satisfied by the function $v \mapsto \|v\|^2$. Hence, a very similar proof to that of Lemma 2.10 will establish the first claim here.

As for the second inequality, since $U_i \overset{i.i.d.}{\sim} \mathcal{N}(0, I_d)$, we have $\sum_{i=1}^m \|U_i\|^2 \sim \chi^2_{md}$. By Laurent and Massart [LM00, Lemma 1] we deduce that

$$\mathbb{P}\left(\left|\frac{1}{m}\sum_{i=1}^m \|U_i\|^2 - \mathbb{E}\,\|U\|^2\right| \ge \sqrt{\frac{2d\log m}{m}} + \frac{2\log m}{m}\right) \le \frac{2}{m},$$

which implies the second claim. □

Proof of Theorem 2.2. Recalling that in the regime of (2.6) event Θ holds with probability at least $1 - 4(\log n)^{-1}$, and working on Θ we have $\hat{b} \in \mathcal{B}$. Thus there exists $M > 0$ depending only on d, α, β, σ_1, σ_2, γ_1, γ_2 such that with probability at least $1 - 33(\log n)^{-1}$, we have

$$\begin{aligned}
\mathcal{L}(\hat{b}) - \mathcal{L}(b^*) &\le 2\sup_{b\in\mathcal{B}}|\mathcal{L}(b) - \mathcal{L}_{n,m}(b)| \\
&\le \left|\frac{1}{m}\sum_{i=1}^m \|U_i\|^2 - \mathbb{E}\,\|U\|^2\right| + \sup_{b\in\mathcal{B}}\left|\frac{1}{n}\sum_{i=1}^n \|T_b(S_i,\varepsilon_i)\|^2 - \mathbb{E}\,\|T_b(S_i,\varepsilon_i)\|^2\right| \\
&\qquad + \sup_{b\in\mathcal{B}}\left|\mathcal{W}_2^2(P^{A(b)}, P^U) - \mathcal{W}_2^2(P_n^{A(b)}, P_m^U)\right| \\
&\le M(\log m)^{\frac{8}{2\wedge\alpha\wedge\beta}}\left(\sqrt{\frac{p}{n}} + \frac{1}{n^{2/d}}\right), \qquad (2.72)
\end{aligned}$$

where the second inequality uses the definition of $\langle\!\langle\cdot,\cdot\rangle\!\rangle_{\mathcal{W}_2}$ and in the final inequality, we used Lemma 2.11 to control the first two terms and Proposition 2.4 for the last term.

On the other hand, by the lower bound developed in (2.31) and Lemma 2.12 we have for $r := \langle\!\langle P^\varepsilon, P^U\rangle\!\rangle_{\mathcal{W}_2}$ that

$$\mathcal{L}(\hat{b}) - \mathcal{L}(b^*) \ge \sqrt{r^2 + \|b^* - \hat{b}\|_\Sigma^2} - r \ge \frac{1}{2}(1 + r^2)^{-1/2}\|b^* - \hat{b}\|_\Sigma^2. \qquad (2.73)$$

Combining (2.72) with (2.73), we obtain that

$$\|b^* - \hat{b}\|_\Sigma \le M(\log m)^{\frac{4}{2\wedge\alpha\wedge\beta}}\left\{\left(\frac{p}{n}\right)^{1/4} + \frac{1}{n^{1/d}}\right\}, \qquad (2.74)$$

with probability at least $1 - 33(\log n)^{-1}$. Here we close the proof. □

2.6 Ancillary results

Lemma 2.12. *For any $a \geq 0$, we have inequality*

$$\sqrt{a+x^2} \leq \begin{cases} \frac{x^2}{2\sqrt{a}} + \sqrt{a} & , \text{if } 0 \leq x \leq 1, \\ (x-1) + \frac{1}{2\sqrt{a}} + \sqrt{a} & , \text{if } x > 1. \end{cases},$$

and

$$\sqrt{a+x^2} \geq \begin{cases} \frac{x^2}{2\sqrt{a+1}} + \sqrt{a} & , \text{if } 0 \leq x \leq 1, \\ \frac{x-1}{\sqrt{a+1}} + \frac{1}{2\sqrt{a+1}} + \sqrt{a} & , \text{if } x > 1. \end{cases}$$

Proof. Write

$$\sqrt{a+x^2} = \int_0^x \frac{t}{\sqrt{a+t^2}} dt + \sqrt{a}.$$

Thus the first inequality can be obtained by utilizing $t/\sqrt{a+t^2} \leq t/\sqrt{a}$ and $t/\sqrt{a+t^2} \leq 1$ in the case of $0 \leq t \leq 1$ and $t \geq 1$ respectively. The second inequality follows by noting that $t/\sqrt{a+t^2} \geq t/\sqrt{a+1}$ when $0 \leq t \leq 1$ and $t/\sqrt{a+t^2} \geq 1/\sqrt{a+1}$ when $t \geq 1$. □

Lemma 2.13. *There exist independent random vectors Z and ε such that $P^Z, P^\varepsilon \in \mathcal{P}_2(\mathbb{R}^d) \cap \mathcal{P}_{\mathrm{ac}}(\mathbb{R}^d)$ such that $\langle\!\langle Z+\varepsilon, U \rangle\!\rangle_{\mathcal{W}_2}^2 = \langle\!\langle Z, U \rangle\!\rangle_{\mathcal{W}_2}^2 + \langle\!\langle \varepsilon, U \rangle\!\rangle_{\mathcal{W}_2}^2$.*

Proof. Consider independent random vectors $Z \sim \mathcal{N}(0, \Sigma)$ and $\varepsilon \sim \mathcal{N}(0, \Gamma)$. By the same argument as in (2.30), we have

$$\begin{aligned} \langle\!\langle Z+\varepsilon, U \rangle\!\rangle_{\mathcal{W}_2} &= \mathrm{Tr}\big((\Sigma+\Gamma)^{1/2}\big) \\ \langle\!\langle Z, U \rangle\!\rangle_{\mathcal{W}_2} &= \mathrm{Tr}\big(\Sigma^{1/2}\big) \\ \langle\!\langle \varepsilon, U \rangle\!\rangle_{\mathcal{W}_2} &= \mathrm{Tr}\big(\Gamma^{1/2}\big). \end{aligned}$$

Hence, the desired result hold if we take $\Sigma = \sigma^2 I_d$ and $\Gamma = \gamma^2 I_d$. □

Proposition 2.5. *Let X be a random vecotor. Then the following properties are equivalent:*

(i) There exists $\sigma > 0$ such that $\mathbb{P}(\|X\| \geq x) \leq 2e^{-\frac{1}{2}(x/\sigma)^\beta}$ for all $x \geq 0$.

(ii) There exists $K_\sigma > 0$ such that $\{\mathbb{E}\|X\|^k\}^{1/k} \leq K_\sigma k^{1/\beta}$.

(iii) There exists $K'_\sigma > 0$ such that $\mathbb{E}\exp\big\{\big((\lambda\|X\|)^\beta\big)\big\} \leq \exp\big\{\big((\lambda K'_\sigma)^\beta\big)\big\}$ for all $|\lambda| \leq 1/K'_\sigma$.

(iv) X follows the (σ, β)-sub-weibull distribution.

The proof follows by Vladimirova et al. [Vla+20, Theorem 2.1].

Proposition 2.6. *For $p_1, p_2 \in \mathbb{N}$, let $X \in \mathbb{R}^{p_1}, Y \in \mathbb{R}^{p_2}$ be two independent sub-Weibull random vectors with parameter (σ_1, α) and (σ_2, β) respectively. Then the following statements holds:*

(i) For matrices $A \in \mathbb{R}^{d\times p_1}$ and $B \in \mathbb{R}^{d\times p_2}$, there exists $\sigma > 0$ depending only on σ_1, σ_2, $\|A\|_{\mathrm{op}}$, $\|B\|_{\mathrm{op}}$ such that $AX + BY \sim (\sigma, \alpha \wedge \beta)$-sub-Weibull.

(ii) There exists $\sigma > 0$ depending only on σ_1, σ_2 such that the concatenation of two random vectors $Z := (X, Y) \in \mathbb{R}^{p_1+p_2}$ is a sub-Weibull random vector with parameter $(\sigma, \alpha \wedge \beta)$.

Proof. (i) Suppose K_{σ_1} and K_{σ_2} are the induced constants of X and Y by the part (ii) of Proposition 2.5. Then it follows that

$$\begin{aligned}\big(\mathbb{E}\,\|AX + BY\|^k\big)^{1/k} &\le (\mathbb{E}\,\|AX\|^k)^{1/k} + (\mathbb{E}\,\|BY\|^k)^{1/k} \\ &\le \|A\|_{\mathrm{op}}(\mathbb{E}\,\|X\|^k)^{1/k} + \|B\|_{\mathrm{op}}(\mathbb{E}\,\|Y\|^k)^{1/k} \\ &\le \|A\|_{\mathrm{op}} \vee \|B\|_{\mathrm{op}} \cdot \big(K_{\sigma_1} k^{1/\alpha} + K_{\sigma_2} k^{1/\beta}\big) \\ &\le 2(\|A\|_{\mathrm{op}} \vee \|B\|_{\mathrm{op}}) \cdot (K_{\sigma_1} \vee K_{\sigma_2}) k^{1/(\alpha\wedge\beta)}.\end{aligned}$$

This proves that $AX + BY$ satisfies part (ii) in the Proposition 2.5 thus the conclusion follows by the equivalence of part (ii) and (iv).

(ii) For any integer $k \ge 1$, we have

$$\begin{aligned}\big(\mathbb{E}\,\|(X, Y)\|^k\big)^{1/k} &\le \big(\mathbb{E}(\|X\| + \|Y\|)^k\big)^{1/k} \\ &\le \big(\mathbb{E}\,\|X\|^k\big)^{1/k} + \big(\mathbb{E}\,\|Y\|^k\big)^{1/k} \le (K_{\sigma_1} \vee K_{\sigma_2}) k^{1/(\alpha\wedge\beta)},\end{aligned}$$

where the sub-Weibull assumption on X and Y have been exploited. The conclusion follows by employing Proposition 2.5. □

Lemma 2.14. *If X is a (σ, β)-sub-Weibull random vector as defined in (2.11), then for any $s > 0$, there exists $C > 0$, depending on s, σ, β, such that $\mathbb{E}\big(\|X\|^s \mathbb{1}\{\|X\| \ge t\}\big) \le Ce^{-\frac{1}{4}(t/\sigma)^\beta}$.*

Proof. We have

$$\begin{aligned}\mathbb{E}\big(\|X\|^s \mathbb{1}\{\|X\| \ge t\}\big) &= \mathbb{E}\Big[\|X\|^s \mathbb{1}\Big\{e^{\frac{1}{4}(\|X\|/\sigma)^\beta} \ge e^{\frac{1}{4}(t/\sigma)^\beta}\Big\}\Big] \\ &\le \mathbb{E}\big\{\|X\|^s e^{\frac{1}{4}(\|X\|/\sigma)^\beta} e^{-\frac{1}{4}(t/\sigma)^\beta}\big\} \\ &\le e^{-\frac{1}{4}(t/\sigma)^\beta}\big\{\mathbb{E}\,\|X\|^{2s}\big\}^{1/2}\Big\{\mathbb{E}\,e^{\frac{1}{2}(\|X\|/\sigma)^\beta}\Big\}^{1/2} \\ &\le 2^{1/2} e^{-\frac{1}{4}(t/\sigma)^\beta}\big\{\mathbb{E}\,\|X\|^{2s}\big\}^{1/2},\end{aligned}$$

where we used the definition of X being (σ, β)-sub-Weibull in final step. The desired bound follows since by Proposition 2.5, we have $\mathbb{E}\|X\|^{2s} \le C$ for some constant C that depends on s, σ, β. □

Lemma 2.15. *Suppose X, Y are independent d-dimensional random vectors with finite second moment. If X follows an absolutely continuous distribution with density function f_X which admits the following anti-concentration inequality for some constant $\gamma_1, \gamma_2 > 0$:*

$$f_X(x) \ge \gamma_1 \exp\big(-\gamma_2 \|x\|^2\big), \quad \forall\, \|x\| \ge \mathbb{E}\,\|Y\|^2.$$

Then there exists a constant K_1 depends on γ_1 and γ_2 such that the density function of $V := X + Y$, write as f_V, satisfying

$$f_V(v) \ge K_1 \exp\big\{(-2\gamma_2 \|v\|^2)\big\}, \quad \forall\, \|v\| \ge 2\,\mathbb{E}\,\|Y\|^2.$$

Proof. Write $M_2 := \mathbb{E}\,\|Y\|^2 < +\infty$. For all $\|v\| \geq 2M_2$, we have

$$\begin{aligned} f_V(v) &= \int f_X(v-y) f_Y(y) dy \geq \int_{\|y\| \leq M_2} \gamma_1 \exp\big\{\big(-\gamma_2 \|v-y\|^2\big)\big\} f_Y(y) dy \\ &\geq \int_{\|y\| \leq M_2} \gamma_1' \exp\big\{\big(-2\gamma_2 \|v\|^2\big)\big\} f_Y(y) dy \geq \gamma_1' \big(1 - \frac{1}{M_2}\big) \exp\big\{\big(-2\gamma_2 \|v\|^2\big)\big\}, \end{aligned}$$

where $\gamma_1' = \gamma_1 \exp(-2\gamma_2 M_2^2)$ and the last inequality is followed by the Markov inequality. Thus the result holds by letting $K_1 = \gamma_1' \big(1 - \frac{1}{M_2}\big)$. □

Lemma 2.16. *Let $\mathcal{X}, \mathcal{Y} \subseteq \mathbb{R}^d$ are Borel sets such that L_2 is bounded on $\mathcal{X} \times \mathcal{Y}$, i.e. $\|L_2\|_\infty := \sup_{(x,y) \in \mathcal{X} \times \mathcal{Y}} L_2(x,y) < +\infty$. Then for any $\mu \in \mathcal{P}_2(\mathcal{X})$ and $\nu \in \mathcal{P}_2(\mathcal{Y})$ we have*

$$\begin{aligned} &\inf\big\{ J_{\mu,\nu}(\varphi,\psi) : (\varphi,\psi) \in \tilde{\Phi} \big\} \\ &\quad = \inf\big\{ J_{\mu,\nu}(\varphi,\psi) : (\varphi,\psi) \in \tilde{\Phi}, -\|L_2\|_\infty \leq \varphi - \|\cdot\|^2/2 \leq 0,\ 0 \leq \psi - \|\cdot\|^2/2 \leq \|L_2\|_\infty \big\}, \end{aligned}$$

where $\tilde{\Phi} := \{(\varphi,\psi) \in L^1(\mathcal{X}) \times L^1(\mathcal{Y}) : \varphi(x) + \psi(y) \geq x^T y,\ \forall (x,y) \in \mathcal{X} \times \mathcal{Y}\}$.

Proof. Note that by the argument same as (2.17) we have

$$\begin{aligned} &\inf\big\{ J_{\mu,\nu}(\varphi,\psi) : (\varphi,\psi) \in \tilde{\Phi} \big\} \\ &\quad = \int_{\mathcal{X}} \frac{\|x\|^2}{2} d\mu(x) + \int_{\mathcal{Y}} \frac{\|y\|^2}{2} d\nu(y) - \sup\big\{ J_{\mu,\nu}(\varphi,\psi) : (\varphi,\psi) \in \Phi_2 \big\}, \end{aligned} \tag{2.75}$$

where $\Phi_2 := \{(\varphi,\psi) \in L^1(\mathcal{X}) \times L^1(\mathcal{Y}) : \varphi(x) + \psi(y) \leq L_2(x,y),\ \forall (x,y) \in \mathbb{R}^d \times \mathbb{R}^d\}$. Note by Villani [Vil21, Remark 1.13], we may restrict the supremum in the right-hand side of (2.75) over some bounded functions:

$$\begin{aligned} &\sup\big\{ J_{\mu,\nu}(\varphi,\psi) : (\varphi,\psi) \in \Phi_2 \big\} \\ &\quad = \sup\big\{ J_{\mu,\nu}(\varphi,\psi) : (\varphi,\psi) \in \Phi_2, 0 \leq \varphi \leq \|L_2\|_\infty,\ -\|L_2\|_\infty \leq \psi \leq 0 \big\}. \end{aligned} \tag{2.76}$$

By Villani [Vil09, Theorem 5.10] we may further impose that φ be c-concave and $\psi = \varphi^c$. Suppose (φ_0, φ_0^c) be a solution to the right-hand side of (2.76). Define $\tilde{\varphi} := \|\cdot\|^2/2 - \varphi_0$, $\tilde{\psi} := \|\cdot\|^2/2 - \varphi_0^c$. Then by (2.75) we have

$$\inf\big\{ J_{\mu,\nu}(\varphi,\psi) : (\varphi,\psi) \in \tilde{\Phi} \big\} = \int_{\mathcal{X}} \tilde{\varphi}(x) d\mu(x) + \int_{\mathcal{Y}} \tilde{\psi}(y) d\nu(y). \tag{2.77}$$

Moreover, note

$$\tilde{\varphi}(x) = \|x\|^2/2 - \varphi_0(x) = \|x\|^2/2 - \inf_{y \in \mathcal{Y}} \{c(x,y) - \varphi_0^c(y)\} = \sup_{y \in \mathcal{Y}} \{x^T y - (\|y\|^2/2 - \varphi_0^c(y))\},$$

which implies that $(\tilde{\varphi}, \tilde{\psi}) \in \tilde{\Phi}$. Combine this with (2.77), we proved that $(\tilde{\varphi}, \tilde{\psi})$ is an optimal solution to the left-hand side of (2.75). Finally, by the boundedness of φ_0 and φ_0^c, we have

$$0 \leq \|x\|^2/2 - \tilde{\varphi}(x) \leq \|L_2\|_\infty \text{ and } -\|L_2\|_\infty \leq \|y\|^2/2 - \tilde{\psi}(y) \leq 0,$$

as desired. □

Theorem 2.5. *[FG15, Theorem 1] Let $X \sim P^X$ be a probability measure on $\mathbb{R}^d$ such that $M_\ell := \mathbb{E}\,\|X\|^\ell < +\infty$ with $\ell \in (2, +\infty)$. If P_n^X is the corresponding empirical distribution, then there exists a constant $C > 0$ depending only on d and ℓ such that for all $n \geq 1$,*

$$\mathbb{E}\left[\mathcal{W}_2^2(P^X, P_n^X)\right] \leq C M_\ell^{2/\ell} \tau_n(d, \ell), \tag{2.78}$$

where

$$\tau_n(d,\ell) := \begin{cases} n^{-\frac{1}{2}} & \text{if } d < 4 \\ n^{-\frac{1}{2}} \log(1+n) & \text{if } d = 4 \\ n^{-\frac{2}{d}} & \text{if } d > 4 \end{cases} + \begin{cases} n^{-\frac{1}{d}} & \text{if } \ell > 4 \\ n^{-\frac{1}{2}} \log(1+n) & \text{if } \ell = 4 \\ n^{\frac{2-\ell}{\ell}} & \text{if } 2 < \ell < 4. \end{cases}$$

2.7 Spatial reference

In this section, we derive the MCQR loss function under reference distribution $U[-1,1]$, which may provide an intuitive example for the verification of Proposition 2.1. In one dimension, the traditional rank and quantile can be understood as a pair of optimal transport maps between the distribution of interest $X \sim P$ and the uniform distribution $U \sim U[0,1]$. When P does not assign mass to sets with Hausdorff dimension 0, the corresponding distribution function F and its inverse map $Q := F^{-1}$ serve as the corresponding optimal transport map. This concept can be generalized to other reference distributions, for instance, $U[-1,1]$. In this case, the spatial distribution function $F_{\text{sp}}(\cdot) := 2F(\cdot) - 1$ takes on the role of F in the previous case. Moreover, the corresponding check function needs to be modified as

$$\rho_\tau^{\text{sp}}(X-\theta) := (1+\tau)(X-\theta) - 2(X-\theta)\,\mathbb{1}\{X-\theta < 0\}, \quad \forall \tau \in [-1,1].$$

Suppose $V \sim U[-1,1]$, then the composite quantile regression optimization becomes

$$\begin{aligned}
\mathbb{E}\int_{-1}^{1} \rho_\tau^{\text{sp}}\big(Y - \beta^\top X - q(\tau)\big) \cdot \frac{1}{2}\, d\tau &= \mathbb{E}\int_{-1}^{1} \big(Y - \beta^\top X - q(\tau)\big)^- d\tau + \int_{-1}^{1}\int_{\tau}^{-1} \frac{1}{2} q(\tau)\, dt d\tau \\
&= \mathbb{E} \max_{t\in[-1,1]} \int_t^1 -\big(Y - bX - q(\tau)\big)\, d\tau + \int_{-1}^{1}\int_t^1 \frac{1}{2} q(\tau)\, d\tau dt \\
&= \mathbb{E} \max_{t\in[-1,1]} \big(-(1-t)(Y-bX) + \phi(t)\big) + \mathbb{E}\,\phi(V) \\
&= \mathbb{E} \max_{t\in[-1,1]} \big(t(Y-bX) + \phi(t)\big) + \mathbb{E}\,\phi(V),
\end{aligned}$$

where $\phi(t) = \int_t^1 q(\tau) d\tau$. Thus, applying the same argument as Lemma 2.2 we can see that the composition quantile regression estimator of b^* is once again

$$b^* = \arg\min \langle\!\langle P^{Y-bX}, P^V \rangle\!\rangle_{\mathcal{W}_2}.$$

This gives some intuition on Proposition 2.1. However, if choose the standard normal distribution as the reference distribution, we may not be able to find a straightforward optimal transport map as F or F_{sp}, but Proposition 2.1 demonstrates the validity of this extension.

2.8 Spatial quantile

The concept of the spatial (or geometric) quantile was initially introduced by [Cha96]. Uniquely characterizing the underlying probability distribution as a special case of M-quantile, as demonstrated in [Kol97, Theorem 2.5], this quantile permits a seamless extension to the regression framework [Cha03] and functional quantile regression [CC14; CC19]. A more recent development involves an extension to the hypersphere, as explored by Konen and Paindaveine [KP23].

The definition of Spatial quantile starts from rewriting the check function $\rho_\tau(\cdot)$ as

$$\rho_\tau(z) = \frac{1}{2}\big(|z| + (2\tau - 1)z\big) = \frac{1}{2}(|z| + vz), \text{ for any } z \in \mathbb{R},$$

with $v = 2\tau - 1$. Thus a natural extension of the check function to the multi-dimensional case is by substituting the absolute value function by the L_1-loss function:

$$\Phi_v(z) := \frac{1}{2}(\|z\| + v^\top z),$$

where $v = \tau u$, and $u \in \mathcal{S}^{d-1}$. This extension of the check function immediately leads to the following definition of spatial quantile:

Definition 2.3. Suppose $Y \sim \mathbb{P}^Y$ is a random variable on $\mathbb{R}^d$ $(d \geq 1)$. Then for any $\tau \in [0, 1]$ and $u \in \mathcal{S}^{d-1}$, the τu-spatial quantile of P^Y is defined as

$$\mathrm{Q}_{\tau u} = \operatorname*{arg\,min}_{y \in \mathbb{R}^d} \mathbb{E}\, \Phi_{\tau u}(Y - y). \tag{2.79}$$

Note the solution of (2.79) are such that

$$\mathbb{E}\Big(\frac{Y - Q_{\tau u}}{\|Y - Q_{\tau u}\|}\Big) = -\tau u.$$

Intuitively speaking, this indicates that $Q_{\tau u}$ defines a point in $\mathbb{R}^d$ such that the average unit vector from it to other random samples should be τu.

The generalization to quantile regression setting is simply by applying the spatial quantile definition to $Y - b^*X - a$, where $X \in \mathbb{R}^p$ is the covariate vector, $b^* \in \mathbb{R}^{d\times p}$ is the regression coefficient and a is the intercept term. Specifically, for fixed $\tau \in [0, 1]$ and $u \in \mathcal{S}^{d-1}$,

$$(a_{\tau u}, b_{\tau,u}) = \operatorname*{arg\,min}_{b \in \mathbb{R}^{d\times p},\ a \in \mathbb{R}^d} \mathbb{E}\, \Phi_{\tau u}(Y - bX - a).$$

Therefore, given observations $(Y_1, X_1), \ldots, (Y_n, X_n)$ satisfying equations

$$Y_i = b^*X_i + \varepsilon_i,\ i = 1, \ldots, n,$$

for some random residue terms ε_i's that are independent with with X_i's, the spatial quantile estimator of b^* can be obtained by

$$(\hat{b}^{(\mathrm{sp})}, \hat{a}^{(\mathrm{sp})}_{\tau u}) = \operatorname*{arg\,min}_{b \in \mathbb{R}^{d\times p},\ a \in \mathbb{R}^p} \frac{1}{n}\sum_{i=1}^n \Phi_{\tau u}(Y_i - bX_i - a).$$

Therefore, the optimizer can be obtained by applying classical convex optimization algorithms.

Chapter 3

Coverage Correlation Coefficient: Beyond Functional Correlation

3.1 Introduction

Chatterjee's correlation has been recently developed [Cha21] to measure the dependence relationship between two univariate random variables X and Y. More precisely, given n independent copies $(X_1, Y_1), \ldots, (X_n, Y_n)$ of (X, Y) with probability distribution $P^{(X,Y)}$, and assuming no ties in $(X_i)_{1\leq i\leq n}$ and $(Y_i)_{1\leq i\leq n}$, we can rearrange the data as $(X_{(1)}, Y_{(1)}), \ldots, (X_{(n)}, Y_{(n)})$ such that $X_{(1)} < \cdots < X_{(n)}$. Chatterjee's correlation is then defined as

$$\xi_n^{(X,Y)} := 1 - \frac{\sum_{i=1}^{n-1} |r_{i+1} - r_i|}{(n^2-1)/3}, \tag{3.1}$$

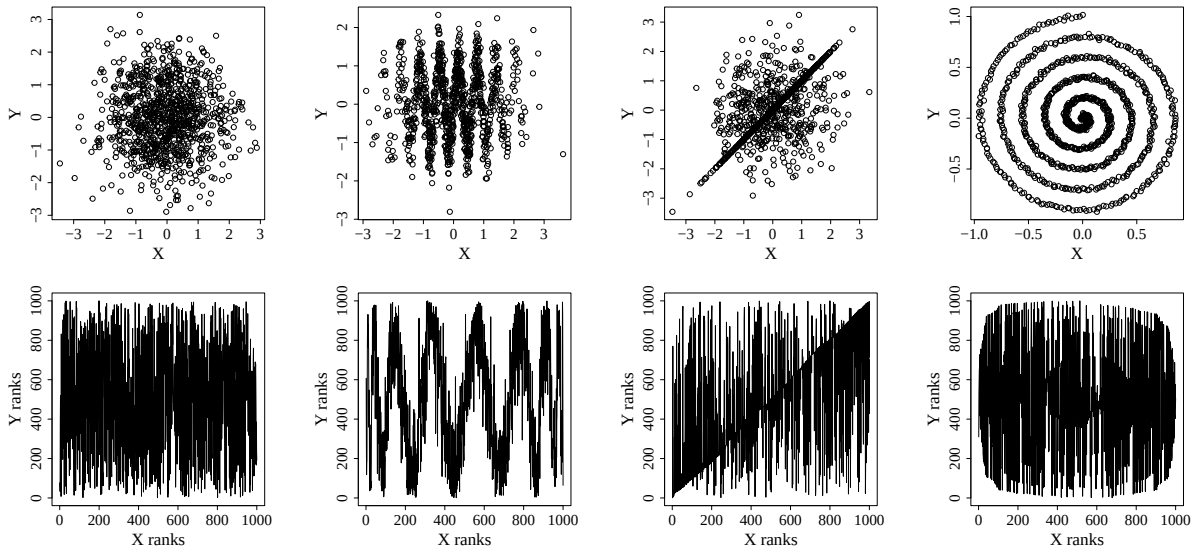
where $r_i := \#\{j : Y_{(j)} \leq Y_{(i)}\}$ is the rank of $Y_{(i)}$. Similarly to Spearman's ρ and Kendall's τ [Ken38; Spe04], Chatterjee's correlation depends only on the ranks of X_i's and Y_i's and is hence invariant under monotone transformations of the data. However, unlike Spearman's and Kendall's correlation, which is mainly used in testing against an alternative of a monotone relationship between X and Y, Chatterjee's correlation is powerful for testing a generic functional relationship between X and Y.

Given its simplicity and nonparametric nature, Chatterjee's correlation has been applied across various practical fields [e.g. Sad22; Suo+24], despite being a relatively recent development. That said, this new correlation is not without its limitations. For instance, the definition in (3.1) is asymmetric in X and Y. Chatterjee [Cha21] claimed that this is a feature of the correlation, which detects dependence of Y as a function of X and not vice versa, and can be symmetrised by taking the maximum of $\xi_n^{X,Y}$ and $\xi_n^{Y,X}$. However, this means that Chatterjee's correlation may not be powerful in detecting dependence between X and Y mediated through their respective functional dependence on some hidden variable H. Another shortcoming of $\xi_n^{(X,Y)}$ is that it only computes correlation of scalar random variables X and Y. Azadkia and Chatterjee [AC21] has generalised Chatterjee's correlation to settings of multivariate X, but the response variable Y remains univariate. Other extensions are also possible [DGS20; AF24] (see Section 3.1.1 a review), but none of them enjoy a simple and distribution-free asymptotic theory as Chatterjee's correlation coefficient.

Before moving on, let us consider an alternative geometric interpretation of Chatterjee's correlation. Figure 3.1 shows a few different (X, Y) distributions and the corresponding Chatterjee's correlation statistics and p-values. The bottom panels plots the ordered X ranks (which are simply $1, 2, \ldots, n$) against the corresponding Y ranks $r_1, \ldots, r_n$. We observe that the numerator in (3.1) is approximately the total segment length in each line plot, or more precisely, it is the total area in the $[0, n]^2$ square covered by

$$\bigcup_{i=1}^{n-1} \Big([i-1, i] \times \big[\min\{r_i, r_{i+1}\}, \max\{r_i, r_{i+1}\} \big] \Big).$$

For independent X and Y (first column of Figure 3.1), the Y-rank against X-rank line plot is highly 'space-filling'. When $\mathbb{E}[Y \mid X]$ is a function of X (second and third column of Figure 3.1), the corresponding line plot has shorter total length and the Chatterjee's correlation statistics are essentially capturing the deficiency in the space covered as compared to the independent case. The last column of Figure 3.1 gives an interesting example where X and Y are defined as functions of U with some perturbations, but the Chatterjee's correlation is unable to tell it apart from the independent case.



$\xi_n^{(X,Y)} = 0.012$, p-value $= 0.55$ | $\xi_n^{(X,Y)} = 0.44$, p-value $< 10^{-16}$ | $\xi_n^{(X,Y)} = 0.24$, p-value $< 10^{-16}$ | $\xi_n^{(X,Y)} = 0.0091$, p-value $= 0.65$

Figure 3.1: Chatterjee's correlation of various (X, Y) pairs using a sample of $n = 1000$ observation pairs. Data generating mechanism are as follows — first column: $X, Y \overset{\text{iid}}{\sim} \mathcal{N}(0, 1)$; second column: $X \sim \mathcal{N}(0, 1)$ and $Y = \sin(10X) + 0.5\epsilon$ where $\epsilon \sim \mathcal{N}(0, 1) \perp\!\!\!\perp (X, Y)$; third column: $X \sim \mathcal{N}(0, 1)$ and $Y = XB + \epsilon(1 - B)$, where $(B, \epsilon) \sim \text{Bernoulli}(1/2) \otimes \mathcal{N}(0, 1) \perp\!\!\!\perp (X, Y)$; fourth column: $X = U \sin(10\pi U) + 0.03\epsilon_X$ and $Y = U \cos(10\pi U) + 0.03\epsilon_Y$, where $(U, \epsilon_X, \epsilon_Y) \sim \text{Unif}[0, 1] \otimes \mathcal{N}(0, 1) \otimes \mathcal{N}(0, 1) \perp\!\!\!\perp (X, Y)$. For each column, the top panel shows the scatter plot and the bottom panel shows the line plot of ordered X ranks against the corresponding Y ranks.

Motivated by the above geometric interpretation of Chatterjee's correlation, we propose a *coverage correlation coefficient* of random vectors $X \in \mathbb{R}^{d_X}$ and $Y \in \mathbb{R}^{d_Y}$ with $d_X, d_Y \in \mathbb{N}$, based

on the (multivariate) ranks of $(X_i)_{1\le i\le n}$ and $(Y_i)_{1\le i\le n}$ that is powerful against the alternative where X and Y possess an implicit dependence. Intuitively speaking, the proposed correlation coefficient measures the uncovered volume in $[0,1]^{d_X+d_Y}$ when small cubes of volume $1/n$, centered at the multivariate ranks of (X_i, Y_i), are used to cover the space.

In the case of d-dimensional Euclidean space, $d \ge 2$, there is no canonical definition of rank since ordering becomes less intuitive when dealing with multivariate variables. Although various concepts have been considered, e.g. depth-based ranks [Tuk75; LS93; ZS00], spatial ranks [MO95; Cha96; Kol97], componentwise ranks [Hod55; Bic65], Mahalanobis ranks [HP02b; HP02a], but none of them enjoy distribution-freeness while the traditional rank notion on real line does. *Monge-Kantorovich (MK) rank* , a concept of multivariate rank proposed in Chernozhukov et al. [Che+17], Hallin [Hal17], and Hallin et al. [Hal+21] provides a new insight of traditional ranks from the perspective of optimal transport map. Many desirable properties of univariate rank are enjoyed by this new concept including distribution-freeness.

The MK ranks for $(X_i)_{i=1}^n$ and $(Y_i)_{i=1}^n$are defined through their optimal transport to respective sets of reference points $\mathcal{U} = \{u_1, \dots, u_n\}$ and $\mathcal{V} = \{v_1, \dots, v_n\}$. Various choices have been used in the literature [DS23; Hal+21; BSS18]. In this work, we consider two types of reference points:

(1) Regular grid: suppose there exists $m_X, m_Y \in \mathbb{N}$ such that $n = m_X^{d_X} = m_Y^{d_Y}$, let $\mathcal{U} = \prod_{i=1}^{d_X}\{1/(m_X+1), 2/(m_X+1)\dots, m_X/(m_X+1)\}$ and $\mathcal{V} = \prod_{i=1}^{d_Y}\{1/(m_Y+1), 2/(m_Y+1)\dots, m_Y/(m_Y+1)\}$.

(2) Uniform random samples: let $\mathcal{U} = \{U_i\}_{i=1}^n$ and $\mathcal{V} = \{W_i\}_{i=1}^n$, where $U_1, \dots, U_n$ and $W_1, \dots, W_n$ are independent i.i.d. samples from uniform distributions $\mathrm{Unif}([0,1]^{d_X})$ and $\mathrm{Unif}([0,1]^{d_Y})$, respectively.

Specifically, define

$$\pi^{X,\star} := \operatorname*{arg\,min}_{\pi\in\mathcal{S}_n} \frac{1}{n}\sum_{i=1}^n \|u_{\pi(i)} - X_i\|^2 \quad \text{and} \quad \pi^{Y,\star} := \operatorname*{arg\,min}_{\pi\in\mathcal{S}_n} \frac{1}{n}\sum_{i=1}^n \|v_{\pi(i)} - Y_i\|^2, \tag{3.2}$$

where $\|\cdot\|$ denotes the Euclidean norm, $\mathcal{S}_n$ denotes the set of all permutations of $[n]$, and $\star \in \{\mathrm{Reg}, \mathrm{Rand}\}$ indicates whether the $\mathcal{U}$ and $\mathcal{V}$ are chosen to be a regular grid ($\star = \mathrm{Reg}$) or uniform random samples ($\star = \mathrm{Rand}$). Then the empirical multivariate rank for X_i and Y_i is

$$\widehat{R}^{X,\star}(X_i) := u_{\pi^{X,\star}(i)} \quad \text{and} \quad \widehat{R}^{Y,\star}(Y_i) := v_{\pi^{Y,\star}(i)}, \tag{3.3}$$

respectively. We further write

$$\widehat{R}_i^\star := (\widehat{R}^{X,\star}(X_i), \widehat{R}^{Y,\star}(Y_i)). \tag{3.4}$$

For any d-dimensional cube P with edge length $l_P \in \mathbb{R}_+$, we can construct a d-torus $\mathbb{T}_P$ by gluing opposite faces together. Then the ℓ_∞-norm on $\mathbb{T}_P$ is equivalent to the following distance on P:

$$d_\infty(u,v) := \inf_{z\in\mathbb{Z}^d} \|u - v - z l_P\|_\infty, \quad \text{for any } u, v \in P. \tag{3.5}$$

Then we define d-dimensional subcube centered at $w \in P$ with the periodic boundary condition and edge length $2r$ as:

$$B_\infty(w, r; P) := \{u \in P : d_\infty(w,u) \le r\}.$$

For any $s \in \mathbb{N}$, when $P = [0,1]^s$, we also write $B^s_\infty(w,r) := B_\infty(w,r;[0,1]^s)$ as shorthand. Write $d := d_X + d_Y$ and $\mathrm{vol} \equiv \mathrm{vol}_d$ for the d-dimensional Lebesgue measure, for $\gamma \in (0,1)$, we define

$$V^\star_{n,\gamma} := 1 - \mathrm{vol}\Big(\bigcup_{i=1}^n B^d_\infty(\widehat{R}^\star_i, \gamma)\Big) \tag{3.6}$$

to be the uncovered volume in the d-dimensional unit cube outside subcubes of radius γ centred at the empirical ranks. We will mostly be interested in setting $\gamma := \frac{1}{2n^{1/d}}$ so that each subcube has volume of $1/n$. We write $V^\star_n = V^\star_{n,\gamma}$ for this specific choice of γ. The empirical coverage correlation coefficient between samples $X_1, \ldots, X_n$ and $Y_1, \ldots, Y_n$ is then defined as

$$\kappa^{X,Y;\star}_n := \frac{V^\star_n - e^{-1}}{1 - e^{-1}} = 1 - \frac{\mathrm{vol}\big(\bigcup_{i=1}^n B^d_\infty(\widehat{R}^\star_i, (1/2)n^{-1/d})\big)}{1 - e^{-1}}, \tag{3.7}$$

with $\star \in \{\mathrm{Reg}, \mathrm{Rand}\}$. It is shown in Theorem 3.1 that when $d_X = d_Y = 1$, for both regular grid reference points and uniform random samples reference points, the coverage correlation coefficient (3.7) consistently estimate the following f-divergence (see Definition 3.1):

$$\kappa^{X,Y} := D_f(P^{(X,Y)} \,\|\, P^X \otimes P^Y) \tag{3.8}$$

with $f(x) = (e^{-x} - e^{-1})/(1 - e^{-1})$ for $x \in \mathbb{R}$.

We summarise several key features of the coverage coefficient of correlation $\kappa^{X,Y;\star}_n$ as follows:

1. It generalises the "space-filling" geometric intuition of Chatterjee's correlation coefficient to multivariate settings, preserving interpretability while extending its application to complex dependence structures between random vectors.

2. When $d_X = d_Y = 1$, it consistenly estimate a population quantity $\kappa^{X,Y}$ that equals to 0 if and only if X is independent with Y, and equals to 1 if and only if $P^{(X,Y)}$ is singular to $P^X \otimes P^Y$ (see Lemma 3.1). Therefore, unlike Chatterjee's correlation coefficient, $\kappa^{X,Y;\star}_n$ is not only capable to detect functional dependency, but also implicit functional correlation. See Section 3.3 for numerical demonstrations.

3. A sub-gaussian concentration inequality is established for $\kappa^{X,Y;\mathrm{Reg}}_n$ (see Proposition 3.2).

4. When X and Y are independent, $\kappa^{X,Y;\mathrm{Rand}}_n$ allows a simple central limit theorem (CLT), converging to a normal distribution with zero mean and an explicit variance formulation that is independent of the underlying distribution of X and Y (see Theorem 3.3).

5. For univariate marginal distributions, we develop an algorithm with $O(n \log n)$ time complexity (see Section 3.3.1), while in higher dimensions, the computational complexity of this algorithm increases polynomially with dimension.

3.1.1 Related works

Dependency measurement has been a timeless problem in statistics. Since the classical concepts of correlation coefficient proposed by Pearson [Pea20], Kendall [Ken38], and Spearman [Spe04], the literature on measuring statistical dependence has expanded considerably, including, maximal

correlation coefficient [Geb41; Koy87], kernel-based method [Gre+05b; Gre+07; SS14; Ram+15; SKR23], coefficient based on pairwise distance [SR09; SRB07; HHG13], copulas based-method [Skl59; SW81; DSS13; LHS13; Fuc24; GJT22; SDS24], coefficient based on Wasserstein distance [NSM21; MS22; Wie22; MS20]. For a comprehensive survey we guide the readers to [TOS22], and a survey more related to this work [Cha24].

A common issue of the methods mentioned above is that most of the coefficients do not allow a simple distribution-free asymptotic theory. Hence, one need to resort to permutation-based techniques to generate p-values, which can be computationally infeasible in the context of multiple testing regime. Chatterjee [Cha21] proposed a rank-based correlation coefficient with a simple form as the classicial coefficients, and thanks to the distribution-freeness of statistical rank, it allows a simple asymptotic theory without any assumptions on the marginal distributions. By leveraging the similar idea, Azadkia and Chatterjee [AC21] generalised the coefficient to the setting of conditional dependence, and the distribution-free asymptotic theory is later obtained in [SDH24]. A line of works on the power analysis of these two correlations coefficients are later developed in [SDH22b; SDH24; CB20], and a modified version of Chatterjee's coefficient which attains near-optimal rates of power under several alternatives is proposed in [LH23].

However, since the Chatterjee's coefficent depends on the rank of the marginal distributions, thus does not allow a direct generalisation to measure the dependency between random vectors. By leveraging the idea of k-nearest neighbor (kNN) (see also [BS19]), Deb, Ghosal, and Sen [DGS20] proposed a kernel-based correlation coefficients under multivariate setting which, in the univariate case, consistently estimate the same population quantity as (3.1). However, the power of the statistic depends on the tuning parameter k, the optimal value of k may vary for different alternatives (see Section 5 of [LH23]). More recently, Ansari and Fuchs [AF22] proposed a direct multivariate extension of Chatterjee's correlation; however, their coefficient does not enjoy a distribution-free asymptotic theory. In this work, the proposed coverage correlation coefficient generalises the geometric principle of Chatterjee's coefficient to the multivariate setting, providing a simple distribution-free asymptotic theory without the need for any tuning parameters.

It is worth to mention that, in the case of univariate marginal distributions, a similar coverage statistic was considered by Rudra, Zhou, and Wright [RZW17] even before the proposal of Chatterjee's correlation coefficient. Although their work recognised the effectiveness of the coverage correaltion coefficient, they didn't develop a consistency theorem or asymptotic theory for the statistic. Moreover, their implementation requires a $O(n^2)$ time complexity, whereas our approach achieves $O(n \log n)$ (see Section 3.3.1).

In stead of focusing on Chatterjee's correlation coefficient, another line of work leverages the concept of MK rank to adapt the classical independence test statistics or correlation coefficients to the multivariate setting. Thanks for the distribution-freeness of the MK rank, the resulting rank-based statistics typically enjoy a distribution-free asymptotic theory. Examples includes Hoeffding's D statistic [GS22], distance covariance [DS21; SDH22a], quadrant statistic, Spearman's ρ and Kendall's τ [Shi+24]. Moreover, Shi et al. [Shi+22] proposed a general framework for designing consistent and distribution-free independence tests using the center-outward multivariate rank [Hal+21]. Please refer to [Han21] for a survey along this direction.

Notation: For probability spaces $(\mathcal{X}, \Sigma_{\mathcal{X}}, \mu)$ and $(\mathcal{Y}, \Sigma_{\mathcal{Y}}, \nu)$, we write the product space as $(\mathcal{X} \times \mathcal{Y}, \Sigma_{\mathcal{X}} \otimes \Sigma_{\mathcal{Y}}, \mu \otimes \nu)$. Let $\mathcal{P}(\mathcal{X})$ be the set of all probability measures on $\mathcal{X}$. For any random variable X, write P^X as the induced Borel probability measure. For any set A, we denote $\binom{A}{k} = \{A' \subseteq A : |A'| = k\}$. The sets of all positive and non-negative real numbers are denoted

by $\mathbb{R}_{>0}$ and $\mathbb{R}_{\geq 0}$ respectively, and the sets of all positive and non-negative rational numbers are denoted by $\mathbb{Q}_{>0}$, $\mathbb{Q}_{\geq 0}$, respectively. For any $n \in \mathbb{N}$, let $[n] = \{1, 2, \ldots, n\}$. We denote $\xrightarrow{\mathrm{w}}$ as weak convergence, $\xrightarrow{\mathrm{p}}$ as convergence in probability and $\xrightarrow{\mathrm{d}}$ as convergence in distribution.

3.2 Theory

In this section, we formally present the theoretical results for the converage correlation coefficient $\kappa_n^{X,Y;\star}$ (see definition in (3.7)). Specifically, when X, Y are both univariate random variables, we show a consistency result for $\kappa_n^{X,Y;\star}$ when $\star \in \{\mathrm{Rand}, \mathrm{Reg}\}$. Moreover, when we choose the regular grid reference distribution, i.e. $\star = \mathrm{Reg}$, a nonparametric concentration inequality is established. For the case of multivariate marginal distributions, a distribution-free asymptotic theorem under the null is obtained for $\kappa_n^{X,Y;\mathrm{Rand}}$.

3.2.1 Univariate case

In this section, we focus on the case that X and Y are univariate distributions. We show that $\kappa_n^{X,Y;\star}$ is a consistent estimator of an f-divergence between $P^{(X,Y)}$ and $P^X \otimes P^Y$ as defined in (3.8). Recall the definition of f-divergence:

Definition 3.1 ([PW25]). Let $f : \mathbb{R} \to \mathbb{R} \cup \{+\infty\}$ be a convex function with $f(1) = 0$. For any two probability measures μ, ν on a space $\mathcal{S}$, let $\mathrm{d}\mu = h\,\mathrm{d}\nu + \mathrm{d}\nu^{\perp}$ be the Lebesgue–Radon–Nikodym decomposition of μ with respect to ν, where h is ν-integrable and $\nu^{\perp}$ is singular with respect to ν. The *f-divergence* between μ and ν is defined as

$$D_f(\mu \,\|\, \nu) = \int_{\mathcal{S}} f \circ h \, d\nu + f'(\infty)\nu^{\perp}(\mathcal{S}),$$

where $f'(\infty) := \lim_{t\to\infty} t^{-1} f(t)$ is the limit of the slope of f at infinity.

Now, we are ready to introduce the consistency result of the coverage correlation coefficient when the marginal distributions are univariate, whose proof can be found at Section 3.4.1.

Theorem 3.1. *Let $(X_1, Y_1), \ldots, (X_n, Y_n)$ be independent and identically distributed from an absolutely continuous distribution $P^{(X,Y)} \in \mathcal{P}(\mathbb{R}^2)$. Let P^X and P^Y be the marginal distributions of X_1 and Y_1 respectively. Define $f : \mathbb{R} \to \mathbb{R}$ as $f(x) = (e^{-x} - e^{-1})/(1 - e^{-1})$. Then, we have*

$$\kappa_n^{X,Y;\star} \xrightarrow{\mathrm{p}} \kappa^{X,Y} := D_f(P^{(X,Y)} \,\|\, P^X \otimes P^Y) \quad \text{as } n \to \infty, \tag{3.9}$$

for $\star \in \{\mathrm{Reg}, \mathrm{Rand}\}$.

Another popular f-divergence for measuring independence is the *mutual information*, defined as the KL-divergence (i.e. let $f(x) = x \log x$ in the definition of f-divergence) between the joint distribution and the product of two marginal distribution. For instance, Berrett and Samworth [BS19] considered a test statistic based on a k-NN estimator of the mutual information; however, due to the lack of asymptotic distribution-freeness, their test relies on permutation-based techniques to yield a p-value. In constrast, as we will see later, $\kappa_n^{X,Y;\star}$ as an estimator of $\kappa^{X,Y}$ enjoys a distribution-free asymptotic theory (see Theorem 3.3).

The following proposition examines whether $\kappa^{X,Y}$, as a dependency measurement, satisfies axioms considered in [Bor+23] (see also [MS19; Rén59]).

Proposition 3.1. *Supoose $(X, Y) \sim \mathcal{P}^{(X,Y)} \in \mathcal{P}(\mathbb{R}^2)$ and P^X, P^Y are the corresponding marginal probability measures. Let $\kappa(X, Y)$ be defined as (3.9), then we have*

(i) $\kappa(X, Y) = 1$ *if and only if* $P^{(X,Y)}$ *is singular to* $P^X \otimes P^Y$, *and* $\kappa(X, Y) = 0$ *if and only if* X *is independent with* Y;

(ii) for a random variable $Z \sim P^Z \in \mathcal{P}(\mathbb{R})$, *if* $X \perp\!\!\!\perp Y|Z$, *then we have* $\kappa(Z, Y) \geq \kappa(X, Y)$;

(iii) for any sequence of random variables X_n *and* Y_n *such that* $P^{(X_n,Y_n)} \xrightarrow{\text{w}} P^{(X,Y)}$, *we have* $\liminf_{n\to\infty} \kappa(X_n, Y_n) \geq \kappa(X, Y)$;

(iv) $\kappa(X, Y) = \kappa(Y, X)$.

The proof of the above are direct applications of basic properties of f-divergence (see Section 3.4.2). Unlike many existing dependency measurement, $\kappa^{X,Y}$ is not maximised when X and Y have a deterministic functional relationship, which is intentional. Because we would like to not only detect the functional correlation between X and Y but also other types of more complicated correlation, for instance, spurious correlation through confounding variables.

Remark 3.1 (Multivariate extension)**.** Note that $\kappa^{X,Y}$ is well-defined for any $P^X \in \mathcal{P}(\mathbb{R}^{d_X})$ and $P^Y \in \mathcal{P}(\mathbb{R}^{d_Y})$ with $d_X, d_Y \geq 1$. Then, the following information gain inequality is an immediate result of Proposition 3.1(ii):

(ii') for any random variable X, X' and Y, we have $\kappa((X, X'), Y) \geq \kappa(X, Y)$.

This inequality is not mentioned in [Bor+23], but as an axiom for denpedency measurement between random vectors in [GJT22].

Moreover, when considering the coverage correlation coefficient under the regular grid reference distribution, a sub-Gaussian concentration inequality can be proved for the vacancy area defined in (3.6). The proof is deferred to Section 3.4.3.

Proposition 3.2. *Let $V_{n,\gamma}^{\text{Reg}}$ be the vacancy area with regular grid reference distribution defined in (3.6). Then we have for any $t \geq 0$, there exists a fixed constant $C > 0$ independent with n and t such that*

$$\mathbb{P}\left(\left|V_{n,\gamma}^{\text{Reg}} - \mathbb{E}(V_{n,\gamma}^{\text{Reg}})\right| \geq t\right) \leq 2e^{-Cnt^2}.$$

Note that Proposition 3.2 implies that $V_{n,\gamma}^{\text{Reg}}$ concentrates around its mean with a $\sqrt{n}$-convergence rate. Moreover, under the assumption of independent marginal distributions, Lemma 3.5 implies that $\mathbb{E}(V_{n,\gamma}^{\text{Reg}}) = (1 - n^{-1/2})^{\sqrt{n}} + O(n^{-1/2})$. Therefore, when $P^{(X,Y)} = P^X \otimes P^Y$, $V_{n,\gamma}^{\text{Reg}}$ concentrates around e^{-1} with a $\sqrt{n}$-convergence rate. We continue discussion on this aspect in the next section.

3.2.2 Multivariate case

Now, we turn to the case of multivariate marginal distributions, i.e. $X \sim P^X \in \mathcal{P}(\mathbb{R}^{d_X})$, $Y \sim P^Y \in \mathcal{P}(\mathbb{R}^{d_Y})$ where d_X, d_Y can be integers larger than 1. Note that the construction of regular grid reference distribution requires there exists $m_X, m_Y \in \mathbb{N}$ such that $n = m_X^{d_X} = m_Y^{d_Y}$, which can be restricted in practice when d_X, d_Y are larger than 1. Moreover, the regular grid on $[0,1]^d$ suffers from the "curse of dimensionality" in the sense that its approximation error to $\mathrm{Unif}([0,1]^d)$ is of order $n^{-2/d}$, while the random samples from $\mathrm{Unif}([0,1]^d)$ is order $n^{-1/2}$ [LP14]. Therefore, we focus on developing theories for $\kappa^{X,Y;\mathrm{Rand}}$ in the case of multivariate marginals.

We first establish the explicit mean and variance expression of $\kappa^{X,Y;\mathrm{Rand}}$ (please refer to Section 3.4.4 for the proof).

Theorem 3.2. *Assuming $X \sim P^X \in \mathcal{P}(\mathbb{R}^{d_X})$ and $Y \sim P^Y \in \mathcal{P}(\mathbb{R}^{d_Y})$, $d_X, d_Y \geq 1$, are independent random vectors. Let $d = d_X + d_Y$, then $\mathbb{E}(V_{n,\gamma}^{\mathrm{Rand}}) = (1-1/n)^n$ and*

$$\mathrm{Var}(V_{n,\gamma}^{\mathrm{Rand}}) = \sum_{i=2}^{n} \binom{n}{i} \Big(1 - \frac{2}{n}\Big)^{n-i} \bigg(\Big(\frac{2}{i+1}\Big)^d n^{-i-1} - n^{-2i} \bigg).$$

Therefore, as $n \to \infty$ we have $\mathbb{E}(V_{n,\gamma}^{\mathrm{Rand}}) \to e^{-1}$ and $\mathrm{Var}(V_{n,\gamma}^{\mathrm{Rand}}) = S_n^2(1 + o(1))$, where

$$S_n^2 = \sum_{i=2}^{n} \binom{n}{i} \Big(1 - \frac{2}{n}\Big)^{n-i} \Big(\frac{2}{i+1}\Big)^d n^{-i-1}. \tag{3.10}$$

By the explicit form of (3.10), it is not hard to see that as $n \to \infty$, $\mathrm{Var}(V_{n,\gamma}^{\mathrm{Rand}})$ scales at the rate of n^{-1}, in fact, we have

$$n\mathrm{Var}(V_{n,\gamma}^{\mathrm{Rand}}) \to e^{-2} \sum_{i=2}^{\infty} \frac{1}{i!} \Big(\frac{2}{i+1}\Big)^d, \quad \text{as } n \to \infty.$$

Now we introduce a totally distribution-free asymptotic theory for $\kappa_n^{X,Y;\mathrm{Rand}}$ when Y is independent with X (the proof is in Section 3.4.5).

Theorem 3.3. *Let $P^X \in \mathcal{P}(\mathbb{R}^{d_X})$, $P^Y \in \mathcal{P}(\mathbb{R}^{d_Y})$ and $\kappa_n^{X,Y;\mathrm{Rand}}$ defined as (3.7). Then when X and Y are independent, we have*

$$n^{1/2} \kappa_n^{X,Y;\mathrm{Rand}} \xrightarrow{\mathrm{d}} \mathcal{N}(0, \sigma^2), \quad \text{as } n \to \infty, \tag{3.11}$$

where $\sigma^2 = (e-1)^{-2} \sum_{i=2}^{\infty} \frac{1}{i!} \big(\frac{2}{i+1}\big)^d$.

The theorem is built upon a class of method called "coverage process" (see [e.g. Hal85; Hal88]). The core idea is to split the hypercube $[0,1]^d$ into numerous small subcubes, and then consider further subcubes, such that the covered area in these further subcubes are independent. Please find the detailed proof in Section 3.4.5.

Based on Theorem 3.3 and the asymptotic variance formula (3.10), we can construct a test for

$$\mathcal{H}_0 : P^{(X,Y)} = P^X \otimes P^Y \quad v.s. \quad \mathcal{H}_1 : P^{(X,Y)} \neq P^X \otimes P^Y, \tag{3.12}$$

as follows: We reject $\mathcal{H}_0$ if

$$n^{1/2}\frac{(1-e^{-1})\kappa_n^{X,Y;\mathrm{Rand}}}{S_n} \geq z_\alpha, \tag{3.13}$$

where z_α is the α-uppder quantile of standard normal distribution. Since the limiting distribution is totally independent with the marginal distribution P^X and P^Y, test (3.13) is distribution-free. Such desirable property is particularly preferred in the context of multiple testing, where one would resort to some correction procedures, for instance, Bonferroni procedure and Benjamini–Hochberg procedure, which requires an extremely small p-value to reject the null hypothesis. Thanks to the distribution-freeness, the proposed test (3.13) can typically achieve a very small p-value under the null hypothesis (see Section 3.3 for an empirical demonstration), while other testing procedures like distance covariance [SRB07] and HSIC [Gre+05b] require numerous permutation tests, which are computationally expensive.

Remark 3.2. Theorem 3.3 establishes a CLT of $\kappa_n^{X,Y;\mathrm{Rand}}$ for all dimensions. In particular, when $d_X = d_Y = 1$, the expression of the asymptotic variance have an explicit value of $(e-1)^{-2}(-4\gamma_0 + 4\mathrm{Ei}(1) - 5) \approx 0.091992$, where γ_0 is Euler's gamma constant and $\mathrm{Ei}(1)$ is the exponential integral evaluate at 1.

3.3 Simulations

In this section, we present some numerical demonstrations on the effectiveness of the coverage correlation coefficient. We first discuss the computation complexity of our $\kappa_n^{X,Y;\star}$ in Section 3.3.1, and then in Section 3.3.2, we compare the power performance of correlation coefficient with some existing methods in different settings.

3.3.1 Computation

The main challenge lies in the implementation of $\kappa_n^{X,Y;\star}$ is the calculation of the uncovered area $V_{n,\gamma}^\star$, or equivalently, the covered area $1 - V_{n,\gamma}^\star$. When $d_X = d_Y = 1$, edges of n subcubes split the unit cube $[0,1]^2$ into $(n+1)^2$ smaller regions that we call *elementary cubes* (see Fig. 3.2 for an example of $n = 2$). To calculate the total covered area, we simply need to sum the areas of all elementary cubes that are covered by at least one subcube. A naive algorithm is to iterate over all n subcubes, and for each subcube, we identify the covered elementary cubes in it. However, in the worst case where all the subcubes stack together, such algorithm is of order $O(n^2)$. An alternative approach leverages the fact that, for every interval on the x-axis, the covered area within that interval depends solely on the corresponding covered length on the y-axis. For instance, in Fig. 3.2, the covered area between (x_2, x_3) is $(y_4 - y_1)(x_3 - x_2)$. Therefore, the problem of computing the total covered area can be reduced into the following query:

Given an interval on x-axis, what is the corresponding covered length on the y-axis?

Since $[0,1]$ is divided into $n+1$ intervals by n points, and each query can be answered via *segment tree* in $O(\log n)$ time [De 00], the entire problem can be solve in $O(n \log n)$ time. In higher dimensional case, the time complexity of this algorithm grows polynomially with the dimension, we suggest to use the Monte Carlo integration to obtain an approximation of the covered area.

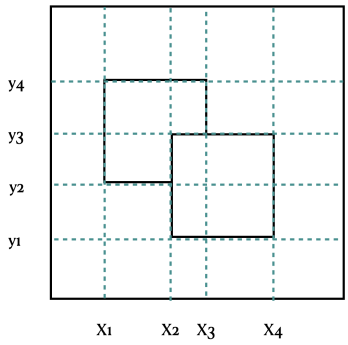


Figure 3.2: Two subcubes split the unit cube into 25 elementary cubes.

3.3.2 Power comparison

In the following, we carry out an empirical power analysis on the proposed independence test (3.13). Our main finding reveals that the proposed test demonstrates performance comparable to Chatterjee's correlation coefficient when signals exhibit oscillatory behavior, while achieving superior performance under alternatives with implicit correlation through confounding variables.

We evaluate the performance of $\kappa_n^{X,Y;\mathrm{Rand}}$ under both univariate and multivariate marginal distributions:

(a) univariate marginal distributions: $d_X = d_Y = 1$ with a fixed sample size $n = 1000$

(b) multivariate marginal distributions: $d_X = 2, d_Y = 1$ with a fixed sample size $n = 2000$

We compare $\kappa_n^{X,Y;\mathrm{Rand}}$ with the following quantities:

(I) Kernel measure of association (KMAc) [DGS20]: Let $(X_1, Y_1), (X_2, Y_2) \overset{\text{iid}}{\sim} P^{(X,Y)} \in \mathcal{P}(\mathbb{R}^d)$. Then generate $(X', Y', \widetilde{Y}')$ as follows: draw $X' \sim P^X$, and then draw $Y', \widetilde{Y}' \overset{\text{iid}}{\sim} P^{Y|X=X'}$. Consider the following measure of dependency:

$$\eta_K(X, Y) := \frac{\mathbb{E}[\mathbb{E}[K(Y', \widetilde{Y}') \mid X']] - \mathbb{E}[K(Y_1, Y_2)]}{\mathbb{E}[K(Y_1, Y_1)] - \mathbb{E}[K(Y_1, Y_2)]},$$

where $K : \mathbb{R}^{d_X} \times \mathbb{R}^{d_Y} \to \mathbb{R}$ is a symmetric, nonnegative kernel function. Although the expression above may not seem to be intuitive as a dependency measurement between X and Y, [DGS20] demonstrates that it is simply a rescaled *maximal mean discrepancy (MMD)* between P^X and the conditional distribution $P^{Y|X}$. In [DGS20], the authors considered a general graph-based estimator of η_K, which is defined as a coefficient of correlation between X and Y. In our implementation, we construct the KMAc by using Gaussian kernel, i.e. $K(x, y) = \exp(-\|x - y\|^2)$, and k-NN with fixed $k = 20$.

(II) Chatterjee's correlation coefficient [Cha21]: since Chatterjee's coefficient is only defined for univariate marginals, we compute its empirical powers and compare them with other methods solely under the case where $d_X = d_Y = 1$.

(III) Distance correlation [SRB07]: Given $(X_1, Y_1), \ldots, (X_n, Y_n) \overset{\text{iid}}{\sim} P^{(X,Y)} \in \mathcal{P}(\mathbb{R}^d)$. Let $a_{ij} = \|X_i - X_j\|$ and $b_{ij} = \|Y_i - Y_j\|$. Then we center them by considering $A_{ij} = a_{ij} - a_{i,+} - a_{+,j} +$

$a_{+,+}$ and $B_{ij} = b_{ij} - b_{i,+} - b_{+,j} + b_{+,+}$, where $a_{i,+} = \frac{1}{n}\sum_{j=1}^n a_{ij}$, $a_{+,+} = \frac{1}{n^2}\sum_{i=1}^n\sum_{j=1}^n a_{ij}$, $a_{+,j} = \frac{1}{n}\sum_{i=1}^n a_{ij}$, and similarly for $b_{i,+}, b_{+,j}, b_{+,+}$. The distance correlation is simply the Pearson's correlation between A_{ij} and B_{ij}'s, i.e.

$$\mathrm{dcor}(\{(X_i,Y_i)\}_{i=1}^n) = \frac{\frac{1}{n^2}\sum_{i,j=1}^n A_{ij}B_{ij}}{\sqrt{\frac{1}{n^2}\sum_{i,j=1}^n A_{ij}}\sqrt{\frac{1}{n^2}\sum_{i,j=1}^n B_{ij}}}.$$

(IV) HSIC [Gre+05b; Gre+07]: Given $(X_1,Y_1),\dots,(X_n,Y_n) \overset{\text{iid}}{\sim} P^{(X,Y)} \in \mathcal{P}(\mathbb{R}^d)$. Consider two kernel functions $K, L : \mathbb{R}^d \to \mathbb{R}$, write $k_{ij} = K(X_i, X_j)$ and $l_{ij} = L(Y_i, Y_j)$. The HSIC test statistic is

$$\mathrm{HSIC}(\{(X_i,Y_i)\}_{i=1}^n) = \frac{1}{n^2}\sum_{i,j=1}^n k_{ij}l_{ij} + \frac{1}{n^4}\sum_{i,j,s,t=1}^n k_{ij}l_{st} - 2\frac{1}{n^3}\sum_{ijs=1}^n k_{ij}l_{is}.$$

We choose both K and L to be be the Gaussian kernel with median heuristic as bandwidth.

For method (I), (III), (IV), we calculate their empirical powers (with 400 replications) under several different settings under both cases (a) and (b), while for Chatterjee's correlation coefficient we only calculate its empirical power under case (a). Additionally, method (III) and (IV) do not allow an distribution-free null distribution, we apply a permutation-based technique with 600 permutations to obtain the p-values.

We introduce some extra notations before stating our findings. We denote $X^{(i)}$, where $i = 1,\dots,d_X$, to be the i-th coordinate of d_X-dimensional random vector X, and $Y^{(j)}$, where $j = 1,\dots,d_Y$, be the j-th coordinate of d_Y-dimensional random vector Y. Without loss of generality, we assume $d_X \geq d_Y$ and let $d = d_X + d_Y$. Let λ to be the noise level taking values in $\{0, 0.2, 0.4, 0.6, 0.8, 1, 1.2, 1.4, 1.6, 1.8, 2\}$. Suppose $(\varepsilon_1, \varepsilon_2) \sim \mathcal{N}(0, I_{d_X}) \otimes \mathcal{N}(0, I_{d_Y})$ are independent with all the other randomness, and we denote $\varepsilon_1^{(j)}$ and $\varepsilon_2^{(j)}$ as the j-th coordinate of them. We summarise our settings and the corresponding findings as follow.

Linear correlation. The following linear correlation model is considered (see Fig. 3.3):

$$Y = 0.5X\mathbf{1} + 7\lambda\varepsilon_2, \tag{3.14}$$

where $\mathbf{1} \in \mathbb{R}^{d_X \times d_Y}$ is a matrix with all entries equal to 1. Since the coverage correlation coefficient inherits the "space-filling" idea from the Chatterjee's coefficient, it is not surprising to see that both $\kappa_n^{X,Y;\mathrm{Rand}}$ and Chatterjee's correlation coefficient have relatively inefficient power performance compare to dCor and HSIC under a smooth alternative, such as linear correlation [CB20; SDH22b; ADN21]. The power of KMAc is notably affacted by the increasing level of noise.

Archimedean spiral and Lissajous. Let $U \sim \mathrm{Unif}([0,1]^{d_X})$. Write $U^{(i)}$ be the i-th coordinate of random vector U. We consider the following two data-generating mechanism:

(a) Archimedean spiral:

$$X^{(i)} = U^{(i)}\sin\big(10\pi U^{(i)}\big) + 0.15\lambda\varepsilon_1^{(i)}, \quad i = 1,\dots,d_X \tag{3.15}$$

$$Y^{(j)} = U^{(j)}\cos\big(10\pi U^{(j)}\big) + 0.15\lambda\varepsilon_2^{(j)}, \quad j = 1,\dots,d_Y. \tag{3.16}$$

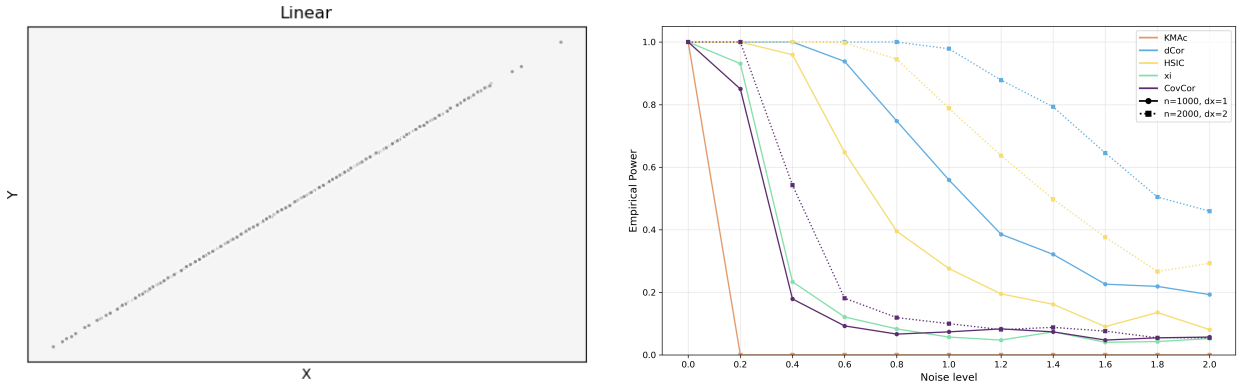


Figure 3.3: Dependency measurements applied to increasing noisy dataset with linear correlation model (3.14). The left-hand side presents a noiseless pattern of the samples with $n = 2000$. The right-hand side is the power curve display the results of various methods under the case of $n = 1000, d_X = d_Y = 1$ and $n = 2000, d_X = 2, d_Y = 1$.

(b) Lissajour curve:

$$X^{(i)} = \sin\big(3U^{(i)} + \pi/2\big) + 0.1\lambda\varepsilon_1^{(i)}, \quad i = 1, \ldots, d_X \tag{3.17}$$

$$Y^{(j)} = \sin\big(4\pi U^{(j)}\big) + 0.1\lambda\varepsilon_2^{(j)}, \quad j = 1, \ldots, d_Y. \tag{3.18}$$

The correlation defined above can be viewed as spurious correlation between X and Y through a confounder U. Fig. 3.4 implies that the coverage correlation coefficient dominates Chatterjee's coefficient in both cases, and for Lissajous type of correlation, our proposed coefficient significantly better than all the other competitors. This is consistent with the finding in Proposition 3.1(i).

Fractional Brownian Motion. Fractional Brownian motion (FBM) is a generalisation of the Brownian motion with dependent increment. Sepcifically, consider a Gaussian process $B^h(t)$ in $[0, T]$ with zero expectation and covariance function

$$\mathbb{E}[B^h(t)B^h(s)] = \frac{1}{2}(|t|^{2h} + |s|^{2h} - |t - s|^{2h}),$$

where $h \in (0, 1)$ is call the Hurst index (see [Nou12]). Let $B_i^h(t)$, for $i = 1, \ldots, d$, be d independent FBM on $[0, T]$ where we fix $h = 0.75$ and $T = n$. Then we generate X and Y through:

$$X = (B_1^h(t), \ldots, B_{d_X}^h) + 0.5\lambda\varepsilon_1 \quad \text{and} \quad Y = (B_{d_X+1}^h(t), \ldots, B_d^h) + 0.5\lambda\varepsilon_2, \tag{3.19}$$

Although model (3.19) implies a stochastic correlation between X and Y, which can not be described by a functional relationship, Fig. 3.5 implies a comparable performance between our method and Chatterjee's coefficient. This is mainly because the stochastic correlation can frequently present a clear correlation pattern for each realisation. For instance, the example on the right-hand side of Fig. 3.5 shows a clear correlation even the correlation itself is generated stochastically.

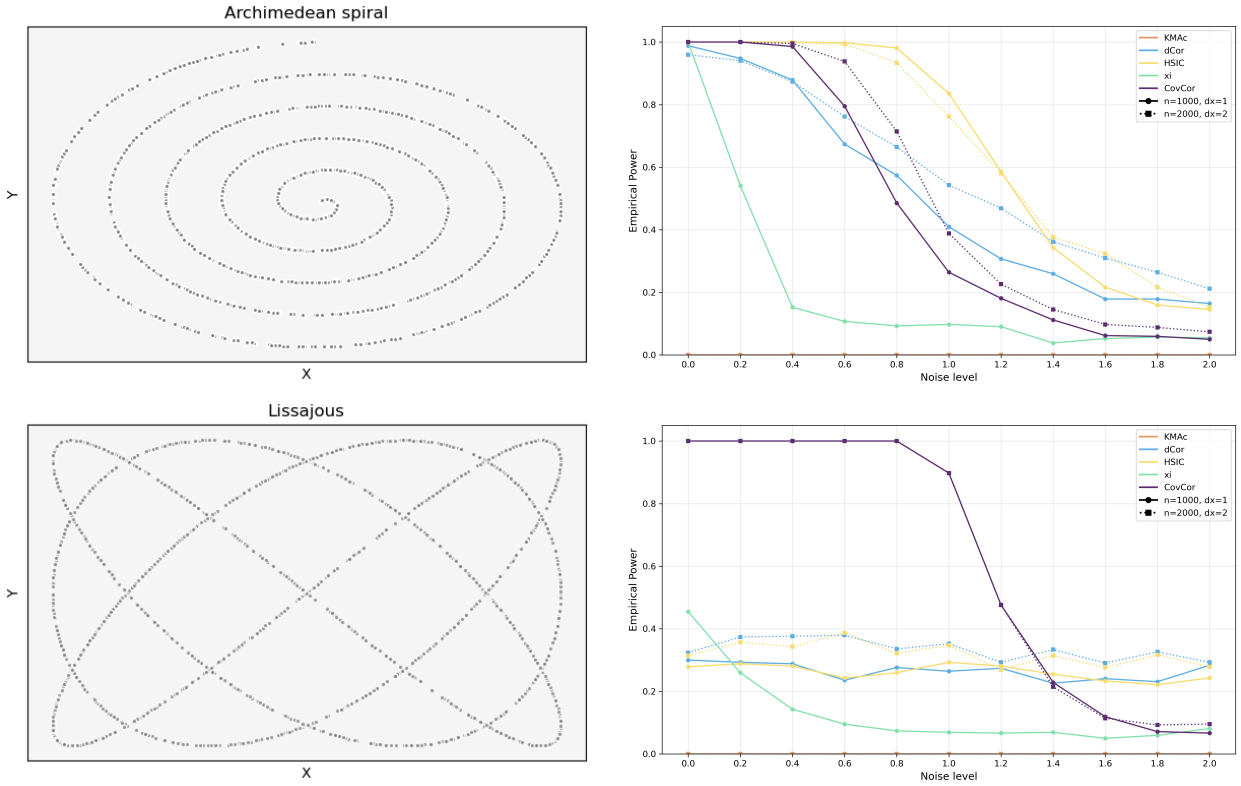


Figure 3.4: Dependency measurements applied to increasing noisy dataset with correlation based on Archimedean spiral and Lissajous curve. The left-hand side presents a noiseless pattern of the samples with $n = 2000$. The right-hand side is the power curve display the results of various methods under the case of $n = 1000, d_X = d_Y = 1$ and $n = 2000, d_X = 2, d_Y = 1$.

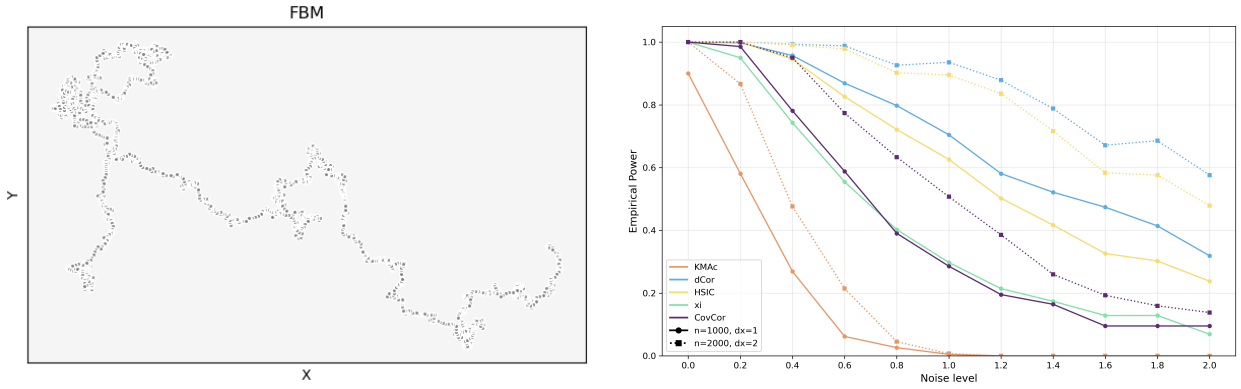


Figure 3.5: Dependency measurements applied to increasing noisy dataset with correlation model (3.19). The left-hand side presents a noiseless pattern of the samples with $n = 2000$. The right-hand side is the power curve display the results of various methods under the case of $n = 1000, d_X = d_Y = 1$ and $n = 2000, d_X = 2, d_Y = 1$.

Local dependency. Let $G_1 \sim \mathcal{N}(0, 0.25I_{d_X})$, $G_2 \sim \mathcal{N}(0, 0.25I_{d_Y})$ and $\varepsilon \sim \mathcal{N}(0, 0.0001I_{d_Y})$ are independently generated. Let $G_1^{(j)}, G_2^{(j)}, \varepsilon^{(j)}$ denote the j-th coordinate of G_1, G_2, ε, respectively.

Let

$$I^{(j)} = \begin{cases} G_2^{(j)}, & \text{if } 0 \le G_1^{(j)} \le 1 \text{ and } 0 \le G_2^{(j)} \le 1 \\ X^{(j)}\mathbf{1} + \varepsilon^{(j)}, & \text{otherwise} \end{cases}, \quad \text{for } j = 1, \dots, d_Y,$$

where $\mathbf{1} \in \mathbb{R}^{d_X \times d_Y}$ is a matrix with all entries equal to 1.. Then we generate (X, Y) through

$$X = G_1 + \varepsilon_1 \quad \text{and} \quad Y^{(j)} = I^{(j)} + \varepsilon_2^{(j)} \tag{3.20}$$

for $j = 1, \dots, d_Y$. Model (3.20) defines a local dependency (see Fig. 3.6) between X and Y. The result in Fig. 3.6 aligns with the theoretical findings in [CB20; SDH22b], the Chatterjee's coefficient does not perform well under such local alternative, while the coverage correlation coefficient performs much better. The 20-NN KMAc does not have any power in this case.

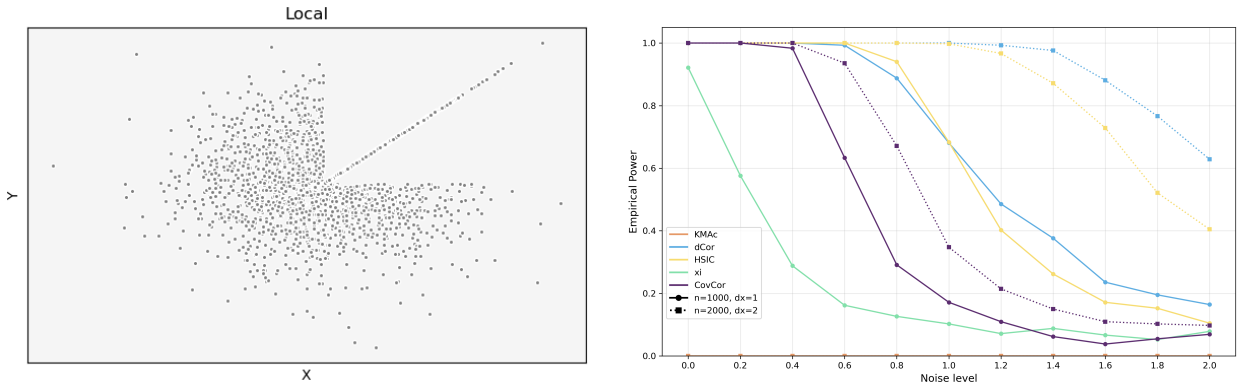


Figure 3.6: Dependency measurements applied to increasing noisy dataset with correlation model (3.20). The left-hand side presents a noiseless pattern of the samples with $n = 2000$. The right-hand side is the power curve display the results of various methods under the case of $n = 1000, d_X = d_Y = 1$ and $n = 2000, d_X = 2, d_Y = 1$.

Sinusoid. Let $U \sim \text{Unif}([-1, 1]^{d_X})$ and $U^{(i)}$ denote the i-th coordinate of it. Then we generate (X, Y) by letting

$$X^{(i)} = U^{(i)}, \quad \text{for } i = 1, \dots, d_X,$$
$$Y^{(j)} = \cos\left(8\pi X^{(j)}\right) + 1.2\lambda\varepsilon_2^{(j)}, \quad \text{for } j = 1, \dots, d_Y.$$

This example is adopted from Chatterjee [Cha21] and Shi, Drton, and Han [SDH22b], demonstrating the strong performance of Chatterjee's correlation coefficient. Fig. 3.7 implies that the coverage correlation coefficient show a comparable performance with Chatterjee's coefficient, while the dCor and HSIC have less power. The 20-NN based KMAc shows no statistical power in this context.

3.4 Proofs

In this section, we present the proofs for all the theoretical results in Section 3.2.

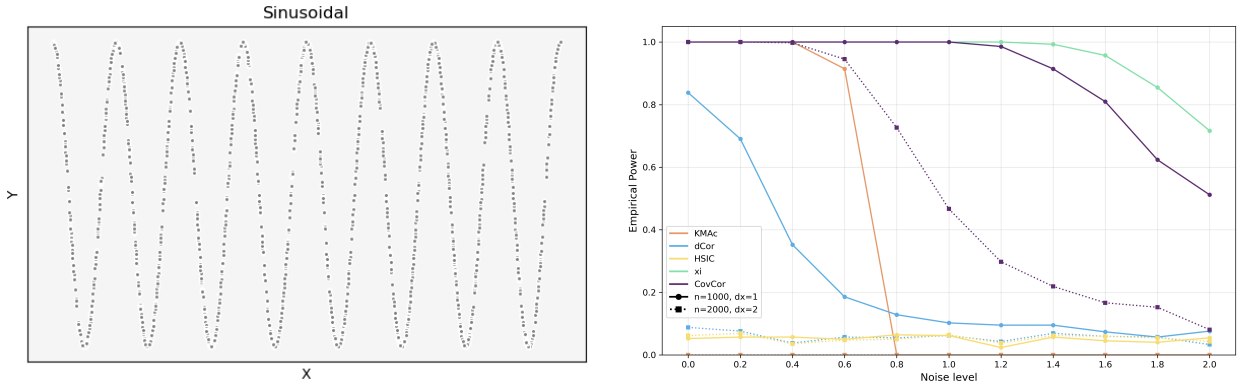


Figure 3.7: Dependency measurements applied to increasing noisy dataset with correlation model (3.15) and (3.16). The left-hand side presents a noiseless pattern of the samples with $n = 2000$. The right-hand side is the power curve display the results of various methods under the case of $n = 1000, d_X = d_Y = 1$ and $n = 2000, d_X = 2, d_Y = 1$.

3.4.1 Proof of Theorem 3.1

To establish the population limit of the empirical coverage correlation coefficient, we will first prove that the desired result holds true when the joint distribution $P^{(X,Y)}$ has a blockwise constant density (as defined in Proposition 3.3) and then use an approximation argument to extend the result to arbitrary continuous distributions.

Proposition 3.3. *For $K, L \in \mathbb{N}$, fix $0 < a_1 < \cdots < a_{K-1} < 1$ and $0 < b_1 < \cdots < b_{L-1} < 1$. Write $a_0 = b_0 = 0$ and $a_K = b_L = 1$. Define a distribution on $[0,1]^2$ as $P^{(X,Y)} := \sum_{k=1}^{K}\sum_{\ell=1}^{L} p_{k,\ell}\mathrm{Unif}\big((a_{k-1}, a_k] \times (b_{\ell-1}, b_\ell]\big)$, where $0 < p_{k,\ell} < 1$ satisfies $\sum_{k\in[K],\ell\in[L]} p_{k,\ell} = 1$. Then for $\delta > 0$ and $V^{\star}_{n,\delta}$ defined as (3.6), given $(X_1, Y_1), \ldots, (X_n, Y_n) \overset{\text{iid}}{\sim} P^{(X,Y)}$, if $n(2\delta)^2 \to \omega \in (0, \infty)$, we have*

$$V^{\star}_{n,\delta} \xrightarrow{\text{p}} D_g(P^{(X,Y)} \,\|\, P^X \otimes P^Y), \quad \text{as } n \to \infty, \tag{3.21}$$

where $g(x) = e^{-\omega x}$ and $\star \in \{\mathrm{Rand}, \mathrm{Reg}\}$.

Proof. For each $k \in [K]$ and $\ell \in [L]$, define $\mathcal{I}_k = \{i \in [n] : a_{k-1} < X_i \le a_k\}$ and $\mathcal{J}_\ell = \{j \in [n] : b_{\ell-1} < Y_j \le b_\ell\}$. Let $N_{k,+} := |\mathcal{I}_k|$, $N_{+,\ell} := |\mathcal{J}_\ell|$ and $N_{k,\ell} := |\mathcal{I}_k \cap \mathcal{J}_\ell|$. We also denote $S_k := \sum_{t=0}^{k} N_{t,+}$, $T_\ell := \sum_{t=0}^{\ell} N_{+,t}$ for $k \in [K]$ and $\ell \in [L]$. By the strong law of large number, let Ω_1 be the almost sure event that $N_{k,\ell} \to np_{k,\ell}$ for all $k \in [K]$ and $\ell \in [L]$. Now we prove the proposition under the cases of regular grid and random uniform samples separately.

(I) When $\star = \mathrm{Rand}$. Given $(U_1, V_1), \ldots, (U_n, V_n) \overset{\text{iid}}{\sim} \mathrm{Unif}([0,1]^2)$, $\widehat{R}^{X,\mathrm{Rand}}$ and $\widehat{R}^{Y,\mathrm{Rand}}$ are the empirical rank maps constructed as in (3.3). We first fix $k \in [K]$ and $\ell \in [L]$. Observe that

$$\{\widehat{R}^{X,\mathrm{Rand}}(X_i) : i \in \mathcal{I}_k\} = \{U_{(S_{k-1}+1)}, \ldots, U_{(S_k)}\}$$
$$\{\widehat{R}^{Y,\mathrm{Rand}}(Y_i) : i \in \mathcal{J}_\ell\} = \{V_{(T_{\ell-1}+1)}, \ldots, V_{(T_\ell)}\}.$$

Then let $\mathcal{F}_{k,\ell}$ be the σ-algebra generated by $U_{(S_{k-1})}$, $U_{(S_k+1)}$, and $V_{(T_{\ell-1})}$, $V_{(T_\ell+1)}$ and $(N_{k,\ell} : k \in$

$[K], \ell \in [L]$). By David and Nagaraja [DN04, Theorem 2.5], we have

$$(\widehat{R}^{X,\mathrm{Rand}}(X_i) : i \in \mathcal{I}_k) \mid \mathcal{F}_{k,\ell} \overset{\text{iid}}{\sim} \mathrm{Unif}[U_{(S_{k-1})}, U_{(S_k+1)}],$$
$$(\widehat{R}^{Y,\mathrm{Rand}}(Y_j) : j \in \mathcal{J}_\ell) \mid \mathcal{F}_{k,\ell} \overset{\text{iid}}{\sim} \mathrm{Unif}[V_{(T_{\ell-1})}, V_{(T_\ell+1)}],$$

where we adopt the convention that $U_{(0)} = U_{(S_0)} = 0$, $U_{(n+1)} = U_{(S_K+1)} = 1$, $V_{(0)} = V_{(T_0)} = 0$ and $V_{(n+1)} = V_{(T_L+1)} = 1$. Furthermore, since

$$\big((X_i, Y_i) : i \in \mathcal{I}_k \cap \mathcal{J}_\ell\big) \mid N_{k,\ell} \overset{\text{iid}}{\sim} \mathrm{Unif}(a_{k-1}, a_k] \otimes \mathrm{Unif}(b_{\ell-1}, b_\ell], \tag{3.22}$$

we have

$$\big((\widehat{R}^{X,\mathrm{Rand}}(X_i), \widehat{R}^{Y,\mathrm{Rand}}(Y_i)) : i \in \mathcal{I}_k \cap \mathcal{J}_\ell\big) \mid \mathcal{F}_{k,\ell} \overset{\text{iid}}{\sim} \mathrm{Unif}[U_{(S_{k-1})}, U_{(S_k+1)}] \otimes \mathrm{Unif}[V_{(T_{\ell-1})}, V_{(T_\ell+1)}].$$

Write $W_{k,\ell} := \mathrm{vol}\big([U_{(S_{k-1})}, U_{(S_k+1)}] \times [V_{(T_{\ell-1})}, V_{(T_\ell+1)}] \setminus \cup_{i\in[n]} B^2_\infty(\widehat{R}^{\mathrm{Rand}}_i, \delta)\big)$ and $\bar{W}_{k,\ell} := \mathrm{vol}\big([U_{(S_{k-1})}, U_{(S_k+1)}] \times [V_{(T_{\ell-1})}, V_{(T_\ell+1)}]\big)$. As $n \to \infty$, the contribution of the covered area by points near the boundary of any square is negligible, hence by $n(2\delta)^d \to \omega$ and Lemma 3.4, conditional on $\mathcal{F}_{k,\ell}$, and Ω_1, we have

$$W_{k,\ell} = (1 + o_p(1)) e^{-\omega p_{k,\ell}/\bar{W}_{k,\ell}} \bar{W}_{k,\ell}.$$

Let Ω_2 be the event that $U_{(S_k)} \to \sum_{t\in[k]} \sum_{\ell\in[L]} p_{t,\ell}$, $U_{(S_k+1)} \to \sum_{t\in[k]} \sum_{\ell\in[L]} p_{t,\ell}$, $V_{(T_\ell)} \to \sum_{t\in[\ell]} \sum_{k\in[K]} p_{k,t}$ and $V_{(T_\ell+1)} \to \sum_{t\in[\ell]} \sum_{k\in[K]} p_{k,t}$ as $n \to \infty$. By the law of large numbers and the limiting behaviour of uniform order statistics, we have $\mathbb{P}(\Omega_2) = 1$. On $\Omega_1 \cap \Omega_2$, we have $\bar{W}_{k,\ell} \to p_{k,+}p_{+,\ell}$, where $p_{k,+} := \sum_{t\in[L]} p_{k,t}$ and $p_{+,\ell} := \sum_{r\in[K]} p_{r,\ell}$. Hence,

$$W_{k,\ell} = (1 + o_p(1)) e^{-\omega p_{k,\ell}/(p_{k,+}p_{+,\ell})} p_{k,+}p_{+,\ell}.$$

The marginal distributions of X_i and Y_i are $P^X = \sum_{k=1}^K p_{k,+}\mathrm{Unif}(a_{k-1}, a_k]$ and $P^Y = \sum_{\ell=1}^L p_{+,\ell}\mathrm{Unif}(b_{\ell-1}, b_\ell]$, so for $g(x) = e^{-\omega x}$, we have

$$\begin{aligned} D_g(P^{(X,Y)} \,\|\, P^X \otimes P^Y) &= \sum_{k\in[K]} \sum_{\ell\in[L]} (a_k - a_{k-1})(b_k - b_{k-1}) e^{-\omega p_{k,\ell}/(p_{k,+}p_{+,\ell})} \frac{p_{k,+}p_{+,\ell}}{(a_k - a_{k-1})(b_k - b_{k-1})} \\ &= \sum_{k\in[K]} \sum_{\ell\in[L]} e^{-\omega p_{k,\ell}/(p_{k,+}p_{+,\ell})} p_{k,+}p_{+,\ell}. \end{aligned} \tag{3.23}$$

Hence,

$$V_{n,\delta} \le \sum_{k\in[K],\ell\in[L]} W_{k,\ell} = (1 + o_p(1)) D_g(P^{(X,Y)} \,\|\, P^X \otimes P^Y). \tag{3.24}$$

On the other hand, since $\sum_{k\in[K],\ell\in[L]} \bar{W}_{k,\ell} = 1 + o_p(1)$, we have

$$V_{n,\delta} \ge 1 - \sum_{k\in[K],\ell\in[L]} (\bar{W}_{k,\ell} - W_{k,\ell}) = (1 + o_p(1)) D_g(P^{(X,Y)} \,\|\, P^X \otimes P^Y). \tag{3.25}$$

The desired result follows by combining the above two inequalities.

(II) When $\star = \mathrm{Reg}$. Given $\mathcal{U} = \mathcal{V} = \{i/(n+1) : i \in [n]\}$, and $\widehat{R}^{X,\mathrm{Reg}}$ and $\widehat{R}^{Y,\mathrm{Reg}}$ are the empirical rank maps constructed as in (3.3). For each fixed $k \in [K]$ and $\ell \in [L]$, we write $\mathcal{R}_k^X = \{\widehat{R}^{X,\mathrm{Reg}}(X_i) : i \in \mathcal{I}_k\}$ and $\mathcal{R}_\ell^Y = \{\widehat{R}^{Y,\mathrm{Reg}}(Y_j) : j \in \mathcal{J}_\ell\}$. Then we have

$$\mathcal{R}_k^X = \{(S_{k-1}+1)/n, \ldots, S_k/n\} \quad \text{and} \quad \mathcal{R}_\ell^Y = \{(T_{\ell-1}+1)/n, \ldots, T_\ell/n\}.$$

Therefore, let $\mathcal{G}_{k,\ell}$ be the σ-algebra generated by $S_{k-1}, S_k, T_{\ell-1}, T_\ell$ and $(N_{k,\ell})_{k\in[K],\ell\in[L]}$, by (3.22), we have

$$\{\widehat{R}^{X,\mathrm{Reg}}(X_i) : i \in \mathcal{I}_k \cap \mathcal{J}_\ell\} | \mathcal{G}_{k,\ell} \sim \mathrm{Unif}\left(\binom{\mathcal{R}_k^X}{N_{k,\ell}}\right),$$
$$\{\widehat{R}^{Y,\mathrm{Reg}}(Y_j) : j \in \mathcal{I}_k \cap \mathcal{J}_\ell\} | \mathcal{G}_{k,\ell} \sim \mathrm{Unif}\left(\binom{\mathcal{R}_\ell^Y}{N_{k,\ell}}\right).$$

Let Ω_3 be the almost sure event such that $N_{k,+}/n \to p_{k,+}$ and $N_{+,\ell}/n \to p_{+,\ell}$ for all $k \in [K]$ and $\ell \in [L]$. Define $Q_{k,\ell} := \mathrm{vol}\big([S_{k-1}/n, (S_k+1)/n] \times [T_{\ell-1}/n, (T_\ell+1)/n] \setminus \cup_{i=1}^n B_\infty^2\big(\widehat{R}_i^{\mathrm{Reg}}, \delta\big)\big)$ and $\bar{Q}_{k,\ell} := \mathrm{vol}([S_{k-1}/n, (S_k+1)/n] \times [T_{\ell-1}/n, (T_\ell+1)/n])$. Then on the event $\Omega_1 \cap \Omega_3$, we have $N_{k,\ell}/N_{k,+} \to p_{k,\ell}/p_{k,+}$ and $N_{k,\ell}/N_{k,+} \to p_{k,\ell}/p_{+,\ell}$, thus by $n(2\delta)^2 \to \omega$, Lemma 3.5 implies that

$$\mathbb{E}(Q_{k,\ell}|\mathcal{G}_{k,\ell}) \to e^{-\omega p_{k,\ell}/\bar{Q}_{k,\ell}}\bar{Q}_{k,\ell}, \quad \text{as } n \to \infty. \tag{3.26}$$

Then on the event $\Omega_1 \cap \Omega_3$, by the concentration inequality given in Proposition 3.4, we obtain that

$$\begin{aligned}\mathbb{E}\Big(\big(Q_{k,\ell} - \mathbb{E}(Q_{k,\ell}|\mathcal{G}_{k,\ell})\big)^2 \mid \mathcal{G}_{k,\ell}\Big) &= \int_0^{+\infty} \mathbb{P}\Big(\big|Q_{k,\ell} - \mathbb{E}(Q_{k,\ell}|\mathcal{G}_{k,\ell})\big| \geq \sqrt{t} \mid \mathcal{G}_{k,\ell}\Big)\,\mathrm{d}t \\ &\leq C\frac{n}{(N_{k,+}+1)^2(N_{+,\ell}+1)^2} \to 0, \quad \text{as } n \to \infty. \end{aligned} \tag{3.27}$$

Combining (3.26) with (3.27), we have on the event $\Omega_1 \cap \Omega_3$, and conditional on $\mathcal{G}_{k,\ell}$

$$Q_{k,\ell} = (1 + o_p(1))e^{-\omega p_{k,\ell}/(p_{k,+}p_{+,\ell})}p_{k,+}p_{+,\ell}, \quad \text{as } n \to \infty.$$

Finally, by notice that $1 - \sum_{k\in[K],\ell\in[L]}(Q_{k,\ell} - \bar{Q}_{k,\ell}) \leq V_{n,\delta} \leq \sum_{k\in[K],\ell\in[L]} Q_{k,\ell}$, the result follows by the same sandwich argument from (3.23)-(3.25).

□

We also need the following the following lemma (see Section 3.6.1 for a proof) to construct an approximation argument.

Lemma 3.1. *Let f be a Lebesgue density on $\mathbb{R}^d$. Suppose $f^+ \geq f$ on $\mathbb{R}^d$ such that $\int f^+ = \alpha > 1$. For any $\lambda > 0$, let $N \sim \mathrm{Poi}(\lambda\alpha)$ and let $M \mid N \sim \mathrm{Bin}(N, 1/\alpha)$. Draw independent samples $X_1, \ldots, X_M \mid M \overset{\mathrm{iid}}{\sim} f$ and $X_{M+1}, \ldots, X_N \mid N - M \overset{\mathrm{iid}}{\sim} (f^+ - f)/(\alpha - 1)$. Let σ be a uniform random permutation of $\{1, \ldots, N\}$ conditional on N. Then*

$$X_{\sigma(1)}, \ldots, X_{\sigma(N)} \mid N \overset{\mathrm{iid}}{\sim} f^+/\alpha.$$

Now, we are ready to prove Theorem 3.1.

Proof of Theorem 3.1. Let $(X,Y) \sim P^{(X,Y)}$ and $f_{X,Y}$ be the density of $P^{(X,Y)}$ with respect to the Lebesgue measure. By replacing (X_i,Y_i) with $(F_X(X_i),F_Y(Y_i))$, where F_X and F_Y are the marginal cumulative distribution functions of (X_i,Y_i), we may assume without loss of generality that $f_{X,Y}$ is supported on $[0,1]^2$ with marginal distributions f_X and f_Y uniform on $[0,1]$.

For any $K,L \in \mathbb{N}$, we define $a_k = k/K$ for $k \in \{0,\dots,K\}$ and $b_\ell = \ell/L$ for $\ell \in \{0,\dots,L\}$. For any given $\epsilon \in (0,0.1)$, since $f_{X,Y}$ is continuous, we choose K,L sufficiently large such that

$$\max_{k\in[K],\ell\in[L]} \left\{ \sup_{x\in(a_{k-1},a_k],y\in(b_{\ell-1},b_\ell]} f_{X,Y}(x,y) - \inf_{x\in(a_{k-1},a_k],y\in(b_{\ell-1},b_\ell]} f_{X,Y}(x,y) \right\} \le \epsilon. \tag{3.28}$$

For each $k \in [K]$ and $\ell \in [L]$, let $f^-_{k,\ell} := \inf_{x\in(a_{k-1},a_k],y\in(b_{\ell-1},b_\ell]} f_{X,Y}(x,y)$ and set $f^-_{X,Y}(x,y) = f^-_{k,\ell}$ if $x \in (a_{k-1},a_k]$ and $y \in (b_{\ell-1},b_\ell]$. We also define

$$p^-_{k,\ell} := \frac{\int_{x\in(a_{k-1},a_k],y\in(b_{\ell-1},b_\ell]} f^-_{X,Y}(x,y)\,\mathrm{d}(x,y)}{\int_{(x,y)\in[0,1]^2} f^-_{X,Y}(x,y)\,\mathrm{d}(x,y)} = \frac{f^-_{k,\ell}}{\sum_{r\in[K]}\sum_{s\in[L]} f^-_{r,s}}.$$

By (3.28), $\beta := \int_{(x,y)\in[0,1]^2} f^-_{X,Y}(x,y)\,\mathrm{d}(x,y) \ge 1-\epsilon$. Let $H_1,\dots,H_n \overset{\text{iid}}{\sim} \mathrm{Unif}[0,1]$ independent of all other randomness in the problem. Then the rejection sample $\{(X_i,Y_i) : H_i \le f^-(X_i,Y_i)/f_{X,Y}(X_i,Y_i)\}$ has cardinality $N^- \sim \mathrm{Bin}(n,\beta)$ and conditionally on N^-, the rejection sample are independent and identically distributed from a joint distribution of the form

$$P^{(X,Y),-} := \sum_{k\in[K]}\sum_{\ell\in[L]} p^-_{k,\ell}\mathrm{Unif}\big((a_{k-1},a_k]\times(b_{\ell-1},b_\ell]\big).$$

Let $P^{X,-}$ and $P^{Y,-}$ be the corresponding marginals of $P^{(X,Y),-}$. Applying Proposition 3.3 conditionally on N^-, and using the fact that $N^-/n \xrightarrow{\text{p}} \beta$, we have the following holds for $\star \in \{\mathrm{Reg},\mathrm{Rand}\}$ with $g^-(x) = e^{-\beta x}$:

$$\begin{aligned} V^\star_n = V^\star_{n,1/(2\sqrt{n})} \le V^\star_{N^-,1/(2\sqrt{n})} &\xrightarrow{\text{p}} D_{g^-}(P^{(X,Y),-} \,\|\, P^{X,-}\otimes P^{Y,-}) \\ &= \sum_{k=1}^K\sum_{\ell=1}^L e^{-\beta p^-_{k,\ell}/(p^-_{k,+}p^-_{+,\ell})} p^-_{k,+}p^-_{+,\ell}, \end{aligned} \tag{3.29}$$

where the final equality follows from a similar calculation as in (3.23) with $p^-_{k,+} := \sum_{\ell\in[L]} p^-_{k,\ell}$ and $p^-_{+,\ell} := \sum_{k\in[K]} p^-_{k,\ell}$. Defining $f^-_X(x) := \int_{y\in[0,1]} f^-_{X,Y}(x,y)\,\mathrm{d}y$ and $f^-_Y(y) := \int_{x\in[0,1]} f^-_{X,Y}(x,y)\,\mathrm{d}x$, then for any $(x,y) \in (a_{k-1},a_k]\times(b_{\ell-1},b_\ell]$, we have by (3.28) that

$$\begin{aligned} KLp^-_{k,\ell} &= \frac{f^-_{X,Y}(x,y)}{\int_{(s,t)\in[0,1]^2} f^-_{X,Y}(s,t)\,\mathrm{d}(s,t)} \ge f_{X,Y}(x,y) - \epsilon, \\ Kp^-_{k,+} &= \frac{f^-_X(x)}{\int_{(s,t)\in[0,1]^2} f^-_{X,Y}(s,t)\,\mathrm{d}(s,t)} \le \frac{1}{1-\epsilon}, \\ Lp^-_{+,\ell} &= \frac{f^-_Y(y)}{\int_{(s,t)\in[0,1]^2} f^-_{X,Y}(s,t)\,\mathrm{d}(s,t)} \le \frac{1}{1-\epsilon}, \end{aligned}$$

where we used the fact that the marginal densities $f_X(x) = f_Y(y) = 1$. Thus, we have

$$\begin{aligned}\sum_{k=1}^{K}\sum_{\ell=1}^{L} e^{-\beta p^-_{k,\ell}/(p^-_{k,+}p^-_{+,\ell})} p^-_{k,+}p^-_{+,\ell} &\le \frac{1}{(1-\epsilon)^2}\int_{(x,y)\in[0,1]^2} e^{-(f_{X,Y}(x,y)-\epsilon)(1-\epsilon)^3}\,\mathrm{d}(x,y)\\ &\le \frac{e^{\epsilon(1+3\|f\|_\infty)}}{(1-\epsilon)^2}\int_{(x,y)\in[0,1]^2} e^{-f_{X,Y}(x,y)}\,\mathrm{d}(x,y)\\ &= \frac{e^{\epsilon(1+3\|f\|_\infty)}}{(1-\epsilon)^2} D_g(P^{(X,Y)} \,\|\, P^X \otimes P^Y). \end{aligned} \tag{3.30}$$

Now, we turn to an asymptotic stochastic lower bound of $V_n^\star$. To this end, we define $f^+_{k,\ell} := \sup_{x\in(a_{k-1},a_k],y\in(b_{\ell-1},b_\ell]} f_{X,Y}(x,y)$ and set $f^+_{X,Y}(x,y) = f^+_{k,\ell}$ if $x \in (a_{k-1},a_k]$ and $y \in (b_{\ell-1},b_\ell]$. We also define

$$p^+_{k,\ell} := \frac{\int_{x\in(a_{k-1},a_k],y\in(b_{\ell-1},b_\ell]} f^+_{X,Y}(x,y)\,\mathrm{d}(x,y)}{\int_{(x,y)\in[0,1]^2} f^+_{X,Y}(x,y)\,\mathrm{d}(x,y)} = \frac{f^+_{k,\ell}}{\sum_{r\in[K]}\sum_{s\in[L]} f^+_{r,s}}.$$

Again by (3.28), $\alpha := \int_{(x,y)\in[0,1]^2} f^+_{X,Y}(x,y)\,\mathrm{d}(x,y) \le 1+\epsilon$. We may assume that $\alpha > 1$, since otherwise $f_{X,Y}$ itself has a blockwise constant structure and the desired result follows directly from Proposition 3.3. Let $N^+ \sim \mathrm{Poi}(\alpha^2 n)$ be independent of other randomness in this problem and set $M \mid N^+ \sim \mathrm{Bin}(N^+, 1/\alpha)$, then $M \sim \mathrm{Poi}(\alpha n)$. We will henceforth work on the asymptotically almost sure event that $M \ge n$. Now draw additional independent samples $X_{n+1},\dots,X_M \mid M \overset{\text{iid}}{\sim} f_{X,Y}$ and $X_{M+1},\dots,X_{N^+} \mid N^+ - M \overset{\text{iid}}{\sim} (f^+_{X,Y} - f_{X,Y})/(\alpha-1)$. We have by Lemma 3.1 that $X_1,\dots,X_{N^+}$, conditional on N^+ and after random permutation, is an i.i.d. sample from $P^{(X,Y),+} := \sum_{k\in[K],\ell\in[L]} p^+_{k,\ell}\mathrm{Unif}\big((a_{k-1},a_k]\times(b_{\ell-1},b_\ell]\big)$ with density $f^+_{X,Y}/\alpha$. Again, applying Proposition 3.3 conditionally on N^+ and using the fact that $N^+/n \xrightarrow{\text{p}} \alpha$, we have the following holds for $\star \in \{\mathrm{Reg},\mathrm{Rand}\}$ with $g^+(x) = e^{-\alpha x}$ that

$$\begin{aligned} V_n^\star = V^\star_{n,1/(2\sqrt{n})} \ge V^\star_{N^+,1/(2\sqrt{n})} &\xrightarrow{\text{p}} D_{g^+}(P^{(X,Y),+} \,\|\, P^{X,+}\otimes P^{Y,+})\\ &= \sum_{k=1}^{K}\sum_{\ell=1}^{L} e^{-\alpha p^+_{k,\ell}/(p^+_{k,+}p^+_{+,\ell})} p^+_{k,+}p^+_{+,\ell}, \end{aligned} \tag{3.31}$$

where $p^+_{k,+} := \sum_{\ell\in[L]} p^+_{k,\ell}$ and $p^+_{+,\ell} := \sum_{k\in[K]} p^+_{k,\ell}$. Defining $f^+_X(x) := \int_{y\in[0,1]} f^+_{X,Y}(x,y)$ and $f^+_Y(y) := \int_{x\in[0,1]} f^+_{X,Y}(x,y)$, then for any $(x,y) \in (a_{k-1},a_k]\times(b_{\ell-1},b_\ell]$, we have by (3.28) that

$$\begin{aligned} KLp^+_{k,\ell} &= \frac{f^+_{X,Y}(x,y)}{\int_{(s,t)\in[0,1]^2} f^+_{X,Y}(s,t)\,\mathrm{d}(s,t)} \le f_{X,Y}(x,y)+\epsilon,\\ Kp^+_{k,+} &= \frac{f^+_X(x)}{\int_{(s,t)\in[0,1]^2} f^+_{X,Y}(s,t)\,\mathrm{d}(s,t)} \ge \frac{1}{1+\epsilon},\\ Lp^+_{+,\ell} &= \frac{f^+_Y(y)}{\int_{(s,t)\in[0,1]^2} f^+_{X,Y}(s,t)\,\mathrm{d}(s,t)} \le \frac{1}{1+\epsilon}. \end{aligned}$$

Thus, we have

$$\begin{aligned}\sum_{k=1}^{K}\sum_{\ell=1}^{L} e^{-\alpha p^+_{k,\ell}/(p^+_{k,+}p^+_{+,\ell})} p^+_{k,+}p^+_{+,\ell} &\geq \frac{1}{(1+\epsilon)^2}\int_{(x,y)\in[0,1]^2} e^{-(f_{X,Y}(x,y)+\epsilon)(1+\epsilon)^3}\,\mathrm{d}(x,y)\\ &\geq \frac{e^{-\epsilon(2+4\|f\|_\infty)}}{(1+\epsilon)^2}\int_{(x,y)\in[0,1]^2} e^{-f_{X,Y}(x,y)}\,\mathrm{d}(x,y)\\ &= \frac{e^{-\epsilon(2+4\|f\|_\infty)}}{(1+\epsilon)^2} D_g(P^{(X,Y)} \,\|\, P^X \otimes P^Y), \end{aligned} \tag{3.32}$$

where we used the fact that $(f_{X,Y}(x,y)+\epsilon)(1+\epsilon)^3 \leq (f_{X,Y}(x,y)+\epsilon)(1+4\epsilon) \leq f_{X,Y}(x,y) + 2\epsilon + 4\epsilon\|f_{X,Y}\|_\infty$ for $\epsilon < 0.1$. Combining (3.29), (3.30), (3.31) and (3.32), and using the fact that ϵ can be chosen arbitrarily close to 0, we have for $\star \in \{\mathrm{Reg}, \mathrm{Rand}\}$

$$V_n^\star \xrightarrow{\mathrm{p}} D_g(P^{(X,Y)} \,\|\, P^X \otimes P^Y).$$

The desired result then follows immediatley from the definition of $\kappa_n^{X,Y;\star}$ in (3.7) and the fact that $f = (g - e^{-1})/(1 - e^{-1})$. □

3.4.2 Proof of Proposition 3.1

Proof. The first part of property (i) can be immediately obtained by the expression of $\kappa^{X,Y}$ and the definition of f-divergence. The second part of (i) follows from Lemma 3.3(i), given the fact that $f(x) = (e^{-x} - e^{-1})/(1 - e^{-1})$ is strictly convex at 1. Property (ii) is a direct application of the data processing inequality of f-divergence (see Lemma 3.3(ii)). Property (iii) follows by Lemma 3.3(iii). Finally, the symmetric property can be seen by the definition of f-divergence. □

3.4.3 Proof of Proposition 3.2

To prove Proposition 3.2, we first prove a bounded difference property of the covered area (Lemma 3.2), then the proposition can be obtained by applying the McDiarmid's inequality.

Lemma 3.2. *Let $(x_1, y_1), \ldots, (x_n, y_n)$ be n points in $\mathbb{R}^2$ and we write $z_i = (x_i, y_i)$ for $i = 1, \ldots, n$ and $z = (z_1, \ldots, z_n)$, $x = (x_1, \ldots, x_n)$, $y = (y_1, \ldots, y_n)$. Assume that $x_i \neq x_j$ and $y_i \neq y_j$ for all $i \neq j$. Let $r_i^x = n^{-1}\sum_{j=1}^n \mathbb{1}\{x_j \leq x_i\}$, $r_i^y = n^{-1}\sum_{j=1}^n \mathbb{1}\{y_j \leq y_i\}$, and $r_i = (r_i^x, r_i^y)$ for $i = 1, \ldots, n$. Let $\mathcal{I}_Q = \{i \in [n] : r_i \in Q\}$ for any fixed subset $Q \subseteq [0,1]^2$. We define a function*

$$f(z) := \frac{\mathrm{vol}\left(\bigcup_{i\in\mathcal{I}_Q} B_\infty^2\left(r_i, \frac{1}{2\sqrt{n}}\right) \cap Q\right)}{\mathrm{vol}(Q)}.$$

Let $(x_1', y_1'), \ldots, (x_n', y_n')$ be another n points in $\mathbb{R}^2$, let $z_i' = (x_i', y_i')$ for $i = 1, \ldots, n$ and $z' = (z_1', \ldots, z_n')$. For any $k \in [n]$, let $z^k \in \mathbb{R}^{2n}$ be the vector obtained by replacing z_k with z_k' in the vector z. Then we have

$$|f(z) - f(z^k)| \leq \frac{5}{n\mathrm{vol}(Q)}.$$

Proof. Note $|f(z) - f(z^k)| = 0$ if $z_k = z'_k$, thus the result holds automatically. Assume that $z_k \neq z'_k$. Define

$$\begin{aligned}
\mathcal{I}_0 &:= \{i \in [n] : (x_i - x_k)(x_i - x'_k) \geq 0 \text{ and } (y_i - y_k)(y_i - y'_k) \geq 0\} \cap \mathcal{I}_Q, \\
\mathcal{I}_1 &:= \{i \in [n] : (x_i - x_k)(x_i - x'_k) < 0 \text{ and } (y_i - y_k)(y_i - y'_k) < 0\} \cap \mathcal{I}_Q, \\
\mathcal{I}_2 &:= \{i \in [n] : (x_i - x_k)(x_i - x'_k) < 0 \text{ and } (y_i - y_k)(y_i - y'_k) \geq 0\} \cap \mathcal{I}_Q, \\
\mathcal{I}_3 &:= \{i \in [n] : (x_i - x_k)(x_i - x'_k) \geq 0 \text{ and } (y_i - y_k)(y_i - y'_k) < 0\} \cap \mathcal{I}_Q.
\end{aligned}$$

Note that when replacing z_k by z'_k in z, both $\{x_i : i \in \mathcal{I}_0\}$ and $\{y_i : i \in \mathcal{I}_0\}$ maintain their original ranks, and both $\{x_i : i \in \mathcal{I}_1\}$ and $\{y_i : i \in \mathcal{I}_1\}$ have a shift of $1/n$ in either direction. The ranks of $\{x_i : i \in \mathcal{I}_2\}$ have a similar shift of $1/n$ while the ranks of $\{y_i : i \in \mathcal{I}_2\}$ remain the same. Conversely, the ranks of $\{x_i : i \in \mathcal{I}_3\}$ remains the same while ranks of $\{y_i : i \in \mathcal{I}_3\}$ have a shift of $1/n$ in either direction. Specifically, we write $r_i^k = (r_i^{x^k}, r_i^{y^k})$ for $i = 1, \ldots, n$, where x^k is the vector obtained by replacing x_k with x'_k in vector x and y^k is obtained by replacing y_k with y'_k in vector y, then we have

$$r_i^k = \begin{cases} r_i, & \text{if } i \in \mathcal{I}_0 \\ r_i + (\pm\frac{1}{n}, \pm\frac{1}{n}), & \text{if } i \in \mathcal{I}_1 \\ r_i + (\pm\frac{1}{n}, 0), & \text{if } i \in \mathcal{I}_2 \\ r_i + (0, \pm\frac{1}{n}), & \text{if } i \in \mathcal{I}_3. \end{cases} \tag{3.33}$$

Let $\mathcal{U}_j = \bigcup_{i \in \mathcal{I}_j} B_\infty^2(r_i, 1/(2n^{-1/2}))$ and $\mathcal{U}_j^k = \bigcup_{i \in \mathcal{I}_j} B_\infty^2(r_i^k, 1/(2n^{-1/2}))$ for $j = 0, 1, 2, 3$, we have the following decompositions

$$\begin{aligned}
\mathcal{C} &:= \bigcup_{i \in \mathcal{I}_Q} B_\infty^2\left(r_i, \frac{1}{2\sqrt{n}}\right) = \left(\bigcup_{j=0}^3 \mathcal{U}_j\right) \cup B_\infty^2\left(r_k, \frac{1}{2\sqrt{n}}\right), \\
\mathcal{C}^k &:= \bigcup_{i \in \mathcal{I}_Q} B_\infty^2\left(r_i^k, \frac{1}{2\sqrt{n}}\right) = \left(\bigcup_{j=0}^3 \mathcal{U}_j^k\right) \cup B_\infty^2\left(r_k^k, \frac{1}{2\sqrt{n}}\right).
\end{aligned}$$

Note by (3.33), we have $\mathrm{vol}\big(\mathcal{U}_0 \Delta \mathcal{U}_0^k\big) = 0$, $\mathrm{vol}\big(\mathcal{U}_1 \Delta \mathcal{U}_1^k\big) \leq 2/n$ and $\mathrm{vol}\big(\mathcal{U}_j \Delta \mathcal{U}_j^k\big) \leq 1/n$ for $j = 2, 3$. Therefore, we have

$$\begin{aligned}
|f(z) - f(z')| &\leq \frac{\mathrm{vol}\Big((\mathcal{C} \Delta \mathcal{C}^k) \cap Q\Big)}{\mathrm{vol}(Q)} \\
&\leq \sum_{j=0}^3 \frac{\mathrm{vol}\big(\mathcal{U}_j \Delta \mathcal{U}_j^k\big)}{\mathrm{vol}(Q)} + \frac{\mathrm{vol}\left(B_\infty^2\left(r_k, \frac{1}{2\sqrt{n}}\right) \Delta B_\infty^2\left(r_k^k, \frac{1}{2\sqrt{n}}\right)\right)}{\mathrm{vol}(Q)} \leq \frac{5}{n\mathrm{vol}(Q)},
\end{aligned}$$

as desired. □

Based on Lemma 3.2, we have the following concentration inequality, which is a stronger version of Proposition 3.2.

Proposition 3.4. *Let* $(X_1, Y_1), \ldots, (X_n, Y_n) \overset{\text{iid}}{\sim} P^{(X,Y)} \in \mathcal{P}(\mathbb{R}^2)$ *and we write* $Z_i = (X_i, Y_i)$ *for* $i = 1, \ldots, n$. *Let* $\widehat{R}_i^{\text{Reg}} = (\widehat{R}^{X,\text{Reg}}(X_i), \widehat{R}^{Y,\text{Reg}}(Y_i))$ *be the corresponding rank transformations under the regular grid reference distribution (see (3.4)). Let* $Q \subseteq [0,1]^2$, *and write* $\mathcal{I}_Q = \{i \in [n] : \widehat{R}_i^{\text{Reg}} \in Q\}$ *and* $N = |\mathcal{I}_Q|$. *We define*

$$V := V(Z_1, \ldots, Z_n) = \frac{\text{vol}\big(Q \setminus \cup_{i \in \mathcal{I}_Q} B_\infty^2(\widehat{R}_i^{\text{Reg}}, \frac{1}{2\sqrt{n}})\big)}{\text{vol}(Q)}.$$

Then we have for any $t \geq 0$ *and* $v \in [n]$,

$$\mathbb{P}\Big(\big|V - \mathbb{E}(V|N = v)\big| \geq t \;\Big|\; N = v\Big) \leq 2e^{-2nt^2 \text{vol}^2(Q)/25}.$$

Proof. Given $N = v$, Lemma 3.2 implies that V is a function satisfies the bounded difference property with parameters $\big(5/(n\text{vol}(Q)), \ldots, 5/(n\text{vol}(Q))\big)$. Thus by the McDiarmid inequality [see, e.g. Wai19, Corollary 2.21], we have

$$\mathbb{P}\Big(\big|V - \mathbb{E}(V|N = v)\big| \geq t \;\Big|\; N = v\Big) \leq 2e^{-2nt^2 \text{vol}^2(Q)/25}, \quad \text{for all } t \geq 0,$$

as desired. □

3.4.4 Proof of Theorem 3.2

Proof. Theorem 3.2 is a direct corollary of Lemma 3.6 by letting $P = Q = [0,1]^d$ and $\delta = 1/2n^{-1/d}$. □

3.4.5 Proof of Theorem 3.3

We first present the following important Lemma follows by the Slutsky's Theorem (see Section 3.6.2 for the proof).

Lemma 3.3. *Let* $(M_n)_n$ *and* $(L)_n$ *be sequences of random variables such that* $M_n \xrightarrow{\text{d}} N(\mu, \alpha^2)$ *and* $L_n \xrightarrow{\text{p}} \beta^2$. *Let* $\mathcal{F}_n$ *be the sigma-algebra generated by* $(M_i)_{i \leq n}$ *and* $(L_i)_{i \leq n}$. *If* $(X_n)_n$ *is a sequence of random variables such that*

$$\mathbb{E} \sup_{-\infty < x < \infty} \left| \mathbb{P}\left(\frac{X_n - M_n}{\sqrt{L_n}} \leq x \;\middle|\; \mathcal{F}_n \right) - \Phi(x) \right| \to 0, \quad n \to +\infty. \tag{3.34}$$

Then we have $X_n \xrightarrow{\text{d}} N(\mu, \alpha^2 + \beta^2)$.

We now proceed to present the proof of Theorem 3.3, which heavily depends on a set of techniques from the theory of coverage process. We guide the readers to [Hal85] for further details.

Proof. By the definition of $\kappa_n^{X,Y;\text{Rand}}$ in (3.7) and the fact that $\mathbb{E}(V_n^{\text{Rand}}) = (1 - 1/n)^n \to e^{-1}$ (see Lemma 3.6), it suffices to prove that

$$n^{1/2}(V_n^{\text{Rand}} - \mathbb{E}V_n^{\text{Rand}}) \xrightarrow{\text{d}} \mathcal{N}\left(0, e^{-2} \sum_{i=2}^{\infty} \frac{1}{i!} \left(\frac{2}{i+1}\right)^d\right).$$

For any $n \in \mathbb{N}$, let $\gamma := \frac{1}{2n^{1/d}}$. Then for any $\lambda \in \mathbb{Q}_{>0}$, there exists a sufficient large n such that $L = \{(\lambda+2)\gamma\}^{-d}$ is an integer. We can then partition $[0,1]^d = \bigcup_{\ell=1}^{L} \mathcal{P}_\ell$ into L d-dimensional small cubes where each small cube $\mathcal{P}_\ell$ has edge length $(\lambda+2)\gamma$, volume $p := (\lambda/2+1)^d n^{-1}$ and that $\mathcal{P}_\ell \cap \mathcal{P}_k$ has Lebesgue measure 0 for all $k \neq \ell$. Inside each $\mathcal{P}_\ell$, we construct a concentric subcube $\mathcal{Q}_\ell$ of edge length $\lambda\gamma$. For any $\ell \in [L]$, define $\mathcal{I}_\ell := \{i \in [n] : \widehat{R}_i^{\mathrm{Rand}} \in \mathcal{P}_\ell\}$ and let $N_\ell := |\mathcal{I}_\ell|$. We write $V_{n,\ell}^{\mathrm{Rand}} := \mathrm{vol}\big(\mathcal{Q}_\ell \setminus \bigcup_{i\in\mathcal{I}_\ell} B_\infty^d(\widehat{R}_i^{\mathrm{Rand}}, \gamma)\big) = \mathrm{vol}\big(\mathcal{Q}_\ell \setminus \bigcup_{i\in\mathcal{I}_\ell} B_\infty(\widehat{R}_i^{\mathrm{Rand}}, \gamma; \mathcal{P}_\ell)\big)$, where the second equality holds since each point in $\mathcal{Q}_\ell$ is at least γ away from the boundary of $\mathcal{P}_\ell$. We will first establish the limiting distribution of $\sum_{\ell=1}^{L} V_{n,\ell}^{\mathrm{Rand}}$ and then control $V_n^{\mathrm{Rand}} - \sum_{\ell=1}^{L} V_{n,\ell}^{\mathrm{Rand}}$.

Recall the definition of $\pi^{X,\mathrm{Rand}}$ and $\pi^{Y,\mathrm{Rand}}$ in (3.2). Under H_0, $\pi^{X,\mathrm{Rand}}$ and $\pi^{Y,\mathrm{Rand}}$ are independent, which implies that $\widehat{R}_i^{\mathrm{Rand}} \overset{\mathrm{iid}}{\sim} \mathrm{Unif}([0,1]^d)$ for $i = 1, \ldots, n$. Then the conditional distribution of $\widehat{R}_i^{\mathrm{Rand}} \mid i \in \mathcal{I}_\ell$, denoted as P_ℓ^R, is uniform in $\mathcal{P}_\ell$. For any $x_1, x_2 \in \mathcal{Q}_1$, define $C(x_1, x_2) := B_\infty(x_1, \gamma; \mathcal{P}_1) \cap B_\infty(x_2, \gamma; \mathcal{P}_1)$ and

$$
\begin{aligned}
v(x_1) &:= 1 - P_1^R\big(B_\infty(x_1, \gamma; \mathcal{P}_1)\big) = 1 - \frac{1}{np}, \\
u(x_1, x_2) &:= 1 - P_1^R\big(B_\infty(x_1, \gamma; \mathcal{P}_1)\big) - P_1^R\big(B_\infty(x_2, \gamma; \mathcal{P}_1)\big) + P_1^R\big(C(x_1, x_2)\big) \\
&= 1 - \frac{2}{np} + \frac{\mathrm{vol}(C(x_1, x_2))}{p}.
\end{aligned}
$$

We first fix an $\ell \in [L]$. By Lemma 3.6 we have for $X_{1,\ell}, X_{2,\ell} \overset{\mathrm{iid}}{\sim} \mathrm{Unif}(\mathcal{Q}_\ell)$ that

$$
\begin{aligned}
\mathbb{E}\big[\mathrm{Var}\big(V_{n,\ell}^{\mathrm{Rand}} \mid N_\ell\big)\big] &= \mathbb{E}\Big[\Big(\mathbb{E}\big((V_{n,\ell}^{\mathrm{Rand}})^2 \mid N_\ell\big) - \big\{\mathbb{E}\big(V_{n,\ell}^{\mathrm{Rand}} \mid N_\ell\big)\big\}^2\Big)\Big] \\
&= \mathrm{vol}^2(\mathcal{Q}_1)\, \mathbb{E}\Big[\mathbb{E}\Big(\big\{u(X_{1,\ell}, X_{2,\ell})\big\}^{N_\ell} \,\Big|\, N_\ell\Big) - \mathbb{E}\Big(\big\{v(X_{1,\ell})v(X_{2,\ell})\big\}^{N_\ell} \,\Big|\, N_\ell\Big)\Big] \\
&= \int_{\mathcal{Q}_1^2} \Big[\Big\{p\big(u(x_1, x_2) - 1\big) + 1\Big\}^n - \Big\{p\big(v(x_1)v(x_2) - 1\big) + 1\Big\}^n\Big]\, \mathrm{d}x_1 \mathrm{d}x_2 \\
&= \int_{\mathcal{Q}_1^2} \big(1 + o(1)\big)\big\{e^{np(u(x_1,x_2)-1)} - e^{np(v(x_1)v(x_2)-1)}\big\}\, \mathrm{d}x_1 \mathrm{d}x_2 \\
&= \int_{\mathcal{Q}_1^2} (1 + o(1))\big\{e^{-2+n\mathrm{vol}(C(x_1,x_2))} - e^{-2+(np)^{-1}}\big\}\, \mathrm{d}x_1 \mathrm{d}x_2 \\
&= (1 + o(1))(\lambda\gamma)^{2d} \int_{[0,1]^{2d}} \big\{e^{-2+(\lambda/2)^d \mathrm{vol}(\widetilde{C}(y_1,y_2))} - e^{-2+(np)^{-1}}\big\}\, \mathrm{d}y_1 \mathrm{d}y_2
\end{aligned}
\tag{3.35}
$$

where we used the fact that $\mathbb{E}\, a^{N_\ell} = \big(1 + p(a-1)\big)^n$ for $N_\ell \sim \mathrm{Bin}(n, p)$ and $a > 0$ in the third equality, Lemma 3.9 in the fourth step and we defined $\widetilde{C}(y_1, y_2) := \bigcap_{i=1}^{2} B_\infty(y_i, 1/\lambda; (-1/\lambda, 1 + 1/\lambda)^d)$ for $(y_1, y_2) \in (0,1)^{2d}$ in the final step. Observe that the integral in the final expression of (3.35) is a constant depending only on λ and d (and does not depend on n). Hence, summing (3.35) over $\ell \in [L]$, we have

$$
\mathbb{E}\left[\sum_{\ell=1}^{L} \mathrm{Var}\big(V_{n,\ell}^{\mathrm{Rand}} \mid N_\ell\big)\right] = \beta^2 n^{-1} + o(n^{-1}), \tag{3.36}
$$

where

$$\beta^2 := 2^{-d}\lambda^{2d}(\lambda+2)^{-d}\int_{[0,1]^{2d}}\left\{e^{-2+(\lambda/2)^d \mathrm{vol}(\widetilde{C}(y_1,y_2))} - e^{-2+(\lambda/2+1)^{-d}}\right\}\mathrm{d}y_1\mathrm{d}y_2. \tag{3.37}$$

Now we turn to bound

$$\mathrm{Var}\left(\sum_{\ell=1}^{L}\mathrm{Var}\big(V_{n,\ell}^{\mathrm{Rand}} \mid N_\ell\big)\right) = \sum_{\ell,k\in[L]}\mathrm{Cov}\big(\mathrm{Var}(V_{n,\ell}^{\mathrm{Rand}} \mid N_\ell), \mathrm{Var}(V_{n,k}^{\mathrm{Rand}} \mid N_k)\big).$$

Fix $\ell, k \in [L]$ for now.

Using Lemma 3.6 and by a similar argument as in the second step of (3.35), we have for independent $X_{1,\ell}, X_{2,\ell} \overset{\mathrm{iid}}{\sim} \mathrm{Unif}(\mathcal{Q}_\ell)$ and $X_{1,k}, X_{2,k} \overset{\mathrm{iid}}{\sim} \mathrm{Unif}(\mathcal{Q}_k)$ that

$$\begin{aligned}
\mathrm{Var}(V_{n,\ell}^{\mathrm{Rand}} \mid N_\ell) &= \mathrm{vol}^2(\mathcal{Q}_1)\,\mathbb{E}\Big(u(X_{1,\ell},X_{2,\ell})^{N_\ell} - \big(v(X_{1,\ell})v(X_{2,\ell})\big)^{N_\ell} \;\Big|\; N_\ell\Big)\\
\mathrm{Var}(V_{n,k}^{\mathrm{Rand}} \mid N_k) &= \mathrm{vol}^2(\mathcal{Q}_1)\,\mathbb{E}\Big(u(X_{1,k},X_{2,k})^{N_k} - \big(v(X_{1,k})v(X_{2,k})\big)^{N_k} \;\Big|\; N_k\Big).
\end{aligned}$$

Consequently, using Fubini's theorem, we have

$$\begin{aligned}
&\mathrm{Cov}\big(\mathrm{Var}(V_{n,\ell}^{\mathrm{Rand}} \mid N_\ell), \mathrm{Var}(V_{n,k}^{\mathrm{Rand}} \mid N_k)\big)\\
&\quad= \int_{\mathcal{Q}_1^2\times\mathcal{Q}_1^2}\Big\{\mathrm{Cov}(u(x_1,x_2)^{N_\ell}, u(x_3,x_4)^{N_k}) - 2\mathrm{Cov}(v(x_1)^{N_\ell}v(x_2)^{N_\ell}, u(x_3,x_4)^{N_k})\\
&\qquad\qquad + \mathrm{Cov}(v(x_1)^{N_\ell}v(x_2)^{N_\ell}, v(x_3)^{N_k}v(x_4)^{N_k}\Big\}\,\mathrm{d}(x_1,x_2,x_3,x_4)\\
&\quad= \mathrm{vol}^4(\mathcal{Q}_1)O(n^{-2}) = O(n^{-6}),
\end{aligned}$$

where in the penultimate step, we have used both Lemma 3.8 and the fact that $\max\{p(u(x_1,x_2)-1), p(u(x_3,x_4)-1), p(v(x_1)v(x_2)-1), p(v(x_3,x_4)-1)\} = O(1/n)$. As a result, we have

$$\mathrm{Var}\left(\sum_{\ell=1}^{L}\mathrm{Var}\big(V_{n,\ell}^{\mathrm{Rand}} \mid N_\ell\big)\right) = O(n^{-4}). \tag{3.38}$$

Combining (3.36) and (3.38), Markov's inequality implies that

$$n\sum_{\ell=1}^{L}\mathrm{Var}\big(V_{n,\ell}^{\mathrm{Rand}} \mid N_\ell\big) \xrightarrow{\mathrm{p}} \beta^2. \tag{3.39}$$

Let $\mathcal{F}_n$ be the sigma-algebra generated by $(N_\ell)_{\ell=1}^L$. Since $V_{n,1}^{\mathrm{Rand}}, \ldots, V_{n,L}^{\mathrm{Rand}}$ are independent and identically distributed conditional on $\mathcal{F}_n$, we may apply the Berry–Esseen theorem [Ber41; Ess42] to obtain

$$\sup_{x\in\mathbb{R}}\left|\mathbb{P}\left(n^{1/2}\sum_{\ell=1}^{L}V_{n,l}^{\mathrm{Rand}} \le n^{1/2}\sum_{\ell=1}^{L}\mathbb{E}(V_{n,l}^{\mathrm{Rand}} \mid N_\ell) + x\Big(n\sum_{\ell=1}^{L}\mathrm{Var}(V_{n,l}^{\mathrm{Rand}} \mid N_\ell)\Big)^{1/2} \;\Big|\; \mathcal{F}_n\right) - \Phi(x)\right| \le CR_n, \tag{3.40}$$

where $C > 0$ is a universal constant and

$$R_n := \frac{\sum_{\ell=1}^{L} \mathbb{E}\big\{|V_{n,\ell}^{\mathrm{Rand}} - \mathbb{E}(V_{n,\ell}^{\mathrm{Rand}} \mid N_\ell)|^3 \;\big|\; N_\ell\big\}}{\big\{\sum_{i=1}^{L} \mathrm{Var}(V_{n,\ell}^{\mathrm{Rand}} \mid N_\ell)\big\}^{3/2}}.$$

Moreover, by (3.39) and the fact that $V_{n,\ell}^{\mathrm{Rand}} \le \mathrm{vol}(\mathcal{P}_\ell) = (\lambda/2+1)^d n^{-1}$ for all $\ell \in [L]$, we have

$$R_n \le (\lambda/2+1)^{3d} n^{-2} \bigg\{\sum_{\ell=1}^{L} \mathrm{Var}(V_{n,l}^{\mathrm{Rand}} | N_\ell)\bigg\}^{-3/2} = O_p(n^{-1/2}). \tag{3.41}$$

If we can further show that for a fixed $\lambda \in \mathbb{Q}$, there exists constants α, depending only on λ and d, such that

$$n^{1/2} \sum_{\ell=1}^{L} \big\{\mathbb{E}(V_{n,\ell}^{\mathrm{Rand}} \mid N_\ell) - \mathbb{E}(V_{n,\ell}^{\mathrm{Rand}})\big\} \xrightarrow{\mathrm{d}} \mathcal{N}(0, \alpha^2), \tag{3.42}$$

then combining (3.40), (3.41), (3.42) with Lemma 3.3, we would have

$$n^{1/2} \sum_{\ell=1}^{L} \{V_{n,l}^{\mathrm{Rand}} - \mathbb{E}\, V_{n,l}^{\mathrm{Rand}}\} \xrightarrow{\mathrm{d}} \mathcal{N}(0, \alpha^2 + \beta^2). \tag{3.43}$$

To prove (3.42), we use the asymptotic normality of the sum of multinomial random variables established in Holst [Hol72, Theorem 1]. Write $f(w) := \mathbb{E}(V_{n,1}^{\mathrm{Rand}} \mid N_1 = w)$ for $w \in \{0, \dots, n\}$ and let $W \sim \mathrm{Poi}(np)$. Define

$$\tau^2 := L\mathrm{Var}\big(f(W)\big) - \frac{L^2}{n}\mathrm{Cov}^2\big(W, f(W)\big).$$

Then Holst [Hol72, Theorem 1] implies that

$$\frac{1}{\tau} \sum_{\ell=1}^{L} \big\{\mathbb{E}(V_{n,\ell}^{\mathrm{Rand}} \mid N_\ell) - \mathbb{E}(V_{n,\ell}^{\mathrm{Rand}})\big\} \xrightarrow{\mathrm{d}} \mathcal{N}(0,1), \quad n \to \infty. \tag{3.44}$$

Hence, to prove (3.42), it suffices to investigate the behavior of $n\tau^2$. Firstly, by Lemma 3.6, we have for $X_{1,1} \sim \mathrm{Unif}(\mathcal{Q}_1)$,

$$\begin{aligned} nL\mathrm{Var}\big(f(W)\big) =& nL\mathrm{vol}^2(\mathcal{Q}_1)\mathrm{Var}\Big(\mathbb{E}\big(\{v(X_{1,1})\}^W \mid W\big)\Big) \\ =& \frac{\lambda^{2d}}{(2\lambda+4)^d}\mathrm{Var}\Big(\big\{1-(\lambda/2+1)^{-d}\big\}^W\Big) = \frac{\lambda^{2d}}{(2\lambda+4)^d} e^{-2}\Big(e^{(\lambda/2+1)^{-d}} - 1\Big), \end{aligned} \tag{3.45}$$

where the final equality follows by the fact that $\mathbb{E}\,\eta^W = e^{np(\eta-1)}$ for any $\eta > 0$. We can similarly derive that

$$\begin{aligned} L\mathrm{Cov}\big(W, f(W)\big) &= L\mathrm{vol}(\mathcal{Q}_\ell)\bigg(\mathbb{E}\Big[W\,\mathbb{E}\big(\{v(X_{1,l})\}^W \mid W\big)\Big] - np\,\mathbb{E}\Big[\mathbb{E}\big(\{v(X_{1,l})\}^W \mid W\big)\Big]\bigg) \\ &= \bigg(\frac{\lambda}{\lambda+2}\bigg)^d \Big(\mathbb{E}\big\{W(1-(\lambda/2+1)^{-d})^W\big\} - (\lambda/2+1)^d\,\mathbb{E}\big\{1-(\lambda/2+1)^{-d}\big\}^W\Big) \\ &= \bigg(\frac{\lambda}{\lambda+2}\bigg)^d \Big(np\big(1-(\lambda/2+1)^{-d}\big)e^{-1} - (\lambda/2+1)^d e^{-1}\Big) = -\bigg(\frac{\lambda}{\lambda+2}\bigg)^d e^{-1}, \end{aligned} \tag{3.46}$$

where we use the fact that $\mathbb{E}(W\eta^W) = np\eta e^{np(\eta-1)}$ for $\eta > 0$ in the penultimate equality. Combining (3.45) and (3.46) we have

$$n\tau^2 = (2\lambda+4)^{-d}\lambda^{2d}e^{-2}\Big(e^{(\lambda/2+1)^{-d}} - 1\Big) - (\lambda+2)^{-2d}\lambda^{2d}e^{-2} =: \alpha^2, \tag{3.47}$$

as desired in (3.42), and the conclusion of (3.43) holds as a result.

Now we will let λ diverge. Suppose $Y_1, Y_2 \overset{\text{iid}}{\sim} \text{Unif}[0,1]^d$, then by (3.37), (3.47) and Lemma 3.7, we have as $\lambda \to \infty$ that

$$\begin{aligned}
\alpha^2 + \beta^2 &= 2^{-d}\lambda^{2d}(\lambda+2)^{-d}e^{-2}\Big(\mathbb{E}\, e^{(\lambda/2)^d \text{vol}\left(\widetilde{C}(Y_1,Y_2)\right)} - 1\Big) - \Big(\frac{\lambda}{\lambda+2}\Big)^{2d}e^{-2} \\
&= e^{-2}\Big(\frac{\lambda}{\lambda+2}\Big)^d\Big(1 + \frac{1}{2!}\big(\frac{2}{3}\big)^d + \frac{1}{3!}\big(\frac{2}{4}\big)^d + \frac{1}{4!}\big(\frac{2}{5}\big)^d + \cdots\Big) - \Big(\frac{\lambda}{\lambda+2}\Big)^{2d}e^{-2} \\
&= e^{-2}\sum_{i=2}^{\infty}\frac{1}{i!}\big(\frac{2}{i+1}\big)^d + o(1).
\end{aligned} \tag{3.48}$$

It remains to analyse the vacancy in $\mathcal{R} := [0,1]^d \setminus \cup_{\ell=1}^L \mathcal{Q}_\ell$. Since $\text{vol}(\mathcal{R}) = 1 - \{\lambda/(\lambda+2)\}^d$, by Lemma 3.6, we have

$$\mathbb{E}\Big(V_n^{\text{Rand}} - \sum_{\ell=1}^L V_{n,\ell}^{\text{Rand}}\Big) = \Big\{1 - \Big(\frac{\lambda}{\lambda+2}\Big)^d\Big\}\Big(1 - \frac{1}{n}\Big)^n.$$

Set $Z_1, Z_2 \overset{\text{iid}}{\sim} \text{Unif}(\mathcal{R})$ and define $V' := \text{vol}(B_\infty^d(Z_1,\gamma) \cap B_\infty^d(Z_2,\gamma))$. By Lemma 3.6 again, we have

$$\begin{aligned}
\mathbb{E}\Big(V_n^{\text{Rand}} - \sum_{\ell=1}^L V_{n,\ell}^{\text{Rand}}\Big)^2 &= \Big\{1 - \Big(\frac{\lambda}{\lambda+2}\Big)^d\Big\}^2 \mathbb{E}\Big\{1 - \frac{2}{n} + V'\Big\}^n \\
&= \Big\{1 - \Big(\frac{\lambda}{\lambda+2}\Big)^d\Big\}^2 \sum_{r=0}^n \binom{n}{r}\Big(1 - \frac{2}{n}\Big)^{n-r}\mathbb{E}(V')^r.
\end{aligned}$$

For any $r \geq 1$, we have

$$\begin{aligned}
\mathbb{E}(V')^r &= \Big\{1 - \Big(\frac{\lambda}{\lambda+2}\Big)^d\Big\}^{-2}\int_{\mathcal{R}^2} \text{vol}\big(B_\infty^d(z_1,\gamma) \cap B_\infty^d(z_2,\gamma)\big)^r \,\mathrm{d}(z_1,z_2) \\
&\leq \Big\{1 - \Big(\frac{\lambda}{\lambda+2}\Big)^d\Big\}^{-2}\int_{\mathcal{R}}\int_{[0,1]^d} n^{-r}\mathbb{1}\{z_1 \in B_\infty^d(z_2,2\gamma)\}\,\mathrm{d}z_2\,\mathrm{d}z_1 \\
&= \Big\{1 - \Big(\frac{\lambda}{\lambda+2}\Big)^d\Big\}^{-1} 2^d n^{-r-1}.
\end{aligned}$$

As $n \to \infty$,

$$\begin{aligned}
\text{Var}\Big(V_n^{\text{Rand}} - \sum_{\ell=1}^L V_{n,\ell}^{\text{Rand}}\Big) &= \Big\{1 - \Big(\frac{\lambda}{\lambda+2}\Big)^d\Big\}^2 \sum_{r=1}^n \binom{n}{r}\Big(1 - \frac{2}{n}\Big)^{n-r}\Big(\mathbb{E}(V')^r - n^{-2r}\Big) \\
&\leq 2^d\Big\{1 - \Big(\frac{\lambda}{\lambda+2}\Big)^d\Big\}\sum_{r=1}^n \binom{n}{r}\Big(1 - \frac{2}{n}\Big)^{n-r} n^{-r-1} \\
&= 2^d\Big\{1 - \Big(\frac{\lambda}{\lambda+2}\Big)^d\Big\}n^{-1}(e^{-1} - e^{-2}) + o(n^{-1}).
\end{aligned}$$

Sending first $n \to \infty$, and then $\lambda \to \infty$, we have by Chebyshev's inequality that

$$n^{1/2}\left\{V_n^{\mathrm{Rand}} - \sum_{\ell=1}^{L} V_{n,\ell}^{\mathrm{Rand}} - \mathbb{E}\big(V_n^{\mathrm{Rand}} - \sum_{\ell=1}^{L} V_{n,\ell}^{\mathrm{Rand}}\big)\right\} \xrightarrow{\mathrm{p}} 0. \tag{3.49}$$

Finally, (3.11) follows by combining (3.43), (3.48) and (3.49). □

3.5 Some general results for the vacancy

In this section, we present some general results related to the vacancy area, which are crucial in supporting the proofs in Section 3.4.

Lemma 3.4. *Suppose $Q \subseteq [0,1]^d$ is a hyperrectangle and $R_1, \ldots, R_m \overset{\mathrm{iid}}{\sim} \mathrm{Unif}(Q)$. For $\delta \in (0, 1/2)$, define*

$$V := \frac{\mathrm{vol}\big(Q \setminus \cup_{i=1}^{m} B_\infty^d(R_i, \delta)\big)}{\mathrm{vol}(Q)}.$$

Suppose that $m(2\delta)^d \to r$, then as $m \to \infty$, we have

$$V \xrightarrow{\mathrm{p}} e^{-r/\mathrm{vol}(Q)}.$$

Proof. Draw $U_1, U_2 \overset{\mathrm{iid}}{\sim} \mathrm{Unif}(Q)$ independent of all other randomness. We have

$$\begin{aligned}
\mathbb{E}(V) &= \mathbb{P}\Big(U_1 \notin \bigcup_{i=1}^{m} B_\infty^d(R_i, \delta)\Big) \\
&= \mathbb{E}\Big[\mathbb{P}\Big(\bigcap_{i=1}^{m}\{R_i \notin B_\infty^d(U_1, \delta)\} \;\Big|\; U_1\Big)\Big] \\
&= \mathbb{E}\left\{1 - \frac{\mathrm{vol}\big(B_\infty^d(U_1, \delta) \cap Q\big)}{\mathrm{vol}(Q)}\right\}^m.
\end{aligned}$$

Let $Q_\delta := \{x \in [0,1]^d : \inf_{y \in Q^{\mathrm{c}}} \|x - y\| \geq \delta\}$. Then

$$\mathbb{E}\left\{1 - \frac{\mathrm{vol}\big(B_\infty^d(U_1, \delta)\big)}{\mathrm{vol}(Q)}\right\}^m \leq \mathbb{E}(V) \leq \mathbb{E}\left\{1 - \frac{\mathrm{vol}\big(B_\infty^d(U_1, \delta)\big)\mathbb{1}_{\{U_1 \in Q_\delta\}}}{\mathrm{vol}(Q)}\right\}^m$$

Since $\mathrm{vol}\big(B_\infty^d(U_1, \delta)\big) = (2\delta)^d$ and $\mathbb{P}(U_1 \in Q_\delta) \to 1$ as $m \to \infty$, we have

$$\lim_{m\to\infty} \mathbb{E}(V) = \lim_{m\to\infty} \{1 - (2\delta)^d/\mathrm{vol}(Q)\}^m = e^{-r/\mathrm{vol}(Q)}.$$

By Chebyshev's inequality, it remains to show that $\mathrm{Var}(V) \to 0$. We have

$$\begin{aligned}\mathbb{E}(V^2) &= \mathbb{P}\Big(U_j \notin \bigcup_{i=1}^{m} B_\infty^d(R_i,\delta)\ \forall\, j \in \{1,2\}\Big)\\ &= \mathbb{E}\bigg[\mathbb{P}\Big(\bigcap_{i=1}^{m}\{R_i \notin B_\infty^d(U_1,\delta)\cup B_\infty^d(U_2,\delta)\}\ \Big|\ U_1,U_2\Big)\bigg]\\ &= \mathbb{E}\bigg[\bigg\{1-\frac{\mathrm{vol}\big(B_\infty^d(U_1,\delta)\cap Q\big)}{\mathrm{vol}(Q)}-\frac{\mathrm{vol}\big(B_\infty^d(U_2,\delta)\cap Q\big)}{\mathrm{vol}(Q)}\\ &\qquad\qquad +\frac{\mathrm{vol}\big(B_\infty^d(U_1,\delta)\cap B_\infty^d(U_2,\delta)\cap Q\big)}{\mathrm{vol}(Q)}\bigg\}^m\bigg].\end{aligned}$$

Note that $\mathrm{vol}(B_\infty^d(U_1,\delta_n)\cap B_\infty^d(U_2,\delta_n))$ is exactly equal to 0 for all large n on an event of probability 1. Thus, by a similar sandwiching argument as above, and the Dominated Convergence Theorem, we have

$$\lim_{m\to\infty}\mathbb{E}(V^2) = \lim_{m\to\infty}\{1-2(2\delta)^d/\mathrm{vol}(Q)\}^m = e^{-2r/\mathrm{vol}(Q)}.$$

Consequently $\mathrm{Var}(V) = \mathbb{E}(V^2) - (\mathbb{E}\,V)^2 \to 0$, when $m\to\infty$ as desired. □

Lemma 3.5. *Let $a_1,a_2,b_1,b_2 \in [0,1]$ such that $a_i < b_i$ for $i=1,2$. Then $Q=[a_1,b_1]\times[a_2,b_2]$ is a rectanlge in $[0,1]^2$. Let $m_1,m_2\in\mathbb{N}$, and $\mathbb{L}=\mathbb{L}_1\times\mathbb{L}_2$, where $\mathbb{L}_i = \{a_i+(b_i-a_i)/(m_i+1), a_i+2(b_i-a_i)/(m_i+1),\ldots,a_i+m_i(b_i-a_i)/(m_i+1)\}$, $i=1,2$. Suppose $k:=k(m_1,m_2)\in\mathbb{N}$ such that $k\le m_1\wedge m_2$ and $k/m_i\to s_i\in(0,1)$, $i=1,2$. Let $G=\{U_1,\ldots,U_k\}\sim \mathrm{Unif}\big(\binom{\mathbb{L}_1}{k}\big)$ and $F=\{V_1,\ldots,V_k\}\sim\mathrm{Unif}\big(\binom{\mathbb{L}_2}{k}\big)$ be independent random sets. Let σ be a uniform random pertumation of $[k]$, and write $R_i=(U_i,V_\sigma(i))$, for $i=1,\ldots,k$. Define*

$$V=\frac{\mathrm{vol}(Q\setminus\cup_{i=1}^k B_\infty^2(R_i,\delta))}{\mathrm{vol}(Q)},$$

where the radius δ satisfies that $k(2\delta)^2\to r\in\mathbb{R}_+$, as $m_1,m_2\to\infty$. Then we have

$$\mathbb{E}\,V=\left(1-\frac{2\delta k}{(b_2-a_2)m_1}\right)^{\frac{2\delta m_1}{b_1-a_1}}+O(m_1^{-1/2})\to e^{-r/\mathrm{vol}(Q)}$$

as $m_1,m_2\to\infty$.

Proof. For any $W=(W_1,W_2)\sim\mathrm{Unif}(Q)$, let N_x denote the number of R_i for which the corresponding value of U_i falls within the δ-neighborhood of W_1. Specifically, $N_x = |\{i\in[k]: d_\infty(U_i,W_1)\le\delta\}|$. Conditioning on W, the random variable N_x follows a Hypergeometric distribution $\mathrm{Hyper}(k;2\delta m_1/(b_1-a_1),m_1)$.[1] Equivalently, by the symmetric property of the Hypergeometric distribution, conditioning on W, we have $N_x\sim\mathrm{Hyper}(2\delta m_1/(b_1-a_1);k,m_1)$. Employing Lemma 3.11, let $\xi_1\sim\mathrm{Bin}(2\delta m_1/(b_1-a_1),k/m_1)$, we have for any $w\in Q$,

$$\|P^{N_x|W=w}-P^{\xi_1}\|_{\mathrm{TV}}\le\frac{2\delta m_1}{(m_1-1)(b_1-a_1)}. \tag{3.50}$$

[1] $\mathrm{Hyper}(k;m,n)$: number of success in k draws without replacement, from a population of n that contains m objects with the desirable characteristics.

Furthermore, let N_{xy} denote the number of R_i covered by $B^2_\infty(W,\delta)$, i.e. $N_{xy} = |\{i \in [k] : d_\infty(W,R_i) < \delta\}|$. Note that given N_x number of R_i with the corresponding U_i falls within the δ-neighborhood of W_1, the value of N_{xy} can be obtained by counting how many out of these points have their V_i values within the δ-neighborhood of W_2. Therefore, conditioning on N_x and W, we have $N_{xy} \sim \mathrm{Hyper}(N_x; 2\delta m_2/(b_2-a_2), m_2)$. Let ξ_2 be a random variable such that conditional on ξ_1, ξ_2 follows a Binomial distribution $\mathrm{Bin}(\xi_1, 2\delta/(b_2-a_2))$. Then for any $v \in [2\delta m_1/(b_1-a_1)]$ and $w \in Q$, the similar the Binomial approximation as (3.50) implies that

$$\|P^{N_{xy}|N_x=v,W=w} - P^{\xi_2|\xi_1=v}\|_{\mathrm{TV}} \le \frac{v-1}{m_2-1}. \tag{3.51}$$

Then for any $w \in Q$ and $A \subseteq \mathbb{N}\cup\{0\}$, consider

$$\begin{aligned}
&\left|P^{N_{xy}|W=w}(A) - P^{\xi_2}(A)\right| \\
=&\Big|\sum_{v\in\mathbb{N}\cup\{0\}} \big(\mathbb{P}(N_{xy}\in A|N_x=v,W=w)\,\mathbb{P}(N_x=v|W=w) - \mathbb{P}(\xi_2\in A|\xi_1=v)\,\mathbb{P}(\xi_1=v)\big)\Big| \\
\le&\Big|\sum_{v\in\mathbb{N}\cup\{0\}} \Big(\mathbb{P}(N_{xy}\in A|N_x=v,W=w) - \mathbb{P}(\xi_2\in A|\xi_1=v)\Big)\,\mathbb{P}(N_x=v|W=w)\Big| \\
&+\Big|\sum_{v\in\mathbb{N}\cup\{0\}} \mathbb{P}(\xi_2\in A|\xi_1=v)\Big(\mathbb{P}(N_x=v|W=w) - \mathbb{P}(\xi_1=v)\Big)\Big| \\
\le&\Big|\sum_{v\in\mathbb{N}\cup\{0\}} \frac{v}{m_2-1}\,\mathbb{P}(N_x=v|W=w)\Big| + \frac{4\delta m_1}{(m_1-1)(b_1-a_1)} \\
=&\frac{2\delta k}{(m_2-1)(b_1-a_1)} + \frac{4\delta m_1}{(m_1-1)(b_1-a_1)},
\end{aligned}$$

where we used (3.50) and (3.51) in the penultimate inequality. Consequently, we have

$$\|P^{N_{xy}|W=w} - P^{\xi_2}\|_{\mathrm{TV}} \le \frac{C\delta m_1}{(m_1\wedge m_2-1)(b_1-a_1)}, \quad \text{as } m_1,m_2\to\infty, \tag{3.52}$$

for some universal constant $C>0$.

Note the marginal distribution of ξ_2 is $\mathrm{Bin}\big(2\delta m_1/(b_1-a_1), 2\delta k/((b_2-a_2)m_1)\big)$, together with (3.52) we have

$$\begin{aligned}
\mathbb{E}\,V &= \mathbb{E}\Big[\mathbb{P}\Big(\bigcap_{i=1}^k \{R_i \notin B^2_\infty(W,\delta)\} \;\Big|\; W\Big)\Big] \\
&= \mathbb{E}\big[\mathbb{P}(N_{xy}=0 \;\big|\; W)\big] \\
&= \mathbb{P}(\xi_2=0) + \frac{C\delta m_1}{(m_1\wedge m_2-1)(b_1-a_1)} \\
&= \Big(1-\frac{2\delta k}{(b_2-a_2)m_1}\Big)^{\frac{2\delta m_1}{b_1-a_1}} + \frac{C\delta m_1}{(m_1\wedge m_2-1)(b_1-a_1)} \to e^{-r/\mathrm{vol}(Q)},
\end{aligned}$$

as $m_1,m_2\to\infty$. □

Lemma 3.6. *Define P is a d-dimensional cube with edge length l and fix $Q \subseteq P$. Let $R_1, \dots, R_n \overset{\text{iid}}{\sim} P^R \in \mathcal{P}(P)$ and for any $\delta > 0$ define*

$$V := \frac{\mathrm{vol}\big(Q \setminus \bigcup_{i=1}^n B_\infty(R_i, \delta; P)\big)}{\mathrm{vol}(Q)}.$$

For any $k \in \mathbb{N}$, let $X_1, \dots, X_k \overset{\text{iid}}{\sim} \mathrm{Unif}(Q)$ be a set of random variables independent with $R_1, \dots, R_n$, then

$$\mathbb{E}\, V^k = \mathbb{E}\bigg[\Big\{1 - P^R\Big(\bigcup_{j=1}^k B_\infty(X_j, \delta; P)\Big)\Big\}^n\bigg].$$

Furthermore, if $P^R = \mathrm{Unif}(P)$,

(i) then we have

$$\mathbb{E}(V^k) = \mathbb{E}\bigg[\Big\{1 - \frac{\mathrm{vol}\Big(\bigcup_{j=1}^k B_\infty(X_i, \delta; P)\Big)}{\mathrm{vol}(P)}\Big\}^n\bigg].$$

(ii) if Q is a subcube of P with the same center but the edge length becomes $l - 2\epsilon$ where $\epsilon < l/2$, then when $\delta < l/4 + \epsilon/2$, we have

$$\begin{aligned}
\mathbb{E}(V) &= \Big\{1 - \frac{(2\delta)^d}{\mathrm{vol}(P)}\Big\}^n, \\
\mathbb{E}(V^2) &= \mathbb{E}\Big[\Big\{1 - \mathrm{vol}(P)^{-1}\Big(2(2\delta)^d - \mathrm{vol}\big(B_\infty(X_1, \delta; P) \cap B_\infty(X_2, \delta; P)\big)\Big)\Big\}^n\Big] \\
&= \sum_{r=1}^n \binom{n}{r}\Big(1 - \frac{2(2\delta)^d}{\mathrm{vol}(P)}\Big)^{n-r}\mathrm{vol}(P)^{-r}\Big(\frac{2(2\delta)^{r+1}}{(r+1)(l-2\epsilon)}\Big)^d + \Big(1 - \frac{2(2\delta)^d}{\mathrm{vol}(P)}\Big)^n.
\end{aligned}$$

Thus

$$\mathrm{Var}(V) = \sum_{r=1}^n \binom{n}{r}\Big(1 - \frac{2(2\delta)^d}{\mathrm{vol}(P)}\Big)^{n-r}\bigg(\Big(\frac{2(2\delta)^{r+1}}{(r+1)(l-2\epsilon)}\Big)^d\mathrm{vol}(P)^{-r} - (2\delta)^{2rd}\mathrm{vol}(P)^{-2r}\bigg)$$

Proof. We observe that

$$\begin{aligned}
\mathbb{E}V^k &= \mathbb{E}\bigg\{\prod_{j=1}^k \mathbb{P}\Big(X_j \notin \bigcup_{i=1}^n B_\infty(R_i, \delta; P) \;\Big|\; R_1, \dots, R_n\Big)\bigg\} \\
&= \mathbb{P}\big(X_j \notin B_\infty(R_i, \delta; P)\ \forall\, i \in [n],\ j \in [k]\big) \\
&= \mathbb{P}\big(R_i \notin B_\infty(X_j, \delta; P)\ \forall\, i \in [n],\ j \in [k]\big) \\
&= \mathbb{E}\bigg\{\prod_{i=1}^n \mathbb{P}\Big(R_i \notin \bigcup_{j=1}^k B_\infty(X_j, \delta; P) \;\Big|\; X_1, \dots, X_k\Big)\bigg\} \\
&= \mathbb{E}\bigg[\Big\{1 - P^R\Big(\bigcup_{j=1}^k B_\infty(X_j, \delta; P)\Big)\Big\}^n\bigg],
\end{aligned}$$

which equals to the following when P^R is the uniform distribution on P,

$$\mathbb{E}(V^k) = \mathbb{E}\left[\left\{1 - \frac{\mathrm{vol}\left(\bigcup_{j=1}^k B_\infty(X_j, \delta; P)\right)}{\mathrm{vol}(P)}\right\}^n\right].$$

Now, we work under the assumption of (ii). If $k = 1$, since $\mathrm{vol}(B_\infty(X_1, \delta; P)) = (2\delta)^d$, we have $\mathbb{E}(V) = \left\{1 - \frac{(2\delta)^d}{\mathrm{vol}(P)}\right\}^n$. If $k = 2$, we have

$$\begin{aligned}\mathrm{vol}\big(B_\infty(X_1, \delta; P) \cup B_\infty(X_2, \delta; P)\big) &= 2\mathrm{vol}\big(B_\infty(X_1, \delta; P)\big) - \mathrm{vol}\big(B_\infty(X_1, \delta; P) \cap B_\infty(X_2, \delta; P)\big) \\ &= 2(2\delta)^d - \mathrm{vol}\big(B_\infty(X_1, \delta; P) \cap B_\infty(X_2, \delta; P)\big).\end{aligned}$$

By Lemma 3.7, it follows that

$$\begin{aligned}\mathbb{E}\, V^2 &= \mathbb{E}\left[\left\{1 - \mathrm{vol}(P)^{-1}\Big(2(2\delta)^d - \mathrm{vol}\big(B_\infty(X_1, \delta; P) \cap B_\infty(X_2, \delta; P)\big)\Big)\right\}^n\right] \\ &= \sum_{r=1}^n \binom{n}{r}\Big(1 - \frac{2(2\delta)^d}{\mathrm{vol}(P)}\Big)^{n-r}\mathrm{vol}(P)^{-r}\Big(\frac{2(2\delta)^{r+1}}{(r+1)(l-2\epsilon)}\Big)^d + \Big(1 - \frac{2(2\delta)^d}{\mathrm{vol}(P)}\Big)^n,\end{aligned}$$

as desired □

Lemma 3.7. *Define P as a d-dimensional cube with edge length l and let $Q \subseteq P$ be a subcube of P with the same center and edge length $l - 2\epsilon$, where $\epsilon \in (0, l/2)$. Let $X_1, X_2 \overset{\text{iid}}{\sim} \mathrm{Unif}(Q)$. Then for $\delta \in (0, l/4 + \epsilon/2)$ and any $s \in \mathbb{N}$, we have*

$$\mathbb{E}\Big\{\mathrm{vol}\big(B_\infty(X_1, \delta; P) \cap B_\infty(X_2, \delta; P)\big)\Big\}^s = \left(\frac{2(2\delta)^{s+1}}{(s+1)(l-2\epsilon)}\right)^d.$$

Proof. First of all, by the translational symmetry of the periodic edge, we observe that

$$\mathrm{vol}\big(B_\infty(X_1, \delta; P) \cup B_\infty(X_2, \delta; P)\big) \overset{\mathrm{d}}{=} \mathrm{vol}\big(B_\infty(X_1, \delta; P) \cup B_\infty(p_0, \delta; P)\big),$$

where $p_0 := (p_{0,1}, \ldots, p_{0,d})^\top$ denotes the center of cube P. For any $\ell \in [d]$, we write $P|_\ell$ as the projected space of P onto the ℓ-th coordinate, i.e. $P = \otimes_{\ell=1}^d P|_\ell$, and $X_1 = (X_{1,1}, \ldots, X_{1,d})^\top$. Since the intersection of two cube remains a cube, when $\delta \in (0, l/4 + \epsilon/2)$ and $\epsilon \in (0, l/2)$ the volume of the intersection area can be expressed as

$$\begin{aligned}\mathrm{vol}\big(B_\infty(X_1, \delta; P) \cap B_\infty(p_0, \delta; P)\big) &= \prod_{\ell=1}^d \mathrm{vol}(B_\infty(X_{1,\ell}, \delta; P|_\ell) \cap B_\infty(p_{0,\ell}, \delta; P|_\ell)) \\ &= \prod_{\ell=1}^d \big(-|X_{1,\ell} - p_{0,\ell}| + 2\delta\big)_+.\end{aligned}$$

Consequently,

$$\begin{aligned}\mathbb{E}\Big\{\mathrm{vol}\big(B_\infty(X_1, \delta; P) \cap B_\infty(X_2, \delta; P)\big)\Big\}^s &= \prod_{\ell=1}^d \int_{p_{0,\ell}-l/2+\epsilon}^{p_{0,\ell}+l/2-\epsilon} \frac{\big\{(-|x_{1,\ell} - p_{0,\ell}| + 2\delta)_+\big\}^s}{l - 2\epsilon}\, dx_{1,\ell} \\ &= \left(\frac{2(2\delta)^{s+1}}{(s+1)(l-2\epsilon)}\right)^d,\end{aligned}$$

as desired. □

3.6 Additional Proofs

3.6.1 Proof of Lemma 3.1

Proof. Consider the homogeneous Poisson point process ν on $D^+ := \{(x, y) : 0 \le y \le f^+(x)\} \subseteq \mathbb{R}^d \times [0, \infty)$ with the intensity constantly equal to λ. Then $\nu(D^+) \overset{\mathrm{d}}{=} N \sim \mathrm{Poi}(\lambda\alpha)$. Suppose $Z_i = (\tilde{X}_i, Y_i)$, $i = 1, \ldots, N$, be the (random) support of ν, where each $\tilde{X}_i \in \mathbb{R}^d$ and $Y_i \in [0, \infty]$, we can write $\nu = \frac{1}{N}\sum_{i=1}^{N} \delta_{Z_i}$.

Define $D := \{(x, y) : 0 \le y \le f(x)\} \subseteq \mathbb{R}^d \times [0, \infty)$, let $\mathcal{I} := \{i \in [N] : Z_i \in D\}$ and $\mathcal{J} := \{i \in [N] : Z_i \in D^+ \setminus D\}$, then $M_1 := |\mathcal{I}| = \nu(D) \sim \mathrm{Poi}(\lambda)$ and $M_2 := |\mathcal{J}| = \nu(D^+ \setminus D) \sim \mathrm{Poi}(\lambda(\alpha - 1))$. By the complete independence property [see, e.g. Kin92, pp. 11] of the poisson process, we have $\frac{1}{M_1}\sum_{i\in\mathcal{I}} \delta_{Z_i}$ and $\frac{1}{M_2}\sum_{i\in\mathcal{J}} \delta_{Z_i}$ are two independent homogeneous Poisson point processes on D and $D^+ \setminus D$ respectively. Moreover, by the mapping Theorem [see, e.g. Kin92, Section 2.3], $\frac{1}{M_1}\sum_{i\in\mathcal{I}} \delta_{\tilde{X}_i}$ and $\frac{1}{M_2}\sum_{i\in\mathcal{J}} \delta_{\tilde{X}_i}$ are Poisson processes on $\mathbb{R}^d$ with intensity function λf and $\lambda(f^+ - f)$ respectively. Hence by the conditional property of Poisson process [see, e.g. Kin92, Section 2.4], we have

$$(X_1, \ldots, X_M) \mid M \overset{\mathrm{d}}{=} (\tilde{X}_i : i \in \mathcal{I}) \mid M_1 \overset{\mathrm{iid}}{\sim} f, \tag{3.53}$$

$$(X_{M+1}, \ldots, X_N) \mid N - M \overset{\mathrm{d}}{=} (\tilde{X}_i : i \in \mathcal{J}) \mid M_2 \overset{\mathrm{iid}}{\sim} (f^+ - f)/(\alpha - 1). \tag{3.54}$$

By observing that $(M, N - M) \overset{\mathrm{d}}{=} (M_1, M_2) \sim \mathrm{Poi}(\lambda) \otimes \mathrm{Poi}(\lambda(\alpha - 1))$ and using the fact that $\tilde{X}_1, \ldots, \tilde{X}_N$ are permutation invariant condition on N, we must have

$$\big(X_{\sigma(1)}, \ldots, X_{\sigma(N)} \mid N\big) \overset{\mathrm{d}}{=} \big(\tilde{X}_{\sigma(1)}, \ldots, \tilde{X}_{\sigma(N)} \mid N\big).$$

Finally, applying a similar argument to that used in (3.53) and (3.54) to Poisson process ν, we obtain that $(\tilde{X}_1, \ldots, \tilde{X}_N) \mid N \overset{\mathrm{iid}}{\sim} f^+/\alpha$, which proves the desired result. □

3.6.2 Proof of Lemma 3.3

Proof. For any $x \in \mathbb{R}$, we have

$$\begin{aligned}
&\sup_{x\in\mathbb{R}} \left| \mathbb{P}\left(X_n \le \mu + x\sqrt{\alpha^2 + \beta^2} \right) - \Phi(x) \right| \\
&\quad = \sup_{x\in\mathbb{R}} \left| \mathbb{E}\left\{ \mathbb{P}\left(\frac{X_n - M_n}{\sqrt{L_n}} \le \frac{\mu - M_n + x\sqrt{\alpha^2 + \beta^2}}{\sqrt{L_n}} \,\middle|\, \mathcal{F}_n \right) \right\} - \Phi(x) \right| \\
&\quad = \sup_{x\in\mathbb{R}} \left| \mathbb{E}\left\{ \Phi\left(\frac{\mu - M_n + x\sqrt{\alpha^2 + \beta^2}}{\sqrt{L_n}} \right) \right\} - \Phi(x) \right| + o(1),
\end{aligned} \tag{3.55}$$

where we use condition (3.34) in the final step. By the Slutsky's theorem we have for each $x \in \mathbb{R}$ that

$$\frac{\mu - M_n + x\sqrt{\alpha^2 + \beta^2}}{\sqrt{L_n}} \xrightarrow{\mathrm{d}} \mathcal{N}\left(\frac{x\sqrt{\alpha^2 + \beta^2}}{\beta}, \frac{\alpha^2}{\beta^2} \right).$$

Consequently, we have for $Z \sim N(0,1)$ independent of all other randomness in the lemma that

$$\mathbb{E}\left\{\Phi\left(\frac{\mu - M_n + x\sqrt{\alpha^2+\beta^2}}{\sqrt{L_n}}\right)\right\} = \mathbb{P}\left(Z - \frac{\mu - M_n + x\sqrt{\alpha^2+\beta^2}}{\sqrt{L_n}} \leq 0\right) = \Phi(x) + o(1).$$

The conclusion holds by combining the above with (3.55), and using Chow and Teicher [CT88, Lemma 3, pp.265]. □

3.7 Auxiliary results

Lemma 3.8. *For $n, L \in \mathbb{N}$, let $p := 1/L$ and suppose $(N_1, \ldots, N_L) \sim \text{Multin}(n; (p, \ldots, p))$. Consider the asymptotic regime where $n \to \infty$ and L is fixed. Suppose $a, b \geq 1$ satisfies $p(a-1) = O(1/n)$ and $p(b-1) = O(1/n)$, then for any and $\ell, k \in [L]$, we have*

$$\text{Cov}(a^{N_\ell}, b^{N_k}) = O(n^{-2}).$$

Proof. We first assume that $\ell \neq k$. We write $\alpha = p(a-1)$ and $\beta = p(b-1)$ for simplicity. Using the moment generating function of Multinomial distribution, we observe that,

$$\mathbb{E}(a^{N_\ell}) = (1+\alpha)^n$$
$$\mathbb{E}(a^{N_\ell} b^{N_k}) = (1+\alpha+\beta)^n.$$

Using the above identities and the Taylor expansion

$$\begin{aligned}
|\text{Cov}(a^{N_\ell}, b^{N_k})| &= |(1+\alpha+\beta)^n - (1+\alpha+\beta+\alpha\beta)^n| \\
&= \alpha\beta \sum_{i=0}^{n-1} (1+\alpha+\beta)^{n-1-i} (\alpha\beta)^i \\
&\leq \alpha\beta(1+\alpha+\beta)^n \sum_{i=0}^{n-1} (\alpha\beta)^i = O(\alpha\beta) = O(n^{-2}).
\end{aligned}$$

It remains to check the case where $\ell = k$. For this, define $\eta = p(ab-1)$,

$$\begin{aligned}
|\text{Cov}(a^{N_\ell}, b^{N_k})| &= \{1+p(ab-1)\}^n - \{1+p(a-1)\}^n \{1+p(b-1)\}^n \\
&= (1+\alpha\beta/p+\alpha+\beta)^n - (1+\alpha+\beta+\alpha\beta)^n \\
&= \alpha\beta(1/p-1) \sum_{i=0}^{n-1} (1+\alpha+\beta+\alpha\beta)^{n-1-i} (\alpha\beta/p - \alpha\beta)^i \\
&\leq \alpha\beta(1/p-1)(1+\alpha)^n(1+\beta)^n \sum_{i=0}^{n-1} (\alpha\beta/p - \alpha\beta)^i = O(\alpha\beta) = O(n^{-2}),
\end{aligned}$$

as desired. □

Lemma 3.9. *Let $\{a_n\}_{n\geq 1}$ be a real sequence such that $a_n = O(n^{-1})$, then we have $(1+a_n)^n = (1+o(1))e^{na_n}$.*

Proof. Since $a_n = O(n^{-1})$, the result immediately follows by

$$(1+a_n)^n = e^{n\log(1+a_n)} = e^{n(a_n+o(a_n))} = (1+o(1))e^{na_n}.$$

□

Lemma 3.10. *Let (X,Y) be a pair of jointly distributed random variables on $\mathbb{R}^{d_1}\times\mathbb{R}^{d_2}$. Let U and V be continuous random variables on $\mathbb{R}^{d_1}$ and $\mathbb{R}^{d_2}$, chosen such that $U \perp\!\!\!\perp V \mid (X,Y)$ and that*

$$U \in \operatorname*{arg\,min}_{\tilde{U}\sim P^U}\{\mathbb{E}\,\|X-\tilde{U}\|^2\} \quad \textit{and} \quad V \in \operatorname*{arg\,min}_{\tilde{V}\sim P^V}\{\mathbb{E}\,\|Y-\tilde{V}\|^2\}. \tag{3.56}$$

Let $P^{(X,Y)}$ be the joint distribution of (X,Y) with marginals P^X and P^Y, and similarly $P^{(U,V)}$, P^U, P^V the joint and marginal distributions of (U,V). Then for any convex function $f:\mathbb{R}\to\mathbb{R}\cup\{\infty\}$ with $f(1)=0$ (cf. Definition 3.1), we have

$$D_f(P^{(X,Y)}\,\|\,P^X\otimes P^Y) = D_f(P^{(U,V)}\,\|\,P^U\otimes P^V).$$

Proof. Observe that U and V are solutions for optimal transport problems (3.56). Let $\pi\in\mathcal{P}(\mathbb{R}^{2d_1})$ be the optimal coupling (joint distribution) of (X,U) and let $\gamma\in\mathcal{P}(\mathbb{R}^{2d_2})$ be the optimal couplings of (Y,V). Let $P^{U|X=x}$ and $P^{V|Y=y}$ be the corresponding conditional distributions of U given $X=x$ and V given $Y=y$ respectively. Note that these conditional distributions are well-defined up to a P^X-measure 0 set of x-values, say $\mathcal{E}_x$, and a P^Y-measure 0 set of y-values, say $\mathcal{E}_y$. The fact that $U \perp\!\!\!\perp V \mid (X,Y)$ means that $P^{U|X=x}\otimes P^{V|Y=y}$ is the conditional distribution of (U,V) given $(X,Y)=(x,y)$. Note again that $P^{U|X=x}\otimes P^{V|Y=y}$ is well-defined up to $(\mathcal{E}_x\times\mathbb{R}^{d_2})\cup(\mathbb{R}^{d_1}\times\mathcal{E}_y)$, which has zero measure with respect to both $P^{(X,Y)}$ and $P^X\otimes P^Y$. Therefore, we have that

$$P^{(U,V)} = \int_{\mathbb{R}^{d_1}\times\mathbb{R}^{d_2}} P^{U|X=x}\otimes P^{V|Y=y}\,\mathrm{d}P^{(X,Y)}(x,y)$$

and

$$P^U\otimes P^V = \int_{\mathbb{R}^{d_1}\times\mathbb{R}^{d_2}} P^{U|X=x}\otimes P^{V|Y=y}\,\mathrm{d}(P^X\otimes P^Y)(x,y).$$

Thus, using the data processing inequality Polyanskiy and Wu [PW25, Theorem 7.4], we have

$$D_f(P^{(X,Y)}\,\|\,P^X\otimes P^Y) \geq D_f(P^{(U,V)}\,\|\,P^U\otimes P^V).$$

Since U is absolutely continuous, by Brenier's theorem [see e.g. Vil21, Theorem 2.12], there exists a convex function $\varphi:\mathbb{R}^{d_1}\to\mathbb{R}$ such that $\mathrm{d}\pi(x,u) = \mathrm{d}P^U(u)\delta_{\{y=\nabla\varphi(u)\}}$. In other words, the optimal transport from U to X is the function $\nabla\varphi$ (which is P^U-almost everywhere uniquely defined), and so $X=\nabla\varphi(U)$. Similarly, we have $Y=\nabla\psi(V)$ for some convex function $\psi:\mathbb{R}^{d_2}\to\mathbb{R}$. Consequently, we have that X and Y are deterministic, so in particular, conditionally independent, given (U,V). This allows us to run a symmetric argument with conditional distribution of (X,Y) given (U,V) to obtain a data processing inequality in the reverse direction, thus establishing the desired equality. □

Lemma 3.11 ([Ehm91, Theorem 2]). *Given $k, n, m \in \mathbb{N}$ such that $k \leq m \leq n$. Let H be a random variable follows a Hypergeometric distribution $\mathrm{Hyper}(k; m, n)$. Let $B \sim \mathrm{Bin}(k, m/n)$, then we have*

$$\|P^H - P^B\|_{\mathrm{TV}} \leq \frac{k-1}{n-1}.$$

Lemma 3.12 ([PW25]). *Suppose $X \sim P^X \in \mathcal{P}(\mathcal{R}^k)$ and $Y \sim P^Y \in \mathcal{P}(\mathcal{R}^l)$, where $k, l \geq 1$. Then for given any convex function $f : (0, \infty) \to \mathbb{R}$ with $f(1) = 0$, we have*

(i) $D_f(P^X \| P^Y) \geq 0$, and if f is strict convex at 1, then the equality holds if and only if $P^X = P^Y$;

(ii) for a random variable $Z \sim P^Z \in \mathcal{P}(\mathbb{R}^m)$, $m \geq 1$, such that $X \perp\!\!\!\perp Y | Z$, then we have $D_f(P^X \| P^Y) \leq D_f(P^X \| P^Z)$.

(iii) For any sequence of random variables X_n and Y_n such that $P^{X_n} \xrightarrow{\mathrm{w}} P^X$ and $P^{Y_n} \xrightarrow{\mathrm{w}} P^Y$, we have

$$\liminf_{n \to \infty} D_f(P^{X_n} \| P^{Y_n}) \geq D_f(P^X \| P^Y).$$

Bibliography

[AC21] Mona Azadkia and Sourav Chatterjee. *A simple measure of conditional dependence*. The Annals of Statistics 49.6 (2021), pp. 3070–3102 (cit. on pp. 20, 24, 73, 77).

[ADN21] Arnab Auddy, Nabarun Deb, and Sagnik Nandy. *Exact detection thresholds and minimax optimality of Chatterjee's correlation coefficient*. arXiv preprint arXiv:2104.15140 (2021) (cit. on pp. 20, 24, 83).

[Ado+23] Urte Adomaityte, Leonardo Defilippis, Bruno Loureiro, and Gabriele Sicuro. *High-dimensional robust regression under heavy-tailed data: Asymptotics and Universality*. arXiv preprint arXiv:2309.16476 (2023) (cit. on p. 31).

[AF22] Jonathan Ansari and Sebastian Fuchs. *A simple extension of Azadkia & Chatterjee's rank correlation to multi-response vectors*. arXiv preprint arXiv:2212.01621 (2022) (cit. on p. 77).

[AF24] J. Ansari and S. Fuchs. *A simple extension of Azadkia & Chatterjee's rank correlation to multi-response vectors*. arXiv preprint arXiv:2212.01621 (2024) (cit. on pp. 20, 73).

[AMS96] Noga Alon, Yossi Matias, and Mario Szegedy. *The space complexity of approximating the frequency moments*. Proceedings of the Twenty-Eighth Annual ACM Symposium on Theory of Computing. 1996, pp. 20–29 (cit. on pp. 15, 31).

[Aur87] Franz Aurenhammer. *Power diagrams: properties, algorithms and applications*. SIAM journal on computing 16.1 (1987), pp. 78–96 (cit. on p. 27).

[AZ22] Pedro Abdalla and Nikita Zhivotovskiy. *Covariance estimation: Optimal dimension-free guarantees for adversarial corruption and heavy tails*. arXiv preprint arXiv:2205.08494 (2022) (cit. on pp. 15, 31).

[BC11] Alexandre Belloni and Victor Chernozhukov. *ℓ_1-penalized quantile regression in high-dimensional sparse models*. Annals of Statistics 39 (2011), pp. 82–130 (cit. on p. 31).

[BD14] Wicher Bergsma and Angelos Dassios. *A consistent test of independence based on a sign covariance related to Kendall's tau*. Bernoulli (2014), pp. 1006–1028 (cit. on p. 19).

[Ber+19] Espen Bernton, Pierre E. Jacob, Mathieu Gerber, and Christian P. Robert. *On parameter estimation with the Wasserstein distance*. Information and Inference: A Journal of the IMA 8 (2019), pp. 657–676 (cit. on p. 37).

[Ber41] Andrew C. Berry. *The accuracy of the Gaussian approximation to the sum of independent variates*. Transactions of the American Mathematical Society 49.1 (1941), pp. 122–136 (cit. on p. 96).

[Bic22] Peter J Bickel. *Measures of independence and functional dependence*. arXiv preprint arXiv:2206.13663 (2022) (cit. on p. 20).

[Bic65] Peter J. Bickel. *On some asymptotically nonparametric competitors of Hotelling's T^2*. The Annals of Mathematical Statistics (1965), pp. 160–173 (cit. on pp. 15, 21, 75).

[BKR61] Julius R Blum, Jack Kiefer, and Murray Rosenblatt. *Distribution free tests of independence based on the sample distribution function*. Sandia Corporation, 1961 (cit. on p. 19).

[Blo50] Nils Blomqvist. *On a measure of dependence between two random variables*. The Annals of Mathematical Statistics (1950), pp. 593–600 (cit. on p. 19).

[BMG14] Munmun Biswas, Minerva Mukhopadhyay, and Anil K. Ghosh. *A distribution-free two-sample run test applicable to high-dimensional data*. Biometrika 101.4 (2014), pp. 913–926 (cit. on p. 22).

[Bor+23] Emanuele Borgonovo, Giuseppe Savaré, Alessio Figalli, Promit Ghosal, and Elmar Plischke. *Convexity and Measures of Statistical Association* (2023) (cit. on pp. 78, 79).

[Bro76] Efim M. Bronshtein. *ε-entropy of convex sets and functions*. Siberian Mathematical Journal 17 (1976), pp. 393–398 (cit. on p. 63).

[BS19] Thomas B. Berrett and Richard J. Samworth. *Nonparametric independence testing via mutual information*. Biometrika 106.3 (2019), pp. 547–566 (cit. on pp. 77, 78).

[BSH24] Eustasio del Barrio, Alberto Gonzalez Sanz, and Marc Hallin. *Nonparametric multiple-output center-outward quantile regression*. Journal of the American Statistical Association (2024), pp. 1–15 (cit. on pp. 15, 27, 28, 32, 33, 35).

[BSS18] Melf Boeckel, Vladimir Spokoiny, and Alexandra Suvorikova. *Multivariate Brenier cumulative distribution functions and their application to non-parametric testing*. arXiv preprint arXiv:1809.04090 (2018) (cit. on p. 75).

[Cat12] Olivier Catoni. *Challenging the empirical mean and empirical variance: a deviation study*. Annales de l'IHP Probabilités et statistiques 48 (2012), pp. 1148–1185 (cit. on p. 31).

[CB20] Sky Cao and Peter J. Bickel. *Correlations with tailored extremal properties*. arXiv preprint arXiv:2008.10177 (2020) (cit. on pp. 24, 77, 83, 86).

[CC14] Anirvan Chakraborty and Probal Chaudhuri. *The spatial distribution in infinite dimensional spaces and related quantiles and depths*. The Annals of Statistics 42.3 (2014) (cit. on pp. 22, 33, 71).

[CC17] Anirvan Chakraborty and Probal Chaudhuri. *Tests for high-dimensional data based on means, spatial signs and spatial ranks*. Annals of Statistics 45.2 (2017), p. 771 (cit. on p. 22).

[CC19] Joydeep Chowdhury and Probal Chaudhuri. *Nonparametric depth and quantile regression for functional data*. Bernoulli 25 (2019), pp. 395–423 (cit. on p. 71).

[CC96] Biman Chakraborty and Probal Chaudhuri. *On a transformation and re-transformation technique for constructing an affine equivariant multivariate median*. Proceedings of the American Mathematical Society 124.8 (1996), pp. 2539–2547 (cit. on p. 22).

[CCG16] Guillaume Carlier, Victor Chernozhukov, and Alfred Galichon. *Vector quantile regression: An optimal transport approach*. Annals of Statistics 44.3 (2016), pp. 1165–1192 (cit. on pp. 15, 28, 32, 33).

[Cha03] Biman Chakraborty. *On multivariate quantile regression*. Journal of Statistical Planning and Inference 110 (2003), pp. 109–132 (cit. on p. 71).

[Cha21] Sourav Chatterjee. *A new coefficient of correlation*. Journal of the American Statistical Association 116.536 (2021), pp. 2009–2022 (cit. on pp. 20, 24, 73, 77, 82, 86).

[Cha24] Sourav Chatterjee. *A survey of some recent developments in measures of association*. Probability and Stochastic Processes: A Volume in Honour of Rajeeva L. Karandikar (2024), pp. 109–128 (cit. on pp. 19, 77).

[Cha96] Probal Chaudhuri. *On a geometric notion of quantiles for multivariate data*. Journal of the American statistical association 91.434 (1996), pp. 862–872 (cit. on pp. 15, 22, 33, 34, 40, 71, 75).

[Che+17] Victor Chernozhukov, Alfred Galichon, Marc Hallin, and Marc Henry. *Monge–Kantorovich depth, quantiles, ranks and signs*. Annals of Statistics 45 (2017), pp. 223–256 (cit. on pp. 7, 15, 23–27, 32, 33, 35, 75).

[CM97] Kyungmee Choi and John Marden. *An approach to multivariate rank tests in multivariate analysis of variance*. Journal of the American Statistical Association 92.440 (1997), pp. 1581–1590 (cit. on p. 22).

[CT88] Y.S. Chow and H. Teicher. *Probability Theory: Independence, Interchangeability, Martingales*. Springer, 1988 (cit. on p. 105).

[De 00] Mark De Berg. *Computational Geometry: Algorithms and Applications*. Springer Science & Business Media, 2000 (cit. on p. 81).

[DG92] David L. Donoho and Miriam Gasko. *Breakdown properties of location estimates based on halfspace depth and projected outlyingness*. The Annals of Statistics (1992), pp. 1803–1827 (cit. on p. 23).

[DGL13] Luc Devroye, László Györfi, and Gábor Lugosi. *A Probabilistic Theory of Pattern Recognition*. Vol. 31. Springer Science & Business Media, 2013 (cit. on p. 56).

[DGS20] Nabarun Deb, Promit Ghosal, and Bodhisattva Sen. *Measuring association on topological spaces using kernels and geometric graphs*. arXiv preprint arXiv:2010.01768 (2020) (cit. on pp. 20, 24, 73, 77, 82).

[DK01] Dominique Drouet Mari and Samuel Kotz. *Correlation and Dependence*. World Scientific, 2001 (cit. on p. 19).

[DK14] Doulaye Dembélé and Philippe Kastner. *Fold change rank ordering statistics: a new method for detecting differentially expressed genes*. BMC Bioinformatics 15 (2014), pp. 1–15 (cit. on p. 15).

[DKP20] Ilias Diakonikolas, Daniel M. Kane, and Ankit Pensia. *Outlier robust mean estimation with subgaussian rates via stability*. Advances in Neural Information Processing Systems 33. 2020, pp. 1830–1840 (cit. on pp. 15, 31).

[DL22] Jules Depersin and Guillaume Lecué. *Robust sub-Gaussian estimation of a mean vector in nearly linear time*. Annals of Statistics 50 (2022), pp. 511–536 (cit. on pp. 15, 31).

[DN04] Herbert A. David and Haikady N. Nagaraja. *Order Statistics*. John Wiley & Sons, 2004 (cit. on p. 88).

[DS21] Nabarun Deb and Bodhisattva Sen. *Multivariate rank-based distribution-free nonparametric testing using measure transportation*. Journal of the American Statistical Association 118 (2021), pp. 1–16 (cit. on pp. 15, 33, 35, 77).

[DS23] Nabarun Deb and Bodhisattva Sen. *Multivariate rank-based distribution-free nonparametric testing using measure transportation*. Journal of the American Statistical Association 118.541 (2023), pp. 192–207 (cit. on pp. 15, 26–29, 75).

[DSS13] Holger Dette, Karl F. Siburg, and Pavel A. Stoimenov. *A copula-based non-parametric measure of regression dependence*. Scandinavian Journal of Statistics 40.1 (2013), pp. 21–41 (cit. on pp. 19, 77).

[Ehm91] Werner Ehm. *Binomial approximation to the Poisson binomial distribution*. Statistics & Probability Letters 11.1 (1991), pp. 7–16 (cit. on p. 107).

[EL11] B. Essama-Nssah and Peter Lambert. *Influence functions for distributional statistics*. Society for the study of economic inequality working paper. ECINEQ WP 236 (2011), p. 2011 (cit. on p. 18).

[ENK16] Anders Eklund, Thomas E. Nichols, and Hans Knutsson. *Cluster failure: Why fMRI inferences for spatial extent have inflated false-positive rates*. Proceedings of the National Academy of Sciences 113 (2016), pp. 7900–7905 (cit. on p. 31).

[Ess42] Carl-Gustav Esseen. *On the Liapunov limit error in the theory of probability*. Ark. Mat. Astr. Fys. 28 (1942), pp. 1–19 (cit. on p. 96).

[FG15] Nicolas Fournier and Arnaud Guillin. *On the rate of convergence in Wasserstein distance of the empirical measure*. Probability Theory and Related Fields 162 (2015), pp. 707–738 (cit. on pp. 38, 70).

[FLW17] Jianqing Fan, Quefeng Li, and Yuyan Wang. *Estimation of high dimensional mean regression in the absence of symmetry and light tail assumptions*. Journal of the Royal Statistical Society Series B: Statistical Methodology 79 (2017), pp. 247–265 (cit. on p. 31).

[Fuc24] Sebastian Fuchs. *Quantifying directed dependence via dimension reduction*. Journal of Multivariate Analysis 201 (2024), p. 105266 (cit. on pp. 19, 77).

[Geb41] Hans Gebelein. *Das statistische problem der korrelation als variationsund eigenwertproblem und sein zusammenhang mit der ausgleichsrechnung*. ZAMM-Journal of Applied Mathematics and Mechanics/Zeitschrift für Angewandte Mathematik und Mechanik 21.6 (1941), pp. 364–379 (cit. on p. 77).

[Gel90] Matthias Gelbrich. *On a formula for the L_2-Wasserstein metric between measures on Euclidean and Hilbert spaces*. Mathematische Nachrichten 147 (1990), pp. 185–203 (cit. on p. 47).

[GJT22] Florian Griessenberger, Robert R. Junker, and Wolfgang Trutschnig. *On a multivariate copula-based dependence measure and its estimation.* Electronic Journal of Statistics 16.1 (2022), pp. 2206–2251 (cit. on pp. 19, 77, 79).

[GM89] S. Graf and R. Daniel Mauldin. *A classification of disintegrations of measures.* Measure and Measurable Dynamics, Contemporary Mathematics 94 (1989), pp. 147–158 (cit. on p. 50).

[Gre+05a] Arthur Gretton, Olivier Bousquet, Alex Smola, and Bernhard Schölkopf. *Measuring statistical dependence with Hilbert-Schmidt norms.* International Conference on Algorithmic Learning Theory. Springer. 2005, pp. 63–77 (cit. on p. 19).

[Gre+05b] Arthur Gretton, Ralf Herbrich, Alexander Smola, Olivier Bousquet, Bernhard Schölkopf, and Aapo Hyvärinen. *Kernel methods for measuring independence.* Journal of Machine Learning Research 6.12 (2005) (cit. on pp. 24, 77, 81, 83).

[Gre+07] Arthur Gretton, Kenji Fukumizu, Choon Teo, Le Song, Bernhard Schölkopf, and Alex Smola. *A kernel statistical test of independence.* Advances in neural information processing systems 20 (2007) (cit. on pp. 19, 24, 77, 83).

[Gre+12] Arthur Gretton, Karsten M. Borgwardt, Malte J. Rasch, Bernhard Schölkopf, and Alexander Smola. *A kernel two-sample test.* The Journal of Machine Learning Research 13.1 (2012), pp. 723–773 (cit. on p. 24).

[GS22] Promit Ghosal and Bodhisattva Sen. *Multivariate ranks and quantiles using optimal transport: Consistency, rates and nonparametric testing.* The Annals of Statistics 50.2 (2022), pp. 1012–1037 (cit. on pp. 26–28, 77).

[Hal+15] Marc Hallin, Zudi Lu, Davy Paindaveine, and Miroslav Šiman. *Local bilinear multiple-output quantile/depth regression.* Bernoulli 21.3 (2015), pp. 1435–1466 (cit. on pp. 22, 32, 33).

[Hal+21] Marc Hallin, Eustasio Del Barrio, Juan Cuesta-Albertos, and Carlos Matrán. *Distribution and quantile functions, ranks and signs in dimension d: A measure transportation approach.* Annals of Statistics 49 (2021), pp. 1139–1165 (cit. on pp. 7, 15, 23–29, 32, 33, 35, 75, 77).

[Hal17] Marc Hallin. *On distribution and quantile functions, ranks and signs in R_d.* Working Papers ECARES ECARES 2017-34. ULB – Universite Libre de Bruxelles, Sept. 2017 (cit. on pp. 15, 75).

[Hal22] Marc Hallin. *Measure transportation and statistical decision theory.* Annual Review of Statistics and Its Application 9.1 (2022), pp. 401–424 (cit. on pp. 32, 33).

[Hal85] Peter Hall. *Three limit theorems for vacancy in multivariate coverage problems.* Journal of Multivariate Analysis 16.2 (1985), pp. 211–236 (cit. on pp. 80, 94).

[Hal88] Peter Hall. *Introduction to the Theory of Coverage Processes.* Wiley, 1988 (cit. on p. 80).

[Han21] Fang Han. *On extensions of rank correlation coefficients to multivariate spaces.* Bernoulli News 28.2 (2021), pp. 7–11 (cit. on pp. 20, 77).

[HDS22] Zhen Huang, Nabarun Deb, and Bodhisattva Sen. *Kernel partial correlation coefficient—a measure of conditional dependence.* Journal of Machine Learning Research 23.216 (2022), pp. 1–58 (cit. on pp. 20, 24).

[HGH12] R. Heller, M. Gorfine, and Y. Heller. *A class of multivariate distribution-free tests of independence based on graphs*. Journal of Statistical Planning and Inference 142.12 (2012), pp. 3097–3106 (cit. on p. 24).

[HH24] Fang Han and Zhihan Huang. *Azadkia–Chatterjee's correlation coefficient adapts to manifold data*. The Annals of Applied Probability 34.6 (2024), pp. 5172–5210 (cit. on p. 20).

[HHG13] Ruth Heller, Yair Heller, and Malka Gorfine. *A consistent multivariate test of association based on ranks of distances*. Biometrika 100.2 (2013), pp. 503–510 (cit. on pp. 19, 77).

[HHH23] Marc Hallin, Daniel Hlubinka, and Šárka Hudecová. *Efficient fully distribution-free center-outward rank tests for multiple-output regression and MANOVA*. Journal of the American Statistical Association 118.543 (2023), pp. 1923–1939 (cit. on pp. 15, 33).

[HL24] Marc Hallin and Hang Liu. *Quantiles and Quantile Regression on Riemannian Manifolds: a measure-transportation-based approach*. arXiv preprint arXiv:2410.15711 (2024) (cit. on p. 28).

[HLL22] Marc Hallin, Davide La Vecchia, and Hang Liu. *Center-outward R-estimation for semiparametric VARMA models*. Journal of the American Statistical Association 117.538 (2022), pp. 925–938 (cit. on p. 15).

[Hod55] Joseph L. Hodges. *A bivariate sign test*. The Annals of Mathematical Statistics 26.3 (1955), pp. 523–527 (cit. on pp. 15, 21, 75).

[Hoe94] Wassily Hoeffding. *A non-parametric test of independence*. The Collected Works of Wassily Hoeffding (1994), pp. 214–226 (cit. on p. 19).

[Hol72] Lars Holst. *Asymptotic normality and efficiency for certain goodness-of-fit tests*. Biometrika 59.1 (1972), pp. 137–145 (cit. on p. 97).

[HP02a] Marc Hallin and Davy Paindaveine. *Multivariate signed ranks: Randles' interdirections or Tyler's angles?* Statistical Data Analysis Based on the L_1-norm and Related Methods. Springer, 2002, pp. 271–282 (cit. on pp. 15, 75).

[HP02b] Marc Hallin and Davy Paindaveine. *Optimal tests for multivariate location based on interdirections and pseudo-Mahalanobis ranks*. The Annals of Statistics 30.4 (2002), pp. 1103–1133 (cit. on pp. 15, 75).

[HPŠ10] Marc Hallin, Davy Paindaveine, and Miroslav Šiman. *Multivariate quantiles and multiple-output regression quantiles: From L_1-optimization to halfspace depth*. Annals of Statistics 38 (2010), pp. 635–669 (cit. on pp. 22, 32, 33).

[HS23] Zhen Huang and Bodhisattva Sen. *Multivariate symmetry: Distribution-free testing via optimal transport*. arXiv preprint arXiv:2305.01839 (2023) (cit. on p. 15).

[Hub04] Peter J. Huber. *Robust Statistics*. John Wiley & Sons, 2004 (cit. on p. 31).

[Hub64] Peter J. Huber. *Robust estimation of a location parameter*. Annals of Mathematical Statistics 35 (1964), pp. 73–101 (cit. on p. 31).

[Hub65] Peter J. Huber. *A robust version of the probability ratio test*. Annals of Mathematical Statistics 36 (1965), pp. 1753–1758 (cit. on p. 31).

[Jak17] Martin Emil Jakobsen. *Distance covariance in metric spaces: non-parametric independence testing in metric spaces (Master's thesis)*. arXiv preprint arXiv:1706.03490 (2017) (cit. on p. 24).

[JH16] Julie Josse and Susan Holmes. *Measuring multivariate association and beyond*. Statistics surveys 10 (2016), p. 132 (cit. on p. 19).

[JVV86] Mark R. Jerrum, Leslie G. Valiant, and Vijay V. Vazirani. *Random generation of combinatorial structures from a uniform distribution*. Theoretical Computer Science 43 (1986), pp. 169–188 (cit. on pp. 15, 31).

[KB78] Roger Koenker and Gilbert Bassett. *Regression quantiles*. Econometrica 46 (1978), pp. 33–50 (cit. on pp. 16, 17, 31, 32).

[KBW20] Ilmun Kim, Sivaraman Balakrishnan, and Larry Wasserman. *Robust multivariate nonparametric tests via projection averaging*. The Annals of Statistics 48.6 (2020), pp. 3417–3441 (cit. on p. 24).

[Ken38] Maurice G. Kendall. *A new measure of rank correlation*. Biometrika 30.1-2 (1938), pp. 81–93 (cit. on pp. 19, 73, 76).

[Kin92] John Frank Charles Kingman. *Poisson Processes*. Vol. 3. Clarendon Press, 1992 (cit. on p. 104).

[KM12a] Linglong Kong and Ivan Mizera. *QUANTILE TOMOGRAPHY: USING QUANTILES WITH MULTIVARIATE DATA*. Statistica Sinica 22 (2012), pp. 1589–1610 (cit. on pp. 32, 33).

[KM12b] Linglong Kong and Ivan Mizera. *Quantile tomography: using quantiles with multivariate data*. Statistica Sinica (2012), pp. 1589–1610 (cit. on pp. 21, 23).

[Kol97] Vladimir I. Koltchinskii. *M-estimation, convexity and quantiles*. The Annals of Statistics (1997), pp. 435–477 (cit. on pp. 15, 22, 33, 34, 71, 75).

[Koy87] Robert A. Koyak. *On measuring internal dependence in a set of random variables*. The Annals of Statistics (1987), pp. 1215–1228 (cit. on p. 77).

[KP23] Dimitri Konen and Davy Paindaveine. *Spatial quantiles on the hypersphere*. The Annals of Statistics 51.5 (2023), pp. 2221–2245 (cit. on pp. 22, 71).

[Kro24] Marius Kroll. *Asymptotic Normality of Chatterjee's Rank Correlation*. arXiv preprint arXiv:2408.11547 (2024) (cit. on p. 20).

[LH22] Zhexiao Lin and Fang Han. *Limit theorems of Chatterjee's rank correlation*. arXiv preprint arXiv:2204.08031 (2022) (cit. on p. 20).

[LH23] Zhexiao Lin and Fang Han. *On boosting the power of Chatterjee's rank correlation*. Biometrika 110.2 (2023), pp. 283–299 (cit. on pp. 20, 24, 77).

[LHS13] David Lopez-Paz, Philipp Hennig, and Bernhard Schölkopf. *The randomized dependence coefficient*. Advances in neural information processing systems 26 (2013) (cit. on pp. 19, 77).

[Liu90] Regina Y. Liu. *On a notion of data depth based on random simplices*. The Annals of Statistics (1990), pp. 405–414 (cit. on p. 23).

[Liu92] Regina Y. Liu. *Data depth and multivariate rank tests.* L_1-statistical Analysis and Related mMethods (1992), pp. 279–294 (cit. on p. 23).

[LM00] Beatrice Laurent and Pascal Massart. *Adaptive estimation of a quadratic functional by model selection.* Annals of Statistics 28 (2000), pp. 1302–1338 (cit. on p. 66).

[LM19] Gábor Lugosi and Shahar Mendelson. *Mean estimation and regression under heavy-tailed distributions: A survey.* Foundations of Computational Mathematics 19 (2019), pp. 1145–1190 (cit. on p. 31).

[LM21] Gábor Lugosi and Shahar Mendelson. *Robust multivariate mean estimation: the optimality of trimmed mean.* Annals of Statistics 49 (2021), pp. 393–410 (cit. on pp. 15, 31).

[LP14] Gunther Leobacher and Friedrich Pillichshammer. *Introduction to quasi-Monte Carlo Integration and Applications.* Springer, 2014 (cit. on p. 80).

[LS93] Regina Y. Liu and Kesar Singh. *A quality index based on data depth and multivariate rank tests.* Journal of the American Statistical Association 88.421 (1993), pp. 252–260 (cit. on pp. 15, 75).

[Lyo13] Russell Lyons. *Distance covariance in metric spaces.* The Annals of Probability (2013), pp. 3284–3305 (cit. on p. 24).

[LZ08] Youjuan Li and Ji Zhu. *ℓ_1-norm quantile regression.* Journal of Computational and Graphical Statistics 17 (2008), pp. 163–185 (cit. on p. 31).

[MN24] Tudor Manole and Jonathan Niles-Weed. *Sharp convergence rates for empirical optimal transport with smooth costs.* Annals of Applied Probability 34 (2024), pp. 1108–1135 (cit. on pp. 40, 53, 56–58, 64, 65).

[MO95] Jyrki Möttönen and Hannu Oja. *Multivariate spatial sign and rank methods.* Journaltitle of Nonparametric Statistics 5.2 (1995), pp. 201–213 (cit. on pp. 15, 22, 75).

[Mon81] Gaspard Monge. *Mémoire sur la théorie des déblais et des remblais.* Mem. Math. Phys. Acad. Royale Sci. (1781), pp. 666–704 (cit. on p. 25).

[Mos46] Frederick Mosteller. *On some useful "inefficient" statistics.* The Annals of Mathematical Statistics 17.4 (1946), pp. 377–408 (cit. on p. 19).

[MOT97] Jyrki Möttönen, Hannu Oja, and Juha Tienari. *On the efficiency of multivariate spatial sign and rank tests.* The Annals of Statistics 25.2 (1997), pp. 542–552 (cit. on p. 22).

[MS19] Tamás F. Móri and Gábor J. Székely. *Four simple axioms of dependence measures.* Metrika 82.1 (2019), pp. 1–16 (cit. on pp. 24, 78).

[MS20] Tamás F. Móri and Gábor J. Székely. *The earth mover's correlation.* arXiv preprint arXiv:2009.04313 (2020) (cit. on pp. 19, 24, 77).

[MS22] Gilles Mordant and Johan Segers. *Measuring dependence between random vectors via optimal transport.* Journal of Multivariate Analysis 189 (2022), p. 104912 (cit. on pp. 19, 24, 77).

[MW47] Henry B. Mann and Donald R. Whitney. *On a test of whether one of two random variables is stochastically larger than the other.* The Annals of Mathematical Statistics (1947), pp. 50–60 (cit. on p. 15).

[MZ20] Shahar Mendelson and Nikita Zhivotovskiy. *Robust covariance estimation under $\ell_4 - \ell_2$ norm equivalence*. Annals of Statistics 48 (2020), pp. 1648–1664 (cit. on pp. 15, 31).

[MZ23] Arshak Minasyan and Nikita Zhivotovskiy. *Statistically optimal robust mean and covariance estimation for anisotropic Gaussians*. arXiv preprint arXiv:2301.09024 (2023) (cit. on pp. 15, 31).

[Neu28] John von Neumann. *Zur theorie der gesellschaftsspiele*. Mathematische Annalen 100 (1928), pp. 295–320 (cit. on p. 36).

[Nou12] Ivan Nourdin. *Selected Aspects of Fractional Brownian Motion*. Vol. 4. Springer, 2012 (cit. on p. 84).

[NSM21] Thomas Giacomo Nies, Thomas Staudt, and Axel Munk. *Transport dependency: Optimal transport based dependency measures*. arXiv preprint arXiv:2105.02073 (2021) (cit. on pp. 19, 77).

[NT13] Nam H. Nguyen and Trac D. Tran. *Exact Recoverability From Dense Corrupted Observations via ℓ_1-Minimization*. IEEE Transactions on Information Theory 59 (2013), pp. 2017–2035 (cit. on p. 31).

[NWD16] Preetam Nandy, Luca Weihs, and Mathias Drton. *Large-sample theory for the Bergsma-Dassios sign covariance*. Electronic Journal of Statistics 10 (2016), pp. 2287–2311 (cit. on p. 19).

[NY83] Arkadij Semenovič Nemirovskij and David Borisovich Yudin. *Problem Complexity and Method Efficiency in Optimization*. Wiley-Interscience, 1983 (cit. on pp. 15, 31).

[Oza+19] Sherjil Ozair, Corey Lynch, Yoshua Bengio, Aaron Van den Oord, Sergey Levine, and Pierre Sermanet. *Wasserstein dependency measure for representation learning*. Advances in Neural Information Processing Systems 32 (2019) (cit. on p. 24).

[PA24] Asif Pervez and Irfan Ali. *Robust regression analysis in analyzing financial performance of public sector banks: A case study of India*. Annals of Data Science 11.2 (2024), pp. 677–691 (cit. on p. 15).

[Pea20] Karl Pearson. *Notes on the history of correlation*. Biometrika 13.1 (1920), pp. 25–45 (cit. on pp. 19, 76).

[Pfi+18] Niklas Pfister, Peter Bühlmann, Bernhard Schölkopf, and Jonas Peters. *Kernel-based tests for joint independence*. Journal of the Royal Statistical Society Series B: Statistical Methodology 80.1 (2018), pp. 5–31 (cit. on p. 19).

[PJL20] Ankit Pensia, Varun Jog, and Po-Ling Loh. *Robust regression with covariate filtering: Heavy tails and adversarial contamination*. arXiv preprint arXiv:2009.12976 (2020) (cit. on p. 31).

[PŠ12] Davy Paindaveine and Miroslav Šiman. *Computing multiple-output regression quantile regions*. Computational Statistics & Data Analysis 56.4 (2012), pp. 840–853 (cit. on p. 22).

[PS66] Madan Lal Puri and Pranab Kumar Sen. *On a class of multivariate multisample rank order tests*. Sankhya A 28.4 (1966), pp. 353–376 (cit. on p. 21).

[PS67] Madan Lal Puri and Pranab Kumar Sen. *A class of rank-order tests for a general linear hypothesis*. Nonparametric Methods in Statistics and Related Topics (1967), p. 109 (cit. on p. 21).

[PS71] Madan Lal Puri and Pranab Kumar Sen. *Nonparametric Methods in Multivariate Analysis*. John Wiley & Sons, New York, 1971 (cit. on p. 21).

[PW16] Yury Polyanskiy and Yihong Wu. *Wasserstein continuity of entropy and outer bounds for interference channels*. IEEE Transactions on Information Theory 62 (2016), pp. 3992–4002 (cit. on p. 39).

[PW25] Yury Polyanskiy and Yihong Wu. *Information theory: From coding to learning*. Cambridge university press, 2025 (cit. on pp. 78, 106, 107).

[Ram+15] Aaditya Ramdas, Sashank Jakkam Reddi, Barnabás Póczos, Aarti Singh, and Larry Wasserman. *On the decreasing power of kernel and distance based nonparametric hypothesis tests in high dimensions*. Proceedings of the AAAI Conference on Artificial Intelligence. Vol. 29. 1. 2015 (cit. on pp. 19, 77).

[Rén59] A. Rényi. *On measures of dependence*. Acta Mathematica Academiae Scientiarum Hungarica 10 (1959), pp. 441–451 (cit. on p. 78).

[RH99] Peter J. Rousseeuw and Mia Hubert. *Regression depth*. Journal of the American Statistical Association 94.446 (1999), pp. 388–402 (cit. on p. 23).

[Ros75] Murray Rosenblatt. *A quadratic measure of deviation of two-dimensional density estimates and a test of independence*. The Annals of Statistics (1975), pp. 1–14 (cit. on p. 19).

[Rot12] Michael Roth. *On the multivariate t-distribution*. Linköping University Electronic Press, 2012 (cit. on p. 21).

[RZW17] Pratyaydipta Rudra, Yihui Zhou, and Fred A. Wright. *A procedure to detect general association based on concentration of ranks*. Stat 6.1 (2017), pp. 88–101 (cit. on p. 77).

[Sad22] Behnam Sadeghi. *Chatterjee correlation coefficient: A robust alternative for classic correlation methods in geochemical studies-(including "TripleCpy" Python package)*. Ore Geology Reviews 146 (2022), p. 104954 (cit. on pp. 20, 73).

[SDH22a] Hongjian Shi, Mathias Drton, and Fang Han. *Distribution-free consistent independence tests via center-outward ranks and signs*. Journal of the American Statistical Association 117.537 (2022), pp. 395–410 (cit. on pp. 15, 27, 29, 77).

[SDH22b] Hongjian Shi, Mathias Drton, and Fang Han. *On the power of Chatterjee's rank correlation*. Biometrika 109.2 (2022), pp. 317–333 (cit. on pp. 20, 24, 77, 83, 86).

[SDH24] Hongjian Shi, Mathias Drton, and Fang Han. *On Azadkia–Chatterjee's conditional dependence coefficient*. Bernoulli 30.2 (2024), pp. 851–877 (cit. on pp. 20, 24, 77).

[SDS24] Christopher Strothmann, Holger Dette, and Karl Friedrich Siburg. *Rearranged dependence measures*. Bernoulli 30.2 (2024), pp. 1055–1078 (cit. on pp. 19, 77).

[Sej+13] Dino Sejdinovic, Bharath Sriperumbudur, Arthur Gretton, and Kenji Fukumizu. *Equivalence of distance-based and RKHS-based statistics in hypothesis testing*. The annals of statistics (2013), pp. 2263–2291 (cit. on p. 24).

[SF20] Takeyuki Sasai and Hironori Fujisawa. *Robust estimation with Lasso when outputs are adversarially contaminated*. arXiv preprint arXiv:2004.05990 (2020) (cit. on p. 31).

[Shi+21] Hongjian Shi, Mathias Drton, Marc Hallin, and Fang Han. *Semiparametrically efficient tests of multivariate independence using center-outward quadrant, Spearman, and Kendall statistics*. arXiv preprint arXiv:2111.15567 (2021) (cit. on pp. 15, 29).

[Shi+22] Hongjian Shi, Marc Hallin, Mathias Drton, and Fang Han. *On universally consistent and fully distribution-free rank tests of vector independence*. The Annals of Statistics 50.4 (2022), pp. 1933–1959 (cit. on pp. 15, 29, 77).

[Shi+24] Hongjian Shi, Mathias Drton, Marc Hallin, and Fang Han. *Distribution-free tests of multivariate independence based on center-outward quadrant, Spearman, Kendall, and van der Waerden statistics*. arXiv preprint arXiv:2111.15567 (2024) (cit. on pp. 33, 77).

[Sid57] Siegel Sidney. *Nonparametric statistics for the behavioral sciences*. The Journal of Nervous and Mental Disease 125.3 (1957), p. 497 (cit. on p. 15).

[Skl59] M. Sklar. *Fonctions de répartition à n dimensions et leurs marges*. Annales de l'ISUP. Vol. 8. 3. 1959, pp. 229–231 (cit. on pp. 19, 77).

[SKR23] Shubhanshu Shekhar, Ilmun Kim, and Aaditya Ramdas. *A permutation-free kernel independence test*. Journal of Machine Learning Research 24.369 (2023), pp. 1–68 (cit. on p. 77).

[SP18] Shashank Singh and Barnabás Póczos. *Minimax distribution estimation in Wasserstein distance*. arXiv preprint arXiv:1802.08855 (2018) (cit. on p. 39).

[SP67] Pranab Kumar Sen and Madan Lal Puri. *On the theory of rank order tests for location in the multivariate one sample problem*. Ann. Math. Statist 38 (1967), pp. 1216–1228 (cit. on p. 21).

[Spe04] Charles Spearman. *The proof and measurement of association between two things*. The American Journal of Psychology 15.1 (1904), pp. 72–101 (cit. on pp. 19, 73, 76).

[SR09] Gábor J Székely and Maria L. Rizzo. *Brownian distance covariance*. The Annals of Applied Statistics (2009), pp. 1236–1265 (cit. on pp. 19, 24, 77).

[SRB07] Gábor J. Székely, Maria L. Rizzo, and Nail K. Bakirov. *Measuring and testing dependence by correlation of distances*. Annals of Statistics 35.6 (2007), pp. 2769–2794 (cit. on pp. 19, 23, 28, 77, 81, 82).

[SS14] Arnab Sen and Bodhisattva Sen. *Testing independence and goodness-of-fit in linear models*. Biometrika 101.4 (2014), pp. 927–942 (cit. on pp. 19, 77).

[Suo+24] Chenqu Suo, Krzysztof Polanski, Emma Dann, Rik GH Lindeboom, Roser Vilarrasa-Blasi, Roser Vento-Tormo, Muzlifah Haniffa, Kerstin B. Meyer, Lisa M. Dratva, Zewen Kelvin Tuong, et al. *Dandelion uses the single-cell adaptive immune receptor repertoire to explore lymphocyte developmental origins*. Nature Biotechnology 42.1 (2024), pp. 40–51 (cit. on pp. 20, 73).

[SW81] Berthold Schweizer and Edward F. Wolff. *On nonparametric measures of dependence for random variables*. The Annals of Statistics 9.4 (1981), pp. 879–885 (cit. on pp. 19, 77).

[Sze+14] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. *Intriguing properties of neural networks*. ICLR. 2014 (cit. on p. 31).

[SZF20] Qiang Sun, Wen-Xin Zhou, and Jianqing Fan. *Adaptive Huber regression*. Journal of the American Statistical Association 115 (2020), pp. 254–265 (cit. on p. 31).

[TM63] John W. Tukey and Donald H. McLaughlin. *Less vulnerable confidence and significance procedures for location based on a single sample: Trimming/Winsorization 1*. Sankhyā: The Indian Journal of Statistics, Series A 25 (1963), pp. 331–352 (cit. on pp. 15, 31).

[TOS22] Dag Tjøstheim, Håkon Otneim, and Bård Støve. *Statistical dependence: Beyond Pearson's ρ*. Statistical science 37.1 (2022), pp. 90–109 (cit. on pp. 19, 77).

[Tuk75] John W. Tukey. *Mathematics and the picturing of data*. Proceedings of the International Congress of Mathematicians. Vol. 2. 1975, pp. 523–531 (cit. on pp. 15, 23, 32, 75).

[Van00] Aad W Van der Vaart. *Asymptotic Statistics*. Vol. 3. Cambridge university press, 2000 (cit. on p. 17).

[VC15] Vladimir N. Vapnik and A. Ya Chervonenkis. *On the uniform convergence of relative frequencies of events to their probabilities*. Measures of complexity: festschrift for alexey chervonenkis. Springer, 2015, pp. 11–30 (cit. on p. 56).

[Ver18] Roman Vershynin. *High-Dimensional Probability: An Introduction with Applications in Data Science*. Cambridge University Press, 2018 (cit. on p. 37).

[VG20] Alexander H.S. Vargo and Anna C. Gilbert. *A rank-based marker selection method for high throughput scRNA-seq data*. BMC Bioinformatics 21 (2020), pp. 1–51 (cit. on p. 15).

[Vil09] Cédric Villani. *Optimal transport: old and new*. Springer, 2009 (cit. on pp. 49, 69).

[Vil21] Cédric Villani. *Topics in optimal transportation*. American Mathematical Society, 2021 (cit. on pp. 43, 46, 47, 53, 69, 106).

[Vla+20] Mariia Vladimirova, Stéphane Girard, Hien Nguyen, and Julyan Arbel. *Sub-Weibull distributions: Generalizing sub-Gaussian and sub-Exponential properties to heavier tailed distributions*. Stat 9 (2020), e318 (cit. on p. 67).

[VW96] Aad W. van der Vaart and John A. Wellner. *Weak Convergence and Empirical Processes*. Springer, New York, 1996 (cit. on p. 37).

[VZ00] Yehuda Vardi and Cun-Hui Zhang. *The multivariate L_1-median and associated data depth*. Proceedings of the National Academy of Sciences 97.4 (2000), pp. 1423–1426 (cit. on p. 23).

[Wai19] Martin J. Wainwright. *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*. Vol. 48. Cambridge University Press, 2019 (cit. on pp. 61–65, 94).

[WDL16] Luca Weihs, Mathias Drton, and Dennis Leung. *Efficient computation of the Bergsma–Dassios sign covariance*. Computational Statistics 31 (2016), pp. 315–328 (cit. on p. 19).

[WGX21] Jie Wang, Rui Gao, and Yao Xie. *Two-sample test using projected Wasserstein distance*. 2021 IEEE International Symposium on Information Theory (ISIT). IEEE. 2021, pp. 3320–3325 (cit. on p. 39).

[WGX22] Jie Wang, Rui Gao, and Yao Xie. *Two-sample test with kernel projected Wasserstein distance*. Proceedings of The 25th International Conference on Artificial Intelligence and Statistics. Vol. 151. Proceedings of Machine Learning Research. PMLR, 28–30 Mar 2022, pp. 8022–8055 (cit. on p. 39).

[Wie22] Johannes C.W. Wiesel. *Measuring association with Wasserstein distances*. Bernoulli 28.4 (2022), pp. 2816–2832 (cit. on pp. 19, 24, 77).

[Wil38] Samuel S. Wilks. *The large-sample distribution of the likelihood ratio for testing composite hypotheses*. The Annals of Mathematical Statistics 9.1 (1938), pp. 60–62 (cit. on p. 19).

[Wil92] Frank Wilcoxon. *Individual comparisons by ranking methods*. Breakthroughs in Statistics: Methodology and Distribution. Springer, 1992, pp. 196–202 (cit. on p. 15).

[WL09] Yichao Wu and Yufeng Liu. *Variable selection in quantile regression*. Statistica Sinica 19 (2009), pp. 801–817 (cit. on p. 31).

[WLJ07] Hansheng Wang, Guodong Li, and Guohua Jiang. *Robust regression shrinkage and consistent variable selection through the LAD-Lasso*. Journal of Business & Economic Statistics 25 (2007), pp. 347–355 (cit. on p. 31).

[WPL15] Lan Wang, Bo Peng, and Runze Li. *A high-dimensional nonparametric multivariate test for mean vector*. Journal of the American Statistical Association 110 (2015), pp. 1658–1669 (cit. on p. 31).

[XW19] Yijun Xiao and William Yang Wang. *Disentangled representation learning with Wasserstein total correlation*. arXiv preprint arXiv:1912.12818 (2019) (cit. on p. 24).

[Yan70] Takemi Yanagimoto. *On measures of association and a related problem*. Annals of the Institute of Statistical Mathematics 22.1 (1970), pp. 57–63 (cit. on p. 19).

[YW24] Xuzhi Yang and Tengyao Wang. *Multiple-output composite quantile regression through an optimal transport lens*. arXiv preprint arXiv:2402.09098 (2024) (cit. on p. 15).

[ZS00] Yijun Zuo and Robert Serfling. *General notions of statistical depth function*. Annals of Statistics (2000), pp. 461–482 (cit. on pp. 15, 23, 75).

[Zuo03] Yijun Zuo. *Projection-based depth functions and associated medians*. The Annals of Statistics 31.5 (2003), pp. 1460–1490 (cit. on p. 23).

[Zuo21] Yijun Zuo. *On general notions of depth for regression* (2021) (cit. on p. 23).

[ZY08] Hui Zou and Ming Yuan. *Composite quantile regression and the oracle model selection theory*. Annals of Statistics 36 (2008), pp. 1108–1126 (cit. on pp. 16, 18, 31, 32, 36).

BIBLIOGRAPHY