Errors-in-variables Fréchet Regression with Low-rank Covariate Approximation

Anonymous Author(s)
Affiliation
Address
email

Abstract

Fréchet regression has emerged as a promising approach for regression analysis involving non-Euclidean response variables. However, its practical applicability has been hindered by its reliance on ideal scenarios with abundant and noiseless covariate data. In this paper, we present a novel estimation method that tackles these limitations by leveraging the low-rank structure inherent in the covariate matrix. Our proposed framework combines the concepts of global Fréchet regression and principal component regression, aiming to improve the efficiency and accuracy of the regression estimator. By incorporating the low-rank structure, our method enables more effective modeling and estimation, particularly in high-dimensional and errors-in-variables regression settings. We provide a theoretical analysis of the proposed estimator's large-sample properties, including a comprehensive rate analysis of bias, variance, and additional variations due to measurement errors. Furthermore, our numerical experiments provide empirical evidence that supports the theoretical findings, demonstrating the superior performance of our approach. Overall, this work introduces a promising framework for regression analysis of non-Euclidean variables, effectively addressing the challenges associated with limited and noisy covariate data, with potential applications in diverse fields.

1 Introduction

Regression analysis is a fundamental statistical methodology to model the relationship between response variables and explanatory variables (covariates). Linear regression, for example, models the (conditional) expected value of the response variable as a linear function of covariates. Regression models enable researchers and analysts to make predictions, gain insights into how input variables influence the outcomes of interest, and validate hypothetical associations between variables in inferential studies. As a result, regression is widely utilized across various scientific domains, including economics, psychology, biology, and engineering [31, 21, 29].

In recent decades, there has been a growing interest in developing statistical methods capable of handling random objects in non-Euclidean spaces. Examples of these include functional data analysis [42], statistical manifold learning [32], statistical network analysis [35], and object-oriented data analysis [40]. In such contexts, the response variable is defined in a metric space that may lack an algebraic structure, making it challenging to apply global, parametric approaches toward regression as in the classical Euclidean setting. To overcome this challenge, (global) Fréchet regression, which models the relationship by fitting the (conditional) barycenters of the responses as a function of covariates, has been introduced [41]. Notably, when the Euclidean metric is considered, Fréchet regression recovers classical Euclidean regression models. For more details on Fréchet regression and its recent developments, we refer readers to [30, 41, 22, 46, 27].

Submitted to 37th Conference on Neural Information Processing Systems (NeurIPS 2023). Do not distribute.

Nevertheless, most existing research on Fréchet regression has focused on ideal scenarios characterized by abundant covariate data that are accurately measured and free of noise. In practical applications, however, high-dimensional data often arise, which are also susceptible to measurement errors and other forms of contamination. These errors can stem from various sources, such as unreliable data collection methods (*e.g.*, low-resolution probes, subjective self-reports) or imperfect data storage and transmission. The high-dimensionality and the presence of measurement errors in covariates pose critical challenges for statistical inference, as regression analysis based on error-prone covariates may result in incorrect associations between variables, yielding misleading conclusions.

To address these limitations, it is crucial to extend the methodology and analysis of Fréchet regression to tackle high-dimensional errors-in-variables problems. In this work, we aim to leverage the low-rank structure in the covariates to enhance the estimation accuracy and computational efficiency of Fréchet regression. Specifically, we explore the extension of principal component regression to handle errors-in-variables regression problems with non-Euclidean response variables.

1.1 Contributions

This paper contributes to advancing the (global) Fréchet regression of non-Euclidean response variables, with a particular focus on high-dimensional, errors-in-variables regression.

Firstly, we propose a novel framework, called the regularized (global) Fréchet regression (Section 3) that combines the ideas from Fréchet regression [41] and the principal component regression [33]. This framework effectively utilizes the low-rank structure in the matrix of (Euclidean) covariates by extracting its principal components via low-rank matrix approximation. Our proposed method is straightforward to implement, not requiring any knowledge about the error-generating mechanism.

Furthermore, we provide a comprehensive theoretical analysis in three main theorems (Section 4) to establish the effectiveness of the proposed framework. Firstly, we prove the consistency of the proposed estimator for the true global Fréchet regression model (Theorem 1). Secondly, we investigate the convergence rate of the estimator's bias and variance (Theorem 2). Lastly, we derive an upper bound for the distance between the estimates obtained using error-free covariates and those with errors-in-variables covariates (Theorem 3). Collectively, these results demonstrate that our approach effectively addresses model mis-specification and achieve more efficient model estimation by incorporating the low-rank structure of covariates, despite the presence of inherent bias due to unobserved measurement errors.

To validate our theoretical findings, we conduct numerical experiments on simulated datasets. Our results demonstrate that the proposed method provides more accurate estimates of the regression parameters, especially in high-dimensional settings. Our experiments emphasize the importance of incorporating the low-rank structure of covariates in Fréchet regression, and provide empirical evidence that aligns with our theoretical analysis.

1.2 Related work

Metric space-valued variables. Nonparametric regression models for Riemannian manifold-valued responses were proposed as a generalization of regression for multivariate outputs by Steinke *et al.* [49, 50]. These works provided a foundation for recent developments in regression analysis of non-Euclidean responses. Later, Hein [30] proposed a Nadaraya-Watson-type kernel estimation of regression model for general metric-space-valued outcomes. Since then, statistical properties of regression models for some special classes of metric space-valued outcomes, such as distribution functions [23, 53, 28] and matrix-valued responses [57, 20], have been investigated. Recently, many researchers have introduced further advances in Fréchet regression, including [41, 10, 38, 46]. In this study, we use the global Fréchet regression proposed by [41] as the basis for our proposed method.

Errors-in-variables regression. Much of earlier work on errors-in-variables problems in the statistical literature can be found in [13], which covers the simulation-extrapolation (SIMEX) [16, 11], the attenuation correction method [37], covariate-adjusted model [47, 19], and the deconvolution kernel method [25, 24, 18]. The regression calibration method [48], instrumental variable modeling [12, 44], and the two-phase study design [9, 4] were also proposed when additional data are available for correcting measurement errors. In the high-dimensional modeling literature, regularization methods for recovering the true covariate structure can also be utilized [39, 7, 17]. However, most of these methods require prior knowledge about the measurement error distributions.

Principal component regression. The principal component regression (PCR) [33] is a statistical technique that regresses response variables on principal component scores of the covariate matrix. The conventional PCR selects a few principal components as the "new" regressors associated with the first leading eigenvalues to explain the highest proportion of variations observed in the original covariate matrix. In functional data analysis, PCR is known to have a shrinkage effect on the model estimate and produce robust prediction performance in functional regression [43, 34]. Recently, Agarwal *et al.* [2] investigated the robustness of PCR in the presence of measurement errors on covariates and the statistical guarantees for learning a good predictive model. Motivated by these findings, we will adopt the PCR framework to improve the estimation and prediction performance of the errors-in-variables Fréchet regression. in this study.

1.3 Organization

This paper is organized as follows. In Section 2, we introduce the notation used throughout the paper, and overview the global Fréchet regression framework. Section 3 presents the problem setup, objectives, and our proposed estimator, which we refer to as the regularized Fréchet regression (Definition 4). In Section 4, we discuss theoretical guarantees on the regularized Fréchet regression method in accurately estimating the global Fréchet regression function. Section 5 presents the results of numerical 'proof-of-concept' experiments that support the theoretical findings. Finally, we conclude this paper with discussions in Section 6. Due to space constraints, detailed proofs of the theorems as well as additional details and discussions of experiments are provided in the Appendix.

2 Preliminaries

2.1 Notation

Let $\mathbb{N}$ denote the set of positive integers and $\mathbb{R}$ denote the set of real numbers. Also, let $\mathbb{R}_+ := \{x \in \mathbb{R} : x \geq 0\}$. For $n \in \mathbb{N}$, we let $[n] := \{1, \dots, n\}$. We mostly use plain letters to denote scalars, vectors, and random variables, but we also use boldface uppercase letters for matrices, and curly letters to denote sets when useful. Note that we may identify a vector with its column matrix representation. For a matrix $\boldsymbol{X}$, we let $\boldsymbol{X}^{-1}$ denote its inverse (if exists) and $\boldsymbol{X}^{\dagger}$ denote the Moore-Penrose pseudoinverse of $\boldsymbol{X}$. Also, we let $\operatorname{rowsp}(\boldsymbol{X})$ and $\operatorname{colsp}(\boldsymbol{X})$ denote the row and column spaces of $\boldsymbol{X}$, respectively. Furthermore, we let $\operatorname{spec}(\boldsymbol{X})$ denote the set of non-zero singular values of $\boldsymbol{X}$, $\sigma_i(\boldsymbol{X})$ denote the i-th largest singular value of $\boldsymbol{X}$, and $\sigma^{(\lambda)}(\boldsymbol{X}) := \inf\{\sigma_i(\boldsymbol{X}) > \lambda : i \in \mathbb{N}\}$ with the convention $\inf \emptyset = \infty$. We let $\mathbf{1}_n = (1, 1, \dots, 1) \in \mathbb{R}^d$ and let $\mathbb{1}$ denote the indicator function. We let $\|\cdot\|$ denote a norm, and set $\|\cdot\| = \|\cdot\|_2$ (the ℓ_2-norm for vectors, and the spectral norm for matrices) by default, unless stated otherwise. For a finite set $\mathcal{D}$, we may identify $\mathcal{D}$ with its empirical measure $\nu_{\mathcal{D}}^{\text{emp}} = \frac{1}{|\mathcal{D}|}\sum_{x \in \mathcal{D}} \delta_x$, where δ_x denotes the Dirac measure supported on $\{x\}$.

Letting $f, g : \mathbb{R} \to \mathbb{R}$, we write $f(x) = O(g(x))$ as $x \to \infty$ if there exist $M > 0$ and $x_0 > 0$ such that $|f(x)| \leq M \cdot g(x)$ for all $x \geq x_0$. Likewise, we write $f(x) = \Omega(g(x))$ if $g(x) = O(f(x))$. Furthermore, we write $f(x) = o(g(x))$ as $x \to \infty$ if $\lim_{x \to \infty} \frac{f(x)}{g(x)} = 0$. For a sequence of random variables X_n, and a sequence a_n, we write $X_n = O_p(a_n)$ as $n \to \infty$ if for any $\varepsilon > 0$, there exists $M \in \mathbb{R}_+$ and $N \in \mathbb{N}$ such that $P\left(\left|\frac{X_n}{a_n}\right| > M\right) < \varepsilon$ for all $n \geq N$. Similarly, we write $X_n = o_p(a_n)$ if $\lim_{n \to \infty} P\left(\left|\frac{X_n}{a_n}\right| > \varepsilon\right) = 0$ for all $\varepsilon > 0$.

2.2 Global Fréchet regression

Let (X, Y) be a random variable that has a joint distribution $F_{X,Y}$ supported on $\mathbb{R}^p \times \mathcal{M}$, where $\mathbb{R}^p$ is the p-dimensional Euclidean space and $\mathcal{M} = (\mathcal{M}, d)$ is a metric space equipped with a distance function $d : \mathcal{M} \times \mathcal{M} \to \mathbb{R}$. We write the marginal distribution of X as F_X, and the conditional distribution of Y given X as $F_{Y|X}$.

Definition 1 (Fréchet regression function)**.** *Let (X, Y) be a random element that takes value in $\mathbb{R}^p \times \mathcal{M}$. The Fréchet regression function of Y on X is a function $\varphi^* : \mathbb{R}^p \to \mathcal{M}$ such that*

$$\varphi^*(x) = \operatorname*{arg\,min}_{y \in \mathcal{M}} \mathbb{E}\big[d^2(Y, y) \,|\, X = x\big], \qquad \forall x \in \operatorname{supp} F_X \subseteq \mathbb{R}^p. \tag{1}$$

We note that $\varphi^*(x)$ is the best predictor of Y given $X = x$, as it minimizes the marginal risk $\mathbb{E}\big[d^2(Y, \varphi^*(X))\big]$ under the squared-distance loss. In the literature, $\varphi^*(x)$ is also known as the conditional Fréchet mean [26] of Y given $X = x$. It is important to recognize that the existence and uniqueness of the Fréchet regression function are closely tied to the geometric characteristics of $\mathcal{M}$, and are not guaranteed in general [3, 8]. Nonetheless, extensive research has been conducted on the existence and uniqueness of Fréchet means in various metric spaces commonly encountered in practical applications. Examples include the unit circle in $\mathbb{R}^2$ [14], Riemannian manifolds [1, 5], Alexandrov spaces with non-positive curvature [52], metric spaces with upper bounded curvature [58], and Wasserstein space [59, 36].

While modeling and estimating the Fréchet regression function φ^* is often of interest, its global (parametric) modeling may not be straightforward, especially when $\mathcal{M}$ lacks a useful algebraic structure, such as an inner product. For instance, in classical linear regression analysis with $\mathcal{M} = \mathbb{R}$, the distribution of $(Y|X = x)$ is normally distributed with a mean of $\varphi^*(x) = \alpha + \beta^\top x$ and variance σ_Y^2, where α and β represent the regression coefficients. Similarly, when $\mathcal{M}$ possesses a linear-algebraic structure, one can specify a class of regression functions that quantifies the association between the expected outcome and covariates in an additive and multiplicative manner. However, the lack of an algebraic structure in general metric spaces may prevent us from characterizing the barycenter $\varphi^*(x)$ in the same way classical regression analysis determines the expected value of outcomes with changing covariates.

To address this challenge, Petersen and Müller [41] recently proposed to exploit algebraic structures in the space of covariates, $\mathbb{R}^p$, instead of $\mathcal{M}$. Specifically, they consider a weighted Fréchet mean as

$$\varphi(x) = \operatorname*{arg\,min}_{y \in \mathcal{M}} \mathbb{E}\big[w(X, x) \cdot d^2(Y, y)\big], \tag{2}$$

where $w : \mathbb{R}^p \times \mathbb{R}^p \to \mathbb{R}$ is an arbitrary weight function such that $w(\xi, x)$ denotes the influence of ξ at x. In particular, we define the global Fréchet regression function with a specific choice of w, following [41].

Definition 2 (Global Fréchet regression function)**.** *Let (X, Y) be a random variable in $\mathbb{R}^p \times \mathcal{M}$. Let $\mu = \mathbb{E}(X)$ and $\Sigma = \operatorname{Var}(X)$. The global Fréchet regression function of Y on X is a function $\varphi_{\text{glo}} : \mathbb{R}^p \to \mathcal{M}$ such that*

$$\varphi_{\text{glo}}(x) = \operatorname*{arg\,min}_{y \in \mathcal{M}} \mathbb{E}\big[w_{\text{glo}}(X, x) \cdot d^2(Y, y)\big] \tag{3}$$

where $w_{\text{glo}}(X, x) = 1 + (X - \mu)^\top \Sigma^{-1}(x - \mu)$.

Note that when $\mathcal{M}$ is an inner product space (e.g., $\mathcal{M} = \mathbb{R}$), the function φ_{glo} restores the standard least squares linear regression. For this reason, φ_{glo} is commonly referred to as the global Fréchet regression model for metric-space-valued outcomes in recent literature [41, 38, 54].

What does it mean by "global" and where does it come from? One might wonder why the term "global" is used to describe φ_{glo} as a Fréchet regression function. The use of the adjective "global" serves to emphasize its distinction from "local" nonparametric regression methods that interpolate data points. Notably, when $\mathcal{M}$ is a Hilbert space, φ_{glo} reduces to the natural linear models. For instance, if $\mathcal{M} = \mathbb{R}$, then it follows that $\varphi_{\text{glo}}(x) = \mathbb{E}\big[w_{\text{glo}}(X, x) \cdot Y\big] = \alpha + \beta^\top(x - \mu)$, where $\alpha = \mathbb{E}[Y]$ and $\beta = \Sigma^{-1} \cdot \mathbb{E}\big[(X - \mu) \cdot Y\big]$. These linear models hold uniformly for the evaluation point x. Similarly, in the case of an L^2 space equipped with the squared-distance metric $d^2(y, y') = \|y - y'\|_2^2$ induced by the L^2 norm, φ_{glo} represents the linear regression model for functional responses. Thus, φ_{glo} establishes a globally defined model that spans the entire space.

3 Problem and methodology

3.1 Problem formulation

Let (X, Y) be a random variable in $\mathbb{R}^p \times \mathcal{M}$ and $F_{X,Y}$ be their joint distribution. Let $\mathcal{D}_n = \{(X_i, Y_i) : i \in [n]\}$ be an independent and identically distributed (IID) sample drawn from $F_{X,Y}$. Note that we may identify the set $\mathcal{D}_n$ with its discrete measure (empirical distribution). We consider the problem of estimating the global Fréchet regression function φ_{glo} (see Definition 2) from data $\mathcal{D}_n$. In this setting, a natural estimator of φ_{glo} would be its sample-analogue estimator. With $\hat{\mu}_{\mathcal{D}_n} = \mathbb{E}_{(X,Y)\sim\mathcal{D}_n}(X) = \frac{1}{n}\sum_{i=1}^n X_i$ and $\hat{\Sigma}_{\mathcal{D}_n} = \mathrm{Var}_{(X,Y)\sim\mathcal{D}_n}(X) = \frac{1}{n}\sum_{i=1}^n (X_i - \hat{\mu}_{\mathcal{D}_n}) \cdot (X_i - \hat{\mu}_{\mathcal{D}_n})^\top$, the sample-analogue estimator $\widehat{\varphi}_{\mathcal{D}_n}$ is defined as

$$\widehat{\varphi}_{\mathcal{D}_n}(x) = \operatorname*{arg\,min}_{y\in\mathcal{M}} \left\{ \frac{1}{n} \sum_{(X_i,Y_i)\in\mathcal{D}_n} \widehat{w}_{\mathcal{D}_n}(X_i, x) \cdot d^2(Y_i, y)] \right\} \tag{4}$$

where $\widehat{w}_{\mathcal{D}_n}(X, x) = 1 + (X - \widehat{\mu}_{\mathcal{D}_n})^\top \widehat{\Sigma}_{\mathcal{D}_n}^{-1} (x - \widehat{\mu}_{\mathcal{D}_n})$. The statistical properties of $\widehat{\varphi}_{\mathcal{D}_n}$, including the asymptotic distribution, a ridge-type variable selection operation, and total variation regularization method have been investigated [41, 38, 54].

In practice, however, we may only be able to access $\widetilde{\mathcal{D}}_n = \{(Z_i, Y_i) : i \in [n]\}$ instead of $\mathcal{D}_n$, where

$$Z_i = X_i + \varepsilon_i, \qquad i = 1, \dots, n \tag{5}$$

denotes an error-prone observation of the covariates X by measurement error ε. This formulation corresponds to the classical errors-in-variables problem.

Objective. Given a dataset, either $\mathcal{D}_n$ or $\widetilde{\mathcal{D}}_n$, our aim is to produce an estimate $\widehat{\varphi}$ of the global Fréchet regression function φ_{glo} so that the prediction error is minimized. Specifically, we evaluate the performance of $\widehat{\varphi}$ by means of the distance in the response space, $d\big(\widehat{\varphi}(x), \varphi_{\text{glo}}(x)\big)$.

3.2 Fréchet regression with covariate principal components

Singular value thresholding. Among various low-rank matrix approximation methods, we consider the (hard) singular value thresholding (SVT). For any $\lambda \in \mathbb{R}_+$, we define the map $\mathrm{SVT}^{(\lambda)} : \mathbb{R}^{n\times p} \to \mathbb{R}^{n\times p}$ that removes all singular values that are less than the threshold λ. To be precise, $\mathrm{SVT}^{(\lambda)}$ can be expressed in terms of the singular value decomposition (SVD) as follows:

$$\boldsymbol{M} = \sum_{i=1}^{\min\{n,p\}} s_i \cdot u_i v_i^\top \text{ is a SVD} \quad \implies \quad \mathrm{SVT}^{(\lambda)}(\boldsymbol{M}) = \sum_{i=1}^{\min\{n,p\}} s_i \cdot \mathbb{1}\{s_i > \lambda\} \cdot u_i v_i^\top. \tag{6}$$

Regularized Fréchet regression. We introduce a variant of the sample-analog estimator of the global Fréchet regression function based on principal components of the sample covariance. To facilitate the description of our proposed estimator, we introduce additional notation here.

Definition 3 (Covariate mean/covariance)**.** *For a probability distribution ν on $\mathbb{R}^p \times \mathcal{M}$, the covariate mean of ν, denoted by μ_ν, and the covariate covariance of ν, denoted by Σ_ν, are defined as*

$$\mu_\nu = \mathbb{E}_{(X,Y)\sim\nu}(X) \qquad \textit{and} \qquad \Sigma_\nu = \mathrm{Var}_{(X,Y)\sim\nu}(X). \tag{7}$$

Recall that a finite set $\mathcal{D} \subset \mathbb{R}^p \times \mathcal{M}$ may be identified with its empirical distribution; it follows that

$$\mu_{\mathcal{D}} = \frac{1}{|\mathcal{D}|} \sum_{(x_i,y_i)\in\mathcal{D}} x_i \qquad \text{and} \qquad \Sigma_{\mathcal{D}} = \frac{1}{|\mathcal{D}|} \sum_{(x_i,y_i)\in\mathcal{D}} (x_i - \mu_{\mathcal{D}}) \cdot (x_i - \mu_{\mathcal{D}})^\top. \tag{8}$$

Definition 4 (Regularized Fréchet regression)**.** *Let ν be a probability distribution on $\mathbb{R}^p \times \mathcal{M}$ and $\lambda \in \mathbb{R}_+$. The λ-regularized Fréchet regression function for ν is a map $\varphi_\nu^{(\lambda)} : \mathbb{R}^p \to \mathcal{M}$ such that*

$$\varphi_\nu^{(\lambda)}(x) = \operatorname*{arg\,min}_{y\in\mathcal{M}} R_\nu^{(\lambda)}(y; x), \quad \textit{where} \quad R_\nu^{(\lambda)}(y; x) = \mathbb{E}_{(X,Y)\sim\nu} \left[w_\nu^{(\lambda)}(X, x) \cdot d^2(Y, y) \right]$$
$$\textit{and} \quad w_\nu^{(\lambda)}(x', x) = 1 + (x' - \mu_\nu)^\top \left[\mathrm{SVT}^{(\lambda)}(\Sigma_\nu) \right]^\dagger (x - \mu_\nu). \tag{9}$$

When $\mathcal{D}_n = \{(X_i, Y_i) \in \mathbb{R}^p \times \mathcal{M} : i \in [n]\}$ is an IID sample from $F_{X,Y}$, the λ-regularized estimator $\varphi^{(\lambda)}_{\mathcal{D}_n}$ subsumes the sample-analogue estimator $\widehat{\varphi}_{\mathcal{D}_n}$ in (4) as a special case where $\lambda = 0$.

Connection to principal component regression. Here we remark that when $\mathcal{M}$ is a Euclidean space, the regularized Fréchet regression function $\varphi^{(\lambda)}_{\nu}$ effectively reduces to the principal component regression. Suppose that $\mathcal{M} = \mathbb{R}$ and $\mathcal{D}_n = \{(x_i, y_i) \in \mathbb{R}^p \times \mathbb{R} : i \in [n]\}$ is a given dataset. Then $\varphi^{(\lambda)}_{\mathcal{D}_n}(x) = \overline{y} + \hat{\beta}_\lambda^\top (x - \mu_{\mathcal{D}_n})$ where $\overline{y} = \frac{1}{n}\sum_{i=1}^n y_i$ and $\hat{\beta}_\lambda = [\mathrm{SVT}^{(\lambda)}(\Sigma_{\mathcal{D}_n})]^\dagger \cdot \big[\frac{1}{n}\sum_{i=1}^n (x_i - \mu_{\mathcal{D}_n}) \cdot (y_i - \overline{y})\big]$. Observe that $\hat{\beta}_\lambda$ is exactly the regression coefficient of principal component regression applied to the centered dataset $\mathcal{D}_n^{\mathrm{ctr}} = \{(x_i - \mu_{\mathcal{D}_n}, y_i - \overline{y}) : i \in [n]\}$ using k principal components where $k = \max_{a \in [p]} \big\{\sigma_a(\Sigma_{\mathcal{D}_n^{\mathrm{ctr}}}) \geq \lambda\big\}$.

4 Main results

In this section, we investigate properties of $\varphi^{(\lambda)}_{\nu}$ for $\lambda \geq 0$, with a focus on two cases: $\nu = \mathcal{D}_n$ and $\nu = \widetilde{\mathcal{D}}_n$, cf. Section 3.1. By denoting the true distribution that generates (X, Y) as ν^*, we can express φ_{glo} as $\varphi^{(0)}_{\nu^*}$. To analyze the discrepancy between the regularized global Fréchet regression estimators and $\varphi_{\mathrm{glo}}(x)$, we examine the relationships depicted in the schematic in Figure 1. Our theoretical findings can be summarized as follows: even in the presence of covariate noises, $\varphi^{(\lambda)}_{\widetilde{\mathcal{D}}_n}$ with a suitable $\lambda > 0$ can effectively eliminate the noise in Z to estimate X, thereby reducing the error in estimating φ_{glo}.

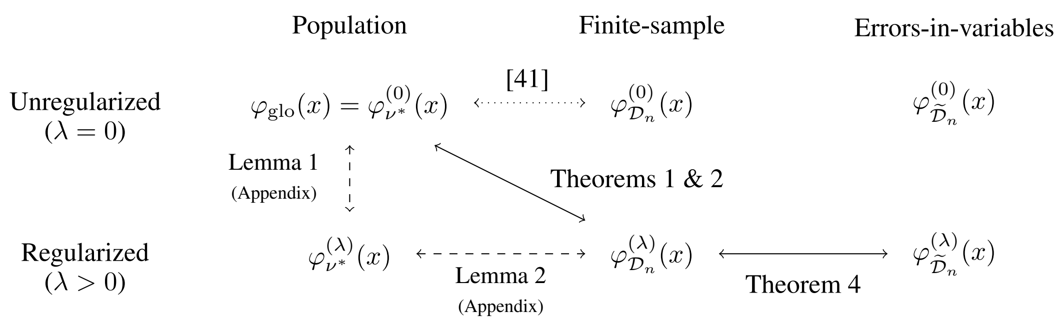

Figure 1: A schematic for the relationship between the regularized Fréchet regression estimators.

4.1 Model assumptions and examples

We impose the following assumptions for our analysis.

(C0) (`Existence`) For any probability distribution ν and any $\lambda \in \mathbb{R}_+$, the object $\varphi^{(\lambda)}_{\nu}(x)$ exists (almost surely) and is unique. In particular, $\inf_{y \in \mathcal{M}:\, d(y, \varphi_{\mathrm{glo}}(x)) > \varepsilon} R(y; x) > R(\varphi_{\mathrm{glo}}(x); x)$ for all $\varepsilon > 0$, where $R(y; x) := R^{(0)}_{\nu^*}(y; x)$.

(C1) (`Growth`) There exist $D_{\mathtt{g}} > 0$, $C_{\mathtt{g}} > 0$ and $\alpha > 1$, possibly depending on x, such that for any probability distribution ν and any $\lambda \in \mathbb{R}_+$,

$$\begin{cases} d\big(y, \varphi^{(\lambda)}_{\nu}(x)\big) < D_{\mathtt{g}} & \implies \quad R^{(\lambda)}_{\nu}(y; x) - R^{(\lambda)}_{\nu}\big(\varphi^{(\lambda)}_{\nu}(x); x\big) \geq C_{\mathtt{g}} \cdot d\big(y, \varphi^{(\lambda)}_{\nu}(x)\big)^{\alpha}, \\ d\big(y, \varphi^{(\lambda)}_{\nu}(x)\big) \geq D_{\mathtt{g}} & \implies \quad R^{(\lambda)}_{\nu}(y; x) - R^{(\lambda)}_{\nu}\big(\varphi^{(\lambda)}_{\nu}(x); x\big) \geq C_{\mathtt{g}} \cdot D_{\mathtt{g}}^{\alpha}. \end{cases} \tag{10}$$

(C2) (`Bounded entropy`) There exists $C_{\mathtt{e}} > 0$, possibly depending on y, such that

$$\limsup_{\delta \to 0} \int_0^1 \sqrt{1 + \log \mathfrak{N}\big(B_d(y, \delta), \delta\varepsilon\big)}\, \mathrm{d}\varepsilon \leq C_{\mathtt{e}}, \tag{11}$$

where $B_d(y, \delta) := \{y' \in \mathcal{M} : d(y, y') \leq \delta\}$ and $\mathfrak{N}(S, \varepsilon)$ is the ε-covering number[1] of S.

[1] A formal definition of covering number is provided in Appendix A; see Definition 6.

Assumption (C0) is common to establish the consistency of an M-estimator [55, Chapter 3.2]; in particular, it ensures the weak convergence of the empirical process $R^{(\lambda)}_{\mathcal{D}_n}$ to the population process $R^{(\lambda)}_{\nu^*}$ implying convergence of their minimizers. Furthermore, the conditions on the curvature (C1) and the covering number (C2) control the behavior of the objecdtives near the minimum in order to obtain rates of convergence; it is worth mentioning that (C2) corresponds to a (locally) bounded entropy for every $y \in \mathcal{M}$, while (P1) in [41] requires the same condition only with $y = \varphi_{\text{glo}}(x)$. These conditions arise from empirical process theory and are also commonly adopted [41, 45, 46].

Here we provide several examples of the space $\mathcal{M}$, in which the conditions (C0), (C1) and (C2) are satisfied. We verify the conditions in Appendix A; see Propositions 1, 2, 3, and 4.

Example 1. *Let $\mathcal{M} = (\mathcal{H}, d_{\text{HS}})$ be a finite-dimensional Hilbert space $\mathcal{H}$ equipped with the Hilbert-Schmidt metric $d_{\text{HS}}(y_1, y_2) = \langle y_1 - y_2, y_1 - y_2\rangle^{1/2}$, e.g., $\mathcal{M} = (\mathbb{R}^r, d_2)$ where d_2 is the ℓ^2-metric.*

Example 2. *Let $\mathcal{M}$ be $\mathcal{W}$, the set of probability distributions G on $\mathbb{R}$ such that $\int_{\mathbb{R}} x^2\, dG(x) < \infty$, equipped with the Wasserstein metric d_W defined as*

$$d_W(G_1, G_2)^2 = \int_0^1 \left(G_1^{-1}(t) - G_2^{-1}(t)\right)^2\, dt,$$

where G_1^{-1} and G_2^{-1} are the quantile functions of G_1 and G_2, respectively. See [41, Section 6].

Example 3. *Let $\mathcal{M} = \left\{M \in \mathbb{R}^{r\times r} : M = M^T, M \succeq 0 \text{ and } M_{ii} = 1, \forall i \in [r]\right\}$ be the set of correlation matrices of size r, equipped with the Frobenius metric, $d_F(M, M') = \|M - M'\|_F$.*

Example 4. *Let $\mathcal{M}$ be a (bounded) Riemannian manifold of dimension r, and let d_g be the geodesic distance induced by the Riemannian metric.*

4.2 Theorem statements

4.2.1 Noiseless covariate setting

First of all, we show the consistency of the λ-regularized Fréchet regression function.

Theorem 1 (Consistency)**.** *Suppose that Assumption (C0) holds. If $\operatorname{diam}(\mathcal{M}) < \infty$, then for any $\lambda \in \mathbb{R}$ such that $0 \le \lambda < \min\{\sigma_i(\Sigma_{\nu^*}) : \sigma_i(\Sigma_{\nu^*}) > 0\}$, and any $x \in \mathbb{R}^p$,*

$$d\big(\varphi^{(\lambda)}_{\mathcal{D}_n}(x), \varphi^{(0)}_{\nu^*}(x)\big) = o_P(1) \qquad \text{as } n \to \infty. \tag{12}$$

If $\lambda < \sigma^{(0)}(\Sigma_{\nu^*}) = \min\{\sigma_i(\Sigma_{\nu^*}) : \sigma_i(\Sigma_{\nu^*}) > 0\}$, then the regularized estimator $\varphi^{(\lambda)}_{\mathcal{D}_n}(x)$ effectively reduces to the same as the sample-analog estimator $\widehat{\varphi}_{\mathcal{D}_n}(x)$ in (4) in the limit $n \to \infty$. Thus, $\varphi^{(\lambda)}_{\mathcal{D}_n}(x)$ inherits the consistency of $\widehat{\varphi}_{\mathcal{D}_n}$. We provide a detailed proof of Theorem 1 in Appendix B.

In addition to the consistency of $\varphi^{(\lambda)}_{\mathcal{D}_n}$ in the small λ limit, we show the rate of its convergence that holds for any fixed $\lambda \in \mathbb{R}_+$.

Definition 5. *For a positive semidefinite matrix Σ, the <u>Mahalanobis seminorm</u> of x induced by Σ is*

$$\|x\|_\Sigma := \left(x^\top \Sigma^\dagger x\right)^{1/2}. \tag{13}$$

Theorem 2 (Rate of convergence)**.** *Suppose that Assumptions (C0)–(C2) hold. If $\operatorname{diam}(\mathcal{M}) < \infty$, then for any $\lambda \in \mathbb{R}$ and $x \in \mathbb{R}^p$ such that $\|x - \mu_{\nu^*}\|_{\Sigma_{\nu^*}} \le \frac{C_g \cdot D_g^\alpha}{\operatorname{diam}(\mathcal{M})^2 \cdot \sqrt{\operatorname{rank} \Sigma_{\nu^*}}}$,*

$$d\Big(\varphi^{(\lambda)}_{\mathcal{D}_n}(x), \varphi^{(0)}_{\nu^*}(x)\Big) = O_P\Big(\mathtt{b}_\lambda(x)^{\frac{1}{\alpha-1}} + n^{-\frac{1}{2(\alpha-1)}}\Big) \quad \text{as} \quad n \to \infty, \tag{14}$$

where $\mathtt{b}_\lambda(x) = \operatorname{rank}\left(\Sigma_{\nu^*} - \Sigma^{(\lambda)}_{\nu^*}\right)^{\frac{1}{2}} \cdot \|x - \mu_{\nu^*}\|_{\Sigma_{\nu^*} - \Sigma^{(\lambda)}_{\nu^*}}$.

We obtain Theorem 2 by showing a "bias" upper bound $d\big(\varphi^{(\lambda)}_{\nu^*}(x), \varphi^{(0)}_{\nu^*}(x)\big) = O\big(\mathtt{b}_\lambda(x)^{\frac{1}{\alpha-1}}\big)$ and a "variance" bound $d\big(\varphi^{(\lambda)}_{\mathcal{D}_n}(x), \varphi^{(\lambda)}_{\nu^*}(x)\big) = O_P\big(n^{-\frac{1}{2(\alpha-1)}}\big)$; see Lemmas 1 and 2 in Appendix C. Here we remark that $\mathtt{b}_\lambda(x)$ is a monotone non-decreasing function of λ, and if $\lambda < \sigma^{(0)}(\Sigma_{\nu^*})$ then $\mathtt{b}_\lambda(x) = 0$. Also, the condition on $\|x - \mu_{\nu^*}\|_{\Sigma_{\nu^*}}$ is introduced for a technical reason, and can be removed when $D_{\mathtt{g}} = \infty$.

Remark 1. Note that Condition (C1) holds with $D_{\mathtt{g}} = \infty$ and $\alpha = 2$ for Examples 1, 2 and 3. Thus, we have $d\big(\varphi^{(\lambda)}_{\mathcal{D}_n}(x), \varphi^{(0)}_{\nu^*}(x)\big) = O_P\big(\mathtt{b}_\lambda(x) + n^{-\frac{1}{2}}\big)$ as $n \to \infty$.

4.2.2 Error-prone covariate setting

Given a set $\mathcal{D}_n = \{(x_i, y_i) : i \in [n]\}$, let $\boldsymbol{X}_{\mathcal{D}_n} := [x_1 \quad \cdots \quad x_n]^\top \in \mathbb{R}^{n\times p}$. We let $\boldsymbol{X} = \boldsymbol{X}_{\mathcal{D}_n}$ and $\boldsymbol{Z} = \boldsymbol{X}_{\widetilde{\mathcal{D}}_n}$ for shorthand, and further, we let $\boldsymbol{X}_{\mathrm{ctr}} = \big(\boldsymbol{I}_n - \frac{1}{n}\mathbf{1}_n\mathbf{1}_n^\top\big)\boldsymbol{X}$ and $\boldsymbol{Z}_{\mathrm{ctr}} = \big(\boldsymbol{I}_n - \frac{1}{n}\mathbf{1}_n\mathbf{1}_n^\top\big)\boldsymbol{Z}$ denote the 'row-centered' matrices.

Theorem 3 (De-noising covariates)**.** *Suppose that Assumptions (C0) and (C1) hold. Then there exists a constant $C > 0$ such that for any $\lambda \in \mathbb{R}_+$, if*

$$x \in \mu_{\mathcal{D}_n} + \operatorname{rowsp} \boldsymbol{X}_{\mathrm{ctr}} \quad \textit{and} \quad \|x - \mu_{\mathcal{D}_n}\|_{\Sigma_{\mathcal{D}_n}} \le \frac{1}{2}\left(\frac{C_g \cdot D_g^\alpha}{2\operatorname{diam}(\mathcal{M})} \cdot \frac{\sigma^{(\lambda)}(\boldsymbol{X}_{\mathrm{ctr}}) \wedge \sigma^{(\lambda)}(\boldsymbol{Z}_{\mathrm{ctr}})}{\|\boldsymbol{Z} - \boldsymbol{X}\|} - 1\right), \tag{15}$$

then

$$d\left(\varphi^{(\lambda)}_{\widetilde{\mathcal{D}}_n}(x), \varphi^{(\lambda)}_{\mathcal{D}_n}(x)\right) \le C \cdot \left(\frac{\|\boldsymbol{Z} - \boldsymbol{X}\|}{\sigma^{(\lambda)}(\boldsymbol{X}_{\mathrm{ctr}}) \wedge \sigma^{(\lambda)}(\boldsymbol{Z}_{\mathrm{ctr}})} \cdot \frac{2 \cdot \|x - \mu_{\mathcal{D}_n}\|_{\Sigma_{\mathcal{D}_n}} + 1}{C_g}\right)^{\frac{1}{\alpha}}. \tag{16}$$

Again, we remark that the condition on $\|x - \mu_{\nu^*}\|_{\Sigma_{\nu^*}}$ in (15) can be removed when $D_{\mathsf{g}} = \infty$. It is worth noting that the quantity $\frac{\|\boldsymbol{Z}-\boldsymbol{X}\|}{\sigma^{(\lambda)}(\boldsymbol{X}_{\mathrm{ctr}}) \wedge \sigma^{(\lambda)}(\boldsymbol{Z}_{\mathrm{ctr}})}$ serves as the reciprocal of the signal-to-noise ratio because $\|\boldsymbol{Z} - \boldsymbol{X}\|$ captures the magnitude of the "noise" in the covariates, while $\min\left\{\sigma^{(\lambda)}(\boldsymbol{X}_{\mathrm{ctr}}),\ \sigma^{(\lambda)}(\boldsymbol{Z}_{\mathrm{ctr}})\right\}$ quantifies the strength of the "signal" remaining in the λ-SVT of the design matrix. Additionally, we observe that the error bound (16) increases proportionally to the normalized deviation of x from the mean, $\mu_{\mathcal{D}_n}$, which is a reasonable outcome. For the complete version of Theorem 3 and its proof, please refer to Appendix D.

4.3 Proof sketches

We outline our proofs for the main theorems, whose details are presented in Appendices B, C and D.

Proof of Theorem 1. We show that $R^{(\lambda)}_{\mathcal{D}_n}(y; x)$ weakly converges to $R^{(0)}_{\nu^*}(y; x)$ in the $\ell^\infty(\mathcal{M})$-sense. According to [55, Theorem 1.5.4], it suffices to show that (1) $R^{(\lambda)}_{\mathcal{D}_n}(y; x) - R^{(0)}_{\nu^*}(y; x) = o_p(1)$ for all $y \in \mathcal{M}$, and (2) $R^{(\lambda)}_{\mathcal{D}_n}$ is asymptotically equicontinuous in probability.

Proof of Theorem 2. We prove upper bounds for the bias and the variance separately.

<u>To control the bias</u> (Lemma 1 in Appendix C), we first show an upper bound for $R\big(\varphi^{(\lambda)}(x); x\big) - R\big(\varphi(x); x\big)$, and then convert it to an upper bound on the distance between the minimizers $d\big(\varphi^{(\lambda)}(x), \varphi(x)\big)$ using the `Growth` condition (C1). We utilize the so-called 'peeling technique' in empirical process theory in this conversion.

<u>To control the variance</u> (Lemma 2 in Appendix C), we follow a similar strategy as in Lemma 1, but with additional technical considerations. We define the 'fluctuation variable' $Z^{(\lambda)}_n(y; x) := R^{(\lambda)}_{\mathcal{D}_n}(y; x) - R^{(\lambda)}_{\nu^*}(y; x)$ parametrized by $y \in \mathcal{M}$, and derive a probabilistic upper bound for $R^{(\lambda)}_{\nu^*}\big(\varphi^{(\lambda)}_{\mathcal{D}_n}(x); x\big) - R^{(\lambda)}_{\nu^*}\big(\varphi^{(\lambda)}_{\nu^*}(x); x\big)$ by establishing a uniform upper bound for $Z^{(\lambda)}_n\big(y; x\big) - Z^{(\lambda)}_n\big(\varphi(x); x\big)$; here, the `Entropy` condition (C2) is used. Again, we use the `Growth` condition (C1) and the peeling technique to obtain a probabilistic upper bound for the distance $d\big(\varphi^{(\lambda)}_{\mathcal{D}_n}(x), \varphi^{(\lambda)}_{\nu^*}(x)\big)$.

Proof of Theorem 3. We express the difference in the objective functions $R^{(\lambda)}_{\widetilde{\mathcal{D}}_n}(y; x) - R^{(\lambda)}_{\mathcal{D}_n}(y; x)$ using the difference in the weights $w^{(\lambda)}_{\widetilde{\mathcal{D}}_n}(y; x) - w^{(\lambda)}_{\mathcal{D}_n}(y; x)$, which can be written in terms of $\boldsymbol{X}$ and $\boldsymbol{Z}$. We use classical matrix perturbation theory to control $R^{(\lambda)}_{\widetilde{\mathcal{D}}_n}(y; x) - R^{(\lambda)}_{\mathcal{D}_n}(y; x)$, and transform it to an upper bound on the distance $d\left(\varphi^{(\lambda)}_{\widetilde{\mathcal{D}}_n}(x), \varphi^{(\lambda)}_{\mathcal{D}_n}(x)\right)$ using the `Growth` condition (C1).

5 Experiments

In this section, we present the results of our numerical simulations, which aim to validate and support the theoretical findings presented earlier. We focus on the problem of global Fréchet regression analysis for one-dimensional distribution functions (Example 2) and conduct a comprehensive set of simulations under various conditions. Our simulations enable us to assess the performance of our proposed methodology and compare it to alternative approaches. Here we briefly summarize the results in Figure 2 and Table 1. See more details about the simulation settings, implementation details and evaluation metrics, as well as further discussions on the results, in Appendix E.

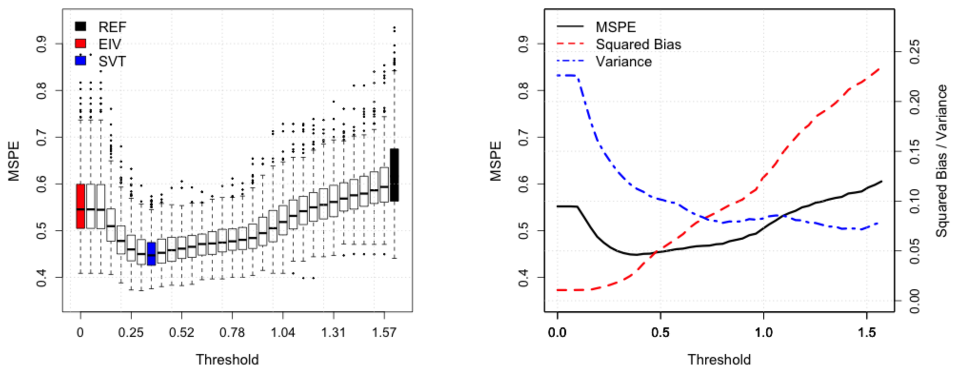

Figure 2: Comparison of the prediction performance of $\varphi^{(0)}_{\mathcal{D}_n}$ (REF), $\varphi^{(0)}_{\widetilde{\mathcal{D}}_n}$ (EIV), and $\varphi^{(\lambda)}_{\widetilde{\mathcal{D}}_n}$ (SVT) (left), and the trade-off between the bias and the variance (right) for $B = 500$, $p = 50$ and $n = 100$.

Table 1: Average performance of $\varphi^{(0)}_{\mathcal{D}_n}$ (REF), $\varphi^{(0)}_{\widetilde{\mathcal{D}}_n}$ (EIV), and $\varphi^{(\lambda)}_{\widetilde{\mathcal{D}}_n}$ (SVT) evaluated in four criteria via $B = 500$ Monte Carlo experiments under nine simulation scenarios (boldface=best).

p	Criterion	$n = 100$			$n = 200$			$n = 400$		
		REF	EIV	SVT	REF	EIV	SVT	REF	EIV	SVT
25	$\lvert\text{Bias}\rvert$	**0.016**	0.084	0.109	**0.011**	0.083	0.095	**0.007**	0.085	0.090
	$\sqrt{\text{Var}}$	0.332	0.290	**0.252**	0.214	0.187	**0.172**	0.145	0.126	**0.120**
	MSE	**0.224**	0.266	0.272	**0.267**	0.290	0.293	**0.285**	0.299	0.301
	MSPE	0.415	0.396	**0.380**	0.350	0.346	**0.343**	**0.325**	0.327	0.326
50	$\lvert\text{Bias}\rvert$	**0.024**	0.085	0.153	**0.015**	0.084	0.115	**0.010**	0.084	0.094
	$\sqrt{\text{Var}}$	0.567	0.495	**0.350**	0.327	0.287	**0.246**	0.211	0.186	**0.176**
	MSE	**0.148**	0.248	0.244	**0.227**	0.268	0.275	**0.267**	0.289	0.291
	MSPE	0.624	0.557	**0.452**	0.411	0.394	**0.378**	0.349	0.346	**0.344**
75	$\lvert\text{Bias}\rvert$	**0.046**	0.094	0.213	**0.019**	0.091	0.151	**0.011**	0.083	0.100
	$\sqrt{\text{Var}}$	1.000	0.884	**0.410**	0.436	0.384	**0.292**	0.270	0.237	**0.215**
	MSE	**0.073**	0.341	0.236	**0.187**	0.251	0.265	**0.247**	0.277	0.281
	MSPE	1.288	1.085	**0.513**	0.493	0.456	**0.411**	0.377	0.367	**0.360**

6 Discussion

This paper has addressed errors-in-variables regression of non-Euclidean response variables through the (global) Fréchet regression framework enhanced by low-rank approximation of covariates. Specifically, we introduce a novel regularized (global) Fréchet regression framework (Section 3), which combines the Fréchet regression with principal component regression. We also provide a comprehensive theoretical analysis in three main theorems (Section 4), and validate our theory through numerical experiments on simulated datasets.

We conclude this paper by proposing several promising directions for future research. First, it would be worthwhile to explore the large sample theory for selecting the optimal threshold parameter λ in the proposed SVT method, in order to characterize the theoretical phase transition of the bias-variance trade-off in the regularized (global) Fréchet regression. Second, we believe that our framework could be extended to errors-in-variables Fréchet regression for response variables in a broader class of metric spaces, e.g., by leveraging the quadruple inequality proposed by Schötz [45, 46]. Lastly, investigating the asymptotic distribution of the proposed SVT estimator would be highly appealing in the statistical literature, as it would enable us to make statistical inferences on the conditional Fréchet mean in non-Euclidean spaces [6, 8] with errors-in-variables covariates.

References

[1] Bijan Afsari. Riemannian L^p center of mass: Existence, uniqueness, and convexity. Proceedings of the American Mathematical Society, 139(2):655–673, 2011.

[2] Anish Agarwal, Devavrat Shah, Dennis Shen, and Dogyoon Song. On robustness of principal component regression. Journal of the American Statistical Association, 116(536):1731–1745, 2021.

[3] Adil Ahidar-Coutrix, Thibaut Le Gouic, and Quentin Paris. Convergence rates for empirical barycenters in metric spaces: Curvature, convexity and extendable geodesics. Probability Theory and Related Fields, 177(1):323–368, 2020.

[4] Gustavo Amorim, Ran Tao, Sarah Lotspeich, Pamela A Shaw, Thomas Lumley, and Bryan E Shepherd. Two-phase sampling designs for data validation in settings with covariate measurement error and continuous outcome. Journal of the Royal Statistical Society Series A: Statistics in Society, 184(4):1368–1389, 2021.

[5] Marc Arnaudon and Laurent Miclo. Means in complete manifolds: Uniqueness and approximation. ESAIM: Probability and Statistics, 18:185–206, 2014.

[6] Dennis Barden, Huiling Le, and Megan Owen. Central limit theorems for Fréchet means in the space of phylogenetic trees. Electronic Journal of Probability, 18(25):1–25, 2013.

[7] Alexandre Belloni, Mathieu Rosenbaum, and Alexandre B Tsybakov. Linear and conic programming estimators in high dimensional errors-in-variables models. Journal of the Royal Statistical Society. Series B (Statistical Methodology), pages 939–956, 2017.

[8] Rabi Bhattacharya and Lizhen Lin. Omnibus CLTs for Fréchet means and nonparametric inference on non-Euclidean spaces. Proceedings of the American Mathematical Society, 145(1):413–428, 2017.

[9] Norman E Breslow and Thomas Lumley. Semiparametric models and two-phase samples: Applications to cox regression. IMS Collections, 9:65–77, 2013.

[10] Louis Capitaine, Jérémie Bigot, Rodolphe Thiébaut, and Robin Genuer. Fréchet random forests for metric space valued regression with non Euclidean predictors. arXiv preprint arXiv:1906.01741, 2019.

[11] Raymond J Carroll, Helmut Küchenhoff, F Lombard, and Leonard A Stefanski. Asymptotics for the SIMEX estimator in nonlinear measurement error models. Journal of the American Statistical Association, 91(433):242–250, 1996.

[12] Raymond J Carroll, David Ruppert, Ciprian M Crainiceanu, Tor D Tosteson, and Margaret R Karagas. Nonlinear and nonparametric regression and instrumental variables. Journal of the American Statistical Association, 99(467):736–750, 2004.

[13] Raymond J Carroll, David Ruppert, Leonard A Stefanski, and Ciprian M Crainiceanu. Measurement Error in Nonlinear Models: A Modern Perspective. Chapman and Hall/CRC, 2006.

[14] Benjamin Charlier. Necessary and sufficient condition for the existence of a Fréchet mean on the circle. ESAIM: Probability and Statistics, 17:635–649, 2013.

[15] Yan Mei Chen, Xiao Shan Chen, and Wen Li. On perturbation bounds for orthogonal projections. Numerical Algorithms, 73(2):433–444, 2016.

[16] John R Cook and Leonard A Stefanski. Simulation-extrapolation estimation in parametric measurement error models. Journal of the American Statistical Association, 89(428):1314–1328, 1994.

[17] Abhirup Datta and Hui Zou. CoCoLasso for high-dimensional error-in-variables regression. The Annals of Statistics, 45(6):2400–2426, 2017.

[18] Aurore Delaigle, Jianqing Fan, and Raymond J Carroll. A design-adaptive local polynomial estimator for the errors-in-variables problem. Journal of the American Statistical Association, 104(485):348–359, 2009.

[19] Aurore Delaigle, Peter Hall, and Wen-xin Zhou. Nonparametric covariate-adjusted regression. The Annals of Statistics, 44(5):2190–2220, 2016.

[20] Shanshan Ding and R Dennis Cook. Matrix variate regressions and envelope models. Journal of the Royal Statistical Society Series B: Statistical Methodology, 80(2):387–408, 2018.

[21] David Donoho. 50 years of data science. Journal of Computational and Graphical Statistics, 26(4):745–766, 2017.

[22] Paromita Dubey and Hans-Georg Müller. Functional models for time-varying random objects. Journal of the Royal Statistical Society Series B: Statistical Methodology, 82(2):275–327, 2020.

[23] Juan José Egozcue, Josep Daunis-I-Estadella, Vera Pawlowsky-Glahn, Karel Hron, and Peter Filzmoser. Simplicial regression. The normal model. Journal of Applied Probability and Statistics, 2012.

[24] Jianqing Fan and Elias Masry. Multivariate regression estimation with errors-in-variables: Asymptotic normality for mixing processes. Journal of Multivariate Analysis, 43(2):237–271, 1992.

[25] Jianqing Fan and Young K Truong. Nonparametric regression with errors in variables. The Annals of Statistics, pages 1900–1925, 1993.

[26] Maurice Fréchet. Les éléments aléatoires de nature quelconque dans un espace distancié. In Annales de l'Institut Henri Poincaré, volume 10, pages 215–310, 1948.

[27] Aritra Ghosal, Wendy Meiring, and Alexander Petersen. Fréchet single index models for object response regression. Electronic Journal of Statistics, 17(1):1074–1112, 2023.

[28] Kyunghee Han, Hans-Georg Müller, and Byeong U Park. Additive functional regression for densities as responses. Journal of the American Statistical Association, 2019.

[29] X He, C Madigan, J Wellner, and B Yu. Statistics at a crossroads: Who is for the challenge. In Workshop report: National Science Foundation. https://www. nsf. gov/mps/dms/documents/Statistics_at_a_Crossroads_Workshop_Report_2019. pdf, 2019.

[30] Matthias Hein. Robust nonparametric regression with metric-space valued output. Advances in Neural Information Processing Systems, 22, 2009.

[31] Anthony JG Hey, Stewart Tansley, Kristin Michele Tolle, et al. The fourth paradigm: Data-intensive scientific discovery, volume 1. Microsoft research Redmond, WA, 2009.

[32] Alan J. Izenman. Modern Multivariate Statistical Techniques: Regression, Classification, and Manifold Learning. Springer, 2008.

[33] Ian T Jolliffe. A note on the use of principal components in regression. Journal of the Royal Statistical Society Series C: Applied Statistics, 31(3):300–303, 1982.

[34] Ioannis Kalogridis and Stefan Van Aelst. Robust functional regression based on principal components. Journal of Multivariate Analysis, 173:393–415, 2019.

[35] Eric D. Kolaczyk. Statistical Analysis of Network Data. Springer, 2009.

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

[37] Hua Liang, Wolfgang Härdle, and Raymond J Carroll. Estimation in a semiparametric partially linear errors-in-variables model. The Annals of Statistics, 27(5):1519–1535, 1999.

[38] Zhenhua Lin and Hans-Georg Müller. Total variation regularized Fréchet regression for metric-space valued data. The Annals of Statistics, 49(6):3510–3533, 2021.

[39] Po-Ling Loh and Martin J Wainwright. High-dimensional regression with noisy and missing data: Provable guarantees with non-convexity. Advances in Neural Information Processing Systems, 24, 2011.

[40] James Stephen Marron and Ian L Dryden. Object Oriented Data Analysis. CRC Press, 2021.

[41] Alexander Petersen and Hans-Georg Müller. Fréchet regression for random objects with Euclidean predictors. The Annals of Statistics, 47(2):691–719, 2019.

[42] J. O. Ramsay and B. W. Silverman. Functional Data Analysis. Springer, 2 edition, 2005.

[43] Philip T Reiss and R Todd Ogden. Functional principal component regression and functional partial least squares. Journal of the American Statistical Association, 102(479):984–996, 2007.

[44] Susanne M Schennach. Instrumental variable estimation of nonlinear errors-in-variables models. Econometrica, 75(1):201–239, 2007.

[45] Christof Schötz. Convergence rates for the generalized Fréchet mean via the quadruple inequality. Electronic Journal of Statistics, 13:4280–4345, 2019.

[46] Christof Schötz. Nonparametric regression in nonstandard spaces. Electronic Journal of Statistics, 16(2):4679–4741, 2022.

[47] Damla Şentürk and Hans-Georg Müller. Covariate-adjusted regression. Biometrika, 92(1):75–89, 2005.

[48] Donna Spiegelman, Aidan McDermott, and Bernard Rosner. Regression calibration method for correcting measurement-error bias in nutritional epidemiology. The American Journal of Clinical Nutrition, 65(4):1179S–1186S, 1997.

[49] Florian Steinke and Matthias Hein. Non-parametric regression between manifolds. Advances in Neural Information Processing Systems, 21, 2008.

[50] Florian Steinke, Matthias Hein, and Bernhard Schölkopf. Nonparametric regression between general Riemannian manifolds. SIAM Journal on Imaging Sciences, 3(3):527–563, 2010.

[51] Gilbert W Stewart. On the perturbation of pseudo-inverses, projections and linear least squares problems. SIAM Review, 19(4):634–662, 1977.

[52] Karl-Theodor Sturm. Probability measures on metric spaces of nonpositive curvature. Heat Kernels and Analysis on Manifolds, Graphs, and Metric Spaces: Lecture Notes from a Quarter Program on Heat Kernels, Random Walks, and Analysis on Manifolds and Graphs, 338:357, 2003.

[53] Renáta Talská, Alessandra Menafoglio, Jitka Machalová, Karel Hron, and E Fišerová. Compositional regression with functional response. Computational Statistics & Data Analysis, 123:66–85, 2018.

[54] Danielle C Tucker, Yichao Wu, and Hans-Georg Müller. Variable selection for global Fréchet regression. Journal of the American Statistical Association, pages 1–15, 2021.

[55] Aad W Vaart and Jon A Wellner. Weak convergence. In Weak Convergence and Empirical Processes, pages 16–28. Springer, 1996.

[56] Roman Vershynin. High-dimensional probability: An introduction with applications in data science, volume 47. Cambridge University Press, 2018.

[57] Cinzia Viroli. On matrix-variate regression analysis. Journal of Multivariate Analysis, 111:296–309, 2012.

[58] Takumi Yokota. Convex functions and barycenter on CAT(1)-spaces of small radii. Journal of the Mathematical Society of Japan, 68(3):1297–1323, 2016.

[59] Y Zemel and V M Panaretos. Fréchet means and procrustes analysis in Wasserstein space. Bernoulli, 25(2):932–976, 2019.

A Verification of the model assumptions

A.1 Additional background

Definition 6 (ε-net and covering number)**.** *Let $(\mathcal{M}, d)$ be a metric space. Let $S \subseteq T$ be a subset and let $\varepsilon > 0$. A subset $\mathcal{N} \subseteq S$ is called an* <u>*ε-net*</u> *of S if every point in S is within distance ε of some point $\mathcal{N}$, i.e.,*

$$\forall x \in S, \ \ \exists x_0 \in \mathcal{N} \text{ such that } d(x, x_0) \le \varepsilon.$$

The <u>*ε-covering number*</u> *of S, denoted by $\mathfrak{N}(S, \varepsilon)$, is the smallest possible cardinality of an ε-net of S, i.e.,*

$$\mathfrak{N}(S, \varepsilon) := \min \left\{ k \in \mathbb{N} : \exists y_1, \dots, y_k \in \mathcal{M} \text{ such that } S \subseteq \bigcup_{i=1}^{k} B_d(y_i, \varepsilon) \right\}, \tag{17}$$

where $B_d(y, \varepsilon) = \{y' \in \mathcal{M} : d(y, y') \le \varepsilon\}$ denotes the closed ε-ball centered at $y \in \mathcal{M}$.

Let $B_2^r(0,1) := \{x \in \mathbb{R}^r : \|x\|_2 \le 1\}$ denote the unit ℓ^2-norm ball in $\mathbb{R}^r$. It is well known[2] that for any $\varepsilon > 0$,

$$\left(\frac{1}{\varepsilon}\right)^r \le \mathfrak{N}\big(B_2^r(0,1), \varepsilon\big) \le \left(\frac{2}{\varepsilon} + 1\right)^r. \tag{18}$$

A.2 Example 1: Euclidean space

Proposition 1. *The space $(\mathcal{H}, d_{\mathrm{HS}})$ defined in Example 1 satisfies Assumptions (C0), (C1), and (C2).*

Proof of Proposition 1. For any probability distribution ν and any $\lambda \in \mathbb{R}_+$, let $y_\nu^{(\lambda)} := \mathbb{E}_\nu\big[w_\nu^{(\lambda)}(X, x) \cdot Y\big]$. Then we observe that

$$\begin{aligned}
R_\nu^{(\lambda)}(y; x) &= \mathbb{E}_\nu\big[w_\nu^{(\lambda)}(X, x) \cdot d^2(Y, y)\big] \\
&= \mathbb{E}_\nu\left[w_\nu^{(\lambda)}(X, x) \cdot \|Y - y\|^2\right] \\
&= \mathbb{E}_\nu\left[w_\nu^{(\lambda)}(X, x) \cdot \|Y - y_\nu^{(\lambda)}\|^2\right] + \|y - y_\nu^{(\lambda)}\|_{\mathrm{HS}}^2 \\
&\qquad + 2\left\langle \underbrace{\mathbb{E}_\nu\left[w_\nu^{(\lambda)}(X, x) \cdot \big(Y - y_\nu^{(\lambda)}\big)\right]}_{=0}, \ y_\nu^{(\lambda)} - y \right\rangle \\
&= R_\nu^{(\lambda)}(y_\nu^{(\lambda)}; x) + \|y - y_\nu^{(\lambda)}\|_{\mathrm{HS}}^2.
\end{aligned}$$

As $R_\nu^{(\lambda)}(y; x)$ is a strictly convex and coercive function, there exists a unique minimizer, $\varphi_\nu^{(\lambda)}$. Thus, Condition (C0) is proved. Furthermore, Condition (C1) is also satisfied with $D_{\mathrm{g}} = \infty$, $C_{\mathrm{g}} = 1$, and $\alpha = 2$.

Lastly, for any $y \in \mathcal{H}$ and any $\varepsilon > 0$,

$$\mathfrak{N}\big(B_{d_{\mathrm{HS}}}(y, \delta), \delta\varepsilon\big) = \mathfrak{N}\big(B_{d_{\mathrm{HS}}}(y, 1), \varepsilon\big) \le \left(\frac{2}{\varepsilon} + 1\right)^r \le C \cdot \varepsilon^{-r}$$

where $r = \dim \mathcal{H}$ and $C > 1$ is a constant that depends on r only; see the covering number upper bound in (18). Thus, the integral (11) is bounded as follows:

$$\begin{aligned}
\int_0^1 \sqrt{1 + \log \mathfrak{N}\big(B_d\big(\varphi(x), \delta\big), \delta\varepsilon\big)} \, \mathrm{d}\varepsilon &\le \int_0^1 \sqrt{1 + \log C - r \log \varepsilon} \, \mathrm{d}\varepsilon \\
&\le \sqrt{1 + \log C} + \sqrt{r} \int_0^1 \sqrt{-\log \varepsilon} \, \mathrm{d}\varepsilon \\
&= \sqrt{1 + \log C} + \sqrt{r} \int_0^\infty e^{-t} \sqrt{t} \, \mathrm{d}t \\
&= \sqrt{1 + \log C} + \frac{\sqrt{r\pi}}{2}
\end{aligned}$$

[2] See [56, Corollary 4.2.13] for example.

using the change of variable $t = -\log \varepsilon$. Therefore, Assumption (C2) holds with $C_{\mathsf{e}} = \sqrt{1 + \log C} + \frac{\sqrt{r\pi}}{2}$.

$\square$

A.3 Example 2: set of probability distributions

Proposition 2. *The space $(\mathcal{W}, d_W)$ defined in Example 2 satisfies Assumptions (C0), (C1), and (C2).*

Proof of Proposition 2. For a probability distribution function $y \in \mathcal{W}$ defined on $\mathbb{R}$, let $\mathcal{Q} = Q(\mathcal{W}) := \{Q(y) : y \in \mathcal{W}\}$ denote the collection of corresponding quantile functions, where $\big(Q(y)\big)(u) = y^{-1}(u)$ for $u \in [0, 1]$.

We note that $f \mapsto \mathbb{E}_\nu\big[w_\nu^{(\lambda)}(X, x)\,\langle Q(Y), f\rangle\big]$ is a bounded linear functional defined on $L^2[0, 1]$ because $\mathbb{E}_\nu|w_\nu^{(\lambda)}(X, x)|^2 \le 2+2p\,\|(x-\mu_\nu)\|^2_{\Sigma_\nu}$ implies that $\mathbb{E}_\nu\big[w_\nu^{(\lambda)}(X, x)|\cdot\|Q(Y)\|_2\big]$ is bounded. It follows from the Riesz representation theorem that there exists $f_x^{(\lambda)} \in L^2[0, 1]$ such that

$$\mathbb{E}_\nu\big[w^{(\lambda)}(X, x)\,\langle Q(Y), g\rangle_2\big] = \langle f_x^{(\lambda)}, g\rangle_2 \tag{19}$$

for all $g \in L^2[0, 1]$. Then, we have

$$R_\nu^{(\lambda)}(y; x) = \mathbb{E}_\nu\big[w_\nu^{(\lambda)}(X, x)\,\|Q(Y) - f_x^{(\lambda)}\|_2^2\big] + \|Q(y) - f_x^{(\lambda)}\|_2^2, \tag{20}$$

which yields that

$$\varphi_\nu^{(\lambda)}(x) = Q^{-1}\left(\operatorname*{arg\,min}_{Q\in\mathcal{Q}} \|Q - f_x^{(\lambda)}\|_2^2\right). \tag{21}$$

The condition (C0) follows from the convexity of $(\mathcal{Q}, \|\cdot\|_2)$. Moreover, the convexity also gives $\big\langle Q(y) - Q(\varphi_\nu^{(\lambda)}(x)), f_x^{(\lambda)}(x) - Q(\varphi_\nu^{(\lambda)}(x))\big\rangle_2 \le 0$ for all $y \in \mathcal{W}$, so that

$$\begin{aligned}
&R_\nu^{(\lambda)}(y; x) - R_\nu^{(\lambda)}(\varphi^{(\lambda)}(x); x)\\
&\quad= \|Q(y) - f_x^{(\lambda)}(x)\|_2^2 - \|Q(\varphi_\nu^{(\lambda)}(x)) - f_x^{(\lambda)}(x)\|_2^2\\
&\quad= \|Q(y) - Q(\varphi_\nu^{(\lambda)}(x))\|_2 - 2\big\langle Q(y) - Q(\varphi_\nu^{(\lambda)}(x)), f_x^{(\lambda)}(x) - Q(\varphi_\nu^{(\lambda)}(x))\big\rangle_2\\
&\quad\ge \|Q(y) - Q(\varphi_\nu^{(\lambda)}(x))\|_2\\
&\quad= d_W^2(y, \varphi_\nu^{(\lambda)}(x)).
\end{aligned} \tag{22}$$

Therefore, the condition (C1) holds for any arbitrary constant $D_{\mathsf{g}} > 0$ with $C_{\mathsf{g}} = 1$ and $\alpha = 2$.

Finally, we refer to [41, Proposition 1] to ensure that for any $\delta > 0$ and any $\varepsilon > 0$,

$$\sup_{y\in\mathcal{W}} \log \mathfrak{N}\big(B_{d_W}(y, \delta), D_{\mathsf{e}}\varepsilon\big) \le \sup_{Q\in\mathcal{Q}} \log \mathfrak{N}\big(B_{d_2}(Q, \delta), \delta\varepsilon\big) \le C \cdot \varepsilon^{-1} \tag{23}$$

holds with an absolute constant $C > 0$. Technically, this fact comes from the covering number bound for a class of uniformly bounded and monotone functions in L^2. This confirms that the entropy condition (C2) holds. $\square$

A.4 Example 3: set of correlation matrices

Proposition 3. *The space $(\mathcal{M}, d_F)$ defined in Example 3 satisfies Assumptions (C0), (C1), and (C2).*

Proof of Proposition 3. First of all, we note that $\mathcal{M}$ is a convex subset of $\mathcal{S}^r := \{X \in \mathbb{R}^{r\times r} : X = X^\top\}$, which is the set of $r \times r$ symmetric matrices. It is because $\mathcal{M} = \mathcal{S}_+^r \cap H$ where $\mathcal{S}_+^r$ denotes the cone of $r \times r$ positive semidefinite matrices and $H := \{X \in \mathcal{S}^r : X_{ii} = 1,\ \forall i \in [r]\}$ denotes an affine subspace of $\mathcal{S}^r$, both of which are convex.

Next, we observe that $\mathcal{S}^r$ equipped with the Frobenius metric d_F is isometrically isomorphic to $\mathbb{R}^{r(r+1)/2}$ equipped with the ℓ^2-metric. Hence, $(\mathcal{M}, d_F)$ satisfies Assumptions (C0), (C1), and (C2), inheriting these properties from the ambient space $\mathcal{S}^r$, which is established by Proposition 1. We note that the inheritance of (C0), (C1) relies on the convexity of $\mathcal{M}$, while (C2) is inherited simply based on the inclusion $\mathcal{M} \subset \mathcal{S}^r$. $\square$

A.5 Example 4: bounded Riemannian manifold

Proposition 4. *The space $(\mathcal{M}, d_g)$ defined in Example 4 satisfies Assumption (C2) provided that the Riemannian metric is equivalent to the ambient Euclidean metric.*

Furthermore, let $T_y\mathcal{M}$ be the tangent space of $\mathcal{M}$ at y, and $\mathrm{Exp}_y : T_y\mathcal{M} \to \mathcal{M}$ be the manifold exponential map at y. Let $g_\nu^{(\lambda)}(u; y, x) := R_\nu^{(\lambda)}\big(\mathbb{E}_y(u), x\big)$ for $u \in T_y\mathcal{M}$ If (C0) holds and the Hessian of $g_\nu^{(\lambda)}\big(u; \varphi_\nu^{(\lambda)}(x), x\big)$ is positive definite, then (C1) for some $D_g > 0$.

Proof of Proposition 3. Since $\mathcal{M}$ has finite dimension and is bounded, the bounded entropy condition (C2) follows from the metric equivalence.

Suppose that (C0) holds, and let $\delta > 0$ be the injectivity radius at $\varphi_\nu^{(\lambda)}(x)$. Consider $y \in \mathcal{M}$ such that $d\big(y, \varphi_\nu^{(\lambda)}(x)\big) < \delta$, and let $u_y = \mathrm{Log}_{\varphi_\nu^{(\lambda)}(x)}(y)$. Then we have

$$R_\nu^{(\lambda)}\big(y; x\big) - R_\nu^{(\lambda)}\big(\varphi_\nu^{(\lambda)}(x); x\big) = g_\nu^{(\lambda)}\big(u_y; \varphi_\nu^{(\lambda)}(x), x\big) - g_\nu^{(\lambda)}\big(0; \varphi_\nu^{(\lambda)}(x), x\big) = u_y^\top \nabla^2 g_\nu^{(\lambda)}(\bar{u}_y)\, u_y$$

for some $\bar{u}_y$ between 0 and u_y. Since $u_y^\top u_y = d\big(y, \varphi_\nu^{(\lambda)}(x)\big)^2$ and $g_\nu^{(\lambda)}$ is continuous, the positive definiteness of $\nabla^2 g_\nu^{(\lambda)}(\bar{u}_y)$ implies (C1) with $\alpha = 1$. □

B Proof of Theorem 1

Proof of Theorem 1. Recall from Definition 4, cf. (9), that for any probability distribution ν on $\mathbb{R}^p$, any $\lambda \in \mathbb{R}_+$, and any $x \in \mathbb{R}^p$, the λ-regularized Fréchet regression function evaluated at x is given as the minimizer of a function $R_\nu^{(\lambda)}$ as

$$\varphi_\nu^{(\lambda)}(x) = \operatorname*{arg\,min}_{y \in \mathcal{M}} R_\nu^{(\lambda)}(y; x)$$

where

$$R_\nu^{(\lambda)}(y; x) = \mathbb{E}_{(X,Y)\sim\nu}\left[w_\nu^{(\lambda)}(X, x) \cdot d^2(Y, y)\right] \quad \text{and}$$
$$w_\nu^{(\lambda)}(x', x) = 1 + (x' - \mu_\nu)^\top \left[\mathrm{SVT}^{(\lambda)}\big(\Sigma_\nu\big)\right]^\dagger (x - \mu_\nu).$$

In this proof, we follow a similar strategy to that in the proof of [41, Theorem 1]. Specifically, it suffices to show $\sup_{y\in\mathcal{M}} \big|R_{\mathcal{D}_n}^{(\lambda)}(y; x) - R_{\nu^*}^{(0)}(y; x)\big|$ converges to zero in probability, due to [55, Corollary 3.2.3]. To this end, we show $R_{\mathcal{D}_n}^{(\lambda)}(y; x)$ weakly converges to $R_{\nu^*}^{(0)}(y; x)$ in the $\ell^\infty(\mathcal{M})$-sense, and then apply [55, Theorem 1.3.6]. Again, according to [55, Theorem 1.5.4], this weak convergence can be proved by showing that

- (S1) $R_{\mathcal{D}_n}^{(\lambda)}(y; x) - R_{\nu^*}^{(0)}(y; x) = o_p(1)$ for all $y \in \mathcal{M}$, and
- (S2) $R_{\mathcal{D}_n}^{(\lambda)}$ is asymptotically equicontinuous in probability, i.e., for any $\varepsilon, \eta > 0$, there exists $\delta > 0$ such that

$$\limsup_n P\left(\sup_{y_1, y_2 \in \mathcal{M}:\, d(y_1, y_2) < \delta} \left|R_{\mathcal{D}_n}^{(\lambda)}(y_1; x) - R_{\mathcal{D}_n}^{(\lambda)}(y_2; x)\right| > \varepsilon\right) < \eta.$$

In what follows, we prove these two statements, (S1) and (S2), thereby completing the proof of Theorem 1.

Step 1: proof of (S1). First of all, we observe that

$$R_{\mathcal{D}_n}^{(\lambda)}(y; x) - R_{\nu^*}^{(0)}(y; x) = \underbrace{\Big(R_{\mathcal{D}_n}^{(\lambda)}(y; x) - R_{\mathcal{D}_n}^{(0)}(y; x)\Big)}_{=:T_1} + \underbrace{\Big(R_{\mathcal{D}_n}^{(0)}(y; x) - R_{\nu^*}^{(0)}(y; x)\Big)}_{=:T_2}. \tag{24}$$

We separately analyze the two terms T_1 and T_2 below to show $T_1 = o_p(1)$ and $T_2 = o_p(1)$ as $n \to \infty$.

(i) $\underline{T_1 = o_p(1)}$.

Let $\mathcal{D}_n = \{(X_i, Y_i) : i \in [n]\}$, and we re-write

$$T_1 = \frac{1}{n}\sum_{i=1}^{n}\left(w^{(\lambda)}_{\mathcal{D}_n}(X_i, x) - w^{(0)}_{\mathcal{D}_n}(X_i, x)\right)\cdot d^2(Y_i, y).$$

Letting $\widehat{\mu}_n = \mu_{\mathcal{D}_n}$, $\widehat{\Sigma}_n = \Sigma_{\mathcal{D}_n}$, and $\widehat{\Sigma}^{(\lambda)}_n = \mathrm{SVT}^{(\lambda)}(\widehat{\Sigma}_n)$ for shorthand, we observe that

$$w^{(\lambda)}_{\mathcal{D}_n}(X_i, x) - w^{(0)}_{\mathcal{D}_n}(X_i, x) = (X_i - \widehat{\mu}_n)^\top\left[\widehat{\Sigma}^{(\lambda),\dagger}_n - \widehat{\Sigma}^{\dagger}_n\right](x - \widehat{\mu}_n).$$

Let $\boldsymbol{X} = [X_1 \quad \cdots \quad X_n]^\top \in \mathbb{R}^{n\times p}$, and note that $\widehat{\Sigma}_n = \frac{1}{n}\left(\boldsymbol{X} - \mathbf{1}_n\widehat{\mu}_n^\top\right)^\top\left(\boldsymbol{X} - \mathbf{1}_n\widehat{\mu}_n^\top\right)$. Then it follows that

$$\frac{1}{n}\sum_{i=1}^{n}(X_i - \widehat{\mu}_n)^\top\left[\widehat{\Sigma}^{(\lambda),\dagger}_n - \widehat{\Sigma}^{\dagger}_n\right] = \frac{1}{n}\mathbf{1}_n^\top\left(\boldsymbol{X} - \mathbf{1}_n\widehat{\mu}_n^\top\right)\left[\widehat{\Sigma}^{(\lambda),\dagger}_n - \widehat{\Sigma}^{\dagger}_n\right]$$

Consider a singular value decomposition of $\boldsymbol{X} - \mathbf{1}_n\widehat{\mu}_n^\top$, namely,

$$\boldsymbol{X} - \mathbf{1}_n\widehat{\mu}_n^\top = \sum_{i=1}^{\min\{n,p\}} s_i \cdot u_i v_i^\top,$$

and observe that $\widehat{\Sigma}_n = \sum_{i=1}^{\min\{n,p\}} s_i^2 \cdot v_i v_i^\top$ is an eigenvalue decomposition of $\widehat{\Sigma}_n$. Letting $\mathcal{V}^{(\lambda)}_n := \operatorname{span}\left\{v_i : i \in [p],\ 0 < s_i \le \sqrt{\lambda}\right\}$ be a subspace of $\mathbb{R}^p$ spanned by the eigenvectors of $\widehat{\Sigma}_n$ corresponding to the nonzero eigenvalues no greater than λ, we have

$$\begin{aligned}
\widehat{\Sigma}^{(\lambda),\dagger}_n - \widehat{\Sigma}^{\dagger}_n &= \sum_{i=1}^{p}\frac{1}{s_i^2}\cdot\mathbb{1}\{s_i > \sqrt{\lambda}\}\cdot v_i v_i^\top - \sum_{i=1}^{p}\frac{1}{s_i^2}\cdot\mathbb{1}\{s_i > 0\}\cdot v_i v_i^\top \\
&= \sum_{i=1}^{p}\frac{1}{s_i^2}\cdot\mathbb{1}\{0 < s_i \le \sqrt{\lambda}\}\cdot v_i v_i^\top \\
&= \widehat{\Sigma}^{\dagger}_n \cdot \Pi_{\mathcal{V}^{(\lambda)}_n} \\
&= n\cdot\left(\boldsymbol{X} - \mathbf{1}_n\widehat{\mu}_n^\top\right)^\dagger\left(\boldsymbol{X} - \mathbf{1}_n\widehat{\mu}_n^\top\right)^{\dagger,\top}\cdot\Pi_{\mathcal{V}^{(\lambda)}_n}
\end{aligned} \tag{25}$$

where $\Pi_{\mathcal{V}^{(\lambda)}_n}$ denotes the projection matrix onto the subspace $\mathcal{V}^{(\lambda)}_n$. Note that $\Pi_{\mathcal{V}^{(\lambda)}_n} = 0$ if and only if $\min\left\{i \in [p] : 0 < s_i \le \sqrt{\lambda}\right\} = \emptyset$.

Therefore, we have

$$\begin{aligned}
T_1 &= \frac{1}{n}\sum_{i=1}^{n}\left(w^{(\lambda)}_{\mathcal{D}_n}(X_i, x) - w^{(0)}_{\mathcal{D}_n}(X_i, x)\right)\cdot d^2(Y_i, y) \\
&\le \frac{\operatorname{diam}(\mathcal{M})^2}{n}\mathbf{1}_n^\top\left(\boldsymbol{X} - \mathbf{1}_n\widehat{\mu}_n^\top\right)\left[\widehat{\Sigma}^{(\lambda),\dagger}_n - \widehat{\Sigma}^{\dagger}_n\right](x - \widehat{\mu}_n) \\
&= \operatorname{diam}(\mathcal{M})^2\cdot\mathbf{1}_n^\top\left(\boldsymbol{X} - \mathbf{1}_n\widehat{\mu}_n^\top\right)^{\dagger,\top}\cdot\Pi_{\mathcal{V}^{(\lambda)}_n}\cdot(x - \widehat{\mu}_n) && \because (25) \\
&= o_p(1).
\end{aligned}$$

The last line follows from the fact that $\sup_{i\in[p]}\left(\sigma_i(\widehat{\Sigma}_n) - \sigma_i(\Sigma_{\nu^*})\right) \to 0$ in probability, and thus, $\Pi_{\mathcal{V}^{(\lambda)}_n} \to 0$ in probability.

(ii) $\underline{T_2 = o_p(1)}$.

Letting $\tilde{R}_n(y;x) = \frac{1}{n}\sum_{i=1}^n w_{\nu^*}^{(0)}(X_i,x)\cdot d^2(Y_i,y)$, we decompose T_2 as follows:

$$\begin{aligned} T_2 &= R_{\mathcal{D}_n}^{(0)}(y;x) - \tilde{R}_n(y;x) + \tilde{R}_n(y;x) - R_{\nu^*}^{(0)}(y;x) \\ &= \underbrace{\frac{1}{n}\sum_{i=1}^n \left\{ w_{\mathcal{D}_n}^{(0)}(X_i,x) - w_{\nu^*}^{(0)}(X_i,x) \right\} \cdot d^2(Y_i,y)}_{=:T_{2A}} \\ &\qquad + \underbrace{\frac{1}{n}\sum_{i=1}^n \left\{ w_{\nu^*}^{(0)}(X_i,x)\cdot d^2(Y_i,y) - \mathbb{E}\left[w_{\nu^*}^{(0)}(X_i,x)\cdot d^2(Y_i,y) \right] \right\}}_{=:T_{2B}} \end{aligned}$$

Note that T_{2B} converges to 0 in probability by the weak law of large numbers.

Now it remains to show $T_{2A} = o_p(1)$. To this end, we note that

$$\begin{aligned} & w_{\mathcal{D}_n}^{(0)}(X_i,x) - w_{\nu^*}^{(0)}(X_i,x) = V_n(x) + X_i^\top W_n(x) \\ & \qquad\qquad \text{where} \quad \begin{cases} V_n(x) = -\widehat{\mu}_n^\top \widehat{\Sigma}_n^\dagger (x-\widehat{\mu}_n) + \mu^\top \Sigma^\dagger (x-\mu), \\ W_n(x) = \widehat{\Sigma}_n^\dagger (x-\widehat{\mu}_n) - \Sigma^\dagger(x-\mu). \end{cases} \end{aligned} \tag{26}$$

Since $\widehat{\mu}_n$ and $\widehat{\Sigma}_n$ respectively converge to μ and Σ in probability, it is possible to verify that $|V_n(x)|, \|W_n(x)\|$ converge to 0 in probability. As a result, T_2 also converges to 0 in probability.

All in all, we have $R_{\mathcal{D}_n}^{(\lambda)}(y;x) - R_{\nu^*}^{(0)}(y;x) = o_p(1)$, and thus, proved (S1).

Step 2: proof of (S2). For any $y_1, y_2 \in \mathcal{M}$,

$$\begin{aligned} \left| R_{\mathcal{D}_n}^{(\lambda)}(y_1;x) - R_{\mathcal{D}_n}^{(\lambda)}(y_2;x) \right| &= \left| \frac{1}{n}\sum_{i=1}^n w_{\mathcal{D}_n}^{(\lambda)}(X_i,x)\cdot \left\{ d^2(Y_i,y_1) - d^2(Y_i,y_2) \right\} \right| \\ &\le \frac{1}{n}\sum_{i=1}^n \left| w_{\mathcal{D}_n}^{(\lambda)}(X_i,x) \right| \cdot |d(Y_i,y_1) + d(Y_i,y_2)| \cdot |d(Y_i,y_1) - d(Y_i,y_2)| \\ &\le 2\,\mathrm{diam}\,(\mathcal{M}) \cdot d(y_1,y_2) \cdot \left(\frac{1}{n}\sum_{i=1}^n \left| w_{\mathcal{D}_n}^{(\lambda)}(X_i,x) \right| \right) \\ &= O_p\left(d(y_1,y_2) \right) \end{aligned}$$

where the O_p term is independent of $y_1, y_2 \in \mathcal{M}$. Therefore,

$$\sup_{y_1,y_2\in\mathcal{M}:\, d(y_1,y_2)<\delta} \left| R_{\mathcal{D}_n}^{(\lambda)}(y_1;x) - R_{\mathcal{D}_n}^{(\lambda)}(y_2;x) \right| = O_p(\delta),$$

which proves (S2).

□

C Proof of Theorem 2

In this section, we prove the two claims in Theorem 2. Specifically, in Section C.1, we present and prove a lemma that controls the bias in the population estimator (Lemma 1), and in Section C.2, we present and prove a lemma that controls the variance of the empirical estimator (Lemma 2).

C.1 Bias in the population estimator

We recall the definition of Mahalanobis seminorm from Definition 5: $\|x\|_\Sigma := \left(x^\top \Sigma^\dagger x \right)^{1/2}$.

Lemma 1. *Suppose that Assumptions (C0) and (C1) hold. If*

$$\|x-\mu_{\nu^*}\|_{\Sigma_{\nu^*}} \le \frac{C_g \cdot D_g^{\alpha}}{\operatorname{diam}(\mathcal{M})^2 \cdot \sqrt{\operatorname{rank}\Sigma_{\nu^*}}}, \tag{27}$$

then for any $\lambda \in \mathbb{R}_+$,

$$d\big(\varphi^{(\lambda)}(x), \varphi(x)\big) \le 2^{K_0} \cdot \mathfrak{b}_\lambda(x)^{\frac{1}{\alpha-1}} = O\Big(\mathfrak{b}_\lambda(x)^{\frac{1}{\alpha-1}}\Big) \tag{28}$$

where

$$K_0 = \left\lfloor \frac{1}{(\alpha-1)\log 2} \cdot \log\left(\frac{4\operatorname{diam}(\mathcal{M})}{C_g \cdot \big(1-2^{-(\alpha-1)}\big)}\right)\right\rfloor + 1 \quad \textit{and}$$

$$\mathfrak{b}_\lambda(x) = \sqrt{\operatorname{rank}\big(\Sigma_{\nu^*} - \Sigma_{\nu^*}^{(\lambda)}\big)} \cdot \|x-\mu_{\nu^*}\|_{\Sigma_{\nu^*} - \Sigma_{\nu^*}^{(\lambda)}}.$$

Proof of Lemma 1. For the sake of brevity, we write $\varphi^{(\lambda)}(x) = \varphi_{\nu^*}^{(\lambda)}(x)$ and $\varphi(x) = \varphi_{\nu^*}^{(0)}(x)$ throughout this proof, dropping the subscript ν^*. Likewise, we simply write $\mu = \mu_{\nu^*}$ and $\Sigma = \Sigma_{\nu^*}$.

Step 1: A naïve upper bound. Observe that for any $\lambda \in \mathbb{R}_+$, $x \in \mathbb{R}^p$, and $y \in \mathcal{M}$,

$$\begin{aligned}
&\big|R(y;x) - R^{(\lambda)}(y;x)\big| \\
&\quad = \Big| \mathbb{E}_{\nu^*}\Big[(X-\mu)^\top \cdot \Big(\Sigma^\dagger - \Sigma^{(\lambda),\dagger}\Big) \cdot (x-\mu) \cdot d^2(Y,y)\Big] \Big| \\
&\quad \le \operatorname{diam}(\mathcal{M})^2 \cdot \mathbb{E}_{\nu^*}\left[\|X-\mu\|_{\Sigma-\Sigma^{(\lambda)}}\right] \cdot \|x-\mu\|_{\Sigma-\Sigma^{(\lambda)}} && \because \text{Cauchy-Schwarz inequality} \\
&\quad \le \operatorname{diam}(\mathcal{M})^2 \cdot \Big(\mathbb{E}_{\nu^*}\|X-\mu\|^2_{\Sigma-\Sigma^{(\lambda)}}\Big)^{1/2} \cdot \|x-\mu\|_{\Sigma-\Sigma^{(\lambda)}} && \because \text{Jensen's inequality} \\
&\quad = \operatorname{diam}(\mathcal{M})^2 \cdot \sqrt{\operatorname{rank}\ \big(\Sigma - \Sigma^{(\lambda)}\big)} \cdot \|x-\mu\|_{\Sigma-\Sigma^{(\lambda)}},
\end{aligned} \tag{29}$$

where the last inequality follows from $\mathbb{E}_{\nu^*}\|X-\mu\|^2_{\Sigma-\Sigma^{(\lambda)}} = \operatorname{rank}\big(\Sigma-\Sigma^{(\lambda)}\big)$.

We observe that the upper bound in (29) is monotone non-decreasing with respect to $\lambda \in \mathbb{R}_+$, and it converges to 0 as $\lambda \to 0$. To see this, for any $\lambda \in \mathbb{R}_+$, we let

$$\mathcal{V}^{(\lambda)} := \operatorname{span}\{v_i : i \in [p],\ 0 < \lambda_i \le \lambda\}$$

where $\Sigma = \sum_{i=1}^p \lambda_i \cdot v_i v_i^\top$ is an eigenvdecomposition of Σ. Letting $\Pi_{\mathcal{V}^{(\lambda)}}$ denote the projection matrix onto the subspace $\mathcal{V}^{(\lambda)}$, we note that $\Sigma - \Sigma^{(\lambda)} = \Pi_{\mathcal{V}^{(\lambda)}}\Sigma\Pi_{\mathcal{V}^{(\lambda)}}$, and that $(\Sigma-\Sigma^{(\lambda)})^\dagger = \Pi_{\mathcal{V}^{(\lambda)}}\Sigma^\dagger\Pi_{\mathcal{V}^{(\lambda)}}$. Thus, $\operatorname{rank}\big(\Sigma-\Sigma^{(\lambda)}\big) = \dim\mathcal{V}^{(\lambda)}$, and furthermore, we notice that $\mathcal{V}^{(\lambda)} = \{0\}$ if and only if $\lambda < \lambda_{\min} := \min\{\lambda_i : \lambda_i > 0\}$. Therefore,

$$\lambda < \lambda_{\min} \qquad \Longrightarrow \qquad R^{(\lambda)}(y;x) - R(y;x) = 0 \qquad \Longrightarrow \qquad \varphi^{(\lambda)}(x) = \varphi(x),\ \ \forall x. \tag{30}$$

The observation (30), together with Assumption (C0), implies that $d\big(\varphi^{(\lambda)}(x), \varphi(x)\big) = o(1)$ as $\lambda \to 0$.

Step 2: Controlling risk difference. Next, we move on to determine the order of $d\big(\varphi^{(\lambda)}(x), \varphi(x)\big)$ — as a function of $\mathfrak{b}_\lambda(x)$ — for a fixed $\lambda \in \mathbb{R}$. We may assume $\lambda > \lambda_{\min}$ for the proof because the lemma is trivial otherwise, cf. (30). Assuming $\lambda > \lambda_{\min}$, we may decompose the difference in the population objective at $\varphi^{(\lambda)}(x)$ and $\varphi(x)$ as follows:

$$\begin{aligned}
R\big(\varphi^{(\lambda)}(x);x\big) - R\big(\varphi(x);x\big) &= \underbrace{\Big\{R\big(\varphi^{(\lambda)}(x);x\big) - R^{(\lambda)}\big(\varphi^{(\lambda)}(x);x\big) + R^{(\lambda)}\big(\varphi(x);x\big) - R\big(\varphi(x);x\big)\Big\}}_{=:\mathfrak{R}_1} \\
&\quad - \underbrace{\Big\{R^{(\lambda)}\big(\varphi(x);x\big) - R^{(\lambda)}\big(\varphi^{(\lambda)}(x);x\big)\Big\}}_{=:\mathfrak{R}_2}.
\end{aligned}$$

We observe that both $\mathfrak{R}_1$ and $\mathfrak{R}_2$ are non-negative, due to the optimality of $\varphi(x)$ and $\varphi^{(\lambda)}(x)$. Then, we obtain an upper bound for $\mathfrak{R}_1$ using a similar argument as in (29). Specifically,

$$\begin{aligned} R\big(\varphi^{(\lambda)}(x);x\big) - R\big(\varphi(x);x\big) &\le \mathfrak{R}_1 \\ &= \mathbb{E}_{\nu^*}\left[\left\{w^{(0)}_{\nu^*}(X,x) - w^{(\lambda)}_{\nu^*}(X,x)\right\}\cdot\left\{d^2\big(Y,\varphi^{(\lambda)}(x)\big) - d^2\big(Y,\varphi(x)\big)\right\}\right] \\ &\le 2\,\mathrm{diam}\,(\mathcal{M})\cdot \mathtt{b}_\lambda(x)\cdot d\big(\varphi^{(\lambda)}(x),\varphi(x)\big). \end{aligned} \tag{31}$$

Step 3: Converting risk difference to bias. Lastly, we convert the upper bound (31) to an upper bound on the distance $d\big(\varphi^{(\lambda)}(x),\varphi(x)\big)$ using Assumption (C1). To this end, we begin by confirming that

$$\begin{aligned} R\big(\varphi^{(\lambda)}(x);x\big) - R\big(\varphi(x);x\big) &= \mathbb{E}_{\nu^*}\left[(X-\mu)^\top\cdot\Sigma^\dagger\cdot(x-\mu)\cdot\left\{d^2\big(Y,\varphi^{(\lambda)}(x)\big) - d^2\big(Y,\varphi(x)\big)\right\}\right] \\ &\le \mathrm{diam}\,(\mathcal{M})^2\cdot\left(\mathbb{E}_{\nu^*}\|X-\mu\|_\Sigma^2\right)^{1/2}\cdot\|x-\mu\|_\Sigma \\ &= \mathrm{diam}\,(\mathcal{M})^2\cdot\sqrt{\mathrm{rank}\,\Sigma}\cdot\|x-\mu\|_\Sigma \\ &\le C_{\mathtt{g}}\cdot D_{\mathtt{g}}^\alpha. \end{aligned}$$

Thereafter, we choose an arbitrary $K\in\mathbb{N}$ and $r\in\mathbb{R}_+$ whose values will be determined later in this proof. Then we obtain the following inequality using the so-called peeling technique:

$$\begin{aligned} &\mathbb{1}\left\{d\big(\varphi^{(\lambda)}(x),\varphi(x)\big) > 2^K\cdot\mathtt{b}_\lambda(x)^r\right\} \\ &\quad= \sum_{k=K}^{\infty}\mathbb{1}\left\{2^k\cdot\mathtt{b}_\lambda(x)^r < d\big(\varphi^{(\lambda)}(x),\varphi(x)\big)\le 2^{k+1}\cdot\mathtt{b}_\lambda(x)^r\right\} \\ &\quad\le \sum_{k=K}^{\infty}\mathbb{1}\left\{2^k\cdot\mathtt{b}_\lambda(x)^r < d\big(\varphi^{(\lambda)}(x),\varphi(x)\big)\le 2^{k+1}\cdot\mathtt{b}_\lambda(x)^r\right\} \\ &\quad\le \sum_{k=K}^{\infty}\frac{R\big(\varphi^{(\lambda)}(x);x\big)-R\big(\varphi(x);x\big)}{C_{\mathtt{g}}\cdot\big(2^k\cdot\mathtt{b}_\lambda(x)^r\big)^\alpha}\cdot\mathbb{1}\left\{d\big(\varphi^{(\lambda)}(x),\varphi(x)\big)\le 2^{k+1}\cdot\mathtt{b}_\lambda(x)^r\right\}. \qquad \because (C1) \end{aligned} \tag{32}$$

Moreover, we decompose the numerator in the fraction appearing in the upper bound (32) as follows:

Combining (31) with (32), we have

$$\begin{aligned} &\mathbb{1}\left\{d\big(\varphi^{(\lambda)}(x),\varphi(x)\big) > 2^K\cdot\mathtt{b}_\lambda(x)^r\right\} \\ &\quad\le \sum_{k=K}^{\infty}\frac{2\,\mathrm{diam}\,(\mathcal{M})\cdot\mathtt{b}_\lambda(x)\cdot d\big(\varphi^{(\lambda)}(x),\varphi(x)\big)}{C_{\mathtt{g}}\cdot\big(2^k\cdot\mathtt{b}_\lambda(x)^r\big)^\alpha}\cdot\mathbb{1}\left\{d\big(\varphi^{(\lambda)}(x),\varphi(x)\big)\le 2^{k+1}\cdot\mathtt{b}_\lambda(x)^r\right\} \\ &\quad\le \frac{4\,\mathrm{diam}\,(\mathcal{M})}{C_{\mathtt{g}}}\cdot\mathtt{b}_\lambda(x)^{1-r(\alpha-1)}\sum_{k=K}^{\infty}\frac{1}{2^{k(\alpha-1)}}. \end{aligned} \tag{33}$$

Note that $C := \frac{4\,\mathrm{diam}\,\mathcal{M}}{C_{\mathtt{g}}} > 0$ is a constant independent of λ. Let $r = 1/(\alpha-1)$, and observe that the upper bound in (33) becomes smaller than 1 for a sufficiently large K. Specifically,

$$K \ge \left\lfloor \frac{1}{(\alpha-1)\log 2}\cdot\log\left(\frac{4\,\mathrm{diam}\,(\mathcal{M})}{C_{\mathtt{g}}\cdot\big(1-2^{-(\alpha-1)}\big)}\right)\right\rfloor + 1 \quad\Longrightarrow\quad \frac{4\,\mathrm{diam}\,(\mathcal{M})}{C_{\mathtt{g}}}\cdot\sum_{k=K}^{\infty}\frac{1}{2^{k(\alpha-1)}} < 1.$$

As a result, the inequality "$d\big(\varphi^{(\lambda)}(x),\varphi(x)\big) > 2^{K_0}\cdot\mathtt{b}_\lambda(x)^r$" in the indicator function must be false, and we conclude that

$$d\big(\varphi^{(\lambda)}(x),\varphi(x)\big) \le 2^{K_0}\cdot\mathtt{b}_\lambda(x)^{\frac{1}{\alpha-1}}.$$

□

C.2 Variance of the empirical estimator

Lemma 2. *Suppose that Assumptions (C0), (C1) and (C2) hold. For any $\lambda \in \mathbb{R}_+$ such that $\lambda \notin \operatorname{spec}\big(\Sigma_{\nu^*}\big)$, it holds that*

$$d\Big(\varphi_{\mathcal{D}_n}^{(\lambda)}(x), \varphi_{\nu^*}^{(\lambda)}(x)\Big) = O_P\Big(n^{-\frac{1}{2(\alpha-1)}}\Big).$$

Proof of Lemma 2. Recall from the definition of λ-regularized Fréchet regression (Definition 4) and (9) that

$$R_{\mathcal{D}_n}^{(\lambda)}(y;x) = \frac{1}{n}\sum_{i=1}^{n} w_{\mathcal{D}_n}^{(\lambda)}(X_i,x)\cdot d^2(Y_i,y) \quad \text{and} \quad R_{\nu^*}^{(\lambda)}(y;x) = \mathbb{E}_{(X,Y)\sim\nu^*}\left[w_{\nu^*}^{(\lambda)}(X,x)\cdot d^2(Y,y)\right].$$

Additionally, we define an auxiliary function $\tilde{R}_n(y;x)$ as the "empirical risk with population weight" such that

$$\tilde{R}_n(y;x) := \frac{1}{n}\sum_{i=1}^{n} w_{\nu^*}^{(\lambda)}(X_i,x)\cdot d^2(Y_i,y).$$

We present the rest of this proof in three steps, outlined as follows. In Step 1, we show the consistency of $\varphi_{\mathcal{D}_n}^{(\lambda)}(x)$, i.e., $d\big(\varphi_{\mathcal{D}_n}^{(\lambda)}(x), \varphi_{\nu^*}^{(\lambda)}(x)\big) = o_P(1)$ as $n\to\infty$. In Step 2, we define the discrepancy variable $Z_n^{(\lambda)}(y;x) := R_{\mathcal{D}_n}^{(\lambda)}(y;x) - R_{\nu^*}^{(\lambda)}(y;x)$ between the finite-sample and the population objectives, cf. (36), and prove a uniform upper bound for $Z_n^{(\lambda)}(y;x)$ that holds in a neighborhood of $\varphi_{\nu^*}^{(\lambda)}(y;x)$. Lastly, in Step 3, we utilize the peeling technique from empirical process theory to obtain the desired rate of convergence.

Step 1: Consistency. We first claim that $d\big(\varphi_{\mathcal{D}_n}^{(\lambda)}(x), \varphi_{\nu^*}^{(\lambda)}(x)\big) = o_P(1)$ by an argument similar to that used in the proof of Theorem 1. Specifcally, it suffices to show that

(S1') $R_{\mathcal{D}_n}^{(\lambda)}(y;x) - R_{\nu^*}^{(\lambda)}(y;x) = o_P(1)$, and

(S2') $R_{\mathcal{D}_n}^{(\lambda)}(\cdot;x) : \mathcal{M}\to\mathbb{R}$ is asymptotically equicontinuous in probability.

Note that we already showed the asymptotic equicontinuity in the proof of Theorem 1; see (S2). Thus, it remains to show the pointwise convergence in probability. To show (S1'), we decompose $R_{\mathcal{D}_n}^{(\lambda)}(y;x) - R_{\nu^*}^{(\lambda)}(y;x)$ as follows.

$$\begin{aligned} R_{\mathcal{D}_n}^{(\lambda)}(y;x) - R_{\nu^*}^{(\lambda)}(y;x) &= \big\{R_{\mathcal{D}_n}^{(\lambda)}(y;x) - \tilde{R}_n(y;x)\big\} + \big\{\tilde{R}_n(y;x) - R_{\nu^*}^{(\lambda)}(y;x)\big\} \\ &= \underbrace{\frac{1}{n}\sum_{i=1}^{n}\big\{w_{\mathcal{D}_n}^{(\lambda)}(X_i,x) - w_{\nu^*}^{(\lambda)}(X_i,x)\big\}\cdot d^2(Y_i,y)}_{:=A_n^{(\lambda)}(y;x)} \\ &\quad + \underbrace{\frac{1}{n}\sum_{i=1}^{n}\Big(w_{\nu^*}^{(\lambda)}(X_i,x)\cdot d^2(Y_i,y) - \mathbb{E}_{\nu^*}\big[w_{\nu^*}^{(\lambda)}(X_i,x)\cdot d^2(Y_i,y)\big]\Big)}_{:=B_n^{(\lambda)}(y;x)}. \end{aligned}$$

Next, we show that $A_n^{(\lambda)}(y;x)$ and $B_n^{(\lambda)}(y;x)$ respectively converge to 0 in probability.

- Letting $\widehat{\mu}_n = \mu_{\mathcal{D}_n}$, $\widehat{\Sigma}_n = \Sigma_{\mathcal{D}_n}$, and $\widehat{\Sigma}_n^{(\lambda)} = \mathrm{SVT}^{(\lambda)}(\widehat{\Sigma}_n)$ for shorthand, we can write

$$w_{\mathcal{D}_n}^{(\lambda)}(X_i,x) - w_{\nu^*}^{(\lambda)}(X_i,x) = V_n^{(\lambda)}(x) + X_i^\top W_n^{(\lambda)}(x),$$

similarly to (26), where

$$\begin{aligned} V_n^{(\lambda)}(x) &= -\widehat{\mu}_n^\top \left[\widehat{\Sigma}_n^{(\lambda)}\right]^\dagger (x-\widehat{\mu}_n) + \mu^\top \left[\Sigma^{(\lambda)}\right]^\dagger (x-\mu), \\ W_n^{(\lambda)}(x) &= \left[\widehat{\Sigma}_n^{(\lambda)}\right]^\dagger (x-\widehat{\mu}_n) - \left[\Sigma^{(\lambda)}\right]^\dagger (x-\mu). \end{aligned} \tag{34}$$

Since $\|\widehat{\mu}_n - \mu\|_2 = O_P(n^{-1/2})$ and $\|\widehat{\Sigma}_n^{(\lambda)} - \Sigma^{(\lambda)}\| = O_P(n^{-1/2})$ (if $\lambda \notin \operatorname{spec}\Sigma$) independent of $\lambda > 0$, we also have $|V_n^{(\lambda)}(x)| = O_P(n^{-1/2})$ and $\|W_n^{(\lambda)}(x)\|_2 = O_P(n^{-1/2})$. This implies that $A_n^{(\lambda)}(y;x) = o_P(1)$.

- Moreover, we note that if $\|x-\mu\|_\Sigma < \infty$, then the random variable $w_{\nu^*}^{(\lambda)}(X,x)$ has finite second moment

$$\begin{aligned} \mathbb{E}_{\nu^*}\left[w_{\nu^*}^{(\lambda)}(X,x)^2\right] &\le 2\left(1 + \mathbb{E}_{\nu^*}\left[\left|(X-\mu)^\top \left[\Sigma^{(\lambda)}\right]^\dagger (x-\mu)\right|^2\right]\right) \\ &\le 2\left(1 + \mathbb{E}_{\nu^*}\left[\left\|X-\mu\right\|_{\Sigma^{(\lambda)}}^2 \cdot \left\|x-\mu\right\|_{\Sigma^{(\lambda)}}^2\right]\right) \\ &\le 2\big\{1 + p\,\|x-\mu\|_\Sigma^2\big\}, \end{aligned} \tag{35}$$

regardless of the value of $\lambda > 0$. When $\operatorname{diam}(\mathcal{M}) < \infty$, the product $w_{\nu^*}^{(\lambda)}(X,x) \cdot d^2(Y,y)$ also has finite second moment. Since $B_n^{(\lambda)}(y;x)$ is the sample mean of IID random variables with mean zero and finite variance, it follows that

$$B_n^{(\lambda)}(y;x) = O_P\left(\sqrt{\frac{\operatorname{Var}\big[\,w_{\nu^*}^{(\lambda)}(X_1,x)\cdot d^2(Y_1,y)\,\big]}{n}}\right) = O_P\big(n^{-1/2}\big).$$

Step 2: Uniform control of the fluctuation in objective discrepancy. For any $\lambda \in \mathbb{R}_+$ and any $(x,y) \in \mathbb{R}^p \times \mathcal{M}$, we let $Z_n^{(\lambda)}(y;x)$ denote the random variable defined as

$$Z_n^{(\lambda)}(y;x) := R_{\mathcal{D}_n}^{(\lambda)}(y;x) - R_{\nu^*}^{(\lambda)}(y;x) \tag{36}$$

We observed that

$$\begin{aligned} &Z_n^{(\lambda)}\big(y;x\big) - Z_n^{(\lambda)}\big(\varphi_{\nu^*}^{(\lambda)}(x);x\big) \\ &\quad= \Big\{R_{\mathcal{D}_n}^{(\lambda)}(y;x) - R_{\nu^*}^{(\lambda)}(y;x)\Big\} - \Big\{R_{\mathcal{D}_n}^{(\lambda)}\big(\varphi_{\nu^*}^{(\lambda)}(x);x\big) - R_{\nu^*}^{(\lambda)}\big(\varphi_{\nu^*}^{(\lambda)}(x);x\big)\Big\} \\ &\quad= \Big[\Big\{R_{\mathcal{D}_n}^{(\lambda)}(y;x) - \tilde{R}_n(y;x)\Big\} - \Big\{R_{\mathcal{D}_n}^{(\lambda)}\big(\varphi_{\nu^*}^{(\lambda)}(x);x\big) - \tilde{R}_n\big(\varphi_{\nu^*}^{(\lambda)}(x);x\big)\Big\}\Big] \\ &\qquad+ \Big[\Big\{\tilde{R}_n(y;x) - R_{\nu^*}^{(\lambda)}(y;x)\Big\} - \Big\{\tilde{R}_n\big(\varphi_{\nu^*}^{(\lambda)}(x);x\big) - R_{\nu^*}^{(\lambda)}\big(\varphi_{\nu^*}^{(\lambda)}(x);x\big)\Big\}\Big] \\ &\quad= \underbrace{\frac{1}{n}\sum_{i=1}^n \Big\{w_{\mathcal{D}_n}^{(\lambda)}(X_i,x) - w_{\nu^*}^{(\lambda)}(X_i,x)\Big\}\cdot \ell_i^{(\lambda)}(y;x)}_{=:\mathfrak{A}_n^{(\lambda)}(y;x)} \\ &\qquad+ \underbrace{\frac{1}{n}\sum_{i=1}^n \Big(w_{\nu^*}^{(\lambda)}(X_i,x)\cdot \ell_i^{(\lambda)}(y;x) - \mathbb{E}_{\nu^*}\Big[w_{\nu^*}^{(\lambda)}(X_i,x)\cdot \ell_i^{(\lambda)}(y;x)\Big]\Big)}_{=:\mathfrak{B}_n^{(\lambda)}(y;x)} \end{aligned} \tag{37}$$

where $\ell_i^{(\lambda)}(y;x) := d^2\big(Y_i,y\big) - d^2\big(Y_i,\varphi_{\nu^*}^{(\lambda)}(x)\big)$.

Next, we analyze the asymptotic behavior of the two terms, $\mathfrak{A}_n^{(\lambda)}(y;x)$ and $\mathfrak{B}_n^{(\lambda)}(y;x)$. Specifically, we establish upper bounds on their magnitudes that hold uniformly over a δ-neighborhood of $\varphi^{(\lambda)}(x) = \varphi_{\nu^*}^{(\lambda)}(x)$, which will be used later in Step 3 of this proof.

- Firstly, we observe that for any $\delta > 0$,

$$\begin{aligned}
&\sup_{y \in B_d\left(\varphi_{\nu^*}^{(\lambda)}(x);\, \delta\right)} \left|\mathfrak{A}_n^{(\lambda)}(y;x)\right| \\
&\leq \frac{1}{n}\sum_{i=1}^n \left|w_{\mathcal{D}_n}^{(\lambda)}(X_i,x) - w_{\nu^*}^{(\lambda)}(X_i,x)\right| \cdot \sup_{y \in B_d\left(\varphi_{\nu^*}^{(\lambda)}(x);\, \delta\right)} \left|d^2(Y_i,y) - d^2(Y_i,\varphi_{\nu^*}^{(\lambda)}(x))\right| \\
&\leq 2\,\mathrm{diam}(\mathcal{M}) \cdot \left\{\frac{1}{n}\sum_{i=1}^n \left\{|V_n^{(\lambda)}(x)| + \|X_i\|_2\, \|W_n^{(\lambda)}(x)\|_2\right\}\right\} \\
&\qquad \times \sup_{y \in B_d\left(\varphi_{\nu^*}^{(\lambda)}(x);\, \delta\right)} d\big(y, \varphi_{\nu^*}^{(\lambda)}(x)\big) \\
&= O_P\left(\delta \cdot n^{-1/2}\right), \qquad (38)
\end{aligned}$$

where we used the property of $V_n^{(\lambda)}(x)$ and $W_n^{(\lambda)}(x)$ discussed in the paragraph following (34). Since the stochastic magnitudes of $V_n^{(\lambda)}(x)$ and $W_n^{(\lambda)}(x)$ are independent of δ, (38) implies that there exists $C_1^{(\lambda)} = C_1^{(\lambda)}(x) > 0$ such that for any $\delta > 0$,

$$\liminf_{n\to\infty} P\left(\sup_{y \in \mathcal{M}}\left\{|\mathfrak{A}_n^{(\lambda)}(y;x)| : d\big(y,\varphi_{\nu^*}^{(\lambda)}(x)\big) < \delta\right\} \leq C_1^{(\lambda)} \cdot \delta \cdot n^{-1/2}\right) = 1. \qquad (39)$$

Furthermore, for any $\gamma, \delta \in \mathbb{R}_+$ such that $0 \leq \gamma < \delta$, let $\mathfrak{E}_n^{(\lambda)}(\gamma,\delta;x)$ be defined as an event such that

$$\mathfrak{E}_n(\gamma,\delta;x) = \left(\sup_{y \in \mathcal{M}}\left\{|\mathfrak{A}_n^{(\lambda)}(y;x)| : d\big(y,\varphi_{\nu^*}^{(\lambda)}(x)\big) \in [\gamma,\delta)\right\} \leq C_1^{(\lambda)} \cdot \delta \cdot n^{-1/2}\right). \qquad (40)$$

For any $\gamma \in [0,\delta]$, we have $\mathfrak{E}_n(0,\delta;x) \subseteq \mathfrak{E}_n(\gamma,\delta;x)$, and thus, $\liminf_{n\to\infty} P\big(\mathfrak{E}_n(\gamma,\delta;x)\big) = 1$.

- Next, we note that

$$\left|w_{\nu^*}^{(\lambda)}(X_i,x) \cdot \ell_i^{(\lambda)}(y;x)\right| \leq 2\,\mathrm{diam}(\mathcal{M}) \cdot d\big(y,\varphi_{\nu^*}^{(\lambda)}(x)\big) \cdot \left|w_{\nu^*}^{(\lambda)}(X_i,x)\right|.$$

Observe that $d\big(y,\varphi_{\nu^*}^{(\lambda)}(x)\big) \leq \mathrm{diam}\,(\mathcal{M}) < \infty$ and recall that $\mathbb{E}_{\nu^*}\left[w_{\nu^*}^{(\lambda)}(X,x)^2\right] \leq 2\big\{1 + p\,\|x-\mu\|_\Sigma^2\big\}$ as shown in Step 1 of this proof, cf. (35). It follows from the uniform entropy condition (C2), Theorem 2.7.11, and Theorem 2.14.2 in [55] that there exists $D_{\mathsf{e}} = D_{\mathsf{e}}(x) > 0$ such that for all $\delta \in [0, D_{\mathsf{e}})$,

$$\begin{aligned}
&\mathbb{E}\left[\sup_{y\in\mathcal{M}}\left\{\left|\mathfrak{B}_n^{(\lambda)}(y;x)\right| : d\big(y,\varphi_{\nu^*}^{(\lambda)}(x)\big) < \delta\right\}\right] \\
&\qquad \leq 2\,\mathrm{diam}(\mathcal{M}) \cdot \delta \cdot n^{-1/2}\sqrt{1 + p\,\|x-\mu\|_\Sigma^2}\int_0^1 \sqrt{1 + \log \mathfrak{N}\big(B_d(\varphi^{(\lambda)}(x);\,\delta), \delta\epsilon\big)}\,\mathrm{d}\epsilon \\
&\qquad \leq C_2^{(\lambda)} \cdot \delta \cdot n^{-1/2}
\end{aligned} \qquad (41)$$

where $C_2^{(\lambda)} = 2\,(C_{\mathsf{e}}+1) \cdot \mathrm{diam}\,(\mathcal{M}) \cdot \sqrt{1 + p\,\|x-\mu\|_\Sigma^2}$ is independent of $\delta > 0$ and $n \geq 1$.

Step 3: Concluding the proof. Lastly, we combine the results from Steps 1-2 to show that, for any $\eta > 0$, there exist $K = K(\eta) > 0$ and $N = N(\eta) \geq 1$ such that $P\Big(d\big(\varphi_{\mathcal{D}_n}^{(\lambda)}(x), \varphi_{\nu^*}^{(\lambda)}(x)\big) > 2^K\, n^{-\beta}\Big) < \eta$ for any $n \geq N$, where $\beta > 0$ is an absolute constant that will be determined later in this proof. We prove this claim using the peeling technique, in a similar manner as we did in the proof of Lemma 1. To avoid cluttered notation, we let $\Delta(x) = d\big(\varphi_{\mathcal{D}_n}^{(\lambda)}(x), \varphi_{\nu^*}^{(\lambda)}(x)\big)$ in the rest of this proof.

For any fixed $K \in \mathbb{N}$ and a sufficiently large $n = n(K) \geq 1$ satisfying $2^K n^{-\beta} < D_* := D_{\mathsf{g}} \wedge D_{\mathsf{e}}$, we observe that

$$P\Big(\Delta(x) > 2^K\, n^{-\beta}\Big) = P\Big(\Delta(x) \geq D_*\Big) + P\Big(2^K\, n^{-\beta} \leq \Delta(x) < D_*\Big) \tag{42}$$

where we used $P(A) \leq P(B^c) + P(A \cap B)$ to get the inequality. As we know that $P\Big(\Delta(x) \geq D_*\Big) = o(1)$ by Step 1 of this proof, we focus on showing an upper bound for the other term, $P\big(2^K\, n^{-\beta} \leq \Delta(x) < D_*\big)$.

Step 3-A: Decomposition of $P\big(2^K\, n^{-\beta} \leq \Delta(x) < D_*\big)$. For each $n, k \in \mathbb{N}$, we define

$$\begin{aligned}
\mathfrak{F}_{n,k} &= \bigcap_{k'=K}^{k} \mathfrak{E}_n^{(\lambda)}\big(2^{k'} n^{-\beta},\, 2^{k'+1} n^{-\beta} \wedge D_*; x\big), \\
\mathfrak{G}_{n,k} &= \Bigg(\bigcap_{k'=K}^{k-1} \mathfrak{E}_n^{(\lambda)}\big(2^{k'} n^{-\beta},\, 2^{k'+1} n^{-\beta} \wedge D_*; x\big)\Bigg) \cap \mathfrak{E}_n^{(\lambda)}\big(2^{k} n^{-\beta},\, 2^{k+1} n^{-\beta} \wedge D_*; x\big)^c,
\end{aligned} \tag{43}$$

where we set $\mathfrak{F}_{n,K-1}$ to be the entire event space so that $\mathfrak{G}_{n,K} = \big(\mathfrak{F}_{n,K}\big)^c$. It is worth mentioning that $\mathfrak{G}_{n,k}$ and $\mathfrak{G}_{n,k'}$ are mutually exclusive for any $k \neq k' \geq K$, and we will use this property when concluding the proof in Step 3-C below.

Now, we observe that

$$\begin{aligned}
&P\Big(2^K\, n^{-\beta} \leq \Delta(x) < D_*\Big) \\
&\quad \leq P\Big(\mathfrak{E}_n^{(\lambda)}\big(2^K n^{-\beta},\, 2^{K+1} n^{-\beta} \wedge D_*; x\big)^c\Big) \\
&\qquad + P\bigg(\Big(2^K\, n^{-\beta} \leq \Delta(x) < D_*\Big) \cap \mathfrak{E}_n^{(\lambda)}\big(2^{k'} n^{-\beta},\, 2^{k'+1} n^{-\beta} \wedge D_*; x\big)\bigg) \\
&\quad = P\big(\mathfrak{G}_{n,K}\big) + P\bigg(\Big(2^K\, n^{-\beta} \leq \Delta(x) < D_*\Big) \cap \mathfrak{F}_{n,K}\bigg) \\
&\quad = P\big(\mathfrak{G}_{n,K}\big) + P\bigg(\Big(2^K\, n^{-\beta} \leq \Delta(x) < 2^{K+1}\, n^{-\beta} \wedge D_*\Big) \cap \mathfrak{F}_{n,K}\bigg) \\
&\qquad + P\bigg(\Big(2^{K+1}\, n^{-\beta} \leq \Delta(x) < D_*\Big) \cap \mathfrak{F}_{n,K}\bigg)
\end{aligned}$$

and that for every $k \geq K$,

$$P\bigg(\Big(2^{k+1}\, n^{-\beta} \leq \Delta(x) < D_*\Big) \cap \mathfrak{F}_{n,k}\bigg) \leq P\bigg(\Big(2^{k+1}\, n^{-\beta} \leq \Delta(x) < D_*\Big) \cap \mathfrak{F}_{n,k+1}\bigg) + P\big(\mathfrak{G}_{n,k+1}\big).$$

As a result, we have

$$P\Big(2^K\, n^{-\beta} \leq \Delta(x) < D_*\Big) = \sum_{k=K}^{\infty} P\big(\mathfrak{G}_{n,k}\big) + \sum_{k=K}^{\infty} \underbrace{P\bigg(\Big(2^k\, n^{-\beta} \leq \Delta(x) < 2^{k+1}\, n^{-\beta} \wedge D_*\Big) \cap \mathfrak{F}_{n,k}\bigg)}_{=: \mathfrak{C}_{n,k}}. \tag{44}$$

Step 3-B: Controlling $\mathfrak{C}_{n,k}$. Next, we show an upper bound for $\mathfrak{C}_{n,k}$. Suppose that $2^k n^{-\beta} \leq \Delta(x) < 2^{k+1}n^{-\beta} \wedge D_*$ and the event $\mathfrak{F}_{n,k}$ occurs. Then it follows from Assumption (C1) that

$$\begin{aligned}
&C_{\mathsf{g}} \cdot \Delta(x)^\alpha \\
&\quad\leq R_{\nu^*}^{(\lambda)}\big(\varphi_{\mathcal{D}_n}^{(\lambda)}(x); x\big) - R_{\nu^*}^{(\lambda)}\big(\varphi_{\nu^*}^{(\lambda)}(x); x\big) \\
&\quad\leq \Big\{R_{\nu^*}^{(\lambda)}\big(\varphi_{\mathcal{D}_n}^{(\lambda)}(x); x\big) - R_{\nu^*}^{(\lambda)}\big(\varphi_{\nu^*}^{(\lambda)}(x); x\big)\Big\} + \underbrace{\Big\{R_{\mathcal{D}_n}^{(\lambda)}\big(\varphi_{\nu^*}^{(\lambda)}(x); x\big) - R_{\mathcal{D}_n}^{(\lambda)}\big(\varphi_{\mathcal{D}_n}^{(\lambda)}(x); x\big)\Big\}}_{\geq 0} \\
&\quad= Z_n^{(\lambda)}\big(\varphi_{\nu^*}^{(\lambda)}; x\big) - Z_n^{(\lambda)}\big(\varphi_{\mathcal{D}_n}^{(\lambda)}(x); x\big) && \text{cf. (36)} \\
&\quad\leq \Big|\mathfrak{A}_n^{(\lambda)}\big(\varphi_{\nu^*}^{(\lambda)}(x); x\big)\Big| + \Big|\mathfrak{B}_n^{(\lambda)}\big(\varphi_{\nu^*}^{(\lambda)}(x); x\big)\Big| && \because (37) \\
&\quad\leq \sup_{y \in \mathcal{M}} \Big\{\Big|\mathfrak{A}_n^{(\lambda)}\big(y; x\big)\Big| + \Big|\mathfrak{B}_n^{(\lambda)}\big(y; x\big)\Big| : 2^k\, n^{-\beta} \leq d\big(y, \varphi_{\nu^*}^{(\lambda)}(x)\big) < 2^{k+1}\, n^{-\beta} \wedge D_*\big)\Big\} \\
&\quad\leq C_1^{(\lambda)} \cdot \big(2^{k+1}n^{-\beta} \wedge D_*\big) \cdot n^{-1/2} + \sup_{y \in \mathcal{M}} \Big\{\big|\mathfrak{B}_n^{(\lambda)}(y; x)\big| : d\big(y, \varphi_{\nu^*}^{(\lambda)}(x)\big) < 2^{k+1}n^{-\beta} \wedge D_*\Big\}. && \because (40)
\end{aligned} \tag{45}$$

Therefore, we obtain that for each $k \geq K$,

$$\begin{aligned}
\mathfrak{C}_{n,k} &= P\bigg(\Big(2^k\, n^{-\beta} \leq \Delta(x) < 2^{k+1}\, n^{-\beta} \wedge D_*\Big) \cap \mathfrak{F}_{n,k}\bigg) \\
&\leq P\bigg(\Big(\Delta(x)^\alpha \geq \big(2^k\, n^{-\beta}\big)^\alpha\Big) \cap \mathfrak{F}_{n,k}\bigg) \\
&\leq \frac{C_1^{(\lambda)} \cdot \big(2^{k+1}n^{-\beta} \wedge D_*\big) \cdot n^{-1/2} + \mathbb{E}\Big[\sup_{y \in \mathcal{M}} \Big\{\big|\mathfrak{B}_n^{(\lambda)}(y; x)\big| : d\big(y, \varphi_{\nu^*}^{(\lambda)}(x)\big) < 2^{k+1}n^{-\beta} \wedge D_*\Big\}\Big]}{C_{\mathsf{g}} \cdot \big(2^k\, n^{-\beta}\big)^\alpha} \\
&\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\because \text{(45) \& Markov's inequality} \\
&\leq \frac{\big(C_1^{(\lambda)} + C_2^{(\lambda)}\big) \cdot \big(2^{k+1}n^{-\beta} \wedge D_*\big) \cdot n^{-1/2}}{C_{\mathsf{g}} \cdot \big(2^k\, n^{-\beta}\big)^\alpha} \qquad \because (41)
\end{aligned} \tag{46}$$

Step 3-C: Concluding Step 3. Combining (42), (44), and (46), we have

$$\begin{aligned}
P\Big(\Delta(x) > 2^K\, n^{-\beta}\Big) \leq \frac{2\big(C_1^{(\lambda)} + C_2^{(\lambda)}\big)}{C_{\mathsf{g}}}\, n^{-\frac{1}{2}+\beta(\alpha-1)} \sum_{k=K}^{\infty} 2^{-k(\alpha-1)} \\
+ \underbrace{P\Big(\Delta(x) \geq D_*\Big)}_{=o(1)\ \because \text{Step 1 of this proof}} + \sum_{k=K}^{\infty} P\Big(\mathfrak{G}_{n,k}\Big).
\end{aligned}$$

Moreover, $\mathfrak{G}_{n,k}$ are mutually exclusive, and thus,

$$\sum_{k=K}^{\infty} P\Big(\mathfrak{G}_{n,k}\Big) = P\left(\bigcup_{k=K}^{\infty} \mathfrak{G}_{n,k}\right) = P\left(\left(\bigcup_{k=K}^{\infty} \mathfrak{E}_n^{(\lambda)}\big(2^k n^{-\beta},\, 2^{k+1}n^{-\beta} \wedge D_*; x\big)\right)^c\right) \to 0 \quad \because (40)$$

Finally, we obtain the desired result by letting $\beta = \frac{1}{2(\alpha-1)}$.

$\square$

D Proof of Theorem 3

In this section, we prove Theorem 4 that establishes an upper bound on $d\big(\varphi_{\widetilde{\mathcal{D}}_n}^{(\lambda)}(x),\, \varphi_{\mathcal{D}_n}^{(\lambda)}(x)\big)$. This section is organized as follows. Firstly, in Section D.1, we present several useful results from matrix perturbation theory as lemmas. Next, in Section D.2, we provide a key lemma (Lemma 6) that establishes the stability of the weight function when there is covariate noise. Lastly, in Section D.3, we state and prove Theorem 4, from which Theorem 3 can be easily derived.

D.1 Useful lemmas

Definition 7. *Let $n, p \in \mathbb{N}$ and let $\boldsymbol{M} \in \mathbb{R}^{n \times p}$. The row projection matrix for $\boldsymbol{M}$, denoted by $\Pi_{\boldsymbol{M}}^{\mathrm{row}} \in \mathbb{R}^{p \times p}$, is a matrix such that*

$$\Pi_{\boldsymbol{M}}^{\mathrm{row}} := \boldsymbol{M}^{\dagger} \cdot \boldsymbol{M}. \tag{47}$$

and the column projection matrix for $\boldsymbol{M}$, denoted by $\Pi_{\boldsymbol{M}}^{\mathrm{col}} \in \mathbb{R}^{n \times n}$, is a matrix such that

$$\Pi_{\boldsymbol{M}}^{\mathrm{col}} := \boldsymbol{M} \cdot \boldsymbol{M}^{\dagger}. \tag{48}$$

We recall from (6) that for any $\lambda \in \mathbb{R}_+$, the singular value thresholding (SVT) operator $\mathrm{SVT}^{(\lambda)}$ is defined such that

$$\boldsymbol{M} = \sum_{i=1}^{\min\{n,p\}} s_i \cdot u_i v_i^{\top} \text{ is a SVD} \qquad \mapsto \qquad \mathrm{SVT}^{(\lambda)}(\boldsymbol{M}) = \sum_{i=1}^{\min\{n,p\}} s_i \cdot \mathbb{1}\{s_i > \lambda\} \cdot u_i v_i^{\top}.$$

In the rest of this section, we let $\boldsymbol{M}^{(\lambda)} := \mathrm{SVT}^{(\lambda)}(\boldsymbol{M})$ for shorthand.

Lemma 3 (Properties of the row/column projection matrices)**.** *Let $n, p \in \mathbb{N}$, and $\boldsymbol{M} \in \mathbb{R}^{n \times p}$. For any $\lambda \in \mathbb{R}_+$, the following statements are true.*

1. *$\Pi_{\boldsymbol{M}^{(\lambda)}}^{\mathrm{row}}$ defines a projection in $\mathbb{R}^p$ and $\operatorname{rank} \Pi_{\boldsymbol{M}^{(\lambda)}}^{\mathrm{row}} = \operatorname{rank} \boldsymbol{M}^{(\lambda)}$.*
2. *$\Pi_{\boldsymbol{M}^{(\lambda)}}^{\mathrm{col}}$ defines a projection in $\mathbb{R}^n$ and $\operatorname{rank} \Pi_{\boldsymbol{M}^{(\lambda)}}^{\mathrm{col}} = \operatorname{rank} \boldsymbol{M}^{(\lambda)}$.*
3. *$\boldsymbol{M} \Pi_{\boldsymbol{M}^{(\lambda)}}^{\mathrm{row}} \boldsymbol{M}^{\dagger} = \Pi_{\boldsymbol{M}^{(\lambda)}}^{\mathrm{col}}$ and $\boldsymbol{M}^{\dagger} \Pi_{\boldsymbol{M}^{(\lambda)}}^{\mathrm{col}} \boldsymbol{M} = \Pi_{\boldsymbol{M}^{(\lambda)}}^{\mathrm{row}}$.*

Proof. Let $r = \operatorname{rank} \boldsymbol{M}$ and consider a compact singular value decomposition (SVD) of $\boldsymbol{M}$:

$$\boldsymbol{M} = \sum_{i=1}^{r} s_i \cdot u_i v_i^{\top}$$

where $s_1, \ldots, s_r$ are non-zero singular values of $\boldsymbol{M}$. Noticing that

$$\boldsymbol{M}^{(\lambda)} = \mathrm{SVT}^{(\lambda)}(\boldsymbol{M}) = \sum_{i=1}^{r} \mathbb{1}\{s_i > \lambda\} \cdot u_i v_i^{\top}$$

and that $\boldsymbol{M}^{\dagger} = \sum_{i=1}^{r} s_i^{-1} \cdot v_i u_i^{\top}$, the three conclusions of the lemma follow straightforwardly from the orthonormality of singular vectors.

- $\Pi_{\boldsymbol{M}^{(\lambda)}}^{\mathrm{row}} = \sum_{i=1}^{r} v_i v_i^{\top} \cdot \mathbb{1}\{s_i > \lambda\}$ is the projection onto the row space of $\boldsymbol{M}^{(\lambda)}$.
- $\Pi_{\boldsymbol{M}^{(\lambda)}}^{\mathrm{col}} = \sum_{i=1}^{r} u_i u_i^{\top} \cdot \mathbb{1}\{s_i > \lambda\}$ is the projection onto the column space of $\boldsymbol{M}^{(\lambda)}$.
- Due to the orthonormality of singular vectors,

$$\begin{aligned} \boldsymbol{M} \Pi_{\boldsymbol{M}^{(\lambda)}}^{\mathrm{row}} \boldsymbol{M}^{\dagger} &= \left(\sum_{i=1}^{r} s_i \cdot u_i v_i^{\top} \right) \left(\sum_{i=1}^{r} v_i v_i^{\top} \cdot \mathbb{1}\{s_i > \lambda\} \right) \left(\sum_{i=1}^{r} s_i^{-1} \cdot v_i u_i^{\top} \right) \\ &= \sum_{i=1}^{r} u_i u_i^{\top} \cdot \mathbb{1}\{s_i > \lambda\} \\ &= \Pi_{\boldsymbol{M}^{(\lambda)}}^{\mathrm{col}}, \end{aligned}$$

  and likewise, $\boldsymbol{M}^{\dagger} \Pi_{\boldsymbol{M}^{(\lambda)}}^{\mathrm{col}} \boldsymbol{M} = \Pi_{\boldsymbol{M}^{(\lambda)}}^{\mathrm{row}}$.

$\square$

In addition, we collect two classical results from matrix perturbation theory and state them as lemmas.

Lemma 4 ([51, Theorem 3.2])**.** *Let $\boldsymbol{X}, \boldsymbol{Z} \in \mathbb{R}^{n\times p}$. Then the following equation is true:*

$$\boldsymbol{Z}^\dagger - \boldsymbol{X}^\dagger = -\boldsymbol{Z}^\dagger \Pi^{\rm col}_{\boldsymbol{Z}}(\boldsymbol{Z}-\boldsymbol{X})\Pi^{\rm row}_{\boldsymbol{X}}\boldsymbol{X}^\dagger + \boldsymbol{Z}^\dagger \Pi^{\rm col}_{\boldsymbol{Z}}\Pi^{{\rm col}\perp}_{\boldsymbol{X}} - \Pi^{{\rm row}\perp}_{\boldsymbol{Z}}\Pi^{\rm row}_{\boldsymbol{X}}\boldsymbol{X}^\dagger \tag{49}$$

where $\Pi^{{\rm col}\perp}_{\boldsymbol{X}} = \boldsymbol{I}_n - \Pi^{\rm col}_{\boldsymbol{X}}$ and $\Pi^{{\rm row}\perp}_{\boldsymbol{Z}} = \boldsymbol{I}_p - \Pi^{\rm row}_{\boldsymbol{Z}}$.

Lemma 5 ([15, Theorems 2.4 & 2.5])**.** *Let $\boldsymbol{X}, \boldsymbol{Z} \in \mathbb{R}^{n\times p}$. Then*

$$\left\|\Pi^{\rm col}_{\boldsymbol{Z}} - \Pi^{\rm col}_{\boldsymbol{X}}\right\| \le \max\left\{\left\|(\boldsymbol{Z}-\boldsymbol{X})\boldsymbol{X}^\dagger\right\|, \left\|(\boldsymbol{Z}-\boldsymbol{X})\boldsymbol{Z}^\dagger\right\|\right\}. \tag{50}$$

Moreover, if $\operatorname{rank}\boldsymbol{X} = \operatorname{rank}\boldsymbol{Z}$, then

$$\left\|\Pi^{\rm col}_{\boldsymbol{Z}} - \Pi^{\rm col}_{\boldsymbol{X}}\right\| \le \min\left\{\left\|(\boldsymbol{Z}-\boldsymbol{X})\boldsymbol{X}^\dagger\right\|, \left\|(\boldsymbol{Z}-\boldsymbol{X})\boldsymbol{Z}^\dagger\right\|\right\}. \tag{51}$$

D.2 Stability of the weights under (small) perturbation in covariates

Let $\mathcal{D}_n = \{(x_i, y_i) \in \mathbb{R}^p \times \mathcal{M} : i \in [n]\}$ and $\widetilde{\mathcal{D}}_n = \{(z_i, y_i) \in \mathbb{R}^p \times \mathcal{M} : i \in [n]\}$ be two sets in $\mathbb{R}^p \times \mathcal{M}$. We may identify these sets with their empirical distributions. Recall the definition of $w^{(\lambda)}_\nu$ from (9): for any probability measure ν on $\mathbb{R}^p \times \mathcal{M}$, any $\lambda \in \mathbb{R}_+$, and any $x, x' \in \mathbb{R}^p$,

$$w^{(\lambda)}_\nu(x', x) = 1 + (x' - \mu_\nu)^\top \left[\mathrm{SVT}^{(\lambda)}(\Sigma_\nu)\right]^\dagger (x - \mu_\nu)$$

where $\mu_\nu = \mathbb{E}_{(X,Y)\sim\nu}(X)$ and $\Sigma_\nu = \mathrm{Var}_{(X,Y)\sim\nu}(X)$, cf. (7). We define the <u>weight vectors</u> induced by $\mathcal{D}_n$ and $\widetilde{\mathcal{D}}_n$ as follows: for any $\lambda \in \mathbb{R}_+$ and any $x \in \mathbb{R}^p$,

$$\begin{aligned}
\vec{w}^{(\lambda)}_{\mathcal{D}_n}(x) &:= \begin{bmatrix} w^{(\lambda)}_{\mathcal{D}_n}(x_1, x) & \cdots & w^{(\lambda)}_{\mathcal{D}_n}(x_n, x)\end{bmatrix} \in \mathbb{R}^n,\\
\vec{w}^{(\lambda)}_{\widetilde{\mathcal{D}}_n}(x) &:= \begin{bmatrix} w^{(\lambda)}_{\widetilde{\mathcal{D}}_n}(z_1, x) & \cdots & w^{(\lambda)}_{\widetilde{\mathcal{D}}_n}(z_n, x)\end{bmatrix} \in \mathbb{R}^n.
\end{aligned} \tag{52}$$

Lemma 6 (Stability of weights)**.** *Let $\mathcal{D}_n = \{(x_i, y_i) \in \mathbb{R}^p \times \mathcal{M} : i \in [n]\}$ and $\widetilde{\mathcal{D}}_n = \{(z_i, y_i) \in \mathbb{R}^p \times \mathcal{M} : i \in [n]\}$. Let $\boldsymbol{X} = \begin{bmatrix} x_1 & \cdots & x_n\end{bmatrix}^\top \in \mathbb{R}^{n\times p}$ and $\boldsymbol{Z} = \begin{bmatrix} z_1 & \cdots & z_n\end{bmatrix}^\top \in \mathbb{R}^{n\times p}$. For any $\lambda \in \mathbb{R}_+$, if $x \in \mathbb{R}^p$ satisfies $x - \mu_{\mathcal{D}_n} \in \operatorname{rowsp}(\boldsymbol{X}_{\rm ctr})$, then*

$$\left\|\vec{w}^{(\lambda)}_{\widetilde{\mathcal{D}}_n}(x) - \vec{w}^{(\lambda)}_{\mathcal{D}_n}(x)\right\| \le \frac{\sqrt{n}\cdot\|\boldsymbol{Z}-\boldsymbol{X}\|}{\min\left\{\sigma^{(\lambda)}(\boldsymbol{X}_{\rm ctr}),\ \sigma^{(\lambda)}(\boldsymbol{Z}_{\rm ctr})\right\}} \cdot \left(2\cdot\left\|x - \mu_{\mathcal{D}_n}\right\|_{\Sigma_{\mathcal{D}_n}} + 1\right) \tag{53}$$

where $\boldsymbol{X}_{\rm ctr} = \left(\boldsymbol{I}_n - \frac{1}{n}\mathbf{1}_n\mathbf{1}_n^\top\right)\boldsymbol{X}$ and $\sigma^{(\lambda)}(\boldsymbol{X}) := \inf\{\sigma_i(\boldsymbol{X}) > \lambda : i \in \mathbb{N}\}$ (likewise for $\boldsymbol{Z}$).

Proof of Lemma 6. This proof consists of three steps. In Step 1, we express the weight discrepancy $\vec{w}^{(\lambda)}_{\widetilde{\mathcal{D}}_n}(x) - \vec{w}^{(\lambda)}_{\mathcal{D}_n}(x)$ as a sum of matrix products using projections. In Step 2, we establish upper bounds on the norm of the expression obtained in Step 1. In Step 3, we collect intermediate results together and conclude the proof.

Step 1: Decomposition of the weight discrepancy. First of all, we rewrite $\vec{w}^{(\lambda)}_{\widetilde{\mathcal{D}}_n}(x) - \vec{w}^{(\lambda)}_{\mathcal{D}_n}(x)$ in a compact matrix representation that is presented in (61) at the end of this step. To this end, we begin by observing that

$$\mu_{\mathcal{D}_n} = \frac{1}{n}\boldsymbol{X}^\top\mathbf{1}_n, \qquad \text{and} \qquad \Sigma_{\mathcal{D}_n} = \frac{1}{n}\left(\boldsymbol{X} - \mathbf{1}_n\mu^\top_{\mathcal{D}_n}\right)^\top\left(\boldsymbol{X} - \mathbf{1}_n\mu^\top_{\mathcal{D}_n}\right) = \frac{1}{n}\boldsymbol{X}^\top_{\rm ctr}\boldsymbol{X}_{\rm ctr}. \tag{54}$$

For given $\lambda \in \mathbb{R}_+$, we let $\boldsymbol{X}^{(\lambda)}_{\rm ctr} := \mathrm{SVT}^{(\lambda)}(\boldsymbol{X}_{\rm ctr})$, and observe that

$$\Sigma^{(\lambda)}_{\mathcal{D}_n} = \Pi^{\rm row}_{\boldsymbol{X}^{(\lambda)}_{\rm ctr}} \cdot \left(\frac{1}{n}\boldsymbol{X}^\top_{\rm ctr}\boldsymbol{X}_{\rm ctr}\right)\cdot \Pi^{\rm row}_{\boldsymbol{X}^{(\lambda)}_{\rm ctr}} = \frac{1}{n}\cdot {\boldsymbol{X}^{(\lambda)}_{\rm ctr}}^\top \cdot \boldsymbol{X}^{(\lambda)}_{\rm ctr}. \tag{55}$$

Then it follows that

$$\left[\Sigma^{(\lambda)}_{\mathcal{D}_n}\right]^\dagger = n\cdot\left[{\boldsymbol{X}^{(\lambda)}_{\rm ctr}}^\top\cdot\boldsymbol{X}^{(\lambda)}_{\rm ctr}\right]^\dagger = n\cdot\left[\boldsymbol{X}^{(\lambda)}_{\rm ctr}\right]^\dagger\cdot\left[{\boldsymbol{X}^{(\lambda)}_{\rm ctr}}^\top\right]^\dagger = n\cdot\Pi^{\rm row}_{\boldsymbol{X}^{(\lambda)}_{\rm ctr}}\cdot\boldsymbol{X}^\dagger_{\rm ctr}\cdot\left(\boldsymbol{X}^\top_{\rm ctr}\right)^\dagger\cdot\Pi^{\rm row}_{\boldsymbol{X}^{(\lambda)}_{\rm ctr}}.$$

Therefore, we have

$$\begin{aligned}\vec{w}^{(\lambda)}_{\mathcal{D}_n}(x) &= \mathbf{1}_n + \left(\boldsymbol{X} - \mathbf{1}_n \mu_{\mathcal{D}_n}^\top\right) \cdot \left[\Sigma^{(\lambda)}_{\mathcal{D}_n}\right]^\dagger \cdot (x - \mu_{\mathcal{D}_n}) \\ &= \mathbf{1}_n + n \cdot \boldsymbol{X}_{\mathrm{ctr}} \cdot \Pi^{\mathrm{row}}_{\boldsymbol{X}^{(\lambda)}_{\mathrm{ctr}}} \cdot \boldsymbol{X}^\dagger_{\mathrm{ctr}} \cdot \left(\boldsymbol{X}^\top_{\mathrm{ctr}}\right)^\dagger \cdot \Pi^{\mathrm{row}}_{\boldsymbol{X}^{(\lambda)}_{\mathrm{ctr}}} \cdot (x - \mu_{\mathcal{D}_n}) \\ &= \mathbf{1}_n + n \cdot \Pi^{\mathrm{col}}_{\boldsymbol{X}^{(\lambda)}_{\mathrm{ctr}}} \cdot \left(\boldsymbol{X}^\top_{\mathrm{ctr}}\right)^\dagger \cdot \Pi^{\mathrm{row}}_{\boldsymbol{X}^{(\lambda)}_{\mathrm{ctr}}} \cdot (x - \mu_{\mathcal{D}_n}),\end{aligned} \tag{56}$$

where the equality in the last line follows from Lemma 3: $\boldsymbol{X}_{\mathrm{ctr}} \Pi^{\mathrm{row}}_{\boldsymbol{X}^{(\lambda)}_{\mathrm{ctr}}} \boldsymbol{X}^\dagger_{\mathrm{ctr}} = \Pi^{\mathrm{col}}_{\boldsymbol{X}^{(\lambda)}_{\mathrm{ctr}}}$.

Likewise, we repeat the above for $\widetilde{\mathcal{D}}_n$ and $\boldsymbol{Z}$ to write

$$\mu_{\widetilde{\mathcal{D}}_n} = \frac{1}{n} \boldsymbol{Z}^\top \mathbf{1}_n \qquad \text{and} \qquad \Sigma_{\widetilde{\mathcal{D}}_n} = \frac{1}{n} \boldsymbol{Z}^\top_{\mathrm{ctr}} \boldsymbol{Z}_{\mathrm{ctr}}.$$

Then, we obtain an expression for $\vec{w}^{(\lambda)}_{\widetilde{\mathcal{D}}_n}(x)$ in a similar form to (56), namely,

$$\vec{w}^{(\lambda)}_{\widetilde{\mathcal{D}}_n}(x) = \mathbf{1}_n + n \cdot \Pi^{\mathrm{col}}_{\boldsymbol{Z}^{(\lambda)}_{\mathrm{ctr}}} \cdot \left(\boldsymbol{Z}^\top_{\mathrm{ctr}}\right)^\dagger \cdot \Pi^{\mathrm{row}}_{\boldsymbol{Z}^{(\lambda)}_{\mathrm{ctr}}} \cdot (x - \mu_{\widetilde{\mathcal{D}}_n}). \tag{57}$$

Thereafter, we define $c_x, \tilde{c}_x \in \mathbb{R}^{n \times 1}$ so that

$$\begin{aligned} c_x &= \|x - \mu_{\mathcal{D}_n}\|_{\Sigma_{\mathcal{D}_n}} = \left(\frac{1}{\sqrt{n}} \boldsymbol{X}^\top_{\mathrm{ctr}}\right)^\dagger \cdot (x - \mu_{\mathcal{D}_n}) \qquad \text{and} \\ \tilde{c}_x &= \|x - \mu_{\mathcal{D}_n}\|_{\Sigma_{\widetilde{\mathcal{D}}_n}} = \left(\frac{1}{\sqrt{n}} \boldsymbol{Z}^\top_{\mathrm{ctr}}\right)^\dagger \cdot \left(x - \mu_{\widetilde{\mathcal{D}}_n}\right). \end{aligned} \tag{58}$$

Then we observe that for any $x \in \mathbb{R}^p$,

$$n \cdot \Pi^{\mathrm{row}}_{\boldsymbol{X}^{(\lambda)}_{\mathrm{ctr}}} \cdot (x - \mu_{\mathcal{D}_n}) = n \cdot \Pi^{\mathrm{row}}_{\boldsymbol{X}^{(\lambda)}_{\mathrm{ctr}}} \cdot \frac{1}{\sqrt{n}} \boldsymbol{X}^\top_{\mathrm{ctr}} \cdot \left(\frac{1}{\sqrt{n}} \boldsymbol{X}^\top_{\mathrm{ctr}}\right)^\dagger \cdot (x - \mu_{\mathcal{D}_n}) = \sqrt{n} \cdot \Pi^{\mathrm{row}}_{\boldsymbol{X}^{(\lambda)}_{\mathrm{ctr}}} \cdot \boldsymbol{X}^\top_{\mathrm{ctr}} \cdot c_x. \tag{59}$$

Likewise,

$$n \cdot \Pi^{\mathrm{row}}_{\boldsymbol{Z}^{(\lambda)}_{\mathrm{ctr}}} \cdot \left(x - \mu_{\widetilde{\mathcal{D}}_n}\right) = n \cdot \Pi^{\mathrm{row}}_{\boldsymbol{Z}^{(\lambda)}_{\mathrm{ctr}}} \cdot \frac{1}{\sqrt{n}} \boldsymbol{Z}^\top_{\mathrm{ctr}} \cdot \left(\frac{1}{\sqrt{n}} \boldsymbol{Z}^\top_{\mathrm{ctr}}\right)^\dagger \cdot (x - \mu_{\widetilde{\mathcal{D}}_n}) = \sqrt{n} \cdot \Pi^{\mathrm{row}}_{\boldsymbol{Z}^{(\lambda)}_{\mathrm{ctr}}} \cdot \boldsymbol{Z}^\top_{\mathrm{ctr}} \cdot \tilde{c}_x. \tag{60}$$

Consequently, for any $x \in \mathbb{R}^p$, we obtain from (56) and (57) with aid of (59) and (60) that

$$\begin{aligned} &\vec{w}^{(\lambda)}_{\widetilde{\mathcal{D}}_n}(x) - \vec{w}^{(\lambda)}_{\mathcal{D}_n}(x) \\ &\quad = \sqrt{n} \cdot \Pi^{\mathrm{col}}_{\boldsymbol{Z}^{(\lambda)}_{\mathrm{ctr}}} \cdot \left(\boldsymbol{Z}^\top_{\mathrm{ctr}}\right)^\dagger \cdot \Pi^{\mathrm{row}}_{\boldsymbol{Z}^{(\lambda)}_{\mathrm{ctr}}} \cdot \boldsymbol{Z}^\top_{\mathrm{ctr}} \cdot \tilde{c}_x - \sqrt{n} \cdot \Pi^{\mathrm{col}}_{\boldsymbol{X}^{(\lambda)}_{\mathrm{ctr}}} \cdot \left(\boldsymbol{X}^\top_{\mathrm{ctr}}\right)^\dagger \cdot \Pi^{\mathrm{row}}_{\boldsymbol{X}^{(\lambda)}_{\mathrm{ctr}}} \cdot \boldsymbol{X}^\top_{\mathrm{ctr}} \cdot c_x \\ &\quad = \sqrt{n} \cdot \Pi^{\mathrm{col}}_{\boldsymbol{Z}^{(\lambda)}_{\mathrm{ctr}}} \cdot \tilde{c}_x - \sqrt{n} \cdot \Pi^{\mathrm{col}}_{\boldsymbol{X}^{(\lambda)}_{\mathrm{ctr}}} \cdot c_x \qquad \because \text{Lemma 3} \\ &\quad = \sqrt{n} \cdot \Pi^{\mathrm{col}}_{\boldsymbol{Z}^{(\lambda)}_{\mathrm{ctr}}} \cdot (\tilde{c}_x - c_x) + \sqrt{n} \cdot \left(\Pi^{\mathrm{col}}_{\boldsymbol{Z}^{(\lambda)}_{\mathrm{ctr}}} - \Pi^{\mathrm{col}}_{\boldsymbol{X}^{(\lambda)}_{\mathrm{ctr}}}\right) \cdot c_x. \end{aligned} \tag{61}$$

By triangle inequality, we obtain the following upper bound:

$$\left\| \vec{w}^{(\lambda)}_{\widetilde{\mathcal{D}}_n}(x) - \vec{w}^{(\lambda)}_{\mathcal{D}_n}(x) \right\| \le \sqrt{n} \cdot \left\| \Pi^{\mathrm{col}}_{\boldsymbol{Z}^{(\lambda)}_{\mathrm{ctr}}} \cdot (\tilde{c}_x - c_x) \right\| + \sqrt{n} \cdot \left\| \left(\Pi^{\mathrm{col}}_{\boldsymbol{Z}^{(\lambda)}_{\mathrm{ctr}}} - \Pi^{\mathrm{col}}_{\boldsymbol{X}^{(\lambda)}_{\mathrm{ctr}}}\right) \cdot c_x \right\|. \tag{62}$$

Step 2: Upper bounding the norm. Next, we establish separate upper bounds for the two terms in (62).

(1) The first term in (62)**.** First of all, we observe from the definition of c_x and $\tilde{c}_x$, cf. (58), that

$$\begin{aligned} \tilde{c}_x - c_x &= \left(\frac{1}{\sqrt{n}} \boldsymbol{Z}^\top_{\mathrm{ctr}}\right)^\dagger \cdot \left(x - \mu_{\widetilde{\mathcal{D}}_n}\right) - \left(\frac{1}{\sqrt{n}} \boldsymbol{X}^\top_{\mathrm{ctr}}\right)^\dagger \cdot (x - \mu_{\mathcal{D}_n}) \\ &= \sqrt{n} \cdot \left({\boldsymbol{Z}^\top_{\mathrm{ctr}}}^\dagger - {\boldsymbol{X}^\top_{\mathrm{ctr}}}^\dagger\right) \cdot (x - \mu_{\mathcal{D}_n}) + \sqrt{n} \cdot \left[\boldsymbol{Z}^\top_{\mathrm{ctr}}\right]^\dagger \cdot \left(\mu_{\widetilde{\mathcal{D}}_n} - \mu_{\mathcal{D}_n}\right). \end{aligned}$$

Then we can upper bound the first term in (62) as follows:

$$\left\|\Pi^{\mathrm{col}}_{\boldsymbol{Z}^{(\lambda)}_{\mathrm{ctr}}}\cdot(\tilde{c}_x - c_x)\right\| = \sqrt{n}\cdot\left\|\Pi^{\mathrm{col}}_{\boldsymbol{Z}^{(\lambda)}_{\mathrm{ctr}}}\cdot\left({\boldsymbol{Z}_{\mathrm{ctr}}^\top}^\dagger - {\boldsymbol{X}_{\mathrm{ctr}}^\top}^\dagger\right)\cdot(x-\mu_{\mathcal{D}_n})\right\| + \sqrt{n}\cdot\left\|\Pi^{\mathrm{col}}_{\boldsymbol{Z}^{(\lambda)}_{\mathrm{ctr}}}\cdot\left[\boldsymbol{Z}_{\mathrm{ctr}}^\top\right]^\dagger\cdot\left(\mu_{\widetilde{\mathcal{D}}_n}-\mu_{\mathcal{D}_n}\right)\right\|. \tag{63}$$

Next, we consider the orthogonal decomposition of $x-\mu_{\mathcal{D}_n}$:

$$x-\mu_{\mathcal{D}_n} = \Pi^{\mathrm{row}}_{\boldsymbol{X}_{\mathrm{ctr}}}\big(x-\mu_{\mathcal{D}_n}\big) + {\Pi^{\mathrm{row}}_{\boldsymbol{X}_{\mathrm{ctr}}}}^{\perp}\cdot\big(x-\mu_{\mathcal{D}_n}\big) = \frac{1}{\sqrt{n}}\boldsymbol{X}_{\mathrm{ctr}}^\top\cdot c_x + {\Pi^{\mathrm{row}}_{\boldsymbol{X}_{\mathrm{ctr}}}}^{\perp}\cdot\big(x-\mu_{\mathcal{D}_n}\big). \tag{64}$$

If $x-\mu_{\mathcal{D}_n}\in\mathrm{rowsp}\left(\boldsymbol{X}_{\mathrm{ctr}}\right)$, then we obtain the following upper bound for the first term in (63):

$$\begin{aligned}
&\sqrt{n}\cdot\left\|\Pi^{\mathrm{col}}_{\boldsymbol{Z}^{(\lambda)}_{\mathrm{ctr}}}\cdot\left({\boldsymbol{Z}_{\mathrm{ctr}}^\top}^\dagger - {\boldsymbol{X}_{\mathrm{ctr}}^\top}^\dagger\right)\cdot(x-\mu_{\mathcal{D}_n})\right\| \\
&\qquad\le \left\|\Pi^{\mathrm{col}}_{\boldsymbol{Z}^{(\lambda)}_{\mathrm{ctr}}}\cdot\left({\boldsymbol{Z}_{\mathrm{ctr}}^\top}^\dagger - {\boldsymbol{X}_{\mathrm{ctr}}^\top}^\dagger\right)\cdot\boldsymbol{X}_{\mathrm{ctr}}^\top\cdot c_x\right\| \\
&\qquad\quad + \sqrt{n}\cdot\left\|\Pi^{\mathrm{col}}_{\boldsymbol{Z}^{(\lambda)}_{\mathrm{ctr}}}\cdot\left({\boldsymbol{Z}_{\mathrm{ctr}}^\top}^\dagger - {\boldsymbol{X}_{\mathrm{ctr}}^\top}^\dagger\right)\cdot\underbrace{{\Pi^{\mathrm{row}}_{\boldsymbol{X}_{\mathrm{ctr}}}}^{\perp}\cdot\big(x-\mu_{\mathcal{D}_n}\big)}_{=0}\right\| && \because (64) \\
&\qquad\le \left\|\Pi^{\mathrm{col}}_{\boldsymbol{Z}^{(\lambda)}_{\mathrm{ctr}}}\cdot\left\{-{\boldsymbol{Z}_{\mathrm{ctr}}^\top}^\dagger\cdot\Pi^{\mathrm{row}}_{\boldsymbol{Z}_{\mathrm{ctr}}}\cdot\left(\boldsymbol{Z}_{\mathrm{ctr}}^\top-\boldsymbol{X}_{\mathrm{ctr}}^\top\right)\cdot\Pi^{\mathrm{col}}_{\boldsymbol{X}_{\mathrm{ctr}}}\cdot{\boldsymbol{X}_{\mathrm{ctr}}^\top}^\dagger\right\}\cdot\boldsymbol{X}_{\mathrm{ctr}}^\top\cdot c_x\right\| && \because \text{Lemma 4} \\
&\qquad\le \left\|\left[{\boldsymbol{Z}^{(\lambda)}_{\mathrm{ctr}}}^\top\right]^\dagger\right\|\cdot\left\|\Pi^{\mathrm{row}}_{\boldsymbol{Z}_{\mathrm{ctr}}}\cdot(\boldsymbol{Z}-\boldsymbol{X})^\top\cdot\Pi^{\mathrm{row}}_{\boldsymbol{1}_n^\perp}\cdot\Pi^{\mathrm{col}}_{\boldsymbol{X}_{\mathrm{ctr}}}\right\|\cdot\|c_x\| \\
&\qquad\le \frac{\|\boldsymbol{Z}-\boldsymbol{X}\|}{\sigma^{(\lambda)}\left(\boldsymbol{Z}_{\mathrm{ctr}}\right)}\cdot\|c_x\|.
\end{aligned}$$

Similarly, the second term in (63) can be bounded by

$$\begin{aligned}
\sqrt{n}\cdot\left\|\Pi^{\mathrm{col}}_{\boldsymbol{Z}^{(\lambda)}_{\mathrm{ctr}}}\cdot\left[\boldsymbol{Z}_{\mathrm{ctr}}^\top\right]^\dagger\cdot\left(\mu_{\widetilde{\mathcal{D}}_n}-\mu_{\mathcal{D}_n}\right)\right\| &\le \frac{1}{\sqrt{n}}\cdot\left\|\left[{\boldsymbol{Z}^{(\lambda)}_{\mathrm{ctr}}}^\top\right]^\dagger\right\|\cdot\left\|\boldsymbol{1}_n^\top\cdot(\boldsymbol{Z}-\boldsymbol{X})\right\| \\
&\le \frac{1}{\sqrt{n}}\cdot\frac{\|\boldsymbol{Z}-\boldsymbol{X}\|}{\sigma^{(\lambda)}\left(\boldsymbol{Z}_{\mathrm{ctr}}\right)}\cdot\|\boldsymbol{1}_n\|.
\end{aligned}$$

All in all, we obtain

$$\sqrt{n}\cdot\left\|\Pi^{\mathrm{col}}_{\boldsymbol{Z}^{(\lambda)}_{\mathrm{ctr}}}\cdot(\tilde{c}_x - c_x)\right\| \le \frac{\|\boldsymbol{Z}-\boldsymbol{X}\|}{\sigma^{(\lambda)}\left(\boldsymbol{Z}_{\mathrm{ctr}}\right)}\cdot\Big(\sqrt{n}\cdot\|c_x\| + \|\boldsymbol{1}_n\|\Big) \tag{65}$$

(2) The second term in (62). Letting $\boldsymbol{E}^{(\lambda)} := \boldsymbol{Z}^{(\lambda)}_{\mathrm{ctr}} - \boldsymbol{X}^{(\lambda)}_{\mathrm{ctr}}$, we observe that

$$\begin{aligned}
\left\|\Pi^{\mathrm{col}}_{\boldsymbol{Z}^{(\lambda)}_{\mathrm{ctr}}} - \Pi^{\mathrm{col}}_{\boldsymbol{X}^{(\lambda)}_{\mathrm{ctr}}}\right\| &\le \max\left\{\left\|\boldsymbol{E}^{(\lambda)}\cdot{\boldsymbol{X}^{(\lambda)}_{\mathrm{ctr}}}^\dagger\right\|,\ \left\|\boldsymbol{E}^{(\lambda)}\cdot{\boldsymbol{Z}^{(\lambda)}_{\mathrm{ctr}}}^\dagger\right\|\right\} && \because \text{Lemma 5} \\
&\le \left\|\boldsymbol{E}^{(\lambda)}\right\|\cdot\max\left\{\left\|{\boldsymbol{X}^{(\lambda)}_{\mathrm{ctr}}}^\dagger\right\|,\ \left\|{\boldsymbol{Z}^{(\lambda)}_{\mathrm{ctr}}}^\dagger\right\|\right\} \\
&\le \frac{\|\boldsymbol{Z}-\boldsymbol{X}\|}{\min\left\{\sigma^{(\lambda)}(\boldsymbol{X}_{\mathrm{ctr}}),\ \sigma^{(\lambda)}(\boldsymbol{Z}_{\mathrm{ctr}})\right\}}.
\end{aligned}$$

All in all, we obtain the following upper bound:

$$\sqrt{n}\left\|\left(\Pi^{\mathrm{col}}_{\boldsymbol{Z}^{(\lambda)}_{\mathrm{ctr}}} - \Pi^{\mathrm{col}}_{\boldsymbol{X}^{(\lambda)}_{\mathrm{ctr}}}\right)\cdot c_x\right\| \le \left\|\Pi^{\mathrm{col}}_{\boldsymbol{Z}^{(\lambda)}_{\mathrm{ctr}}} - \Pi^{\mathrm{col}}_{\boldsymbol{X}^{(\lambda)}_{\mathrm{ctr}}}\right\|\cdot\|c_x\| \le \frac{\|\boldsymbol{Z}-\boldsymbol{X}\|}{\min\left\{\sigma^{(\lambda)}(\boldsymbol{X}_{\mathrm{ctr}}),\ \sigma^{(\lambda)}(\boldsymbol{Z}_{\mathrm{ctr}})\right\}}\cdot\sqrt{n}\cdot\|c_x\|. \tag{66}$$

Step 3: Concluding the proof. We conclude this proof by inserting the upper bounds (65) and (66) from Step 2 into the upper bound (62) in Step 1. Specifically, we obtain

$$\begin{aligned}
\left\|\vec{w}^{(\lambda)}_{\widetilde{\mathcal{D}}_n}(x) - \vec{w}^{(\lambda)}_{\mathcal{D}_n}(x)\right\| &\le \frac{\|\boldsymbol{Z}-\boldsymbol{X}\|}{\sigma^{(\lambda)}\left(\boldsymbol{Z}_{\mathrm{ctr}}\right)}\cdot\Big(\sqrt{n}\cdot\|c_x\| + \|\boldsymbol{1}_n\|\Big) + \frac{\|\boldsymbol{Z}-\boldsymbol{X}\|}{\min\left\{\sigma^{(\lambda)}(\boldsymbol{X}_{\mathrm{ctr}}),\ \sigma^{(\lambda)}(\boldsymbol{Z}_{\mathrm{ctr}})\right\}}\cdot\sqrt{n}\cdot\|c_x\| \\
&\le \frac{\|\boldsymbol{Z}-\boldsymbol{X}\|}{\min\left\{\sigma^{(\lambda)}(\boldsymbol{X}_{\mathrm{ctr}}),\ \sigma^{(\lambda)}(\boldsymbol{Z}_{\mathrm{ctr}})\right\}}\cdot\big(2\sqrt{n}\cdot\|c_x\| + \|\boldsymbol{1}_n\|\big).
\end{aligned}$$

Lastly, we note that $\|c_x\| = \sqrt{\left(x-\mu_{\mathcal{D}_n}\right)^\top\Sigma^\dagger_{\mathcal{D}_n}\left(x-\mu_{\mathcal{D}_n}\right)} = \left\|x-\mu_{\mathcal{D}_n}\right\|_{\Sigma_{\mathcal{D}_n}}$ and $\|\boldsymbol{1}_n\| = \sqrt{n}$. □

D.3 Completing the proof of Theorem 3

Recall that given a set $\mathcal{D}_n = \{(x_i, y_i) : i \in [n]\}$, we let $\boldsymbol{X}_{\mathcal{D}_n} := \begin{bmatrix} x_1 & \cdots & x_n \end{bmatrix}^\top \in \mathbb{R}^{n\times p}$. In addition, we let

$$\forall y \in \mathcal{M}, \ \vec{d}^2_{\mathcal{D}_n}(y) := \begin{bmatrix} d^2(y_1, y) & \cdots & d^2(y_n, y) \end{bmatrix} \in \mathbb{R}^n. \tag{67}$$

Recall that we let $\boldsymbol{X} = \boldsymbol{X}_{\mathcal{D}_n}$ and $\boldsymbol{Z} = \boldsymbol{X}_{\widetilde{\mathcal{D}}_n}$ for shorthand, and further, we let $\boldsymbol{X}_{\mathrm{ctr}} = \big(\boldsymbol{I}_n - \frac{1}{n}\mathbf{1}_n\mathbf{1}_n^\top\big)\boldsymbol{X}$ and $\boldsymbol{Z}_{\mathrm{ctr}} = \big(\boldsymbol{I}_n - \frac{1}{n}\mathbf{1}_n\mathbf{1}_n^\top\big)\boldsymbol{Z}$ denote the 'row-centered' matrices. Here we present and prove the complete version of Theorem 3.

Theorem 4 (De-noising covariates)**.** *Suppose that Assumptions (C0) and (C1) hold. For any $\lambda \in \mathbb{R}_+$, if $x \in \mu_{\mathcal{D}_n} + \operatorname{rowsp} \boldsymbol{X}_{\mathrm{ctr}}$ and*

$$\|x - \mu_{\mathcal{D}_n}\|_{\Sigma_{\mathcal{D}_n}} \le \frac{1}{2}\left(\frac{C_g \cdot D_g^\alpha}{2\operatorname{diam}(\mathcal{M})} \cdot \frac{\min\left\{\sigma^{(\lambda)}(\boldsymbol{X}_{\mathrm{ctr}}),\ \sigma^{(\lambda)}(\boldsymbol{Z}_{\mathrm{ctr}})\right\}}{\|\boldsymbol{Z} - \boldsymbol{X}\|} - 1\right), \tag{68}$$

then

$$\begin{aligned} &d\left(\varphi^{(\lambda)}_{\widetilde{\mathcal{D}}_n}(x), \varphi^{(\lambda)}_{\mathcal{D}_n}(x)\right) \\ &\quad\le \left(\frac{\|\boldsymbol{Z} - \boldsymbol{X}\|}{\min\left\{\sigma^{(\lambda)}(\boldsymbol{X}_{\mathrm{ctr}}),\ \sigma^{(\lambda)}(\boldsymbol{Z}_{\mathrm{ctr}})\right\}} \cdot \frac{2\cdot \left\|x - \mu_{\mathcal{D}_n}\right\|_{\Sigma_{\mathcal{D}_n}} + 1}{C_g} \cdot \frac{\left\|\vec{d}^2_{\mathcal{D}_n}(\tilde{\varphi}_n)\right\| + \left\|\vec{d}^2_{\mathcal{D}_n}(\varphi_n)\right\|}{\sqrt{n}}\right)^{\frac{1}{\alpha}}. \end{aligned} \tag{69}$$

Proof of Theorem 4. First of all, we recall from (52) that

$$\vec{w}^{(\lambda)}_{\mathcal{D}_n}(x) = \left[w^{(\lambda)}_{\mathcal{D}_n}(x_1, x) \quad \cdots \quad w^{(\lambda)}_{\mathcal{D}_n}(x_n, x)\right] \qquad \text{and} \qquad \vec{w}^{(\lambda)}_{\widetilde{\mathcal{D}}_n}(x) = \left[w^{(\lambda)}_{\widetilde{\mathcal{D}}_n}(z_1, x) \quad \cdots \quad w^{(\lambda)}_{\widetilde{\mathcal{D}}_n}(z_n, x)\right].$$

In addition, recall that we let for any $y \in \mathcal{M}$,

$$\vec{d}^2_{\mathcal{D}_n}(y) = \begin{bmatrix} d^2(y_1, y) & \cdots & d^2(y_n, y) \end{bmatrix} \in \mathbb{R}^n.$$

Thereafter, we observe that for any $y \in \mathcal{M}$ and any $x \in \big(\mu_{\mathcal{D}_n} + \operatorname{rowsp} \boldsymbol{X}_{\mathrm{ctr}}\big)$,

$$\begin{aligned} \left|R^{(\lambda)}_{\widetilde{\mathcal{D}}_n}(y; x) - R^{(\lambda)}_{\mathcal{D}_n}(y; x)\right| &= \frac{1}{n}\left|\sum_{i=1}^n \left(w^{(\lambda)}_{\widetilde{\mathcal{D}}_n}(z_i, x) - w^{(\lambda)}_{\mathcal{D}_n}(x_i, x)\right)\cdot d^2(y_i, y)\right| \\ &= \frac{1}{n}\left|\left\langle \vec{w}^{(\lambda)}_{\widetilde{\mathcal{D}}_n}(x) - \vec{w}^{(\lambda)}_{\mathcal{D}_n}(x),\ \vec{d}^2_{\mathcal{D}_n}(y)\right\rangle\right| \\ &\overset{(a)}{\le} \frac{1}{n}\left\|\vec{w}^{(\lambda)}_{\widetilde{\mathcal{D}}_n}(x) - \vec{w}^{(\lambda)}_{\mathcal{D}_n}(x)\right\| \cdot \left\|\vec{d}^2_{\mathcal{D}_n}(y)\right\| \\ &\overset{(b)}{\le} \frac{\|\boldsymbol{Z} - \boldsymbol{X}\|}{\min\left\{\sigma^{(\lambda)}(\boldsymbol{X}_{\mathrm{ctr}}),\ \sigma^{(\lambda)}(\boldsymbol{Z}_{\mathrm{ctr}})\right\}} \cdot \left(2\cdot \left\|x - \mu_{\mathcal{D}_n}\right\|_{\Sigma_{\mathcal{D}_n}} + 1\right)\cdot \frac{\left\|\vec{d}^2_{\mathcal{D}_n}(y)\right\|}{\sqrt{n}} \end{aligned} \tag{70}$$

where (a) is due to Cauchy-Schwarz inequality, and (b) follows from Lemma 6.

Using shorthand notation $R_n = R^{(\lambda)}_{\mathcal{D}_n}$, $\tilde{R}_n = R^{(\lambda)}_{\widetilde{\mathcal{D}}_n}$, $\varphi_n = \varphi^{(\lambda)}_{\mathcal{D}_n}(x)$, and $\tilde{\varphi}_n = \varphi^{(\lambda)}_{\widetilde{\mathcal{D}}_n}(x)$, we observe that

$$\begin{aligned} &R_n(\tilde{\varphi}_n) - R_n(\varphi_n) \\ &\quad= R_n(\tilde{\varphi}_n) - \tilde{R}_n(\tilde{\varphi}_n) + \tilde{R}_n(\tilde{\varphi}_n) - R_n(\varphi_n) \\ &\quad\overset{(a)}{\le} R_n(\tilde{\varphi}_n) - \tilde{R}_n(\tilde{\varphi}_n) + \tilde{R}_n(\varphi_n) - R_n(\varphi_n) \\ &\quad\overset{(b)}{\le} \frac{\|\boldsymbol{Z} - \boldsymbol{X}\|}{\min\left\{\sigma^{(\lambda)}(\boldsymbol{X}_{\mathrm{ctr}}),\ \sigma^{(\lambda)}(\boldsymbol{Z}_{\mathrm{ctr}})\right\}} \cdot \left(2\cdot \left\|x - \mu_{\mathcal{D}_n}\right\|_{\Sigma_{\mathcal{D}_n}} + 1\right)\cdot \frac{\left\|\vec{d}^2_{\mathcal{D}_n}(\tilde{\varphi}_n)\right\| + \left\|\vec{d}^2_{\mathcal{D}_n}(\varphi_n)\right\|}{\sqrt{n}} \end{aligned} \tag{71}$$

where (a) follows from the optimality of $\tilde{\varphi}_n$, i.e., $\tilde{R}_n(\varphi_n) \geq \tilde{R}_n(\tilde{\varphi}_n)$, and (b) is due to (70).

Finally, we note that if

$$\|x - \mu_{\mathcal{D}_n}\|_{\Sigma_{\mathcal{D}_n}} \leq \frac{1}{2}\left(\frac{C_{\mathrm{g}} \cdot D_{\mathrm{g}}^{\alpha}}{2\,\mathrm{diam}\,(\mathcal{M})} \cdot \frac{\min\left\{\sigma^{(\lambda)}(\boldsymbol{X}_{\mathrm{ctr}}),\ \sigma^{(\lambda)}(\boldsymbol{Z}_{\mathrm{ctr}})\right\}}{\|\boldsymbol{Z} - \boldsymbol{X}\|} - 1\right),$$

then the upper bound in (71) certifies that $R_n(\tilde{\varphi}_n) - R_n(\varphi_n) < C_{\mathrm{g}} \cdot D_{\mathrm{g}}^{\alpha}$. Thus, we can use Assumption (C1) to convert the risk bound (71) to derive a distance bound between the minimizers:

$$d\left(\tilde{\varphi}_n, \varphi_n\right) \leq \left(\frac{R_n(\tilde{\varphi}_n) - R_n(\varphi_n)}{C_{\mathrm{g}}}\right)^{\frac{1}{\alpha}},$$

which completes the proof. □

E Details on the experiments

Experimental setup. We consider combinations of $p \in \{25, 50, 75\}$ and $n \in \{100, 200, 400\}$. The datasets $\mathcal{D}_n = \{(X_i, Y_i) : i \in [n]\}$ and $\widetilde{\mathcal{D}}_n = \{(Z_i, Y_i) : i \in [n]\}$ are generated as follows.

(True covariate X) Let $X_i \sim \mathcal{N}_p\big(\mathbf{0}_p, \Sigma\big)$ be IID multivariate Gaussian with mean $\mathbf{0}_p$ and covariance Σ such that $\mathrm{spec}\,(\Sigma) = \{\lambda_j > 0 : j \in [p]\}$ is an exponentially decreasing sequence such that $\mathrm{tr}\,(\Sigma) = \sum_{j=1}^{p} \lambda_j = p$. In particular, for each p, we consider an exponentially decreasing sequence $1 = a_1 > \cdots > a_p = 10^{-3}$, and then set $\lambda_j = p \cdot a_j / (\sum_{j'=1} a_{j'})$ for each $j \in [p]$. Note that $\sum_{j=1}^{\lfloor p/3 \rfloor} \lambda_j \big/ \sum_{j'=1}^{p} \lambda_{j'} \approx 0.9$ for all $p \in \{25, 50, 75\}$, and thus, Σ is effectively low-rank.

(Noisy covariate Z) Let $Z = X + \varepsilon$, where $\varepsilon \sim \mathcal{N}_p\big(\mathbf{0}_p, \sigma_\varepsilon^2 \cdot \mathrm{diag}\,(\mathbf{1}_p)\big)$. Note that in this setting, we have the signal-to-noise ratio $\mathbb{E}(\|X\|_2^2)/\mathbb{E}(\|\varepsilon\|_2^2) = 1/\sigma_\varepsilon^2$. We set $\sigma_\varepsilon^2 = 0.05^2$.

(Response Y) Given $X = x$, let Y be the distribution function of $\mathcal{N}\big(\mu_{\alpha,\beta}(x) + \eta, \tau^2\big)$, where

- $\mu_{\alpha,\beta}(x) = \alpha + \beta^\top x$ with $\alpha = 1$ and $\beta = p^{-1/2} \cdot \mathbf{1}_p$,
- $\eta \sim \mathcal{N}\big(0, \sigma_\eta^2\big)$,
- $\tau^2 \sim \mathcal{IG}(s_1, s_2)$, an inverse gamma distribution with shape s_1 and scale s_2.

We note that $\mathbb{E}(\tau^2) = \frac{s_2}{s_1 - 1}$ and $\mathrm{Var}(\tau^2) = \frac{s_2^2}{(s_1-1)^2(s_1-2)}$. In particular, when $\tau^2 = 0$, this setting corresponds to the classical linear regression model for scalar responses. We set $\sigma_\eta^2 = 0.5^2$, and $(s_1, s_2) = (18, 17)$. In this setting, we have

- $\mathbb{E}\big(\mu_{\alpha,\beta}(X)\big) = 1$ and $\mathrm{Var}\big(\mu_{\alpha,\beta}(X)\big) = \beta^\top \Sigma \beta \approx 1$ for all $p \in \{25, 50, 75\}$,
- $\mathbb{E}(\tau^2) = 1$ and $\mathrm{Var}(\tau^2) = 0.25^2$.

Evaluation metrics. For the assessment of simulation results, we perform $B = 500$ Monte Carlo experiments, i.e., we draw $\mathcal{D}_n^{(b)}$ and $\widetilde{\mathcal{D}}_n^{(b)}$ independent copies of $\mathcal{D}_n$ and $\widetilde{\mathcal{D}}_n$, respectively, for $b = 1, \ldots, B$.

(Model estimation) Being motivated by the standard regression analysis, we evaluate the accuracy and efficiency of the Fréchet regression function estimator with

$$\mathrm{Bias}^2(\varphi_\nu^{(\lambda)}(x)) = d_W\left(\overline{\varphi}_\nu^{(\lambda)}(x), \varphi_{\nu^*}^{(0)}(x)\right)^2 \quad \text{and} \quad \mathrm{Var}(\varphi_\nu^{(\lambda)}(x)) = \frac{1}{B}\sum_{b=1}^{B} d_W\left(\varphi_{\nu^{(b)}}^{(\lambda)}(x), \overline{\varphi}_\nu^{(\lambda)}(x)\right)^2,$$

where $\nu \in \{\mathcal{D}_n, \widetilde{\mathcal{D}}_n\}$. We note that the above representation is a generalization of the standard bias and variance of the regression function estimator in Euclidean spaces. For the global assessment of the estimation performance, we use the average criterion

$$\overline{\mathrm{Bias}}^2(\varphi_\nu^{(\lambda)}) = \frac{1}{M}\sum_{m=1}^{M} \mathrm{Bias}^2(\varphi_\nu^{(\lambda)}(x_m)) \quad \text{and} \quad \overline{\mathrm{Var}}(\varphi_\nu^{(\lambda)}) = \frac{1}{M}\sum_{m=1}^{M} \mathrm{Var}(\varphi_\nu^{(\lambda)}(x_m)),$$

where $\mathcal{G}_M = \{x_m : m = 1, \ldots, M\}$ a set of fixed evaluation points. In our simulation, we generated $x_1, \ldots, x_M$ from $\mathcal{N}_p\big(\mathbf{0}_p, \Sigma\big)$ with $M = 500$ and the same evaluation set was used throughout the Monte Carlo experiments. In Table 1, we have reported $|\mathrm{Bias}|(\varphi_\nu^{(\lambda)}) = \big[\overline{\mathrm{Bias}}^2(\varphi_\nu^{(\lambda)})\big]^{1/2}$ and $\sqrt{\mathrm{Var}}(\varphi_\nu^{(\lambda)}) = \big[\overline{\mathrm{Var}}(\varphi_\nu^{(\lambda)})\big]^{1/2}$ to have them on the same scale of the metric distance.

(In-sample regression fit) In addition to the above bias and variance, we assess the model error by validating the global Fréchet regression fits of the estimated model with the mean squared error

$$\mathrm{MSE}(\varphi_\nu^{(\lambda)}) = \frac{1}{n}\sum_{i=1}^{n} d_W\big(Y_i, \varphi_\nu^{(\lambda)}(X_i)\big)^2.$$

The MSE is the average of squared metric-distance residuals from the observed responses, which is often unitized to measure the model adequacy in the classical regression analysis. Similarly, the overall performance $\overline{\mathrm{MSE}}(\varphi_\nu^{(\lambda)}) = B^{-1}\sum_{b=1}^{B}\mathrm{MPE}(\varphi_{\nu^{(b)}}^{(\lambda)})$ is reported in Table 1.

(Out-of-sample prediction) For $N = 1000$, generate $(X_1^{\mathrm{new}}, Y_1^{\mathrm{new}}, Z_1^{\mathrm{new}}), \ldots, (X_N^{\mathrm{new}}, Y_N^{\mathrm{new}}, Z_N^{\mathrm{new}})$ from (X, Y, Z), and set $\mathcal{D}_N^{\mathrm{new}} = \{(X_i^{\mathrm{new}}, Y_i^{\mathrm{new}}) : i = 1, \ldots, N\}$ and $\widetilde{\mathcal{D}}_N^{\mathrm{new}} = \{(Z_i^{\mathrm{new}}, Y_i^{\mathrm{new}}) : i = 1, \ldots, N\}$ which independent of $\mathcal{D}_n$ and $\widetilde{\mathcal{D}}_n$, respectively. We measure the our-of-sample prediction performance with the mean squared prediction error

$$\mathrm{MSPE}(\varphi_\nu^{(\lambda)}) = \frac{1}{N}\sum_{i=1}^{N} d_W\big(Y_i^{\mathrm{new}}, \varphi_\nu^{(\lambda)}(X_i^{\mathrm{new}})\big)^2,$$

where $\nu \in \{\mathcal{D}_n, \widetilde{\mathcal{D}}_n\}$. We evaluate the average performance with $\overline{\mathrm{MSPE}}(\varphi_\nu^{(\lambda)}) = B^{-1}\sum_{b=1}^{B}\mathrm{MSPE}(\varphi_{\nu^{(b)}}^{(\lambda)})$.

(The choice of threshold) For simplicity, we chose a universal threshold value as

$$\hat{\lambda}_n = \underset{\lambda\in\Lambda}{\arg\min}\, \overline{\mathrm{MSPE}}(\varphi_{\widetilde{\mathcal{D}}_n}^{(\lambda)}),$$

where Λ is a fine grid on $\big(0, \sqrt{\lambda_1 \cdot p/n}\big)$. Then the same threshold $\hat{\lambda}_n$ was used to evaluate $\mathrm{Bias}^2(\varphi_{\widetilde{\mathcal{D}}^{(b)}}^{(\lambda)}(x))$, $\mathrm{Var}(\varphi_{\widetilde{\mathcal{D}}^{(b)}}^{(\lambda)}(x))$, and $\mathrm{MSE}(\varphi_{\widetilde{\mathcal{D}}^{(b)}}^{(\lambda)}(x))$ for all $b = 1, \ldots, B$. Therefore, we claim that the performance of the SVT estimator reported in Table 1 has further room for improvement if one substitute $\hat{\lambda}_n^{(b)} = \arg\min_{\lambda\in\Lambda}\mathrm{MSPE}(\varphi_{\nu^{(b)}}^{(\lambda)})$ for each Monte Carlo experiment. Although suboptimal results are reported, we note that the proposed SVT outperforms both the oracle estimator and the naive EIV estimator in our simulation study. In practice, one may employ cross-validation for better performance. For the MSPE in Table 1, we reported $\min_{\lambda\in\Lambda}\overline{\mathrm{MSPE}}(\varphi_{\widetilde{\mathcal{D}}_n}^{(\lambda)})$.

Discussion on the simulation results. The results of our numerical study demonstrate that the proposed SVT method consistently improves the estimation and prediction performance, particularly in the errors-in-variables setting. Figure 2 illustrates how the proposed SVT estimator outperforms the naive errors-in-variables (EIV) estimator that corresponds to the SVT with zero thresholding. The naive EIV has an intrinsic model bias, known as the attenuation effect [13], because it regresses responses on error-prone covariates. We note that the naive EIV misspecifies the association structure between responses and the true covariates, and eventually, it leads to statistical inference on the misspecified model. Although the naive EIV analysis has the same efficiency as the global Frćhet regression analysis attains [41], this is not the interest of the original study designed by the error-free sample.

As shown in Theorem 2, the proposed SVT estimator is biased from thresholding singular values in the covariate matrix. However, unlike the naive EIV approach, the SVT estimator benefits from a shrinkage estimation effect such that the error-prone covariates are projected on a low-rank space and the global Fréchet regression model has a reduced dimension of effective covariates. Therefore, the SVT approach gains estimation efficiency by having a smaller estimation variance in the finite sample.

Motivated by these observations, we conducted a finite-sample study to evaluate the estimation and prediction performance of the SVT estimator. Table 1 summarizes our numerical experiments.

As discussed earlier, the EIV consistently showed intrinsic bias, which cannot be improved by increasing the sample size. The SVT method has a greater bias than the naive EIV, but the variance is always smaller. This bias-variance trade-off, as a consequence, significantly improved the prediction performance of the SVT compared to the naive EIV.

Remarkably, it turned out that the SVT estimator achieved a smaller mean squared prediction error (MSPE) than even the oracle estimator (REF) obtained from the error-free sample. The REF estimator showed the smallest mean squared error (MSE) because it had a considerably small bias. However, in our simulation study, the REF overfitted the training sample and showed poor performance in prediction. It is also worth mentioning that the naive EIV estimator showed better prediction performance than the REF estimator. This phenomenon happened because the true covariate matrix is nearly singular in our simulation setup, and the REF suffered from the multicollinearity between covariate components. In addition, measurement errors introduced non-ignorable minimum singular values in the errors-in-variables covariate matrix with the magnitude of the variance of measurement errors. As a result, the naive EIV could unintentionally avoid the multicollinearity issue and have a shrinkage effect like a ridge regression.

These findings provide empirical evidence of the effectiveness and superiority of our approach, reinforcing its practical relevance and potential impact in non-Euclidean regression analysis.