Multiple-output composite quantile regression through an optimal transport lens

Xuzhi Yang X.YANG64@LSE.AC.UK
London School of Economics and Political Science

Tengyao Wang T.WANG59@LSE.AC.UK
London School of Economics and Political Science

Editors: Shipra Agrawal and Aaron Roth

Abstract

Composite quantile regression has been used to obtain robust estimators of regression coefficients in linear models with good statistical efficiency. By revealing an intrinsic link between the composite quantile regression loss function and the Wasserstein distance from the residuals to the set of quantiles, we establish a generalization of the composite quantile regression to the multiple-output settings. Theoretical convergence rates of the proposed estimator are derived both under the setting where the additive error possesses only a finite ℓ-th moment (for $\ell > 2$) and where it exhibits a sub-Weibull tail. In doing so, we develop novel techniques for analyzing the M-estimation problem that involves Wasserstein-distance in the loss. Numerical studies confirm the practical effectiveness of our proposed procedure.

Keywords: quantile regression, optimal transport, multivariate quantiles, robust estimation

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 (Eklund et al., 2016; Wang et al., 2015; Szegedy et al., 2014). Compared to earlier works in this area pioneered by Tukey and McLaughlin (1963) and Huber (1964, 1965), modern treatment of this topic focuses more on handling multivariate data. For instance, in the area of robust mean estimation, Diakonikolas et al. (2020); Lugosi and Mendelson (2021); Depersin and Lecué (2022); Minasyan and Zhivotovskiy (2023) have proposed various extensions of univariate robust mean procedures such as the trimmed mean estimator (Tukey and McLaughlin, 1963) and median of means estimator (Nemirovskij and Yudin, 1983; Jerrum et al., 1986; Alon et al., 1996) to the multivariate setting. We witness a similar surge in research interest in the area of robust covariance estimation (Mendelson and Zhivotovskiy, 2020; Abdalla and Zhivotovskiy, 2022; Minasyan and Zhivotovskiy, 2023).

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

$$Y = b^* X + \varepsilon, \tag{1}$$

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 (Catoni, 2012; Lugosi and Mendelson, 2019) and outlier contamination (Szegedy et al., 2014; Huber, 2004). When $d = 1$, both directions have thrived in recent years (Nguyen and Tran, 2013; Fan et al., 2017; Sun et al., 2020; Sasai and Fujisawa, 2020; Pensia et al., 2020; Adomaityte et al., 2023). 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 (Koenker and Bassett, 1978; Wang et al., 2007; Li and Zhu, 2008; Zou and Yuan, 2008; Wu and Liu, 2009; Belloni and Chernozhukov, 2011). 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 (Koenker and Bassett, 1978). 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 (2008) 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

$$(\hat{q}_1, \ldots, \hat{q}_K, \tilde{b}) = \underset{q_1,\ldots,q_K \in \mathbb{R},\ b \in \mathbb{R}^{d\times p}}{\arg\min} \sum_{i=1}^{n} \sum_{k=1}^{K} \rho_{\tau_k}(Y_i - bX_i - q_k), \tag{2}$$

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 (2008) 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. (2017) and Hallin et al. (2021). 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 (Tukey, 1975), 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 (2022) 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 $(Y_i - bX_i : i = 1, \ldots, n)$ 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 (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 (Hallin et al., 2010; Kong and Mizera, 2012; Hallin et al., 2015; Carlier et al., 2016; del Barrio et al., 2022), the majority have concentrated on estimating the quantile contours rather than focusing on the robust estimation of the regression coefficients. See Section 2 for a more detailed discussion of our proposed method.

Then in Section 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 5). Then a faster convergence rate is established when we assume a noise distribution with a sub-Weibull tail (see Theorem 8). 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 5 and Theorem 8 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 3 for detailed descriptions of the Theorems and proof sketches.

1.1. Related works

Various definitions of multiple-output quantile regression have been proposed in the past, including the depth-based directional method (Hallin et al., 2010; Kong and Mizera, 2012; Hallin et al., 2015), the M-quantile (Koltchinskii, 1997), the spatial quantile (Chaudhuri, 1996; Chakraborty and Chaudhuri, 2014), 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 (Hallin et al., 2021). 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. (2017) and Hallin et al. (2021) 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 (Chernozhukov et al., 2017; Hallin et al., 2021). Several applications in multivariate statistics have been established successfully (Deb and Sen, 2021; del Barrio et al., 2022; Hallin et al., 2023; Shi et al., 2024). We refer to a comprehensive survey Hallin (2022) and references therein. Building upon this groundwork, Carlier et al. (2016) and del Barrio et al. (2022) proposed two notions of multiple-output quantile regression, though concentrating primarily on the estimation of conditional quantile functions rather than the regression coefficients themselves.

1.2. Notation

For $n \in \mathbb{N}$, write $[n] := \{1, \ldots, n\}$. 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 $\|M\|_{\mathrm{F}} := \big(\mathrm{Tr}(M^\top M)\big)^{1/2}$. 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 $X_1, \ldots, X_n$ are n independent copies of X and δ_x denote the Dirac measure on x.

2. The MCQR construction

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.

2.1. Univariate CQR revisited

Since $q_1, \ldots, q_K$ in (2) have the interpretation of quantiles associated with $\tau_1, \ldots, \tau_K$, 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) with this additional constraint can be viewed as the empirical version of the following optimization problem

$$\operatorname*{arg\,min}_{q\in\mathcal{M},\, b\in\mathbb{R}^{1\times p}} \mathbb{E}\Big\{\rho_T\big(Y - bX - q(T)\big)\Big\} = \operatorname*{arg\,min}_{q\in\mathcal{M},\, b\in\mathbb{R}^{1\times p}} \mathbb{E}\Big\{\int_0^1 \rho_\tau\big(Y - bX - q(\tau)\big)\mathrm{d}\tau\Big\}, \tag{3}$$

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 (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 (3). As we will see from Lemma 2 and Proposition 3, this is actually the unique solution to the problem.

Lemma 1 *Under the linear model* (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 A.3) of the above lemma reveals that if $\tau_1, \ldots, \tau_K$ converges to a distribution P^Z with support $\mathcal{Z}$ rather than to $U[0, 1]$, then a similar result to Lemma 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 (Chaudhuri, 1996; Koltchinskii, 1997), 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 1.1, as well as empirical comparisons in Section 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 (3) in Lemma 2 below. To state the lemma, we define the *Wasserstein product* between $P, Q \in \mathcal{P}_2(\mathbb{R}^d)$ as

$$\langle\!\langle P, Q \rangle\!\rangle_{\mathcal{W}_2} := \sup_{\gamma \in \mathcal{C}(P,Q)} \int \langle x, y \rangle \, \mathrm{d}\gamma(x, y), \tag{4}$$

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 *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 A.4. Writing $\mathcal{L}(b; U) := \langle\!\langle Y - bX, U \rangle\!\rangle_{\mathcal{W}_2}$, Lemma 2 and Equation (3) imply that, the optimizer in b for the population CQR loss function in (3) is equal to $\arg\min_{b \in \mathbb{R}^{d\times p}} \mathcal{L}(b; U)$ when $d = 1$.

2.2. Multiple-output CQR via optimal transport

With the help of Lemma 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 A.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 C for an intuitive illustration).

Proposition 3 *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 (Chernozhukov et al., 2017; Deb and Sen, 2021) and the spherical uniform distribution (Hallin et al., 2021; del Barrio et al., 2022). 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 3 motivates the following natural estimator of b^* based on the Wasserstein product of the empirical distributions.

Definition 4 *Given i.i.d. covariate-response pairs* $(X_1, Y_1), \ldots, (X_n, Y_n)$ *generated as in* (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*

$$\hat{b} \in \underset{b \in \mathbb{R}^{d \times p}}{\arg\min}\, \mathcal{L}_{n,m}(b), \text{ where } \mathcal{L}_{n,m}(b) := \langle\!\langle P_n^{Y-bX}, P_m^U \rangle\!\rangle_{\mathcal{W}_2}. \tag{5}$$

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 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) when $d = 1$. Instead, we employ it as the reference distribution to redefine the distribution function and the quantile function (refer to Section C for an example). Thus, even when $d = 1$ with a uniform reference distribution, the plug-in estimator in (5) does not reduce to the univariate CQR estimator (2). This can also be seen from the proof of Lemma 2. Therefore, our proposed MCQR estimator (5) is different from the univariate CQR estimator that is studied in Zou and Yuan (2008) but shares the same loss function at the population level. See also Figure 3(*a*) and Figure 3(*b*) for an interesting difference in their robustness to contamination in one dimension.

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 (5) can be written as

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

where the exchange of the minimum and maximum is allowed as the objective is linear (von Neumann, 1928). 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 (5) is

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

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.

3. Theoretical guarantees

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 5, 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 8 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 5 *Suppose $(X,Y),(X_1,Y_1),\dots,(X_n,Y_n)$ are i.i.d. pairs generated according to (1), $U_1,\dots,U_m \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,I_d)$. 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 (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 5 is that if taking n and m to be large enough such that

$$C\big(n^{-\frac{1}{4}} + n^{-\frac{1}{d\vee p}} + n^{-\frac{\ell-2}{2\ell}}\big)\log m < 1, \tag{6}$$

then we have

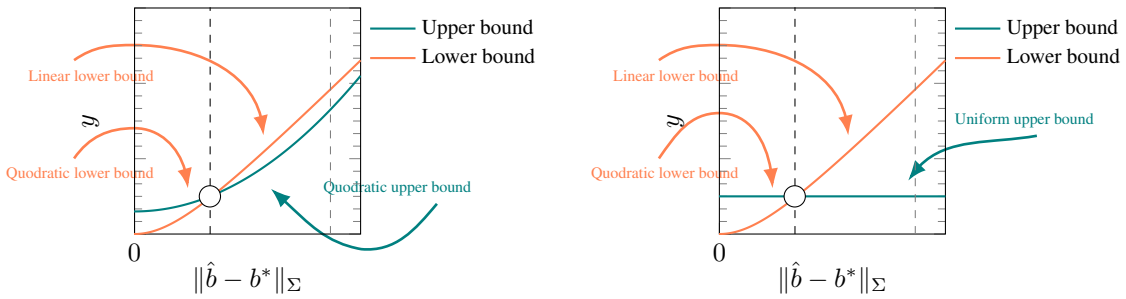
$$\|\hat{b} - b^*\|_\Sigma^2 \leq C\big(n^{-\frac{1}{4}} + n^{-\frac{1}{d\vee p}} + n^{-\frac{\ell-2}{2\ell}}\big)\log m \tag{7}$$

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. (2019) 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 (van der Vaart and Wellner, 1996, 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 5, we first introduce the following lemmas.

Lemma 6 *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*

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



(*a*) The upper and lower bound constructed in the Proof of Theorem 5.

(*b*) The upper and lower bound constructed in the proof of Theorem 8.

Figure 1: Illustration of proofs.

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. Vershynin, 2018, 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 A.7 for the proof. We remark that the lower bound in Lemma 6 is sharp, as can be seen from Lemma S23.

Lemma 7 *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 A.8 for the proof.

Proof sketch of Theorem 5 We start with the basic inequality:

$$\mathcal{L}(\hat{b}) - \mathcal{L}(b^*) \le \mathcal{L}(\hat{b}) - \mathcal{L}_{n,m}(\hat{b}) + \mathcal{L}_{n,m}(b^*) - \mathcal{L}(b^*). \tag{8}$$

The proof strategy involves establishing a lower bound for the left-hand side of (8) with respect to $\|\hat{b} - b^*\|_\Sigma$ and an upper bound for the right-hand side of (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 (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 6 and the explicit form for $\langle\!\langle (b^* - b)X, U \rangle\!\rangle_{\mathcal{W}_2}$ we can show that

$$\mathcal{L}(b) - \mathcal{L}(b^*) \ge \sqrt{r^2 + \|b^* - b\|_\Sigma^2} - r, \tag{9}$$

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 1(*a*) for an illustration).

To upper bound the right-hand side of (8), by applying Lemma 7 we have for each $b \in \mathbb{R}^{d\times p}$,

$$|\mathcal{L}(b) - \mathcal{L}_{n,m}(b)| \le \left(\frac{1}{m}\sum_{i=1}^m \|U_i\|^2\right)^{1/2} \mathcal{W}_2(P^{Y-bX}, P_n^{Y-bX}) + \left(\mathbb{E}\,\|Y - bX\|^2\right)^{1/2} \mathcal{W}_2(P^U, P_m^U). \tag{10}$$

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. Fournier and Guillin, 2015) (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 (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 (9). See Section A.6 for a complete 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*

$$\mathbb{E}\exp\left\{\frac{1}{2}(\|\Sigma^{-1/2}X\|/\sigma_1)^\alpha\right\} \le 2 \quad \textit{and} \quad \mathbb{E}\exp\left\{\frac{1}{2}(\|\varepsilon\|/\sigma_2)^\beta\right\} \le 2 \tag{11}$$

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

$$f_\varepsilon(e) \ge \gamma_1 \exp\left(-\gamma_2\|e\|^2\right), \quad \textit{for } \|e\| \ge 1. \tag{12}$$

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

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

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 S24(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 (12) is a relaxation of the so-called (γ_1, γ_2)-regularity defined in Polyanskiy and Wu (2016). The merit of employing this relaxation becomes apparent when examining Lemma S27, where it is demonstrated that the convolution of two independent probability densities adhering to (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 8 *Under the same setup of Theorem 5 and suppose that Assumptions 2 and 3 are satisfied. For m, n large enough such that (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*

$$\|b^* - \hat{b}\|_\Sigma^2 \le M\big((p/n)^{1/2} + n^{-2/d}\big)(\log m)^{\frac{8}{2\wedge\alpha\wedge\beta}}. \tag{13}$$

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 (14) and Proposition S17), and its minimax optimality has been established in Singh and Póczos (2018). Compared to (7), this improved bound in (13) removes the dependence on p in the exponent. Moreover, unlike the convergence rate result established for the projected Wasserstein distance in Wang et al. (2021, 2022), 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 A.9 for a complete proof.

Proof sketch of Theorem 8 Assume the setting of Theomem 5, error bound (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 (8) indicates the following uniform bound

$$\begin{aligned}\mathcal{L}(\hat{b}) - \mathcal{L}(b^*) &\le 2 \sup_{b\in\mathcal{B}} |\mathcal{L}(b) - \mathcal{L}_{n,m}(b)| \\ &\le \left|\frac{1}{m}\sum_{i=1}^{m} \|U_i\|^2 - \mathbb{E}\,\|U\|^2\right| + \sup_{b\in\mathcal{B}}\left|\frac{1}{n}\sum_{i=1}^{n} \|Y_i - bX_i\|^2 - \mathbb{E}\,\|Y - bX\|^2\right| \\ &\quad + \sup_{b\in\mathcal{B}}\left|\mathcal{W}_2^2(P^{Y-bX}, P^U) - \mathcal{W}_2^2(P_n^{Y-bX}, P_m^U)\right|. \quad (14)\end{aligned}$$

Utilizing the same lower bound for the left-hand side as in (9), it remains to derive an upper bound for the right-hand side of the above inequality. While the initial two terms of (14) can be effectively controlled through the application of statistical concentration arguments, as elucidated in Lemma S21, achieving control over the last term demands much more effort. Motivated by the duality argument presented in Manole and Niles-Weed (2024, Theorem 13), we establish a non-asymptotic *uniform* error bound for the empirical 2-Wasserstein distance (Proposition S17; see also Figure 1(*b*) for an illustration), which forms the key ingredient of the proof. ■

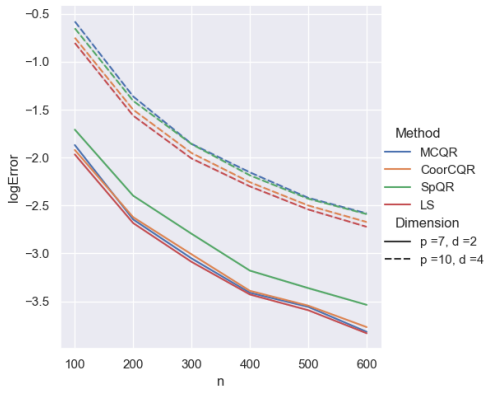
4. Numerical experiments

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 2.3. The competitors used in the simulation studies include the ordinary least squares estimator (LS), the spatial quantile regression (SpQR) with zero quantile level (Chaudhuri, 1996), and coordinate-wise CQR (CoorCQR), i.e. independently applying CQR to each component of the response variable. We refer readers to Section D for more details about SpQR.

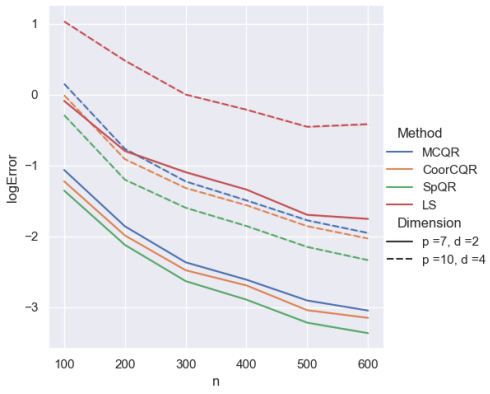
In each experiment, we draw i.i.d. data $(X_1, Y_1), \ldots, (X_n, Y_n)$ according to model (1), where the regression coefficients $b^* \in \mathbb{R}^{d\times p}$ has independent $\mathcal{N}(5, 5)$ entries and is kept fixed for all repetitions. Covariates $X_i \in \mathbb{R}^p$, $i = 1, \ldots, n$, 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 $\Sigma' = (0.9^{|i-j|})_{i,j} \in \mathbb{R}^{d\times d}$

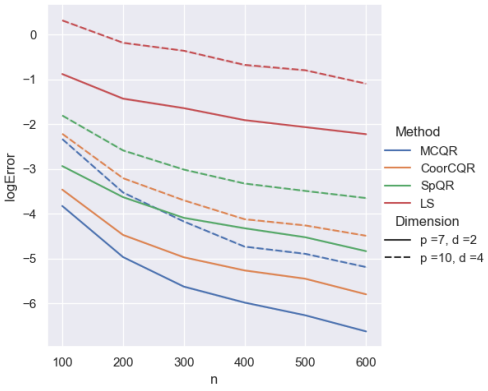
1. the Pareto distribution $\mathrm{Pareto}(k, \alpha, s)$ has density function $f(x) \propto \frac{\alpha s^{\alpha+1}}{(x-k)^{\alpha+1}}$ for all $x \ge 1+k$, with shape parameter $\alpha > 0$, location parameter $k \in \mathbb{R}$ and scale parameter $s > 0$. Here $\mathrm{Pareto}(-2, 2, 1)$ has mean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



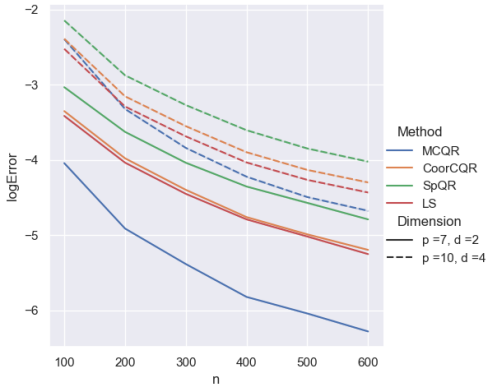
(*a*) Gaussian noise



(*b*) Multivariate t_2 noise



(*c*) Pareto copula noise



(*d*) Banana-shaped noise

Figure 2: 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 4 for various sample size n, covariate dimension p and response dimension d and four different noise distributions (panels (a) to (d)).

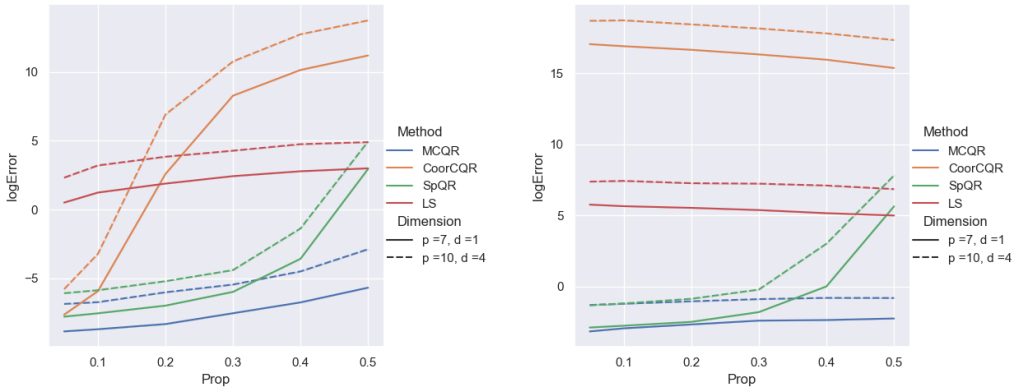
(1d) ε follows a centered Banana-shaped distribution, i.e. $\varepsilon_i \overset{d}{=} (B_{d-1}, \|B_{d-1}\|^2 - \frac{2}{d+2}) + 0.3B_d$, where B_d is uniformly distributed in the unit ball in $\mathbb{R}^d$

Figure 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 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 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 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.

Acknowledgments

This research is funded by EPSRC grant EP/T02772X/1.

References

Pedro Abdalla and Nikita Zhivotovskiy. Covariance estimation: Optimal dimension-free guarantees for adversarial corruption and heavy tails. *arXiv preprint arXiv:2205.08494*, 2022.

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.

Noga Alon, Yossi Matias, and Mario Szegedy. The space complexity of approximating the frequency moments. In *Proceedings of the Twenty-Eighth Annual ACM Symposium on Theory of Computing*, pages 20–29, 1996.

Alexandre Belloni and Victor Chernozhukov. ℓ_1-penalized quantile regression in high-dimensional sparse models. *Annals of Statistics*, 39:82–130, 2011.

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*, 8:657–676, 2019.

Efim M. Bronshtein. ε-entropy of convex sets and functions. *Siberian Mathematical Journal*, 17: 393–398, 1976.

Guillaume Carlier, Victor Chernozhukov, and Alfred Galichon. Vector quantile regression: An optimal transport approach. *Annals of Statistics*, 44:1165–1192, 2016.

Olivier Catoni. Challenging the empirical mean and empirical variance: a deviation study. *Annales de l'IHP Probabilités et statistiques*, 48:1148–1185, 2012.

Anirvan Chakraborty and Probal Chaudhuri. The spatial distribution in infinite dimensional spaces and related quantiles and depths. *Annals of Statistics*, 32:1203–1231, 2014.

Biman Chakraborty. On multivariate quantile regression. *Journal of Statistical Planning and Inference*, 110:109–132, 2003.

Probal Chaudhuri. On a geometric notion of quantiles for multivariate data. *Journal of the American Statistical Association*, 91:862–872, 1996.

Victor Chernozhukov, Alfred Galichon, Marc Hallin, and Marc Henry. Monge–Kantorovich depth, quantiles, ranks and signs. *Annals of Statistics*, 45:223–256, 2017.

Joydeep Chowdhury and Probal Chaudhuri. Nonparametric depth and quantile regression for functional data. *Bernoulli*, 25:395–423, 2019.

Nabarun Deb and Bodhisattva Sen. Multivariate rank-based distribution-free nonparametric testing using measure transportation. *Journal of the American Statistical Association*, 118:1–16, 2021.

Eustasio del Barrio, Alberto Gonzalez Sanz, and Marc Hallin. Nonparametric multiple-output center-outward quantile regression. *arXiv preprint arXiv:2204.11756*, 2022.

Jules Depersin and Guillaume Lecué. Robust sub-Gaussian estimation of a mean vector in nearly linear time. *Annals of Statistics*, 50:511–536, 2022.

Luc Devroye, László Györfi, and Gábor Lugosi. *A Probabilistic Theory of Pattern Recognition*, volume 31. Springer Science & Business Media, 2013.

Ilias Diakonikolas, Daniel M. Kane, and Ankit Pensia. Outlier robust mean estimation with subgaussian rates via stability. In *NeurIPS*, pages 1830–1840, 2020.

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*, 113:7900–7905, 2016.

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*, 79:247–265, 2017.

Nicolas Fournier and Arnaud Guillin. On the rate of convergence in Wasserstein distance of the empirical measure. *Probability Theory and Related Fields*, 162:707–738, 2015.

Matthias Gelbrich. On a formula for the ℓ_2-Wasserstein metric between measures on Euclidean and Hilbert spaces. *Mathematische Nachrichten*, 147:185–203, 1990.

S. Graf and R. Daniel Mauldin. A classification of disintegrations of measures. *Measure and Measurable Dynamics, Contemporary Mathematics*, 94:147–158, 1989.

Marc Hallin. Measure transportation and statistical decision theory. *Annual Review of Statistics and Its Application*, 9:401–424, 2022.

Marc Hallin, Davy Paindaveine, and Miroslav Šiman. Multivariate quantiles and multiple-output regression quantiles: From ℓ_1 optimization to halfspace depth. *Annals of Statistics*, 38:635–703, 2010.

Marc Hallin, Zudi Lu, Davy Paindaveine, and Miroslav Šiman. Local bilinear multiple-output quantile/depth regression. *Bernoulli*, 21, 2015.

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*, 49:1139–1165, 2021.

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*, 118:1923–1939, 2023.

Peter J. Huber. Robust estimation of a location parameter. *Annals of Mathematical Statististics*, 35: 73–101, 1964.

Peter J. Huber. A robust version of the probability ratio test. *Annals of Mathematical Statististics*, 36:1753–1758, 1965.

Peter J. Huber. *Robust Statistics*. John Wiley & Sons, 2004.

Mark R. Jerrum, Leslie G. Valiant, and Vijay V. Vazirani. Random generation of combinatorial structures from a uniform distribution. *Theoretical Computer Science*, 43:169–188, 1986.

Roger Koenker and Gilbert Bassett. Regression quantiles. *Econometrica*, 46:33–50, 1978.

Vladimir I Koltchinskii. M-estimation, convexity and quantiles. *Annals of Statistics*, 25:435–477, 1997.

Dimitri Konen and Davy Paindaveine. Spatial quantiles on the hypersphere. *Annals of Statistics*, 51:2221–2245, 2023.

Linglong Kong and Ivan Mizera. Quantile tomography: Using quantiles with multivariate data. *Statistica Sinica*, 22:1589–1610, 2012.

Beatrice Laurent and Pascal Massart. Adaptive estimation of a quadratic functional by model selection. *Annals of Statistics*, 28:1302–1338, 2000.

Youjuan Li and Ji Zhu. ℓ_1-norm quantile regression. *Journal of Computational and Graphical Statistics*, 17:163–185, 2008.

Gábor Lugosi and Shahar Mendelson. Mean estimation and regression under heavy-tailed distributions: A survey. *Foundations of Computational Mathematics*, 19:1145–1190, 2019.

Gábor Lugosi and Shahar Mendelson. Robust multivariate mean estimation: the optimality of trimmed mean. *Annals of Statistics*, 49:393–410, 2021.

Tudor Manole and Jonathan Niles-Weed. Sharp convergence rates for empirical optimal transport with smooth costs. *Annals of Applied Probability*, 34:1108–1135, 2024.

Shahar Mendelson and Nikita Zhivotovskiy. Robust covariance estimation under $\ell_4 - \ell_2$ norm equivalence. *Annals of Statistics*, 48:1648–1664, 2020.

Arshak Minasyan and Nikita Zhivotovskiy. Statistically optimal robust mean and covariance estimation for anisotropic Gaussians. *arXiv preprint arXiv:2301.09024*, 2023.

Arkadij Semenovič Nemirovskij and David Borisovich Yudin. *Problem Complexity and Method Efficiency in Optimization*. Wiley-Interscience, 1983.

Nam H. Nguyen and Trac D. Tran. Exact recoverability from dense corrupted observations via ℓ_1-minimization. *IEEE Transactions on Information Theory*, 59:2017–2035, 2013.

Ankit Pensia, Varun Jog, and Po-Ling Loh. Robust regression with covariate filtering: Heavy tails and adversarial contamination. *arXiv preprint arXiv:2009.12976*, 2020.

Yury Polyanskiy and Yihong Wu. Wasserstein continuity of entropy and outer bounds for interference channels. *IEEE Transactions on Information Theory*, 62:3992–4002, 2016.

Takeyuki Sasai and Hironori Fujisawa. Robust estimation with lasso when outputs are adversarially contaminated. *arXiv preprint arXiv:2004.05990*, 2020.

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.

Shashank Singh and Barnabás Póczos. Minimax distribution estimation in Wasserstein distance. *arXiv preprint arXiv:1802.08855*, 2018.

Qiang Sun, Wen-Xin Zhou, and Jianqing Fan. Adaptive Huber regression. *Journal of the American Statistical Association*, 115:254–265, 2020.

Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. In *ICLR*, 2014.

John W. Tukey. Mathematics and the picturing of data. In *Proceedings of the International Congress of Mathematicians*, volume 2, pages 523–531, 1975.

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*, 25:331–352, 1963.

Aad W. van der Vaart and John A. Wellner. *Weak Convergence and Empirical Processes*. Springer, New York, 1996.

Vladimir N Vapnik and A Ya Chervonenkis. On the uniform convergence of relative frequencies of events to their probabilities. In *Measures of complexity: Festschrift for Alexey Chervonenkis*, pages 11–30. Springer, 2015.

Roman Vershynin. *High-Dimensional Probability: An Introduction with Applications in Data Science*. Cambridge University Press, 2018.

Cédric Villani. *Optimal Transport: Old and New*. Springer, 2009.

Cédric Villani. *Topics in Optimal Transportation*. American Mathematical Society, 2021.

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: e318, 2020.

John von Neumann. Zur theorie der gesellschaftsspiele. *Mathematische Annalen*, 100:295–320, 1928.

Martin J. Wainwright. *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*, volume 48. Cambridge University Press, 2019.

Hansheng Wang, Guodong Li, and Guohua Jiang. Robust regression shrinkage and consistent variable selection through the LAD-Lasso. *Journal of Business & Economic Statistics*, 25:347–355, 2007.

Jie Wang, Rui Gao, and Yao Xie. Two-sample test using projected Wasserstein distance. In *2021 IEEE International Symposium on Information Theory (ISIT)*, pages 3320–3325. IEEE, 2021.

Jie Wang, Rui Gao, and Yao Xie. Two-sample test with kernel projected Wasserstein distance. In *AISTATS*, volume 151, pages 8022–8055, 2022.

Lan Wang, Bo Peng, and Runze Li. A high-dimensional nonparametric multivariate test for mean vector. *Journal of the American Statistical Association*, 110:1658–1669, 2015.

Yichao Wu and Yufeng Liu. Variable selection in quantile regression. *Statistica Sinica*, 19:801–817, 2009.

Hui Zou and Ming Yuan. Composite quantile regression and the oracle model selection theory. *Annals of Statistics*, 36:1108–1126, 2008.

Supplement to "Multiple-output composite quantile regression through an optimal transport lens"

We presents the proofs of all main results in Appendix A. Specifically, Appendix A.3 - A.5 contain proof of the theoretical results in Section 2. Then the proof of Lemma 6 and Lemma 7 are included in Appendix A.7 and A.8, respectively. The consistency Theorem 5 is proved in Appendix A.6, while the proof regarding the convergence rate, as specified in Theorem 8, is showed in A.9. All ancillary results are included in Appendix B.

Appendix C provides an example of a check function under a $U[-1,1]$ reference distribution. A brief introduction about spatial quantile is provided in Appendix D

Appendix A. Proofs

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

A.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

$$\frac{1}{2}\mathcal{W}_2^2(P,Q) = \inf_{\gamma \in \mathcal{C}(P,Q)} \int L_2(x,y) d\gamma(x,y) =: I_2(P,Q). \tag{S15}$$

Our proof depends on the following Kantorovich duality (see e.g., Villani, 2021, Theorem 1.3)

$$I_2(P,Q) = \sup_{\varphi,\psi \in \Phi_2} J_{P,Q}(\varphi,\psi), \tag{S16}$$

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 $\tilde{\Phi} := \{(\varphi,\psi) \in L^1(P) \times L^1(Q) : \varphi(x) + \psi(y) \ge x^T y\}$ that

$$\int \frac{\|x\|^2}{2} dP(x) + \int \frac{\|y\|^2}{2} dQ(y) - \sup_{\varphi,\psi \in \Phi_2} J_{P,Q}(\varphi,\psi) = \inf_{\varphi,\psi \in \tilde{\Phi}} J_{P,Q}(\varphi,\psi) := \tilde{I}_2(P,Q). \tag{S17}$$

Thus solve the problem of (S16) degenerates to solve the problem of $\tilde{I}_2(P,Q)$.

For any $\varphi \in L^1(P)$, define its *Legendre transform* as $\varphi^*(y) := \sup_{x \in \mathbb{R}^d}(x^T y - \varphi(x))$. 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 $\varphi(x) \ge \sup_{y \in \mathbb{R}^d}\big(x^T y - \varphi^*(y)\big) = \varphi^{**}(x)$, $\forall x \in \mathbb{R}^d$, 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. Villani, 2021, 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 S9 *(Villani, 2021, 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 S10 (Kantorovich–Rubinstein theorem) *Suppose $\mathcal{X}$ is a subset of $\mathbb{R}^d$, define the diameter of $\mathcal{X}$ as $\mathrm{diam}(\mathcal{X}) := \sup_{x,y\in\mathcal{X}} \|x - y\|$. Let $\mathrm{Lip}(\mathcal{X})$ denote the space of all Lipschitz function on $\mathcal{X}$ and for any f within this space define*

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

Then

$$\mathcal{W}_1(P, Q) = \sup\left\{ \int f(x)dP(x) - \int f(y)dQ(y) : f \in \mathrm{L}^1(|P - Q|),\ f \in \mathrm{Lip}_1(\mathcal{X}) \right\}, \quad \text{(S18)}$$

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

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 S11 (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*

$$\mathrm{IPM}(P, Q; \mathcal{F}) = \sup_{f\in\mathcal{F}} \Big\{ \int f(x)dP(x) - \int f(y)dQ(y) \Big\}. \quad \text{(S19)}$$

A.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

$$\mathcal{R}_n(\mathcal{F}, P) := \mathbb{E}\Big(\sup_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^{n} \xi_i f(X_i)\Big), \tag{S20}$$

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 $\mathcal{B}^p_{x,r} := \{y \in \mathbb{R}^p : \|y\| \le r\}$ and we omit r when $r = 1$: $\mathcal{B}^p_{x,1} := \mathcal{B}^p_x$. The matrix operator norm is denoted by $\|\cdot\|_{\mathrm{op}}$, so that $\|A\|_{\mathrm{op}} := \sup_{x:\|x\|=1} \|Ax\|$.

A.3. Proof for Lemma 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 (1) we have for any $x \in \mathbb{R}^p$,

$$(b^*, q^*_\varepsilon(\tau)) \in \underset{b\in\mathbb{R}^{1\times d}, q\in\mathbb{R}}{\arg\min}\ \mathbb{E}[\rho_\tau(Y - bX - q) \mid X = x]. \tag{S21}$$

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 (S21) 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

$$\mathbb{E}[\rho_\tau(Y - b^*X - q^*_\varepsilon(\tau))] \le \mathbb{E}[\rho_\tau(Y - bX - q)]. \tag{S22}$$

Because the quantile function $q^*_\varepsilon \in \mathcal{M}$, thus (S22) 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. ■

A.4. Proof for Lemma 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

$$\begin{aligned}
\inf_{q\in\mathcal{M}} \mathbb{E}\Big\{\int_0^1 \rho_\tau\big(Y - bX - q(\tau)\big)d\tau\Big\} + \frac{1}{2}\mathbb{E}\,Y & \\
&= \inf_{q\in\mathcal{M}}\Big\{\mathbb{E}\int_0^1 (Y - q(\tau) - bX)^+ d\tau + \int_0^1 (1-\tau)q(\tau)d\tau\Big\} \\
&= \inf_{q\in\mathcal{M}}\Big\{\mathbb{E}\max_{t\in[0,1]}\int_0^t (Y - q(\tau) - bX)\,d\tau + \int_0^1\int_\tau^1 q(\tau)\,du\,d\tau\Big\} \\
&= \inf_{\phi\in\mathcal{C}}\Big\{\mathbb{E}\max_{t\in[0,1]}(t(Y-bX) - \phi(t)) + \mathbb{E}\phi(U)\Big\} \\
&= \inf_{\phi\in\mathcal{C}}\mathbb{E}\big\{\phi^*(Y-bX) + \mathbb{E}\phi(U)\big\}, \qquad \text{(S23)}
\end{aligned}$$

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 (S23) and ϕ_0^* its Legendre conjugate, then by Villani (2021, 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 (2021, 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.

$$\mathbb{E}\tilde\psi_0(Y-bX) + \mathbb{E}\tilde\phi_0(U) = \sup_{\substack{\tilde\phi,\tilde\psi\in L^1(\mathbb{R}) \\ \tilde\phi(x)+\tilde\psi(y)\leq(x-y)^2/2}} \mathbb{E}\tilde\psi(Y-bX) + \mathbb{E}\tilde\phi(U). \qquad \text{(S24)}$$

By the strong duality theorem (Villani, 2021, Theorem 1.3), we have

$$\begin{aligned}
\frac{1}{2}\mathcal{W}_2^2\big(P^{Y-bX}, P^U\big) &= \mathbb{E}\tilde\psi_0(Y-bX) + \mathbb{E}\tilde\phi_0(U) \\
&= \mathbb{E}\Big\{\frac{1}{2}(Y-bX)^2 - \phi_0^*(Y-bX)\Big\} + \mathbb{E}\Big\{\frac{1}{2}U^2 - \phi_0(U)\Big\}, \qquad \text{(S25)}
\end{aligned}$$

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 (S23). ■

A.5. Proof for Proposition 3

Proof By Brenier's Theorem, there exists a P^U-a.e. unique (and P^ε-a.e. 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

$$\begin{aligned}\int \|u - v\|_2^2 \, dP^{(U,Y-bX)}(u, v) &= \int \|u - (e + z)\|_2^2 \, d(P^{(U,\varepsilon)} \otimes P^{(b^*-b)X})(u, e, z) \\ &= \int \|\phi(u) - u\|_2^2 \, dP^U(u) + \int \|z\|_2^2 \, dP^{(b^*-b)X}(z) \\ &= \mathcal{W}_2^2(P^U, P^\varepsilon) + \mathbb{E}\,\|(b^* - b)X\|_2^2. \end{aligned} \tag{S26}$$

Therefore, we have

$$\begin{aligned}\mathcal{L}(b; U) - \mathcal{L}(b^*; U) &= -\frac{1}{2}\mathcal{W}_2^2(P^U, P^{Y-bX}) + \frac{1}{2}\mathcal{W}_2^2(P^U, P^\varepsilon) + \frac{1}{2}\,\mathbb{E}\,\|(b^* - b)X\|_2^2 \\ &= \frac{1}{2}\int \|u - v\|_2^2 \, dP^{(U,Y-bX)}(u, v) \\ &\qquad - \frac{1}{2}\inf_{Q \in \mathcal{C}(P^U, P^{Y-bX})} \int \|u - v\|_2^2 \, dQ(u, v) \geq 0. \end{aligned} \tag{S27}$$

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 P^U-a.e. unique, thus the equality can only be achieved in (S27) if $P^{(U,Y-bX)}$ is the optimal coupling. In such a case, by the Knott-Smith optimality criterion (Villani, 2021, 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 almost all $(u, v) \in \mathrm{Supp}(P^{(U,Y-bX)})$, 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. ■

A.6. Proof for Theorem 5

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 (5), we have the following basic inequality:

$$\mathcal{L}(\hat{b}) - \mathcal{L}(b^*) \leq \mathcal{L}(\hat{b}) - \mathcal{L}_{n,m}(\hat{b}) + \mathcal{L}_{n,m}(b^*) - \mathcal{L}(b^*). \tag{S28}$$

By the explicit formula for the 2-Wasserstein distance between two elliptical distributions (see Gelbrich, 1990, Theorem 2.1), we have

$$\begin{aligned}\langle\!\langle P^{(b^*-b)X}, P^U\rangle\!\rangle_{\mathcal{W}_2} &= \frac{1}{2}\Big\{\mathbb{E}\,\|(b^*-b)X\|^2 + \mathbb{E}\,\|U\|^2 - \mathcal{W}_2^2(P^{(b^*-b)X}, P^U)\Big\}\\ &= \frac{1}{2}\Big\{\mathbb{E}\,\|(b^*-b)X\|^2 + \mathbb{E}\,\|U\|^2 - \big\|\big((b^*-b)\Sigma(b^*-b)^T\big)^{1/2} - I_d\big\|_{\mathrm{F}}^2\Big\}\\ &= \mathrm{Tr}\big\{\big((b^*-b)\Sigma(b^*-b)^T\big)^{1/2}\big\} \qquad \text{(S29)}\\ &\geq \mathrm{Tr}^{1/2}\big\{(b^*-b)\Sigma(b^*-b)^T\big\} = \|b^*-b\|_\Sigma. \qquad \text{(S30)}\end{aligned}$$

Hence, writing $r := \langle\!\langle P^\varepsilon, P^U\rangle\!\rangle_{\mathcal{W}_2}$, we have by Lemma 6 that for any $b \in \mathbb{R}^{d\times p}$,

$$\begin{aligned}\mathcal{L}(b) - \mathcal{L}(b^*) &= \langle\!\langle P^{(b^*-b)X+\varepsilon}, P^U\rangle\!\rangle_{\mathcal{W}_2} - \langle\!\langle P^\varepsilon, P^U\rangle\!\rangle_{\mathcal{W}_2}\\ &\geq \sqrt{r^2 + \langle\!\langle P^{(b^*-b)X}, P^U\rangle\!\rangle_{\mathcal{W}_2}^2} - r \geq \sqrt{r^2 + \|b^*-b\|_\Sigma^2} - r. \qquad \text{(S31)}\end{aligned}$$

On the other hand, by Lemma 7, we have

$$\begin{aligned}|\mathcal{L}(b) - \mathcal{L}_{n,m}(b)| &= \Big|\langle\!\langle P^{Y-bX}, P^U\rangle\!\rangle_{\mathcal{W}_2} - \langle\!\langle P_n^{Y-bX}, P_m^U\rangle\!\rangle_{\mathcal{W}_2}\Big|\\ &\leq \alpha_m \mathcal{W}_2(P^{Y-bX}, P_n^{Y-bX}) + (\mathbb{E}\,\|Y-bX\|^2)^{1/2}\mathcal{W}_2(P^U, P_m^U), \qquad \text{(S32)}\end{aligned}$$

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 (S32) 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

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

Thus,

$$\mathcal{W}_2(P^{Y-bX}, P_n^{Y-bX}) \leq \|b^*-b\|_\Sigma \mathcal{W}_2(P^S, P_n^S) + \mathcal{W}_2(P^\varepsilon, P_n^\varepsilon) =: I_n(\|b^*-b\|_\Sigma). \qquad \text{(S33)}$$

For the second term on the right-hand side of (S32), define $s^2 := \mathbb{E}\,\|\varepsilon\|^2$, we have

$$\begin{aligned}(\mathbb{E}\,\|Y-bX\|^2)^{1/2} &= (\mathbb{E}\,\|(b^*-b)X+\varepsilon\|^2)^{1/2} \leq \{2\mathbb{E}\|(b^*-b)X\|^2 + 2\mathbb{E}\|\varepsilon\|^2\}^{1/2}\\ &= \big\{2\|b^*-b\|_\Sigma^2 + 2s^2\big\}^{1/2}. \qquad \text{(S34)}\end{aligned}$$

Combining (S32), (S33) and (S34), we obtain that

$$|\mathcal{L}(b) - \mathcal{L}_{n,m}(b)| \leq \alpha_m I_n(\|b^*-b\|_\Sigma) + \big\{2\|b^*-b\|_\Sigma^2 + 2s^2\big\}^{1/2}\mathcal{W}_2(P^U, P_m^U). \qquad \text{(S35)}$$

Since (S31) and (S35) holds for arbitrary $b \in \mathbb{R}^{d\times p}$, we have by (S28) 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 S22 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:

$$\begin{aligned}&\frac{(2\|b^* - \hat{b}\|_\Sigma - 1) \wedge \|b^* - \hat{b}\|_\Sigma^2}{(\|b^* - \hat{b}\|_\Sigma \vee 1)} \\ &\quad \le C(2 + 2s)\big(\alpha_m \mathcal{W}_2(P^S, P_n^S) + (\sqrt{2} + 2s\sqrt{2})\mathcal{W}_2(P^U, P_m^U) + 2\alpha_m \mathcal{W}_2(P^\varepsilon, P_n^\varepsilon)\big). \quad \text{(S36)}\end{aligned}$$

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 S29 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

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

Then combining this with (S36), and working on the event $E^{(n,m)}$, there exists some constant $\tilde{M} > 0$ depending only on d, ℓ, p such that

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

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. ■

A.7. Proof for Lemma 6

Proof By the Brenier's Theorem (Villani, 2009, 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 $t = \frac{\langle\!\langle \varepsilon, U\rangle\!\rangle^2_{\mathcal{W}_2}}{\langle\!\langle \varepsilon, U\rangle\!\rangle^2_{\mathcal{W}_2} + \langle\!\langle Z, U\rangle\!\rangle^2_{\mathcal{W}_2}}$, we have

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

as desired. ■

A.8. Proof for Lemma 7

Proof Let $\mathcal{X}_1, \mathcal{X}_2, \mathcal{Y}_1, \mathcal{Y}_2$ denote four copies of $\mathcal{X}$. By Lemma S12, 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

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

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 S12 *For $L \in \mathbb{N}$, write $V = \{1, \dots, L\}$. 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 $\left(\prod_{i=1}^L \mathcal{X}_i, \bigotimes_{i=1}^L \Omega_i\right)$ 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 $1, 2, \dots, L$. We now prove by induction that for any $\ell \in \{1, \dots, L\}$, there exists a measure $\rho_{1,\dots,\ell}$ on $\mathcal{X}_1 \times \cdots \times \mathcal{X}_\ell$ such that $\rho_{1,\dots,\ell}|_{\mathcal{X}_i} = \nu_i$ for all $i \in \{1, \dots, \ell\}$ and $\rho_{1,\dots,\ell}|_{\mathcal{X}_i \times \mathcal{X}_j} = \xi_{i,j}$ for all $(i,j) \in E \cap \{1, \dots, \ell\}^2$.

The base case of the induction is trivially true as we can take $\rho_1 = \nu_1$. Now assume that we have successfully constructed $\rho_{1,\dots,\ell-1}$ for some $\ell \in \{2, \dots, L\}$. Let ℓ' be the only neighbour of ℓ in $\{1, \dots, \ell - 1\}$ (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. Graf and Mauldin, 1989), 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 $i \leq \ell - 1$ 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. ■

A.9. Proof for Theorem 8

Define event $\Theta := \{\|\hat{b} - b^*\|_\Sigma < 1\}$, then in the regime of (6) we have $\mathbb{P}(\Theta) \geq 1 - 4(\log n)^{-1}$. 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 8 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

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

in Proposition S17. The proof of Proposition S17 relies on rewriting the Wasserstein distances using the Kantorovich dual formulation. Specifically, writing $\tilde{\Phi}_b := \{(f, g) \in L^1(P_n^{A(b)}) \times L^1(P_m^U) : v^T u \leq f(v) + g(u),\ \forall (v, u) \in \mathrm{Supp}(P_n^{A(b)}) \times \mathrm{Supp}(P_m^U)\}$, then for any fixed $b \in \mathcal{B}$, by Theorem S9 and Lemma S28, there exists a conjugate pair $(\tilde{\varphi}_{b;n,m}, \tilde{\varphi}^*_{b;n,m})$ such that

$$(\tilde{\varphi}^*_{b;n,m}, \tilde{\varphi}_{b;n,m}) = \operatorname*{arg\,min}_{(f,g)\in\tilde{\Phi}_b} \int f \, dP_n^{A(b)} + \int g \, dP_m^U, \tag{S37}$$

$$\frac{1}{2}\mathcal{W}_2^2(P_n^{A(b)}, P_m^U) = \int \|v\|^2/2 - \tilde{\varphi}^*_{b;n,m}(v) \, dP_n^{A(b)}(v) + \int \|u\|^2/2 - \tilde{\varphi}_{b;n,m}(u) \, dP_m^U(u),$$

and

$$\|u\|^2/2 \leq \tilde{\varphi}_{b;n,m}(u) \leq \|u\|^2/2 + L_{b;n,m}, \quad \|v\|^2/2 - L_{b;n,m} \leq \tilde{\varphi}^*_{b;n,m}(v) \leq \|v\|^2/2, \tag{S38}$$

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

Before stating Proposition S17, 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 S17.

Proposition S13 *Let $\tilde{\varphi}$ and $\tilde{\varphi}^*$ be defined as in (S37) 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 $(u, v) \in \mathbb{R}^d \times \mathbb{R}^d$, $v^T u \leq \varphi_{b;n,m}(u) + \varphi^*_{b;n,m}(v)$;

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

(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 $u \in \mathrm{Supp}(P_m^U)$

$$\varphi_{b;n,m}(u) \le \sup_{v\in\mathbb{R}^d}\big\{v^Tu - v^Tu + \tilde{\varphi}_{b;n,m}(u)\big\} = \tilde{\varphi}_{b;n,m}(u). \tag{S39}$$

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

$$\begin{aligned}\varphi^*_{b;n,m}(v) &\le \sup_{u\in\mathbb{R}^d}\big\{v^Tu - v^Tu + \zeta_{b;n,m}(v)\big\}\\ &= \zeta_{b;n,m}(v) \le \tilde{\varphi}^*_{b;n,m}(v) \vee \big(\frac{\|v\|^2}{2} - \|c\|_\infty\big) \le \tilde{\varphi}^*_{b;n,m}(v).\end{aligned} \tag{S40}$$

Assume any of (S39) or (S40) 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 (S38) and preceding constructions we have for $u \in \mathbb{R}^d$

$$\|u\|^2/2 - \varphi_{b;n,m}(u) \ge \inf_{v\in\mathbb{R}^d}\big\{L_2(u,v) - L_{b;n,m}\big\} = -L_{b;n,m}. \tag{S41}$$

Moreover, we have

$$\begin{aligned}\|u\|^2/2 - \varphi_{b;n,m}(u) &\le -\big(\|u\|^2/2 - \zeta_{b;n,m}(u)\big)\\ &= -\inf_{u'\in\mathrm{Supp}(P_n^{A(b)})}\big(L(u,u') - (\|u'\|^2/2 - \tilde{\varphi}_{b;n,m}(u'))\big) \wedge L_{b;n,m} \le 0,\end{aligned} \tag{S42}$$

where the last step follows by the fact that $\|u'\|^2/2 - \tilde{\varphi}_{b;n,m}(u') \le 0$, for all $u' \in \mathrm{Supp}(P_n^{A(b)})$. Here, we proved that $-L_{b;n,m} \le \frac{\|u\|^2}{2} - \varphi_{b;n,m}(u) \le 0$ and the result holds. For any $v \in \mathbb{R}^d$, by (S42) we have

$$\|v\|^2/2 - \varphi^*_{b;n,m}(v) = \inf_{u\in\mathbb{R}^d}\Big(L_2(u,v) - (\|u\|^2/2 - \varphi_{b;n,m}(u))\Big) \ge 0. \tag{S43}$$

Moreover, by (S41) it follows that

$$\|v\|^2/2 - \varphi^*_{b;n,m}(v) \le -(\|v\|^2/2 - \varphi_{b;n,m}(v)) \le L_{b;n,m}. \tag{S44}$$

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

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

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

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 (Villani, 2021, 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 (2024, Lemma 10), but for the sake of completeness, we provide it here.

Lemma S14 *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 (S17)). 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*

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

Proof We focus on φ_0 and the same argument can be used for φ_0^*. Firstly, by Villani (2021, 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

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

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

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

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

For all $j \geq 0$, 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 $\{P_{j,k}\}_{k=1,\ldots,N}$, 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 $\mathcal{C}_{m,u}(I) := \{f \in \mathcal{C}(I) : \exists m, u > 0, \text{ s.t. } |f(x) - f(y)| \leq m\|x - y\|, \ |f(x)| \leq u, \ \forall x, y \in I\}$ to be the class of m-Lipschitz convex functions on I bounded in value by u. Given a sequence M and U, define

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

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 $\|A(b)\| \leq \|S\| + \|\varepsilon\|$ 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 S15 *Define $E_{2,n} := \sup_{b\in\mathcal{B}} E_{b;2,n}$. Then for any fixed $n, m \geq 1$ 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 $E_{1,m} \leq 2 + \sqrt{\frac{\log m}{m}}$ with probability at least $1 - 2(\log m)^{-1}$ and $E_{2,n} \leq 2 + \sqrt{\frac{\log n}{n}}$ with probability at least $1 - 2(\log n)^{-1}$.*

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

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

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 S16 *Let* $J_n = \left\lfloor \frac{1}{2}\log_3\left(\frac{\log n}{16\gamma_2 d}\right)\right\rfloor$, $I_m = \left\lfloor \frac{1}{2}\log_3\left(\frac{\log m}{8d}\right)\right\rfloor$ *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*

$$M_j = \begin{cases} C_0'\ell_j, & 0 \le j \le J_n \\ C_1' l_{n,m}\ell_j, & j > J_n \end{cases}, \quad U_j = \begin{cases} \tilde{C}_0\ell_j^3, & 0 \le j \le J_n \\ \tilde{C}_1 l_{n,m}\ell_j^3, & j > J_n \end{cases}, \tag{S45}$$

and R *and* T *are chosen as*

$$R_i = \begin{cases} C_2'\ell_i, & 0 \le i \le I_m \\ C_3' l_{n,m}\ell_i, & i > I_m \end{cases}, \quad T_i = \begin{cases} \tilde{C}_2\ell_i^3, & 0 \le i \le I_m \\ \tilde{C}_3 l_{n,m}\ell_i^3, & i > I_m \end{cases}. \tag{S46}$$

Proof Note Lemma S14 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 S13(iv) and Lemma S15, we can apply Manole and Niles-Weed (2024, 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

$$\|y\| \le C_0(2dE_{1,m})^{1/2}\Big\{(\|v\|+1) \vee \sup_{w:\|v-w\|\le 2}\Big[\log\Big(\frac{1}{P_n^{A(b)}(\mathcal{B}^d_{w,3})}\Big)\Big]^{1/2}\Big\}. \tag{S47}$$

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 (12), for any $0 \le j \le J_n$, $v \in P_j$ and w such that $\|w - v\| \le 2$, we have

$$\begin{aligned} P^\varepsilon(\mathcal{B}^d_{w,2}) \ge \int_{\mathcal{B}^d_{w,2}\setminus\mathcal{B}^d_0} \gamma_1\exp(-\gamma_2\|e\|^2)\,de &\ge \frac{\pi^{d/2}(2^d-1)}{\Gamma(d/2+1)}\gamma_1\exp\big(-2\gamma_2\|z\|^2 - 50\gamma_2\big) \\ &\ge 2K_1\exp(-2\gamma_2\ell_j^2), \end{aligned} \tag{S48}$$

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

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

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.

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 (2024, 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.

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

$$\Upsilon_1 := \Big\{ \sup_{B\in\mathcal{B}^d} |P_n^\varepsilon \otimes P_n^S(B \times \mathcal{B}_{0,}^p) - P^\varepsilon \otimes P^S(B \times \mathcal{B}_0^p)| < \tilde{u} \Big\}.$$

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 $P_n^\varepsilon \otimes P_n^S(\mathcal{B}^d_{w,2} \times \mathcal{B}_0^d) \ge K_1' \exp(-2\gamma_2\ell_j^2)$. Thus consider the event

$$\Upsilon_2 := \bigcap_{j=0}^{J_n} \Big\{ \inf_{v\in P_j} \inf_{w:\|v-w\|\le 2} P_n^\varepsilon \otimes P_n^S(\mathcal{B}^d_{w,2} \times \mathcal{B}_0^p) \ge K_1' \exp\big(-2\gamma_2\ell_j^2\big) \Big\},$$

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. Devroye et al., 2013, Corollary 13.2), by the VC-inequality (see Vapnik and Chervonenkis, 2015, Theorem 2) we have

$$\mathbb{P}(\Upsilon_1^c) \lesssim n^{d+2} \exp(-n\tilde{u}^2/32) \le n^{2-4d} \le n^{-2}, \tag{S49}$$

whence $\mathbb{P}(\Upsilon_2) \ge 1 - n^{-2}$. Thus working on $\Upsilon_2 \cap \Gamma_1$, by $\|A(b)_i\| \le \|S_i\| + \|\varepsilon_i\|$ for all $b \in \mathcal{B}$ and $i \in [n]$, we have $P_n^{A(b)}(\mathcal{B}^d_{w,3}) \ge K_1' \exp(-2\gamma_2\ell_j^2)$, and combining this with (S47), we conclude that for any $1 \le k \le N$ and $0 \le j \le J_n$, there exits some sufficiently large constant $C_0' > 0$ depends on $d, \gamma_1, \gamma_2, \sigma_1, \alpha$ such that

$$\begin{aligned}\sup_{y\in\partial\varphi^*_{b;n,m}(P_{j,k})} \|y\| &\le C_0(2dE_{1,m})^{1/2}\Big(\ell_j + 1 + \sqrt{2}\ell_j\gamma_2^{1/2} + \sqrt{\log(1/K_1')}\Big) \\ &\le C_0' E_{1,m}^{1/2}\ell_j \le C_0'\Big(2 + \sqrt{\frac{\log m}{m}}\Big)^{1/2}\ell_j \lesssim C_0'\ell_j := M_j. \end{aligned} \tag{S50}$$

When $j > J_n$, by Lemma S13(iii) and Manole and Niles-Weed (2024, Proposition 16), we only need to bound $L_{b;n,m}$. Note $L_{b;n,m} \le L_{n,m} := 2\max_{i\in[n]} \|\Sigma^{-1/2}X_i\|^2 + 2\max_{i\in[n]}\|\varepsilon_i\|^2 + 2\max_{j\in[m]}\|U_j\|^2$. 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 S24 and union bound, it follows that

$$\begin{aligned}\mathbb{P}\big(\Upsilon_3^c\big) &\le \mathbb{P}\big(\max_{i\in[n]}\|\Sigma^{-1/2}X_i\|^2 \ge \sigma_1^2(4\log n)^{2/\alpha}\big) + \mathbb{P}\big(\max_{i\in[n]}\|\varepsilon_i\|^2 \ge \sigma_2^2(4\log n)^{2/\beta}\big) \\ &\quad + \mathbb{P}\big(\max_{j\in[m]}\|U_j\|^2 \ge 8d\log m\big) \le 4n^{-1} + 2m^{-1}. \end{aligned} \tag{S51}$$

Therefore on the event Υ_3, by Manole and Niles-Weed (2024, 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 $1 \le k \le N$,

$$\sup_{y\in\partial\varphi^*_{b;n,m}(P_{j,k})} \|y\| \le C_1(\ell_j + r_{n,m}) \le C_1' l_{n,m}\ell_j =: M_j, \quad \text{for all } j > J_n. \tag{S52}$$

Putting (S50) and (S52) 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

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

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 (2024, Theorem 11) and Lemma S14 again we obtain that for $0 \leq i \leq I_m$, there exists constants $C_2 > 0$ depends on d, α, β and $C_2' > 0$ depends on d, σ, α, β such that

$$\begin{aligned} \sup_{z\in\partial\varphi_{b;n,m}(u)} \|z\| &\leq C_2 \sigma (2E_{2,n})^{1/(\alpha\wedge\beta)} (2\ell_i + 1 + \sqrt{\log(1/K_2)}) \\ &\leq C_2' E_{2,n}^{1/(\alpha\wedge\beta)} \ell_i \leq C_2' \Big(2 + \sqrt{\frac{\log n}{n}}\Big)^{1/(\alpha\wedge\beta)} \ell_i \lesssim C_2' \ell_i := R_i. \end{aligned} \tag{S53}$$

When $i > I_m$, since we still have $|\|u\|^2/2 - \varphi_{b;n,m}| \leq L_{b;n,m} \leq L_{n,m}$ by Lemma S13(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 $1 \leq k \leq N$.

$$\sup_{z\in\partial\varphi_{b;n,m}(P_{i,k})} \|z\| \leq C_3(\ell_i + r_{n,m}) \leq C_3' l_{n,m} \ell_i := R_i \quad \text{for } i > I_m. \tag{S54}$$

Thus combine (S53) and (S54) 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

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

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 $\mathbb{P}(\Upsilon) \geq 1 - n^{-2} - 4n^{-1} - 2m^{-1} - m^{-2} - 2(\log m)^{-1} - 2(\log n)^{-1} \geq 1 - 12(\log n)^{-1}$ when $m \geq n$, which completes the proof. ■

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

Proposition S17 *There exists a constant $C > 0$ depending on $d, \alpha, \beta, \sigma_1, \sigma_2, \gamma_1, \gamma_2$ such that for any fixed $n, m \geq 1$, the following inequality holds*

$$\sup_{b\in\mathcal{B}}\left|\mathcal{W}_2^2(P^{A(b)}, P^U) - \mathcal{W}_2^2(P_n^{A(b)}, P_m^U)\right| \leq C(\log m)^{\frac{8}{2\wedge\alpha\wedge\beta}}\left(\sqrt{\frac{p}{n}} + \frac{1}{n^{2/d}}\right) \tag{S55}$$

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

Proof Note part (i) and part (iii) of Lemma S13 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

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

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

$$\begin{aligned}
&\frac{1}{2}\mathcal{W}_2^2(P_n^{A(b)}, P_m^U) - \frac{1}{2}\mathcal{W}_2^2(P^{A(b)}, P^U)\\
&\qquad\leq \underbrace{\int\varphi^*_{b;n,m}(v)d\big(P^{A(b)} - P_n^{A(b)}\big)(v) + \int\varphi_{b;n,m}(u)d\big(P^U - P_m^U\big)(u)}_{=:E_{b;n,m}}\\
&\qquad\quad + \underbrace{\int\frac{\|v\|^2}{2}d\big(P_n^{A(b)} - P^{A(b)}\big)(v) + \int\frac{\|u\|^2}{2}d\big(P_m^U - P^U\big)(u)}_{=:F_{b;n,m}}.
\end{aligned} \tag{S56}$$

On the other hand, define $\Psi_b := \{(f,g) \in L^1(P^{A(b)})\times L^1(P^U) : v^Tu \leq f(v)+g(u),\ \forall(v,u) \in \mathrm{Supp}(P^{A(b)}) \times \mathrm{Supp}(P^U)\}$, then Theorem S9 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

$$\begin{aligned}\frac{1}{2}\mathcal{W}_2^2(P_n^{A(b)}, P_m^U) - \frac{1}{2}\mathcal{W}_2^2(P^{A(b)}, P^U) \\ \geq \underbrace{\int \psi_b^*(v) d\big(P^{A(b)} - P_n^{A(b)}\big)(v) + \int \psi_b(u) d\big(P^U - P_m^U\big)(u)}_{=:G_{b;n,m}} \\ + \int \frac{\|v\|^2}{2} d\big(P_n^{A(b)} - P^{A(b)}\big)(v) + \int \frac{\|u\|^2}{2} d\big(P_m^U - P^U\big)(u). \end{aligned} \tag{S57}$$

Write the first two terms and the last two terms of (S56) as $E_{n,m}$ and $F_{n,m}$ respectively, and write the first two terms of (S57) as $G_{n,m}$. Then combining (S56) and (S57), for ϑ_k defined in (S68), we have

$$\begin{aligned}\sup_{b\in\mathcal{B}}\Big|\frac{1}{2}\mathcal{W}_2^2(P_n^{A(b)}, P_m^U) - \frac{1}{2}\mathcal{W}_2^2(P^{A(b)}, P^U)\Big| &\leq \sup_{b\in\mathcal{B}}|E_{b;n,m}| + 2\sup_{b\in\mathcal{B}}|F_{b;n,m}| + \sup_{b\in\mathcal{B}}|G_{b;n,m}| \\ &\lesssim (\log m)^{\frac{6}{2\wedge\alpha\wedge\beta}}\big(\vartheta_n + \sqrt{\frac{p}{n}} + \sqrt{\frac{\log n}{n}} + \vartheta_m + \sqrt{\frac{\log m}{m}}\big), \\ &\leq (\log m)^{\frac{8}{2\wedge\alpha\wedge\beta}}\left(\sqrt{\frac{p}{n}} + \frac{1}{n^{2/d}}\right). \end{aligned} \tag{S58}$$

with probability at least $1 - 29(\log n)^{-1}$, where we have used Lemmas S18, S21 and S20 to bound each of the three terms in the penultimate inequality. ■

Lemma S18 *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*

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

where ϑ_n is defined as (S68).

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 S25(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 $\Omega_1 := \{\max_{1\leq i\leq n}\|(\varepsilon_i, S_i)\| \leq \kappa\}$, and by Proposition S24(i), we have

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

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

$$T_b(s,e) := (b^* - b)\Sigma^{1/2}s + e. \tag{S59}$$

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

$$\begin{aligned}
\int_{\mathbb{R}^d \setminus E_b} \varphi^*_{b;n,m}(v)\, d\big(P^{A(b)} - P_n^{A(b)}\big)(v) &= \int_{\mathbb{R}^{d+p}\setminus \mathcal{B}_{0,\kappa}^{d+p}} \varphi^*_{b;n,m} \circ T_b(e,s)\, dP^{\varepsilon} \otimes P^S(e,s) \\
&\overset{(a)}{\le} \int_{\mathbb{R}^{d+p}\setminus \mathcal{B}_{0,\kappa}^{d+p}} \left(\frac{\|T_b(e,s)\|^2}{2} + r_{n,m} \right) d(P^{\varepsilon} \otimes P^S)(e,s) \\
&\le \int_{\mathbb{R}^{d+p}\setminus \mathcal{B}_{0,\kappa}^{d+p}} \left(\frac{\|(e,s)\|^2}{2} + r_{n,m} \right) d(P^{\varepsilon} \otimes P^S)(e,s) \\
&\overset{(b)}{\le} C_4 e^{-\frac{1}{4}(\frac{\kappa}{\rho})^{\alpha\wedge\beta}} + \frac{2r_{n,m}}{n^2} \lesssim \frac{C_4}{n},
\end{aligned} \tag{S60}$$

where we use part (iii) of Proposition S13 to obtain (a) and Lemma S26 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

$$\mathcal{F} := \Big\{ (s,e) \mapsto (\varphi \circ T_b)(s,e)\, \mathbb{1}_{\mathcal{B}_0^{d+p}}(s,e) : b \in \mathcal{B},\ \varphi \in \mathrm{Lip}_{1,1}(\mathcal{B}_0^d) \Big\}. \tag{S61}$$

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 S15. Let Υ be the event with probability $1 - 12(\log n)^{-1}$ on which Proposition S16 holds. Then, from Proposition S16, we have $\varphi^*_{b;n,m}|_{\mathcal{B}_{0,\kappa}^d}$ is Lipschitz continuous with parameter M_{j_n} and upper bound U_{j_n}, for M_j and U_j are defined in (S45). Specifically, since $j_n > J_n$, from (S45), there exists $C_5 > 0$, depending only on $d, \alpha, \beta, \sigma_1, \sigma_2, \gamma_2$, such that $M_{j_n} \vee U_{j_n} \le C_5(\log m)^{\frac{5}{2\wedge\alpha\wedge\beta}}$. 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

$$\begin{aligned}
&\int_{E_b} \frac{\varphi^*_{b;n,m}(v)}{C_5\kappa(\log m)^{\frac{5}{2\wedge\alpha\wedge\beta}}}\, d\big(P^{A(b)} - P_n^{A(b)}\big)(v) \\
&\qquad = \int_{\mathcal{B}_{0,\kappa}^{d+p}} \frac{\varphi^*_{b;n,m}(T_b(s,e))}{C_5\kappa(\log m)^{\frac{5}{2\wedge\alpha\wedge\beta}}}\, d(P^{\varepsilon}\otimes P^S - P_n^{\varepsilon}\otimes P_n^S)(e,s) \\
&\qquad = \int_{\mathcal{B}_{0,1}^{d+p}} \frac{\varphi^*_{b;n,m}(\kappa T_b(s,e))}{C_5(\log m)^{\frac{5}{2\wedge\alpha\wedge\beta}}}\, d(P^{\varepsilon}\otimes P^S - P_n^{\varepsilon}\otimes P_n^S)(e,s) \\
&\qquad \le \sup_{f\in\mathcal{F}} \Big\{ \int f(s,e)\, d(P^{\varepsilon}\otimes P^S - P_n^{\varepsilon}\otimes P_n^S)(e,s) \Big\}.
\end{aligned} \tag{S62}$$

By Lemma S19 and Wainwright (2019, 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

$$\sup_{f\in\mathcal{F}}\Big|\int f\, d(P^{\varepsilon}\otimes P^{S} - P_n^{\varepsilon}\otimes P_n^{S})\Big| \le 2C'\Big(\vartheta_n + \sqrt{\frac{p}{n}}\Big) + \sqrt{\frac{2\log n}{n}}. \tag{S63}$$

Combining (S60), (S62) and (S63), we have on event $\Upsilon\cap\Omega_1\cap\Omega_2$ that

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

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 $\Omega_3 := \{\max_{1\le i\le m}\|U_i\| \le \gamma\}$. Then by a union bound we have $\mathbb{P}(\Omega_3^c) \le m\,\mathbb{P}\big(\|U_1\| \ge \gamma\big) \le 2m\exp(-\frac{1}{2}\frac{\gamma^2}{2d}) \le \frac{2}{m}$. Working on Ω_3 we deduce that for some absolute constant $C_6 > 0$,

$$\begin{aligned}\int_{\mathbb{R}^d\setminus\mathcal{B}^d_{0,\gamma}} \varphi_{b;n,m}(u)\, d(P^U - P_m^U)(u) &= \int_{\mathbb{R}^d\setminus\mathcal{B}^d_{0,\gamma}} \varphi_{b;n,m}(u)\, dP^U(u)\\ &\overset{(c)}{\le} \int_{\mathbb{R}^d\setminus\mathcal{B}^d_{0,\gamma}} \frac{\|u\|^2}{2}\, dP^U(u) \overset{(d)}{\le} \frac{C_6}{m^2}.\end{aligned} \tag{S64}$$

In the above, we use part (iii) in the Proposition S13 to obtain (c) and Lemma S26 in inequality (d).

Define

$$\mathcal{H} = \{g\,\mathbb{1}_{\mathcal{B}_0^d} : g \in \mathrm{Lip}_{1,1}(\mathcal{B}_0^d)\}. \tag{S65}$$

Let $i_m := (I_m + 1) + \lceil\frac{1}{2}\log_3(8d\log m)\rceil$. Observe that $3^{i_m}\sqrt{d} \ge \gamma$ thus we have $\mathcal{B}^d_{0,\gamma} \subset \bigcup_{i=0}^{i_m}\bigcup_{k=1}^{N} P_{i,k}$. Since $\varphi_{b;n,m}\in\mathcal{C}_{R,T}$ on Υ according to Proposition S16, we have that $\varphi_{b;n,m}|_{\mathcal{B}^d_{0,\gamma}}$ is bounded and Lipschitz continuous with upper bound T_{i_m} and Lipshictz constant R_{i_m} as defined in (S46). Moreover, by the explicit display of (S46), there exists a constant C_7 depends on $d, \sigma_1, \sigma_2, \alpha, \beta, \gamma_2$ such that $R_{i_m}\vee T_{i_m} \le C_7(\log m)^{\frac{5}{2\wedge\alpha\wedge\beta}}$. Therefore, on Υ, we have

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

and consequently,

$$\frac{1}{C_7\gamma(\log m)^{\frac{5}{2\wedge\alpha\wedge\beta}}}\int_{\mathcal{B}^d_{0,\gamma}} \varphi_{b;n,m}(u)\, d(P^U - P_m^U)(u) \le \sup_{h\in\mathcal{H}}\Big\{\int h(u)\, d(P^U - P_m^U)(u)\Big\}. \tag{S66}$$

Then applying Lemma S19 and Wainwright (2019, 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

$$\sup_{h\in\mathcal{H}}\left|\int h\, d(P^U - P_m^U)\right| \le 2\vartheta_m + \sqrt{\frac{2\log m}{m}}. \tag{S67}$$

Consequently, combining (S64), (S66) and (S67), and working on the event $\Upsilon \cap \Omega_3 \cap \Omega_4$, we obtain

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

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 $\mathbb{P}(\Upsilon \cap \Omega_1 \cap \Omega_2 \cap \Omega_3 \cap \Omega_4) \ge 1 - 18(\log n)^{-1}$. ■

Lemma S19 *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*

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

where

$$\vartheta_k := \begin{cases} k^{-2/d}, & \text{if } d \ge 5,\\ k^{-1/2}\log k, & \text{if } d = 4,\\ k^{-1/2}, & \text{if } d \le 3.\end{cases} \tag{S68}$$

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 (1976, 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 (2019, 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

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

which implies

$$\log N(2\delta, \mathcal{F}, \|\cdot\|_{L^2(P)}) \le \log(N_0N_1) \le C_8\left(\frac{4}{\delta}\right)^{d/2} + dp\log\left(1+\frac{2}{\delta}\right) \le C_8\left(\frac{4}{\delta}\right)^{d/2} + \frac{2dp}{\delta}. \tag{S69}$$

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. Wainwright, 2019, 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 S20 *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 S27 and the anti-concentration inequality of ε given in (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

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

Then by recalling that $P^U \sim (\sqrt{2d}, 2)$-sub-Weibull, we apply Manole and Niles-Weed (2024, 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 $v \in \mathbb{R}^d$.

3. In the original Theorem 11 of Manole and Niles-Weed (2024), 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.

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 S18, we have $\mathbb{P}(\Omega_1) \ge 1 - 2/n$. Then on Ω_1 we can obtain that

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

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 S26 in the final inequality. It therefore remains to control

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

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

$$\mathbb{E}(G) = \int_0^{+\infty}\mathbb{P}(G\ge t)\,dt \le n^{-1/2} + C\int_{n^{-1/2}}^{+\infty}\frac{(\log m)^{\frac{2}{2\wedge\alpha\wedge\beta}}}{nt^2}dt \lesssim \frac{(\log m)^{\frac{2}{2\wedge\alpha\wedge\beta}}}{\sqrt{n}}. \tag{S70}$$

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. Wainwright, 2019, (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 that $v_i = v_i'$ for any $i \ne 1$. We have

$$\begin{aligned}
\sup_{b\in\mathcal{B}}\Big|\frac1n\sum_{i=1}^n\bar\psi_b^*(v_i)\Big| - \sup_{b\in\mathcal{B}}\Big|\frac1n\sum_{i=1}^n\bar\psi_b^*(v_i')\Big| &\le \sup_{b\in\mathcal{B}}\left\{\Big|\frac1n\sum_{i=1}^n\bar\psi_b^*(v_i)\Big| - \Big|\frac1n\sum_{i=1}^n\bar\psi_b^*(v_i')\Big|\right\} \\
&\le \frac1n\sup_{b\in\mathcal{B}}\Big|\bar\psi_b^*(v_1) - \bar\psi_b^*(v_1')\Big| \le \frac{2C(\log m)^{\frac{1}{2\wedge\alpha\wedge\beta}}}{n},
\end{aligned}$$

estbalishing, the bounded difference property for G. Thus by McDiarmid's inequality (see e.g. Wainwright, 2019, Corollary 2.21), we obtain that the event

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

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

Thus, working on the event $\Omega_1 \cap \Lambda_1$, we deduce from (S70) that

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

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 (2024, Theorem 11) implies that $\|\nabla\psi_b(u)\| \le \tilde{C}(1 + \|u\|)^{\frac{2}{\alpha\wedge\beta}}$ for some $\tilde{C} > 0$ depending on $d, \sigma_1, \sigma_2, \alpha, \beta$, and it follows that $|\psi_b(u)| \le \tilde{C}(1 + \|u\|)^{\frac{2}{\alpha\wedge\beta}+1}$.

Define $\Omega_2 := \{\max_{1\le i\le m}\|U_i\| \le \gamma\}$. From the proof of Lemma S18 again, we have $\mathbb{P}(\Omega_2) \ge 1 - 2/n$. Working on Ω_2, we have

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

for some constant $\tilde{C} > 0$ depending on $d, \alpha, \beta, \sigma_1, \sigma_2$. Now, defining $\tilde{G} := \sup_{b\in\mathcal{B}}\big|\int_{\mathcal{B}_{0,\gamma}} \psi_b d(P^U - P_m^U)\big|$, 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

$$\sup_{b\in\mathcal{B}}\Big|\int \psi_b(u)d(P^U - P_m^U)(u)\Big| \le \tilde{G} + \frac{\tilde{C}'}{m} \le \mathbb{E}\tilde{G} + \frac{\bar{C}(\log m)^{\frac{2}{2\wedge\alpha\wedge\beta}}}{\sqrt{m}} + \frac{\tilde{C}}{m} \le \frac{\bar{C}(\log m)^{\frac{2}{2\wedge\alpha\wedge\beta}}}{\sqrt{m}}, \tag{S71}$$

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

Lemma S21 *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*

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

Proof Observe that the only property of ψ_b^* that we used in the first part of the proof of Lemma S20 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 S20 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 (2000, Lemma 1) we deduce that

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

which implies the second claim. ■

Proof of Theorem 8 Recalling that in the regime of (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

$$\begin{aligned}\mathcal{L}(\hat{b}) - \mathcal{L}(b^*) &\le 2\sup_{b\in\mathcal{B}}|\mathcal{L}(b) - \mathcal{L}_{n,m}(b)| \\ &\le \Big|\frac{1}{m}\sum_{i=1}^{m}\|U_i\|^2 - \mathbb{E}\,\|U\|^2\Big| + \sup_{b\in\mathcal{B}}\Big|\frac{1}{n}\sum_{i=1}^{n}\|T_b(S_i,\varepsilon_i)\|^2 - \mathbb{E}\,\|T_b(S_i,\varepsilon_i)\|^2\Big| \\ &\qquad + \sup_{b\in\mathcal{B}}\Big|\mathcal{W}_2^2(P^{A(b)},P^U) - \mathcal{W}_2^2(P_n^{A(b)},P_m^U)\Big| \\ &\le M(\log m)^{\frac{8}{2\wedge\alpha\wedge\beta}}\bigg(\sqrt{\frac{p}{n}} + \frac{1}{n^{2/d}}\bigg), \end{aligned} \tag{S72}$$

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 S21 to control the first two terms and Proposition S17 for the last term.

On the other hand, by the lower bound developed in (S31) and Lemma S22 we have for $r := \langle\!\langle P^{\varepsilon}, P^U\rangle\!\rangle_{\mathcal{W}_2}$ that

$$\mathcal{L}(\hat{b}) - \mathcal{L}(b^*) \ge \sqrt{r^2 + \|b^*-\hat{b}\|_{\Sigma}^2} - r \ge \frac{1}{2}(1+r^2)^{-1/2}\|b^*-\hat{b}\|_{\Sigma}^2. \tag{S73}$$

Combining (S72) with (S73), we obtain that

$$\|b^*-\hat{b}\|_{\Sigma} \le M(\log m)^{\frac{4}{2\wedge\alpha\wedge\beta}}\bigg\{\bigg(\frac{p}{n}\bigg)^{1/4} + \frac{1}{n^{1/d}}\bigg\}, \tag{S74}$$

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

Appendix B. Ancillary results

Lemma S22 *For any $a \ge 0$, we have inequality*

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

and

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

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 $t/\sqrt{a+t^2} \le t/\sqrt{a}$ and $t/\sqrt{a+t^2} \le 1$ in the case of $0 \le t \le 1$ and $t \ge 1$ respectively. The second inequality follows by noting that $t/\sqrt{a+t^2} \ge t/\sqrt{a+1}$ when $0 \le t \le 1$ and $t/\sqrt{a+t^2} \ge 1/\sqrt{a+1}$ when $t \ge 1$. ■

Lemma S23 *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 $\langle\!\langle Z+\varepsilon, U\rangle\!\rangle^2_{\mathcal{W}_2} = \langle\!\langle Z, U\rangle\!\rangle^2_{\mathcal{W}_2} + \langle\!\langle \varepsilon, U\rangle\!\rangle^2_{\mathcal{W}_2}$.*

Proof Consider independent random vectors $Z \sim \mathcal{N}(0,\Sigma)$ and $\varepsilon \sim \mathcal{N}(0,\Gamma)$. By the same argument as in (S30), we have

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

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

Proposition S24 *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 $\mathbb{E}\exp\left((\lambda\|X\|)^\beta\right) \leq \exp\left((\lambda K'_\sigma)^\beta\right)$ for all $|\lambda| \leq 1/K'_\sigma$.

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

The proof follows by Vladimirova et al. (2020, Theorem 2.1).

Proposition S25 *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 S24. Then it follows that

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

This proves that $AX+BY$ satisfies part (ii) in the Proposition S24 thus the conclusion follows by the equivalence of part (ii) and (iv).

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

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

where the sub-Weibull assumption on X and Y have been exploited. The conclusion follows by employing Proposition S24. ■

Lemma S26 *If X is a (σ,β)-sub-Weibull random vector as defined in (11), then for any $s>0$, there exists $C>0$, depending on s,σ,β, such that $\mathbb{E}\big(\|X\|^s\,\mathbb{1}\{\|X\|\geq t\}\big)\leq Ce^{-\frac{1}{4}(t/\sigma)^\beta}$.*

Proof We have

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

where we used the definition of X being (σ,β)-sub-Weibull in final step. The desired bound follows since by Proposition S24, we have $\mathbb{E}\|X\|^{2s}\leq C$ for some constant C that depends on s,σ,β. ■

Lemma S27 *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 $\gamma_1,\gamma_2>0$:*

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

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

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

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

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

where $\gamma_1'=\gamma_1\exp(-2\gamma_2M_2^2)$ 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 S28 *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 $\tilde{\Phi}:=\{(\varphi,\psi)\in L^1(\mathcal{X})\times L^1(\mathcal{Y}):\varphi(x)+\psi(y)\geq x^Ty,\ \forall(x,y)\in\mathcal{X}\times\mathcal{Y}\}$.

Proof Note that by the argument same as (S17) we have

$$\begin{aligned}&\inf\big\{J_{\mu,\nu}(\varphi,\psi):(\varphi,\psi)\in\tilde{\Phi}\big\}\\&\quad=\int_{\mathcal{X}}\frac{\|x\|^2}{2}d\mu(x)+\int_{\mathcal{Y}}\frac{\|y\|^2}{2}d\nu(y)-\sup\big\{J_{\mu,\nu}(\varphi,\psi):(\varphi,\psi)\in\Phi_2\big\},\end{aligned}\tag{S75}$$

where $\Phi_2:=\{(\varphi,\psi)\in L^1(\mathcal{X})\times L^1(\mathcal{Y}):\varphi(x)+\psi(y)\le L_2(x,y),\ \forall(x,y)\in\mathbb{R}^d\times\mathbb{R}^d\}$. Note by Villani (2021, Remark 1.13), we may restrict the supremum in the right-hand side of (S75) over some bounded functions:

$$\begin{aligned}&\sup\big\{J_{\mu,\nu}(\varphi,\psi):(\varphi,\psi)\in\Phi_2\big\}\\&\quad=\sup\big\{J_{\mu,\nu}(\varphi,\psi):(\varphi,\psi)\in\Phi_2,0\le\varphi\le\|L_2\|_\infty,\ -\|L_2\|_\infty\le\psi\le 0\big\}.\end{aligned}\tag{S76}$$

By Villani (2009, 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 (S76). Define $\tilde{\varphi}:=\|\cdot\|^2/2-\varphi_0$, $\tilde{\psi}:=\|\cdot\|^2/2-\varphi_0^c$. Then by (S75) we have

$$\inf\big\{J_{\mu,\nu}(\varphi,\psi):(\varphi,\psi)\in\tilde{\Phi}\big\}=\int_{\mathcal{X}}\tilde{\varphi}(x)d\mu(x)+\int_{\mathcal{Y}}\tilde{\psi}(y)d\nu(y).\tag{S77}$$

Moreover, note

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

which implies that $(\tilde{\varphi},\tilde{\psi})\in\tilde{\Phi}$. Combine this with (S77), we proved that $(\tilde{\varphi},\tilde{\psi})$ is an optimal solution to the left-hand side of (S75). Finally, by the boundedness of φ_0 and φ_0^c, we have

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

as desired. ■

Theorem S29 *(Fournier and Guillin, 2015, 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 $n\ge 1$,*

$$\mathbb{E}\left[\mathcal{W}_2^2(P^X,P_n^X)\right]\le CM_\ell^{2/\ell}\tau_n(d,\ell),\tag{S78}$$

where

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

Appendix C. Spatial reference distribution

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 3. 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

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

where $\phi(t) = \int_t^1 q(\tau)d\tau$. Thus, applying the same argument as Lemma 2 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 3. 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 3 demonstrates the validity of this extension.

Appendix D. Spatial quantile

The concept of the spatial (or geometric) quantile was initially introduced by Chaudhuri (1996). Uniquely characterizing the underlying probability distribution as a special case of M-quantile, as demonstrated in (Koltchinskii, 1997, Theorem 2.5), this quantile permits a seamless extension to the regression framework (Chakraborty, 2003) and functional quantile regression (Chakraborty and Chaudhuri, 2014; Chowdhury and Chaudhuri, 2019). A more recent development involves an extension to the hypersphere, as explored by Konen and Paindaveine (2023).

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 S30 *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*

$$\mathrm{Q}_{\tau u} = \operatorname*{arg\,min}_{y \in \mathbb{R}^d} \mathbb{E}\, \Phi_{\tau u}(Y - y). \tag{S79}$$

Note the solution of (S79) 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.