Generalized Tensor Decomposition for Understanding Multi-Output Regression under Combinatorial Shifts

Andong Wang
RIKEN AIP
andong.wang@riken.jp

Yuning Qiu
RIKEN AIP
yuning.qiu@riken.jp

Mingyuan Bai
RIKEN AIP
mingyuan.bai@riken.jp

Zhong Jin
China University of Petroleum-Beijing at Karamay
zhongjin@cupk.edu.cn

Guoxu Zhou*
Guangdong University of Technology
gx.zhou@gdut.edu.cn

Qibin Zhao*
RIKEN AIP
qibin.zhao@riken.jp

Abstract

In multi-output regression, we identify a previously neglected challenge that arises from the inability of training distribution to cover all combinations of input features, leading to combinatorial distribution shift (CDS). To the best of our knowledge, this is the first work to formally define and address this problem. We tackle it through a novel tensor decomposition perspective, proposing the Functional t-Singular Value Decomposition (Ft-SVD) theorem which extends the classical tensor SVD to infinite and continuous feature domains, providing a natural tool for representing and analyzing multi-output functions. Within the Ft-SVD framework, we formulate the multi-output regression problem under CDS as a low-rank tensor estimation problem under the missing not at random (MNAR) setting, and introduce a series of assumptions about the true functions, training and testing distributions, and spectral properties of the ground-truth embeddings, making the problem more tractable. To address the challenges posed by CDS in multi-output regression, we develop a tailored Double-Stage Empirical Risk Minimization (ERM-DS) algorithm that leverages the spectral properties of the embeddings and uses specific hypothesis classes in each frequency component to better capture the varying spectral decay patterns. We provide rigorous theoretical analyses that establish performance guarantees for the ERM-DS algorithm. This work lays a preliminary theoretical foundation for multi-output regression under CDS.

1 Introduction

In the realm of Multi-Output Regressions (MOR), systems are designed to predict multiple related outputs from a set of input features, unlike single-output regressions which neglect the relationship between targets. MOR is effective for various scientific problems [55] such as river quality prediction [17], natural gas demand forecasting [5] and drug efficacy prediction [30]. In many MOR applications, it is interactions among multiple features that the predictions are typically derived from. For example, to predict multiple outcomes such as risk scores and health outcomes of patients, the input features usually include healthcare costs and demographic variables (e.g., race) [40]. Similarly, predictions

[*] Qibin Zhao and Guoxu Zhou are the corresponding authors.

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

to new climate scenarios such as increased temperature and extreme weather events require new representative concentration pathways [12]. To enhance the generalization capability, it is often necessary to increase the diversity of training samples. However, in some scenarios, data from specific interactions remain inaccessible, such as risk scores of multiple diseases involving novel combinations of factors including age, physiological indicators, and biochemical markers (see Figure 1). Therefore, unseen feature combinations impose a significant challenge to precise predictions, especially when involving multiple outputs.

To this end, in this paper, we address this issue by reducing it to only a couple of input features. This issue is known as the Combinatorial Distribution Shift (CDS) problem [46]. Under CDS, the training data are *i.i.d.* sampled from the training distribution $\mathcal{D}_{\text{train}}$ with corresponding label $\mathbf{z}_i$, where $(x_i, y_i) \subset \mathcal{X} \times \mathcal{Y}$ are the input feature combinations from the feature domains $\mathcal{X}$ and $\mathcal{Y}$, and $\mathbf{z}_i = \underline{h}^\star(x_i, y_i) \in \mathbb{R}^K$ denotes the K outputs determined by the vector-valued ground truth function $\underline{h}^\star$. Then the problem reduces to predict the new interactions of input combinations (x, y) sample from $\mathcal{D}_{\text{test}}$, where the probability density or mass of test samples (x, y) under $\mathcal{D}_{\text{train}}$ may be zero. Therefore, the goal of the MOR problem becomes: *how can we generalize to the new feature combinations that have never appeared in the training distribution?*

It should be noticed that the MOR community has low awareness of CDS and current methods struggle to address these challenges. Traditional models dependent on static datasets, fail to capture the variability in practical applications, making them ineffective with new data combinations. This research gap underscores the urgent need for dynamic MOR models that better adapt to the evolving input landscapes seen in real-world deployments.

To address this bottleneck, we formulate the issue as a generalized tensor completion problem, arranging the predictions into a third-order tensor $\underline{\mathbf{Z}}(x, y, :) = \underline{h}^\star(x, y) \in \mathbb{R}^K$. Consequently, the prediction of test data can be intuitively interpreted as vector-valued predictions for Missing Not At Random (MNAR) [38] scenarios in multivariate function (see Eq. (4) for more details). While low-rank tensor decomposition techniques, particularly tensor Singular Value Decomposition (t-SVD) [27, 26, 61], have proven effective in recovering missing entries for discrete data, adapting these methods to the complex domain of possibly continuous multivariate functions requires a fundamental re-establishment of the existing t-SVD framework to address the inherent challenges posed by this transition.

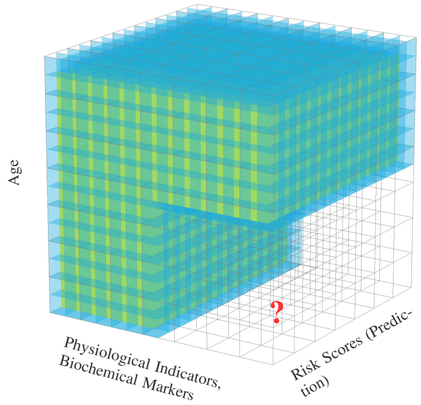

Figure 1: Risk score prediction given novel combinations of features, such as interactions between age, physiological indicators, and biochemical markers that are unseen in the training data, as represented by the empty bottom-right corner.

Our contributions. This paper proposes a novel tensor decomposition perspective for multi-output regression under CDS. It analyzes spectral properties that enable handling such shifts, and introduces an algorithm with theoretical guarantees to improve prediction accuracy. The key contributions are:

- To the best of our knowledge, we are the first to identify and define the problem of multi-output regression under CDS, revealing the challenges that arise when the training distribution fails to cover all combinations of input features.
- We propose a theoretical framework of Functional Tensor Singular Value Decomposition (Ft-SVD), which extends the classical t-SVD to infinite and continuous feature domains, providing a natural tool for representing and analyzing multi-output functions.
- Under the Ft-SVD framework, we formulate the multi-output regression problem under CDS as a low-rank tensor estimation problem under the MNAR setting, and introduce a series of assumptions about the true functions, training and testing distributions, and spectral properties of the embeddings, making the problem more tractable.
- We develop a tailored Double-Stage Empirical Risk Minimization (ERM-DS) algorithm that uses specific hypothesis classes in each sub-domain to better capture the spectral decay patterns across different sub-domains, and provide theoretical guarantees for the algorithm's performance under CDS.

Brief related work. Our work is related to the areas of multi-output regression, tensor completion, and learning under distribution shift. Multi-output regression has been studied extensively, with approaches ranging from multi-task learning [13, 6] to shared representation learning [39], linear models [11], kernel methods [2], and neural networks [9]. However, these methods often assume that the training and test data come from the same distribution. Tensor completion has seen advancements with various decomposition techniques [24, 50, 41] and methods for handling non-random missing patterns [54, 14], but they do not consider the specific challenges of multi-output regression under CDS. Various types of distribution shifts, such as covariate shift [45], concept drift [19], and domain adaptation [44], have been studied, with recent work addressing distribution shift in multi-output regression by aligning distributions through invariant representations [33]. Recent functional tensor decomposition advancements include spectral tensor-train for high-dimensional function evaluations [7], Tucker-neural network for data recovery [37], Bayesian CP/Tucker approach for uncertainty quantification [18], guaranteed functional CP decomposition for functional data analysis [23], and functional Tensor-Train for efficient multivariate function representation [20]. Equivariant disentangled transformation for domain generalization under combination shift was studied through category theory [59]. Simchowitz et al. [46] proposed the ERM-DS framework for robust learning under CDS. Our work extends the ERM-DS framework to multi-output regression by representing vector-valued functions as embeddings in a Hilbert t-module and leveraging tensor algebra to capture interdependencies among multiple outputs.

Notations. We use lowercase boldface, and uppercase boldface letters to denote vectors, *e.g.*, $\mathbf{a} \in \mathbb{R}^m$, and matrices, *e.g.*, $\mathbf{A} \in \mathbb{R}^{m\times n}$, respectively. A 3-way tensor of size $1 \times 1 \times K$ is also named a *t-scalar*, denoted by underlined lowercase, *e.g.*, $\underline{t}$, a 3-way tensor of size $d \times 1 \times K$ is called a *t-vector* and denoted by underlined lowercase boldface, *e.g.*, $\underline{\mathbf{t}}$, whereas a 3-way tensor of size $m \times n \times K$ is also called a *t-matrix* and denoted by underlined uppercase, *e.g.*, $\underline{\mathbf{T}}$. For any tensor $\underline{\mathbf{T}} \in \mathbb{R}^{m\times n\times K}$, we use $\mathbf{T}^{(i)}$ to denote its i-th frontal slice. For $\mathbf{A} \in \mathbb{R}^{m\times m}$, $\sigma_j(\mathbf{A}) \geq 0$ denotes its j-th largest singular value; for symmetric $\mathbf{A}$, $\lambda_j(\mathbf{A})$ denotes its j-th largest eigenvalue.

2 Functional t-Singular Value Decomposition for Multi-output Regression

2.1 A t-SVD Perspective of Multi-output Regression

Consider a simpler case where the feature domains $\mathcal{X}$ and $\mathcal{Y}$ are finite sets. Here, we can represent the vector-valued ground truth $\underline{h}^\star$ as a tensor $\underline{\mathbf{Z}} \in \mathbb{R}^{|\mathcal{X}|\times|\mathcal{Y}|\times K}$, where each tube $\underline{\mathbf{Z}}(i_x, j_y, :) = \underline{h}^\star(x, y) \in \mathbb{R}^K$ corresponds to the K outputs for the input feature combination (x, y) indexed by (i_x, j_y). Then, the multi-output regression problem for a new feature combination (x', y') becomes a tensor completion task, where the goal is to estimate the missing tube of $\underline{\mathbf{Z}}$ at the index $(i_{x'}, j_{y'})$.

To tackle general tensor completion problems, the framework of t-SVD [26, 27] offers an ideal tool [35, 32, 58]. The motivation behind t-SVD stems from the observation that under certain linear transformations M, tensors may exhibit stronger low-rank characteristics than in their original domain. This enhancement of low-rankness often arises due to intrinsic correlations within the data, which these transformations can exploit more effectively. Recent research has focused on using an orthogonal matrix $\mathbf{M}$ to define the transform M due to its advantageous properties [34, 51], a convention that this paper also adopts[2]. Given an *orthogonal* matrix $\mathbf{M} \in \mathbb{R}^{K\times K}$, we define the associated linear transform $M(\cdot)$ and its inverse $M^{-1}(\cdot)$ on any tensor $\underline{\mathbf{T}} \in \mathbb{R}^{m\times n\times K}$ as follows:

$$M(\underline{\mathbf{T}}) := \underline{\mathbf{T}} \times_3 \mathbf{M}, \quad \text{and} \quad M^{-1}(\underline{\mathbf{T}}) := \underline{\mathbf{T}} \times_3 \mathbf{M}^{-1}, \tag{1}$$

where $\times_3$ denotes the mode-3 tensor-matrix product [25]. In real applications, the choice of the transformation matrix $\mathbf{M}$ is often guided by the inherent characteristics of the signal being modeled. Popular choices include the Discrete Cosine Transform (DCT) matrix for smooth and periodic signals [34, 57, 32], the Discrete Wavelet Transform (DWT) matrix for multi-resolution analysis [49], data-dependent transformations that adapt to the dataset's specific characteristics [58], and graph spectral projection matrices for data structured in non-Euclidean spaces [16]. By selecting a transformation matrix that aligns with the signal properties, t-SVD can more effectively uncover the low-rank structure and enable efficient representation and processing of the data.

[2] We restrict $\mathbf{M}$ to be a orthogonal matrix for simplicity of discussions. But our results still hold with simple extensions if necessary for unitary $\mathbf{M}$ used in [27].

Based on the linear transform M, the t-product is specifically defined.

Definition 1 (t-product [25]). *The t-product of tensors* $\underline{\mathbf{A}} \in \mathbb{R}^{m \times n \times K}$ *and* $\underline{\mathbf{B}} \in \mathbb{R}^{n \times k \times K}$ *under the transform* M *in Eq.* (1) *is denoted by* $\underline{\mathbf{A}} *_M \underline{\mathbf{B}}$ *and defined as the tensor* $\underline{\mathbf{C}} \in \mathbb{R}^{m \times k \times K}$ *such that* $M(\underline{\mathbf{C}}) = M(\underline{\mathbf{A}}) \odot M(\underline{\mathbf{B}})$ *in the transformed domain, where* $\odot$ *denotes the tensor frontal-slice-wise product (see Definition 4).*

This paper also follows the definitions of t-transpose, t-identity tensor, t-orthogonal tensor, and f-diagonal tensor given by [25]. Based on these definitions, the t-SVD is introduced as follows:

Definition 2 (t-SVD, tubal rank [25, 27]). *The tensor singular value decomposition (t-SVD) of* $\underline{\mathbf{T}} \in \mathbb{R}^{m \times n \times K}$ *under the transform* M *in Eq.* (1) *is given by:*

$$\underline{\mathbf{T}} = \underline{\mathbf{U}} *_M \underline{\mathbf{S}} *_M \underline{\mathbf{V}}^\top, \tag{2}$$

where $\underline{\mathbf{U}} \in \mathbb{R}^{m \times m \times K}$ *and* $\underline{\mathbf{V}} \in \mathbb{R}^{n \times n \times K}$ *are t-orthogonal tensors,* $\underline{\mathbf{S}} \in \mathbb{R}^{m \times n \times K}$ *is an f-diagonal tensor, and* $(\cdot)^\top$ *denotes the t-transpose. The tubal rank of tensor* $\underline{\mathbf{T}}$ *is defined as the number of non-zero tubes in* $\underline{\mathbf{S}}$ *in Eq.* (2)*, i.e.,*

$$r_{\mathrm{t}}(\underline{\mathbf{T}}) := |\{i : \underline{\mathbf{S}}(i, i, :) \neq \mathbf{0}, i \leq \min\{m, n\}\}|.$$

Based on t-SVD, extensive tensor completion models [32, 58, 52] provide robust support for MOR with discrete combinatorial features. However, in many machine learning settings, feature domains $\mathcal{X}$ and $\mathcal{Y}$ are infinite and potentially continuous sets, posing significant challenges to the traditional t-SVD. Here, t-SVD becomes inapplicable due to the potentially non-discrete nature of these domains, necessitating a novel approach to effectively handle infinite and continuous feature domains. This leads us to consider: *how can we extend the powerful t-SVD framework to address multi-output regression problems in the context of infinite and continuous feature domains?*

2.2 Functional t-Singular Value Decomposition

To address the above challenge, we propose a novel theoretical framework that extends the foundational principles of t-SVD for infinite and continuous feature domains. By introducing a new theorem, we enable the representation of data and functions defined on these domains while preserving the key properties of t-SVD. This extension allows us to employ t-SVD in a principled way for learning vector-valued functions and tackling related problems, opening new possibilities for applying such methods to a wider range of learning tasks.

Theorem 1 (Functional t-Singular Value Decomposition). *Let* $F : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}^K$ *be a square-integrable vector-valued function, where* $\mathcal{X} \subset \mathbb{R}^{D_1}$ *and* $\mathcal{Y} \subset \mathbb{R}^{D_2}$*. Then, there exist sets of functions* $\{\underline{\phi}_i\}_{i=1}^{\infty} \subset L^2(\mathcal{X}; \mathbb{R}^K)$ *and* $\{\underline{\psi}_i\}_{i=1}^{\infty} \subset L^2(\mathcal{Y}; \mathbb{R}^K)$*, and a sequence of t-scalars* $\{\underline{\sigma}_i\}_{i=1}^{\infty} \subset \mathbb{N}^K$ *with* $\lim_{i \to \infty} \underline{\sigma}_i = \underline{0}$*, satisfying*[3] *the functional t-Singular Value Decomposition (Ft-SVD):*

$$F(x, y) = \sum_{i=1}^{\infty} \underline{\phi}_i(x) *_M \underline{\sigma}_i *_M \underline{\psi}_i(y), \tag{3}$$

where the convergence is in the L^2 *sense. The functions* $\underline{\phi}_i$ *and* $\underline{\psi}_i$ *are called the left and right t-singular functions, respectively, and the t-scalars* $\underline{\sigma}_i$ *are called the t-singular values. The orthonormality conditions* $\int_{\mathcal{X}} \underline{\phi}_i(x) *_M \underline{\phi}_j(x) dx = \delta_{ij} M^{-1}(\underline{1})$ and $\int_{\mathcal{Y}} \psi_i(y) *_M \psi_j(y) dy = \delta_{ij} M^{-1}(\underline{1})$ *hold, where* $\underline{1} \in \mathbb{R}^{1 \times 1 \times K}$ *is the t-scalar with all entries equal to 1, and* δ_{ij} *is the Kronecker delta.*

The proof is provided in Appendix B.2.1. The Ft-SVD theorem provides a principled conceptual framework for decomposing a multivariate function, which maps pairs of inputs from domains $\mathcal{X}$ and $\mathcal{Y}$ into a vector in $\mathbb{R}^K$. Essentially, this theorem states that such a function can be expressed as an infinite series of products of three components: t-singular functions from $\mathcal{X}$ and $\mathcal{Y}$ ($\underline{\phi}_i$ and $\underline{\psi}_i$, respectively), and a series of t-singular values ($\underline{\sigma}_i$). This decomposition is analogous to breaking down a complex multivariate relationship into simpler, interpretable modes of variation, where each mode is scaled by its importance, signified by the corresponding t-singular value $\underline{\sigma}_i$.

[3] With a slight abuse of notation, we define the t-product between two vectors in $\mathbb{R}^K$ by treating them as $\mathbb{R}^{1 \times 1 \times K}$ t-scalars. For the sake of simplicity and clarity, we do not explicitly distinguish between $\mathbb{R}^K$ and $\mathbb{R}^{1 \times 1 \times K}$ when the context is clear and there is no risk of confusion.

Unlike t-SVD, which factorizes a tensor using a finite number of t-rank-one components, representing the original function exactly using Ft-SVD may require an infinite number of t-rank-one components in Eq. (3). However, in most practical applications, an approximation with a finite number of components is often desired due to computational constraints and the need for a more compact representation. The natural follow-up inquiry is: *how can we achieve a finite approximation of the original function using Ft-SVD*?

To address this question, we can leverage principles akin to the Eckart-Young theorem for t-SVD [27], which identifies the best low-tubal-rank tensor approximation by minimizing the Frobenius norm error. Building upon these principles, we demonstrate that an optimal rank-r approximation also exists for functions represented using Ft-SVD (see the proof in Appendix B.2.2):

Theorem 2 (r-term truncated Ft-SVD)**.** *Let $F : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}^K$ be a square-integrable function with the Ft-SVD given by Theorem 1. For any $r \in \mathbb{N}$, the r-term truncated Ft-SVD is defined as*

$$F_r(x, y) := \sum_{i=1}^{r} \underline{\phi}_i(x) *_M \underline{\sigma}_i *_M \underline{\psi}_i(y).$$

Then, $F_r(x, y)$ is the best r-term approximation to $F(x, y)$ in the L^2 sense within Ft-SVD framework. Moreover, the approximation error is given by: $\|F(x, y) - F_r(x, y)\|^2_{L^2(\mathcal{X}\times\mathcal{Y};\mathbb{R}^K)} = \sum_{i=r+1}^{\infty} \|\underline{\sigma}_i\|^2$.

This theorem highlights the optimality of the truncated Ft-SVD in approximating vector-valued functions, making it a potentially useful tool for function compression, denoising, and other applications involving low-rank approximations of functions. However, the error term, an infinite sum $\sum_{i=r+1}^{\infty} \|\underline{\sigma}_i\|_2^2$, may still be difficult to bound in the worst case. This raises the question: *are there situations where the error terms are well-controlled?*

In many practical scenarios, the inherent spatial or temporal correlations within data often lead to significant smoothness in functions, a trait crucial for applications in image processing, machine learning, computer vision, climate modeling, and time series analysis, and fluid dynamics [15, 36, 3, 4, 29, 42]. This smoothness property can be leveraged to show that the r-term approximation error of Ft-SVD is indeed well-controlled. The following Theorem 3 shows that if the function components belong to a Sobolev space $H^s(\mathcal{Y})$, which captures their smoothness, then the t-singular values of the function decay rapidly. This rapid decay effectively controls the approximation error, keeping it bounded and manageable. The proof can be found in Appendix B.2.3.

Theorem 3 (Spectral decay of smooth functions)**.** *Let $\mathcal{Y} \subset \mathbb{R}^{D_2}$ be a domain satisfying the strong local Lipschitz condition, and let $F : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}^K$ be a vector-valued function with components $F^{(i)}(x, y)$, for all $i \in [K]$. Suppose there exists a constant $s > 0$ such that $F^{(i)} \in L^2(\mathcal{X}, H^s(\mathcal{Y}))$ for all $i \in [K]$, where $H^s(\mathcal{Y})$ denotes the s-order Sobolev space on $\mathcal{Y}$. Then, the singular values $\underline{\sigma}_i$ of F satisfy the polynomial decay rate $\|\underline{\sigma}_i\|^2 \le O(i^{-1-\frac{2s}{D_2}})$. Moreover, the optimal approximation error of the truncated Ft-SVD with rank r is upper bounded by $O((r+1)^{-\frac{2s}{D_2}})$ in the L_2 sense.*

The rapid spectral decay in the Ft-SVD framework, as guaranteed by the Sobolev smoothness assumption on the output functions, has significant implications for the generalization performance of multi-output regression models. By promoting solutions with rapid spectral decay, the Ft-SVD framework can effectively constrain the complexity of the learned function, striking a balance between fitting the training data and maintaining simplicity. The significance of the spectral decay property in the Ft-SVD framework becomes even more apparent when considering multi-output regression under CDS. In the next section, we will delve into how the approximate low-rankness of the multi-output ground truth, as implied by the rapid spectral decay, provides a key insight into the generalization under CDS. Specifically, we will explore how the spectral decay of the ground truth in the Ft-SVD framework plays a crucial role in enabling effective generalization under combinatorial shifts.

3 An Ft-SVD-based Framework For Multi-output Regression under CDS

3.1 Generalized Tensor Completion with MNAR for MOR under CDS

We propose a theoretical framework based on Ft-SVD to formulate the CDS problem in multi-output regression. The key idea is to represent multi-output functions as t-bilinear embeddings in a Hilbert t-Module, a generalization of Hilbert spaces for handling vector-valued functions. We introduce a series of assumptions on the ground-truth functions, the training and test distributions, and the spectral properties of the embeddings. These assumptions allow us to characterize the multi-output regression problem under CDS as a low-rank tensor estimation problem under MNAR settings.

To set the stage, we first define the concept of a Hilbert t-Module, which serves as the foundation for our t-bilinear representation.

Definition 3 (Hilbert t-Module)**.** *Let $\mathcal{R}$ be the ring of K-dimensional real vectors with t-product. A Hilbert t-Module is a module $\mathcal{M}$ over $\mathcal{R}$ equipped with an $\mathcal{R}$-valued inner product $\langle \cdot, \cdot \rangle_{\mathcal{M}}$, which is complete with respect to the ℓ_2-norm induced by the $\mathcal{R}$-valued inner product.*[4]

Intuitively, a Hilbert t-Module extends the classical Hilbert space to accommodate vector-valued functions, allowing us to perform inner products and norm calculations in a way that respects the tensor structure. This provides a natural framework for representing multi-output functions. With the Hilbert t-Module in place, we can now state our key assumption on the ground-truth functions:

Assumption 1 (t-Bilinear representation)**.** *We have the following assumptions on the ground truth:*

(I) There is a Hilbert t-Module $(\mathcal{M}, \langle \cdot, \cdot \rangle_{\mathcal{M}})$ and two embeddings $\underline{\mathbf{f}}^\star : \mathcal{X} \to \mathcal{M}$ and $\underline{\mathbf{g}}^\star : \mathcal{Y} \to \mathcal{M}$ satisfying that $\underline{h}^\star(x, y) := \langle \underline{\mathbf{f}}^\star(x), \underline{\mathbf{g}}^\star(y) \rangle_{\mathcal{M}}$ is the Bayes optimal predictor on $\mathcal{D}_{\mathrm{train}}$ and $\mathcal{D}_{\mathrm{test}}$, i.e., $\mathbb{E}_{\mathcal{D}_{\mathrm{train}}}[\mathbf{z}|x, y] = \mathbb{E}_{\mathcal{D}_{\mathrm{test}}}[\mathbf{z}|x, y] = \underline{h}^\star(x, y)$. Hereafter, we call $\underline{h}^\star(x, y)$ the ground truth.

(II) We also assume that for some $B > 0$, such that $\mathbb{P}_{(x,y,\mathbf{z})\sim\mathcal{D}_{\mathrm{train}}}[\|\mathbf{z}\|^2 \le B^2] = 1$, and $\max\left\{\sup_{x\in\mathcal{X}} \|\underline{\mathbf{f}}^\star(x)\|_{\mathcal{M}}, \sup_{y\in\mathcal{Y}} \|\underline{\mathbf{g}}^\star(y)\|_{\mathcal{M}}\right\} \le B$.

This assumption postulates that the ground-truth functions admit a t-bilinear representation in terms of two embeddings in a Hilbert t-Module, and the inner product of these embeddings gives the Bayes optimal predictor for both the training and testing distributions. The t-bilinear form is a natural extension of the bilinear form in Ref. [46], and it allows us to capture the multi-dimensional structure of the problem. While the t-bilinear representation is expressive, we need additional assumptions on the training and test distributions to make the problem tractable under CDS:

Assumption 2 (Coverage of training and test distribution, Assumption 2.2 in Ref. [46])**.** *There exist constants $\kappa_{\mathrm{tst}}, \kappa_{\mathrm{trn}} > 0$ and marginal distributions $\mathcal{D}_{\mathcal{X},1}, \mathcal{D}_{\mathcal{X},2}$ over $\mathcal{X}$, and $\mathcal{D}_{\mathcal{Y},1}, \mathcal{D}_{\mathcal{Y},2}$ over $\mathcal{Y}$, with product measures $\mathcal{D}_{i\otimes j} := \mathcal{D}_{\mathcal{X},i} \otimes \mathcal{D}_{\mathcal{Y},j}$, such that for all $(x, y) \in \mathcal{X} \times \mathcal{Y}$, the following Radon–Nikodym derivative conditions hold: (I) Training coverage: $\frac{\mathrm{d}\mathcal{D}_{i\otimes j}(x,y)}{\mathrm{d}\mathcal{D}_{\mathrm{train}}(x,y)} \le \kappa_{\mathrm{trn}}$ for $(i, j) \in \{(1, 1), (1, 2), (2, 1)\}$, and (II) Test coverage: $\frac{\mathrm{d}\mathcal{D}_{\mathrm{test}}(x,y)}{\sum_{i,j\in\{1,2\}} \mathrm{d}\mathcal{D}_{i\otimes j}(x,y)} < \kappa_{\mathrm{tst}}$.*

In words, this assumption requires that the training distribution covers the key feature combinations in $\mathcal{D}_{1\otimes 1}, \mathcal{D}_{1\otimes 2}$ and $\mathcal{D}_{2\otimes 1}$, while the test distribution is allowed to include unseen combinations in $\mathcal{D}_{2\otimes 2}$. The constants $\kappa_{\mathrm{trn}}, \kappa_{\mathrm{tst}}$ quantify the degree of coverage[5].

While Assumption 2 characterizes the relationship between the training and test distributions, it does not directly control the impact of distribution shifts on the model's performance. For this, we introduce a further assumption on the covariate shift:

Assumption 3 (Controlled covariate shifts)**.** *There exists a $\kappa_{\mathrm{cov}} \ge 1$ such that, for any $\underline{\mathbf{v}} \in \mathcal{M}$, the following inequalities hold: $\mathbb{E}_{x\sim\mathcal{D}_{\mathcal{X},2}}[\|\langle \underline{\mathbf{f}}^\star(x), \underline{\mathbf{v}} \rangle_{\mathcal{M}}\|^2] \le \kappa_{\mathrm{cov}} \cdot \mathbb{E}_{x\sim\mathcal{D}_{\mathcal{X},1}}[\|\langle \underline{\mathbf{f}}^\star(x), \underline{\mathbf{v}} \rangle_{\mathcal{M}}\|^2]$ and $\mathbb{E}_{y\sim\mathcal{D}_{\mathcal{Y},2}}[\|\langle \underline{\mathbf{g}}^\star(y), \underline{\mathbf{v}} \rangle_{\mathcal{M}}\|^2] \le \kappa_{\mathrm{cov}} \cdot \mathbb{E}_{y\sim\mathcal{D}_{\mathcal{Y},1}}[\|\langle \underline{\mathbf{g}}^\star(y), \underline{\mathbf{v}} \rangle_{\mathcal{M}}\|^2]$.*

This assumption can be seen as a t-Module variant of Assumption 2.3 in Ref. [46]. It essentially bounds the worst-case impact of covariate shift on the model's performance, ensuring that the error on the unseen distribution $\mathcal{D}_{2\otimes 2}$ is controlled by the error on the training distribution (up to a constant κ_{cov})[6]. This is a key ingredient that allows us to provide generalization guarantees under CDS.

Finally, to leverage the spectral structure of the embeddings, we make the following assumptions:

Assumption 4 (Polynomial spectral decay)**.** *Consider the t-covariances $\underline{\mathbf{\Sigma}}_{\underline{\mathbf{f}}^\star} := \mathbb{E}_{\mathcal{D}_{\mathcal{X},1}}[\underline{\mathbf{f}}^\star *_M (\underline{\mathbf{f}}^\star)^\top]$ and $\underline{\mathbf{\Sigma}}_{\underline{\mathbf{g}}^\star} := \mathbb{E}_{\mathcal{D}_{\mathcal{Y},1}}[\underline{\mathbf{g}}^\star *_M (\underline{\mathbf{g}}^\star)^\top]$. We have the following assumptions:*

[4] The ℓ_2-norm used here differs from the traditional C*-algebra norm in the definition of Hilbert C*-modules, as it does not satisfy the C*-equality $\|x\|^2 = \|\langle x, x \rangle_{\mathcal{M}}\|_{\mathcal{R}}$. Despite this difference, we still refer to this structure as a Hilbert t-Module to emphasize its similarities with Hilbert C*-modules. See Definition 12 for more details.

[5] We give a probabilistic version of this assumption in Assumption 6.

[6] A slacked version of Assumption 3 can be found in Assumption 7.

(I) Balanced embeddings: The ground truth embeddings $\underline{\mathbf{f}}^\star$ and $\underline{\mathbf{g}}^\star$ reside in an appropriate basis of $\mathcal{M}$ such that $\underline{\boldsymbol{\Sigma}}_{\underline{\mathbf{f}}^\star} = \underline{\boldsymbol{\Sigma}}_{\underline{\mathbf{g}}^\star} =: \underline{\boldsymbol{\Sigma}}^\star_{1\otimes 1}$; we also assume $\lambda_1(M(\underline{\boldsymbol{\Sigma}}^\star_{1\otimes 1})^{(i)}) > 0$ holds for all $i \in [K]$.

(II) Polynomial spectral decay: Each of the K frequency components $M(\underline{\boldsymbol{\Sigma}}^\star_{1\otimes 1})^{(i)}$ of the t-covariance $\underline{\boldsymbol{\Sigma}}^\star_{1\otimes 1}$ exhibits a polynomial singular value decay pattern, but potentially with different decay rates $\gamma_i > 0$, i.e., $\sigma_j(M(\underline{\boldsymbol{\Sigma}}^\star_{1\otimes 1})^{(i)}) \le Cj^{-(1+\gamma_i)}, \forall j \in \mathbb{N}, \forall i \in [K]$.

Assumption 5 (Small low-rank approximation error of $(\underline{\mathbf{f}}^\star, \underline{\mathbf{g}}^\star)$ on $\mathcal{D}_{\text{train}}$)**.** *Let $\mathbf{P}_k^{(i)}$ denote the projection onto the top-k eigenspace of $M(\underline{\boldsymbol{\Sigma}}^\star_{1\otimes 1})^{(i)}$, which represents the i-th frequency component ($\forall i \in [K]$) of $\underline{\boldsymbol{\Sigma}}^\star_{1\otimes 1}$ induced by the operator $M(\cdot)$ in Eq. (1). Define $\text{ApxErr}_k^{(i)}(x, y) := (M(\underline{h}^\star(x, y))^{(i)} - \langle \mathbf{P}_k^{(i)} M(\underline{\mathbf{f}}^\star(x))^{(i)}, \mathbf{P}_k^{(i)} M(\underline{\mathbf{g}}^\star(y))^{(i)}\rangle)^2$ as the error between the ground truth $\underline{h}^\star(x, y)$ and the optimal rank-k approximations of the ground truth embeddings $(\underline{\mathbf{f}}^\star, \underline{\mathbf{g}}^\star)$ over the training data $\mathcal{D}_{\text{train}}$ for all rank $k \in \mathbb{N}$ in each i-th frequency component. We assume that*

$$\mathbb{E}_{\mathcal{D}_{\text{train}}}[\text{ApxErr}_k^{(i)}(x, y)] \le \kappa_{\text{apx}} \cdot \mathbb{E}_{\mathcal{D}_{1\otimes 1}}[\text{ApxErr}_k^{(i)}(x, y)], \ \forall i \in [K].$$

Assumptions 4 and 5 postulate that the ground-truth embeddings have a favorable spectral structure. Specifically, they assume that the embeddings are balanced in an appropriate basis, and their spectrum decays polynomially (possibly at different rates for different frequency components). Fig. 2 illustrates the varying rates of spectral decay exhibited by the DCT frequency components of the *Akiyo* video data[7]. This empirical observation serves as a compelling motivation for Assumption 4-(II). Assumption 5 implies that the training distribution $\mathcal{D}_{\text{train}}$ is sufficiently representative of the true function $\underline{h}^\star$ in each frequency component. It guarantees that if we find embeddings that well approximate $\underline{h}^\star$ on the training data, they will also perform well on the reference distribution $\mathcal{D}_{1\otimes 1}$.

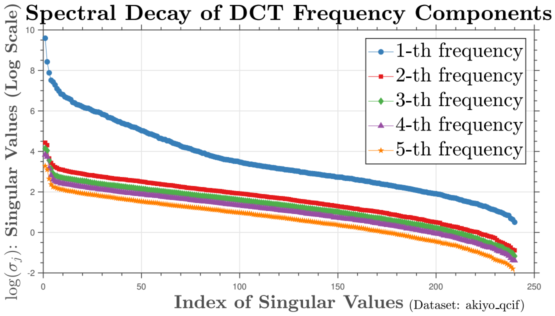

Figure 2: An illustration of the varying rates of spectral decay across different frequency components following Discrete Cosine Transform (DCT). For this example, we consider a discrete tensor $\underline{\mathbf{T}} \in \mathbb{R}^{240\times 320\times 5}$, which comprises the initial five frames of the reshaped *Akiyo* video sequence. This tensor serves as an instance of $\underline{\mathbf{f}}^\star$ in Assumption 4-(II), with the transform M represented by the DCT operation.

Formulating MOR under CDS as generalized tensor completion with MNAR. Based on the assumptions, we can formulate the MOR problem under CDS as a tensor completion problem under the MNAR setting. Specifically, given the training data $\{(x_i, y_i, \mathbf{z}_i)\}_{i=1}^n$ drawn from $\mathcal{D}_{\text{train}}$, we aim to estimate the underlying embeddings $\underline{\mathbf{f}}^\star$ and $\underline{\mathbf{g}}^\star$ by solving the following problem:

$$\min_{\underline{\mathbf{f}}, \underline{\mathbf{g}} \in \mathcal{M}} \frac{1}{n} \sum\nolimits_{i=1}^n \ell(\langle \underline{\mathbf{f}}(x_i), \underline{\mathbf{g}}(y_i)\rangle_{\mathcal{M}}, \mathbf{z}_i), \tag{4}$$

where $\ell(\cdot, \cdot)$ is a suitable loss function. Problem (4) can be interpreted as a tensor completion problem with MNAR: (1) The t-bilinear form $\langle \underline{\mathbf{f}}^\star(x), \underline{\mathbf{g}}^\star(y)\rangle_{\mathcal{M}}$ encodes the function $\underline{h}^\star(x, y)$ as a (generalized) tensor. (2) The training data only covers a subset of the feature combinations due to Assumption 2. (3) The goal is to estimate the complete tensor from the partially observed tensor entries. By solving Problem (4), we obtain the estimated embeddings $\hat{\underline{\mathbf{f}}}$ and $\hat{\underline{\mathbf{g}}}$, which can be used to make predictions on a new test point (x, y) via the t-bilinear form $\langle \hat{\underline{\mathbf{f}}}(x), \hat{\underline{\mathbf{g}}}(y)\rangle_{\mathcal{M}}$.

3.2 Algorithms for Robust Generalization of Multi-output Regression under CDS

How does ERM-based multi-output regression perform under CDS? A natural approach to training a multi-output regression model is to fix a target rank $r \in \mathbb{N}$ and compute an Empirical Risk

[7] The original video can be accessed at http://trace.eas.asu.edu/yuv/index.html.

Minimizer (ERM) by

$$(\hat{\underline{\mathbf{f}}}_{\text{erm}}, \hat{\underline{\mathbf{g}}}_{\text{erm}}) \in \underset{\underline{\mathbf{f}} \in \mathcal{F}_r, \underline{\mathbf{g}} \in \mathcal{G}_r}{\operatorname{argmin}} \frac{1}{n} \sum_{i=1}^{n} \|\underline{\mathbf{f}}(x_i)^\top *_M \underline{\mathbf{g}}(y_i) - \mathbf{z}_i\|^2, \tag{5}$$

where $(\cdot)^\top$ denotes the t-transpose (see Definition 6), and function classes $\mathcal{F}_r, \mathcal{G}_r$ are given in Assumption 8. Assumption 8 ensures that the function classes $\mathcal{F}_r$ and $\mathcal{G}_r$ are sufficiently expressive to approximate the true embeddings $\underline{\mathbf{f}}^\star$ and $\underline{\mathbf{g}}^\star$, while also having controlled complexity in terms of their covering numbers. With the assumption in place, we can obtain the following generalization guarantee for the ERM solution (see the proof in Appendix D.2):

Theorem 4 (Excess Risk Bound for ERM under CDS). *Suppose the learned embeddings $(\hat{\underline{\mathbf{f}}}_{\text{erm}}, \hat{\underline{\mathbf{g}}}_{\text{erm}})$ are $\boldsymbol{\alpha}$-conditioned and $(\boldsymbol{\epsilon}_{\text{trn}}, \boldsymbol{\epsilon}_{\mathcal{D}_{1\otimes 1}})$-accurate embeddings satisfying*[8] *$\breve{\epsilon}^{(i)}_{\mathcal{D}_{1\otimes 1}} \leq \breve{\sigma}_1^{\star,(i)}/(40r)$ for all $i \in [K]$. Then, under Assumptions 1 to 5 and 8, the excess risk of the ERM solution on $\mathcal{D}_{\text{test}}$ can be bounded up to a constant factor $c = \operatorname{poly}(\kappa_{\text{cov}}, \kappa_{\text{tst}}, \kappa_{\text{trn}})$ with probability at least $1-\delta$:*

$$\underbrace{\alpha r^2 \sum_i (\breve{\sigma}_r^{\star,(i)})^2 + r^4 \textsf{\textbf{tail}}_2^\star(r) + r^2 (\textsf{\textbf{tail}}_1^\star(r))^2 + \frac{\alpha r^6 \textsf{\textbf{tail}}_2^\star(r)^2}{\sigma^2}}_{\text{approximation error}} + \underbrace{r^4 \Delta_n + \frac{\alpha r^6}{\sigma^2} \Delta_n^2}_{\text{statistical error}},$$

where $\alpha := \max_i\{\alpha_i\}$, $\sigma := \min_i\{\breve{\sigma}_r^{\star,(i)}\} > 0$, $\textsf{\textbf{tail}}_q^\star(r) := \sum_{i=1}^K \sum_{j>r} (\breve{\sigma}_j^{\star,(i)})^q$, $q \geq 1$, $\Delta_n = B^4(\mathscr{N}(r, 2B/n) + \log\frac{2}{\delta})/n$ with $\mathscr{N}(r,\epsilon) = \mathcal{N}(\mathcal{F}_r, \epsilon/(2B), \|\cdot\|_\infty) \cdot \mathcal{N}(\mathcal{G}_r, \epsilon/(2B), \|\cdot\|_\infty)$. Here, $\breve{\sigma}_j^{\star,(i)} := \sigma_j\left(M(\underline{\mathbf{\Sigma}}^\star_{1\otimes 1})^{(i)}\right)$ denotes the j-th largest singular value of the i-th frequency component $M(\underline{\mathbf{\Sigma}}^\star_{1\otimes 1})^{(i)}$ of the t-covariance operator $\underline{\mathbf{\Sigma}}^\star_{1\otimes 1}$, and $\mathcal{N}(\mathcal{F}, \epsilon, \|\cdot\|_\infty)$ denotes the covering number of a function class at scale ϵ.

Extending the framework of Ref. [46], this theorem bounds the excess risk of the ERM solution for multi-output regression under CDS, incorporating both approximation and statistical errors. Our key contribution, the Ft-SVD for infinite-dimensional tensor completion, addresses multi-output MOR while considering spectral decay across frequency components. However, Theorem 4 has two primary limitations: (1) It assumes that the learned embeddings are $\boldsymbol{\alpha}$-conditioned and $(\boldsymbol{\epsilon}_{\text{trn}}, \boldsymbol{\epsilon}_{\mathcal{D}_{1\otimes 1}})$-accurate, which may not be achievable with ERM. (2) The theorem requires that the minimum singular value $\breve{\sigma}_r^{\star,(i)}$ is strictly positive for each frequency component $i \in [K]$. This condition may not hold for t-embeddings that exhibit varying spectral decay patterns across different frequency components. In practice, multi-output ground truths often demonstrate diverse spectral decay patterns, making it challenging for the hypothesis classes $\mathcal{F}_r$ and $\mathcal{G}_r$ to capture precise low-rank structures across different frequency components (See Fig. 2 for example).

ERM-DS: Addressing the limitations of single-stage ERM. To address the weaknesses of single-stage ERM, we propose the Double-Stage Empirical Risk Minimization (ERM-DS) algorithm. ERM-DS uses a two-stage training process with hypothesis classes tailored to each frequency component, better capturing varying spectral decay patterns and balancing model complexity with generalization ability. Our work extends the ERM-DS framework, originally proposed by Simchowitz et al. for robust learning under CDS [46], to address the challenging setting of multi-output regression. By representing vector-valued functions as embeddings in a Hilbert t-module and employing tensor algebra, we capture the complex interdependencies among multiple outputs, enabling robust generalization.

Unlike Assumption 8, the ERM-DS algorithm considers *fine-grained* hypothesis classes satisfying Assumption 9, which defines function classes specifically for each frequency component to better capture the distinct decay behaviors exhibited by the ground truth in each frequency component. This allows ERM-DS to adapt to the varying spectral properties of the true embeddings across different frequency components, leading to improved generalization performance[9]. The ERM-DS algorithm consists of four main steps:

[8] The $\boldsymbol{\alpha}$-conditioned embeddings and $(\boldsymbol{\epsilon}_{\text{trn}}, \boldsymbol{\epsilon}_{\mathcal{D}_{1\otimes 1}})$-accurate embeddings are defined in Definition 17 and Definition 18, respectively. We require $\breve{\epsilon}^{(i)}_{\mathcal{D}_{1\otimes 1}} \leq \breve{\sigma}_1^{\star,(i)}/(40r)$, where $\breve{\epsilon}^{(i)}_{\mathcal{D}_{1\otimes 1}}$ is the i-th element of vector $\boldsymbol{\epsilon}_{\mathcal{D}_{1\otimes 1}}$ in the definition of $(\boldsymbol{\epsilon}_{\text{trn}}, \boldsymbol{\epsilon}_{\mathcal{D}_{1\otimes 1}})$-accurate embeddings.

[9] Please refer to Appendix D.3 for the details of Assumption 9 and the ERM-DS algorithm.

1. **Overparameterized Training**: Train an overparameterized model $(\tilde{\underline{\mathbf{f}}}, \tilde{\underline{\mathbf{g}}})$ to approximate the unknown true predictive functions $\underline{\mathbf{f}}^\star$ and $\underline{\mathbf{g}}^\star$ by choosing their frequency components separately via K parallel ERM sub-problems (Eq. (18)).
2. **t-Covariance Estimation**: Estimate the t-covariances of the embeddings using additional unlabeled examples to capture the important directions of variation (Eq. (19)).
3. **Dimension Reduction**: Compute low-rank projections using the estimated covariances and obtain reduced-rank embeddings by projecting the overparameterized embeddings onto the low-rank t-subspace (Eq. (20)).
4. **Distillation**: Fine-tune the reduced-rank embeddings by approximating their frequency components separately in the transformed domain, minimizing a combination of empirical risk and consistency with the reduced-rank embeddings (Eq. (21)).

The theoretical guarantees for the performance of the ERM-DS algorithm are provided in Theorem 5:

Theorem 5 (Excess Risk Bound for ERM-DS under CDS)**.** *Under appropriate conditions on algorithm parameters and sample sizes*[10]*, for any* $\delta > 0$*, the ERM-DS solution achieves an excess risk bound of* $c'(\epsilon^2 + \sum_{i=1}^{K}(1+\gamma_i^{-2})r_{i,\mathrm{cut}}^{-2\gamma_i})$ *with probability at least* $1-\delta$*. Here,* $c' = \mathrm{poly}(\kappa_{\mathrm{cov}}, \kappa_{\mathrm{tst}}, \kappa_{\mathrm{trn}})$ *is a problem-dependent constant,* $\epsilon > 0$ *is the desired accuracy level,* $\{r_{i,\mathrm{cut}}\}_{i=1}^{K}$ *are cutoff rank parameters in ERM-DS, and* $\{\gamma_i\}_{i=1}^{K}$ *are parameters related to spectral decay in Assumption 4.*

The proof is given in Appendix D.3. Theorem 5 provides a generalization error bound for the ERM-DS algorithm in multi-output regression under CDS. The expected squared error between predicted and true outputs can be bounded by two main terms: the desired accuracy level ϵ^2, adjustable by algorithm parameters and a term depending on the spectral decay properties of the true embeddings in each frequency component, represented by $\sum_{i=1}^{K}(1+\gamma_i^{-2})r_{i,\mathrm{cut}}^{-2\gamma}$, which captures approximation error due to low-rank structure and target function complexity.

Numerical Experiments. As the first theoretical work on the MOR problem under CDS, this paper proposes the novel Ft-SVD theoretical framework, along with related assumptions and algorithm design. The experiments conducted serve solely as a conceptual validation using synthetic data. We consider the settings when $\mathcal{X}$ and $\mathcal{Y}$ are finite, in which case MOR under CDS naturally degenerates to tensor completion with MNAR tubes. We construct a ground truth tensor $\underline{\mathbf{X}}$ composed of factor matrices[11] $\underline{\mathbf{A}}_1, \underline{\mathbf{B}}_1, \underline{\mathbf{A}}_2, \underline{\mathbf{B}}_2$, where $\mathrm{range}(\underline{\mathbf{A}}_2) \subset \mathrm{range}(\underline{\mathbf{A}}_1)$ and $\mathrm{range}(\underline{\mathbf{B}}_2) \subset \mathrm{range}(\underline{\mathbf{B}}_1)$. The singular values of $M(\underline{\mathbf{A}}_1)^{(i)}, M(\underline{\mathbf{B}}_1)^{(i)}$ follow a power-law decay $\sigma_j^{[1]} = j^{-(1+\gamma)/2}$, while those of $M(\underline{\mathbf{A}}_2)^{(i)}, M(\underline{\mathbf{B}}_2)^{(i)}$ are set to $\sigma_j^{[2]} = c_j \kappa \sigma_j^{[1]}$, where $c_j \in (0,1)$ is a uniform random variable and κ controls the covariate shift. To simulate CDS, we perform two-stage sampling on $\underline{\mathbf{X}}$: first, we randomly sample entries with rate $\mathrm{sr}_{\mathrm{all}}$; second, we sample the top block $\underline{\mathbf{X}}_{11}$ $(\underline{\mathbf{A}}_1, \underline{\mathbf{B}}_1)$ with a lower rate sr_{11}, leaving the bottom block $\underline{\mathbf{X}}_{22}$ $(\underline{\mathbf{A}}_2, \underline{\mathbf{B}}_2)$ as the unobserved test set. We compare the proposed ERM-DS algorithm with the single-stage one, evaluating their test risks under different CDS intensities (by varying the κ parameter that controls the covariate shift) and different percentages of training data. As shown in Fig. 3, the ERM-DS algorithm consistently outperforms the single-stage ERM across all settings. On the left, we observe that as κ increases, indicating more severe covariate shift, the test risk of both algorithms rises, but the ERM-DS algorithm maintains a significant advantage over the single-stage ERM, with the performance gap widening for larger κ values. On the right, the results demonstrate that the ERM-DS algorithm achieves lower test risks compared to the single-stage ERM under varying training data sizes, highlighting its robustness even with limited training samples. Overall, Fig. 3 clearly shows the ERM-DS algorithm achieves lower test risks and stronger generalization capabilities in handling the CDS problem across different covariate shift intensities and training data availabilities.

4 Extensions, Conclusion and Limitation

Extension to higher-order Ft-SVD. Our Ft-SVD framework, while designed for 3-order tensors, is not strictly limited to two-dimensional input cases. It is versatile enough to handle multi-dimensional

[10] Detailed conditions are provided in Theorem 6, Appendix D.3.

[11] Details of the experiments are shown in Appendix A.3.

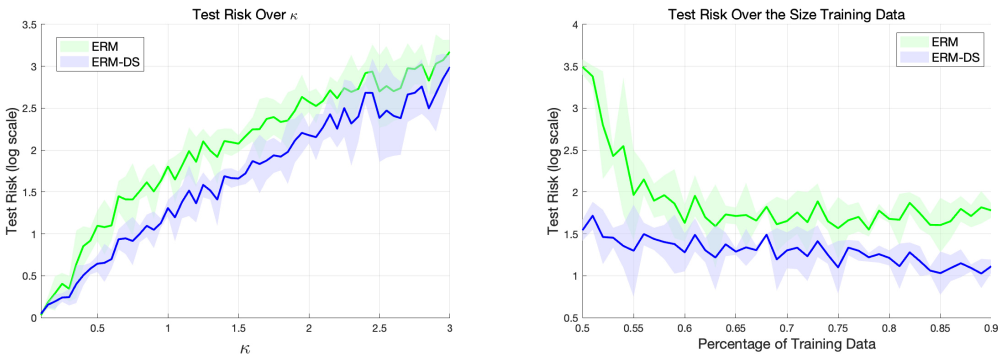

Figure 3: Test risk under different experimental settings. (left) Comparison of test risk over κ (covariate shift intensity) for single and double training approaches. (right) Test risk over the percentage of training data for ERM and ERM-DS.

inputs that can be divided into two distinct sets. Extending Ft-SVD to higher-order cases is indeed non-trivial and we show a preliminary conjecture of a higher-order extension motivated by Ref. [53]:

Let $F : \prod_{i=1}^{N} X_i \to \mathbb{R}^K$ be a square-integrable vector-valued function[12], where $X_i \subset \mathbb{R}^{D_i}$ for $i = 1, \dots, N$. Then there exist sets of functions $\{U_i^n\}_{i=1}^{\infty} \subset L^2(X_n; \mathbb{R}^K)$ for each $n = 1, \dots, N$, and a core function $S : \prod_{i=1}^{N} \mathbb{N} \to \mathbb{R}^K$, such that F can be represented as $F(x_1, \dots, x_N) = \sum_{i_1,\dots,i_N=1}^{\infty} S(i_1, \dots, i_N) *_M \prod_{*\, n=1}^{N} U_{i_n}^n(x_n)$. Here, $\prod_*$ denotes the sequential t-product, which applies the t-product operation sequentially to the functions $U_{i_n}^n(x_n)$. The convergence of this infinite sum is in the L^2 sense. We also have

- Orthogonality: $\forall n \in [N]$, the functions $\{U_i^n\}$ satisfy: $\int_{X_n} U_i^n(x) *_M (U_j^n(x))^\top dx = \delta_{ij} M^{-1}(\underline{1})$.
- Core properties: The core function S has two key characteristics:
  - *i)* All-orthogonality: For all $1 \leq n \leq N$ and all $\alpha \neq \beta$, we have $\int_{\prod_{i \neq n} X_i} S(\dots, \alpha, \dots)^\top *_M S(\dots, \beta, \dots) \prod_{i \neq n} dx_i = \underline{0}$. This means that slices of the core tensor are t-orthogonal.
  - *ii)* Ordering: For all $n = 1, \dots, N$, we have $\|S_{x_n=1}\|_{L^2} \geq \|S_{x_n=2}\|_{L^2} \geq \cdots$, where $\|S_{x_n=\alpha}\|_{L^2}$ denotes the L^2 norm of S with its n-th mode fixed at α. This property ensures a unique ordering of the components.

This decomposition generalizes the concept of Ft-SVD to multilinear functions. It is interesting to generalize the various higher-order variants of t-SVD [22, 31, 43, 1] to functional settings.

Conclusion. The paper addresses the challenge in multi-output regression where training distribution does not cover all input feature combinations, leading to CDS. We introduce a new methodology within a generalized tensor decomposition framework, named Ft-SVD, to tackle this challenge by treating the problem as a tensor completion task under the missing-not-at-random setting. The paper highlights the role of spectral decay of the true embeddings in enhancing model generalization and, through detailed analysis, establishes how multi-output models can manage combinatorial shifts, improving prediction accuracy for new and unseen input combinations.

Limitation. This paper introduces a tensor spectral theory framework to address MOR under CDS, marking an early theoretical exploration in this field. However, several limitations are recognized. First, spectral methods may not fully capture the complexity of real-world data, and the robustness of controlled experimental results remains uncertain. We encourage future research to refine these methods, aiming to develop more effective solutions. Furthermore, an open problem remains in extending the framework to accommodate more than two combinatorial features. Initial investigations suggest that the approach in Ref. [46] does not readily generalize to cases with more than two feature combinations, potentially requiring new mathematical tools to address this challenge.

[12] Note: In the domain of F, $\prod_{i=1}^{N} X_i$ denotes the Cartesian product of the spaces $X_1, \dots, X_N$, not to be confused with the product operations used in the decomposition.

Acknowledgments

Our sincere appreciation goes to both the area chair and reviewers for their unwavering dedication and meticulous attention given to this paper. This research was supported in part by National Natural Science Foundation of China under Grants 62103110, 62203124, 62073087, and 62071132, in part by RIKEN Incentive Research Project 100847-202301062011, and in part by JSPS KAKENHI Grant Numbers JP24K20849, JP24K03005, JP23K28109.

References

[1] S. Ahmadi-Asl, V. Leplat, A.-H. PHAN, and A. Cichocki. A new tensor network: Tubal tensor train network and its applications, 2024. 4

[2] M. A. Alvarez, L. Rosasco, and N. D. Lawrence. Kernels for vector-valued functions: A review. *Foundations and Trends in Machine Learning*, 4(3):195–266, 2012. 1, A.2

[3] H. Antil and S. Bartels. Spectral approximation of fractional pdes in image processing and phase field modeling. *Computational Methods in Applied Mathematics*, 17(4):661–678, 2017. 2.2, B.2.3

[4] H. Antil and C. N. Rautenberg. Sobolev spaces with non-muckenhoupt weights, fractional elliptic operators, and applications. *SIAM Journal on Mathematical Analysis*, 51(3):2479–2503, 2019. 2.2, B.2.3

[5] H. Aras and N. Aras. Forecasting residential natural gas demand. *Energy Sources*, 26:463–472, 04 2004. 1

[6] B. Bakker and T. Heskes. Task clustering and gating for bayesian multitask learning. *Journal of Machine Learning Research*, 4:83–99, 2003. 1, A.2

[7] D. Bigoni, A. P. Engsig-Karup, and Y. M. Marzouk. Spectral tensor-train decomposition. *SIAM Journal on Scientific Computing*, 38(4):A2405–A2439, 2016. 1

[8] B. Blackadar. *Operator algebras: theory of C*-algebras and von Neumann algebras*, volume 122. Springer Science & Business Media, 2006. C.1.1

[9] H. Borchani, G. Varando, C. Bielza, and P. Larrañaga. A survey on multi-output regression. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 5(5):216–233, 2015. 1, A.2

[10] S. Boucheron, O. Bousquet, and G. Lugosi. Theory of classification: A survey of some recent advances. *ESAIM: probability and statistics*, 9:323–375, 2005. D.18

[11] L. Breiman. Bagging predictors. *Machine Learning*, 24(2):123–140, 1996. 1, A.2

[12] C. J. Carlson et al. Climate change will force new animal encounters — and boost viral outbreaks. *Nature*, 605:20, 2022. 1

[13] R. Caruana. Multitask learning. *Machine Learning*, 28(1):41–75, 1997. 1, A.2

[14] J. Chen, C. Lu, Y. Li, H. Zhang, and J. Zhang. Nonconvex tensor completion from noisy data with a guaranteed convergence rate. In *International Conference on Machine Learning*, pages 1489–1498. PMLR, 2020. 1, A.2

[15] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(4):834–848, 2017. 2.2

[16] L. Deng, X.-Y. Liu, H. Zheng, X. Feng, and Y. Chen. Graph spectral regularized tensor completion for traffic data imputation. *IEEE Transactions on Intelligent Transportation Systems*, 23(8):10996–11010, 2021. 2.1

[17] S. Džeroski, D. Demšar, and J. Grbović. Predicting Chemical Parameters of River Water Quality from Bioindicator Data. *Applied Intelligence*, 13(1):7–17, July 2000. 1

[18] S. Fang et al. Functional bayesian tucker decomposition for continuous-indexed tensor data. *arXiv preprint arXiv:2311.04829*, 2023. 1

[19] J. Gama, I. Žliobaitė, A. Bifet, M. Pechenizkiy, and A. Bouchachia. A survey on concept drift adaptation. *ACM Computing Surveys (CSUR)*, 46(4):1–37, 2014. 1, A.2

[20] A. Gorodetsky, S. Karaman, and Y. Marzouk. A continuous analogue of the tensor-train decomposition. *Computer Methods in Applied Mechanics and Engineering*, 347:59–84, 2019. 1

[21] M. Griebel and G. Li. On the decay rate of the singular values of bivariate functions. *SIAM Journal on Numerical Analysis*, 56(2):974–993, 2018. B.2.3

[22] J. Han. p-order tensor products with invertible linear transforms. *arXiv preprint arXiv:2005.11477*, 2020. 4

[23] R. Han, P. Shi, and A. R. Zhang. Guaranteed functional tensor singular value decomposition. *Journal of the American Statistical Association*, 119(546):995–1007, 2024. 1

[24] F. L. Hitchcock. The expression of a tensor or a polyadic as a sum of products. *Journal of Mathematics and Physics*, 6(1-4):164–189, 1927. 1, A.2

[25] E. Kernfeld, M. Kilmer, and S. Aeron. Tensor–tensor products with invertible linear transforms. *Linear Algebra and its Applications*, 485:545–570, 2015. 2.1, 1, 2.1, 2, 6, 7, 8, C.1.1

[26] M. E. Kilmer, K. Braman, et al. Third-order tensors as operators on matrices: A theoretical and computational framework with applications in imaging. *SIAM J MATRIX ANAL A*, 34(1):148–172, 2013. 1, 2.1, B.1, 9

[27] M. E. Kilmer, L. Horesh, H. Avron, and E. Newman. Tensor-tensor algebra for optimal representation and compression of multiway data. *Proceedings of the National Academy of Sciences*, 118(28):e2015851118, 2021. 1, 2.1, 2, 2, 2.2, B.2.2

[28] T. G. Kolda and B. W. Bader. Tensor decompositions and applications. *SIAM Review*, 51(3):455–500, 2009. B.1

[29] J. Lai, Z. Li, D. Huang, and Q. Lin. The optimality of kernel classifiers in sobolev space. In *The Twelfth International Conference on Learning Representations*, 2024. 2.2, B.2.3

[30] H. Li, W. Zhang, Y. Chen, Y. Guo, G.-Z. Li, and X. Zhu. A novel multi-target regression framework for time-series prediction of drug efficacy. *Scientific Reports*, 7(1):40652, Jan. 2017. 1

[31] S. Liu, X.-L. Zhao, J. Leng, B.-Z. Li, J.-H. Yang, and X. Chen. Revisiting high-order tensor singular value decomposition from basic element perspective. *IEEE Transactions on Signal Processing*, 2024. 4

[32] X. Liu, S. Aeron, V. Aggarwal, and X. Wang. Low-tubal-rank tensor completion using alternating minimization. *IEEE TIT*, 66(3):1714–1737, 2020. 2.1, 2.1, 2.1

[33] M. Long, Z. Cao, J. Wang, and M. I. Jordan. Conditional adversarial domain adaptation. In *Advances in Neural Information Processing Systems*, pages 1640–1650, 2018. 1, A.2

[34] C. Lu. Transforms based tensor robust pca: Corrupted low-rank tensors recovery via convex optimization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1145–1152, 2021. 2.1, 2.1

[35] C. Lu, J. Feng, W. Liu, Z. Lin, S. Yan, et al. Tensor robust principal component analysis with a new tensor nuclear norm. *IEEE TPAMI*, 2019. 2.1, 4

[36] V. Lucarini and M. D. Chekroun. Theoretical tools for understanding the climate crisis from Hasselmann's programme and beyond. *Nature Reviews Physics*, 5(12):744–765, 2023. 2.2

[37] Y. Luo et al. Low-rank tensor function representation for multi-dimensional data recovery. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023. 1

[38] W. Ma and G. H. Chen. Missing not at random in matrix completion: The effectiveness of estimating missingness probabilities under a low nuclear norm assumption. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. 1

[39] I. Misra, A. Shrivastava, A. Gupta, and M. Hebert. Cross-stitch networks for multi-task learning. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 3994–4003, 2016. 1, A.2

[40] Z. Obermeyer, B. Powers, C. Vogeli, and S. Mullainathan. Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, 366(6464):447–453, 2019. 1

[41] I. V. Oseledets. Tensor-train decomposition. *SIAM Journal on Scientific Computing*, 33(5):2295–2317, 2011. 1, A.2

[42] K. Park and A. Stinchcombe. Deep reinforcement learning of viscous incompressible flow. *Journal of Computational Physics*, 452:110916, 2022. 2.2, B.2.3

[43] W. Qin, H. Wang, F. Zhang, J. Wang, X. Luo, and T. Huang. Low-rank high-order tensor completion with applications in visual data. *IEEE Transactions on Image Processing*, 31:2433–2448, 2022. 4

[44] J. Quionero-Candela, M. Sugiyama, A. Schwaighofer, and N. D. Lawrence. Dataset shift in machine learning. *The MIT Press*, 2009. 1, A.2

[45] H. Shimodaira. Improving predictive inference under covariate shift by weighting the log-likelihood function. *Journal of Statistical Planning and Inference*, 90(2):227–244, 2000. 1, A.2

[46] M. Simchowitz, A. Gupta, and K. Zhang. Tackling combinatorial distribution shift: A matrix completion perspective. In *The Thirty Sixth Annual Conference on Learning Theory*, pages 3356–3468. PMLR, 2023. 1, 1, 3.1, 2, 3.1, 3.2, 4, A.2, C.1.3, C.1.3, 6, C.1.3, C.2.2, D, D.1, D.1.2, D.1.4, D.11, D.3, D.3, 20, D.3, D.3, D.3, D.3.2, D.3.2

[47] J. Šimša. The best l_2-approximation by finite sums of functions with separable variables. *Aequationes Mathematicae*, 43:248–263, 1992. B.2.2

[48] F. Smithies. The eigen-values and singular values of integral equations. *Proceedings of the London Mathematical Society*, 2(1):255–279, 1938. B.2.2

[49] G. Song, M. K. Ng, and X. Zhang. Robust tensor completion using transformed tensor singular value decomposition. *NUMER LINEAR ALGEBR*, 27(3):e2299, 2020. 2.1

[50] L. R. Tucker. Some mathematical notes on three-mode factor analysis. *Psychometrika*, 31(3):279–311, 1966. 1, A.2

[51] A. Wang, C. Li, M. Bai, Z. Jin, G. Zhou, and Q. Zhao. Transformed low-rank parameterization can help robust generalization for tensor neural networks. In *Advances in Neural Information Processing Systems*, volume 36, 2023. 2.1

[52] A. Wang, G. Zhou, Z. Jin, and Q. Zhao. Tensor recovery via $*_l$-spectral k-support norm. *IEEE Journal of Selected Topics in Signal Processing*, 15(3):522–534, 2021. 2.1

[53] Y. Wang and Y. Yang. Hot-svd: higher order t-singular value decomposition for tensors based on tensor–tensor product. *Computational and Applied Mathematics*, 41(8):394, 2022. 4

[54] W. Xiao, J. Sun, and J. Zhou. Tensor completion with side information: A riemannian manifold approach. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 5051–5058, 2019. 1, A.2

[55] D. Xu, Y. Shi, I. W. Tsang, Y.-S. Ong, C. Gong, and X. Shen. Survey on multi-output learning. *IEEE transactions on neural networks and learning systems*, 31(7):2409–2429, 2019. 1

[56] K. Yosida. *Functional analysis*, volume 123. Springer Science & Business Media, 2012. B.2.1

[57] X. Zhang, M. Bai, and M. Ng. Nonconvex-tv based image restoration with impulse noise removal. *SIAM Journal on Imaging Sciences*, 10(3):1627–1667, 2017. 2.1

[58] X. Zhang and M. K.-P. Ng. Low rank tensor completion with poisson observations. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021. 2.1, 2.1, 2.1

[59] Y. Zhang, J. Wang, X. Xie, and M. Sugiyama. Equivariant disentangled transformation for domain generalization under combination shift. *arXiv preprint arXiv:2208.02011*, 2022. 1

[60] Z. Zhang and S. Aeron. Exact tensor completion using t-svd. *IEEE TSP*, 65(6):1511–1526, 2017. 10

[61] Z. Zhang, G. Ely, S. Aeron, et al. Novel methods for multilinear data completion and de-noising based on tensor-svd. In *CVPR*, pages 3842–3849, 2014. 1

Supplementary Materials for
Generalized Tensor Decomposition for Multi-Output Regression with Combinatorial Distribution Shifts

Contents

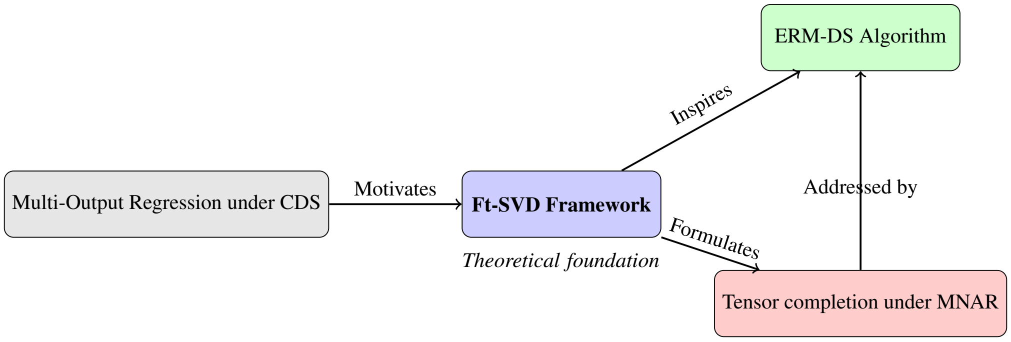

Figure 4: Overview of the core research content in this paper.

The appendix provides additional materials, detailed proofs, and extended discussions to support the main content of the paper *Generalized Tensor Decomposition for Multi-Output Regression with Combinatorial Distribution Shifts*. An overview of the core research content in this paper is shown in Fig. 4. We summarize the notations in Table 1 which provides a comprehensive list of the main symbols, along with their descriptions. The symbols are categorized based on their roles, such as input and output spaces, distributions, embeddings, error terms, hypothesis classes, and algorithmic parameters.

The purpose of this appendix is to offer a comprehensive and rigorous treatment of the theoretical foundations, algorithmic details, and experimental results that underpin the proposed Functional t-Singular Value Decomposition (Ft-SVD) framework for multi-output regression under CDS.

The appendix is organized as follows:

- In Appendix A, we discuss potential applications of the proposed Ft-SVD framework, provide a detailed review of related work, and present details in the experiments.
- Appendix B delves into the theoretical foundations of the Ft-SVD framework. We start by introducing the preliminaries of t-Singular Value Decomposition (t-SVD) and then present the Functional t-SVD, along with the proofs of its key properties (Theorems 1 and Theorem 2). We also discuss the low-tubal-rank approximability of smooth functions and provide the proof of Theorem 3.
- Appendix C provides further explanations and insights into the Ft-SVD framework for multi-output regression under CDS. We offer more details on the Hilbert t-Module, t-bilinear embeddings, and the assumptions used in our analysis. We also present a more in-depth discussion of the proposed algorithms, including the Empirical Risk Minimization (ERM) and the Double-Stage ERM (ERM-DS).
- In Appendix D, we present a comprehensive analysis of the algorithms proposed in the main paper. We start by deriving an error decomposition that forms the basis for our theoretical guarantees. We then provide detailed proofs for the performance bounds of the ERM and ERM-DS algorithms, along with the necessary lemmas and change-of-measure results. This appendix also includes a discussion on the risk bound for t-bilinear combinatorial extrapolation, which is a key component of our analysis.

Broader effects. The paper provides a new tensor spectral perspective for the multi-output regression problem due to combinatorial distribution shift. It focuses solely on these technical aspects and does not have potential societal impacts.

Table 1: Symbol table

Symbol	Description
$\mathcal{X}$	Input space for the first feature
$\mathcal{Y}$	Input space for the second feature
$\mathbb{R}^K$	Output space, where K is the number of outputs
$M(\cdot)$	Linear transform induced by an orthogonal matrix $\mathbf{M} \in \mathbb{R}^{K \times K}$ in Eq. (1)
$M(\underline{\mathbf{T}})^{(i)}$ or $\breve{\underline{\mathbf{T}}}^{(i)}$	i-th frequency component of $\underline{\mathbf{T}}$ in the M-domain
$\mathcal{M}$	The Hilbert t-module (Definition 12) used in Assumption 1
$\langle \cdot, \cdot \rangle_{\mathcal{M}}$	Inner product in the Hilbert t-module $\mathcal{M}$
$\Vert \cdot \Vert_{\mathcal{M}}$	Norm induced by the inner product in the Hilbert t-module $\mathcal{M}$ (in ℓ_2 sense)
$\underline{\mathbf{f}}^\star, \underline{\mathbf{g}}^\star$	Ground truth embeddings, where $\underline{\mathbf{f}}^\star : \mathcal{X} \to \mathcal{M}$ and $\underline{\mathbf{g}}^\star : \mathcal{Y} \to \mathcal{M}$
$\underline{h}^\star$	Ground truth function, where $\underline{h}^\star(x, y) = \langle \underline{\mathbf{f}}^\star(x), \underline{\mathbf{g}}^\star(y) \rangle_{\mathcal{M}}$
$\mathcal{D}_{\text{train}}$	Training distribution over $\mathcal{X} \times \mathcal{Y} \times \mathbb{R}^K$
$\mathcal{D}_{\text{test}}$	Testing distribution over $\mathcal{X} \times \mathcal{Y} \times \mathbb{R}^K$
$\mathcal{D}_{\mathcal{X},1}, \mathcal{D}_{\mathcal{X},2}$	Marginal distributions over $\mathcal{X}$
$\mathcal{D}_{\mathcal{Y},1}, \mathcal{D}_{\mathcal{Y},2}$	Marginal distributions over $\mathcal{Y}$
$\mathcal{D}_{i \otimes j}$	Product measure of $\mathcal{D}_{\mathcal{X},i}$ and $\mathcal{D}_{\mathcal{Y},j}$, where $i, j \in \{1, 2\}$
$\underline{\mathbf{\Sigma}}^\star_{1 \otimes 1}$	t-covariance operator of the ground truth embeddings on $\mathcal{D}_{1 \otimes 1}$
$M(\underline{\mathbf{\Sigma}}^\star_{1 \otimes 1})^{(i)}$	i-th frequency component of the t-covariance operator $\underline{\mathbf{\Sigma}}^\star_{1 \otimes 1}$
$\breve{\sigma}_j^{\star,(i)}$	j-th largest singular value of $M(\underline{\mathbf{\Sigma}}^\star_{1 \otimes 1})^{(i)}$
γ_i	Polynomial decay rate of the singular values of $M(\underline{\mathbf{\Sigma}}^\star_{1 \otimes 1})^{(i)}$
$\underline{\mathbf{P}}^\star_{\mathbf{k}}$	t-projection operator of multi-rank-$\mathbf{k}$ defined by $\underline{\mathbf{\Sigma}}^\star_{1 \otimes 1}$
$\underline{\mathbf{f}}^\star_{\mathbf{k}}, \underline{\mathbf{g}}^\star_{\mathbf{k}}$	Ground truth embeddings truncated at multi-rank $\mathbf{k}$
$\textsf{\textbf{tail}}_q^{(i)\star}(k)$	q-th power tail sum of singular values of $M(\underline{\mathbf{\Sigma}}^\star_{1 \otimes 1})^{(i)}$, starting from index k
$\textsf{\textbf{tail}}_q^\star(\mathbf{k})$	q-th power tail sum of singular values of $M(\underline{\mathbf{\Sigma}}^\star_{1 \otimes 1})$, starting from multi-index $\mathbf{k}$
$\hat{\underline{\mathbf{f}}}_{\text{erm}}, \hat{\underline{\mathbf{g}}}_{\text{erm}}$	Learned embeddings by ERM, where $\hat{\underline{\mathbf{f}}}_{\text{erm}} : \mathcal{X} \to \mathcal{M}$ and $\hat{\underline{\mathbf{g}}}_{\text{erm}} : \mathcal{Y} \to \mathcal{M}$
$\hat{\underline{\mathbf{f}}}_{\text{ds}}, \hat{\underline{\mathbf{g}}}_{\text{ds}}$	Learned embeddings by ERM-DS, where $\hat{\underline{\mathbf{f}}}_{\text{ds}} : \mathcal{X} \to \mathcal{M}$ and $\hat{\underline{\mathbf{g}}}_{\text{ds}} : \mathcal{Y} \to \mathcal{M}$
$\mathcal{R}(\underline{\mathbf{f}}, \underline{\mathbf{g}}; \mathcal{D})$	Excess risk of embeddings $(\underline{\mathbf{f}}, \underline{\mathbf{g}})$ under distribution $\mathcal{D}$
$\Delta_0(\underline{\mathbf{f}}, \underline{\mathbf{g}}, \mathbf{k})$	Weighted error term for embeddings $(\underline{\mathbf{f}}, \underline{\mathbf{g}})$ at multi-rank $\mathbf{k}$
$\Delta_1(\underline{\mathbf{f}}, \underline{\mathbf{g}}, \mathbf{k})$	Unweighted error term for embeddings $(\underline{\mathbf{f}}, \underline{\mathbf{g}})$ at multi-rank $\mathbf{k}$
$\Delta_2(\underline{\mathbf{f}}, \underline{\mathbf{g}}, \mathbf{k})$	Factor recovery error term for embeddings $(\underline{\mathbf{f}}, \underline{\mathbf{g}})$ at multi-rank $\mathbf{k}$ on $\mathcal{D}_{2 \otimes 2}$
$\Delta_{\text{apx}}(\mathbf{k})$	Approximation error at multi-rank $\mathbf{k}$
$\Delta_{\text{train}}(\mathbf{k})$	Error on the training distribution at multi-rank $\mathbf{k}$
$\boldsymbol{\alpha}$	Conditioning vector for the learned embeddings
$\boldsymbol{\epsilon}_{\text{trn}}$	Accuracy vector for the learned embeddings on $\mathcal{D}_{\text{train}}$
$\boldsymbol{\epsilon}_{\mathcal{D}_{1 \otimes 1}}$	Accuracy vector for the learned embeddings on $\mathcal{D}_{1 \otimes 1}$
$\breve{\epsilon}_{\text{trn}}^{(i)}$	Accuracy for the i-th frequency component of the learned embeddings on $\mathcal{D}_{\text{train}}$
$\breve{\epsilon}_{\mathcal{D}_{1 \otimes 1}}^{(i)}$	Accuracy for the i-th frequency component of the learned embeddings on $\mathcal{D}_{1 \otimes 1}$
$\sigma_{i,\text{cut}}$	Cut-off singular value parameter for the i-th frequency component in ERM-DS
$r_{i,\text{cut}}$	Cut-off rank parameter for the i-th frequency component in ERM-DS algorithm
$\boldsymbol{\sigma}_{\text{cut}} := (\sigma_{i,\text{cut}})_{i=1}^K$	Parameter vector of cut-off singular values in the ERM-DS algorithm
$\mathbf{r}_{\text{cut}} := (r_{i,\text{cut}})_{i=1}^K$	Parameter vector of cut-off ranks in the ERM-DS algorithm
$\textsc{Err}_{\text{DT}}^{(i)}(r_{i,\text{cut}}, \sigma_{i,\text{cut}})$	Error term for the i-th frequency component in the ERM-DS algorithm
$\mathcal{F}_k, \mathcal{G}_k$	Hypothesis classes for the t-embeddings at rank k
$\breve{\mathcal{F}}_k^{(i)}, \breve{\mathcal{G}}_k^{(i)}$	Hypothesis classes for i-th frequency component of the t-embeddings at rank k
$\mathscr{N}(k, \epsilon)$	Covering number for the hypothesis classes $(\mathcal{F}_k, \mathcal{G}_k)$ at scale ϵ
$\breve{\mathscr{N}}^{(i)}(k, \epsilon)$	Covering number for the hypothesis classes $(\breve{\mathcal{F}}_k^{(i)}, \breve{\mathcal{G}}_k^{(i)})$ at scale ϵ
Δ_n	Statistical error term depending on the sample size n
$a \lesssim_\star b$	$a \leq c \cdot b$ for some c at most polynomial in the problem constants $\kappa_{\text{cov}}, \kappa_{\text{trn}}, \kappa_{\text{apx}}$

A Prevalence of MoR under CDS, Related Work, and Details of Experiments

A.1 Prevalence of MoR under CDS across Domains

The challenge of Multi-output Regression under Combinatorial Distribution Shifts (MoR under CDS) is pervasive across a wide spectrum of real-world applications. The following examples illustrate the ubiquity of this problem in diverse domains:

- **Healthcare**: In medical diagnostics, predicting a patient's risk for multiple diseases (e.g., heart disease, diabetes) based on various patient attributes (e.g., age, blood pressure, heart rate) constitutes a multi-output regression problem. However, training data often comes from specific patient cohorts, covering only a subset of possible attribute combinations. When new patient groups (e.g., different age brackets or geographic regions) emerge, they introduce novel attribute combinations, leading to combinatorial distribution shifts. Diagnostic systems must adapt to these new patient profiles to accurately predict disease risks.
- **Marketing**: Predicting how different user segments respond to various marketing strategies (e.g., ad types, promotions) based on user attributes (e.g., age, income) and strategy attributes (e.g., ad themes, delivery channels) is a multi-output regression task. Training data typically covers common user-strategy combinations, but when new user groups or marketing strategies appear, they create unseen attribute combinations, resulting in combinatorial distribution shifts. Marketing prediction models need to transfer knowledge to these new scenarios to forecast user responses accurately.
- **Materials Science**: Predicting the performance of material formulations across multiple metrics (e.g., hardness, conductivity) based on composition ratios is a multi-output regression problem. Due to experimental constraints, training data usually includes only a subset of known formulations. When researchers explore novel composition ratios, they encounter new formulation combinations, leading to combinatorial distribution shifts. Material performance prediction systems must generalize to these unseen formulations to accelerate materials discovery.
- **Transportation**: Forecasting multiple traffic flow indicators (e.g., vehicle flow, pedestrian flow) for different areas and time periods based on area attributes (e.g., road network, land use) and temporal attributes (e.g., day of the week, holidays) is a multi-output regression task. Training data may cover only typical area-time combinations, but when new areas develop or novel temporal patterns emerge, they create unseen attribute combinations, causing combinatorial distribution shifts. Traffic prediction models must adapt to these new scenarios for effective transportation planning and management.
- **Environmental Monitoring**: Predicting concentrations of various pollutants (e.g., PM2.5, ozone) at different locations and seasons based on location attributes (e.g., terrain, land use) and meteorological conditions (e.g., temperature, humidity) is a multi-output regression problem. Training data often comes from specific monitoring stations and time periods, covering a limited set of location-weather combinations. When new monitoring stations are deployed or unusual weather patterns occur, they introduce new attribute combinations, leading to combinatorial distribution shifts. Environmental monitoring systems must generalize to these unseen scenarios to comprehensively assess environmental quality.

These examples summarized in Figure 5 underscore the prevalence of MoR under CDS across various domains. In these scenarios, multiple prediction tasks share the same input space, but training distribution covers only a subset of attribute combinations. The emergence of new attribute combinations constitutes a combinatorial distribution shift. Transferring knowledge from observed combinations to novel ones is a common challenge in these applications, and addressing this problem is crucial for developing robust intelligent systems.

Figure 5: Examples of MoR under CDS across domains

Domain	**Input**	**Output**	**CDS**
Healthcare	Patient attributes	Disease risks	New patient groups
Marketing	User & strategy attr.	User responses	New groups/strategies
Materials	Composition ratios	Material properties	Novel formulations
Transport	Area & temporal attr.	Traffic flow	New areas/patterns
Environment	Location & weather	Pollutants	New stations/weather

A.2 Detailed Related Work

Multi-Output Regression. Multi-output regression has been extensively studied in machine learning, with early works focusing on multi-task learning [13, 6] and recent approaches extending to shared representation learning [39], linear models [11], kernel methods [2], and neural networks [9]. However, most of these methods assume that the training and test data come from the same distribution, which may not hold under distribution shifts. In contrast, our work specifically addresses the problem of multi-output regression under CDS by proposing a novel theoretical framework based on Ft-SVD and an extended version of the ERM-DS algorithm.

Tensor Completion. Tensor completion has seen significant advancements with the development of various tensor decomposition techniques [24, 50, 41] and methods that can handle non-random missing patterns [54, 14]. While these works provide a foundation for addressing tensor completion, they do not consider the specific challenges posed by multi-output regression under CDS, such as dealing with infinite and continuous feature domains. Our work aims to address this gap by formulating multi-output regression under CDS as a low-rank tensor estimation problem under the Missing Not At Random (MNAR) setting and developing a tailored ERM-DS algorithm with theoretical guarantees.

Distribution Shift. Various types of distribution shifts have been studied in machine learning, including covariate shift [45], concept drift [19], and domain adaptation [44]. Some recent works have addressed distribution shift in multi-output regression by aligning the distributions through learning invariant representations [33]. However, these methods typically require access to labeled data from the target domain, which may not always be available. In contrast, our work focuses on the specific case of CDS, where the training distribution covers certain marginal distributions of the input features, but the test distribution involves unseen combinations of these features. We address this challenge by leveraging the Ft-SVD framework and the ERM-DS algorithm, which do not require labeled data from the test distribution.

Simchowitz et al. [46] proposed the Double-Stage Empirical Risk Minimization (ERM-DS) framework to address single-output learning under CDS. Our work extends the ERM-DS framework to the multi-output regression setting by representing vector-valued functions as embeddings in a Hilbert t-module and leveraging tensor algebra to capture the interdependencies among multiple outputs. The proposed Ft-SVD theoretical framework, the formulation of multi-output regression under CDS as a low-rank tensor estimation problem, and the tailored ERM-DS algorithm with theoretical guarantees aim to contribute to the existing literature on multi-output regression, tensor completion, and distribution shift, by providing a principled approach to address the specific challenges posed by CDS in multi-output regression.

A.3 Details of Experiments

As the pioneering theoretical study tackling the multi-output regression problem under combinatorial distribution shift, this paper proposes the novel Ft-SVD theoretical framework, along with related assumptions and algorithm design. To validate the proposed approach, we conducted a series of experiments on synthetic tensor data, serving solely as a conceptual proof-of-concept. Considering finite feature domains $\mathcal{X}$ and $\mathcal{Y}$, the MOR problem under CDS naturally degenerates to tensor completion with missing-not-at-random (MNAR) tubes. We conducted a series of experiments using synthetic tensor data. The primary focus of our analysis was to compare the empirical risk of three algorithms: Single ERM in Eq. (5), Overparameterized Training (Step 1 of ERM-DS), and the proposed ERM-DS. We generated a synthetic tensor $\underline{\mathbf{X}} \in \mathbb{R}^{(m_1+m_2)\times(d_1+d_2)\times K}$. The tensor $\underline{\mathbf{X}}$ was constructed using a combination of factor matrices $\underline{\mathbf{A}}_1 \in \mathbb{R}^{m_1\times r\times K}$, $\underline{\mathbf{B}}_1 \in \mathbb{R}^{d_1\times r\times K}$, $\underline{\mathbf{A}}_2 \in \mathbb{R}^{m_2\times r\times K}$, and $\underline{\mathbf{B}}_2 \in \mathbb{R}^{d_2\times r\times K}$, where range$(\underline{\mathbf{A}}_2) \subset$ range$(\underline{\mathbf{A}}_1)$ and range$(\underline{\mathbf{B}}_2) \subset$ range$(\underline{\mathbf{B}}_1)$. The singular value tubes of $\underline{\mathbf{A}}_1$ and $\underline{\mathbf{B}}_1$ followed a power-law decay with rate γ, i.e., $M(\underline{\sigma}_1^{[i]}) = \text{repmat}(i^{-(1+\gamma)/2}, K) \in \mathbb{R}^{1\times 1\times K}$, and the singular value tubes of $\underline{\mathbf{A}}_2$ and $\underline{\mathbf{B}}_2$ are set to be $M(\underline{\sigma}_2^{[i]}) = \text{repmat}(c_i \kappa \sigma_1^{[i]}, K) \in \mathbb{R}^{1\times 1\times K}$, and $c_i \in (0,1)$ is the random variable following the uniform distribution. Therefore, the tensor data can be divided into four parts, namely, $\underline{\mathbf{X}}_{11} \in \mathbb{R}^{m_1\times d_1\times K}$, $\underline{\mathbf{X}}_{12} \in \mathbb{R}^{m_1\times d_2\times K}$, $\underline{\mathbf{X}}_{21} \in \mathbb{R}^{m_2\times d_1\times K}$ and $\underline{\mathbf{X}}_{22} \in \mathbb{R}^{m_2\times d_2\times K}$:

$$\begin{bmatrix} \underline{\mathbf{X}}_{11} & \underline{\mathbf{X}}_{12} \\ \underline{\mathbf{X}}_{12} & \underline{\mathbf{X}}_{22} \end{bmatrix} = \begin{bmatrix} \underline{\mathbf{A}}_1 \\ \underline{\mathbf{A}}_2 \end{bmatrix} *_M \begin{bmatrix} \underline{\mathbf{\Sigma}}_{11} & \\ & \underline{\mathbf{\Sigma}}_{22} \end{bmatrix} *_M \begin{bmatrix} \underline{\mathbf{B}}_1 \\ \underline{\mathbf{B}}_2 \end{bmatrix}^\top \tag{6}$$

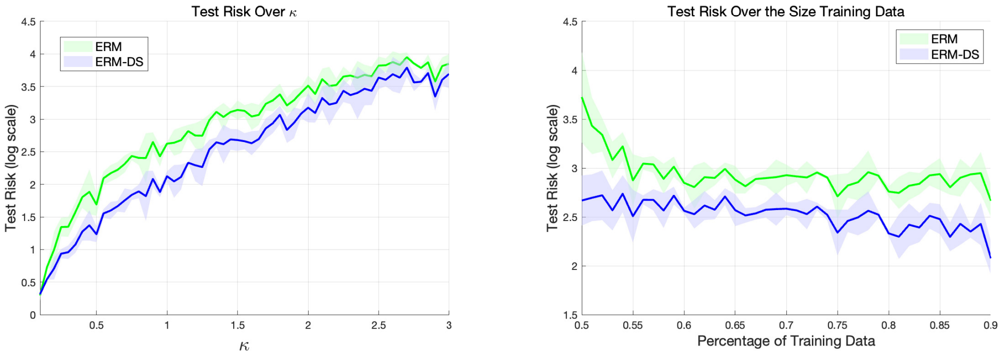

Figure 6: Test risk on a larger-scale tensor of size $400 \times 400 \times 10$ using DCT transformation. (left) Comparison of test risk over κ (covariate shift intensity) for single and double training approaches. (right) Test risk over the percentage of training data for ERM and ERM-DS.

In this case, the factor tensor can be represented as the features in Eq., that is

$$\underline{\mathbf{f}}^{\star}(x_{i'}) = \underline{\mathbf{A}}(i') \text{ and } \underline{\mathbf{g}}^{\star}(y_{j'}) = \underline{\mathbf{B}}(j'), \tag{7}$$

and the tube $\underline{\mathbf{X}}(i', j', :) = \langle \underline{\mathbf{f}}^{\star}(x_{i'}), \underline{\mathbf{g}}^{\star}(y_{j'}) \rangle$ is regarded as the output. Therefore, the data sample can be regarded as $(i', j', \underline{z}_{i'j'})$ where $\underline{z}_{i'j'} = \underline{\mathbf{X}}(i', j', :))$, $i' \in \mathbb{N}^{m_1+m_2}$ and $j' \in \mathbb{N}^{d_1+d_2}$.

To construct $\mathcal{D}_{\text{train}}$, the training data are uniformly sampled from three parts, namely, $\underline{\mathbf{X}}_{11}, \underline{\mathbf{X}}_{12}$ and $\underline{\mathbf{X}}_{12}$. The sampling process was performed in two steps. First, we randomly sampled entries from the subtensor $\underline{\mathbf{X}}_{11}$ with sampling rate $\text{sr}^1_{\text{train}}$ and from the subtensors $\underline{\mathbf{X}}_{12}$ and $\underline{\mathbf{X}}_{21}$ with sampling rates of $\text{sr}^2_{\text{train}}$. The block $\underline{\mathbf{X}}_{22}$ was left unobserved to be the test set. All experiments are implemented in MATLAB on a Linux server equipped with dual Intel E5 2640v4 and 128GB of RAM. The demo code can be found at `https://github.com/pingzaiwang/FtSVD4MORCDS`.

To generate the training and test data, we follow the above construction procedure by letting $m_1 = d_2 = 80$ and $m_2 = d_2 = 120$, and the rank $r = 15$. The singular value decay γ is fixed to 1. The sample ratio of $\underline{\mathbf{X}}_{11}$ is simply fixed to be $\text{sr}^1_{\text{train}} = 10\%$. The transform matrix $\mathbf{M}$ is adopted the DFT matrix. In the single training ERM and over-parameter training step, the estimated rank is set to $2r$ and $4r$, respectively. For simplicity, we let $\kappa \in [0.1, 3]$ and $\text{sr}^2_{\text{train}} \in [50\%, 90\%]$. For each experimental setting, we randomly ran the experiments for four times and computed the average test risk and standard deviation.

In Figure 3 and Figure 6, we depict the test risk and standard deviation over κ and the percentage of training data. On the left, we observe that as κ increases, indicating a more severe covariate shift, the test risk of both algorithms rises, but the ERM-DS algorithm maintains a significant advantage over the single-stage ERM, with the performance gap widening for larger κ values. On the right, the results demonstrate that the ERM-DS algorithm achieves lower test risks compared to the single-stage ERM under varying training data sizes, highlighting its robustness even with limited training samples.

Overall, the ERM-DS algorithm achieves lower test risks and stronger generalization capabilities compared to the single-stage ERM approach. It is crucial to note that these experiments serve only as a conceptual validation of the proposed Ft-SVD theoretical framework, aiming to preliminarily verify the effectiveness of the theoretical methods in handling CDS. Future work is still needed to further develop more effective and practical algorithms for improved performance on real-world data.

B Functional t-Singular Value Decomposition

This section provides a comprehensive exploration of the theoretical underpinnings of the Ft-SVD framework. It begins with an overview of the basics of t-SVD, setting the stage for the introduction of Ft-SVD. This section includes detailed proofs of key properties, as outlined in Theorems 1 and 2. Additionally, the appendix examines the low-rank approximability of smooth functions and presents the proof of Theorem 3, highlighting the theoretical contributions and foundational aspects of Ft-SVD.

B.1 Preliminaries of t-Singular Value Decomposition

We first introduce some basic notations and concepts to lay the foundation for the subsequent discussions.

Basic notations. For any positive integer $n \in \mathbb{N}$, let $[n] := \{1, \cdots, n\}$ be the set of integers from 1 to n. We use lowercase bold letters (e.g., $\mathbf{a} \in \mathbb{R}^m$) for vectors, and uppercase bold letters (e.g., $\mathbf{A} \in \mathbb{R}^{m \times n}$) for matrices. Following the standard notations in [26], we refer to a third-order tensor of size $1 \times 1 \times K$ as a t-scalar, denoted by an underlined lowercase letter, e.g., $\underline{x}$; a third-order tensor of size $d \times 1 \times K$ as a t-vector, denoted by an underlined lowercase bold letter, e.g., $\underline{\mathbf{x}}$; and a third-order tensor of size $m \times n \times K$ as a t-matrix, denoted by an underlined uppercase letter, e.g., $\underline{\mathbf{X}}$. For $\mathbf{A} \in \mathbb{R}^{K \times K}$, $\sigma_i(\mathbf{A}) \geq 0$ denotes its i-th largest singular value; for symmetric $\mathbf{A}$, $\lambda_i(\mathbf{A})$ denotes its i-th largest eigenvalue, and if $\mathbf{A} \succeq 0$, $\mathbf{A}^{1/2}$ its matrix square-root.

Given a tensor $\underline{\mathbf{T}}$, its ℓ_p-norm is defined as $\|\underline{\mathbf{T}}\|_p := \|\texttt{vec}(\underline{\mathbf{T}})\|_p$, and its F-norm is defined as $\|\underline{\mathbf{T}}\|_{\mathrm{F}} := \|\texttt{vec}(\underline{\mathbf{T}})\|_2$, where $\texttt{vec}(\cdot)$ denotes the vectorization operation of a tensor [28]. We also use $\|\cdot\|$ to represent the ℓ_2-norm of vectors, F-norm of matrices and tensors for notation simplicity. For $\underline{\mathbf{T}} \in \mathbb{R}^{m \times n \times K}$, we use $\mathbf{T}^{(i)}$ or $\underline{\mathbf{T}}(:,:,i)$ to denote its i-th frontal slice. The inner product between two tensors $\underline{\mathbf{A}}$ and $\underline{\mathbf{B}}$ is defined as $\langle \underline{\mathbf{A}}, \underline{\mathbf{B}} \rangle := \texttt{vec}(\underline{\mathbf{A}})^\top \texttt{vec}(\underline{\mathbf{B}})$. $|\cdot|$ denotes the absolute value of a scalar or the cardinality of a set, and $\circ$ denotes the function composition operation. For simplicity, let

$$\breve{\underline{\mathbf{T}}} := M(\underline{\mathbf{T}})$$

denote the tensor obtained by applying the M transform to $\underline{\mathbf{T}}$, as defined in Eq. (1). **In the following, we will frequently use $\breve{(\cdot)}$ and $\breve{(\cdot)}^{(i)}$ to represent $M(\cdot)$ and $M(\cdot)^{(i)}$, respectively. Here, $\breve{(\cdot)}$ denotes the tensor or t-embedding after the M transform, while $\breve{(\cdot)}^{(i)}$ refers to its i-th frontal slice, i.e., the i-th frequency component (or the i-th sub-domain).**

Concepts related to t-SVD. Due to space limitations, some concepts related to t-SVD were omitted in the main text. We provide additional explanations here, as these concepts are crucial for understanding the properties and operational rules of t-SVD.

Definition 4 (Frontal-slice-wise product [35])**.** *The frontal-slice-wise product of any two tensors $\underline{\mathbf{A}} \in \mathbb{R}^{m \times n \times K}$ and $\underline{\mathbf{B}} \in \mathbb{R}^{n \times k \times K}$, denoted by $\underline{\mathbf{A}} \odot \underline{\mathbf{B}}$, is defined as a tensor $\underline{\mathbf{T}}$ such that*

$$\underline{\mathbf{T}}(:,:,i) = \underline{\mathbf{A}}(:,:,i) \cdot \underline{\mathbf{B}}(:,:,i),\ i \in [K],$$

where $\cdot$ denotes the standard matrix multiplication. The frontal-slice-wise product performs matrix multiplication on each frontal slice of the tensors, resulting in a new tensor.

Definition 5 (M-block-diagonal matrix)**.** *The M-block-diagonal matrix of any tensor $\underline{\mathbf{T}} \in \mathbb{R}^{m \times n \times K}$, denoted by $\bar{\mathbf{T}}$, is the block diagonal matrix whose diagonal blocks are the frontal slices of $\breve{\underline{\mathbf{T}}} := M(\underline{\mathbf{T}})$:*

$$\begin{aligned} \bar{\mathbf{T}} &:= \texttt{bdiag}(\breve{\underline{\mathbf{T}}}) \\ &:= \begin{bmatrix} \breve{\underline{\mathbf{T}}}^{(1)} & & & \\ & \breve{\underline{\mathbf{T}}}^{(2)} & & \\ & & \ddots & \\ & & & \breve{\underline{\mathbf{T}}}^{(K)}) \end{bmatrix} \in \mathbb{R}^{mK \times nK}. \end{aligned}$$

This concept arranges the slices of a tensor in the frequency domain into a block diagonal matrix, facilitating the theoretical analysis of t-SVD.

We further provide some definitions and properties related to t-SVD:

Definition 6 ([25])**.** *The t-transpose of a tensor* $\underline{\mathbf{T}} \in \mathbb{R}^{m\times n\times K}$ *under the* M *transform (as shown in Eq. (1)), denoted by* $\underline{\mathbf{T}}^\top$*, satisfies*

$$M(\underline{\mathbf{T}}^\top)^{(i)} = \left(M(\underline{\mathbf{T}})^{(i)}\right)^\top,\ i \in [K].$$

In other words, the t-transpose performs a transpose on each slice in the frequency domain and then transforms back to the time domain. This operation is one of the foundations of t-SVD theory.

Definition 7 ([25])**.** *The t-identity tensor* $\underline{\mathbf{I}} \in \mathbb{R}^{m\times m\times K}$ *under the* M *transform satisfies that each frontal slice of* $M(\underline{\mathbf{I}})$ *is an* $K \times K$ *identity matrix, i.e.,*

$$M(\underline{\mathbf{I}})^{(i)} = \mathbf{I},\ i \in [K].$$

It is easy to verify that $\underline{\mathbf{T}} *_M \underline{\mathbf{I}} = \underline{\mathbf{T}}$ and $\underline{\mathbf{I}} *_M \underline{\mathbf{T}} = \underline{\mathbf{T}}$ hold for appropriate dimensions. The t-identity tensor plays a role similar to the identity matrix in t-SVD.

Definition 8 ([25])**.** *A tensor* $\underline{\mathbf{Q}} \in \mathbb{R}^{d\times d\times d_3}$ *is called t-orthogonal under the* M *transform if it satisfies*

$$\underline{\mathbf{Q}}^\top *_M \underline{\mathbf{Q}} = \underline{\mathbf{Q}} *_M \underline{\mathbf{Q}}^\top = \underline{\mathbf{I}}.$$

T-orthogonality is an important property of tensor transformations, ensuring that the inner product and norm of tensors remain invariant before and after the transformation.

Definition 9 ([26])**.** *A tensor is called f-diagonal if all its frontal slices are diagonal matrices.*

F-diagonal tensors play a central role in t-SVD decomposition, similar to the singular value matrix in matrix SVD.

We present some basic facts about t-product:

(a) The t-product of two t-scalars is commutative:

$$\underline{a} *_M \underline{b} = M^{-1}\left(M(\underline{a}) \odot M(\underline{b})\right) = M^{-1}\left(M(\underline{b}) \odot M(\underline{a})\right) = \underline{b} *_M \underline{a}.$$

(b) The ℓ_2-norm of a t-scalar is defined as

$$\|\underline{a}\|^2 = \|M(\underline{a})\|^2 = \|M(\underline{a}) \odot M(\underline{a})\|_1 = \|M(\underline{a} *_M \underline{a})\|_1. \tag{8}$$

(c) The square of the difference between two t-scalars satisfies $\|\underline{a} - \underline{b}\|^2 \le 2(\|\underline{a}\|^2 + \|\underline{b}\|^2)$.

(d) The square of the sum of three t-scalars satisfies $\|\underline{a} + \underline{b} + \underline{c}\|^2 \le 3(\|\underline{a}\|^2 + \|\underline{b}\|^2 + \|\underline{c}\|^2)$.

(e) For any $\underline{\mathbf{f}}, \underline{\mathbf{g}} \in \mathbb{R}^{m\times 1\times K}$, we can upper bound $\|\underline{\mathbf{f}}^\top *_M \underline{\mathbf{g}}\|^2$ by $\|\underline{\mathbf{f}}\|^2\|\underline{\mathbf{g}}\|^2$ due to:

$$\begin{aligned}
\|\underline{\mathbf{f}}^\top *_M \underline{\mathbf{g}}\|^2 &= \|M(\underline{\mathbf{f}}^\top *_M \underline{\mathbf{g}})\|^2 && \text{(Orthogonality of } \mathbf{M}\text{)}\\
&= \|\breve{\underline{\mathbf{f}}}^\top \odot \breve{\underline{\mathbf{g}}}\|^2 && \text{(Definition of } M(\cdot)\text{)}\\
&= \sum_{i=1}^{K}\left((\breve{\underline{\mathbf{f}}}^{(i)})^\top \breve{\underline{\mathbf{g}}}^{(i)}\right)^2\\
&\le \sum_{i=1}^{K}\left(\|\breve{\underline{\mathbf{f}}}^{(i)}\|\|\breve{\underline{\mathbf{g}}}^{(i)}\|\right)^2 && \text{(Cauchy-Schwartz inequality)}\\
&\le \left(\sum_{i=1}^{K}\|\breve{\underline{\mathbf{f}}}^{(i)}\|^2\right)\left(\sum_{i=1}^{K}\|\breve{\underline{\mathbf{g}}}^{(i)}\|^2\right)\\
&= \|\breve{\underline{\mathbf{f}}}\|^2\|\breve{\underline{\mathbf{g}}}\|^2 = \|\underline{\mathbf{f}}\|^2\|\underline{\mathbf{g}}\|^2. && \text{(Orthogonality of } \mathbf{M}\text{)}
\end{aligned}$$

The above inequalities characterize the properties of t-product from different perspectives and serve as the foundation for subsequent theoretical analyses.

Tensor Tubal-rank and tensor multi-rank as measures of low-tankness in the transformed domain. The tensor singular value decomposition (t-SVD) of $\underline{\mathbf{T}} \in \mathbb{R}^{m \times n \times K}$ under the transform M in Eq. (1) is given by:

$$\underline{\mathbf{T}} = \underline{\mathbf{U}} *_M \underline{\mathbf{S}} *_M \underline{\mathbf{V}}^\top, \tag{9}$$

where $\underline{\mathbf{U}} \in \mathbb{R}^{m \times m \times K}$ and $\underline{\mathbf{V}} \in \mathbb{R}^{n \times n \times K}$ are t-orthogonal tensors, $\underline{\mathbf{S}} \in \mathbb{R}^{m \times n \times K}$ is an f-diagonal tensor, and $(\cdot)^\top$ denotes the t-transpose.

According to Definition 2, the tubal rank of tensor $\underline{\mathbf{T}}$ is the number of non-zero tubes in $\underline{\mathbf{S}}$ in Eq. (2), i.e.,

$$r_{\text{t}}(\underline{\mathbf{T}}) := |\{i : \underline{\mathbf{S}}(i,i,:) \neq \mathbf{0}, i \leq \min\{m,n\}\}|.$$

The tubal rank is also equal to the maximum rank of the frontal slices of $\underline{\breve{\mathbf{T}}}$, i.e.,

$$r_{\text{t}}(\underline{\mathbf{T}}) = \max_i \{\text{rank}(M(\underline{\mathbf{T}})^{(i)})\},$$

which means that the tubal rank is a complexity measure of the low-rankness in the transformed domain. Within the context of t-SVD, we have another rank definition:

Definition 10 (Tensor multi-rank [60])**.** *The multi-rank of* $\underline{\mathbf{T}} \in \mathbb{R}^{m \times n \times K}$ *under the transform* M *in Eq.* (1) *is defined as the vector of matrix ranks of all the frontal slices of* $M(\underline{\mathbf{T}})$*, i.e.,*

$$\mathbf{r}_{\text{mul}}(\underline{\mathbf{T}}) := (\text{rank}(M(\underline{\mathbf{T}})^{(1)}), \ldots, \text{rank}(M(\underline{\mathbf{T}})^{(K)}))^\top \in \mathbb{N}^K.$$

The relationship between the tubal rank and multi-rank is given by

$$r_{\text{t}}(\underline{\mathbf{T}}) = \|\mathbf{r}_{\text{mul}}(\underline{\mathbf{T}})\|_\infty,$$

where $\|\cdot\|_\infty$ denotes the ℓ_∞-norm of a vector.

The multi-rank can be seen as a fine-grained low-rankness measure in the transformed domain compared to the tubal rank. It captures the notion of structural simplicity by assigning a rank value to each frontal slice of the transformed tensor. A low rank value for a frontal slice suggests that the corresponding frequency component of the data can be well-approximated by a low-dimensional subspace, indicating a strong correlation or pattern within that slice. Conversely, a high rank value implies that the corresponding frequency component is more complex and requires a higher-dimensional subspace to capture its information content. The multi-rank provides a more detailed characterization of the low-rank structure of a tensor in the transformed domain. While the tubal rank gives a single value representing the overall low-rankness of the tensor, the multi-rank offers a vector of rank values, each corresponding to a specific frontal slice. This allows for a more nuanced analysis of the tensor's complexity along its different frequency components.

B.2 Functional t-Singular Value Decomposition

B.2.1 Proof of the Ft-SVD Theorem

The Functional t-Singular Value Decomposition (Ft-SVD) in Theorem 1 is an extension of the traditional t-SVD framework to infinite and continuous feature domains. This theoretical approach enables the representation of data and functions defined on these domains while preserving the core properties of t-SVD. By generalizing t-SVD to functional settings, Ft-SVD facilitates the development of efficient and robust algorithms for learning vector-valued functions, addressing complex interdependencies among multiple outputs, and solving related problems in a unified and principled manner. We provide the proof of Theorem 1 as follows.

Proof of Theorem 1. The key idea of proving the Functional t-SVD is to define appropriate Hilbert spaces and linear operators, and then use the spectral theorem for compact operators to establish the existence and properties of the Functional t-SVD.

Step 1: Defining the t-Linear Operators. We first define a $\mathbb{R}^{1\times1\times K}$-valued inner product, referred to as the t-inner product, along with the corresponding t-linear operators. Let $L^2(\mathcal{X}; \mathbb{R}^{1\times1\times K})$ and $L^2(\mathcal{Y}; \mathbb{R}^{1\times1\times K})$ be the set of square-integrable vector-valued functions from $\mathcal{X}$ and $\mathcal{Y}$ to $\mathbb{R}^{1\times1\times K}$, respectively. The t-inner product is defined as:

$$\langle F, G\rangle_{L^2(\mathcal{X};\mathbb{R}^{1\times 1\times K})} := \int_{\mathcal{X}} F(x) *_M G(x)dx,$$

where $*_M$ denotes the t-product.

Define the t-linear operator $\mathcal{T} : L^2(\mathcal{Y};\mathbb{R}^{1\times 1\times K}) \to L^2(\mathcal{X};\mathbb{R}^{1\times 1\times K})$ as:

$$(\mathcal{T}G)(x) := \int_{\mathcal{Y}} F(x,y) *_M G(y)dy,$$

and its adjoint operator $\mathcal{T}^* : L^2(\mathcal{X};\mathbb{R}^{1\times 1\times K}) \to L^2(\mathcal{Y};\mathbb{R}^{1\times 1\times K})$ as:

$$(\mathcal{T}^*H)(y) := \int_{\mathcal{X}} F(x,y) *_M H(x)dx.$$

To show that $\mathcal{T}^*$ is indeed the adjoint of $\mathcal{T}$, we verify that $\langle \mathcal{T}G, H\rangle_{L^2(\mathcal{X};\mathbb{R}^{1\times 1\times K})} = \langle G, \mathcal{T}^*H\rangle_{L^2(\mathcal{Y};\mathbb{R}^{1\times 1\times K})}$:

$$\begin{aligned}
\langle \mathcal{T}G, H\rangle_{L^2(\mathcal{X};\mathbb{R}^{1\times 1\times K})} &= \int_{\mathcal{X}} (\mathcal{T}G)(x) *_M H(x)dx \\
&= \int_{\mathcal{X}} \left(\int_{\mathcal{Y}} F(x,y) *_M G(y)dy\right) *_M H(x)dx \\
&= \int_{\mathcal{Y}} G(y) *_M \left(\int_{\mathcal{X}} F(x,y) *_M H(x)dx\right) dy \\
&= \int_{\mathcal{Y}} G(y) *_M (\mathcal{T}^*H)(y)dy \\
&= \langle G, \mathcal{T}^*H\rangle_{L^2(\mathcal{Y};\mathbb{R}^{1\times 1\times K})},
\end{aligned}$$

where we used the commutativity of the t-product for t-scalars to interchange the order of integration and the t-product.

Then, we establish the self-adjointness of the operators $\mathcal{T}\mathcal{T}^*$ and $\mathcal{T}^*\mathcal{T}$. Consider the operators $\mathcal{T}\mathcal{T}^* : L^2(\mathcal{X};\mathbb{R}^{1\times 1\times K}) \to L^2(\mathcal{X};\mathbb{R}^{1\times 1\times K})$ and $\mathcal{T}^*\mathcal{T} : L^2(\mathcal{Y};\mathbb{R}^{1\times 1\times K}) \to L^2(\mathcal{Y};\mathbb{R}^{1\times 1\times K})$. To show that $\mathcal{T}\mathcal{T}^*$ is self-adjoint, we verify that $\langle \mathcal{T}\mathcal{T}^*F, G\rangle_{L^2(\mathcal{X};\mathbb{R}^{1\times 1\times K})} = \langle F, \mathcal{T}\mathcal{T}^*G\rangle_{L^2(\mathcal{X};\mathbb{R}^{1\times 1\times K})}$:

$$\langle \mathcal{T}\mathcal{T}^*F, G\rangle_{L^2(\mathcal{X};\mathbb{R}^{1\times 1\times K})} = \langle \mathcal{T}^*F, \mathcal{T}^*G\rangle_{L^2(\mathcal{Y};\mathbb{R}^{1\times 1\times K})} = \langle F, \mathcal{T}\mathcal{T}^*G\rangle_{L^2(\mathcal{X};\mathbb{R}^{1\times 1\times K})}, \tag{10}$$

where we used the adjoint property of $\mathcal{T}^*$ in the first equality. The self-adjointness of $\mathcal{T}^*\mathcal{T}$ can be shown similarly.

Step 2: Establishing Spectral Properties in the Transformed Domain. The next step is to transform the operators to the transformed domain defined by transform M in Eq. (1) and establish their spectral properties. Define $\breve{\mathcal{T}} := M(\mathcal{T}) : L^2(\mathcal{Y};\mathbb{R}^{1\times 1\times K}) \to L^2(\mathcal{X};\mathbb{R}^{1\times 1\times K})$ and $\breve{\mathcal{T}}^* := M(\mathcal{T}^*) : L^2(\mathcal{X};\mathbb{R}^{1\times 1\times K}) \to L^2(\mathcal{Y};\mathbb{R}^{1\times 1\times K})$ as the transformed operators for any vector-valued functions $G \in L^2(\mathcal{Y};\mathbb{R}^{1\times 1\times K})$ and $H \in L^2(\mathcal{X};\mathbb{R}^{1\times 1\times K})$, respectively:

$$M(\mathcal{T}G)(y) := M\Big((\mathcal{T}G)(y)\Big) \quad \text{and} \quad M(\mathcal{T}^*H)(x) := M\Big((\mathcal{T}^*H)(x)\Big),$$

where M is the invertible linear transform in Eq. (1). We similarly define operators $M(\mathcal{T}\mathcal{T}^*) : L^2(\mathcal{X};\mathbb{R}^{1\times 1\times K}) \to L^2(\mathcal{X};\mathbb{R}^{1\times 1\times K})$ and $M(\mathcal{T}^*\mathcal{T}) : L^2(\mathcal{Y};\mathbb{R}^{1\times 1\times K}) \to L^2(\mathcal{Y};\mathbb{R}^{1\times 1\times K})$ by

$$M(\mathcal{T}\mathcal{T}^*H)(x) := M\Big((\mathcal{T}\mathcal{T}^*H)(x)\Big) \quad \text{and} \quad M(\mathcal{T}^*\mathcal{T}G)(y) := M\Big((\mathcal{T}^*\mathcal{T}G)(y)\Big),$$

We proceed to show that:

$$\breve{\mathcal{T}}\breve{\mathcal{T}}^* = M(\mathcal{T}\mathcal{T}^*) \quad \text{and} \quad \breve{\mathcal{T}}^*\breve{\mathcal{T}} = M(\mathcal{T}^*\mathcal{T}).$$

To verify these identities, consider any $H \in L^2(\mathcal{X}; \mathbb{R}^{1\times1\times K})$:

$$\begin{aligned} M((\mathcal{T}\mathcal{T}^*)H)(x) &= M\big((\mathcal{T}\mathcal{T}^*)H(x)\big) \\ &= M\left(\int_{\mathcal{X}}\left(\int_{\mathcal{Y}} F(x,y) *_M F(x',y)\,dy\right) *_M H(x')\,dx'\right) \\ &\overset{(i)}{=} \int_{\mathcal{X}}\left(\int_{\mathcal{Y}} \breve{F}(x,y) \odot \breve{F}(x',y)\,dy\right) \odot \breve{H}(x')\,dx' \\ &= \big((\breve{\mathcal{T}}\breve{\mathcal{T}}^*)\breve{H}\big)(x), \end{aligned}$$

where $\odot$ denotes the frontal slice-wise product, and step (i) applies the definition of the t-product, with the notation $\breve{F}(x,y) = M(F(x,y))$. This verifies that $M(\mathcal{T}\mathcal{T}^*) = \breve{\mathcal{T}}\breve{\mathcal{T}}^*$. A similar approach confirms $M(\mathcal{T}^*\mathcal{T}) = \breve{\mathcal{T}}^*\breve{\mathcal{T}}$.

We now characterize the frequency-specific behavior of $\mathcal{T}\mathcal{T}^*$ and $\mathcal{T}^*\mathcal{T}$ by performing eigendecompositions on each of their K frequency "component" (i.e., the "frontal slices" of operators $\breve{\mathcal{T}}\breve{\mathcal{T}}^*$ and $\breve{\mathcal{T}}^*\breve{\mathcal{T}}$, respectively). Specifically, for any $H \in L^2(\mathcal{X}; \mathbb{R}^{1\times1\times K})$, we define the operator $M(\mathcal{T}\mathcal{T}^*)^{(i)} : L^2(\mathcal{X}; \mathbb{R}) \to L^2(\mathcal{X}; \mathbb{R})$ by

$$(M(\mathcal{T}\mathcal{T}^*)^{(i)}\breve{H}^{(i)})(x) := M\left((\mathcal{T}\mathcal{T}^* H)(x)\right)^{(i)},\ i \in [K]$$

as the "i-th frequency componen" of $\mathcal{T}\mathcal{T}^*$ in the transformed domain induced by transform $M(\cdot)$.

Thus, we have:

$$\left(M(\mathcal{T}\mathcal{T}^*)^{(i)}\breve{H}^{(i)}\right)(x) = \int_{\mathcal{X}}\int_{\mathcal{Y}} \breve{F}(x,y)^{(i)} \cdot \breve{F}^*(x',y)^{(i)} \cdot \breve{H}(x')^{(i)}\,dy\,dx',$$

where $\cdot$ denotes the standard multiplication.

Next, we show that $M(\mathcal{T}\mathcal{T}^*)^{(i)}$ is a compact and self-adjoint operator on the Hilbert space $L^2(\mathcal{X}; \mathbb{R})$.

(1) *Compactness*: Note that one condition of Theorem 1 is that the vector-valued function $F(x,y)$ is square-integrable, implying that $M(F(x,y)^{(i)})$ satisfies the Hilbert-Schmidt condition:

$$\begin{aligned} \int_{\mathcal{X}}\int_{\mathcal{Y}} M(F(x,y)^{(i)})^2 dxdy &\le \int_{\mathcal{X}}\int_{\mathcal{Y}} \sum_{i=1}^{K} M(F(x,y)^{(i)})^2 dxdy \\ &= \int_{\mathcal{X}}\int_{\mathcal{Y}} \|M(F(x,y))\|^2 dxdy \\ &\overset{(i)}{=} \int_{\mathcal{X}}\int_{\mathcal{Y}} \|F(x,y)\|^2 dxdy \\ &\overset{(ii)}{<} \infty \end{aligned}$$

where $\|\cdot\|$ denotes the ℓ_2-norm of t-scalars (i.e., vectors), *(i)* holds since $\mathbf{M}$ is an orthogonal matrix in Eq. (1), and *(ii)* is due to the square-integrability of $F(x,y)$.

(2) *Self-adjointness*: It can be readily verified that $M(\mathcal{T}\mathcal{T}^*)^{(i)}$ is also a self-adjoint operator on the Hilbert space $L^2(\mathcal{X}; \mathbb{R})$, following equations analogous to Eq. (10), and thus we omit it.

Similarly, the operators $M(\mathcal{T}^*\mathcal{T})^{(i)}$ also satisfies compactness and self-adjointness.

Then, by the spectral theory of compact linear operator in Hilbert spaces [56], there exist orthonormal eigenfunctions $\{u_{i,j}\}_{j=1}^{\infty}$ and $\{v_{i,j}\}_{j=1}^{\infty}$ and non-descending non-negative eigenvalues $\{\omega_{i,j}\}_{j=1}^{\infty}$ such that:

$$M(\mathcal{T}\mathcal{T}^*)^{(i)}u_{i,j} = \omega_{i,j}u_{i,j} \quad \text{and} \quad M(\mathcal{T}^*\mathcal{T})^{(i)}v_{i,j} = \omega_{i,j}v_{i,j} \quad \forall (i,j) \in [K] \times \mathbb{N}. \tag{11}$$

Then, for each $i \in [K]$, we can express $\breve{F}(x,y)^{(i)}$ using the eigenfunctions and eigenvalues as follows:

$$\breve{F}(x,y)^{(i)} = \sum_{j=1}^{\infty} \omega_{i,j}^{1/2} \cdot u_{i,j}(x) \cdot v_{i,j}(y), \quad \forall i \in [K],$$

which is equivalent to the slice-wise product formulation:

$$M(F)(x,y) = \sum_{j=1}^{\infty} \underline{\breve{\varphi}}_j(x) \odot \underline{\breve{\varsigma}}_j \odot \underline{\breve{\psi}}_j(y),$$

where $\underline{\breve{\varphi}}_j(x)$, $\underline{\breve{\varsigma}}_j$, $\underline{\breve{\psi}}_j(y) \in \mathbb{R}^{1\times 1\times K}$ satisfy

$$\underline{\breve{\varphi}}_j(x)^{(i)} = u_{i,j}(x), \quad \underline{\breve{\varsigma}}_j^{(i)} = \omega_{i,j}^{1/2}, \quad \text{and} \quad \underline{\breve{\psi}}_j(y)^{(i)} = v_{i,j}(y), \ \ \forall (i,j) \in [K] \times \mathbb{N}.$$

Step 3: Transforming Back to the Original Domain. In this final step, we apply the inverse transform M^{-1} from Eq. (1) to return to the original domain and construct the Functional t-SVD of F. The decomposition of F is given by:

$$F(x,y) = \sum_{j=1}^{\infty} \underline{\phi}_j(x) *_M \underline{\sigma}_j *_M \underline{\psi}_j(y), \tag{12}$$

where the t-singular values and t-singular functions are defined as follows:

$$\underline{\sigma}_j = M^{-1}(\underline{\breve{\varsigma}}_j), \quad \underline{\phi}_j(x) = M^{-1}(\underline{\breve{\varphi}}_j(x)), \quad \text{and} \quad \underline{\psi}_j(y) = M^{-1}(\underline{\breve{\psi}}_j(y)).$$

The orthonormality conditions

$$\int_{\mathcal{X}} \underline{\phi}_i(x) *_M \underline{\phi}_j(x) dx = \delta_{ij} M^{-1}(\underline{1}), \quad \int_{\mathcal{Y}} \underline{\psi}_i(y) *_M \underline{\psi}_j(y) dy = \delta_{ij} M^{-1}(\underline{1})$$

result from the orthonormality of $\{u_{i,j}\}_{j=1}^{\infty}$ and $\{v_{i,j}\}_{j=1}^{\infty}$ (from Eq. (11)) and the properties of the transform $M(\cdot)$ specified in Eq. (1). This concludes the proof.

□

B.2.2 Proof of the r-Term Truncated Ft-SVD Theorem

Theorem 2 provides an extent of the Eckart-Young theorem for t-SVD [27] to the proposed Ft-SVD framework. It highlights the optimality of the truncated Ft-SVD in approximating vector-valued functions, making it a valuable tool for applications such as function compression, denoising, and other tasks that benefit from low-rank approximations of functions. The proof is given as follows.

Proof of Theorem 2. Let $G \in L^2(\mathcal{X} \times \mathcal{Y}; \mathbb{R}^{1\times 1\times K})$ be an arbitrary r-term t-product expansion:

$$G(x,y) = \sum_{i=1}^{r} a_i(x) *_M \underline{\lambda}_i *_M b_i(y), \tag{13}$$

where $a_i \in L^2(\mathcal{X}; \mathbb{R}^{1\times 1\times K}), b_i \in L^2(\mathcal{Y}; \mathbb{R}^{1\times 1\times K}), \underline{\lambda}_i \in \mathbb{R}^{1\times 1\times K}$.

Our goal is to show that

$$\|F - F_r\|_{L^2(\mathcal{X}\times\mathcal{Y};\mathbb{R}^{1\times 1\times K})} \leq \|F - G\|_{L^2(\mathcal{X}\times\mathcal{Y};\mathbb{R}^{1\times 1\times K})}$$

for all $G \in \mathcal{K}_r$, where $\mathcal{K}_r$ is the set of all r-term t-product expansions.

Recall that $\breve{F}^{(i)} = M(F)(:,:,i)$ denotes the i-th ($i \in [K]$) frequency component of F in the transformed domain induced by Eq. (1). According to the construction of the functional t-singular value decomposition in the proof of Theorem 1, we have:

$$\breve{F}^{(i)}(x,y)=\sum_{j=1}^{\infty}\underline{\breve{\phi}}_j^{(i)}(x)\cdot\underline{\breve{\sigma}}_j^{(i)}\cdot\underline{\breve{\psi}}_j^{(i)}(y),\quad \breve{G}^{(i)}(x,y)=\sum_{j=1}^{r}\breve{a}_j^{(i)}(x)\cdot\underline{\breve{\lambda}}_j^{(i)}\cdot\breve{b}_j^{(i)}(y),$$

where the sequences $(\underline{\breve{\sigma}}_j^{(i)})$ and $(\underline{\breve{\lambda}}_j^{(i)})$ are in non-ascending order, and the sets $\{\underline{\breve{\phi}}_j^{(i)}\}_{j=1}^{\infty}, \{\underline{\breve{\psi}}_j^{(i)}\}_{j=1}^{\infty}, \{\breve{a}_j^{(i)}\}_{j=1}^{r}$, and $\{\breve{b}_j^{(i)}\}_{j=1}^{r}$ are orthonormal bases in their respective spaces:

$$\begin{aligned}\int_{\mathcal{X}}\underline{\breve{\phi}}_j^{(i)}(x)\underline{\breve{\phi}}_k^{(i)}(x)dx=\delta_{jk},&\quad \int_{\mathcal{Y}}\underline{\breve{\psi}}_j^{(i)}(y)\underline{\breve{\psi}}_k^{(i)}(y)dy=\delta_{jk},\forall(j,k)\in\mathbb{N}^2,\\ \int_{\mathcal{X}}\breve{a}_j^{(i)}(x)\breve{a}_k^{(i)}(x)dx=\delta_{jk},&\quad \int_{\mathcal{Y}}\breve{b}_j^{(i)}(y)\breve{b}_k^{(i)}(y)dy=\delta_{jk},\forall(j,k)\in[r]^2.\end{aligned}$$

Now, let's consider the approximation error:

$$\begin{aligned}\|F-G\|^2_{L^2(\mathcal{X}\times\mathcal{Y};\mathbb{R}^{1\times1\times K})}&=\int_{\mathcal{X}}\int_{\mathcal{Y}}\|F(x,y)-G(x,y)\|_2^2dydx\\ &=\int_{\mathcal{X}}\int_{\mathcal{Y}}\|M(F(x,y)-G(x,y))\|_2^2dydx\\ &=\sum_{i=1}^{K}\int_{\mathcal{X}}\int_{\mathcal{Y}}|\breve{F}^{(i)}(x,y)-\breve{G}^{(i)}(x,y)|^2dydx\\ &=\sum_{i=1}^{K}\|\breve{F}^{(i)}-\breve{G}^{(i)}\|^2_{L^2(\mathcal{X}\times\mathcal{Y};\mathbb{R})}\\ &\geq\sum_{i=1}^{K}\|\sum_{j=1}^{\infty}\underline{\breve{\phi}}_j^{(i)}\cdot\underline{\breve{\sigma}}_j^{(i)}\cdot\underline{\breve{\psi}}_j^{(i)}-\sum_{j=1}^{r}\underline{\breve{\phi}}_j^{(i)}\cdot\underline{\breve{\sigma}}_j^{(i)}\cdot\underline{\breve{\psi}}_j^{(i)}\|^2_{L^2(\mathcal{X}\times\mathcal{Y};\mathbb{R})}\\ &=\sum_{i=1}^{K}\|\sum_{j=r+1}^{\infty}\underline{\breve{\phi}}_j^{(i)}\cdot\underline{\breve{\sigma}}_j^{(i)}\cdot\underline{\breve{\psi}}_j^{(i)}\|^2_{L^2(\mathcal{X}\times\mathcal{Y};\mathbb{R})}\\ &=\sum_{i=1}^{K}\sum_{j=r+1}^{\infty}(\underline{\breve{\sigma}}_j^{(i)})^2=\sum_{j=r+1}^{\infty}\|\underline{\sigma}_j\|_2^2.\end{aligned}$$

The last equality holds due to the orthonormality of the bases $\{\underline{\breve{\phi}}_j^{(i)}\}_{j=1}^{\infty}$ and $\{\underline{\breve{\psi}}_j^{(i)}\}_{j=1}^{\infty}$. The inequality in the sixth line becomes an equality when

$$\begin{aligned}&\underline{\breve{\phi}}_j^{(i)}(x)=\breve{a}_j^{(i)}(x),\quad \underline{\breve{\sigma}}_j^{(i)}=\underline{\breve{\lambda}}_j^{(i)},\quad \underline{\breve{\psi}}_j^{(i)}(y)=\breve{b}_j^{(i)}(y),\quad \forall j\in[r],i\in[K],\\ \Rightarrow\quad &\underline{\phi}_j(x)=a_j(x),\quad \underline{\sigma}_j=\underline{\lambda}_j,\quad \underline{\psi}_j(y)=b_j(y),\quad \forall j\in[r].\end{aligned}$$

This choice of $a_j, \underline{\lambda}_j$, and b_j corresponds to the best L^2-approximation by finite sums of functions with separable variables [48, 47].

Therefore, we have shown that $\|F-F_r\|^2_{L^2(\mathcal{X}\times\mathcal{Y};\mathbb{R}^{1\times1\times K})}=\sum_{j=r+1}^{\infty}\|\underline{\sigma}_j\|_2^2$, which is the minimum approximation error among all r-term t-product expansions. This completes the proof. □

B.2.3 Smooth Functions Exhibit Favorable Low-rank Approximability under Ft-SVD

The Sobolev spaces are a family of function spaces that play a crucial role in the theory of partial differential equations, signal processing and machine learning [3, 4, 29, 42]. They are defined by combining conditions on the function itself and its derivatives up to a certain order.

In our settings, let $\mathcal{Y}\subset\mathbb{R}^{D_2}$ be a domain satisfying the strong local Lipschitz condition, and let $f:\mathcal{X}\times\mathcal{Y}\to\mathbb{R}^{1\times1\times K}$ be a vector-valued function with components $f^{(i)}(x,y)$, $i=1,\ldots,K$.

Suppose there exists a constant $s > 0$ such that $f^{(i)} \in L^2(\mathcal{X}, H^s(\mathcal{Y}))$ for all i, where $H^s(\mathcal{Y})$ denotes the s-order Sobolev space on $\mathcal{Y}$. The function class $L^2(\mathcal{X}, H^s(\mathcal{Y}))$ is a type of mixed-norm Sobolev space that arises when dealing with functions defined on a product domain $\mathcal{X} \times \mathcal{Y}$, where $\mathcal{X}$ is a domain in $\mathbb{R}^{D_1}$ and $\mathcal{Y}$ is a domain in $\mathbb{R}^{D_2}$ satisfying the strong local Lipschitz condition. The strong local Lipschitz condition on a domain $\mathcal{Y} \subset \mathbb{R}^{D_2}$ requires that, locally around each boundary point $y_0 \in \partial\mathcal{Y}$, there exists a Lipschitz parametrization of the boundary, ensuring a certain level of regularity and avoiding pathological features like cusps, which is crucial for establishing various results in the theory of Sobolev spaces.

For a vector-valued function $f : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}^{1\times 1\times K}$ with components $f^{(i)}(x, y)$, $i = 1, \dots, K$, the condition $f^{(i)} \in L^2(\mathcal{X}, H^s(\mathcal{Y}))$ means that each component function $f^{(i)}$ belongs to the mixed-norm Sobolev space $L^2(\mathcal{X}, H^s(\mathcal{Y}))$. The space $L^2(\mathcal{X}, H^s(\mathcal{Y}))$ is defined as the set of all vector-valued functions $g : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}^{1\times 1\times K}$ with components $g^{(i)}(x, y)$, such that for each $i = 1, \dots, K$, the following norm is finite:

$$\|g^{(i)}\|_{L^2(\mathcal{X},H^s(\mathcal{Y}))} := \left(\int_{\mathcal{X}} \|g^{(i)}(x, \cdot)\|^2_{H^s(\mathcal{Y})}\, dx \right)^{1/2},$$

where $\|g^{(i)}(x, \cdot)\|_{H^s(\mathcal{Y})}$ denotes the Sobolev norm of order s for the function $g^{(i)}(x, \cdot)$ considered as a function of the y variable alone, with x treated as a parameter. More precisely, the Sobolev norm $\|g^{(i)}(x, \cdot)\|_{H^s(\mathcal{Y})}$ is defined as:

$$\|g^{(i)}(x, \cdot)\|_{H^s(\mathcal{Y})} := \left(\sum_{|\alpha|\le s} \int_{\mathcal{Y}} |D^\alpha_y g^{(i)}(x, y)|^2\, dy \right)^{1/2},$$

where D^α_y denotes the weak (or distributional) derivative of order α with respect to the y variable, and the sum is taken over all multi-indices $\alpha = (\alpha_1, \alpha_2, \dots, \alpha_{D_2})$ with $|\alpha| = \alpha_1 + \alpha_2 + \cdots + \alpha_{D_2} \le s$.

Theorem 3 quantifies the relationship between the smoothness of a vector-valued function and the decay of its t-singular values, as well as the approximation error of its truncated t-SVD. It provides a theoretical foundation for using t-SVD to approximate smooth vector-valued functions and highlights the role of the function's smoothness and the domain's dimensionality in the approximation accuracy. The proof of Theorem 3 is given as follows:

Proof of Theorem 3. Let $f : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}^{1\times 1\times K}$ be a vector-valued function with components $f^{(i)}(x, y)$, $i = 1, \dots, K$. The condition $f^{(i)} \in L^2(\mathcal{X}, H^s(\mathcal{Y}))$ implies that each component function $f^{(i)}$ belongs to the mixed-norm Sobolev space $L^2(\mathcal{X}, H^s(\mathcal{Y}))$.

Consider the i-th frequency component $\breve{f}^{(i)}(x, y)$ of f in the transformed domain induced by the transform M defined in Eq. (1). By the definition of $M(\cdot)$, we have

$$\breve{f}^{(i)}(x, y) = \sum_{k=1}^{K} m_{ik} f^{(k)}(x, y), \tag{14}$$

which is a linear combination of the component functions $f^{(k)}(x, y)$ in the original domain. Since each $f^{(k)} \in L^2(\mathcal{X}, H^s(\mathcal{Y}))$, it follows that $\breve{f}^{(i)} \in L^2(\mathcal{X}, H^s(\mathcal{Y}))$ as well.

Next, we apply Theorem 3.2 from Ref. [21] to the function $\breve{f}^{(i)}$. This theorem states that for a function $g \in L^2(\mathcal{X}, H^s(\mathcal{Y}))$, the singular values $\sigma_j(g)$ satisfy the decay estimate

$$\sigma^2_j(g) \le \mathtt{diam}(\mathcal{Y})^{2s} C_{ext}(\mathcal{Y}, s) C_{em}(D_2, s) \cdot \|g\|_{L^2(\mathcal{X},H^s(\mathcal{Y}))} \cdot j^{-1-\frac{2s}{D_2}}, \tag{15}$$

where $C_{ext}(\mathcal{Y}, s)$ is the extension constant that depends on $\mathcal{Y}$ and s only, and $C_{em}(D_2, s)$ is the embedding constant from $\ell_{D_2/(D_2+2s),1}$ to $\ell_{D_2/(D_2+2s),\infty}$ given by

$$C_{em}(D_2, s) = 2^{D_2/2}\pi^{-s}\Gamma(s + 1/2)/\Gamma(1 - s), \tag{16}$$

with $\Gamma(\cdot)$ denoting the Gamma function.

Applying the estimate (15) to each frequency component $\breve{f}^{(i)}$, we obtain

$$\sigma_j^2(\breve{f}^{(i)}) \leq \mathtt{diam}(\mathcal{Y})^{2s} C_{ext}(\mathcal{Y}, s) C_{em}(D_2, s) \cdot \|\breve{f}^{(i)}\|_{L^2(\mathcal{X}, H^s(\mathcal{Y}))} \cdot j^{-1-\frac{2s}{D_2}}, \quad \forall i \in [K], j \in \mathbb{N}. \tag{17}$$

Now, recall from the construction of the functional t-SVD in the proof of Theorem 1 that the t-scalars $\underline{\sigma}_j$ satisfy

$$\|\underline{\sigma}_j\|^2 = \sum_{i=1}^{K} \sigma_j^2(\breve{f}^{(i)}).$$

Combining this with the previous estimate, we deduce that

$$\begin{aligned}\|\underline{\sigma}_j\|^2 &= \sum_{i=1}^{K} \sigma_j^2(\breve{f}^{(i)}) \\ &\leq \mathtt{diam}(\mathcal{Y})^{2s} C_{ext}(\mathcal{Y}, s) C_{em}(D_2, s) \cdot \sum_{i=1}^{K} \|\breve{f}^{(i)}\|_{L^2(\mathcal{X}, H^s(\mathcal{Y}))} \cdot j^{-1-\frac{2s}{D_2}}.\end{aligned}$$

Finally, according to Theorem 2, the rank-r approximation error of f can be bounded as follows:

$$\begin{aligned}&\min_{\tilde{f} \in \mathcal{F}_r} \|f - \tilde{f}\|^2_{L^2(\mathcal{X} \times H^s(\mathcal{Y}))} \\ &= \sum_{j=r+1}^{\infty} \|\underline{\sigma}_j\|^2 \\ &\leq \sum_{j=r+1}^{\infty} \mathtt{diam}(\mathcal{Y})^{2s} C_{ext}(\mathcal{Y}, s) C_{em}(D_2, s) \cdot \sum_{i=1}^{K} \|\breve{f}^{(i)}\|_{L^2(\mathcal{X}, H^s(\mathcal{Y}))} \cdot j^{-1-\frac{2s}{D_2}} \\ &\leq \sum_{j=r+1}^{\infty} \mathtt{diam}(\mathcal{Y})^{2s} C_{ext}(\mathcal{Y}, s) C_{em}(D_2, s) \cdot \sum_{i=1}^{K} \|\breve{f}^{(i)}\|_{L^2(\mathcal{X}, H^s(\mathcal{Y}))} \frac{D_2}{2s} (r+1)^{-\frac{2s}{D_2}} \\ &= \mathtt{diam}(\mathcal{Y})^{2s} C_{ext}(\mathcal{Y}, s) C_{em}(D_2, s) \cdot \sum_{i=1}^{K} \|\breve{f}^{(i)}\|_{L^2(\mathcal{X}, H^s(\mathcal{Y}))} \frac{D_2}{2s} (r+1)^{-\frac{2s}{D_2}},\end{aligned}$$

which completes the proof. □

C Additional Explanations of Ft-SVD Framework for MOR under CDS

C.1 More Explanations about the Hilbert t-Module, t-Bilinear Embeddings, and Assumptions

C.1.1 About the Definition of Hilbert t-Module

Definition 11 (Ring of K-dimensional vectors with t-product)**.** *Let $\mathcal{R} := \mathbb{R}^{1\times 1\times K}$ be the set of all real t-scalars of K-dimensionality which is isomorphic to $\mathbb{R}^K$. For any $\underline{a}, \underline{b} \in \mathcal{R}$, we define their t-product as:*

$$\underline{a} *_M \underline{b} = M^{-1}(M(\underline{a}) \odot M(\underline{b}))$$

where $M : \mathcal{R} \to \mathcal{R}$ is the linear transform induced by a given $K \times K$ orthogonal matrix $\mathbf{M}$ in Eq. (1), and the slice-wise-product $\odot$ degenerates to the element-wise product here.

$(\mathcal{R}, +, *_M)$ forms a ring with the t-product and standard vector addition, which we call the ring of K-dimensional real vectors with t-product. Specifically, for any $\underline{a}, \underline{b}, \underline{c} \in \mathcal{R}$ and $\alpha, \beta \in \mathbb{R}$, the following properties hold:

1. *(Associativity of addition) $(\underline{a} + \underline{b}) + \underline{c} = \underline{a} + (\underline{b} + \underline{c})$.*
2. *(Commutativity of addition) $\underline{a} + \underline{b} = \underline{b} + \underline{a}$.*
3. *(Additive identity) There exists an element $\mathbf{0} \in \mathcal{R}$ such that $\underline{a} + \mathbf{0} = \underline{a}$ for all $\underline{a} \in \mathcal{R}$.*
4. *(Additive inverses) For each $\underline{a} \in \mathcal{R}$, there exists an element $-\underline{a} \in \mathcal{R}$ such that $\underline{a} + (-\underline{a}) = \mathbf{0}$.*
5. *(Associativity of multiplication) $(\underline{a} *_M \underline{b}) *_M \underline{c} = \underline{a} *_M (\underline{b} *_M \underline{c})$.*
6. *(Distributivity of multiplication over addition) $\underline{a} *_M (\underline{b} + \underline{c}) = \underline{a} *_M \underline{b} + \underline{a} *_M \underline{c}$ and $(\underline{a} + \underline{b}) *_M \underline{c} = \underline{a} *_M \underline{c} + \underline{b} *_M \underline{c}$.*
7. *(Multiplicative identity) There exists an element $\mathbf{1} \in \mathcal{R}$ such that $\underline{a} *_M \mathbf{1} = \underline{a}$ for all $\underline{a} \in \mathcal{R}$.*
8. *(Scalar multiplication) $(\alpha\beta)\underline{a} = \alpha(\beta\underline{a})$, $\alpha(\underline{a} + \underline{b}) = \alpha\underline{a} + \alpha\underline{b}$, and $(\alpha + \beta)\underline{a} = \alpha\underline{a} + \beta\underline{a}$.*

Definition 12 (Hilbert t-Module over $\mathcal{R}$)**.** *Let $\mathcal{R}$ be the ring of K-dimensional real vectors with t-product as defined above. A Hilbert t-Module over $\mathcal{R}$ is a module $\mathcal{M}$ over $\mathcal{R}$ equipped with an $\mathcal{R}$-valued inner product $\langle \cdot, \cdot \rangle_{\mathcal{M}} : \mathcal{M} \times \mathcal{M} \to \mathcal{R}$ satisfying the following conditions:*

1. *(Conjugate symmetry) $\langle \underline{\mathbf{f}}, \underline{\mathbf{g}} \rangle_{\mathcal{M}} = \overline{\langle \underline{\mathbf{g}}, \underline{\mathbf{f}} \rangle_{\mathcal{M}}}$ for all $\underline{\mathbf{f}}, \underline{\mathbf{g}} \in \mathcal{M}$, where the overline denotes the element-wise complex conjugate.*[13]
2. *(Linearity in the second argument) $\langle \underline{\mathbf{f}}, \underline{\mathbf{g}}_1 + \underline{\mathbf{g}}_2 \rangle_{\mathcal{M}} = \langle \underline{\mathbf{f}}, \underline{\mathbf{g}}_1 \rangle_{\mathcal{M}} + \langle \underline{\mathbf{f}}, \underline{\mathbf{g}}_2 \rangle_{\mathcal{M}}$ and $\langle \underline{\mathbf{f}}, \underline{\mathbf{g}} *_M \underline{a} \rangle_{\mathcal{M}} = \langle \underline{\mathbf{f}}, \underline{\mathbf{g}} \rangle_{\mathcal{M}} *_M \underline{a}$ for all $\underline{\mathbf{f}}, \underline{\mathbf{g}}, \underline{\mathbf{g}}_1, \underline{\mathbf{g}}_2 \in \mathcal{M}$ and $\underline{a} \in \mathcal{R}$.*
3. *(Positivity) $\langle \underline{\mathbf{f}}, \underline{\mathbf{f}} \rangle_{\mathcal{M}} \geq 0$ (element-wise) for all $\underline{\mathbf{f}} \in \mathcal{M}$, with equality if and only if $\underline{\mathbf{f}} = 0$.*

   The Hilbert t-Module $\mathcal{M}$ is required to be complete with respect to the ℓ_2-norm induced by the $\mathcal{R}$-valued inner product:

$$\|\underline{\mathbf{f}}\|_{\mathcal{M}} := \|\underline{\mathbf{f}}\|_{\ell_2} = \sqrt{\sum_{i=1}^{K} (\langle \underline{\mathbf{f}}, \underline{\mathbf{f}} \rangle_{\mathcal{M}})_i^2},$$

   where $(\langle \underline{\mathbf{f}}, \underline{\mathbf{f}} \rangle_{\mathcal{M}})_i$ denotes the i-th component of the vector $\langle \underline{\mathbf{f}}, \underline{\mathbf{f}} \rangle_{\mathcal{M}}$.

Remark. *The ℓ_2-norm used here differs from the traditional C*-algebra norm in the definition of Hilbert C*-modules, as it does not satisfy the C*-equality $\|\underline{\mathbf{f}}\|^2 = \|\langle \underline{\mathbf{f}}, \underline{\mathbf{f}} \rangle_{\mathcal{M}}\|_{\mathcal{R}}$ [25, 8]. Despite this difference, we still refer to this structure as a Hilbert t-Module to emphasize its similarities with Hilbert C*-modules.*

[13] In the case of real Hilbert t-Modules, this condition reduces to the usual symmetry property $\langle \underline{\mathbf{f}}, \underline{\mathbf{g}} \rangle_{\mathcal{M}} = \langle \underline{\mathbf{g}}, \underline{\mathbf{f}} \rangle_{\mathcal{M}}$.

The benefits of using ℓ_2-norm for multi-output regression instead of the standard C* norm The choice of the ℓ_2-norm for Hilbert t-modules in the context of multi-output regression offers several advantages over the standard C* norm:

1. Geometric interpretation and learning theory: The ℓ_2-norm, as the Euclidean distance between vectors, provides a natural and intuitive way to measure the similarity between vector-valued functions. This geometric interpretation aligns well with the learning theory for vector-valued functions, where the goal is often to find a function that minimizes the expected risk or empirical risk, which are typically defined using the ℓ_2-norm.
2. Computational efficiency and scalability: The ℓ_2-norm is computationally more efficient than the C* norm, especially for high-dimensional vector-valued functions. This computational advantage is crucial for large-scale multi-output regression problems, where the number of outputs and the dimensionality of the input space can be large. The ℓ_2-norm allows for the development of more scalable and efficient learning algorithms.
3. Theoretical foundations and connections: The ℓ_2-norm is closely related to the theory of Hilbert spaces, which provides a rich set of tools and results for studying vector-valued functions. By using the ℓ_2-norm, Hilbert t-modules can leverage the well-established theory of Hilbert spaces, including concepts such as orthogonality, projection, and spectral decomposition. These theoretical foundations can guide the design and analysis of multi-output regression algorithms.
4. Practical applications and existing methods: The ℓ_2-norm is widely used in various practical applications, such as signal processing, computer vision, and control systems. Many existing multi-output regression methods, such as multi-task learning, multi-label classification, and multi-target regression, are based on the ℓ_2-norm. By adopting the ℓ_2-norm, Hilbert t-modules can readily benefit from and contribute to these existing methods and applications.

Although the ℓ_2-norm may not satisfy all the properties of the standard C* norm, it offers significant benefits for multi-output regression. The use of the ℓ_2-norm in Hilbert t-modules opens up new opportunities for the development of efficient, scalable, and theoretically grounded multi-output regression methods. It also facilitates the integration of multi-output regression with other areas of machine learning and signal processing, where the ℓ_2-norm is widely used. The name "Hilbert t-module" highlights the connection to the well-established theory of Hilbert spaces while acknowledging the specific choice of the ℓ_2-norm and the potential for further generalization and exploration in the context of multi-output regression.

Now we have a Hilbert t-Module $\mathcal{M}$ over the ring $\mathcal{R}$ of K-dimensional real vectors with t-product. The elements of $\mathcal{M}$ can be thought of as vector-valued functions or tensors, and we consider a linear transform M induced by an orthogonal matrix **M** that allows us to represent these elements in the frequency domain. We also define the associated Hilbert space $\mathcal{H}_j$ as follows:

Definition 13 (Hilbert space $\mathcal{H}_j$)**.** *For each $j = 1, \dots, K$, let $\mathcal{H}_j$ be a Hilbert space over the field of real numbers $\mathbb{R}$. We assume that for any element $\underline{\mathbf{f}} \in \mathcal{M}$, its j-th frequency component $\breve{\underline{\mathbf{f}}}^{(i)} := M(\underline{\mathbf{f}})^{(i)}$ belongs to $\mathcal{H}_j$, i.e., $\breve{\underline{\mathbf{f}}}^{(i)} \in \mathcal{H}_j$. The Hilbert space $\mathcal{H}_j$ is equipped with an inner product $\langle \cdot, \cdot \rangle_{\mathcal{H}_j} : \mathcal{H}_j \times \mathcal{H}_j \to \mathbb{R}$ satisfying the following properties:*

1. (Conjugate symmetry) $\langle \mathbf{x}, \mathbf{y} \rangle_{\mathcal{H}_j} = \langle \mathbf{y}, \mathbf{x} \rangle_{\mathcal{H}_j}$ for all $\mathbf{x}, \mathbf{y} \in \mathcal{H}_j$.

2. (Linearity in the second argument) $\langle \mathbf{x}, \mathbf{y}_1 + \mathbf{y}_2 \rangle_{\mathcal{H}_j} = \langle \mathbf{x}, \mathbf{y}_1 \rangle_{\mathcal{H}_j} + \langle \mathbf{x}, \mathbf{y}_2 \rangle_{\mathcal{H}_j}$ and $\langle \mathbf{x}, \alpha\mathbf{y} \rangle_{\mathcal{H}_j} = \alpha \langle \mathbf{x}, \mathbf{y} \rangle_{\mathcal{H}_j}$ for all $\mathbf{x}, \mathbf{y}, \mathbf{y}_1, \mathbf{y}_2 \in \mathcal{H}_j$ and $\alpha \in \mathbb{R}$.

3. (Positivity) $\langle \mathbf{x}, \mathbf{x} \rangle_{\mathcal{H}_j} \geq 0$ for all $\mathbf{x} \in \mathcal{H}_j$, with equality if and only if $\mathbf{x} = 0$.

The Hilbert space $\mathcal{H}_j$ is complete with respect to the norm induced by the inner product: $\|\mathbf{x}\|_{\mathcal{H}_j} = \sqrt{\langle \mathbf{x}, \mathbf{x} \rangle_{\mathcal{H}_j}}$ for all $\mathbf{x} \in \mathcal{H}_j$.

In essence, $\mathcal{H}_j$ is a standard Hilbert space associated with the j-th frequency component of the elements in the Hilbert t-Module $\mathcal{M}$. The inner product and norm in $\mathcal{H}_j$ satisfy the usual properties of a Hilbert space, and the completeness ensures that limits of Cauchy sequences in $\mathcal{H}_j$ also belong to $\mathcal{H}_j$.

This definition allows us to work with the frequency components of the elements in the Hilbert t-Module using the tools and properties of Hilbert spaces. In particular, it enables us to apply the

Cauchy-Schwarz inequality in each $\mathcal{H}_j$, which is a key step in proving the inequality between the norm of the inner product and the product of the norms in the Hilbert t-Module.

We also have a Cauchy-Schwarz-like inequality in Hilbert t-Modules as follows.

Lemma C.1 (Cauchy-Schwarz inequality in Hilbert t-Modules)**.** *It holds for any* $\underline{\mathbf{f}}, \underline{\mathbf{g}} \in \mathcal{M}$ *that*

$$\|\langle \underline{\mathbf{f}}, \underline{\mathbf{g}} \rangle_{\mathcal{M}}\|^2 \leq \|\underline{\mathbf{f}}\|_{\mathcal{M}}^2 \cdot \|\underline{\mathbf{g}}\|_{\mathcal{M}}^2.$$

Proof. Let $\underline{\mathbf{f}}, \underline{\mathbf{g}} \in \mathcal{M}$ be two elements of the Hilbert t-Module. We begin by expanding the left-hand side of the inequality:

$$\begin{aligned}
&\|\langle \underline{\mathbf{f}}, \underline{\mathbf{g}} \rangle_{\mathcal{M}}\|^2 \\
&= \sum_{i=1}^{K} \left((\langle \underline{\mathbf{f}}, \underline{\mathbf{g}} \rangle_{\mathcal{M}})_i \right)^2 && \text{(Definition of the } \ell_2\text{-norm on } \mathcal{R}\text{)} \\
&= \sum_{i=1}^{K} \left(\sum_{j=1}^{K} m_{ij} \langle \underline{\mathbf{f}}_j, \underline{\mathbf{g}}_j \rangle_{\mathcal{H}_j} \right)^2 && \text{(Definition of the inner product on } \mathcal{M}\text{)} \\
&\leq \sum_{i=1}^{K} \left(\sum_{j=1}^{K} m_{ij}^2 \right) \left(\sum_{j=1}^{K} \langle \underline{\mathbf{f}}_j, \underline{\mathbf{g}}_j \rangle_{\mathcal{H}_j}^2 \right) && \text{(Cauchy-Schwarz inequality on } \mathbb{R}^K\text{)} \\
&= \sum_{i=1}^{K} \left(\sum_{j=1}^{K} \langle \underline{\mathbf{f}}_j, \underline{\mathbf{g}}_j \rangle_{\mathcal{H}_j}^2 \right) && \text{(Orthogonality of } \mathbf{M}\text{)} \\
&\leq \sum_{i=1}^{K} \left(\sum_{j=1}^{K} \|\underline{\mathbf{f}}_j\|_{\mathcal{H}_j}^2 \cdot \|\underline{\mathbf{g}}_j\|_{\mathcal{H}_j}^2 \right) && \text{(Cauchy-Schwarz inequality on } \mathcal{H}_j\text{)} \\
&= \left(\sum_{j=1}^{K} \|\underline{\mathbf{f}}_j\|_{\mathcal{H}_j}^2 \right) \cdot \left(\sum_{j=1}^{K} \|\underline{\mathbf{g}}_j\|_{\mathcal{H}_j}^2 \right) \\
&= \|\underline{\mathbf{f}}\|_{\mathcal{M}}^2 \cdot \|\underline{\mathbf{g}}\|_{\mathcal{M}}^2, && \text{(Definition of the norm on } \mathcal{M}\text{)}
\end{aligned}$$

which completes the proof. □

C.1.2 One the Expressive Power of Hilbert t-Module and embeddings

The Functional t-SVD theorem states that for a square-integrable vector-valued function $F : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}^{1\times 1\times K}$, there exist orthonormal sets of functions $\{\underline{\phi}_i\}_{i=1}^{\infty} \subset L^2(\mathcal{X}; \mathbb{R}^{1\times 1\times K})$ and $\{\underline{\psi}_i\}_{i=1}^{\infty} \subset L^2(\mathcal{Y}; \mathbb{R}^{1\times 1\times K})$, and a sequence of t-scalars $\{\underline{\sigma}_i\}_{i=1}^{\infty} \subset \mathbb{R}^{1\times 1\times K}$, such that

$$F(x, y) = \sum_{i=1}^{\infty} \underline{\phi}_i(x) *_M \underline{\sigma}_i *_M \underline{\psi}_i(y).$$

Now, consider the Ground Truth Representation assumption, which posits the existence of a Hilbert t-Module $(\mathcal{M}, \langle \cdot, \cdot \rangle_{\mathcal{M}})$ and two embeddings $\underline{\mathbf{f}}^{\star} : \mathcal{X} \mapsto \mathcal{M}$ and $\underline{\mathbf{g}}^{\star} : \mathcal{Y} \mapsto \mathcal{M}$, such that the true function $\underline{h}^{\star}(x, y)$ can be expressed as the inner product of these embeddings:

$$\underline{h}^{\star}(x, y) = \left\langle \underline{\mathbf{f}}^{\star}(x), \underline{\mathbf{g}}^{\star}(y) \right\rangle_{\mathcal{M}}.$$

Assuming that the true function $\underline{h}^{\star}$ is square-integrable, i.e., $\underline{h}^{\star} \in L^2(\mathcal{X} \times \mathcal{Y}; \mathbb{R}^{1\times 1\times K})$, the Functional t-SVD theorem implies that it admits a representation of the form

$$\underline{h}^{\star}(x, y) = \sum_{i=1}^{\infty} \underline{\phi}_i(x) *_M \underline{\sigma}_i *_M \underline{\psi}_i(y).$$

In this context, the Hilbert t-Module $\mathcal{M}$ can be interpreted as the space spanned by the left singular functions $\{\underline{\phi}_i\}_{i=1}^{\infty}$, and the embeddings $\underline{\mathbf{f}}^{\star}$ and $\underline{\mathbf{g}}^{\star}$ can be defined as

$$\underline{\mathbf{f}}^{\star}(x) = \sum_{i=1}^{\infty} \underline{\phi}_i(x) *_M M^{-1}(M(\underline{\sigma}_i)^{\frac{1}{2}}),$$
$$\underline{\mathbf{g}}^{\star}(y) = \sum_{i=1}^{\infty} \underline{\psi}_i(y) *_M M^{-1}(M(\underline{\sigma}_i)^{\frac{1}{2}}),$$

where the square root operation $(\cdot)^{\frac{1}{2}}$ is performed element-wisely. With these definitions, the inner product of the embeddings in the Hilbert t-Module $\mathcal{M}$ recovers the true function:

$$\left\langle \underline{\mathbf{f}}^{\star}(x), \underline{\mathbf{g}}^{\star}(y) \right\rangle_{\mathcal{M}} = \sum_{i=1}^{\infty} \underline{\phi}_i(x) *_M \underline{\sigma}_i *_M \underline{\psi}_i(y) = \underline{h}^{\star}(x, y).$$

This demonstrates that, under the square-integrability assumption, the Functional t-SVD theorem provides a constructive way to obtain a Hilbert t-Module and embeddings that satisfy the t-bilinear representation assumption (Assumption 1). The theorem reveals a deep connection between the spectral decomposition of vector-valued functions and the geometric structure of Hilbert t-Modules, which can be exploited to develop efficient algorithms for multi-output regression and analysis. The use of the t-product ($*_M$) highlights the role of the tensor structure in this framework, and the consistent notation for the singular functions ($\underline{\phi}_i$ and $\underline{\psi}_i$) emphasizes their fundamental importance in the construction of the Hilbert t-Module and the associated embeddings.

C.1.3 More Explanations for the Assumptions

Explanation of Assumption 1. Assumption 1 is the cornerstone of the proposed theoretical framework for multi-output regression under CDS. It establishes a connection between the ground truth function and the geometry of a Hilbert t-Module, which is a generalization of a Hilbert space that accommodates vector-valued functions.

Part (I) of the assumption postulates that the ground truth function $\underline{h}^{\star}(x, y)$, which maps input pairs (x, y) to vector-valued outputs, admits a t-bilinear representation in a Hilbert t-Module $(\mathcal{M}, \langle \cdot, \cdot \rangle_{\mathcal{M}})$. This means that there exist two embeddings, $\underline{\mathbf{f}}^{\star} : \mathcal{X} \to \mathcal{M}$ and $\underline{\mathbf{g}}^{\star} : \mathcal{Y} \to \mathcal{M}$, such that $\underline{h}^{\star}(x, y)$ can be expressed as the inner product of these embeddings, i.e., $\underline{h}^{\star}(x, y) = \langle \underline{\mathbf{f}}^{\star}(x), \underline{\mathbf{g}}^{\star}(y) \rangle_{\mathcal{M}}$. The t-bilinear form is a natural extension of the bilinear form used (in single-output learning) [46], and it allows us to capture the multi-dimensional structure of the problem.

The assumption further states that this t-bilinear representation is the Bayes optimal predictor on both the training distribution $\mathcal{D}_{\text{train}}$ and the test distribution $\mathcal{D}_{\text{test}}$. This means that among all measurable functions, the t-bilinear representation minimizes the expected loss on both distributions. This assumption provides a clear learning objective and a benchmark for the performance of any learning algorithm.

Part (II) of the assumption imposes regularity and boundedness conditions on the output variable $\mathbf{z}$ and the embeddings $\underline{\mathbf{f}}^{\star}(x)$ and $\underline{\mathbf{g}}^{\star}(y)$. It assumes that the squared ℓ_2-norm of $\mathbf{z}$ is bounded by B^2 almost surely under the training distribution $\mathcal{D}_{\text{train}}$, and the Hilbert t-Module norms of the embeddings are uniformly bounded by the same constant B. These conditions are crucial for deriving theoretical guarantees and ensuring the well-posedness of the learning problem. They prevent pathological cases and ensure that the learning problem is tractable.

Intuitively, Assumption 1 tells us that the ground truth function has a simple and interpretable structure (t-bilinear representation) that is optimal on both the training and testing distributions. Moreover, the output and the embeddings are well-behaved (bounded). This sets the stage for the learning problem and provides a clear target for any learning algorithm.

Explanation of Assumption 2. Assumption 2 characterizes the relationship between the training and testing distributions in the context of CDS. In real-world applications, it is common for the testing distribution to differ from the training distribution. CDS is a specific type of distribution shift where the testing distribution includes novel combinations of features that are not present in the training distribution. Following Ref. [46], we can also relax Assumption 2 through a probabilistic manner as follows:

Assumption 6 (Coverage of training and test distribution (probabilistic version), Assumption 2.2b in Ref. [46])**.** *There exist constants* $\kappa_{\text{tst}}, \kappa_{\text{trn}} > 0$*,* $\eta_{\text{tst}}, \eta_{\text{trn}} \in (0, 1]$ *and* marginal distributions $\mathcal{D}_{\mathcal{X},1}, \mathcal{D}_{\mathcal{X},2}$ *over* $\mathcal{X}$*, and* $\mathcal{D}_{\mathcal{Y},1}, \mathcal{D}_{\mathcal{Y},2}$ *over* $\mathcal{Y}$*, with product measures* $\mathcal{D}_{i\otimes j} := \mathcal{D}_{\mathcal{X},i} \otimes \mathcal{D}_{\mathcal{Y},j}$*, such that the following conditions in Radon–Nikodym derivatives hold for all* $(x, y) \in \mathcal{X} \times \mathcal{Y}$*:*

$$\mathbb{P}_{\mathcal{D}_{i\otimes j}}\left[\frac{\mathrm{d}\mathcal{D}_{i\otimes j}(x,y)}{\mathrm{d}\mathcal{D}_{\text{train}}(x,y)} > \kappa_{\text{trn}}\right] \le \eta_{\text{trn}}, \quad (i,j) \in \{(1,1),(1,2),(2,1)\} \qquad \text{(Training coverage)}$$

$$\mathbb{P}_{\mathcal{D}_{\text{test}}}\left[\frac{\mathrm{d}\mathcal{D}_{\text{test}}(x,y)}{\sum_{i,j\in\{1,2\}} \mathrm{d}\mathcal{D}_{i\otimes j}(x,y)} > \kappa_{\text{tst}}\right] \le \eta_{\text{tst}}. \qquad \text{(Test coverage)}$$

Assumption 6 formalizes this notion by introducing marginal distributions $\mathcal{D}_{\mathcal{X},1}, \mathcal{D}_{\mathcal{X},2}$ over the input space $\mathcal{X}$ and $\mathcal{D}_{\mathcal{Y},1}, \mathcal{D}_{\mathcal{Y},2}$ over the output space $\mathcal{Y}$. The product measures $\mathcal{D}_{i\otimes j} := \mathcal{D}_{\mathcal{X},i} \otimes \mathcal{D}_{\mathcal{Y},j}$ represent different combinations of these marginal distributions.

The assumption requires that the training distribution $\mathcal{D}_{\text{train}}$ covers the key feature combinations ((1,1), (1,2), (2,1)) of these product measures, while allowing the testing distribution $\mathcal{D}_{\text{test}}$ to include unseen combinations (e.g., (2,2)). The coverage is quantified by the constants $\kappa_{\text{trn}}, \kappa_{\text{tst}}, \eta_{\text{trn}}, \eta_{\text{tst}}$, which provide some flexibility in the assumption.

Intuitively, this assumption tells us that the training distribution should be diverse enough to cover the key aspects of the marginal distributions, but it doesn't need to cover all possible combinations. The testing distribution, on the other hand, can include novel combinations that are not present in the training distribution. This is a realistic assumption in many practical scenarios.

Explanation of Assumption 3. Assumption 3 is a technical assumption that controls the impact of covariate shift on the model's performance. Covariate shift refers to the situation where the distribution of the input features changes between the training and testing distributions. Motivated by Assumptionn2.3b in Ref. [46] we can also consider a slacked version of Assumption 3 as follows.

Assumption 7 (Controlled covariate shifts, slacked version)**.** *There exists a* $\kappa_{\text{cov}} \ge 1$ *and* $\eta_{\text{cov}} \in (0, 1]$ *such that, for any* $\underline{\mathbf{v}} \in \mathcal{M}$*, the following inequalities hold:*

$$\mathbb{E}_{x\sim\mathcal{D}_{\mathcal{X},2}}[\|\langle\underline{\mathbf{f}}^\star(x), \underline{\mathbf{v}}\rangle_{\mathcal{M}}\|^2] \le \kappa_{\text{cov}} \cdot \mathbb{E}_{x\sim\mathcal{D}_{\mathcal{X},1}}[\|\langle\underline{\mathbf{f}}^\star(x), \underline{\mathbf{v}}\rangle_{\mathcal{M}}\|^2] + \eta_{\text{cov}}\|\underline{\mathbf{v}}\|_{\mathcal{M}}^2$$
$$\mathbb{E}_{y\sim\mathcal{D}_{\mathcal{Y},2}}[\|\langle\underline{\mathbf{g}}^\star(y), \underline{\mathbf{v}}\rangle_{\mathcal{M}}\|^2] \le \kappa_{\text{cov}} \cdot \mathbb{E}_{y\sim\mathcal{D}_{\mathcal{Y},1}}[\|\langle\underline{\mathbf{g}}^\star(y), \underline{\mathbf{v}}\rangle_{\mathcal{M}}\|^2] + \eta_{\text{cov}}\|\underline{\mathbf{v}}\|_{\mathcal{M}}^2.$$

The assumption states that for any t-vector $\underline{\mathbf{v}}$ in the Hilbert t-Module $\mathcal{M}$, the expected squared norm of the t-inner product between $\underline{\mathbf{v}}$ and the embeddings $\underline{\mathbf{f}}^\star(x)$ and $\underline{\mathbf{g}}^\star(y)$ under the unseen distributions $\mathcal{D}_{\mathcal{X},2}$ and $\mathcal{D}_{\mathcal{Y},2}$ is bounded by a constant factor κ_{cov} times the same quantity under the training distributions $\mathcal{D}_{\mathcal{X},1}$ and $\mathcal{D}_{\mathcal{Y},1}$, plus a small constant η_{cov} times the squared Hilbert t-Module norm of $\underline{\mathbf{v}}$.

Intuitively, this assumption ensures that the behavior of the embeddings doesn't change too much under covariate shift. It bounds the worst-case impact of the shift on the model's performance. This is a key ingredient in providing generalization guarantees under CDS.

Without this assumption, the model's performance on the testing distribution could be arbitrarily bad, even if it performs well on the training distribution. The assumption ensures that the performance degradation is controlled, which is essential for learning under distribution shift.

Explanation of Assumption 4. Assumption 4 postulates favorable spectral properties of the ground truth embeddings $\underline{\mathbf{f}}^\star$ and $\underline{\mathbf{g}}^\star$. The spectral properties refer to the behavior of the singular values of the embeddings' t-covariance operators.

Part (I) of the assumption, termed "Balanced embeddings," posits that the embeddings are balanced in an appropriate basis of the Hilbert t-Module, such that their t-covariances $\underline{\mathbf{\Sigma}}_{\underline{\mathbf{f}}^\star}$ and $\underline{\mathbf{\Sigma}}_{\underline{\mathbf{g}}^\star}$ equal a common t-covariance $\underline{\mathbf{\Sigma}}_{1\otimes 1}^\star$. This assumption might initially appear restrictive, but mathematically, it is very mild. Specifically, we can always find a pair of invertible t-linear transformations $\underline{\mathbf{T}}_f$ and $\underline{\mathbf{T}}_g$

such that:

$$\underline{\mathbf{T}}_f *_M \underline{\mathbf{\Sigma}}_{\underline{\mathbf{f}}^\star} *_M \underline{\mathbf{T}}_f^\top = \underline{\mathbf{T}}_g *_M \underline{\mathbf{\Sigma}}_{\underline{\mathbf{g}}^\star} *_M \underline{\mathbf{T}}_g^\top =: \underline{\mathbf{\Sigma}}_{1\otimes 1}^\star.$$

In other words, by applying suitable transformations to the embeddings $\underline{\mathbf{f}}^\star$ and $\underline{\mathbf{g}}^\star$, their t-covariances can be aligned to a common t-covariance $\underline{\mathbf{\Sigma}}_{1\otimes 1}^\star$ in a transformed basis of the Hilbert t-Module. Practically, the assumption of balanced embeddings can be viewed as a form of normalization or standardization. By aligning their t-covariances, we ensure that the embeddings have similar scales and orientations in the Hilbert t-Module, facilitating the learning process and simplifying theoretical analysis.

Part (II) of the assumption, termed "Polynomial spectral decay," assumes that each frequency component of $\underline{\mathbf{\Sigma}}_{1\otimes 1}^\star$ in the transformed domain induced by the transform $M(\cdot)$ exhibits a polynomial singular value decay, potentially with different decay rates γ_i for different components.

Intuitively, this assumption tells us that the information in the embeddings is concentrated in a few principal directions (corresponding to the large singular values), and the importance of the remaining directions decays polynomially. This is a common assumption in many learning problems, and it allows us to effectively approximate the embeddings using low-rank structures.

The polynomial decay assumption is more flexible than the often-used exponential decay assumption, as it allows different frequency components to decay at different rates. This is particularly useful in the context of multi-output regression, where different outputs may have different spectral properties.

Explanation of Assumption 5. Assumption 5 essentially states that the error between the optimal rank-k embeddings and the ground truth function $\underline{h}^\star$ over the training data $\mathcal{D}_{\text{train}}$ is bounded by a constant factor κ_{apx} times the error over a reference distribution $\mathcal{D}_{1\otimes 1}$. This assumption ensures that the training data provides a good approximation of the ground truth function in each frequency sub-domain.

Intuitively, this assumption implies that the training data $\mathcal{D}_{\text{train}}$ is sufficiently representative of the true function $\underline{h}^\star$ in each frequency sub-domain induced by the transformation $M(\cdot)$. It guarantees that if we find embeddings that well approximate $\underline{h}^\star$ on the training data, they will also perform well on the reference distribution $\mathcal{D}_{1\otimes 1}$. More specifically, $\text{ApxErr}_k^{(i)}(x, y)$ measures the squared difference between the inner product of the optimal rank-k embeddings ($\mathbf{P}_k M(\underline{\mathbf{f}}^\star(x))^{(i)}$ and $\mathbf{P}_k M(\underline{\mathbf{g}}^\star(y))^{(i)}$) and the i-th frequency component of the true function $(M(\underline{h}^\star(x, y))^{(i)})$ for a given pair of inputs (x, y). By taking the expectation of this error over the training distribution $\mathcal{D}_{\text{train}}$ and the reference distribution $\mathcal{D}_{1\otimes 1}$, we obtain a measure of how well the rank-k embeddings approximate the true function in each frequency sub-domain.

The assumption states that the expected approximation error over the training data is bounded by a constant factor κ_{apx} times the expected error over the reference distribution, for all frequency sub-domains $i \in [K]$. This means that the training data provides a good approximation of the true function in each sub-domain, relative to the reference distribution.

The significance of this assumption lies in its importance for generalization analysis under CDS. By ensuring that the training data adequately represents the true function within each frequency sub-domain, it enables the learning of embeddings that effectively capture essential characteristics of the true function. This, in turn, allows the learned model to generalize well to unseen combinations of input features, provided that the spectral properties of the true function are favorable (e.g., exhibiting polynomial decay as assumed in Assumption 4), thereby accommodating the multi-output nature of the problem.

C.2 More Details of the Algorithms

C.2.1 Limitations of ERM for multi-output regression under CDS.

In the ERM model (5), the function classes $\mathcal{F}_r, \mathcal{G}_r$ are given as follows:

Assumption 8 (Hypothesis classes in the ERM Model (5))**.** *For each $k \in \mathbb{N}$, there exist function classes $\mathcal{F}_k \subseteq \{\mathcal{X} \to \mathbb{R}^{k\times 1\times K}\}$ and $\mathcal{G}_k \subseteq \{\mathcal{Y} \to \mathbb{R}^{k\times 1\times K}\}$ satisfy:*

(a) There exist some $(\underline{\mathbf{f}}, \underline{\mathbf{g}}) \in \mathcal{F}_k \times \mathcal{G}_k$ *such that* $\langle \underline{\mathbf{f}}(x), \underline{\mathbf{g}}(y) \rangle_t := (\underline{\mathbf{f}}(x))^\top *_M \underline{\mathbf{g}}(y) = \langle \underline{\mathbf{f}}_k^\star(x), \underline{\mathbf{g}}_k^\star(y) \rangle_{\mathcal{M}}$ *for all* $(x, y) \in \mathcal{X} \times \mathcal{Y}$*.*

(b) The function classes $\mathcal{F}_k$ *and* $\mathcal{G}_k$ *have their* ϵ*-covering numbers bounded as follows:* $\log \mathcal{N}(\mathcal{F}_k, \epsilon, \|\cdot\|_\infty) \leq N_{\mathcal{F}_k}(\epsilon)$ *and* $\log \mathcal{N}(\mathcal{G}_k, \epsilon, \|\cdot\|_\infty) \leq N_{\mathcal{G}_k}(\epsilon)$*, where* $N_{\mathcal{F}_k}$ *and* $N_{\mathcal{G}_k}$ *are non-decreasing functions of* k *and decreasing functions of* $\epsilon > 0$*.*

(c) There exists some $B > 0$*, such that* $\max \left\{ \sup_{x \in \mathcal{X}} \|\underline{\mathbf{f}}(x)\|_{\mathcal{M}}, \sup_{y \in \mathcal{Y}} \|\underline{\mathbf{g}}(y)\|_{\mathcal{M}} \right\} \leq B$*.*

This assumption specifies the hypothesis classes $\mathcal{F}_r$ and $\mathcal{G}_r$ used in the ERM problem (5). It ensures that these classes are sufficiently expressive to approximate the true embeddings $\underline{\mathbf{f}}^\star$ and $\underline{\mathbf{g}}^\star$, while also having controlled complexity in terms of their covering numbers. With these assumptions in place, we can obtain the generalization guarantee for the ERM solution in Theorem 4. However, the single-stage ERM approach has several limitations when applied to multi-output regression under CDS:

- *Lack of adaptivity to spectral decay patterns*: The hypothesis classes $\mathcal{F}_r$ and $\mathcal{G}_r$ used in the single-stage ERM approach do not explicitly take into account the potentially varying spectral decay patterns of the ground truth embeddings along different frequency components of the transform $M(\cdot)$. As a result, the learned embeddings may not adequately capture the distinct decay behaviors in different frequency sub-domains, leading to suboptimal approximations.
- *Global learning and complexity control*: The single-stage ERM approach learns the embeddings $\hat{\underline{\mathbf{f}}}_{\text{erm}}$ and $\hat{\underline{\mathbf{g}}}_{\text{erm}}$ globally, without considering the specific characteristics of different frequency sub-domains. This global learning approach may not be able to adapt to the nuances and variations in the decay patterns across sub-domains. Moreover, the complexity of the hypothesis classes $\mathcal{F}_r$ and $\mathcal{G}_r$ is controlled globally through their covering numbers, which may not provide a fine-grained enough complexity control to capture the differences in the decay patterns.
- *Suboptimal generalization guarantees*: Due to the limitations mentioned above, the single-stage ERM approach may result in suboptimal generalization guarantees for multi-output regression under CDS. The learned embeddings may not generalize well to new feature combinations, particularly when the spectral decay patterns vary significantly across different frequency sub-domains.

C.2.2 Double-Stage ERM (ERM-DS) for improved multi-output regression under CDS.

To address these weaknesses, we propose a Double-Stage Empirical Risk Minimization (ERM-DS) algorithm. This approach employs a two-stage training process with more specific hypothesis classes to better capture the varying spectral decay patterns across sub-domains. By striking a balance between model complexity and generalization ability, ERM-DS aims to overcome the limitations of the single-stage ERM approach and provide improved generalization guarantees for multi-output regression under CDS.

The framework of Double-Stage Empirical Risk Minimization is first proposed by Ref. [46] to address robust learning under CDS (for single-output learning), our work extends this framework to the more complex setting of multi-output regression with t-bilinear embeddings. By learning vector functions as embeddings in a Hilbert t-module and leveraging tensor algebra, we capture the intricate dependencies among multiple outputs.

To more accurately capture the varying spectral decay patterns of the ground truth along each frequency component of the transform $M(\cdot)$, the algorithm learns the t-bilinear embeddings $\underline{\mathbf{f}}^\star$ and $\underline{\mathbf{g}}^\star$ by approximating their frequency components $\breve{\underline{\mathbf{f}}}^{\star,(i)}$ and $\breve{\underline{\mathbf{g}}}^{\star,(i)}$ separately in each i-th frequency sub-domain. This is achieved by employing function classes specifically characterized for each sub-domain, allowing the algorithm to adapt to the distinct decay behaviors exhibited by the ground truth in each frequency range. By treating each sub-domain independently and using specialized function classes, the algorithm can better capture the nuances and variations in the decay patterns, leading to a more precise approximation of the t-bilinear embeddings $\underline{\mathbf{f}}^\star$ and $\underline{\mathbf{g}}^\star$.

Assumption 9 (Function Approximation in Transformed sub-frequency components Separately)**.** *In any sub-frequency component* $i \in [K]$*, for each* $k \in \mathbb{N}$*, there exist function classes* $\breve{\mathcal{F}}_k^{(i)} \subseteq \{\mathcal{X} \to \mathbb{R}^k\}$ *and* $\breve{\mathcal{G}}_k^{(i)} \subseteq \{\mathcal{Y} \to \mathbb{R}^k\}$ *satisfy:*

(a) $\sup_{f \in \breve{\mathcal{F}}_k^{(i)}} \sup_{x \in \mathcal{X}} \|f(x)\| \leq B$ *and* $\sup_{g \in \breve{\mathcal{G}}_k^{(i)}} \sup_{y \in \mathcal{Y}} \|g(y)\| \leq B$.

(b) There exist some $(f, g) \in \breve{\mathcal{F}}_k^{(i)} \times \breve{\mathcal{G}}_k^{(i)}$ *such that* $\langle f(x), g(y) \rangle = \langle \breve{\underline{\mathbf{f}}}_k^{\star,(i)}(x), \breve{\underline{\mathbf{g}}}_k^{\star,(i)}(y) \rangle$ *for all* $(x, y) \in \mathcal{X} \times \mathcal{Y}$*, where* $\breve{\underline{\mathbf{f}}}_k^{\star,(i)}(x), \breve{\underline{\mathbf{g}}}_k^{\star,(i)}(y)$ *are the projections of* $\breve{\underline{\mathbf{f}}}^{\star,(i)}(x)$ *and* $\breve{\underline{\mathbf{g}}}^{\star,(i)}(y)$ *in the top-k eigenspaces of* $M(\underline{\mathbf{\Sigma}}_{1\otimes 1}^{\star})^{(i)}$*, respectively.*

(c) The function classes $\breve{\mathcal{F}}_k^{(i)}$ *and* $\breve{\mathcal{G}}_k^{(i)}$ *have their* ϵ*-covering numbers bounded as follows:* $\log \mathcal{N}(\breve{\mathcal{F}}_k^{(i)}, \epsilon, \|\cdot\|_\infty) \leq M_{\breve{\mathcal{F}}_k^{(i)}}(\epsilon)$ *and* $\log \mathcal{N}(\breve{\mathcal{G}}_k^{(i)}, \epsilon, \|\cdot\|_\infty) \leq M_{\breve{\mathcal{G}}_k^{(i)}}(\epsilon)$*, where* $M_{\breve{\mathcal{F}}_k^{(i)}}$ *and* $M_{\breve{\mathcal{G}}_k^{(i)}}$ *are non-decreasing functions of* k *and decreasing functions of* $\epsilon > 0$.

The algorithm consists of the following key steps:

Step 1: Overparameterized Training. In the first stage, ERM-DS trains an overparameterized model $(\tilde{\underline{\mathbf{f}}}, \tilde{\underline{\mathbf{g}}})$ with high capacity to approximate the unknown true predictive functions $\underline{\mathbf{f}}^{\star}$ and $\underline{\mathbf{g}}^{\star}$. This allows the algorithm to handle the uncertainty in the true predictive functions, which are unknown and potentially intricate. The over-parameterized model is trained by minimizing the empirical risk on labeled training data $\{(x_{1,j}, y_{1,j}, \mathbf{z}_{1,j})\}_{j=1}^{n_1}$ sampled *i.i.d.* from $\mathcal{D}_{\text{train}}$. Instead of training $(\tilde{\underline{\mathbf{f}}}, \tilde{\underline{\mathbf{g}}})$ within a single optimization problem as in Eq. (5), we consider choosing their frequency components in the transformed domains induced by $M(\cdot)$, separately via K parallel ERM sub-problems as follows:

$$(\breve{\tilde{\underline{\mathbf{f}}}}^{(i)}, \breve{\tilde{\underline{\mathbf{g}}}}^{(i)}) \in \underset{(f,g) \in \breve{\mathcal{F}}_p^{(i)} \times \breve{\mathcal{G}}_p^{(i)}}{\operatorname{argmin}} \frac{1}{n_1} \sum_{j=1}^{n_1} (\langle f(x_{1,j}), g(y_{1,j}) \rangle - M(\mathbf{z}_{1,j})^{(i)})^2, \quad \forall i \in [K]. \tag{18}$$

In this equation, $(\breve{\tilde{\underline{\mathbf{f}}}}^{(i)}, \breve{\tilde{\underline{\mathbf{g}}}}^{(i)})$ represents the i-th frequency component of the overparameterized model $(\tilde{\underline{\mathbf{f}}}, \tilde{\underline{\mathbf{g}}})$ in the transformed domain, and $\breve{\mathcal{F}}_p^{(i)}$ and $\breve{\mathcal{G}}_p^{(i)}$ denote the function classes for f and g, respectively. The objective is to minimize the squared difference between the inner product of $f(x_{1,j})$ and $g(y_{1,j})$ and the i-th component of the transformed target variable $M(\mathbf{z}_{1,j})$, averaged over the training data.

By solving these K parallel ERM sub-problems, the algorithm learns the frequency components of the overparameterized model in the transformed domains. This approach allows for a more flexible and potentially more accurate approximation of the unknown true predictive functions, compared to training the model within a single optimization problem.

Step 2: t-Covariance Estimation. To leverage the assumption that the true predictive functions have a low-rank structure (Assumption 4), the algorithm estimates the t-covariances of the embeddings using n_2 additional unlabeled examples $\{(x_{2,j}, y_{2,j})\}_{j=1}^{n_2}$ sampled from $\mathcal{D}_{1\otimes 1}$:

$$\hat{\underline{\mathbf{\Sigma}}}_{\tilde{\underline{\mathbf{f}}}} := \frac{1}{n_2} \sum_{i=1}^{n_2} \tilde{\underline{\mathbf{f}}}(x_{2,j}) *_M \tilde{\underline{\mathbf{f}}}(x_{2,j})^\top, \quad \hat{\underline{\mathbf{\Sigma}}}_{\tilde{\underline{\mathbf{g}}}} := \frac{1}{n_2} \sum_{i=1}^{n_2} \tilde{\underline{\mathbf{g}}}(y_{2,j}) *_M \tilde{\underline{\mathbf{g}}}(y_{2,j})^\top. \tag{19}$$

These covariance estimates capture the important directions of variation in the learned embeddings and will be used for dimension reduction in the next step.

Step 3: Dimension Reduction. Using the estimated covariances, the algorithm computes low-rank projections $(\hat{\underline{\mathbf{Q}}}_{\hat{\mathbf{r}}}, \hat{\mathbf{r}})$ using the `DimReduce` function in Algorithm 1 with a target ranks $\mathbf{r}_{\text{cut}} \in \mathbb{N}^K$ and cutoff parameters $\boldsymbol{\sigma}_{\text{cut}} \in \mathbb{R}^K$ in each frequency domain:

$$(\hat{\underline{\mathbf{Q}}}_{\hat{\mathbf{r}}}, \hat{\mathbf{r}}) \leftarrow \texttt{DimReduce}(\hat{\underline{\mathbf{\Sigma}}}_{\tilde{\underline{\mathbf{f}}}} + \mu \underline{\mathbf{I}}_p, \hat{\underline{\mathbf{\Sigma}}}_{\tilde{\underline{\mathbf{g}}}} + \mu \underline{\mathbf{I}}_p, \mathbf{r}_{\text{cut}}, \boldsymbol{\sigma}_{\text{cut}}), \tag{20}$$

where μ is a regularization parameter and $\underline{\mathbf{I}}$ is the t-identity tensor. The reduced-rank embeddings are then obtained by projecting the overparameterized embeddings onto the low-rank t-subspace: $\underline{h}_{\text{red}}(x, y) := \tilde{\underline{\mathbf{f}}}(x)^\top *_M \hat{\underline{\mathbf{Q}}}_{\hat{\mathbf{r}}} *_M \tilde{\underline{\mathbf{g}}}(y)$.

The high capacity of the overparameterized model makes it prone to overfitting, especially when the training data is limited. However, the true predictive functions can be effectively approximated using

Algorithm 1 `DimReduce`$(\underline{\mathbf{X}}, \underline{\mathbf{Y}}, \mathbf{r}, \boldsymbol{\sigma}')$

```
1: Input: Two t-covariances tensors $\underline{\mathbf{X}}, \mathbf{Y} \in \mathbb{R}^{p\times p\times K}$, $\mathbf{r} = (r_1, \dots, r_K)^\top \in \mathbb{N}^K$, $\boldsymbol{\sigma}' = (\sigma'_1, \dots, \sigma'_K)^\top \in \mathbb{R}^K$.
2: Initialize $\underline{\mathbf{Q}} = \underline{\mathbf{0}} \in \mathbb{R}^{p\times p\times K}$
3: for $i = 1$ to $K$ do
4:     Compute $\mathbf{W}_i := \mathbf{X}_i^{\frac{1}{2}}(\mathbf{X}_i^{\frac{1}{2}}\mathbf{Y}_i\mathbf{X}_i^{\frac{1}{2}})^{-\frac{1}{2}}\mathbf{X}_i^{\frac{1}{2}}$, where $\mathbf{X}_i = M(\underline{\mathbf{X}})^{(i)}, \mathbf{Y}_i = M(\underline{\mathbf{Y}})^{(i)}$.
5:     Compute $\boldsymbol{\Sigma}_i := \mathbf{W}_i^{\frac{1}{2}}\mathbf{Y}_i\mathbf{W}_i^{\frac{1}{2}}$.
6:     Set $s_i \leftarrow \max\left\{s \in [r_i] : \sigma_s(\boldsymbol{\Sigma}_i) \geq \sigma'_i, \sigma_s(\boldsymbol{\Sigma}_i) - \sigma_{s+1}(\boldsymbol{\Sigma}_i) \geq \frac{\sigma_s(\boldsymbol{\Sigma}_i)}{r_i}\right\}$.
7:     Let $\mathbf{P}_{s_i}$ denote the projection onto the top $s_i$ eigenvectors of $\boldsymbol{\Sigma}_i$.
8:     Compute $\mathbf{Q}_{s_i} := \mathbf{W}_i^{-\frac{1}{2}}\mathbf{P}_{s_i}\mathbf{W}_i^{\frac{1}{2}}$ and set $\underline{\mathbf{Q}}^{(i)} \leftarrow \mathbf{Q}_{s_i}$.
9: end for
10: Return $(M(\underline{\mathbf{Q}}), \mathbf{s})$ where $\mathbf{s} = (s_1, \dots, s_K) \in \mathbb{N}^K$.
```

fewer dimensions due to their low-rank structure. By projecting the embeddings onto a low-rank subspace, the algorithm reduces the model's complexity and mitigates overfitting. This dimensionality reduction step helps strike a balance between model capacity and generalization ability. It enables the model to focus on the most important patterns in the data while discarding less informative directions. This step is crucial for improving the model's generalization performance and preventing it from memorizing noise or irrelevant details in the training data.

Step 4: Distillation. In the final stage, the algorithm fine-tunes the reduced-rank embeddings $(\hat{\underline{\mathbf{f}}}_{\text{ds}}, \hat{\underline{\mathbf{g}}}_{\text{ds}})$ by approximating their frequency components $(\breve{\hat{\underline{\mathbf{f}}}}^{(i)}_{\text{ds}}, \breve{\hat{\underline{\mathbf{g}}}}^{(i)}_{\text{ds}})$ separately in the transformed domain induced by $M(\cdot)$. This is achieved by minimizing a combination of two loss functions:

$$(\breve{\hat{\underline{\mathbf{f}}}}^{(i)}_{\text{ds}}, \breve{\hat{\underline{\mathbf{g}}}}^{(i)}_{\text{ds}}) \in \operatorname*{argmin}_{(f,g)\in\breve{\mathcal{F}}^{(i)}_{\hat{r}_i}\times\breve{\mathcal{G}}^{(i)}_{\hat{r}_i}} \hat{L}^{(i)}_{(3)}(f,g) + \nu_i\hat{L}^{(i)}_{(4)}(f,g), \tag{21}$$

where $\boldsymbol{\nu} = (\nu_1, \cdots, \nu_K)^\top \in \mathbb{R}^K$ is a vector of regularization parameters and the function classes $\breve{\mathcal{F}}^{(i)}_{\hat{r}_i}$ and $\breve{\mathcal{G}}^{(i)}_{\hat{r}_i}$ have reduced complexity compared to $\breve{\mathcal{F}}^{(i)}_p$ and $\breve{\mathcal{G}}^{(i)}_p$.

The first loss $\hat{L}^{(i)}_{(3)}$ is the empirical risk in the i-th frequency domain on n_3 additional labeled examples $\{(x_{3,j}, y_{3,j}, \mathbf{z}_{3,j})\}_{j=1}^{n_3}$ sampled from $\mathcal{D}_{\text{train}}$:

$$\hat{L}_{(3)}(f,g) = \frac{1}{n_3}\sum_{j=1}^{n_3}(\langle f(x_{3,i}), g(y_{3,i})\rangle - M(\mathbf{z}_{3,i})^{(i)})^2.$$

By minimizing this loss, the algorithm ensures that the embeddings accurately capture the relationships between the input features and the target outputs.

The second loss $\hat{L}^{(i)}_{(4)}$ is a regularization term that encourages consistency with the reduced-rank embeddings $\underline{h}_{\text{red}}$ on n_4 unlabeled examples $\{(x_{4,j}, y_{4,j})\}_{j=1}^{n_4}$ sampled from $\mathcal{D}_{1\otimes 1}$:

$$\hat{L}^{(i)}_{(4)}(f,g) = \frac{1}{n_4}\sum_{i=1}^{n_4}(\langle f(x_{4,i}), g(y_{4,i})\rangle - M(\underline{h}_{\text{red}}(x_{4,i}, y_{4,i}))^{(i)})^2.$$

By learning to match the reduced-rank embeddings $\underline{h}_{\text{red}}$ on these unlabeled samples, the final model becomes more robust to distribution shifts.

Intuitively, the distillation step allows the model to learn from both labeled and unlabeled data, leveraging the information captured by the reduced-rank embeddings to improve its generalization ability. The regularization term acts as a guide, encouraging the fine-tuned embeddings to maintain the important patterns and structures learned during the previous steps while adapting to new feature combinations. By combining the empirical risk and the regularization term, ERM-DS strikes a balance between fitting the labeled training data and being robust to distribution shifts. This helps the model generalize well to unseen feature combinations and ensures stable performance even when the test distribution differs from the training distribution.

C.2.3 The Role of Balanced t-Covariance Operator: Capturing Knowledge under CDS

The t-covariance operator $\underline{\boldsymbol{\Sigma}}^{\star}_{1\otimes 1}$ is central to the Ft-SVD framework for multi-output regression under combinatorial distribution shift, capturing the spectral properties and low-rank structure of the predictive functions across K frequency components. Defined as the common t-covariance of balanced embeddings $\underline{\mathbf{f}}^{\star}$ and $\underline{\mathbf{g}}^{\star}$ on $\mathcal{D}_{1\otimes 1}$, $\underline{\boldsymbol{\Sigma}}^{\star}_{1\otimes 1}$ reflects the intrinsic low-dimensional structure and shared patterns across feature combinations. Its spectral decay, indicating the eigenvalue decay of its frontal slices in the transformed domain, guides the Ft-SVD framework in aligning embeddings with principal eigenfunctions, enabling effective knowledge transfer under distribution shift.

Estimated from $\mathcal{D}_{1\otimes 1} := \mathcal{D}_{\mathcal{X},1} \otimes \mathcal{D}_{\mathcal{Y},1}$, the t-covariance operator supports three main steps the ERM-DS algorithm:

- **Estimation**: Using unlabeled samples from $\mathcal{D}_{1\otimes 1}$ to identify significant variation directions in embeddings.
- **Dimension Reduction**: Employing spectral decay to compute low-rank projections that balance capacity and generalization.
- **Distillation**: Constructing a regularization term to align fine-tuned and reduced-rank embeddings, enhancing robustness.

At its core, $\underline{\boldsymbol{\Sigma}}^{\star}_{1\otimes 1}$ ***acts as a form of transferable "knowledge" for multi-output problems under combinatorial distribution shifts, capturing and transferring crucial patterns across feature combinations.*** Its spectral properties and principal eigenfunctions encapsulate the low-dimensional structure, guiding the model to focus on the most transferable features and achieve generalization in shifted distributions. Through $\underline{\boldsymbol{\Sigma}}^{\star}_{1\otimes 1}$, the Ft-SVD framework and ERM-DS enable effective knowledge transfer and address the complexities of combinatorial distribution shift.

D Analysis of the Proposed Algorithms

In this section, we embark on a comprehensive theoretical analysis of the ERM and ERM-DS algorithms for multi-output regression under the challenging setting of CDS. Our primary goal is to establish rigorous generalization guarantees for these algorithms.

Challenges. In the our settings of multi-output regression under CDS, several key challenges arise. Firstly, the algorithms must be able to effectively capture the complex relationships between the input features and the multiple output variables, even when the training data only covers a limited range of feature combinations. Secondly, the spectral decay patterns of the true embeddings, which encode the intrinsic structure of the data, may vary across different sub-domains (i.e., frequency components), necessitating algorithms that can adapt to these variations. Finally, the presence of distribution shift between the training and test domains can significantly impact the generalization performance, requiring algorithms that can mitigate the effects of this shift.

Section organization. To address these challenges and provide a comprehensive analysis of the ERM and ERM-DS algorithms, we organize this section into three key subsections.

- First, we introduce a technical tool that allows us to decompose the excess risk of the algorithms into various interpretable terms, such as the approximation error, statistical error, and distribution shift terms in Appendix D.1. This decomposition forms the foundation of our subsequent analysis and provides valuable insights into the factors influencing the generalization performance.
- Second, we focus on the theoretical analysis of the ERM algorithm in Appendix D.2. We present the main theorem that provides generalization guarantees for the ERM solution under certain conditions and outline the key steps in proving this theorem. Through this analysis, we highlight the limitations of the ERM approach, such as its lack of adaptivity to the spectral decay patterns and suboptimal generalization guarantees, motivating the need for a more sophisticated algorithm.
- Third, we introduce the ERM-DS algorithm as a response to the limitations of the ERM approach in Appendix D.3. We provide an overview of the key steps of the algorithm, including overparameterization, covariance estimation, dimension reduction, and distillation, and explain how each step contributes to the improved generalization performance. Through a high-level sketch of the proof for the main theorem, we highlight the key ideas and techniques used in each step of the algorithm and discuss how the error terms are controlled by carefully selecting the algorithm parameters and sample sizes.

Addtional notations. To facilitate the analysis, we introduce some additional notations. We use $a \lesssim b$ to denote $a \leq c \cdot b$ for some absolute constant c; we use $a \lesssim_\star b$ to denote $a \leq c \cdot b$ for some c that is at most polynomial in the problem constants $\kappa_{\mathrm{cov}}, \kappa_{\mathrm{trn}}, \kappa_{\mathrm{apx}}$.

Given a probability distribution $\mathcal{D}$ on $(x, y) \in \mathcal{X} \times \mathcal{Y}$ pairs, we define the excess risk of the embeddings $(\hat{\underline{\mathbf{f}}}, \hat{\underline{\mathbf{g}}})$ as $\mathcal{R}(\hat{\underline{\mathbf{f}}}, \hat{\underline{\mathbf{g}}}; \mathcal{D}) := \mathbb{E}_{(x,y)\sim\mathcal{D}}[\|\langle \hat{\underline{\mathbf{f}}}(x), \hat{\underline{\mathbf{g}}}(y)\rangle_{\mathcal{M}} - \underline{h}^\star(x,y)\|^2]$, which quantifies the expected squared difference between the product of the embeddings and the true relationship $\underline{h}^\star(x, y)$ within a Hilbert t-module $\mathcal{M}$. We often omit the function dependence on (x, y) in expectations for brevity. For all embeddings $\underline{\mathbf{f}}, \underline{\mathbf{g}} \in \mathcal{M}$ and distribution $\mathcal{D}$, we decompose the risk as:

$$
\begin{aligned}
\mathcal{R}(\underline{\mathbf{f}}, \underline{\mathbf{g}}; \mathcal{D}) &:= \mathbb{E}_{(x,y)\in\mathcal{D}}[\|\underline{\mathbf{f}}(x)^\top *_M \underline{\mathbf{g}}(y) - \underline{h}^\star(x,y)\|^2] \\
&= \mathbb{E}_{(x,y)\in\mathcal{D}}[\|M(\underline{\mathbf{f}}(x)^\top *_M \underline{\mathbf{g}}(y) - \underline{h}^\star(x,y))\|^2] \\
&= \mathbb{E}_{(x,y)\in\mathcal{D}}[\|\breve{\underline{\mathbf{f}}}(x)^\top \odot \breve{\underline{\mathbf{g}}}(y) - \breve{\underline{h}}^\star(x,y)\|^2] \\
&= \mathbb{E}_{(x,y)\in\mathcal{D}}\left[\sum_{i=1}^{K}\left((\breve{\underline{\mathbf{f}}}^{(i)}(x))^\top \breve{\underline{\mathbf{g}}}^{(i)}(y) - \breve{\underline{h}}^{\star(i)}(x,y)\right)^2\right] \\
&= \sum_{i=1}^{K} \underbrace{\mathbb{E}_{(x,y)\in\mathcal{D}}\left[\left((\breve{\underline{\mathbf{f}}}^{(i)}(x))^\top \breve{\underline{\mathbf{g}}}^{(i)}(y) - \breve{\underline{h}}^{\star(i)}(x,y)\right)^2\right]}_{=:\breve{\mathcal{R}}_i(\underline{\mathbf{f}},\underline{\mathbf{g}};\mathcal{D})}.
\end{aligned}
$$

In words, the risk $\mathcal{R}(\underline{\mathbf{f}}, \underline{\mathbf{g}}; \mathcal{D})$ measures the expected squared difference between the predicted output $\underline{\mathbf{f}}(x)^\top *_M \underline{\mathbf{g}}(y)$ and the true output $\underline{h}^\star(x, y)$ under the distribution $\mathcal{D}$. It quantifies how well the embeddings $\underline{\mathbf{f}}$ and $\underline{\mathbf{g}}$ approximate the true relationship $\underline{h}^\star$ on average. **The key idea in the risk decomposition is to transform the risk into the frequency domain using the transform $M(\cdot)$ and then decompose it into a sum of risks over individual frequency components, such that the tools developed in Ref. [46] can be applied.** This is achieved by expressing the risk in terms of the frontal slices of the transformed embeddings $\breve{\underline{\mathbf{f}}}$ and $\breve{\underline{\mathbf{g}}}$, and the transformed true output $\breve{\underline{h}}^\star$. By defining the risk $\breve{\mathcal{R}}_i(\underline{\mathbf{f}}, \underline{\mathbf{g}}; \mathcal{D})$ for each frequency component i, we can analyze the performance of the embeddings in a more fine-grained manner. This decomposition allows us to study the behavior of the embeddings in different frequency components and identify the sources of error that contribute to the overall risk. Finally, we can express the total risk $\mathcal{R}(\hat{\underline{\mathbf{f}}}, \hat{\underline{\mathbf{g}}}; \mathcal{D}_{\text{test}})$ as the sum of the risks $\breve{\mathcal{R}}_i(\hat{\underline{\mathbf{f}}}, \hat{\underline{\mathbf{g}}}; \mathcal{D}_{\text{test}})$ over all frequency components.

We also define some key quantities related to the spectral properties of the embeddings and the t-covariance operators, including the full-multi-rank embeddings, the t-singular values $\underline{\sigma}_j(\underline{\mathbf{f}}, \underline{\mathbf{g}})$ and $\sigma_j(\breve{\underline{\mathbf{f}}}^{(i)}, \breve{\underline{\mathbf{g}}}^{(i)})$, the tail sums $\textsf{tail}_q^{(i)\star}$, $\textsf{tail}_q^\star(k)$, and $\textsf{tail}_q^\star(\mathbf{k})$ as follows.

- Full-multi-rank embeddings: Given a vector $\mathbf{r} = (r_1, \ldots, r_K)^\top \in \mathbb{N}^K$, we say that a pair of $\mathbb{R}^{\|\mathbf{r}\|_\infty \times 1 \times K}$-embeddings $(\underline{\mathbf{f}}, \underline{\mathbf{g}})$ are *full-multi-rank*-$\mathbf{r}$ if the r_i-th singular value $\sigma_{r_i}(\breve{\hat{\underline{\mathbf{f}}}}^{(i)}, \breve{\hat{\underline{\mathbf{g}}}}^{(i)})$ is strictly positive for each frequency component $i \in [K]$. This condition ensures that the embeddings have sufficient expressive power and that their spectral properties are well-defined in each frequency component.

- t-singular values $\underline{\sigma}_j(\underline{\mathbf{f}}, \underline{\mathbf{g}})$ and $\sigma_j(\breve{\underline{\mathbf{f}}}^{(i)}, \breve{\underline{\mathbf{g}}}^{(i)})$: The t-singular values $\underline{\sigma}_j(\underline{\mathbf{f}}, \underline{\mathbf{g}})$ are defined as the t-singular values of the product of the square roots of the covariance operators $\mathbb{E}_{\mathcal{D}_{\mathcal{X},1}}[\underline{\mathbf{f}} *_M \underline{\mathbf{f}}^\top]$ and $\mathbb{E}_{\mathcal{D}_{\mathcal{Y},1}}[\underline{\mathbf{g}} *_M \underline{\mathbf{g}}^\top]$:

  $$\underline{\sigma}_j(\underline{\mathbf{f}}, \underline{\mathbf{g}}) := \underline{\sigma}_j\left(\mathbb{E}_{\mathcal{D}_{\mathcal{X},1}}[\underline{\mathbf{f}} *_M \underline{\mathbf{f}}^\top]^{\frac{1}{2}} *_M \mathbb{E}_{\mathcal{D}_{\mathcal{Y},1}}[\underline{\mathbf{g}} *_M \underline{\mathbf{g}}^\top]^{\frac{1}{2}}\right).$$

  Similarly, $\sigma_j(\breve{\underline{\mathbf{f}}}^{(i)}, \breve{\underline{\mathbf{g}}}^{(i)})$ denotes the singular values of the product of the square roots of the covariance operators in the i-th frequency component:

  $$\sigma_j(\breve{\underline{\mathbf{f}}}^{(i)}, \breve{\underline{\mathbf{g}}}^{(i)}) := \sigma_j\left(\mathbb{E}_{\mathcal{D}_{\mathcal{X},1}}[\breve{\underline{\mathbf{f}}}^{(i)}(\breve{\underline{\mathbf{f}}}^{(i)})^\top]^{\frac{1}{2}} \cdot \mathbb{E}_{\mathcal{D}_{\mathcal{Y},1}}[\breve{\underline{\mathbf{g}}}^{(i)}(\breve{\underline{\mathbf{g}}}^{(i)})^\top]^{\frac{1}{2}}\right).$$

  These singular values capture the interactions between the left and right embeddings in each individual frequency component, providing a more fine-grained characterization of their relationship.

- Tail sums $\textsf{tail}_q^{(i)\star}(k_i)$, $\textsf{tail}_q^\star(k)$, and $\textsf{tail}_q^\star(\mathbf{k})$: The tail sums measure the decay of the singular values in different frequency components and provide a way to quantify the complexity of the embeddings.

  (a) $\textsf{tail}_q^{(i)\star}(k_i)$ denotes the sum of the q-th powers of the singular values beyond index k_i in the i-th frequency component:

  $$\textsf{tail}_q^{(i)\star}(k_i) := \sum_{j_i > k_i} \left(\breve{\sigma}_{j_i}^{\star,(i)}\right)^q, \quad q \geq 1,$$

  where $\breve{\sigma}_j^{\star,(i)} := \lambda_j(M(\underline{\mathbf{\Sigma}}_{1\otimes 1}^\star)^{(i)})$ are the singular values of the covariance operator $\underline{\mathbf{\Sigma}}_{1\otimes 1}^\star$ in the i-th frequency component.

  (b) $\textsf{tail}_q^\star(k)$ denotes the sum of the q-th powers of the singular values beyond index k across all frequency components:

  $$\textsf{tail}_q^\star(k) := \sum_{i=1}^{K} \sum_{j > k} \left(\breve{\sigma}_j^{\star,(i)}\right)^q, \quad q \geq 1.$$

(c) Given a vector $\mathbf{k} = (k_1, \dots, k_K)^\top \in \mathbb{N}^K$, $\textsf{tail}^\star_q(\mathbf{k})$ is a more general version of $\textsf{tail}^\star_q(k)$, where the summation is taken beyond index k_i in each frequency component i:

$$\textsf{tail}^\star_q(\mathbf{k}) := \sum_{i=1}^{K} \sum_{j_i > k_i} \left(\breve{\sigma}^{\star,(i)}_{j_i} \right)^q, \quad q \geq 1.$$

These tail sums capture the complexity of the embeddings by measuring the decay of their singular values. A rapidly decaying spectrum (i.e., small tail sums) indicates that the embeddings have a low-rank structure and can be well-approximated by a small number of principal components. Conversely, slowly decaying tail sums suggest that the embeddings are more complex and require a larger number of components to capture their variability.

Furthermore, we introduce the multi-rank-$\mathbf{k}$ t-projection operator $\underline{\mathbf{P}}^\star_{\mathbf{k}}$, which projects an embedding onto the top principal directions in each frequency component.

Definition 14 (Multi-rank-$\mathbf{k}$ t-projection operator)**.** *Let* $\mathbf{k} = (k_1, \cdots, k_K)^\top \in \mathbb{N}^K$. *The multi-rank-*$\mathbf{k}$ *projection operator* $\underline{\mathbf{P}}^\star_{\mathbf{k}} : \mathcal{M} \to \mathcal{M}$ *on the range of* $\underline{\mathbf{\Sigma}}^\star_{1\otimes 1}$ *satisfies*

$$(M(\underline{\mathbf{P}}^\star_{\mathbf{k}} \underline{\mathbf{f}}))^{(i)} = \breve{\mathbf{P}}_{k_i} \breve{\underline{\mathbf{f}}}^{(i)}, \forall i \in [K], \tag{22}$$

where $\breve{\mathbf{P}}_{k_i}$ *is the orthogonal projection operator onto the top-*k_i *eigenspaces of* $\operatorname{range}(M(\underline{\mathbf{\Sigma}}^\star_{1\otimes 1})^{(i)})$. $\underline{\mathbf{P}}^\star_{\mathbf{k}}$ *projects the embedding* $\underline{\mathbf{f}}$ *onto the top* k_i *principal directions in each frequency component, serving the purpose of dimensionality reduction and extracting key information.*

Definition 14 introduces the concept of the multi-rank-$\mathbf{k}$ t-projection operator $\underline{\mathbf{P}}^\star_{\mathbf{k}}$, which plays a crucial role in dimensionality reduction and information extraction within the Hilbert t-module framework. This operator is designed to capture the most significant components of an embedding $\underline{\mathbf{f}}$ by projecting it onto the top k_i principal directions in each frequency component.

The $\underline{\mathbf{P}}^\star_{\mathbf{k}}$ operator offers a principled approach to tackle the CDS challenge by exploiting the low-rank structure of the multi-output functions in the transformed domain. By projecting the embeddings onto the top k_i principal directions in each frequency component, $\underline{\mathbf{P}}^\star_{\mathbf{k}}$ effectively captures the most significant patterns and correlations present in the data, even in the presence of CDS. The key idea is that the low-rank structure of the multi-output functions, as revealed by the spectral decomposition of the covariance operator $\underline{\mathbf{\Sigma}}^\star_{1\otimes 1}$, remains relatively stable across different feature combinations. In other words, the principal directions along which the functions exhibit the highest variability or energy concentration are likely to be shared among different combinations of input features.

Lemma D.1 (Properties of $\underline{\mathbf{P}}^\star_{\mathbf{k}}$)**.** *Let* $\mathbf{k} = (k_1, \cdots, k_K)^\top \in \mathbb{N}^K$. *The multi-rank-*$\mathbf{k}$ *projection operator* $\underline{\mathbf{P}}^\star_{\mathbf{k}} : \mathcal{M} \to \mathcal{M}$ *has the following properties:*

1. *Self-adjointness:* $\langle \underline{\mathbf{P}}^\star_{\mathbf{k}} \underline{\mathbf{f}}, \underline{\mathbf{g}} \rangle_{\mathcal{M}} = \langle \underline{\mathbf{f}}, \underline{\mathbf{P}}^\star_{\mathbf{k}} \underline{\mathbf{g}} \rangle_{\mathcal{M}}$.
2. *Idempotence:* $\underline{\mathbf{P}}^\star_{\mathbf{k}} \underline{\mathbf{P}}^\star_{\mathbf{k}} \underline{\mathbf{f}} = \underline{\mathbf{P}}^\star_{\mathbf{k}} \underline{\mathbf{f}}$.

Proof. By definition, we have

$$\left(M(\langle \underline{\mathbf{P}}^\star_{\mathbf{k}} \underline{\mathbf{f}}, \underline{\mathbf{g}} \rangle_{\mathcal{M}}) \right)^{(i)} = \langle (M(\underline{\mathbf{P}}^\star_{\mathbf{k}} \underline{\mathbf{f}}))^{(i)}, \breve{\underline{\mathbf{g}}}^{(i)} \rangle_{\mathcal{H}_i} = \langle \breve{\mathbf{P}}_{k_i} \breve{\underline{\mathbf{f}}}^{(i)}, \breve{\underline{\mathbf{g}}}^{(i)} \rangle_{\mathcal{H}_i} = \langle \breve{\underline{\mathbf{f}}}^{(i)}, \breve{\mathbf{P}}_{k_i} \breve{\underline{\mathbf{g}}}^{(i)} \rangle_{\mathcal{H}_i},$$

holds for all $i \in [K]$ where $\langle \cdot, \cdot \rangle_{\mathcal{H}_i}$ denotes the inner product of Hilbert space defined in Definition 13. Self-adjointness and idempotence reflect the basic properties of projection operators. □

Definition 15 below introduces the concepts of the balanced t-embeddings and balancing t-operator. The main idea behind these concepts is to balance the left and right t-embeddings in a way that preserves their spectral properties while simplifying the analysis.

Definition 15 (Balanced t-embeddings, balancing t-operator)**.** *Given a fixed vector of integers* $\mathbf{r} = (r_1, \cdots, r_K)^\top \in \mathbb{N}^K$. *For a pair of* $\mathbb{R}^{\|\mathbf{r}\|_\infty \times 1 \times K}$*-embeddings* $(\hat{\underline{\mathbf{f}}}, \hat{\underline{\mathbf{g}}})$*, if their t-covariance tensors* $\mathbb{E}_{\mathcal{D}_{\mathcal{X}},1}[\hat{\underline{\mathbf{f}}} *_M \hat{\underline{\mathbf{f}}}^\top] = \mathbb{E}_{\mathcal{D}_{\mathcal{Y}},1}[\hat{\underline{\mathbf{g}}} *_M \hat{\underline{\mathbf{g}}}^\top]$*, then* $(\hat{\underline{\mathbf{f}}}, \hat{\underline{\mathbf{g}}})$ *is said to be balanced t-embeddings.*

For full-multi-rank-$\mathbf{r}$ $\mathbb{R}^{\|\mathbf{r}\|_\infty \times 1 \times K}$*-embeddings* $(\hat{\underline{\mathbf{f}}}, \hat{\underline{\mathbf{g}}})$*, an operator* $\underline{\mathbf{T}} : \mathbb{R}^{\|\mathbf{r}\|_\infty \times 1 \times K} \to \mathbb{R}^{\|\mathbf{r}\|_\infty \times 1 \times K}$ *is defined as a balancing t-operator if it satisfies: (a) the pair of t-embeddings* $(\tilde{\underline{\mathbf{f}}}, \tilde{\underline{\mathbf{g}}}) = (\underline{\mathbf{T}}^{-1} \hat{\underline{\mathbf{f}}}, \underline{\mathbf{T}} \hat{\underline{\mathbf{g}}})$ *is* *balanced, and (b)* $\sigma_{r_i}(M(\mathbb{E}_{\mathcal{D}_{\mathcal{X}},1}[\tilde{\underline{\mathbf{f}}} *_M \tilde{\underline{\mathbf{f}}}^\top])^{(i)}) = \sigma_{r_i}(M(\mathbb{E}_{\mathcal{D}_{\mathcal{Y}},1}[\tilde{\underline{\mathbf{g}}} *_M \tilde{\underline{\mathbf{g}}}^\top])^{(i)}) = \sigma_{r_i}(\breve{\hat{\underline{\mathbf{f}}}}^{(i)}, \breve{\hat{\underline{\mathbf{g}}}}^{(i)})$ *for all* $i \in [K]$.

The key property of the balanced t-embeddings is that the r_i-th singular values of the t-covariance tensors $\mathbb{E}_{\mathcal{D}_{\mathcal{X}},1}[\tilde{\underline{\mathbf{f}}} *_M \tilde{\underline{\mathbf{f}}}^\top]$ and $\mathbb{E}_{\mathcal{D}_{\mathcal{Y}},1}[\tilde{\underline{\mathbf{g}}} *_M \tilde{\underline{\mathbf{g}}}^\top]$ in each frequency component $i \in [K]$ are equal to the r_i-th singular value of the original embeddings $(\breve{\hat{\underline{\mathbf{f}}}}^{(i)}, \breve{\hat{\underline{\mathbf{g}}}}^{(i)})$. Intuitively, the balancing operator helps to align the spectral properties of the left and right embeddings, making them more compatible and easier to analyze. By balancing the t-embeddings, we can focus on their essential characteristics and ignore any irrelevant or redundant information that may complicate the analysis.

The balancing t-operator plays a crucial role in the definition of aligned $\mathbf{k}$-proxies.

Definition 16 (Aligned proxies for t-embeddings)**.** *Given a fixed vector of integers* $\mathbf{r} = (r_1, \cdots, r_K)^\top \in \mathbb{N}^K$. *We say* $\iota_{\mathbf{r}} : \mathbb{R}^{\|\mathbf{r}\|_\infty \times 1 \times K} \to \mathcal{M}$ *is a t-isometric inclusion if it preserves t-inner products, i.e.,* $\underline{\mathbf{v}}^\top *_M \underline{\mathbf{w}} = \langle \iota_{\mathbf{r}}(\underline{\mathbf{v}}), \iota_{\mathbf{r}}(\underline{\mathbf{w}}) \rangle_{\mathcal{M}}$. *Let* $\hat{\underline{\mathbf{f}}} : \mathcal{X} \to \mathbb{R}^{\|\mathbf{r}\|_\infty \times 1 \times K}$ *and* $\hat{\underline{\mathbf{g}}} : \mathcal{Y} \to \mathbb{R}^{\|\mathbf{r}\|_\infty \times 1 \times K}$ *be full-multi-rank-*$\mathbf{r}$. *We say* $(\underline{\mathbf{f}}, \underline{\mathbf{g}})$ *are aligned* $\mathbf{k}$*-proxies for* $(\hat{\underline{\mathbf{f}}}, \hat{\underline{\mathbf{g}}})$ *if: (a)* $\underline{\mathbf{f}} = (\iota_{\mathbf{r}} \circ \underline{\mathbf{T}}^{-1})\hat{\underline{\mathbf{f}}}$, $\underline{\mathbf{g}} = (\iota_{\mathbf{r}} \circ \underline{\mathbf{T}})\hat{\underline{\mathbf{g}}}$, *where* $\iota_{\mathbf{r}} : \mathbb{R}^{\|\mathbf{r}\|_\infty \times 1 \times K} \to \mathcal{M}$ *is a t-isometric inclusion, and* $\underline{\mathbf{T}}$ *is the balancing t-operator in Definition 15, and (b) for*[14] $\underline{\mathbf{P}}^\star_{\mathbf{k}}$ *being the multi-rank-*$\mathbf{k}$ *t-projection operator defined by* $\underline{\mathbf{\Sigma}}^\star_{1\otimes 1}$, *we have*

$$\mathrm{range}(\underline{\mathbf{P}}^\star_{\mathbf{k}}) \subseteq \mathrm{range}(\mathbb{E}_{\mathcal{D}_{\mathcal{X}},1}[\underline{\mathbf{f}} *_M \underline{\mathbf{f}}^\top]). \tag{23}$$

The concepts of t-isometric inclusion and aligned $\mathbf{k}$-proxies are essential for establishing the generalization bounds for the ERM-DS algorithm. These concepts allow us to relate the learned embeddings to the true embeddings and control the approximation error.

Definition 17 ($\boldsymbol{\alpha}$-Conditioned embeddings)**.** *Given vectors* $\mathbf{r} = (r_i)_{i=1}^K \in \mathbb{N}^K$ *and* $\boldsymbol{\alpha} = (\alpha_i)_{i=1}^K \in \mathbb{R}^K$ *where* $\alpha_i \geq 1$ *for all* $i \in [K]$, *we say* $\mathbb{R}^{\|\mathbf{r}\|_\infty \times 1 \times K}$*-embeddings* $(\hat{\underline{\mathbf{f}}}, \hat{\underline{\mathbf{g}}})$ *are* $\boldsymbol{\alpha}$*-conditioned with multi-rank* $\mathbf{r}$ *if*

$$\sigma_{r_i}(\breve{\hat{\underline{\mathbf{f}}}}^{(i)}, \breve{\hat{\underline{\mathbf{g}}}}^{(i)})^2 \geq \sigma^\star_{r_i}(\breve{\underline{\mathbf{f}}}^{\star,(i)}, \breve{\underline{\mathbf{g}}}^{\star,(i)})^2/\alpha_i, \quad \forall i \in [K],$$

where $\sigma_{r_i}(\breve{\hat{\underline{\mathbf{f}}}}^{(i)}, \breve{\hat{\underline{\mathbf{g}}}}^{(i)})$ *denotes the* r_i*-th singular value of the learned embeddings* $(\hat{\underline{\mathbf{f}}}, \hat{\underline{\mathbf{g}}})$ *in the* i*-th frequency component, and* $\sigma^\star_{r_i}(\breve{\underline{\mathbf{f}}}^{\star,(i)}, \breve{\underline{\mathbf{g}}}^{\star,(i)})$ *is the corresponding singular value of the true embeddings* $(\underline{\mathbf{f}}^\star, \underline{\mathbf{g}}^\star)$.

The $\boldsymbol{\alpha}$-conditioned property ensures that the learned embeddings $(\hat{\underline{\mathbf{f}}}, \hat{\underline{\mathbf{g}}})$ have singular values that are not too small compared to those of the true embeddings $(\underline{\mathbf{f}}^\star, \underline{\mathbf{g}}^\star)$, up to a constant factor α_i in each frequency component i. This condition guarantees that the learned embeddings capture the essential spectral properties of the true embeddings and have sufficient expressive power to approximate the true relationship between the input features and the output variables. The constants $\alpha_i \geq 1$ provide flexibility in the condition, allowing for some discrepancy between the singular values of the learned and true embeddings. Smaller values of α_i indicate a tighter condition, requiring the learned embeddings to be more closely aligned with the true embeddings in terms of their spectral properties. The $\boldsymbol{\alpha}$-conditioned property is important in the analysis of multi-output regression algorithms under CDS because it helps control the approximation error in the generalization bounds. By ensuring that the learned embeddings have sufficient expressive power, this condition enables the algorithms to effectively capture the intrinsic structure of the data and achieve good generalization performance.

Definition 18 ($(\boldsymbol{\epsilon}_{\text{trn}}, \boldsymbol{\epsilon}_{\mathcal{D}_{1\otimes 1}})$-accurate embeddings)**.** *Given vectors* $\mathbf{r} = (r_i)_{i=1}^K \in \mathbb{N}^K$, $\boldsymbol{\epsilon}_{\text{trn}} = (\breve{\epsilon}^{(i)}_{\text{trn}})_{i=1}^K \in \mathbb{R}^K$, $\boldsymbol{\epsilon}_{\mathcal{D}_{1\otimes 1}} = (\breve{\epsilon}^{(i)}_{\mathcal{D}_{1\otimes 1}})_{i=1}^K \in \mathbb{R}^K$ *where* $\breve{\epsilon}^{(i)}_{\text{trn}}, \breve{\epsilon}^{(i)}_{\mathcal{D}_{1\otimes 1}} \geq 0$ *for all* $i \in [K]$, *we say* $\mathbb{R}^{\|\mathbf{r}\|_\infty \times 1 \times K}$*-embeddings* $(\hat{\underline{\mathbf{f}}}, \hat{\underline{\mathbf{g}}})$ *are* $(\boldsymbol{\epsilon}_{\text{trn}}, \boldsymbol{\epsilon}_{\mathcal{D}_{1\otimes 1}})$*-accurate with tensor multi-rank* $\mathbf{r}$ *if*

$$\breve{\mathcal{R}}^{(i)}(\breve{\hat{\underline{\mathbf{f}}}}^{(i)}, \breve{\underline{\mathbf{g}}}^{(i)}; \mathcal{D}_{\text{train}}) \leq (\breve{\epsilon}^{(i)}_{\text{trn}})^2, \quad \forall i \in [K],$$

$$\inf_{r' \geq r_i} \breve{\mathcal{R}}^{(i)}_{[r']}(\breve{\hat{\underline{\mathbf{f}}}}^{(i)}, \breve{\hat{\underline{\mathbf{g}}}}^{(i)}; \mathcal{D}_{1\otimes 1}) \leq (\breve{\epsilon}^{(i)}_{\mathcal{D}_{1\otimes 1}})^2, \quad \forall i \in [K],$$

[14] In case of non-uniqueness, any choice of the t-projections works.

where $\breve{\mathcal{R}}^{(i)}_{[s_i]}(\breve{\hat{\underline{\mathbf{f}}}}^{(i)}, \breve{\hat{\underline{\mathbf{g}}}}^{(i)}; \mathcal{D}_{1\otimes 1})$ *is the excess risk relative to* $\breve{\underline{h}}^{\star,(i)}_{s_i} = \langle \breve{\underline{\mathbf{f}}}^{\star,(i)}_{s_i}, \breve{\underline{\mathbf{g}}}^{\star,(i)}_{s_i} \rangle$*, evaluated on the top-block distribution* $\mathcal{D}_{1\otimes 1}$*:*

$$\breve{\mathcal{R}}^{(i)}_{[s_i]}(\breve{\hat{\underline{\mathbf{f}}}}^{(i)}, \breve{\hat{\underline{\mathbf{g}}}}^{(i)}; \mathcal{D}_{1\otimes 1}) := \mathbb{E}_{(x,y)\sim\mathcal{D}_{1\otimes 1}}[(\langle \breve{\hat{\underline{\mathbf{f}}}}^{(i)}(x), \breve{\hat{\underline{\mathbf{g}}}}^{(i)}(y)\rangle_{\mathcal{H}^i} - \breve{\underline{h}}^{\star,(i)}_{s_i}(x,y))^2]. \tag{24}$$

The $(\boldsymbol{\epsilon}_{\text{trn}}, \boldsymbol{\epsilon}_{\mathcal{D}_{1\otimes 1}})$-accurate property ensures that the learned embeddings $(\hat{\underline{\mathbf{f}}}, \hat{\underline{\mathbf{g}}})$ closely approximate the true embeddings $(\underline{\mathbf{f}}^\star, \underline{\mathbf{g}}^\star)$ in terms of the excess risk on both the training distribution $\mathcal{D}_{\text{train}}$ and the top-block distribution $\mathcal{D}_{1\otimes 1}$. The first condition, $\breve{\mathcal{R}}^{(i)}(\breve{\hat{\underline{\mathbf{f}}}}^{(i)}, \breve{\hat{\underline{\mathbf{g}}}}^{(i)}; \mathcal{D}_{\text{train}}) \le (\breve{\epsilon}^{(i)}_{\text{trn}})^2$, requires that the excess risk of the learned embeddings on the training distribution is bounded by $(\breve{\epsilon}^{(i)}_{\text{trn}})^2$ in each frequency component i. This condition ensures that the learned embeddings fit the training data well and capture the underlying patterns in the data. The second condition, $\inf_{r'\ge r_i} \breve{\mathcal{R}}^{(i)}_{[r']}(\breve{\hat{\underline{\mathbf{f}}}}^{(i)}, \breve{\hat{\underline{\mathbf{g}}}}^{(i)}; \mathcal{D}_{1\otimes 1}) \le (\breve{\epsilon}^{(i)}_{\mathcal{D}_{1\otimes 1}})^2$, requires that the excess risk of the learned embeddings on the top-block distribution, relative to the best rank-r' approximation of the true embeddings for any $r' \ge r_i$, is bounded by $(\breve{\epsilon}^{(i)}_{\mathcal{D}_{1\otimes 1}})^2$ in each frequency component i. This condition ensures that the learned embeddings generalize well to the top-block distribution and can effectively capture the important spectral components of the true embeddings. The constants $\breve{\epsilon}^{(i)}_{\text{trn}}$ and $\breve{\epsilon}^{(i)}_{\mathcal{D}_{1\otimes 1}}$ quantify the degree of accuracy, with smaller values indicating a better approximation. The $(\boldsymbol{\epsilon}_{\text{trn}}, \boldsymbol{\epsilon}_{\mathcal{D}_{1\otimes 1}})$-accurate property is important in the analysis of multi-output regression algorithms under CDS because it helps control the estimation error in the generalization bounds. By ensuring that the learned embeddings closely approximate the true embeddings on the relevant distributions, this condition enables the algorithms to achieve good generalization performance and mitigate the impact of distribution shift.

D.1 Error Decomposition

In this section, we present a technical method that breaks down the excess risk of the algorithms into distinct, interpretable components, such as the approximation error, statistical variance, and terms capturing the effects of distribution shift. The main idea in this section builds on Ref. [46]. For completeness, we provide a detailed overview of the error decomposition framework used to analyze the generalization performance of multi-output regression algorithms under CDS. Our objective is to establish an upper bound on the population risk $\mathcal{R}(\hat{\underline{\mathbf{f}}}, \hat{\underline{\mathbf{g}}}; \mathcal{D}_{\text{test}})$ for the test distribution in terms of the training distribution risk and additional error terms that account for distribution shift.

To achieve this, we introduce key error terms that capture different aspects of the shift between training and test data, defining these terms to address the challenges of CDS in multi-output regression. We then present a series of lemmas that bound the test risk in terms of these error terms, which collectively lead to the final error decomposition. This culminates in Lemma D.2, summarizing the decomposition and providing insights into the algorithm's generalization behavior under CDS.

D.1.1 Key Error Terms

Definition 19. *Given functions* $\underline{\mathbf{f}} : \mathcal{X} \to \mathcal{M}$ *and* $\underline{\mathbf{g}} : \mathcal{Y} \to \mathcal{M}$ *and* $\mathbf{k} = (k_1, \cdots, k_K)^\top \in \mathbb{N}^K$*, we define several key error terms as follows:*

$$\begin{aligned}
\Delta_0(\underline{\mathbf{f}}, \underline{\mathbf{g}}, \mathbf{k}) &:= \max\left\{\mathbb{E}_{\mathcal{D}_{1\otimes 1}}\left[\|\langle \underline{\mathbf{f}}^\star_{\mathbf{k}}, \underline{\mathbf{g}}^\star_{\mathbf{k}} - \underline{\mathbf{g}}\rangle_{\mathcal{M}}\|^2\right], \mathbb{E}_{\mathcal{D}_{1\otimes 1}}\left[\|\langle \underline{\mathbf{f}}^\star_{\mathbf{k}} - \underline{\mathbf{f}}, \underline{\mathbf{g}}^\star_{\mathbf{k}}\rangle_{\mathcal{M}}\|^2\right]\right\} && \text{(weighted error)}\\
\Delta_1(\underline{\mathbf{f}}, \underline{\mathbf{g}}, \mathbf{k}) &:= \max\left\{\mathbb{E}_{\mathcal{D}_{\mathcal{X},1}}\|\underline{\mathbf{f}}^\star_{\mathbf{k}} - \underline{\mathbf{f}}\|^2_{\mathcal{M}}, \mathbb{E}_{\mathcal{D}_{\mathcal{Y},1}}\|\underline{\mathbf{g}}^\star_{\mathbf{k}} - \underline{\mathbf{g}}\|^2_{\mathcal{M}}\right\} && \text{(unweighted error)}\\
\Delta_2(\underline{\mathbf{f}}, \underline{\mathbf{g}}, \mathbf{k}) &:= \max\left\{\mathbb{E}_{\mathcal{D}_{\mathcal{X},2}}\|\underline{\mathbf{f}}^\star_{\mathbf{k}} - \underline{\mathbf{f}}\|^2_{\mathcal{M}}, \mathbb{E}_{\mathcal{D}_{\mathcal{Y},2}}\|\underline{\mathbf{g}}^\star_{\mathbf{k}} - \underline{\mathbf{g}}\|^2_{\mathcal{M}}\right\} && \text{($\mathcal{D}_{2\otimes 2}$-recovery error)}\\
\Delta_{\text{apx}}(\mathbf{k}) &:= \mathcal{R}(\underline{\mathbf{f}}^\star_{\mathbf{k}}, \underline{\mathbf{g}}^\star_{\mathbf{k}}; \mathcal{D}_{1\otimes 1}) && \text{(approximation error)}\\
\Delta_{\text{train}}(\mathbf{k}) &:= \mathcal{R}(\underline{\mathbf{f}}^\star_{\mathbf{k}}, \underline{\mathbf{g}}^\star_{\mathbf{k}}; \mathcal{D}_{\text{train}}). && \text{(error on training distribution)}
\end{aligned}$$

When it is clear from the context, we will use the shorthand notation $\Delta_0, \Delta_1, \Delta_2$*,* Δ_{apx}*, and* Δ_{train}*, respectively, for convenience.*

These error terms are essential for assessing the performance of embeddings $\underline{\mathbf{f}}$ and $\underline{\mathbf{g}}$ across different distributions:

- Weighted Error $\Delta_0(\underline{\mathbf{f}}, \underline{\mathbf{g}}, \mathbf{k})$: Measures discrepancies between true embeddings $(\underline{\mathbf{f}}^\star_{\mathbf{k}}, \underline{\mathbf{g}}^\star_{\mathbf{k}})$ and learned embeddings $(\underline{\mathbf{f}}, \underline{\mathbf{g}})$ under $\mathcal{D}_{1\otimes 1}$, focusing on the weighted differences along each direction.
- Unweighted Error $\Delta_1(\underline{\mathbf{f}}, \underline{\mathbf{g}}, \mathbf{k})$: Captures unweighted differences between true and learned embeddings under $\mathcal{D}_1$, assessing how well $(\underline{\mathbf{f}}, \underline{\mathbf{g}})$ approximate $(\underline{\mathbf{f}}^\star_{\mathbf{k}}, \underline{\mathbf{g}}^\star_{\mathbf{k}})$ without specific weighting.
- $\mathcal{D}_{2\otimes 2}$-Recovery Error $\Delta_2(\underline{\mathbf{f}}, \underline{\mathbf{g}}, \mathbf{k})$: Represents discrepancies under $\mathcal{D}_{2\otimes 2}$, evaluating differences between true and learned embeddings in this setting.
- Approximation Error $\Delta_{\mathrm{apx}}(\mathbf{k})$: Measures how well the true embeddings $(\underline{\mathbf{f}}^\star_{\mathbf{k}}, \underline{\mathbf{g}}^\star_{\mathbf{k}})$ capture the actual relationships under $\mathcal{D}_{1\otimes 1}$.
- Training Error $\Delta_{\mathrm{train}}(\mathbf{k})$: Indicates the risk of true embeddings $(\underline{\mathbf{f}}^\star_{\mathbf{k}}, \underline{\mathbf{g}}^\star_{\mathbf{k}})$ on the training distribution $\mathcal{D}_{\mathrm{train}}$, reflecting their alignment with the training data.

D.1.2 Error Decomposition Lemma

Lemma D.2 (Error decomposition on $\mathcal{D}_{\mathrm{test}}$)**.** *Suppose Assumption 2 and Assumption 3 hold. For any vector* $\mathbf{k} = (k_1, \cdots, k_K)^\top$ *satisfying* $\mathbf{k} \le \mathbf{r}$ *element-wisely with some fixed vector of integers* $\mathbf{r} = (r_1, \cdots, r_K)^\top \in \mathbb{N}^K$*, and any aligned* $\mathbf{k}$*-proxies* $(\underline{\mathbf{f}}, \underline{\mathbf{g}})$ *of the* $\mathbb{R}^{\|\mathbf{r}\|_\infty \times 1 \times K}$*-embeddings* $(\hat{\underline{\mathbf{f}}}, \hat{\underline{\mathbf{g}}})$*, denote* $\Delta_0 = \Delta_0(\underline{\mathbf{f}}, \underline{\mathbf{g}}, \mathbf{k})$ *and* $\Delta_1 = \Delta_1(\underline{\mathbf{f}}, \underline{\mathbf{g}}, \mathbf{k})$*. Let* $\sigma \le \min_{i\in[K]} \sigma_{r_i}(\breve{\hat{\underline{\mathbf{f}}}}^{(i)}, \breve{\hat{\underline{\mathbf{g}}}}^{(i)})$ *be a lower bound on* $\sigma_{r_i}(\breve{\hat{\underline{\mathbf{f}}}}^{(i)}, \breve{\hat{\underline{\mathbf{g}}}}^{(i)})$*, which satisfies* $\sigma_i^2 \in (0, \textsf{tail}_2^{(i)\star}(k_i) + \Delta_0 + \Delta_{\mathrm{train}}]$ *for all* $i \in [K]$*. Then,*

$$\mathcal{R}(\underline{\mathbf{f}}, \underline{\mathbf{g}}; \mathcal{D}_{\mathrm{test}}) \lesssim_\star \Delta_1^2 + \frac{1}{\sigma^2} \left(\Delta_{\mathrm{apx}} + \Delta_0 + \Delta_{\mathrm{train}}\right)^2 .$$

The proof closely follows that of Proposition 4.1b in Ref. [46] and proceeds by incrementally bounding the risk on the test distribution $\mathcal{D}_{\mathrm{test}}$ through a sequence of lemmas. In these lemmas, rather than applying Assumption 2 and Assumption 3 directly, we employ their relaxed forms as given in Assumption 6 and Assumption 7. This adjustment allows us to handle cases where coverage and covariate shift conditions hold only in a probabilistic or approximate sense, thus broadening the applicability of the result.

First, we apply Lemma D.3 to decompose the risk on $\mathcal{D}_{\mathrm{test}}$ into the risk on the bottom-right block distribution $\mathcal{D}_{2\otimes 2}$ and the risk on the training distribution $\mathcal{D}_{\mathrm{train}}$.

Lemma D.3 (Error decomposition on $\mathcal{D}_{\mathrm{test}}$)**.** *Under Assumption 6 and Assumption 8, the following holds for any* $\underline{\mathbf{f}} : \mathcal{X} \to \mathcal{M}$ *and* $\underline{\mathbf{g}} : \mathcal{Y} \to \mathcal{M}$*:*

$$\mathcal{R}(\underline{\mathbf{f}}, \underline{\mathbf{g}}; \mathcal{D}_{\mathrm{test}}) \le \kappa_{\mathrm{tst}} \left(\mathcal{R}(\underline{\mathbf{f}}, \underline{\mathbf{g}}; \mathcal{D}_{2\otimes 2}) + 3\kappa_{\mathrm{trn}} \mathcal{R}(\underline{\mathbf{f}}, \underline{\mathbf{g}}; \mathcal{D}_{\mathrm{train}})\right) + 4B^4(\eta_{\mathrm{tst}} + 3\kappa_{\mathrm{tst}}\eta_{\mathrm{trn}}).$$

Next, we focus on bounding the risk on $\mathcal{D}_{2\otimes 2}$ using Lemma D.4. This lemma reveals that the dominant term in the risk decomposition is the weighted error Δ_0, while the unweighted errors Δ_1 and Δ_2 contribute only quadratically:

Lemma D.4 (Error decomposition on $\mathcal{D}_{2\otimes 2}$)**.** *Under Assumption 6, Assumption 7 and Assumption 8, for any* $f : \mathcal{X} \to \mathcal{M}$*,* $g : \mathcal{Y} \to \mathcal{M}$*, and* $\mathbf{k} \in \mathbb{N}^K$*:*

$$\mathcal{R}(\underline{\mathbf{f}}, \underline{\mathbf{g}}; \mathcal{D}_{2\otimes 2}) \lesssim \kappa_{\mathrm{cov}}^2(\Delta_0 + (\Delta_1)^2 + \Delta_{\mathrm{apx}}) + (\Delta_2)^2 + \kappa_{\mathrm{cov}}\kappa_{\mathrm{trn}}\Delta_{\mathrm{train}} + B^4\kappa_{\mathrm{cov}}(\eta_{\mathrm{cov}} + \eta_{\mathrm{trn}}),$$

where above we suppress error term dependence on $\underline{\mathbf{f}}, \underline{\mathbf{g}}, \mathbf{k}$*.*

To bound the term Δ_2, we leverage the assumption that $(\underline{\mathbf{f}}, \underline{\mathbf{g}})$ are aligned $\mathbf{k}$-proxies of $(\hat{\underline{\mathbf{f}}}, \hat{\underline{\mathbf{g}}})$, along with the construction in Definition 16. This allows us to apply Lemma D.5 and obtain a bound on Δ_2 in terms of the training error Δ_{train}, the weighted error Δ_0, and the approximation error $\Delta_{\mathrm{apx}}(\mathbf{k})$:

Going forward, recall the construction in Definition 16 that

$$\sigma_{r_i}(\breve{\hat{\underline{\mathbf{f}}}}^{(i)}, \breve{\hat{\underline{\mathbf{g}}}}^{(i)}) := \sigma_{r_i}(M(\mathbb{E}_{\mathcal{D}_{\mathcal{X},1}}[\underline{\mathbf{f}} *_M \underline{\mathbf{f}}^\top])^{(i)}), \quad \forall i \in [K].$$

where the positivity is a consequence of the assumption that $(\hat{\underline{\mathbf{f}}}, \hat{\underline{\mathbf{g}}})$ are full-multi-rank. We may now bound Δ_2.

Lemma D.5 (Decomposition of Δ_2)**.** *Suppose* $(\underline{\mathbf{f}}, \underline{\mathbf{g}})$ *are aligned* $\mathbf{k}$*-proxies of* $(\hat{\underline{\mathbf{f}}}, \hat{\underline{\mathbf{g}}})$*. Then, we have*

$$(\min_i \{\sigma_{r_i}(\breve{\hat{\underline{\mathbf{f}}}}^{(i)}, \breve{\hat{\underline{\mathbf{g}}}}^{(i)})\})\Delta_2 \lesssim \Delta_{\text{train}} + \kappa_{\text{cov}}(\Delta_0 + \Delta_{\text{apx}}(\mathbf{k})) + B^2(\eta_{\text{cov}} + \eta_{\text{tst}}),$$

where above we suppress dependence on $\underline{\mathbf{f}}, \underline{\mathbf{g}}, \mathbf{k}$ *in all error terms.*

Lastly, it is noteworthy that the multi-rank-$\mathbf{k}$ approximation error under the $\mathcal{D}_{1\otimes 1}$ distribution is precisely the tail term $\textsf{\textbf{tail}}_2^\star(\mathbf{k})$. We can exploit this crucial fact by replacing the $\Delta_{\text{apx}}(\mathbf{k})$ term in the previous error decompositions with $\textsf{\textbf{tail}}_2^\star(\mathbf{k})$.

Lemma D.6. *We have* $\Delta_{\text{apx}}(\mathbf{k}) = \mathcal{R}(\underline{\mathbf{f}}_{\mathbf{k}}^\star, \underline{\mathbf{g}}_{\mathbf{k}}^\star; \mathcal{D}_{1\otimes 1}) = \textsf{\textbf{tail}}_2^\star(\mathbf{k})$.

Proof of Lemma D.2. First, we apply Lemma D.3 to decompose the risk on $\mathcal{D}_{\text{test}}$ into the risks on $\mathcal{D}_{2\otimes 2}$ and $\mathcal{D}_{\text{train}}$. We then further decompose the risk on $\mathcal{D}_{2\otimes 2}$ into various error terms using Lemma D.4. Next, we bound Δ_2 by leveraging Lemma D.5 and the assumption that $(\underline{\mathbf{f}}, \underline{\mathbf{g}})$ is an aligned $\mathbf{k}$-agent of $(\hat{\underline{\mathbf{f}}}, \hat{\underline{\mathbf{g}}})$. Substituting the bound on Δ_2 back into the risk decomposition on $\mathcal{D}_{2\otimes 2}$, and applying Lemma D.6 to replace $\Delta_{\text{apx}}(\mathbf{k})$ with $\textsf{\textbf{tail}}_2^\star(\mathbf{k})$. Finally, we combine the error terms using the condition on σ to obtain the final result. □

D.1.3 Proof of Lemma D.3

Lemma D.7. *Under Assumption 6, we have for any* $(i, j) \in \{(1,1),(1,2),(2,1)\}$

$$\mathcal{R}(\underline{\mathbf{f}}, \underline{\mathbf{g}}; \mathcal{D}_{i\otimes j}) \leq 4B^4\eta_{\text{trn}} + \kappa_{\text{trn}}\mathcal{R}(\underline{\mathbf{f}}, \underline{\mathbf{g}}; \mathcal{D}_{\text{train}}).$$

Proof of Lemma D.7. To bound the risk under $\mathcal{D}_{i\otimes j}$ in terms of the training distribution, we introduce a density-based event and apply it to control expectations.

Define the event $\mathcal{E}_{\text{train}, i\otimes j}$ for any $(i, j) \in \{(1,1),(1,2),(2,1)\}$ as follows:

$$\mathcal{E}_{\text{train}, i\otimes j} := \left\{ \frac{\mathrm{d}\mathcal{D}_{i\otimes j}(x,y)}{\mathrm{d}\mathcal{D}_{\text{train}}(x,y)} \leq \kappa_{\text{trn}} \right\}.$$

This event limits the density ratio between $\mathcal{D}_{i\otimes j}$ and $\mathcal{D}_{\text{train}}$, ensuring that $\mathcal{D}_{i\otimes j}$ does not deviate excessively from the training distribution.

Now, for any function $\underline{h} : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}^{1\times 1\times K}$ with $\|\underline{h}(x,y)\| \leq M$, we proceed to bound the expectation under $\mathcal{D}_{i\otimes j}$:

$$\mathbb{E}_{\mathcal{D}_{i\otimes j}}[\|\underline{h}(x,y)\|^2] \leq M^2\mathbb{P}_{\mathcal{D}_{i\otimes j}}[\neg\mathcal{E}_{\text{train}, i\otimes j}] + \mathbb{E}_{\mathcal{D}_{i\otimes j}}[\|\underline{h}(x,y)\|^2\mathbb{I}\{\mathcal{E}_{\text{train}, i\otimes j}\}].$$

The first term represents cases where $\mathcal{E}_{\text{train}, i\otimes j}$ does not hold, bounded by $M^2\mathbb{P}_{\mathcal{D}_{i\otimes j}}[\neg\mathcal{E}_{\text{train}, i\otimes j}]$. For the second term, we leverage $\mathcal{E}_{\text{train}, i\otimes j}$ to express the expectation under $\mathcal{D}_{\text{train}}$:

$$\mathbb{E}_{\mathcal{D}_{i\otimes j}}[\|\underline{h}(x,y)\|^2\mathbb{I}\{\mathcal{E}_{\text{train}, i\otimes j}\}] = \mathbb{E}_{\mathcal{D}_{\text{train}}}\left[\|\underline{h}(x,y)\|^2\mathbb{I}\{\mathcal{E}_{\text{train}, i\otimes j}\} \cdot \frac{\mathrm{d}\mathcal{D}_{i\otimes j}(x,y)}{\mathrm{d}\mathcal{D}_{\text{train}}(x,y)}\right].$$

Since $\mathcal{E}_{\text{train}, i\otimes j}$ holds in this term, the density ratio is bounded by κ_{trn}, giving:

$$\mathbb{E}_{\mathcal{D}_{i\otimes j}}[\|\underline{h}(x,y)\|^2] \leq M^2\mathbb{P}_{\mathcal{D}_{i\otimes j}}[\neg\mathcal{E}_{\text{train}, i\otimes j}] + \mathbb{E}_{\mathcal{D}_{\text{train}}}[\kappa_{\text{trn}}\|\underline{h}(x,y)\|^2].$$

Substituting $\mathbb{P}_{\mathcal{D}_{i\otimes j}}[\neg\mathcal{E}_{\text{train}, i\otimes j}] \leq \eta_{\text{trn}}$ yields:

$$\mathbb{E}_{\mathcal{D}_{i\otimes j}}[\|\underline{h}(x,y)\|^2] \leq M^2\eta_{\text{trn}} + \kappa_{\text{trn}}\mathbb{E}_{\mathcal{D}_{\text{train}}}[\|\underline{h}(x,y)\|^2].$$

Now, let $\underline{h}(x,y) = \langle \underline{\mathbf{f}}(x), \underline{\mathbf{g}}(y)\rangle_{\mathcal{M}} - \underline{h}^\star(x,y)$. Note that

$$\begin{aligned} \|\underline{h}(x,y)\|^2 &= \|\langle \underline{\mathbf{f}}(x), \underline{\mathbf{g}}(y)\rangle_{\mathcal{M}} - \underline{h}^\star(x,y)\|^2 \\ &\leq 2(\|\langle \underline{\mathbf{f}}(x), \underline{\mathbf{g}}(y)\rangle_{\mathcal{M}}\|^2 + \|\langle \underline{\mathbf{f}}^\star(x), \underline{\mathbf{g}}^\star(y)\rangle_{\mathcal{M}}\|^2) \\ &\leq 4B^4. \end{aligned} \tag{25}$$

Then, $\|\underline{h}(x,y)\|^2$ is bounded by $4B^4$, giving:

$$\mathbb{E}_{\mathcal{D}_{i\otimes j}}[\|\underline{h}(x,y)\|^2] \leq 4B^4\eta_{\text{trn}} + \kappa_{\text{trn}}\mathcal{R}(\underline{\mathbf{f}}, \underline{\mathbf{g}}; \mathcal{D}_{\text{train}}),$$

which completes the proof. □

Proof of Lemma D.3. To bound the excess risk under Assuption 6, we decompose the expected norm of a function under $\mathcal{D}_{\text{test}}$ by defining a density-based event and applying it to control expectations.

First, we define the density-based event $\mathcal{E}_{\text{test}}$ to control the ratio between the density of $\mathcal{D}_{\text{test}}$ and a mixture of reference distributions $\mathcal{D}_{i\otimes j}$:

$$\mathcal{E}_{\text{test}} := \left\{ \frac{\mathrm{d}\mathcal{D}_{\text{test}}(x,y)}{\sum_{i,j\in\{1,2\}} \mathrm{d}\mathcal{D}_{i\otimes j}(x,y)} \leq \kappa_{\text{tst}} \right\}.$$

This event ensures that $\mathcal{D}_{\text{test}}$ does not deviate excessively from the reference distributions, allowing us to manage expectations over $\mathcal{D}_{\text{test}}$ using properties of the distributions $\mathcal{D}_{i\otimes j}$.

Next, we bound the expected norm of any function $\underline{h}(x,y) : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}^{1\times 1\times K}$ with $\|\underline{h}(x,y)\|^2 \leq M^2$ under $\mathcal{D}_{\text{test}}$. We decompose this expectation as follows:

$$\mathbb{E}_{\mathcal{D}_{\text{test}}}[\|\underline{h}(x,y)\|^2] \leq M^2\mathbb{P}_{\mathcal{D}_{\text{test}}}[\neg\mathcal{E}_{\text{test}}] + \mathbb{E}_{\mathcal{D}_{\text{test}}}[\mathbb{I}\{\mathcal{E}_{\text{test}}\}\|\underline{h}(x,y)\|^2].$$

The first term accounts for cases where $\mathcal{E}_{\text{test}}$ does not hold, bounded by $M^2\mathbb{P}_{\mathcal{D}_{\text{test}}}[\neg\mathcal{E}_{\text{test}}]$. The second term, where $\mathcal{E}_{\text{test}}$ holds, allows us to use the density ratio to write:

$$\mathbb{E}_{\mathcal{D}_{\text{test}}}[\mathbb{I}\{\mathcal{E}_{\text{test}}\}\|\underline{h}(x,y)\|^2] = \int_{(x,y)} \|\underline{h}(x,y)\|^2 \cdot \left(\sum_{i,j\in\{1,2\}} \mathrm{d}\mathcal{D}_{i\otimes j}(x,y)\right) \cdot \frac{\mathrm{d}\mathcal{D}_{\text{test}}(x,y)}{\sum_{i,j\in\{1,2\}} \mathrm{d}\mathcal{D}_{i\otimes j}(x,y)}\mathbb{I}\{\mathcal{E}_{\text{test}}\}.$$

Since $\mathcal{E}_{\text{test}}$ holds in this term, the density ratio is bounded by κ_{tst}, leading to:

$$\mathbb{E}_{\mathcal{D}_{\text{test}}}[\|\underline{h}(x,y)\|^2] \leq M^2\mathbb{P}_{\mathcal{D}_{\text{test}}}[\neg\mathcal{E}_{\text{test}}] + \kappa_{\text{tst}}\int_{(x,y)} \|\underline{h}(x,y)\|^2 \cdot \left(\sum_{i,j\in\{1,2\}} \mathrm{d}\mathcal{D}_{i\otimes j}(x,y)\right).$$

We rewrite the integral using the expectations over $\mathcal{D}_{i\otimes j}$, yielding:

$$\mathbb{E}_{\mathcal{D}_{\text{test}}}[\|\underline{h}(x,y)\|^2] \leq M^2\mathbb{P}_{\mathcal{D}_{\text{test}}}[\neg\mathcal{E}_{\text{test}}] + \kappa_{\text{tst}}\sum_{i,j=1}^{2} \mathbb{E}_{\mathcal{D}_{i\otimes j}}[\|\underline{h}(x,y)\|^2].$$

Finally, noting that $\mathbb{P}_{\mathcal{D}_{\text{test}}}[\neg\mathcal{E}_{\text{test}}] \leq \eta_{\text{tst}}$, we have:

$$\mathbb{E}_{\mathcal{D}_{\text{test}}}[\|\underline{h}(x,y)\|^2] \leq M^2\eta_{\text{tst}} + \kappa_{\text{tst}}\sum_{i,j=1}^{2} \mathbb{E}_{\mathcal{D}_{i\otimes j}}[\|\underline{h}(x,y)\|^2].$$

Now, let $\underline{h}(x,y) = \langle \underline{\mathbf{f}}(x), \underline{\mathbf{g}}(y)\rangle_{\mathcal{M}} - \underline{h}^\star(x,y)$, where $\|\underline{h}(x,y)\|^2$ is bounded by $4B^4$ by Eq. (25). This gives:

$$\mathcal{R}(f,g;\mathcal{D}_{\text{test}}) \leq 4B^4\eta_{\text{tst}} + \kappa_{\text{tst}}\sum_{i,j=1}^{2} \mathbb{E}_{\mathcal{D}_{i\otimes j}}[\|\underline{h}(x,y)\|^2].$$

By applying Lemma D.7, we further bound the sum over $\mathcal{D}_{i\otimes j}$ for all pairs except $(2,2)$:

$$\sum_{i,j\neq(2,2)} \mathcal{R}(\underline{\mathbf{f}},\underline{\mathbf{g}};\mathcal{D}_{i\otimes j}) \leq 12B^4\eta_{\text{trn}} + 3\kappa_{\text{trn}}\mathcal{R}(\underline{\mathbf{f}},\underline{\mathbf{g}};\mathcal{D}_{\text{train}}).$$

Therefore, combining these results, we conclude:

$$\mathcal{R}(\underline{\mathbf{f}},\underline{\mathbf{g}};\mathcal{D}_{\text{test}}) \leq \kappa_{\text{tst}}\left(\mathcal{R}(\underline{\mathbf{f}},\underline{\mathbf{g}};\mathcal{D}_{2\otimes 2}) + 3\kappa_{\text{trn}}\mathcal{R}(f,g;\mathcal{D}_{\text{train}})\right) + 4B^4(\eta_{\text{tst}} + 3\kappa_{\text{tst}}\eta_{\text{trn}}).$$

□

D.1.4 Proof of Lemma D.4

Lemma D.4 establishes bounds on the risk under $\mathcal{D}_{2\otimes 2}$, highlighting that the primary contribution in the risk decomposition comes from the weighted error term Δ_0. In contrast, the unweighted errors Δ_1 and Δ_2 have only a quadratic impact. Before proceeding with the proof of Lemma D.4, we first present a lemma that assists in expanding the risk term $\mathcal{R}(\underline{\mathbf{f}},\underline{\mathbf{g}};\mathcal{D}_{2\otimes 2})$ by breaking it down into more manageable components.

Lemma D.8. *For any function $\underline{h}^\star : \mathcal{X}\times\mathcal{Y}\to\mathbb{R}^{1\times 1\times K}$, and for $\underline{\mathbf{f}}_1,\underline{\mathbf{f}}_2 \in \mathcal{X}\to\mathcal{M}$ and $\underline{\mathbf{g}_1},\underline{\mathbf{g}_2} \in \mathcal{Y}\to\mathcal{M}$, we have*

$$\begin{aligned}&\|\langle\underline{\mathbf{f}}_1,\underline{\mathbf{g}_1}\rangle_{\mathcal{M}} - \underline{h}^\star\|^2\\ &\leq 2\|\langle\underline{\mathbf{f}}_2,\underline{\mathbf{g}_2}\rangle_{\mathcal{M}} - \underline{h}^\star\|^2 + 6\|\langle\underline{\mathbf{f}}_1-\underline{\mathbf{f}}_2,\underline{\mathbf{g}_2}\rangle_{\mathcal{M}}\|^2 + 6\|\langle\underline{\mathbf{f}}_2,\underline{\mathbf{g}_1}-\underline{\mathbf{g}_2}\rangle_{\mathcal{M}}\|^2 + 6\|\underline{\mathbf{f}}_1-\underline{\mathbf{f}}_2\|^2_{\mathcal{M}}\|\underline{\mathbf{g}_1}-\underline{\mathbf{g}_2}\|^2_{\mathcal{M}}.\end{aligned}$$

Proof of Lemma D.8. To simplify notation, let $\underline{h}_1 = \langle\underline{\mathbf{f}}_1,\underline{\mathbf{g}_1}\rangle_{\mathcal{M}}$ and $\underline{h}_2 = \langle\underline{\mathbf{f}}_2,\underline{\mathbf{g}_2}\rangle_{\mathcal{M}}$. Our goal is to bound the expression $\|\underline{h}_1-\underline{h}^\star\|^2$ using $\underline{h}_2$ and differences between $\underline{\mathbf{f}}_1,\underline{\mathbf{f}}_2$ and $\underline{\mathbf{g}_1},\underline{\mathbf{g}_2}$.

We start by expanding $\|\underline{h}_1-\underline{h}^\star\|^2 - \|\underline{h}_2-\underline{h}^\star\|^2$ as follows:

$$\|\underline{h}_1-\underline{h}^\star\|^2 - \|\underline{h}_2-\underline{h}^\star\|^2 = \langle\underline{h}_1-\underline{h}^\star+\underline{h}_2-\underline{h}^\star,\underline{h}_1-\underline{h}_2\rangle.$$

This identity allows us to split the difference in terms of the inner products of $\underline{h}_1-\underline{h}_2$ and each of the terms $\underline{h}_1-\underline{h}^\star$ and $\underline{h}_2-\underline{h}^\star$.

Next, applying the Cauchy-Schwarz inequality, we get:

$$\|\underline{h}_1-\underline{h}^\star\|^2 - \|\underline{h}_2-\underline{h}^\star\|^2 \leq \|\underline{h}_1-\underline{h}_2\|^2 + 2\|\underline{h}_2-\underline{h}^\star\|\cdot\|\underline{h}_1-\underline{h}_2\|.$$

Expanding and rearranging terms, we conclude that:

$$\|\underline{h}_1-\underline{h}^\star\|^2 \leq 2\|\underline{h}_1-\underline{h}_2\|^2 + 2\|\underline{h}_2-\underline{h}^\star\|^2.$$

To bound $\|\underline{h}_1-\underline{h}_2\|^2$, we expand this term as:

$$\begin{aligned}\|\underline{h}_1-\underline{h}_2\|^2 &= \|\langle\underline{\mathbf{f}}_1,\underline{\mathbf{g}_1}\rangle_{\mathcal{M}} - \langle\underline{\mathbf{f}}_2,\underline{\mathbf{g}_2}\rangle_{\mathcal{M}}\|^2\\ &= \|\langle\underline{\mathbf{f}}_1-\underline{\mathbf{f}}_2,\underline{\mathbf{g}_2}\rangle_{\mathcal{M}} + \langle\underline{\mathbf{f}}_2,\underline{\mathbf{g}_1}-\underline{\mathbf{g}_2}\rangle_{\mathcal{M}} + \langle\underline{\mathbf{f}}_1-\underline{\mathbf{f}}_2,\underline{\mathbf{g}_1}-\underline{\mathbf{g}_2}\rangle_{\mathcal{M}}\|^2.\end{aligned}$$

Using the Cauchy-Schwarz inequality on each term separately, we obtain:

$$\|\underline{h}_1-\underline{h}_2\|^2 \leq 3\|\langle\underline{\mathbf{f}}_1-\underline{\mathbf{f}}_2,\underline{\mathbf{g}_2}\rangle_{\mathcal{M}}\|^2 + 3\|\langle\underline{\mathbf{f}}_2,\underline{\mathbf{g}_1}-\underline{\mathbf{g}_2}\rangle_{\mathcal{M}}\|^2 + 3\|\langle\underline{\mathbf{f}}_1-\underline{\mathbf{f}}_2,\underline{\mathbf{g}_1}-\underline{\mathbf{g}_2}\rangle_{\mathcal{M}}\|^2.$$

Finally, applying the Cauchy-Schwarz inequality again to the last term, we have:

$$\|\underline{h}_1-\underline{h}_2\|^2 \leq 3\|\langle\underline{\mathbf{f}}_1-\underline{\mathbf{f}}_2,\underline{\mathbf{g}_2}\rangle_{\mathcal{M}}\|^2 + 3\|\langle\underline{\mathbf{f}}_2,\underline{\mathbf{g}_1}-\underline{\mathbf{g}_2}\rangle_{\mathcal{M}}\|^2 + 3\|\underline{\mathbf{f}}_1-\underline{\mathbf{f}}_2\|^2_{\mathcal{M}}\|\underline{\mathbf{g}_1}-\underline{\mathbf{g}_2}\|^2_{\mathcal{M}}.$$

Combining this with our earlier bound on $\|\underline{h}_1-\underline{h}^\star\|^2$, we arrive at the final result:

$$\|\underline{h}_1-\underline{h}^\star\|^2 \leq 2\|\langle\underline{\mathbf{f}}_2,\underline{\mathbf{g}_2}\rangle_{\mathcal{M}} - \underline{h}^\star\|^2 + 6\|\langle\underline{\mathbf{f}}_1-\underline{\mathbf{f}}_2,\underline{\mathbf{g}_2}\rangle_{\mathcal{M}}\|^2 + 6\|\langle\underline{\mathbf{f}}_2,\underline{\mathbf{g}_1}-\underline{\mathbf{g}_2}\rangle_{\mathcal{M}}\|^2 + 6\|\underline{\mathbf{f}}_1-\underline{\mathbf{f}}_2\|^2_{\mathcal{M}}\|\underline{\mathbf{g}_1}-\underline{\mathbf{g}_2}\|^2_{\mathcal{M}}.$$

This completes the proof. □

Proof of Lemma D.4. The proof approach follows the structure of Lemma L2.b in Ref. [46]. The primary distinction is that our proof deals with t-embeddings within the Hilbert t-Module $\mathcal{M}$ for multi-output regression, while Lemma L2.b in Ref. [46] addresses single-output learning.

Step 1: Tackling covariate shift under $\mathcal{D}_{2\otimes 2}$. To analyze the impact of covariate shift under $\mathcal{D}_{2\otimes 2}$, let's set $\underline{\mathbf{f}}_1 = \underline{\mathbf{f}}$, $\underline{\mathbf{g}}_1 = \underline{\mathbf{g}}$, $\underline{\mathbf{f}}_2 = \underline{\mathbf{f}}^\star_{\mathbf{k}}$, and $\underline{\mathbf{g}}_2 = \underline{\mathbf{g}}^\star_{\mathbf{k}}$. With this setup, we can apply Lemma D.8 to bound the difference in risk as follows:

$$
\begin{aligned}
&\mathbb{E}_{\mathcal{D}_{2\otimes 2}}[\|\langle \underline{\mathbf{f}}, \underline{\mathbf{g}} \rangle_{\mathcal{M}} - \underline{h}^\star\|^2] - 2\mathbb{E}_{\mathcal{D}_{2\otimes 2}}[\|\langle \underline{\mathbf{f}}^\star_{\mathbf{k}}, \underline{\mathbf{g}}^\star_{\mathbf{k}} \rangle_{\mathcal{M}} - \underline{h}^\star\|^2] \\
&\le 6\left(\mathbb{E}_{\mathcal{D}_{2\otimes 2}}\|\langle \underline{\mathbf{f}} - \underline{\mathbf{f}}^\star_{\mathbf{k}}, \underline{\mathbf{g}}^\star_{\mathbf{k}} \rangle_{\mathcal{M}}\|^2 + \mathbb{E}_{\mathcal{D}_{2\otimes 2}}\|\langle \underline{\mathbf{f}}^\star_{\mathbf{k}}, \underline{\mathbf{g}} - \underline{\mathbf{g}}^\star_{\mathbf{k}} \rangle_{\mathcal{M}}\|^2 + \mathbb{E}_{\mathcal{D}_{2\otimes 2}}[\|\underline{\mathbf{f}} - \underline{\mathbf{f}}^\star_{\mathbf{k}}\|^2_{\mathcal{M}}\|\underline{\mathbf{g}} - \underline{\mathbf{g}}^\star_{\mathbf{k}}\|^2_{\mathcal{M}}] \right) \\
&\overset{(i)}{\le} 6\kappa_{\text{cov}}\left(\mathbb{E}_{\mathcal{D}_{2\otimes 1}}\|\langle \underline{\mathbf{f}} - \underline{\mathbf{f}}^\star_{\mathbf{k}}, \underline{\mathbf{g}}^\star_{\mathbf{k}} \rangle_{\mathcal{M}}\|^2 + \mathbb{E}_{\mathcal{D}_{1\otimes 2}}\|\langle \underline{\mathbf{f}}^\star_{\mathbf{k}}, \underline{\mathbf{g}} - \underline{\mathbf{g}}^\star_{\mathbf{k}} \rangle_{\mathcal{M}}\|^2 \right) \\
&\quad + 6\mathbb{E}_{\mathcal{D}_{2\otimes 2}}\|\underline{\mathbf{f}} - \underline{\mathbf{f}}^\star_{\mathbf{k}}\|^2_{\mathcal{M}}\|\underline{\mathbf{g}} - \underline{\mathbf{g}}^\star_{\mathbf{k}}\|^2_{\mathcal{M}} + 48B^2\eta_{\text{cov}} \\
&\overset{(ii)}{=} 6\kappa_{\text{cov}}\left(\mathbb{E}_{\mathcal{D}_{2\otimes 1}}\|\langle \underline{\mathbf{f}} - \underline{\mathbf{f}}^\star_{\mathbf{k}}, \underline{\mathbf{g}}^\star_{\mathbf{k}} \rangle_{\mathcal{M}}\|^2 + \mathbb{E}_{\mathcal{D}_{1\otimes 2}}\|\langle \underline{\mathbf{f}}^\star_{\mathbf{k}}, \underline{\mathbf{g}} - \underline{\mathbf{g}}^\star_{\mathbf{k}} \rangle_{\mathcal{M}}\|^2 \right) \\
&\quad + 6\mathbb{E}_{\mathcal{D}_{\mathcal{X},2}}\|\underline{\mathbf{f}} - \underline{\mathbf{f}}^\star_{\mathbf{k}}\|^2_{\mathcal{M}} \cdot \mathbb{E}_{\mathcal{D}_{\mathcal{Y},2}}\|\underline{\mathbf{g}} - \underline{\mathbf{g}}^\star_{\mathbf{k}}\|^2_{\mathcal{M}} + 48B^4\eta_{\text{cov}} \\
&\overset{(iii)}{\le} 6\kappa_{\text{cov}}\left(\mathbb{E}_{\mathcal{D}_{2\otimes 1}}\|\langle \underline{\mathbf{f}} - \underline{\mathbf{f}}^\star_{\mathbf{k}}, \underline{\mathbf{g}}^\star_{\mathbf{k}} \rangle_{\mathcal{M}}\|^2 + \mathbb{E}_{\mathcal{D}_{1\otimes 2}}\|\langle \underline{\mathbf{f}}^\star_{\mathbf{k}}, \underline{\mathbf{g}} - \underline{\mathbf{g}}^\star_{\mathbf{k}} \rangle_{\mathcal{M}}\|^2 \right) + 6(\Delta_2)^2 + 48B^4\eta_{\text{cov}},
\end{aligned}
\tag{26}
$$

where each step is justified as follows:

- In (i), we apply Lemma D.9 to bound $\mathbb{E}_{\mathcal{D}_{2\otimes 2}}\|\langle \underline{\mathbf{f}} - \underline{\mathbf{f}}^\star_{\mathbf{k}}, \underline{\mathbf{g}}^\star_{\mathbf{k}} \rangle_{\mathcal{M}}\|^2$ and $\mathbb{E}_{\mathcal{D}_{2\otimes 2}}\|\langle \underline{\mathbf{f}}^\star_{\mathbf{k}}, \underline{\mathbf{g}} - \underline{\mathbf{g}}^\star_{\mathbf{k}} \rangle_{\mathcal{M}}\|^2$. This is possible since $\|\underline{\mathbf{g}} - \underline{\mathbf{g}}^\star_{\mathbf{k}}\|_{\mathcal{M}} \le \|\underline{\mathbf{g}}\|_{\mathcal{M}} + \|\underline{\mathbf{g}}^\star_{\mathbf{k}}\|_{\mathcal{M}} \le 2B$ and similarly, $\|\underline{\mathbf{f}} - \underline{\mathbf{f}}^\star_{\mathbf{k}}\|_{\mathcal{M}} \le \|\underline{\mathbf{f}}\|_{\mathcal{M}} + \|\underline{\mathbf{f}}^\star_{\mathbf{k}}\|_{\mathcal{M}} \le 2B$.
- In (ii), we utilize the fact that $\mathcal{D}_{2\otimes 2} = \mathcal{D}_{\mathcal{X},2} \otimes \mathcal{D}_{\mathcal{Y},2}$ is a product measure, allowing us to separate the expectations over $\mathcal{D}_{\mathcal{X},2}$ and $\mathcal{D}_{\mathcal{Y},2}$.
- In (iii), we introduce the term $(\Delta_2)^2 = (\Delta_2(f, g, k))^2$ to further simplify the expression.

Thus, Equation (26) provides the bound on the risk under $\mathcal{D}_{2\otimes 2}$, accounting for covariate shifts in the multi-output setting.

Step 2: Expanding Terms under $\mathcal{D}_{1\otimes 2}$ and $\mathcal{D}_{2\otimes 1}$. In this step, we expand the first two terms from Equation (26) to further analyze their contributions.

Starting with $\langle \underline{\mathbf{f}} - \underline{\mathbf{f}}^\star_{\mathbf{k}}, \underline{\mathbf{g}}^\star_{\mathbf{k}} \rangle_{\mathcal{M}}$, observe that:

$$
\langle \underline{\mathbf{f}} - \underline{\mathbf{f}}^\star_{\mathbf{k}}, \underline{\mathbf{g}}^\star_{\mathbf{k}} \rangle_{\mathcal{M}} = \langle \underline{\mathbf{f}}, \underline{\mathbf{g}}^\star_{\mathbf{k}} \rangle_{\mathcal{M}} - \underline{h}^\star_{\mathbf{k}} = \langle \underline{\mathbf{f}}, \underline{\mathbf{g}} \rangle_{\mathcal{M}} - \underline{h}^\star_{\mathbf{k}} + \langle \underline{\mathbf{f}}, \underline{\mathbf{g}}^\star_{\mathbf{k}} - \underline{\mathbf{g}} \rangle_{\mathcal{M}} = \langle \underline{\mathbf{f}}, \underline{\mathbf{g}} \rangle_{\mathcal{M}} - \underline{h}^\star_{\mathbf{k}} + \langle \underline{\mathbf{f}}^\star_{\mathbf{k}}, \underline{\mathbf{g}}^\star_{\mathbf{k}} - \underline{\mathbf{g}} \rangle_{\mathcal{M}} + \langle \underline{\mathbf{f}} - \underline{\mathbf{f}}^\star_{\mathbf{k}}, \underline{\mathbf{g}}^\star_{\mathbf{k}} - \underline{\mathbf{g}} \rangle_{\mathcal{M}}.
$$

With this, we can bound:

$$
\begin{aligned}
&\mathbb{E}_{\mathcal{D}_{2\otimes 1}}\|\langle \underline{\mathbf{f}} - \underline{\mathbf{f}}^\star_{\mathbf{k}}, \underline{\mathbf{g}}^\star_{\mathbf{k}} \rangle_{\mathcal{M}}\|^2 \\
&\le 3\mathbb{E}_{\mathcal{D}_{2\otimes 1}}[\|\langle \underline{\mathbf{f}}, \underline{\mathbf{g}} \rangle_{\mathcal{M}} - \underline{h}^\star_{\mathbf{k}}\|^2] + 3\mathbb{E}_{\mathcal{D}_{2\otimes 1}}\|\langle \underline{\mathbf{f}}^\star_{\mathbf{k}}, \underline{\mathbf{g}}^\star_{\mathbf{k}} - \underline{\mathbf{g}} \rangle_{\mathcal{M}}\|^2 + 3\mathbb{E}_{\mathcal{D}_{2\otimes 1}}\|\langle \underline{\mathbf{f}} - \underline{\mathbf{f}}^\star_{\mathbf{k}}, \underline{\mathbf{g}}^\star_{\mathbf{k}} - \underline{\mathbf{g}} \rangle_{\mathcal{M}}\|^2 \\
&\overset{(i)}{\le} 3\mathbb{E}_{\mathcal{D}_{2\otimes 1}}[\|\langle \underline{\mathbf{f}}, \underline{\mathbf{g}} \rangle_{\mathcal{M}} - \underline{h}^\star_{\mathbf{k}}\|^2] + 3\kappa_{\text{cov}}\mathbb{E}_{\mathcal{D}_{1\otimes 1}}\|\langle \underline{\mathbf{f}}^\star_{\mathbf{k}}, \underline{\mathbf{g}}^\star_{\mathbf{k}} - \underline{\mathbf{g}} \rangle_{\mathcal{M}}\|^2 + 3\mathbb{E}_{\mathcal{D}_{2\otimes 1}}\|\underline{\mathbf{f}}^\star_{\mathbf{k}} - \underline{\mathbf{f}}\|^2_{\mathcal{M}}\|\underline{\mathbf{g}}^\star_{\mathbf{k}} - \underline{\mathbf{g}}\|^2_{\mathcal{M}} + 12B^4\eta_{\text{cov}},
\end{aligned}
$$

where in (i) we use Lemma D.9 to bound terms involving covariate shifts.

Next, we expand and bound the product of norms term as follows:

$$
\begin{aligned}
\mathbb{E}_{\mathcal{D}_{2\otimes 1}}\|\underline{\mathbf{f}}^\star_{\mathbf{k}} - \underline{\mathbf{f}}\|^2_{\mathcal{M}}\|\underline{\mathbf{g}}^\star_{\mathbf{k}} - \underline{\mathbf{g}}\|^2_{\mathcal{M}} &= \mathbb{E}_{\mathcal{D}_{\mathcal{X},2}}\|\underline{\mathbf{f}}^\star_{\mathbf{k}} - \underline{\mathbf{f}}\|^2_{\mathcal{M}} \cdot \mathbb{E}_{\mathcal{D}_{\mathcal{Y},1}}\|\underline{\mathbf{g}}^\star_{\mathbf{k}} - \underline{\mathbf{g}}\|^2_{\mathcal{M}} \\
&\le \frac{1}{2\kappa_{\text{cov}}}\left(\mathbb{E}_{\mathcal{D}_{\mathcal{X},2}}\|\underline{\mathbf{f}}^\star_{\mathbf{k}} - \underline{\mathbf{f}}\|^2_{\mathcal{M}}\right)^2 + \frac{\kappa_{\text{cov}}}{2}\left(\mathbb{E}_{\mathcal{D}_{\mathcal{Y},1}}\|\underline{\mathbf{g}}^\star_{\mathbf{k}} - \underline{\mathbf{g}}\|^2_{\mathcal{M}}\right)^2 \\
&\le \frac{1}{2\kappa_{\text{cov}}}(\Delta_2)^2 + \frac{\kappa_{\text{cov}}}{2}(\Delta_1)^2,
\end{aligned}
$$

where we have used the definitions of Δ_2 and Δ_1 from Definition 19.

Summing up these results, we obtain:

$$\begin{aligned}
&\mathbb{E}_{\mathcal{D}_{2\otimes 1}}\|\langle \underline{\mathbf{f}} - \underline{\mathbf{f}}^\star_{\mathbf{k}}, \underline{\mathbf{g}}^\star_{\mathbf{k}}\rangle_{\mathcal{M}}\|^2 \\
&\le 3\mathbb{E}_{\mathcal{D}_{2\otimes 1}}[\|\langle \underline{\mathbf{f}}, \underline{\mathbf{g}}\rangle_{\mathcal{M}} - \underline{h}^\star_{\mathbf{k}}\|^2] + 3\kappa_{\text{cov}}\left(\mathbb{E}_{\mathcal{D}_{1\otimes 1}}\|\langle \underline{\mathbf{f}}^\star_{\mathbf{k}}, \underline{\mathbf{g}}^\star_{\mathbf{k}} - \underline{\mathbf{g}}\rangle_{\mathcal{M}}\|^2 + \frac{1}{2}(\Delta_1)^2\right) + \frac{3}{2\kappa_{\text{cov}}}(\Delta_2)^2 + 12B^4\eta_{\text{cov}}.
\end{aligned}$$

Using a similar argument, we can bound the other term as:

$$\begin{aligned}
&\mathbb{E}_{\mathcal{D}_{1\otimes 2}}\|\langle \underline{\mathbf{f}}^\star_{\mathbf{k}}, \underline{\mathbf{g}} - \underline{\mathbf{g}}^\star_{\mathbf{k}}\rangle_{\mathcal{M}}\|^2 \\
&\le 3\mathbb{E}_{\mathcal{D}_{1\otimes 2}}[\|\langle \underline{\mathbf{f}}, \underline{\mathbf{g}}\rangle_{\mathcal{M}} - \underline{h}^\star_{\mathbf{k}}\|^2] + 3\kappa_{\text{cov}}\left(\mathbb{E}_{\mathcal{D}_{1\otimes 1}}\|\langle \underline{\mathbf{f}} - \underline{\mathbf{f}}^\star_{\mathbf{k}}, \underline{\mathbf{g}}^\star_{\mathbf{k}}\rangle_{\mathcal{M}}\|^2 + \frac{1}{2}(\Delta_1)^2\right) + \frac{3}{2\kappa_{\text{cov}}}(\Delta_2)^2 + 12B^4\eta_{\text{cov}}.
\end{aligned}$$

Now, define:

$$\Delta_{\text{off}} = \mathbb{E}_{\mathcal{D}_{1\otimes 2}}[\|\langle \underline{\mathbf{f}}, \underline{\mathbf{g}}\rangle_{\mathcal{M}} - \underline{h}^\star_{\mathbf{k}}\|^2] + \mathbb{E}_{\mathcal{D}_{2\otimes 1}}[\|\langle \underline{\mathbf{f}}, \underline{\mathbf{g}}\rangle_{\mathcal{M}} - \underline{h}^\star_{\mathbf{k}}\|^2].$$

Thus, we can combine these results as:

$$\begin{aligned}
&\mathbb{E}_{\mathcal{D}_{2\otimes 1}}\|\langle \underline{\mathbf{f}} - \underline{\mathbf{f}}^\star_{\mathbf{k}}, \underline{\mathbf{g}}^\star_{\mathbf{k}}\rangle_{\mathcal{M}}\|^2 + \mathbb{E}_{\mathcal{D}_{1\otimes 2}}\|\langle \underline{\mathbf{f}}^\star_{\mathbf{k}}, \underline{\mathbf{g}} - \underline{\mathbf{g}}^\star_{\mathbf{k}}\rangle_{\mathcal{M}}\|^2 \\
&\le 3\mathbb{E}_{\mathcal{D}_{1\otimes 2}}[\|\langle \underline{\mathbf{f}}, \underline{\mathbf{g}}\rangle_{\mathcal{M}} - \underline{h}^\star_{\mathbf{k}}\|^2] + 3\mathbb{E}_{\mathcal{D}_{2\otimes 1}}[\|\langle \underline{\mathbf{f}}, \underline{\mathbf{g}}\rangle_{\mathcal{M}} - \underline{h}^\star_{\mathbf{k}}\|^2] \\
&\quad + 3\kappa_{\text{cov}}\left(2\mathbb{E}_{\mathcal{D}_{1\otimes 1}}\|\langle \underline{\mathbf{f}} - \underline{\mathbf{f}}^\star_{\mathbf{k}}, \underline{\mathbf{g}}^\star_{\mathbf{k}}\rangle_{\mathcal{M}}\|^2 + (\Delta_1)^2\right) + \frac{3}{\kappa_{\text{cov}}}(\Delta_2)^2 + 24B^4\eta_{\text{cov}} \\
&= 3\Delta_{\text{off}} + 3\kappa_{\text{cov}}(2\Delta_0 + (\Delta_1)^2) + \frac{3}{\kappa_{\text{cov}}}(\Delta_2)^2 + 24B^4\eta_{\text{cov}}. \qquad (27)
\end{aligned}$$

Step 3: Consolidating Results. Combining Equation (27) and Equation (26), we obtain:

$$\begin{aligned}
&\mathbb{E}_{\mathcal{D}_{2\otimes 2}}[\|\langle \underline{\mathbf{f}}, \underline{\mathbf{g}}\rangle_{\mathcal{M}} - \underline{h}^\star\|^2] - 2\mathbb{E}_{\mathcal{D}_{2\otimes 2}}[\|\langle \underline{\mathbf{f}}^\star_{\mathbf{k}}, \underline{\mathbf{g}}^\star_{\mathbf{k}}\rangle_{\mathcal{M}} - \underline{h}^\star\|^2] \\
&\le 6\kappa_{\text{cov}}\left(\mathbb{E}_{\mathcal{D}_{2\otimes 1}}\|\langle \underline{\mathbf{f}} - \underline{\mathbf{f}}^\star_{\mathbf{k}}, \underline{\mathbf{g}}^\star_{\mathbf{k}}\rangle_{\mathcal{M}}\|^2 + \mathbb{E}_{\mathcal{D}_{1\otimes 2}}\|\langle \underline{\mathbf{f}}^\star_{\mathbf{k}}, \underline{\mathbf{g}} - \underline{\mathbf{g}}^\star_{\mathbf{k}}\rangle_{\mathcal{M}}\|^2\right) + 6(\Delta_2)^2 + 48B^4\eta_{\text{cov}} \\
&\le 18\kappa_{\text{cov}}\Delta_{\text{off}} + 18\kappa_{\text{cov}}^2(2\Delta_0 + (\Delta_1)^2) + 24(\Delta_2)^2 + (144\kappa_{\text{cov}} + 48)B^4\eta_{\text{cov}}.
\end{aligned}$$

Rearranging terms, we get:

$$\begin{aligned}
\mathbb{E}_{\mathcal{D}_{2\otimes 2}}[\|\langle \underline{\mathbf{f}}, \underline{\mathbf{g}}\rangle_{\mathcal{M}} - \underline{h}^\star\|^2] &\le 18\kappa_{\text{cov}}^2(2\Delta_0 + (\Delta_1)^2) + 24(\Delta_2)^2 + (144\kappa_{\text{cov}} + 48)B^4\eta_{\text{cov}} \\
&\quad + 18\kappa_{\text{cov}}\Delta_{\text{off}} + 2\mathbb{E}_{\mathcal{D}_{2\otimes 2}}[\|\langle \underline{\mathbf{f}}^\star_{\mathbf{k}}, \underline{\mathbf{g}}^\star_{\mathbf{k}}\rangle_{\mathcal{M}} - \underline{h}^\star\|^2]. \qquad (28)
\end{aligned}$$

Step 4: Final Bound. To finalize, we need to bound Equation (28). By Lemma D.10, we have:

$$\mathbb{E}_{\mathcal{D}_{2\otimes 2}}[\|\langle \underline{\mathbf{f}}^\star_{\mathbf{k}}, \underline{\mathbf{g}}^\star_{\mathbf{k}}\rangle_{\mathcal{M}} - \underline{h}^\star\|^2] = \mathcal{R}(\underline{\mathbf{f}}^\star_{\mathbf{k}}, \underline{\mathbf{g}}^\star_{\mathbf{k}}; \mathcal{D}_{2\otimes 2}) \le \kappa_{\text{cov}}^2\mathcal{R}(\underline{\mathbf{f}}^\star_{\mathbf{k}}, \underline{\mathbf{g}}^\star_{\mathbf{k}}; \mathcal{D}_{1\otimes 1}) + 2\kappa_{\text{cov}}\eta_{\text{cov}}B^4.$$

Similarly, for Δ_{off}, by Lemma D.10:

$$\begin{aligned}
\Delta_{\text{off}} &:= \mathbb{E}_{\mathcal{D}_{1\otimes 2}}[\|\langle \underline{\mathbf{f}}, \underline{\mathbf{g}}\rangle_{\mathcal{M}} - \underline{h}^\star_{\mathbf{k}}\|^2] + \mathbb{E}_{\mathcal{D}_{2\otimes 1}}[\|\langle \underline{\mathbf{f}}, \underline{\mathbf{g}}\rangle_{\mathcal{M}} - \underline{h}^\star_{\mathbf{k}}\|^2] \\
&\le 2\mathbb{E}_{\mathcal{D}_{1\otimes 2}}[\|\langle \underline{\mathbf{f}}, \underline{\mathbf{g}}\rangle_{\mathcal{M}} - \underline{h}^\star\|^2] + 2\mathbb{E}_{\mathcal{D}_{2\otimes 1}}[\|\langle \underline{\mathbf{f}}, \underline{\mathbf{g}}\rangle_{\mathcal{M}} - \underline{h}^\star\|^2] \\
&\quad + 2\underbrace{\mathbb{E}_{\mathcal{D}_{1\otimes 2}}[\|\underline{h}^\star_{\mathbf{k}} - \underline{h}^\star\|^2]}_{=\mathcal{R}(\underline{\mathbf{f}}^\star_{\mathbf{k}},\underline{\mathbf{g}}^\star_{\mathbf{k}};\mathcal{D}_{1\otimes 2})} + 2\underbrace{\mathbb{E}_{\mathcal{D}_{2\otimes 1}}[\|\underline{h}^\star_{\mathbf{k}} - \underline{h}^\star\|^2]}_{=\mathcal{R}(\underline{\mathbf{f}}^\star_{\mathbf{k}},\underline{\mathbf{g}}^\star_{\mathbf{k}};\mathcal{D}_{2\otimes 1})} \\
&\le 2\mathbb{E}_{\mathcal{D}_{1\otimes 2}}[\|\langle \underline{\mathbf{f}}, \underline{\mathbf{g}}\rangle_{\mathcal{M}} - \underline{h}^\star\|^2] + 2\mathbb{E}_{\mathcal{D}_{2\otimes 1}}[\|\langle \underline{\mathbf{f}}, \underline{\mathbf{g}}\rangle_{\mathcal{M}} - \underline{h}^\star\|^2] \\
&\quad + 4\kappa_{\text{cov}}\mathcal{R}(\underline{\mathbf{f}}^\star_{\mathbf{k}}, \underline{\mathbf{g}}^\star_{\mathbf{k}}; \mathcal{D}_{1\otimes 1}) + 4\eta_{\text{cov}}B^4.
\end{aligned}$$

Thus,

$$\begin{aligned}
&18\kappa_{\text{cov}}\Delta_{\text{off}} + 2\mathbb{E}_{\mathcal{D}_{2\otimes 2}}[\|\langle \underline{\mathbf{f}}^\star_{\mathbf{k}}, \underline{\mathbf{g}}^\star_{\mathbf{k}}\rangle_{\mathcal{M}} - \underline{h}^\star\|^2] \\
&\le (4\cdot 18+2)\kappa_{\text{cov}}^2\mathcal{R}(\underline{\mathbf{f}}^\star_{\mathbf{k}}, \underline{\mathbf{g}}^\star_{\mathbf{k}}; \mathcal{D}_{1\otimes 1}) + (4\cdot 18+4)\kappa_{\text{cov}}\eta_{\text{cov}}B^4 \\
&\quad + (2\cdot 18)\kappa_{\text{cov}}\left(\mathbb{E}_{\mathcal{D}_{1\otimes 2}}[\|\langle \underline{\mathbf{f}}, \underline{\mathbf{g}}\rangle_{\mathcal{M}} - \underline{h}^\star\|^2] + \mathbb{E}_{\mathcal{D}_{2\otimes 1}}[\|\langle \underline{\mathbf{f}}, \underline{\mathbf{g}}\rangle_{\mathcal{M}} - \underline{h}^\star\|^2]\right) \\
&= 74\kappa_{\text{cov}}^2\mathcal{R}(\underline{\mathbf{f}}^\star_{\mathbf{k}}, \underline{\mathbf{g}}^\star_{\mathbf{k}}; \mathcal{D}_{1\otimes 1}) + 76\kappa_{\text{cov}}\eta_{\text{cov}}B^4 + 36\kappa_{\text{cov}}\left(\mathcal{R}(\underline{\mathbf{f}}, \underline{\mathbf{g}}; \mathcal{D}_{1\otimes 2}) + \mathcal{R}(\underline{\mathbf{f}}, \underline{\mathbf{g}}; \mathcal{D}_{2\otimes 1})\right).
\end{aligned}$$

Using Lemma D.7:

$$\mathcal{R}(\underline{\mathbf{f}}, \underline{\mathbf{g}}; \mathcal{D}_{1\otimes 2}) + \mathcal{R}(\underline{\mathbf{f}}, \underline{\mathbf{g}}; \mathcal{D}_{2\otimes 1}) \le 2\kappa_{\text{trn}}\mathcal{R}(\underline{\mathbf{f}}, \underline{\mathbf{g}}; \mathcal{D}_{\mathrm{train}}) + 8B^4\eta_{\text{trn}},$$

we get:

$$\begin{aligned}
&18\kappa_{\text{cov}}\Delta_{\text{off}} + 2\mathbb{E}_{\mathcal{D}_{2\otimes 2}}[\|\langle \underline{\mathbf{f}}^\star_{\mathbf{k}}, \underline{\mathbf{g}}^\star_{\mathbf{k}}\rangle_{\mathcal{M}} - \underline{h}^\star\|^2] \\
&\le 74\kappa_{\text{cov}}^2\mathcal{R}(\underline{\mathbf{f}}^\star_{\mathbf{k}}, \underline{\mathbf{g}}^\star_{\mathbf{k}}; \mathcal{D}_{1\otimes 1}) + 72\kappa_{\text{cov}}\kappa_{\text{trn}}\mathcal{R}(\underline{\mathbf{f}}, \underline{\mathbf{g}}; \mathcal{D}_{\mathrm{train}}) + 288\eta_{\text{trn}}\kappa_{\text{cov}}B^4 + 76\kappa_{\text{cov}}\eta_{\text{cov}}B^4.
\end{aligned}$$

Summing everything up, we obtain:

$$\begin{aligned}
\mathbb{E}_{\mathcal{D}_{2\otimes 2}}[\|\langle \underline{\mathbf{f}}, \underline{\mathbf{g}}\rangle_{\mathcal{M}} - \underline{h}^\star\|^2] &\le 18\kappa_{\text{cov}}^2(2\Delta_0 + (\Delta_1)^2) + 24(\Delta_2)^2 + 72\kappa_{\text{cov}}\kappa_{\text{trn}}\underbrace{\mathcal{R}(\underline{\mathbf{f}}, \underline{\mathbf{g}}; \mathcal{D}_{\mathrm{train}})}_{=\Delta_{\mathrm{train}}} \\
&\quad + 74\kappa_{\text{cov}}^2\underbrace{\mathcal{R}(\underline{\mathbf{f}}^\star_{\mathbf{k}}, \underline{\mathbf{g}}^\star_{\mathbf{k}}; \mathcal{D}_{1\otimes 1})}_{=\Delta_{\mathrm{apx}}} + 268\kappa_{\text{cov}}B^4\eta_{\text{cov}} + 288\eta_{\text{trn}}\kappa_{\text{cov}}B^4.
\end{aligned}$$

Dropping constants, we conclude:

$$\begin{aligned}
\mathcal{R}(\underline{\mathbf{f}}, \underline{\mathbf{g}}; \mathcal{D}_{2\otimes 2}) &= \mathbb{E}_{\mathcal{D}_{2\otimes 2}}[\|\langle \underline{\mathbf{f}}, \underline{\mathbf{g}}\rangle_{\mathcal{M}} - \underline{h}^\star\|^2] \\
&\lesssim \kappa_{\text{cov}}^2(\Delta_0 + (\Delta_1)^2 + \Delta_{\mathrm{apx}}) + (\Delta_2)^2 + \kappa_{\text{cov}}\kappa_{\text{trn}}\Delta_{\mathrm{train}} + B^4\kappa_{\text{cov}}(\eta_{\text{cov}} + \eta_{\text{trn}}).
\end{aligned}$$

□

D.1.5 Key Change-of-Measure Lemmas

We begin by establishing some important change-of-measure results.

Lemma D.9 (Change of Measure: Factor Estimation Error)**.** *Under Assumption 7 and Assumption 8, the following holds for any* $i, j \in \{1, 2\}$ *and any* $(\tilde{\underline{\mathbf{f}}}, \tilde{\underline{\mathbf{g}}})$*,*

- $\mathbb{E}_{\mathcal{D}_{i\otimes 2}}[\|\langle \tilde{\underline{\mathbf{f}}}, \underline{\mathbf{g}}^\star_{\mathbf{k}}\rangle_{\mathcal{M}}\|^2] \le \kappa_{\text{cov}}\mathbb{E}_{\mathcal{D}_{i\otimes 1}}[\|\langle \tilde{\underline{\mathbf{f}}}, \underline{\mathbf{g}}^\star_{\mathbf{k}}\rangle_{\mathcal{M}}\|^2] + B^2\eta_{\text{cov}}$
- $\mathbb{E}_{\mathcal{D}_{2\otimes j}}[\|\langle \underline{\mathbf{f}}^\star_{\mathbf{k}}, \tilde{\underline{\mathbf{g}}}\rangle_{\mathcal{M}}\|^2] \le \kappa_{\text{cov}}\mathbb{E}_{\mathcal{D}_{1\otimes j}}[\|\langle \underline{\mathbf{f}}^\star_{\mathbf{k}}, \tilde{\underline{\mathbf{g}}}\rangle_{\mathcal{M}}\|^2] + B^2\eta_{\text{cov}}.$

The same holds if $\underline{\mathbf{g}}^\star_{\mathbf{k}}$ *(resp.* $\underline{\mathbf{f}}^\star_{\mathbf{k}}$*) are replaced by* $\underline{\mathbf{g}}^\star_{>\mathbf{k}} := \underline{\mathbf{g}}^\star - \underline{\mathbf{g}}^\star_{\mathbf{k}}$ *(resp.* $\underline{\mathbf{f}}^\star_{>\mathbf{k}} := \underline{\mathbf{f}}^\star - \underline{\mathbf{f}}^\star_{\mathbf{k}}$*) or* $\underline{\mathbf{g}}^\star$ *(resp.* $\underline{\mathbf{f}}^\star$*).*

Proof of Lemma D.9. We start by proving the first part under Assumption 7; the extension to the second part follows a similar approach. Specifically, we have:

$$\begin{aligned}
\mathbb{E}_{\mathcal{D}_{i\otimes 2}}[\|\langle \tilde{\underline{\mathbf{f}}}, \underline{\mathbf{g}}^\star_{\mathbf{k}}\rangle_{\mathcal{M}}\|^2] &= \mathbb{E}_{\mathcal{D}_{\mathcal{X},i}}[\mathbb{E}_{\mathcal{D}_{\mathcal{Y},2}}[\|\langle \tilde{\underline{\mathbf{f}}}, \underline{\mathbf{g}}^\star_{\mathbf{k}}\rangle_{\mathcal{M}}\|^2]] && \text{(by Fubini's theorem)} \\
&= \mathbb{E}_{\mathcal{D}_{\mathcal{X},i}}[\mathbb{E}_{\mathcal{D}_{\mathcal{Y},2}}[\|\langle \tilde{\underline{\mathbf{f}}}, \underline{\mathbf{P}}^\star_{\mathbf{k}}\underline{\mathbf{g}}^\star_{\mathbf{k}}\rangle_{\mathcal{M}}\|^2]] \\
&\le \mathbb{E}_{\mathcal{D}_{\mathcal{X},i}}[\kappa_{\text{cov}}\mathbb{E}_{\mathcal{D}_{\mathcal{Y},1}}[\|\langle \underline{\mathbf{P}}^\star_{\mathbf{k}}\tilde{\underline{\mathbf{f}}}, \underline{\mathbf{g}}^\star\rangle_{\mathcal{M}}\|^2] + \eta_{\text{cov}}\|\underline{\mathbf{P}}^\star_{\mathbf{k}}\tilde{\underline{\mathbf{f}}}\|^2_{\mathcal{M}}] && \text{(by Assumption 7)} \\
&\le \mathbb{E}_{\mathcal{D}_{\mathcal{X},i}}[\kappa_{\text{cov}}\mathbb{E}_{\mathcal{D}_{\mathcal{Y},1}}[\|\langle \underline{\mathbf{P}}^\star_{\mathbf{k}}\tilde{\underline{\mathbf{f}}}, \underline{\mathbf{g}}^\star\rangle_{\mathcal{M}}\|^2] + \eta_{\text{cov}}\|\tilde{\underline{\mathbf{f}}}\|^2_{\mathcal{M}}] && \text{(since } \underline{\mathbf{P}}^\star_{\mathbf{k}} \text{ is a projection)} \\
&\le \mathbb{E}_{\mathcal{D}_{\mathcal{X},i}}[\kappa_{\text{cov}}\mathbb{E}_{\mathcal{D}_{\mathcal{Y},1}}[\|\langle \underline{\mathbf{P}}^\star_{\mathbf{k}}\tilde{\underline{\mathbf{f}}}, \underline{\mathbf{g}}^\star\rangle_{\mathcal{M}}\|^2] + B^2\eta_{\text{cov}}] && \text{(by Assumption 8)} \\
&= \kappa_{\text{cov}}\mathbb{E}_{\mathcal{D}_{\mathcal{X},i}\otimes\mathcal{D}_{\mathcal{Y},1}}[\|\langle \underline{\mathbf{P}}^\star_{\mathbf{k}}\tilde{\underline{\mathbf{f}}}, \underline{\mathbf{g}}^\star\rangle_{\mathcal{M}}\|^2] + B^2\eta_{\text{cov}} && \text{(by Fubini's theorem)} \\
&= \kappa_{\text{cov}}\mathbb{E}_{\mathcal{D}_{\mathcal{X},i}\otimes\mathcal{D}_{\mathcal{Y},1}}[\|\langle \tilde{\underline{\mathbf{f}}}, \underline{\mathbf{g}}^\star_{\mathbf{k}}\rangle_{\mathcal{M}}\|^2] + B^2\eta_{\text{cov}}.
\end{aligned}$$

The second part of the lemma follows similarly. To derive analogous bounds for $\underline{\mathbf{g}}^\star - \underline{\mathbf{g}}^\star_{\mathbf{k}}$, we note that $\underline{\mathbf{g}}^\star - \underline{\mathbf{g}}^\star_{\mathbf{k}} = (\underline{\mathcal{I}} - \underline{\mathbf{P}}^\star_k)\underline{\mathbf{g}}^\star$, where $\underline{\mathcal{I}} - \underline{\mathbf{P}}^\star_k$ is also a projection operator. A similar argument applies for bounding $\underline{\mathbf{f}}^\star - \underline{\mathbf{f}}^\star_{\mathbf{k}}$.

Finally, the bounds for $\underline{\mathbf{f}}^\star$ and $\underline{\mathbf{g}}^\star$ are simpler to establish, as they do not require commuting the projection operator. □

Lemma D.10 (Change of Measure: Approximation Error)**.** *The following bounds hold:*

- *The risks on the product distributions $\mathcal{D}_{1\otimes 2}$ and $\mathcal{D}_{2\otimes 1}$ are bounded by*

$$\mathcal{R}(\underline{\mathbf{f}}^\star_{\mathbf{k}}, \underline{\mathbf{g}}^\star_{\mathbf{k}}; \mathcal{D}_{1\otimes 2}) \vee \mathcal{R}(\underline{\mathbf{f}}^\star_{\mathbf{k}}, \underline{\mathbf{g}}^\star_{\mathbf{k}}; \mathcal{D}_{2\otimes 1}) \le \kappa_{\mathrm{cov}}\Delta_{\mathrm{apx}}(\mathbf{k}) + \eta_{\mathrm{cov}}B^2.$$

- *The risk on the product distribution $\mathcal{D}_{2\otimes 2}$ is bounded by*

$$\mathcal{R}(\underline{\mathbf{f}}^\star_{\mathbf{k}}, \underline{\mathbf{g}}^\star_{\mathbf{k}}; \mathcal{D}_{2\otimes 2}) \le \kappa_{\mathrm{cov}}^2\Delta_{\mathrm{apx}}(\mathbf{k}) + 2\kappa_{\mathrm{cov}}\eta_{\mathrm{cov}}B^2.$$

Proof of Lemma D.10. Introduce the shorthand $\underline{\mathbf{f}}^\star_{>\mathbf{k}} := \underline{\mathbf{f}}^\star - \underline{\mathbf{f}}^\star_{\mathbf{k}}$ and $\underline{\mathbf{g}}^\star_{>\mathbf{k}} := \underline{\mathbf{g}}^\star - \underline{\mathbf{g}}^\star_{\mathbf{k}}$. Note that

$$\underline{\mathbf{f}}^\star_{>\mathbf{k}} = (\underline{\mathcal{I}} - \underline{\mathbf{P}}^\star_{\mathbf{k}})\underline{\mathbf{f}}^\star, \quad \underline{\mathbf{g}}^\star_{>\mathbf{k}} = (\underline{\mathcal{I}} - \underline{\mathbf{P}}^\star_{\mathbf{k}})\underline{\mathbf{g}}^\star,$$

which implies that both $\underline{\mathbf{f}}^\star_{>\mathbf{k}}$ and $\underline{\mathbf{g}}^\star_{>\mathbf{k}}$ are B-bounded by Assumption 8.

Since $\underline{\mathbf{P}}^\star_{\mathbf{k}}$ is an orthogonal projection, $\underline{\mathcal{I}} - \underline{\mathbf{P}}^\star_{\mathbf{k}}$ is also an orthogonal projection, meaning it is self-adjoint and idempotent. Therefore, we can write:

$$\begin{aligned}
\underline{h}^\star - \langle \underline{\mathbf{f}}^\star_{\mathbf{k}}, \underline{\mathbf{g}}^\star_{\mathbf{k}} \rangle_{\mathcal{M}} &= \langle \underline{\mathbf{f}}^\star, \underline{\mathbf{g}}^\star \rangle_{\mathcal{M}} - \langle \underline{\mathbf{P}}^\star_{\mathbf{k}}\underline{\mathbf{f}}^\star, \underline{\mathbf{g}}^\star \rangle_{\mathcal{M}} \\
&= \langle (\underline{\mathcal{I}} - \underline{\mathbf{P}}^\star_{\mathbf{k}})\underline{\mathbf{f}}^\star, \underline{\mathbf{g}}^\star \rangle_{\mathcal{M}} \\
&= \langle (\underline{\mathcal{I}} - \underline{\mathbf{P}}^\star_{\mathbf{k}})^\top (\underline{\mathcal{I}} - \underline{\mathbf{P}}^\star_{\mathbf{k}})\underline{\mathbf{f}}^\star, \underline{\mathbf{g}}^\star \rangle_{\mathcal{M}} \\
&= \langle (\underline{\mathcal{I}} - \underline{\mathbf{P}}^\star_{\mathbf{k}})\underline{\mathbf{f}}^\star, (\underline{\mathcal{I}} - \underline{\mathbf{P}}^\star_{\mathbf{k}})\underline{\mathbf{g}}^\star \rangle_{\mathcal{M}} \\
&= \langle \underline{\mathbf{f}}^\star_{>\mathbf{k}}, \underline{\mathbf{g}}^\star_{>\mathbf{k}} \rangle_{\mathcal{M}}.
\end{aligned} \tag{29}$$

Thus, by applying Lemma D.9 and noting that $\underline{\mathbf{f}}^\star_{>\mathbf{k}}$ is B-bounded, we have:

$$\begin{aligned}
\mathcal{R}(\underline{\mathbf{f}}^\star_{\mathbf{k}}, \underline{\mathbf{g}}^\star_{\mathbf{k}}; \mathcal{D}_{1\otimes 2}) &= \mathbb{E}_{\mathcal{D}_{1\otimes 2}}[\|\underline{h}^\star - \langle \underline{\mathbf{f}}^\star_{\mathbf{k}}, \underline{\mathbf{g}}^\star_{\mathbf{k}} \rangle_{\mathcal{M}}\|^2] \\
&= \mathbb{E}_{\mathcal{D}_{1\otimes 2}}[\|\langle \underline{\mathbf{f}}^\star_{>\mathbf{k}}, \underline{\mathbf{g}}^\star_{>\mathbf{k}} \rangle_{\mathcal{M}}\|^2] \\
&\le \kappa_{\mathrm{cov}}\mathbb{E}_{\mathcal{D}_{1\otimes 1}}[\|\langle \underline{\mathbf{f}}^\star_{>\mathbf{k}}, \underline{\mathbf{g}}^\star_{>\mathbf{k}} \rangle_{\mathcal{M}}\|^2] + \eta_{\mathrm{cov}}B^2.
\end{aligned}$$

Similarly, we can bound $\mathcal{R}(\underline{\mathbf{f}}^\star_{\mathbf{k}}, \underline{\mathbf{g}}^\star_{\mathbf{k}}; \mathcal{D}_{2\otimes 1})$ as:

$$\mathcal{R}(\underline{\mathbf{f}}^\star_{\mathbf{k}}, \underline{\mathbf{g}}^\star_{\mathbf{k}}; \mathcal{D}_{2\otimes 1}) \le \kappa_{\mathrm{cov}}\mathbb{E}_{\mathcal{D}_{1\otimes 1}}[\|\langle \underline{\mathbf{f}}^\star_{>\mathbf{k}}, \underline{\mathbf{g}}^\star_{>\mathbf{k}} \rangle_{\mathcal{M}}\|^2] + \eta_{\mathrm{cov}}B^2.$$

Finally, applying Lemma D.9 twice, we find:

$$\begin{aligned}
\mathcal{R}(\underline{\mathbf{f}}^\star_{\mathbf{k}}, \underline{\mathbf{g}}^\star_{\mathbf{k}}; \mathcal{D}_{2\otimes 2}) &= \mathbb{E}_{\mathcal{D}_{2\otimes 2}}[\|\underline{h}^\star - \langle \underline{\mathbf{f}}^\star_{\mathbf{k}}, \underline{\mathbf{g}}^\star_{\mathbf{k}} \rangle_{\mathcal{M}}\|^2] \\
&= \mathbb{E}_{\mathcal{D}_{2\otimes 2}}[\|\langle \underline{\mathbf{f}}^\star_{>\mathbf{k}}, \underline{\mathbf{g}}^\star_{>\mathbf{k}} \rangle_{\mathcal{M}}\|^2] \\
&\le \kappa_{\mathrm{cov}}\mathbb{E}_{\mathcal{D}_{1\otimes 2}}[\|\langle \underline{\mathbf{f}}^\star_{>\mathbf{k}}, \underline{\mathbf{g}}^\star_{>\mathbf{k}} \rangle_{\mathcal{M}}\|^2] + \eta_{\mathrm{cov}}B^2 \\
&\le \kappa_{\mathrm{cov}}^2\mathbb{E}_{\mathcal{D}_{1\otimes 1}}[\|\langle \underline{\mathbf{f}}^\star_{>\mathbf{k}}, \underline{\mathbf{g}}^\star_{>\mathbf{k}} \rangle_{\mathcal{M}}\|^2] + (\kappa_{\mathrm{cov}}\eta_{\mathrm{cov}} + \eta_{\mathrm{cov}})B^2 \\
&\le \kappa_{\mathrm{cov}}^2\mathbb{E}_{\mathcal{D}_{1\otimes 1}}[\|\langle \underline{\mathbf{f}}^\star_{>\mathbf{k}}, \underline{\mathbf{g}}^\star_{>\mathbf{k}} \rangle_{\mathcal{M}}\|^2] + 2\kappa_{\mathrm{cov}}\eta_{\mathrm{cov}}B^2 \qquad \text{(since } \kappa_{\mathrm{cov}} \ge 1) \\
&= \kappa_{\mathrm{cov}}^2\mathbb{E}_{\mathcal{D}_{1\otimes 1}}[\|\underline{h}^\star - \langle \underline{\mathbf{f}}^\star_{\mathbf{k}}, \underline{\mathbf{g}}^\star_{\mathbf{k}} \rangle_{\mathcal{M}}\|^2] + 2\kappa_{\mathrm{cov}}\eta_{\mathrm{cov}}B^2 \\
&= \kappa_{\mathrm{cov}}^2\mathcal{R}(\underline{\mathbf{f}}^\star_{\mathbf{k}}, \underline{\mathbf{g}}^\star_{\mathbf{k}}; \mathcal{D}_{1\otimes 1}) + 2\kappa_{\mathrm{cov}}\eta_{\mathrm{cov}}B^2.
\end{aligned}$$

This completes the proof. □

D.1.6 Proof of Lemma D.5

Proof of Lemma D.5. The lemma states that if $(\underline{\mathbf{f}}, \underline{\mathbf{g}})$ are aligned $\mathbf{k}$-proxies of $(\hat{\underline{\mathbf{f}}}, \hat{\underline{\mathbf{g}}})$, then the following inequality holds:

$$(\min_i \{\sigma_{r_i}(\breve{\hat{\underline{\mathbf{f}}}}^{(i)}, \breve{\underline{\mathbf{g}}}^{(i)})\})\Delta_2 \lesssim \Delta_{\text{train}} + \kappa_{\text{cov}}(\Delta_0 + \Delta_{\text{apx}}(\mathbf{k}))$$

where $\Delta_2 := \max \left\{ \mathbb{E}_{\mathcal{D}_{\mathcal{X},2}} \|\underline{\mathbf{f}}^\star_{\mathbf{k}} - \underline{\mathbf{f}}\|^2_{\mathcal{M}}, \mathbb{E}_{\mathcal{D}_{\mathcal{Y},2}} \|\underline{\mathbf{g}}^\star_{\mathbf{k}} - \underline{\mathbf{g}}\|^2_{\mathcal{M}} \right\}$.

Lemma D.5 expresses that under the aligned proxies condition, the deviation Δ_2 between $(\underline{\mathbf{f}}, \underline{\mathbf{g}})$ and their rank-$\mathbf{k}$ approximations $(\underline{\mathbf{f}}^\star_{\mathbf{k}}, \underline{\mathbf{g}}^\star_{\mathbf{k}})$ under distributions $\mathcal{D}_{\mathcal{X},2}, \mathcal{D}_{\mathcal{Y},2}$ can be controlled by the error Δ_{train} on the training distribution, the weighted embedding estimation error Δ_0, and $\mathbf{k}$-approximation error $\Delta_{\text{apx}}(\mathbf{k})$, as long as the estimated embeddings $(\hat{\underline{\mathbf{f}}}, \hat{\underline{\mathbf{g}}})$ have positive minimum singular values on the corresponding modes. This lemma quantifies the effect of the aligned proxies condition.

The main idea of the proof is to consider $\mathbb{E}_{\mathcal{D}_{\mathcal{X},2}} \|\underline{\mathbf{f}}^\star_{\mathbf{k}} - \underline{\mathbf{f}}\|^2_{\mathcal{M}}$ and $\mathbb{E}_{\mathcal{D}_{\mathcal{Y},2}} \|\underline{\mathbf{g}}^\star_{\mathbf{k}} - \underline{\mathbf{g}}\|^2_{\mathcal{M}}$ separately and relate them to $\Delta_0, \Delta_{\text{apx}}(\mathbf{k}), \Delta_{\text{train}}$ using the properties of aligned proxies. Since the arguments for the two parts are symmetric, we focus on how to handle $\mathbb{E}_{\mathcal{D}_{\mathcal{Y},2}} \|\underline{\mathbf{g}}^\star_{\mathbf{k}} - \underline{\mathbf{g}}\|^2_{\mathcal{M}}$.

Our aim is to bound Δ_2. We focus on bounding $\mathbb{E}_{\mathcal{D}_{\mathcal{Y},2}} \|\underline{\mathbf{g}}^\star_{\mathbf{k}} - \underline{\mathbf{g}}\|^2_{\mathcal{M}}$, for the bound on $\mathbb{E}_{\mathcal{D}_{\mathcal{X},2}} \|\underline{\mathbf{f}}^\star_{\mathbf{k}} - \underline{\mathbf{f}}\|^2_{\mathcal{M}}$ is analogous. Further, let us recall what it means for $(\underline{\mathbf{f}}, \underline{\mathbf{g}})$ to be aligned $\mathbf{k}$-proxies. This means that (a) $\underline{\mathbf{f}} = (\iota_{\mathbf{r}} \circ \underline{\mathbf{T}}^{-1})\hat{\underline{\mathbf{f}}}$, $\underline{\mathbf{g}} = (\iota_{\mathbf{r}} \circ \underline{\mathbf{T}})\hat{\underline{\mathbf{g}}}$, where $\iota_{\mathbf{r}} : \mathbb{R}^{\|\mathbf{r}\|_\infty \times 1 \times K} \to \mathcal{M}$ is an isometric inclusion, and $\underline{\mathbf{T}}$ is the balancing operator for tensors, and (b) for $\underline{\mathbf{P}}^\star_{\mathbf{k}}$ being the multi-rank-$\mathbf{k}$ projection defined by $\underline{\mathbf{\Sigma}}^\star_{1\otimes 1}$, we have $\forall i \in [K]$:

$$\text{range}(\underline{\mathbf{P}}^\star_{\mathbf{k}}) \subseteq \text{range}(\mathbb{E}_{\mathcal{D}_{\mathcal{X},1}}[\underline{\mathbf{f}} *_M \underline{\mathbf{f}}^\top]) \Rightarrow \text{range}(M(\underline{\mathbf{P}}^\star_{\mathbf{k}})^{(i)}) \subseteq \text{range}(\mathbb{E}_{\mathcal{D}_{\mathcal{X},1}}[\breve{\underline{\mathbf{f}}}^{(i)}(\breve{\underline{\mathbf{f}}}^{(i)})^\top]). \quad (30)$$

Let $\mathcal{V} := \text{range}(\mathbb{E}_{\mathcal{D}_{\mathcal{X},1}}[\breve{\underline{\mathbf{f}}}^{(i)}(\breve{\underline{\mathbf{f}}}^{(i)})^\top])$. Since $\hat{\underline{\mathbf{f}}}, \hat{\underline{\mathbf{g}}}$ are full-multi-rank-$\mathbf{r}$, $\mathcal{V} = \text{range}(M(\iota_{\mathbf{r}})^{(i)}) = \text{range}(\mathbb{E}_{\mathcal{D}_{\mathcal{Y},1}}[\breve{\underline{\mathbf{g}}}^{(i)}(\breve{\underline{\mathbf{g}}}^{(i)})^\top])$. Moreover, $\text{range}(\mathbb{E}_{\mathcal{D}_{\mathcal{Y},1}}[\breve{\underline{\mathbf{g}}}^{\star,(i)}_{k_i}(\breve{\underline{\mathbf{g}}}^{\star,(i)}_{k_i})^\top]) = \text{range}(M(\underline{\mathbf{P}}^\star_{\mathbf{k}})^{(i)}) \subseteq \mathcal{V}_{r_i}$. Hence, by Lemma D.11 , $\breve{\underline{\mathbf{g}}}^{(i)}(y), \breve{\underline{\mathbf{g}}}^{\star,(i)}_{k_i}(y) \in \mathcal{V}$ almost surely, and thus, $\breve{\underline{\mathbf{g}}}^{\star,(i)}_{k_i}(y) - \breve{\underline{\mathbf{g}}}^{(i)}(y) \in \mathcal{V}$ with probability one. In addition, since $\mathcal{V}$ has dimension r_i, it follows that for any $\boldsymbol{\mathsf{v}} \in \mathcal{V}$, and since $\sigma_{r_i}(\mathbb{E}_{\mathcal{D}_{\mathcal{X},1}}[\breve{\underline{\mathbf{f}}}^{(i)}(\breve{\underline{\mathbf{f}}}^{(i)})^\top]) = \sigma_{r_i}(\breve{\hat{\underline{\mathbf{f}}}}^{(i)}, \breve{\hat{\underline{\mathbf{g}}}}^{(i)}) > 0$,

$$\mathbf{v}^\top \mathbb{E}_{\mathcal{D}_{\mathcal{X},1}}[\breve{\underline{\mathbf{f}}}^{(i)}(\breve{\underline{\mathbf{f}}}^{(i)})^\top]\mathbf{v} \geq \|\boldsymbol{\mathsf{v}}\|^2 \cdot \sigma_{r_i}(\breve{\hat{\underline{\mathbf{f}}}}^{(i)}, \breve{\hat{\underline{\mathbf{g}}}}^{(i)}), \quad \forall i \in [K].$$

Therefore

$$\begin{aligned}
\|\underline{\mathbf{g}}^\star_{\mathbf{k}} - \underline{\mathbf{g}}\|^2_{\mathcal{M}} &= \sum_{i=1}^{K} \|\breve{\underline{\mathbf{g}}}^{\star,(i)}_{k_i} - \breve{\underline{\mathbf{g}}}^{(i)}\|^2 && \text{(Transformed to } M\text{-domain)} \\
&\leq \sum_{i=1}^{K} \frac{1}{\sigma_{r_i}(\breve{\hat{\underline{\mathbf{f}}}}^{(i)}, \breve{\hat{\underline{\mathbf{g}}}}^{(i)})} (\breve{\underline{\mathbf{g}}}^{\star,(i)}_{k_i} - \breve{\underline{\mathbf{g}}}^{(i)})^\top \mathbb{E}_{\mathcal{D}_{\mathcal{X},1}}[\breve{\underline{\mathbf{f}}}^{(i)}(\breve{\underline{\mathbf{f}}}^{(i)})^\top](\breve{\underline{\mathbf{g}}}^{\star,(i)}_{k_i} - \breve{\underline{\mathbf{g}}}^{(i)}) \\
&= \frac{1}{\min_i \sigma_{r_i}(\breve{\hat{\underline{\mathbf{f}}}}^{(i)}, \breve{\hat{\underline{\mathbf{g}}}}^{(i)})} \sum_{i=1}^{K} \mathbb{E}_{\mathcal{D}_{\mathcal{X},1}}[\langle(\breve{\underline{\mathbf{g}}}^{\star,(i)}_{k_i} - \breve{\underline{\mathbf{g}}}^{(i)}), \breve{\underline{\mathbf{f}}}^{(i)}\rangle^2] \\
&= \frac{1}{\min_i \sigma_{r_i}(\breve{\hat{\underline{\mathbf{f}}}}^{(i)}, \breve{\hat{\underline{\mathbf{g}}}}^{(i)})} \mathbb{E}_{\mathcal{D}_{\mathcal{X},1}}[\langle(\underline{\mathbf{g}}^\star_{\mathbf{k}} - \underline{\mathbf{g}}), \underline{\mathbf{f}}\rangle^2_{\mathcal{M}}].
\end{aligned}$$

Therefore,

$$\begin{aligned}
\mathbb{E}_{\mathcal{D}_{\mathcal{Y},2}} \|\underline{\mathbf{g}}^\star_{\mathbf{k}} - \underline{\mathbf{g}}\|^2_{\mathcal{M}} &\leq \mathbb{E}_{\mathcal{D}_{\mathcal{Y},2}} \left[\mathbb{E}_{\mathcal{D}_{\mathcal{X},1}}[\langle \underline{\mathbf{g}}^\star_{\mathbf{k}} - \underline{\mathbf{g}}, \underline{\mathbf{f}}\rangle^2_{\mathcal{M}}] \right] \\
&= \frac{1}{\min_i \sigma_{r_i}(\breve{\hat{\underline{\mathbf{f}}}}^{(i)}, \breve{\hat{\underline{\mathbf{g}}}}^{(i)})} \mathbb{E}_{\mathcal{D}_{1\otimes 2}}[\langle \underline{\mathbf{f}}, \underline{\mathbf{g}}^\star_{\mathbf{k}} - \underline{\mathbf{g}}\rangle^2_{\mathcal{M}}].
\end{aligned}$$

In other words, we bound $\mathbb{E}_{\mathcal{D}_{\mathcal{Y},2}}\|\underline{\mathbf{g}}_{\mathbf{k}}^{\star} - \underline{\mathbf{g}}\|_{\mathcal{M}}^2$ by relating an expectation involving $\mathcal{D}_{\mathcal{X},1}$. Now, we can further expand

$$\begin{aligned}\langle \underline{\mathbf{f}}, \underline{\mathbf{g}}_{\mathbf{k}}^{\star} - \underline{\mathbf{g}}\rangle_{\mathcal{M}} &= \langle \underline{\mathbf{f}}, \underline{\mathbf{g}}_{\mathbf{k}}^{\star}\rangle_{\mathcal{M}} - \langle \underline{\mathbf{f}}, \underline{\mathbf{g}}\rangle_{\mathcal{M}} = \langle \underline{\mathbf{f}}, \underline{\mathbf{g}}_{\mathbf{k}}^{\star}\rangle_{\mathcal{M}} - \underline{h}^{\star} - (\langle \underline{\mathbf{f}}, \underline{\mathbf{g}}\rangle_{\mathcal{M}} - \underline{h}^{\star}) \\ &= \langle \underline{\mathbf{f}} - \underline{\mathbf{f}}_{\mathbf{k}}^{\star}, \underline{\mathbf{g}}_{\mathbf{k}}^{\star}\rangle_{\mathcal{M}} - (\underline{h}^{\star} - \langle \underline{\mathbf{f}}_{\mathbf{k}}^{\star}, \underline{\mathbf{g}}_{\mathbf{k}}^{\star}\rangle_{\mathcal{M}}) - (\langle \underline{\mathbf{f}}, \underline{\mathbf{g}}\rangle_{\mathcal{M}} - \underline{h}^{\star}).\end{aligned}$$

Hence,

$$\begin{aligned}&\min_i \sigma_{r_i}(\breve{\hat{\underline{\mathbf{f}}}}^{(i)}, \breve{\hat{\underline{\mathbf{g}}}}^{(i)})\mathbb{E}_{\mathcal{D}_{\mathcal{Y},2}}\|\underline{\mathbf{g}}_{\mathbf{k}}^{\star} - \underline{\mathbf{g}}\|_{\mathcal{M}}^2 \\ &\leq 3\mathbb{E}_{\mathcal{D}_{1\otimes 2}}\|\langle \underline{\mathbf{f}} - \underline{\mathbf{f}}_{\mathbf{k}}^{\star}, \underline{\mathbf{g}}_{\mathbf{k}}^{\star}\rangle_{\mathcal{M}}\|^2 + 3\mathbb{E}_{\mathcal{D}_{1\otimes 2}}\|\underline{h}^{\star} - \langle \underline{\mathbf{f}}_{\mathbf{k}}^{\star}, \underline{\mathbf{g}}_{\mathbf{k}}^{\star}\rangle_{\mathcal{M}}\|^2 + 3\mathbb{E}_{\mathcal{D}_{1\otimes 2}}\|\langle \underline{\mathbf{f}}, \underline{\mathbf{g}}\rangle_{\mathcal{M}} - \underline{h}^{\star}\|^2 \\ &\leq 3\mathbb{E}_{\mathcal{D}_{1\otimes 2}}\|\langle \underline{\mathbf{f}} - \underline{\mathbf{f}}_{\mathbf{k}}^{\star}, \underline{\mathbf{g}}_{\mathbf{k}}^{\star}\rangle_{\mathcal{M}}\|^2 + 3\mathcal{R}(\underline{\mathbf{f}}_{\mathbf{k}}^{\star}, \underline{\mathbf{g}}_{\mathbf{k}}^{\star}, \mathcal{D}_{1\otimes 2}) + 3\mathcal{R}(\underline{\mathbf{f}}, \underline{\mathbf{g}}, \mathcal{D}_{1\otimes 2}).\end{aligned}$$

By Lemma D.9 and the fact that $\|f - f_k^{\star}\|_{\mathcal{M}}$ is $2B$-bounded,

$$\mathbb{E}_{\mathcal{D}_{1\otimes 2}}\|\langle \underline{\mathbf{f}} - \underline{\mathbf{f}}_{\mathbf{k}}^{\star}, \underline{\mathbf{g}}_{\mathbf{k}}^{\star}\rangle_{\mathcal{M}}\|^2 \leq \kappa_{\text{cov}}\mathbb{E}_{\mathcal{D}_{1\otimes 1}}\|\langle \underline{\mathbf{f}} - \underline{\mathbf{f}}_{\mathbf{k}}^{\star}, \underline{\mathbf{g}}_{\mathbf{k}}^{\star}\rangle_{\mathcal{M}}\|^2 + 4B^4\eta_{\text{cov}} = \kappa_{\text{cov}}\Delta_0 + 4B^4\eta_{\text{cov}}.$$

By Lemma D.10, $\mathcal{R}(\underline{\mathbf{f}}_{\mathbf{k}}^{\star}, \underline{\mathbf{g}}_{\mathbf{k}}^{\star}; \mathcal{D}_{1\otimes 2}) \leq \kappa_{\text{cov}}\mathcal{R}(\underline{\mathbf{f}}_{\mathbf{k}}^{\star}, \underline{\mathbf{g}}_{\mathbf{k}}^{\star}; \mathcal{D}_{1\otimes 1}) + \eta_{\text{cov}}B^4 = \kappa_{\text{cov}}\Delta_{\text{apx}} + \eta_{\text{cov}}B^4$. Finally, by applying Lemma D.7,

$$\mathcal{R}(\underline{\mathbf{f}}, \underline{\mathbf{g}}; \mathcal{D}_{1\otimes 2}) \leq 4B^4\eta_{\text{tst}} + \kappa_{\text{trn}}\mathcal{R}(\underline{\mathbf{f}}, \underline{\mathbf{g}}; \mathcal{D}_{\text{train}}) = 4B^4\eta_{\text{tst}} + \kappa_{\text{trn}}\Delta_{\text{train}}.$$

Thus, $\min_i\{\sigma_{r_i}(\breve{\hat{\underline{\mathbf{f}}}}^{(i)}, \breve{\hat{\underline{\mathbf{g}}}}^{(i)})\} \cdot \mathbb{E}_{\mathcal{D}_{\mathcal{Y},2}}\|\underline{\mathbf{g}}_{\mathbf{k}}^{\star} - \underline{\mathbf{g}}\|_{\mathcal{M}}^2 \leq 3\kappa_{\text{cov}}(\Delta_0 + \Delta_{\text{apx}}) + 3\kappa_{\text{trn}}\Delta_{\text{train}} + 12B^4\eta_{\text{cov}} + \eta_{\text{tst}}$. Considering the symmetric argument for $\mathbb{E}_{\mathcal{D}_{\mathcal{X},2}}\|\underline{\mathbf{f}}_{\mathbf{k}}^{\star} - \underline{\mathbf{f}}\|_{\mathcal{M}}^2$ and taking the maximum of the two bounds, we arrive at the conclusion of the lemma:

$$(\min_i\{\sigma_{r_i}(\breve{\hat{\underline{\mathbf{f}}}}^{(i)}, \breve{\hat{\underline{\mathbf{g}}}}^{(i)})\})\Delta_2 \lesssim \Delta_{\text{train}} + \kappa_{\text{cov}}(\Delta_0 + \Delta_{\text{apx}}(\mathbf{k})) + B^4(\eta_{\text{cov}} + \eta_{\text{tst}}).$$

This completes the proof. □

Lemma D.11 (Lemma L.7 in Ref. [46]). *Let $\mathcal{D}_{\mathcal{X}}$ be a distribution over $\mathcal{X}$, let $\mathbf{\Sigma} = \mathbb{E}_{\mathcal{D}_{\mathcal{X}}}[ff^{\top}]$, and let $\mathbf{P}$ be the orthogonal projection on $\mathrm{range}(\mathbf{\Sigma})$. Then $\mathbf{P}f = f$ $\mathcal{D}_{\mathcal{X}}$-almost surely; that is, $\mathbb{P}_{\mathcal{D}_{\mathcal{X}}}[f(x) \in \mathrm{range}(\mathbf{\Sigma})] = 1$.*

D.1.7 Proof of LemmeD.6

Proof of Lemma D.6. Since the t-projection operators are self-adjoint and idempotent, we begin by expressing the risk term as follows:

$$\mathcal{R}(\underline{\mathbf{f}}_{\mathbf{k}}^{\star}, \underline{\mathbf{g}}_{\mathbf{k}}^{\star}; \mathcal{D}_{1\otimes 1}) = \mathbb{E}_{\mathcal{D}_{1\otimes 1}}\left[\|\underline{\mathbf{f}}_{\mathbf{k}}^{\star}(x)^{\top} *_M \underline{\mathbf{g}}_{\mathbf{k}}^{\star}(y) - \langle \underline{\mathbf{f}}^{\star}(x), \underline{\mathbf{g}}^{\star}(y)\rangle_{\mathcal{M}}\|^2\right].$$

Rewriting this in terms of projection operators, we have

$$\begin{aligned}\mathcal{R}(\underline{\mathbf{f}}_{\mathbf{k}}^{\star}, \underline{\mathbf{g}}_{\mathbf{k}}^{\star}; \mathcal{D}_{1\otimes 1}) &= \mathbb{E}_{\mathcal{D}_{1\otimes 1}}\left[\|\langle \underline{\mathbf{P}}_{\mathbf{k}}^{\star}\underline{\mathbf{f}}^{\star}(x), \underline{\mathbf{P}}_{\mathbf{k}}^{\star}\underline{\mathbf{g}}^{\star}(y)\rangle_{\mathcal{M}} - \langle \underline{\mathbf{f}}^{\star}(x), \underline{\mathbf{g}}^{\star}(y)\rangle_{\mathcal{M}}\|^2\right] \\ &= \mathbb{E}_{\mathcal{D}_{1\otimes 1}}\left[\|\langle \underline{\mathbf{P}}_{\mathbf{k}}^{\star}\underline{\mathbf{f}}^{\star}(x), \underline{\mathbf{g}}^{\star}(y)\rangle_{\mathcal{M}} - \langle \underline{\mathbf{f}}^{\star}(x), \underline{\mathbf{g}}^{\star}(y)\rangle_{\mathcal{M}}\|^2\right] \\ &= \mathbb{E}_{\mathcal{D}_{1\otimes 1}}\left[\|\langle (\underline{\mathcal{I}} - \underline{\mathbf{P}}_{\mathbf{k}}^{\star})\underline{\mathbf{f}}^{\star}(x), \underline{\mathbf{g}}^{\star}(y)\rangle_{\mathcal{M}}\|^2\right].\end{aligned}$$

Next, switching to the M-domain by transforming into individual components, we find

$$\mathcal{R}(\underline{\mathbf{f}}_{\mathbf{k}}^{\star}, \underline{\mathbf{g}}_{\mathbf{k}}^{\star}; \mathcal{D}_{1\otimes 1}) = \mathbb{E}_{\mathcal{D}_{1\otimes 1}}\left[\sum_{i=1}^{K}\langle (\mathbf{I} - \breve{\mathbf{P}}^{(i)})\breve{\underline{\mathbf{f}}}^{\star,(i)}(x), \breve{\underline{\mathbf{g}}}^{\star,(i)}(y)\rangle_{\mathcal{H}}^2\right].$$

Since expectation and summation are interchangeable, we can separate the components as

$$\mathcal{R}(\underline{\mathbf{f}}_{\mathbf{k}}^{\star}, \underline{\mathbf{g}}_{\mathbf{k}}^{\star}; \mathcal{D}_{1\otimes 1}) = \sum_{i=1}^{K}\mathbb{E}_{\mathcal{D}_{1\otimes 1}}\left[\langle (\mathbf{I} - \breve{\mathbf{P}}^{(i)})\breve{\underline{\mathbf{f}}}^{\star,(i)}(x), \breve{\underline{\mathbf{g}}}^{\star,(i)}(y)\rangle_{\mathcal{H}}^2\right].$$

Using the trace operator to express this in matrix form, we get

$$\mathcal{R}(\underline{\mathbf{f}}^\star_{\mathbf{k}}, \underline{\mathbf{g}}^\star_{\mathbf{k}}; \mathcal{D}_{1\otimes 1}) = \sum_{i=1}^{K} \mathtt{Tr}\left((\mathbf{I} - \breve{\mathbf{P}}^{(i)}) \cdot \mathbb{E}_{\mathcal{D}_{\mathcal{X},1}} \left[\breve{\underline{\mathbf{f}}}^{\star,(i)}(x) \breve{\underline{\mathbf{f}}}^{\star,(i)}(x)^\top \right] \cdot (\mathbf{I} - \breve{\mathbf{P}}^{(i)}) \mathbb{E}_{\mathcal{D}_{\mathcal{Y},1}} \left[\breve{\underline{\mathbf{g}}}^{\star,(i)}(y) \breve{\underline{\mathbf{g}}}^{\star,(i)}(y)^\top \right] \right).$$

Since $\mathbb{E}_{\mathcal{D}_{\mathcal{X},1}} \left[\breve{\underline{\mathbf{f}}}^{\star,(i)}(x) \breve{\underline{\mathbf{f}}}^{\star,(i)}(x)^\top \right] = M(\underline{\mathbf{\Sigma}}^\star_{1\otimes 1})^{(i)}$ and $\mathbb{E}_{\mathcal{D}_{\mathcal{Y},1}} \left[\breve{\underline{\mathbf{g}}}^{\star,(i)}(y) \breve{\underline{\mathbf{g}}}^{\star,(i)}(y)^\top \right] = M(\underline{\mathbf{\Sigma}}^\star_{1\otimes 1})^{(i)}$, we obtain

$$\mathcal{R}(\underline{\mathbf{f}}^\star_{\mathbf{k}}, \underline{\mathbf{g}}^\star_{\mathbf{k}}; \mathcal{D}_{1\otimes 1}) = \sum_{i=1}^{K} \mathtt{Tr}\left((\mathbf{I} - \breve{\mathbf{P}}^{(i)}) \cdot M(\underline{\mathbf{\Sigma}}^\star_{1\otimes 1})^{(i)} \cdot (\mathbf{I} - \breve{\mathbf{P}}^{(i)}) M(\underline{\mathbf{\Sigma}}^\star_{1\otimes 1})^{(i)} \right).$$

Finally, by summing over the eigenvalues of the transformed covariance matrices, we find

$$\mathcal{R}(\underline{\mathbf{f}}^\star_{\mathbf{k}}, \underline{\mathbf{g}}^\star_{\mathbf{k}}; \mathcal{D}_{1\otimes 1}) = \sum_{i=1}^{K} \sum_{j > k_i} \lambda_j \left(M(\underline{\mathbf{\Sigma}}^\star_{1\otimes 1})^{(i)} \right)^2 := \mathsf{tail}^\star_2(\mathbf{k}).$$

This completes the proof. □

D.2 Excess Risk Bounds for the ERM Model

In this section, we analyze the ERM algorithm for multi-output regression under CDS. Our primary objective is to establish a rigorous upper bound on the excess risk $\mathcal{R}(\hat{\underline{\mathbf{f}}}_{\text{erm}}, \hat{\underline{\mathbf{g}}}_{\text{erm}}; \mathcal{D}_{\text{test}})$ of the ERM solution $(\hat{\underline{\mathbf{f}}}_{\text{erm}}, \hat{\underline{\mathbf{g}}}_{\text{erm}})$ obtained by Model (5).

Recall that Theorem 4 states the following: Given vectors $\mathbf{r} = (r_i)_{i=1}^K \in \mathbb{N}^K$, $\boldsymbol{\alpha} = (\alpha_i)_{i=1}^K \in \mathbb{R}^K$, $\boldsymbol{\epsilon}_{\text{trn}} = (\breve{\epsilon}^{(i)}_{\text{trn}})_{i=1}^K \in \mathbb{R}^K$, $\boldsymbol{\epsilon}_{\mathcal{D}_{1\otimes 1}} = (\breve{\epsilon}^{(i)}_{\mathcal{D}_{1\otimes 1}})_{i=1}^K \in \mathbb{R}^K$ where $\alpha_i \geq 1$, and $\breve{\epsilon}^{(i)}_{\text{trn}}$, $\breve{\epsilon}^{(i)}_{\mathcal{D}_{1\otimes 1}} \geq 0$ for all $i \in [K]$, suppose the learned embeddings $(\hat{\underline{\mathbf{f}}}_{\text{erm}}, \hat{\underline{\mathbf{g}}}_{\text{erm}})$ are $\boldsymbol{\alpha}$-conditioned and $(\boldsymbol{\epsilon}_{\text{trn}}, \boldsymbol{\epsilon}_{\mathcal{D}_{1\otimes 1}})$-accurate, satisfying $\breve{\epsilon}^{(i)}_{\mathcal{D}_{1\otimes 1}} \leq \breve{\sigma}_1^{\star,(i)}/(40r)$ for all $i \in [K]$. Then, under Assumptions 1 to 5 and 8, the excess risk of the ERM solution on $\mathcal{D}_{\text{test}}$ can be bounded up to a constant factor $c = \text{poly}(\kappa_{\text{cov}}, \kappa_{\text{tst}}, \kappa_{\text{trn}})$ with probability at least $1 - \delta$:

$$\underbrace{\alpha r^2 \sum_i (\breve{\sigma}_r^{\star,(i)})^2 + r^4 \mathsf{tail}^\star_2(r) + r^2 (\mathsf{tail}^\star_1(r))^2 + \frac{\alpha r^6 (\mathsf{tail}^\star_2(r))^2}{\sigma^2}}_{\text{approximation error}} + \underbrace{r^4 \Delta_n + \frac{\alpha r^6}{\sigma^2} \Delta_n^2}_{\text{statistical error}},$$

where $\alpha := \max_i\{\alpha_i\}$, $\sigma := \min_i\{\breve{\sigma}_r^{\star,(i)}\} > 0$, $\Delta_n = B^4(\mathscr{N}(r, 2B/n) + \log \frac{2}{\delta})/n$ with $\mathscr{N}(r, \epsilon) = \mathcal{N}(\mathcal{F}_r, \epsilon/(2B), \|\cdot\|_\infty) \cdot \mathcal{N}(\mathcal{G}_r, \epsilon/(2B), \|\cdot\|_\infty)$.

To prove Theorem 4, we begin by applying Lemma D.2 with the multi-rank parameter $\mathbf{r} = (r, \dots, r)^\top \in \mathbb{R}^K$, where r is the tubal-rank parameter of the embeddings in the ERM formulation (5). This lemma provides a general decomposition of the excess risk into several key error terms, each capturing a different aspect of the learning problem:

$$\mathcal{R}(\hat{\underline{\mathbf{f}}}_{\text{erm}}, \hat{\underline{\mathbf{g}}}_{\text{erm}}; \mathcal{D}_{\text{test}}) \lesssim_\star \left(\Delta_1^2 + \frac{1}{\sigma^2} \left(\Delta_{\text{apx}} + \Delta_0 + \Delta_{\text{train}} \right)^2 \right).$$

This decomposition lays the groundwork for our subsequent analysis, as it allows us to focus on bounding each error term separately.

Next, we introduce a set of conditions on the ERM solution to ensure its quality and enable more refined bounds. Specifically, we assume that the learned embeddings $(\hat{\underline{\mathbf{f}}}_{\text{erm}}, \hat{\underline{\mathbf{g}}}_{\text{erm}})$ are $\boldsymbol{\alpha}$-conditioned and $(\boldsymbol{\epsilon}_{\text{trn}}, \boldsymbol{\epsilon}_{\mathcal{D}_{1\otimes 1}})$-accurate, as formalized in Definition 17 and Definition 18, respectively.

Moreover, we impose a constraint on the rank r of the embeddings, requiring that $r \leq \breve{\sigma}_1^{\star,(i)}/(40\breve{\epsilon}^{(i)}_{\mathcal{D}_{1\otimes 1}})$ for all $i \in [K]$. Under these assumptions, we can leverage the powerful Lemma

D.22 to obtain a more refined bound on the excess risk:

$$\begin{aligned}
&\mathcal{R}(\hat{\underline{\mathbf{f}}}_{\text{erm}}, \hat{\underline{\mathbf{g}}}_{\text{erm}}; \mathcal{D}_{\text{test}}) \\
&= \sum_{i=1}^{K} \breve{\mathcal{R}}^{(i)}(\breve{\hat{\underline{\mathbf{f}}}}^{(i)}_{\text{erm}}, \breve{\hat{\underline{\mathbf{g}}}}^{(i)}_{\text{erm}}; \mathcal{D}_{\text{test}}) \\
&\lesssim_{\star} \sum_{i=1}^{K} \left(r^4 (\breve{\epsilon}^{(i)}_{\mathcal{D}_{1\otimes 1}})^2 + \alpha_i r^2 (\breve{\sigma}^{\star,(i)}_{r+1})^2 + (\textsf{tail}^{(i)\star}_1(r))^2 \right) + \alpha_i \left(\frac{r^6 (\breve{\epsilon}^{(i)}_{\mathcal{D}_{1\otimes 1}})^4 + (\breve{\epsilon}^{(i)}_{\text{trn}})^4 + (\textsf{tail}^{(i)\star}_2(r))^2}{(\breve{\sigma}^{\star,(i)}_r)^2} \right) \\
&\lesssim_{\star} r^4 \cdot \mathrm{err}^2_{1\otimes 1} + \max_i\{\alpha_i\} \cdot r^2 \sum_{i=1}^{K} \left((\breve{\sigma}^{\star,(i)}_{r+1})^2 + (\textsf{tail}^{(i)\star}_1(r)) \right) \\
&\qquad\qquad + \max_i\{\alpha_i\} \left(\frac{r_i^6 \sum_i (\breve{\epsilon}^{(i)}_{\mathcal{D}_{1\otimes 1}})^4 + \sum_i (\breve{\epsilon}^{(i)}_{\text{trn}})^4 + \sum_i (\textsf{tail}^{(i)\star}_2(r))^2}{(\min_i \breve{\sigma}^{\star,(i)}_r)^2} \right) \\
&\lesssim_{\star} r^4 \cdot \mathrm{err}^2_{1\otimes 1} + \max_i\{\alpha_i\} \cdot r^2 \|\underline{\sigma}^{\star}_{r+1}\|^2 + \textsf{tail}^{\star}_1(r)^2 + \max_i\{\alpha_i\} \left(\frac{r_i^6 \cdot \mathrm{err}^4_{1\otimes 1} + \mathrm{err}^4_{\text{trn}} + (\textsf{tail}^{\star}_2(r))^2}{(\min_i \breve{\sigma}^{\star,(i)}_r)^2} \right),
\end{aligned}$$

where $\mathrm{err}_{\text{trn}} := \mathcal{R}(\hat{\underline{\mathbf{f}}}_{\text{erm}}, \hat{\underline{\mathbf{g}}}_{\text{erm}}; \mathcal{D}_{\text{train}})$ and $\mathrm{err}_{1\otimes 1} := \mathcal{R}(\hat{\underline{\mathbf{f}}}_{\text{erm}}, \hat{\underline{\mathbf{g}}}_{\text{erm}}; \mathcal{D}_{1\otimes 1})$ denote the accuracy on $\mathcal{D}_{\text{train}}$ and $\mathcal{D}_{1\otimes 1}$, respectively. This bound reveals the intricate dependencies of the excess risk on various problem parameters, such as the accuracy of the learned embeddings ($\mathrm{err}_{1\otimes 1}$ and $\mathrm{err}_{\text{trn}}$), the conditioning of the true embeddings (α_i), and the spectral properties of the true embeddings ($\underline{\sigma}^{\star}_{r+1}$, $\textsf{tail}^{\star}_1(r)$, and $\textsf{tail}^{\star}_2(r)$). However, to make this bound more explicit, we still need to control the error terms $\mathrm{err}_{\text{trn}}$ and $\mathrm{err}_{1\otimes 1}$.

To this end, we use Lemma D.12 below to establish high-probability bounds on the excess risks $\mathcal{R}(\hat{\underline{\mathbf{f}}}_{\text{erm}}, \hat{\underline{\mathbf{g}}}_{\text{erm}}; \mathcal{D}_{\text{train}})$ and $\mathcal{R}(\hat{\underline{\mathbf{f}}}_{\text{erm}}, \hat{\underline{\mathbf{g}}}_{\text{erm}}; \mathcal{D}_{1\otimes 1})$ of the empirical risk minimizers $(\hat{\underline{\mathbf{f}}}_{\text{erm}}, \hat{\underline{\mathbf{g}}}_{\text{erm}})$. These bounds are expressed in terms of the true risk $\mathcal{R}(\underline{\mathbf{f}}^{\star}_r, \underline{\mathbf{g}}^{\star}_r; \mathcal{D}_{\text{train}})$, the tail sum $\textsf{tail}^{\star}_2(r)$, and the covering number $\mathscr{N}(r, 2B/n)$, which quantifies the complexity of the hypothesis class. By carefully balancing these terms, Lemma D.12 provides the key ingredients needed to control $\mathrm{err}_{\text{trn}}$ and $\mathrm{err}_{1\otimes 1}$.

Lemma D.12. *Under Assumptions 1 to 5 and 8, let $(\tilde{\underline{\mathbf{f}}}, \tilde{\underline{\mathbf{g}}}) \in \mathcal{F}_p \times \mathcal{G}_p$ be empirical risk minimizers on n i.i.d. samples $(x_i, y_i, \mathbf{z}_i) \sim \mathcal{D}_{\text{train}}$. Then, for any $\delta \in (0,1)$, the followings hold with probability at least $1-\delta$:*

$$\begin{aligned}
\mathcal{R}(\hat{\underline{\mathbf{f}}}_{\text{erm}}, \hat{\underline{\mathbf{g}}}_{\text{erm}}; \mathcal{D}_{\text{train}}) &\leq 2\mathcal{R}(\underline{\mathbf{f}}^{\star}_r, \underline{\mathbf{g}}^{\star}_r; \mathcal{D}_{\text{train}}) + \frac{368B^4(\mathscr{N}(r, 2B/n) + \log\frac{2}{\delta})}{n} \\
\mathcal{R}(\hat{\underline{\mathbf{f}}}_{\text{erm}}, \hat{\underline{\mathbf{g}}}_{\text{erm}}; \mathcal{D}_{\text{train}}) &\leq 2\kappa_{\text{apx}} \textsf{tail}^{\star}_2(r) + \frac{368B^4(\mathscr{N}(r, 2B/n) + \log\frac{2}{\delta})}{n} \\
\mathcal{R}(\hat{\underline{\mathbf{f}}}_{\text{erm}}, \hat{\underline{\mathbf{g}}}_{\text{erm}}; \mathcal{D}_{1\otimes 1}) &\leq \kappa_{\text{trn}} \left(2\kappa_{\text{apx}} \textsf{tail}^{\star}_2(r) + \frac{368B^4(\mathscr{N}(r, 2B/n) + \log\frac{2}{\delta})}{n} \right).
\end{aligned}$$

where $\mathscr{N}(r, \varepsilon) = \mathcal{N}(\mathcal{F}_r, \varepsilon/(2B), \|\cdot\|_{\infty}) \cdot \mathcal{N}(\mathcal{G}_r, \varepsilon/(2B), \|\cdot\|_{\infty})$.

Proof of Lemma D.12. The lemma is proved as follows:

- The first inequality is a direct consequence of Lemma D.19. Here, we take the function class $\Phi = \mathcal{F}_r \times \mathcal{G}_r$. Moreover, by Assumption 8,

$$\sup_{\underline{\mathbf{f}} \in \mathcal{F}_r, \underline{\mathbf{g}} \in \mathcal{G}_r} \sup_{x,y} \|\langle \underline{\mathbf{f}}(x), \underline{\mathbf{g}}(y) \rangle_{\mathcal{M}} - \langle \underline{\mathbf{f}}^{\star}(x), \underline{\mathbf{g}}^{\star}(y) \rangle_{\mathcal{M}}\| \leq 2B^2,$$

  Then, we can choose $U = 2B^2$ and $\epsilon = 2B^2/n$ in Lemma D.19. Note that, according to Lemma D.20, it holds that

$$\mathcal{N}(\Phi, 2B^2/n, \|\cdot\|_{\infty}) \leq \mathcal{N}(\mathcal{F}_r, 2B/n, \|\cdot\|_{\infty}) \cdot \mathcal{N}(\mathcal{G}_r, 2B/n, \|\cdot\|_{\infty}).$$

- The second inequality uses Assumption 8 to bound $\mathcal{R}(\underline{\mathbf{f}}_p^\star, \underline{\mathbf{g}}_p^\star; \mathcal{D}_{\mathrm{train}}) \leq \kappa_{\mathrm{apx}} \mathcal{R}(\underline{\mathbf{f}}_p^\star, \underline{\mathbf{g}}_p^\star; \mathcal{D}_{1\otimes 1})$, and noting the fact that $\mathcal{R}(\underline{\mathbf{f}}_p^\star, \underline{\mathbf{g}}_p^\star; \mathcal{D}_{1\otimes 1}) = \textsf{tail}_2^\star(p)$ by Lemma D.6.
- The third inequality uses Assumption 2, incurring an addition factor of κ_{trn}.

□

Now, we can proceed to derive a more explicit bound on the excess risk.

Proof of Theorem 4. By combining the results of Lemmas D.2, D.22, and D.12, and introducing some convenient shorthand notations ($\sigma = \min_i \breve{\sigma}_r^{\star,(i)} > 0$, $\alpha = \max_i \alpha_i$, and $\Delta_n = \frac{B^4(\mathscr{N}(r, 2B/n) + \log\frac{2}{\delta})}{n}$), we arrive at the following bound:

$$
\begin{aligned}
&\mathcal{R}(\hat{\underline{\mathbf{f}}}_{\text{erm}}, \hat{\underline{\mathbf{g}}}_{\text{erm}}; \mathcal{D}_{\mathrm{test}}) \\
&\lesssim_\star r^4 \cdot \mathrm{err}_{1\otimes 1}^2 + \max_i\{\alpha_i\} \cdot r^2 \|\underline{\sigma}_{r+1}^\star\|^2 + (\textsf{tail}_1^\star(r))^2 + \max_i\{\alpha_i\} \left(\frac{r_i^6 \cdot \mathrm{err}_{1\otimes 1}^4 + \mathrm{err}_{\text{trn}}^4 + (\textsf{tail}_2^\star(r))^2}{(\min_i \breve{\sigma}_r^{\star,(i)})^2} \right) \\
&\lesssim_\star r^4 \cdot \textsf{tail}_2^\star(r) + \alpha r^2 \|\underline{\sigma}_{r+1}^\star\|^2 + r^2 \cdot \textsf{tail}_1^\star(r)^2 + r^4 \frac{B^4(\mathscr{N}(r, 2B/n) + \log\frac{2}{\delta})}{n} \\
&\qquad + \alpha \frac{r^6 \cdot \textsf{tail}_2^\star(r)^2}{\sigma^2} + \alpha \frac{r^6}{\sigma^2} \left(\frac{B^4(\mathscr{N}(r, 2B/n) + \log\frac{2}{\delta})}{n} \right)^2 \\
&\lesssim_\star \underbrace{r^4 \cdot \textsf{tail}_2^\star(r) + \alpha r^2 \sum_i (\breve{\sigma}^{\star,(i)})^2 + r^2 \cdot \textsf{tail}_1^\star(r)^2 + \frac{\alpha r^6 \cdot \textsf{tail}_2^\star(r)^2}{\sigma^2}}_{\text{approximation error}} + \underbrace{r^4 \Delta_n + \frac{\alpha r^6}{\sigma^2} \Delta_n^2}_{\text{statistical error}}.
\end{aligned}
$$

□

Limitations of the ERM Model. The ERM Model (5) faces several challenges in multi-output regression under CDS:

- Inflexibility in adapting to spectral decay patterns: The hypothesis classes $\mathcal{F}_r$ and $\mathcal{G}_r$ in single-stage ERM do not account for varying spectral decay patterns across different frequency components induced by the transform $M(\cdot)$. Consequently, the learned embeddings may fail to capture distinct decay characteristics within each frequency sub-domain, resulting in suboptimal approximations.
- Lack of localized learning and complexity control: The ERM approach learns the embeddings $\hat{\underline{\mathbf{f}}}_{\text{erm}}$ and $\hat{\underline{\mathbf{g}}}_{\text{erm}}$ globally, disregarding the unique properties of each frequency sub-domain. This global learning strategy can miss the nuanced decay variations across sub-domains. Additionally, controlling the complexity of $\mathcal{F}_r$ and $\mathcal{G}_r$ through global covering numbers may be too coarse to effectively capture these variations.
- Limited generalization performance: Due to these constraints, the single-stage ERM approach may offer limited generalization guarantees for multi-output regression under CDS. The embeddings may struggle to generalize to new feature combinations, especially when spectral decay patterns differ significantly between frequency sub-domains.

D.3 Excess Risk Bounds for ERM-DS

The ERM-DS algorithm is introduced to address the limitations inherent in the ERM approach. This section provides a comprehensive analysis of the ERM-DS algorithm. The algorithm consists of four main stages: overparameterization, t-covariance estimation, dimension reduction, and distillation. We analyze each stage separately and then combine the results to obtain an overall generalization bound. We first give a detailed version for Theorem 5 as follows.

Theorem 6. *For any $r_{i,\text{cut}} \gtrsim_\star \text{poly}(C/\sigma_1^\star)$ and $\epsilon, \delta > 0$, there exists a choice of $\sigma_{\text{cut}} > 0$, $p \lesssim_\star (\min_i r_{i,\text{cut}})^c$ for some universal $c > 0$, and sample sizes $n_{1:4}$ satisfy Condition 8, such that with $\nu_i = r_{i,\text{cut}}^4$ and $\mu = B^2/n_1$, the ERM-DS algorithm satisfies with probability at least $1-\delta$:*

$$\mathbb{E}_{(x,y)\in\mathcal{D}_{\text{train}}}[\|\hat{\underline{\mathbf{f}}}_{\text{ds}}(x)^\top *_M \hat{\underline{\mathbf{g}}}_{\text{ds}}(y) - \underline{h}^\star(x,y)\|^2] \lesssim_\star \epsilon^2 + C^2 \sum_{i=1}^{K} (1+\gamma_i^{-2}) r_{i,\text{cut}}^{-2\gamma_i}.$$

Our proof follows the core idea and approach of Theorem 6 in Ref. [46]. The main steps of the proof are provided as follows.

Analysis of Step 1 (Overparameterization). We begin with a precise analysis of the first phase of the double-stage ERM. We first show that after the first step, we have the following upper bound on the excess risk on $\mathcal{D}_{\text{train}}$ and $\mathcal{D}_{1\otimes 1}$.

Lemma D.13 (Guarantee for ERM-DS, Step 1)**.** *Let $(\tilde{\underline{\mathbf{f}}}, \tilde{\underline{\mathbf{g}}})$ be the obtained t-embeddings obtained via Step 1 of ERM-DS on n_1 i.i.d. samples $(x_{1,j}, y_{1,j}, \mathbf{z}_{1,j}) \sim \mathcal{D}_{\text{train}}$. Then, for any $\delta \in (0,1)$, the followings hold with probability at least $1-\delta$:*

$$\begin{aligned}
\mathcal{R}(\tilde{\underline{\mathbf{f}}}, \tilde{\underline{\mathbf{g}}}; \mathcal{D}_{\text{train}}) &\le 2\mathcal{R}(\underline{\mathbf{f}}_p^\star, \underline{\mathbf{g}}_p^\star; \mathcal{D}_{\text{train}}) + \frac{368B^4(\sum_i \breve{\mathscr{N}}^{(i)}(p, 2B/n_1) + K\log\frac{2K}{\delta})}{n_1}\\
\mathcal{R}(\tilde{\underline{\mathbf{f}}}, \tilde{\underline{\mathbf{g}}}; \mathcal{D}_{\text{train}}) &\le 2\kappa_{\text{apx}}\textsf{tail}_2^\star(p) + \frac{368B^4(\sum_i \breve{\mathscr{N}}^{(i)}(p, 2B/n_1) + K\log\frac{2K}{\delta})}{n_1}\\
\mathcal{R}(\tilde{\underline{\mathbf{f}}}, \tilde{\underline{\mathbf{g}}}; \mathcal{D}_{1\otimes 1}) &\le \kappa_{\text{trn}}\left(2\kappa_{\text{apx}}\textsf{tail}_2^\star(p) + \frac{368B^4(\sum_i \breve{\mathscr{N}}^{(i)}(p, 2B/n_1) + K\log\frac{2K}{\delta})}{n_1}\right).
\end{aligned}$$

where $\breve{\mathscr{N}}^{(i)}(p,\epsilon) = \mathcal{N}(\breve{\mathcal{F}}_p^{(i)}, \epsilon/(2B), \|\cdot\|_\infty) \cdot \mathcal{N}(\breve{\mathcal{G}}_p^{(i)}, \epsilon/(2B), \|\cdot\|_\infty)$.

This lemma establishes upper bounds on the excess risk of the initial embeddings $(\tilde{\underline{\mathbf{f}}}, \tilde{\underline{\mathbf{g}}})$ obtained in Step 1 of ERM-DS for both the training distribution $\mathcal{D}_{\text{train}}$ and distribution $\mathcal{D}_{1\otimes 1}$. It refines Lemma D.12. The main difference from Lemma D.12 lies in the covering numbers $\breve{\mathscr{N}}^{(i)}(p, 2B/n_1)$, which here relate specifically to the function classes $\breve{\mathcal{F}}_p^{(i)}$ and $\breve{\mathcal{G}}_p^{(i)}$ in Assumption 9, rather than to $\mathcal{F}$ and $\mathcal{G}$ in Assumption 8. The proof is similar and leverages techniques from statistical learning theory, along with the specific properties of the function classes $\breve{\mathcal{F}}_p^{(i)}$ and $\breve{\mathcal{G}}_p^{(i)}$.

Proof of Lemma D.13. As $\mathcal{R}(\tilde{\underline{\mathbf{f}}}, \tilde{\underline{\mathbf{g}}}; \mathcal{D}) = \sum_{i=1}^K \breve{\mathcal{R}}^{(i)}(\tilde{\breve{\underline{\mathbf{f}}}}^{(i)}, \tilde{\breve{\underline{\mathbf{g}}}}^{(i)}; \mathcal{D})$, we prove the result by upper bounding the excess risk in all the transformed sub-domains. For all $i \in [K]$, let $(\tilde{\breve{\underline{\mathbf{f}}}}^{(i)}, \tilde{\breve{\underline{\mathbf{g}}}}^{(i)}) \in \breve{\mathcal{F}}_p^{(i)} \times \breve{\mathcal{G}}_p^{(i)}$ be empirical risk minimizers on n_1 i.i.d. samples $(x_j, y_j, \breve{\mathbf{z}}_j^{(i)}) \sim \mathcal{D}_{\text{train}}$. Then, we will prove that for any $\delta \in (0,1)$, the followings hold with probability at least $1-\delta/K$:

$$\begin{aligned}
\breve{\mathcal{R}}^{(i)}(\tilde{\breve{\underline{\mathbf{f}}}}^{(i)}, \tilde{\breve{\underline{\mathbf{g}}}}^{(i)}; \mathcal{D}_{\text{train}}) &\le 2\breve{\mathcal{R}}^{(i)}(\breve{\underline{\mathbf{f}}}_p^{\star,(i)}, \breve{\underline{\mathbf{g}}}_p^{\star,(i)}; \mathcal{D}_{\text{train}}) + \frac{368B^4(\breve{\mathscr{N}}^{(i)}(p, 2B/n_1) + \log\frac{2K}{\delta})}{n_1}\\
\breve{\mathcal{R}}^{(i)}(\tilde{\breve{\underline{\mathbf{f}}}}^{(i)}, \tilde{\breve{\underline{\mathbf{g}}}}^{(i)}; \mathcal{D}_{\text{train}}) &\le 2\kappa_{\text{apx}}\textsf{tail}_2^{(i)\star}(p) + \frac{368B^4(\breve{\mathscr{N}}^{(i)}(p, 2B/n_1) + \log\frac{2K}{\delta})}{n_1}\\
\breve{\mathcal{R}}^{(i)}(\tilde{\breve{\underline{\mathbf{f}}}}^{(i)}, \tilde{\breve{\underline{\mathbf{g}}}}^{(i)}; \mathcal{D}_{1\otimes 1}) &\le \kappa_{\text{trn}}\left(2\kappa_{\text{apx}}\textsf{tail}_2^{(i)\star}(p) + \frac{368B^4(\breve{\mathscr{N}}^{(i)}(p, 2B/n_1) + \log\frac{2K}{\delta})}{n_1}\right).
\end{aligned} \tag{31}$$

- The first inequality is a direct consequence of Lemma D.19. Here, we take the function class $\Phi = \breve{\mathcal{F}}_p^{(i)} \times \breve{\mathcal{G}}_p^{(i)}$. Moreover, by Assumption 9,

$$\sup_{\breve{\underline{\mathbf{f}}}^{(i)} \in \breve{\mathcal{F}}_p^{(i)}, \breve{\underline{\mathbf{g}}}^{(i)} \in \breve{\mathcal{G}}_p^{(i)}} \sup_{x,y} \|\langle \breve{\underline{\mathbf{f}}}^{(i)}(x), \breve{\underline{\mathbf{g}}}^{(i)}(y)\rangle - \langle \breve{\underline{\mathbf{f}}}^{\star,(i)}(x), \breve{\underline{\mathbf{g}}}^{\star,(i)}(y)\rangle\| \le 2B^2,$$

Then, we can choose $U = 2B^2$ and $\epsilon = 2B^2/n_1$ in Lemma D.19. Note that, according to Lemma D.21, it holds that

$$\mathcal{N}(\Phi, 2B^2/n_1, \|\cdot\|_\infty) \leq \mathcal{N}(\breve{\mathcal{F}}_p^{(i)}, B/n_1, \|\cdot\|_\infty) \cdot \mathcal{N}(\breve{\mathcal{G}}_p^{(i)}, B/n_1, \|\cdot\|_\infty).$$

- The second inequality uses Assumption 9 to bound $\breve{\mathcal{R}}^{(i)}(\breve{\underline{\mathbf{f}}}_p^{\star,(i)}, \breve{\underline{\mathbf{g}}}_p^{\star,(i)}; \mathcal{D}_{\text{train}}) \leq \kappa_{\text{apx}} \breve{\mathcal{R}}^{(i)}(\breve{\underline{\mathbf{f}}}_p^{\star,(i)}, \breve{\underline{\mathbf{g}}}_p^{\star,(i)}; \mathcal{D}_{1\otimes 1})$, and noting the fact that $\breve{\mathcal{R}}^{(i)}(\breve{\underline{\mathbf{f}}}_p^{\star,(i)}, \breve{\underline{\mathbf{g}}}_p^{\star,(i)}; \mathcal{D}_{1\otimes 1}) = \textsf{\textbf{tail}}_2^{(i)\star}(p)$.
- The third inequality uses Assumption 2, incurring an addition factor of κ_{trn}.

□

Analysis of Step 2 (t-Covariance Estimation) and Step 3 (Dimension Reduction). Next, we analyze the effect of the t-covariance estimation and dimension reduction steps in ERM-DS. Let $(\tilde{\underline{\mathbf{f}}}, \tilde{\underline{\mathbf{g}}})$ be the empirical risk minimizers from Step 1, and let $\hat{\underline{\mathbf{Q}}}_{\hat{\mathbf{r}}}$ be the balancing projection onto the t-eigenvectors of $M(\hat{\underline{\mathbf{\Sigma}}}_{\tilde{\underline{\mathbf{g}}}})$, where $\hat{\mathbf{r}} = (\hat{r}_1, \dots, \hat{r}_K)^\top \in \mathbb{N}^K$. We define the following effective error term in each transformed sub-domain:

$$\begin{aligned}&\breve{\tilde{\epsilon}}^{(i)}(p, n_1, \delta)^2 \\ &:= \kappa_{\text{trn}} \left(2\kappa_{\text{apx}} \textsf{\textbf{tail}}_2^{(i)\star}(p) + \frac{(368+2)B^4(\breve{\mathscr{N}}^{(i)}(p, 2B/n_1) + \log \frac{6K}{\delta})}{n_1} \right), \forall i \in [K].\end{aligned}$$

To select the rank $\hat{\mathbf{r}}$ for dimension reduction, we use parameters $\boldsymbol{\sigma}_{\text{cut}} = (\sigma_{i,\text{cut}})_{i=1}^K \in \mathbb{R}^K$ and $\mathbf{r}_{\text{cut}} = (r_{i,\text{cut}})_{i=1}^K \in \mathbb{N}^K$. Motivated by Definition F.1 in Ref. [46], we consider the following "good" event $\mathcal{E}_{\text{spec}}(\hat{\mathbf{r}}, \boldsymbol{\sigma}_{\text{cut}}, \mathbf{r}_{\text{cut}})$, under which the choices of $\boldsymbol{\sigma}_{\text{cut}}$ and $\mathbf{r}_{\text{cut}}$ are sufficient for finding suitable ranks $\hat{r}_i$ for all the i-th frequency components ($\forall i \in [K]$).

Definition 20 (Good spectral event in all the transformed sub-domains, adapted from Definition F.1 in Ref. [46])**.** *For parameters $\boldsymbol{\sigma}_{\text{cut}} = (\sigma_{i,\text{cut}})_{i=1}^K \in \mathbb{R}^K$ and $\mathbf{r}_{\text{cut}} = (r_{i,\text{cut}})_{i=1}^K \in \mathbb{N}^K$ used in Step 3 of ERM-DS, we define $\mathcal{E}_{\text{spec}}(\hat{\mathbf{r}}, \boldsymbol{\sigma}_{\text{cut}}, \mathbf{r}_{\text{cut}})$ as the event that the following inequalities hold in each frequency component $i \in [K]$:*

$$\begin{aligned}&\breve{\sigma}^\star_{\hat{r}_i} \geq \frac{3}{4}\sigma_{i,\text{cut}}, \quad \sigma^\star_{\hat{r}_i+1} \leq 3\sigma_{i,\text{cut}}, \quad \breve{\sigma}^\star_{\hat{r}_i} - \breve{\sigma}^\star_{\hat{r}_i+1} \geq \frac{\breve{\sigma}^\star_{\hat{r}_i}}{3r_{i,\text{cut}}} \\ &\textsf{\textbf{tail}}_2^{(i)\star}(\hat{r}_i) \leq \textsf{\textbf{tail}}_2^{(i)\star}(r_{i,\text{cut}}) + 9\sigma_{i,\text{cut}}^2 r_{i,\text{cut}}, \quad \textsf{\textbf{tail}}_1^{(i)\star}(\hat{r}_i)^2 \leq 18 r_{i,\text{cut}}^2 \sigma_{i,\text{cut}}^2 + 2\textsf{\textbf{tail}}_1^{(i)\star}(r_{i,\text{cut}})^2.\end{aligned}$$

The following lemma shows that under appropriate conditions, the excess risk of the reduced-rank embeddings $(\tilde{\underline{\mathbf{f}}}, \hat{\underline{\mathbf{Q}}}_{\hat{\mathbf{r}}} *_M \tilde{\underline{\mathbf{g}}})$ can be bounded in terms of the effective error $\breve{\tilde{\epsilon}}^{(i)}(p, n_1, \delta)$. The proof involves a delicate balance between the eigenvalues of the estimated and ground truth covariance operators, as well as careful control of the approximation and estimation errors.

Lemma D.14 (Guarantee for Double-Training, Step 2 & Step 3)**.** *Suppose the parameters $\boldsymbol{\sigma}_{\text{cut}} = (\sigma_{i,\text{cut}})_{i=1}^K$ and $\mathbf{r}_{\text{cut}} = (r_{i,\text{cut}})_{i=1}^K$ satisfy $\sigma_{i,\text{cut}} \in [2\sigma^\star_{i,r_{\text{cut}}}, \frac{2}{3e}\sigma^\star_{i,1}]$, for all $i \in [K]$. Assume $n_1 \geq p \geq 2$, $n_1 \geq \max_i\{B^2/\sigma_{i,\text{cut}}^2\}$, $\mu = B^2/n_1$, $n_2 \geq 722 \max_i\{r_{i,\text{cut}}^2\} n_1^9 \log(24pK/\delta)$, and $\breve{\tilde{\epsilon}}^{(i)}(p, n_1, \delta)^2 \leq \sigma_{i,\text{cut}}^2/(64 r_{i,\text{cut}}^2)$ for all $i \in [K]$. Then, with probability at least $1 - \frac{2}{3}\delta$, we have the following upper bound on the excess risk of $\underline{h}_{\text{red}}(x, y)$ to the multi-rank-$\hat{\mathbf{r}}$ ground truth $\underline{h}^\star_{\hat{\mathbf{r}}} := (\underline{\mathbf{f}}^\star_{\hat{\mathbf{r}}})^\top *_M \underline{\mathbf{g}}^\star_{\hat{\mathbf{r}}}$ on $\mathcal{D}_{1\otimes 1}$:*

$$\mathcal{R}_{[\hat{\mathbf{r}}]}(\tilde{\underline{\mathbf{f}}}, \hat{\underline{\mathbf{Q}}}_{\hat{\mathbf{r}}} *_M \tilde{\underline{\mathbf{g}}}; \mathcal{D}_{1\otimes 1}) \leq 3000 \sum_{i=1}^K r_{i,\text{cut}}^2 \breve{\tilde{\epsilon}}^{(i)}(p, n_1, \delta)^2.$$

Moreover, on this event, $\sup_{x,y} |\langle \breve{\tilde{\underline{\mathbf{f}}}}^{(i)}(x), \breve{\hat{\underline{\mathbf{Q}}}}^{(i)}_{\hat{r}_i} \breve{\tilde{\underline{\mathbf{g}}}}^{(i)}(y) \rangle| \leq \sqrt{2n_1} B^2$ and $\mathcal{E}_{\text{spec}}(\hat{\mathbf{r}}, \boldsymbol{\sigma}_{\text{cut}}, \mathbf{r}_{\text{cut}})$ holds.

The derivation of Lemma D.14 utilizes Proposition F.2 from Ref. [46] across all K subdomains in parallel. This approach extends the proposition's application to our specific multi-output regression context. As the adaptation follows directly, the detailed proof is omitted.

Analysis of Step 4: Distillation After Step 4 of ERM-DS, we obtain embedding $(\hat{\underline{\mathbf{f}}}_{\text{ds}}, \hat{\underline{\mathbf{g}}}_{\text{ds}})$ which satisfy the following lemma.

Lemma D.15. *Suppose it holds that* $\|\breve{\hat{\underline{\mathbf{Q}}}}^{(i)}_{\hat{r}_i}\|_{\text{op}} \leq \sqrt{2n_1}$, $\forall i \in [K]$. *Then, with probability at least* $1 - \delta/(3K)$*, the regularized risk in the* i*-th frequency component satisfy*

$$\breve{\mathcal{R}}^{(i)}(\breve{\hat{\underline{\mathbf{f}}}}^{(i)}_{\text{ds}}, \breve{\hat{\underline{\mathbf{g}}}}^{(i)}_{\text{ds}}; \mathcal{D}_{\text{train}}) + \frac{\breve{\nu}^{(i)}}{2} \breve{\mathcal{R}}^{(i)}_{[r_i]}(\breve{\hat{\underline{\mathbf{f}}}}^{(i)}_{\text{ds}}, \breve{\hat{\underline{\mathbf{g}}}}^{(i)}_{\text{ds}}; \mathcal{D}_{1\otimes 1})$$
$$\leq 2\kappa_{\text{apx}} \textsf{tail}_2^{(i)\star}(r_i) + 3\breve{\nu}^{(i)} \breve{\mathcal{R}}^{(i)}_{[r_i]}(\breve{\tilde{\underline{\mathbf{f}}}}^{(i)}, \breve{\hat{\underline{\mathbf{Q}}}}_{r_i} \cdot \breve{\tilde{\underline{\mathbf{g}}}}^{(i)}; \mathcal{D}_{1\otimes 1}) + 368\left(1 + \frac{2\breve{\nu}^{(i)} n_1 n_3}{n_4}\right) \frac{B^4(\mathscr{N}_{r_i, n_4} + \log(12K/\delta))}{n_3}.$$

This lemma provides upper bounds on the training and testing risks of the final embeddings $(\hat{\underline{\mathbf{f}}}_{\text{ds}}, \hat{\underline{\mathbf{g}}}_{\text{ds}})$ in terms of the tail sum $\textsf{tail}_2^{(i)\star}(r_i)$, the excess risk $\breve{\mathcal{R}}^{(i)}_{[r_i]}(\breve{\tilde{\underline{\mathbf{f}}}}^{(i)}, \breve{\hat{\underline{\mathbf{Q}}}}_{r_i} \cdot \breve{\tilde{\underline{\mathbf{g}}}}^{(i)}; \mathcal{D}_{1\otimes 1})$, and the covering number $\mathscr{N} r_i, n_4$. The proof of this lemma relies on a careful decomposition of the excess risk and the application of concentration inequalities. The proof of Lemma D.15 is provided in Appendix D.3.1.

We now provide a sufficient condition for parameter settings and sample sizes to improve the bound in Lemma D.15.

Lemma D.16. *Suppose that the parameters in ERM-DS are chosen as* $\mu = B^2/n_1$*, and other parameters* $(p, \boldsymbol{\sigma}_{\text{cut}}, \mathbf{r}_{\text{cut}})$*, the sample sizes* $n_1, \ldots, n_4$*, and parameter* $\boldsymbol{\nu} = (\breve{\nu}^{(i)})_{i=1}^K \in \mathbb{R}^K$ *satisfy that for some* $C \lesssim_\star 1$, $\forall i \in [K]$

- $\sigma_{i,\text{cut}} \in [2\breve{\sigma}^\star_{r_{i,\text{cut}}}, \frac{2}{3e}\breve{\sigma}^\star_1]$, $\textsf{tail}_2^\star(p) \leq \frac{\sigma^2_{i,\text{cut}}}{Cr^2_{i,\text{cut}}}$, *and* $p \geq 2$;
- $n_1 \geq p + C\max\{1, B^4\}\max_i\{\sigma^{-2}_{i,\text{cut}} r^2_{i,\text{cut}}(\breve{\mathscr{N}}^{(i)}(p, 2B/n_1) + \log(K/\delta))\}$,
- $n_2 \geq 722\max_i\{r^2_{i,\text{cut}}\}n_1^9\log(24pK/\delta)$,
- $n_4 \geq n_1 n_3 \max\{2\breve{\nu}^{(i)}, 3\}$.

Then, with probability at least $1 - \delta$*, the event* $\mathcal{E}_{\text{spec}}(\hat{\mathbf{r}}, \boldsymbol{\sigma}_{\text{cut}}, \mathbf{r}_{\text{cut}})$ *in Definition 20 holds and*

$$\breve{\mathcal{R}}^{(i)}(\breve{\hat{\underline{\mathbf{f}}}}^{(i)}_{\text{ds}}, \breve{\hat{\underline{\mathbf{g}}}}^{(i)}_{\text{ds}}; \mathcal{D}_{\text{train}}) + \breve{\nu}^{(i)} \breve{\mathcal{R}}^{(i)}_{[\hat{r}_i]}(\breve{\hat{\underline{\mathbf{f}}}}^{(i)}_{\text{ds}}, \breve{\hat{\underline{\mathbf{g}}}}^{(i)}_{\text{ds}}; \mathcal{D}_{1\otimes 1})$$
$$\lesssim_\star \textsf{tail}_2^{(i)\star}(r_{i,\text{cut}}) + r_{i,\text{cut}}\sigma^2_{i,\text{cut}} + \breve{\nu}^{(i)} r^2_{i,\text{cut}} \textsf{tail}_2^{(i)\star}(p)$$
$$+ \frac{B^4(\mathscr{N}(\hat{r}_i, 2B/n_4) + \log(K/\delta))}{n_3} + \frac{\breve{\nu}^{(i)} r^2_{i,\text{cut}} B^4(\mathscr{N}(p, 2B/n_1) + \log(K/\delta))}{n_1}.$$

Lemma D.16 follows from the application of Proposition F.1 in Ref. [46] to our context. The steps align closely with those in the referenced proposition, so the detailed proof is omitted here.

To ensure the successful application of the ERM-DS algorithm, we must carefully select the parameters $\breve{\nu}^{(i)}$ to control the error terms. However, before diving into this crucial step, let us first establish a refined set of sufficient conditions for the algorithm parameters and sample sizes, which will serve as the foundation for our analysis.

First, let us focus on the algorithm parameters, as outlined in Condition 7. We require that the parameters $r_{i,\text{cut}}$, $\sigma_{i,\text{cut}}$, and p satisfy a series of intricate relationships. These relationships, expressed in parts (a), (b), and (c) of the condition, ensure that the algorithm's components are well-balanced and can effectively capture the underlying structure of the data.

Condition 7 (Algorithm parameters)**.** *Let* $c_{i,1}$ *be some unspecified parameter satisfying* $1 \leq c_{i,1} \lesssim_\star 1$. *We stipulate that the algorithm parameters* $(\sigma_{i,\text{cut}}, r_{i,\text{cut}}, p)$ *satisfy* *(a)* $r_{i,\mathrm{cut}} \geq c_{i,1}$ *and* $\textsf{tail}_2^{(i)\star}(r_{i,\mathrm{cut}}) \leq \frac{1}{c_{i,1}} r_{i,\mathrm{cut}}^2 (\sigma_{i,\mathrm{cut}})^2$*;*

(b) $\textsf{tail}_2^{(i)\star}(p) \leq \frac{1}{c_{i,1}} \frac{\sigma_{i,\mathrm{cut}}^2}{r_{i,\mathrm{cut}}^5}$*;*

(c) $\sigma_{i,\mathrm{cut}} \in [2\breve{\sigma}^\star_{r_{i,\mathrm{cut}}}, \frac{2}{3e}\breve{\sigma}^\star_1]$.

Next, we turn our attention to the sample size requirements, as specified in Condition 8.

Condition 8 (Sample size conditions)**.** *Let* $c_{i,2}$ *be some unspecified parameter satisfying* $c_{i,2} \lesssim_\star 1$. *We stipulate that, given* $\delta \in (0, 1)$*,*

- *The supervised sample sizes of* n_1, n_3 *satisfy*

$$n_1 \geq p + B^4 \max_i \{c_{i,2}(\breve{\mathscr{N}}^{(i)}(p, 2B/n_1) + \log \frac{K}{\delta}) r_{i,\mathrm{cut}}^4 \sigma_{i,\mathrm{cut}}^{-2}\},$$
$$n_3 \geq \max_i \{c_{i,2} B^4 (\mathscr{N}(r_{i,\mathrm{cut}}, B/n_4) + \log \frac{K}{\delta}) \sigma_{i,\mathrm{cut}}^{-2}\}$$

- *The unsupervised sample sizes* n_2, n_4 *satisfy*

$$n_2 \geq 722 n_1^9 \log(24pK/\delta) \max_i \{r_{i,\mathrm{cut}}^2\}, \quad n_4 \geq n_1 n_3 \max_i \{r_{i,\mathrm{cut}}^4\}.$$

With these conditions in place, we can now state the main result of our analysis, as presented in Theorem 9. This theorem provides a powerful guarantee for the performance of the ERM-DS algorithm, as measured by the risk function $\mathcal{R}(\underline{\hat{\mathbf{f}}}_{\mathrm{ds}}, \underline{\hat{\mathbf{g}}}_{\mathrm{ds}}; \mathcal{D}_{\mathrm{test}})$. The theorem asserts that, with high probability, this risk can be bounded by a sum of terms that depend on the parameters $r_{i,\mathrm{cut}}$, $\sigma_{i,\mathrm{cut}}$, and the tail sums of the singular values, $\textsf{tail}_1^{(i)\star}$ and $\textsf{tail}_2^{(i)\star}$.

Theorem 9. *Suppose the parameters of ERM-DS satisfy* $\boldsymbol{\sigma}_{\mathrm{cut}}, \mathbf{r}_{\mathrm{cut}}, p$*, sample sizes* $n_1, \ldots, n_4$*, and* $\breve{\nu}^{(i)} = r_{i,\mathrm{cut}}^4$*,* $\mu = B^2/n_1$ *and fix a probability of error* $\delta \in (0, 1)$*. Then, as long as* $\sigma_{\mathrm{cut}}, r_{\mathrm{cut}}, p$ *satisfy Condition 7 and* $n_{1:4}$ *satisfy Condition 8, it holds with probability at least* $1 - \delta$*,*

$$\mathcal{R}(\underline{\hat{\mathbf{f}}}_{\mathrm{ds}}, \underline{\hat{\mathbf{g}}}_{\mathrm{ds}}; \mathcal{D}_{\mathrm{test}}) \lesssim_\star \sum_{i=1}^{K} \left(r_{i,\mathrm{cut}}^2 \sigma_{i,\mathrm{cut}}^2 + \textsf{tail}_1^{(i)\star}(r_{i,\mathrm{cut}})^2 + \frac{\textsf{tail}_2^{(i)\star}(r_{i,\mathrm{cut}})^2}{(\sigma_{i,\mathrm{cut}})^2} \right).$$

The proof of this theorem relies on a careful analysis of the error terms and their relationships to the algorithm parameters and sample sizes. By leveraging the conditions we have established, along with advanced techniques from tensor algebra and empirical process theory, we can derive the stated bound on the risk function.

Proof of Theorem 9. Our goal is to bound the generalization error of the ERM-DS algorithm in each subdomain after transformation:

$$\breve{\mathcal{R}}^{(i)}(\breve{\underline{\hat{\mathbf{f}}}}_{\mathrm{ds}}^{(i)}, \breve{\underline{\hat{\mathbf{g}}}}_{\mathrm{ds}}^{(i)}; \mathcal{D}_{\mathrm{test}}) \lesssim_\star \underbrace{r_{i,\mathrm{cut}}^2 \sigma_{i,\mathrm{cut}}^2 + \textsf{tail}_1^{(i)\star}(r_{i,\mathrm{cut}})^2 + \frac{\textsf{tail}_2^{(i)\star}(r_{i,\mathrm{cut}})^2}{(\sigma_{i,\mathrm{cut}})^2}}_{=:\mathrm{E{\scriptsize RR}}_{\mathrm{DT}}^{(i)}(r_{i,\mathrm{cut}}, \sigma_{i,\mathrm{cut}})}, \quad \forall i \in [K].$$

The key to the proof is to leverage the properties of each step in the algorithm and recursively bound the generalization error. While the overall proof strategy follows the ideas from [46], we extend these results to the multi-output setting using tensor algebra, which requires handling the additional complexity introduced by the tensor structure and the Hilbert t-Module framework.

Step 1: Based on Lemma D.22, if the learned embeddings $(\breve{\underline{\hat{\mathbf{f}}}}_{\mathrm{ds}}^{(i)}, \breve{\underline{\hat{\mathbf{g}}}}_{\mathrm{ds}}^{(i)})$ have risk $(\breve{\epsilon}_{\mathrm{trn}}^{(i)})^2$ on the training set $\mathcal{D}_{\mathrm{train}}$ and truncated risk $(\breve{\epsilon}_{\mathcal{D}_{1\otimes 1}}^{(i)})^2$ on distributio $\mathcal{D}_{1\otimes 1}$, and satisfy $\breve{\epsilon}_{\mathcal{D}_{1\otimes 1}}^{(i)} \leq \min\{\breve{\sigma}_1^{\star,(i)}/40\hat{r}_i, \breve{\sigma}_{\hat{r}_i}^{\star,(i)}/4\}$, then their generalization error on the test distribution $\mathcal{D}_{\mathrm{test}}$ can be bounded as (with $\alpha_i \le 2$) for all $i \in [K]$:

$$\begin{aligned}
&\breve{\mathcal{R}}^{(i)}(\breve{\hat{\underline{\mathbf{f}}}}_{\text{ds}}^{(i)}, \breve{\hat{\underline{\mathbf{f}}}}_{\text{ds}}^{(i)}; \mathcal{D}_{\text{test}}) \\
&\lesssim_\star \left\{\hat{r}_i^4 (\breve{\epsilon}_{\mathcal{D}_{1\otimes 1}}^{(i)})^2 + \textsf{\textbf{tail}}_1^{(i)\star}(\hat{r}_i)^2 + \hat{r}_i^2 (\breve{\sigma}_{\hat{r}_i+1}^{\star,(i)})^2\right\} + \left\{\frac{(\hat{r}_i^3 (\breve{\epsilon}_{\mathcal{D}_{1\otimes 1}}^{(i)})^2 + (\breve{\epsilon}_{\text{trn}}^{(i)})^2 + \textsf{\textbf{tail}}_2^{(i)\star}(\hat{r}_i))^2}{(\breve{\sigma}_{\hat{r}_i}^{\star,(i)})^2}\right\}.
\end{aligned}$$

Step 2: Recall that in the ERM-DS algorithm, we choose $\hat{r}_i \le r_{i,\text{cut}}$ and $\breve{\nu}^{(i)} = r_{i,\text{cut}}^4$. In this case, as long as $\breve{\epsilon}_{\mathcal{D}_{1\otimes 1}}^{(i)} \le \min\{\breve{\sigma}_1^{\star,(i)}/40 r_{i,\text{cut}}, \breve{\sigma}_{r_{i,\text{cut}}}^{\star,(i)}/4\}$, we can obtain for all $i \in [K]$:

$$\begin{aligned}
&\breve{\mathcal{R}}^{(i)}(\breve{\hat{\underline{\mathbf{f}}}}_{\text{ds}}^{(i)}, \breve{\hat{\underline{\mathbf{g}}}}_{\text{ds}}^{(i)}; \mathcal{D}_{\text{test}}) \\
&\lesssim_\star \left\{\breve{\nu}^{(i)} (\breve{\epsilon}_{\mathcal{D}_{1\otimes 1}}^{(i)})^2 + \textsf{\textbf{tail}}_1^{(i)\star}(\hat{r}_i)^2 + \hat{r}_i^2 (\breve{\sigma}_{\hat{r}_i+1}^{\star,(i)})^2\right\} + \left\{\frac{(\breve{\nu}^{(i)} (\breve{\epsilon}_{\mathcal{D}_{1\otimes 1}}^{(i)})^2 + (\breve{\epsilon}_{\text{trn}}^{(i)})^2 + \textsf{\textbf{tail}}_2^{(i)\star}(\hat{r}_i))^2}{(\breve{\sigma}_{\hat{r}_i}^{\star,(i)})^2}\right\}.
\end{aligned}$$

Step 3: Conditioned on the occurrence of the good spectral event $\mathcal{E}_{\text{spec}}(\hat{\mathbf{r}}, \boldsymbol{\sigma}_{\text{cut}}, \mathbf{r}_{\text{cut}})$ (Definition 20), we can further constrain the generalization error using the spectral properties provided by this event:

$$\begin{aligned}
\breve{\mathcal{R}}^{(i)}(\breve{\hat{\underline{\mathbf{f}}}}_{\text{ds}}^{(i)}, \breve{\hat{\underline{\mathbf{g}}}}_{\text{ds}}^{(i)}; \mathcal{D}_{\text{test}}) \lesssim_\star &\left\{\breve{\nu}^{(i)} (\breve{\epsilon}_{\mathcal{D}_{1\otimes 1}}^{(i)})^2 + r_{i,\text{cut}}^2 \sigma_{i,\text{cut}}^2 + \textsf{\textbf{tail}}_1^{(i)\star}(r_{i,\text{cut}})^2\right\} \\
&+ \left\{\frac{(\breve{\nu}^{(i)} (\breve{\epsilon}_{\mathcal{D}_{1\otimes 1}}^{(i)})^2 + (\breve{\epsilon}_{\text{trn}}^{(i)})^2 + r_{i,\text{cut}} (\sigma_{i,\text{cut}})^2 + \textsf{\textbf{tail}}_2^{(i)\star}(r_{i,\text{cut}}))^2}{(\sigma_{i,\text{cut}})^2}\right\}.
\end{aligned}$$

Step 4: Substituting the specific forms of $(\breve{\epsilon}_{\text{trn}}^{(i)})^2$ and $(\breve{\epsilon}_{\mathcal{D}_{1\otimes 1}}^{(i)})^2$ from Lemma D.16, and leveraging the high-probability event described in the same lemma, we can deduce:

$$\begin{aligned}
&\breve{\nu}^{(i)} (\breve{\epsilon}_{\mathcal{D}_{1\otimes 1}}^{(i)})^2 \le \breve{\nu}^{(i)} (\breve{\epsilon}_{\mathcal{D}_{1\otimes 1}}^{(i)})^2 + (\breve{\epsilon}_{\text{trn}}^{(i)})^2 \\
&\lesssim_\star \textsf{\textbf{tail}}_2^{(i)\star}(r_{i,\text{cut}}) + r_{i,\text{cut}} \sigma_{i,\text{cut}}^2 + r_{i,\text{cut}}^6 \textsf{\textbf{tail}}_2^{(i)\star}(p) \\
&\quad + \underbrace{\frac{B^4(\mathscr{N}(\hat{r}_i, 2B/n_4) + \log(K/\delta))}{n_3} + \frac{\breve{\nu}^{(i)} r_{i,\text{cut}}^2 B^4(\mathscr{N}(p, 2B/n_1) + \log(K/\delta))}{n_1}}_{\le 2\sigma_{i,\text{cut}}^2 \text{ under Condition 8}} \\
&\lesssim \textsf{\textbf{tail}}_2^{(i)\star}(r_{i,\text{cut}}) + r_{i,\text{cut}} \sigma_{i,\text{cut}}^2 + r_{i,\text{cut}}^6 \textsf{\textbf{tail}}_2^{(i)\star}(p).
\end{aligned}$$

This step involves balancing the algorithm parameters, sample sizes, and spectral decay rates across all sub-domains to ensure that the error terms are well-controlled.

Step 5: Substituting the above inequality into the bound from Step 3 yields the result stated in the theorem:

$$\breve{\mathcal{R}}^{(i)}(\breve{\hat{\underline{\mathbf{f}}}}_{\text{ds}}^{(i)}, \breve{\hat{\underline{\mathbf{g}}}}_{\text{ds}}^{(i)}; \mathcal{D}_{\text{test}}) \lesssim_\star r_{i,\text{cut}}^2 \sigma_{i,\text{cut}}^2 + \textsf{\textbf{tail}}_1^{(i)\star}(r_{i,\text{cut}})^2 + \frac{\textsf{\textbf{tail}}_2^{(i)\star}(r_{i,\text{cut}})^2}{(\sigma_{i,\text{cut}})^2},$$

provided that the small tail assumption $\textsf{\textbf{tail}}_2^{(i)\star}(p) \le \frac{(\sigma_{i,\text{cut}})^2}{r_i^5}$ is satisfied and that the conditions on the algorithm parameters and sample sizes in Lemma D.16 hold, which are ensured by Conditions 7 and 8. □

Lemma D.17. *Suppose Assumption 4 holds, and that the algorithm parameters $\sigma_{i,\text{cut}}, r_{i,\text{cut}}, p$ satisfy $\sigma_{i,\text{cut}} \le \frac{2}{3e}\breve{\sigma}_1^{\star,(i)}$, and the following (feasible) constraints*

$$r_{i,\text{cut}} \ge \max\{c_{i,1}, 3eC(\breve{\sigma}_1^{\star,(i)})^{-1}\}, \quad p \ge c_{i,1}^{-\frac{1}{1+2\gamma_i}} r_{i,\text{cut}}^{\frac{7+5\gamma_i}{1+2\gamma_i}}, \quad \sigma_{i,\text{cut}} \ge 2Cr_{i,\text{cut}}^{-(1+\gamma_i)}.$$

Then, Condition 7 holds and

$$\textsc{Err}_{\text{DT}}^{(i)}(r_{i,\text{cut}}, \sigma_{i,\text{cut}}) \lesssim \sigma_{i,\text{cut}}^2 r_{i,\text{cut}}^2 + C^2(1+\gamma_i^{-2}) r_{\text{cut}}^{-2\gamma_i}. \tag{32}$$

The proof closely follows that of Lemma E.2 in Ref. [46] and is therefore omitted. This lemma establishes the relationship between the algorithm parameters $(\sigma_{i,\text{cut}}, r_{i,\text{cut}}, p)$ and the error term $\text{E}{\small\text{RR}}_{\text{DT}}^{(i)}(r_{i,\text{cut}}, \sigma_{i,\text{cut}})$, which plays a central role in bounding the excess risk of the algorithm.

Proof of Theorem 6. Our goal is to achieve the desired accuracy level ϵ. To this end, we let $\sigma_{i,\text{cut}} = \max\{2Cr_{i,\text{cut}}^{-(1+\gamma_i)}, \epsilon/(K^{1/2}r_{i,\text{cut}})\}$. We then invoke Equation (32), absorb absolute constants into the $\lesssim_\star$ notation, and arrive at the following bound:

$$\mathcal{R}(\hat{\underline{\mathbf{f}}}_{\text{ds}}, \hat{\underline{\mathbf{g}}}_{\text{ds}}; \mathcal{D}_{\text{test}}) \lesssim_\star \epsilon^2 + \sum_{i=1}^{K}(1+\gamma_i^{-2})r_{i,\text{cut}}^{-2\gamma_i}. \tag{33}$$

We confirm that both Condition 7 and Condition 8 hold. Specifically, Condition 7 is satisfied as $r_{i,\text{cut}}$ is large enough (as a polynomial in $C/\breve{\sigma}_1^{\star,(i)}$) and p remains within a constant power of $r_{i,\text{cut}}$. Suitable sample sizes n_1, n_2, n_3, n_4 ensure Condition 8, leading to the result by Theorem 9. □

D.3.1 Lemmas From Statistical Learning Theory

Lemma D.18 (Bernstein inequality [10])**.** *Let $Z_1, \dots, Z_n \in \mathbb{R}$ be i.i.d. random variables with $|Z_i| \le M$ and $\mathrm{Var}[Z_i] \le \sigma^2$. Then, with probability at least $1-\delta$,*

$$\left|\frac{1}{n}\sum_{i=1}^{n} Z_i - \mathbb{E}[Z_i]\right| \le \sqrt{\frac{2\sigma^2\log(1/\delta)}{n}} + \frac{M\log(1/\delta)}{3n}.$$

Lemma D.19. *Let Φ be a function class with functions $\phi : \mathcal{W} \to \mathbb{R}^k$, and let $\phi^\star(w)$ be a target function defined on $\mathcal{W}$. Assume there exists a constant $U > 0$ such that $\sup_{w\in\mathcal{W}}\sup_{\phi\in\Phi}\|\phi(w) - \phi^\star(w)\|_2 \le U$, and let $\mathcal{N}(\Phi, \epsilon, \|\cdot\|_\infty)$ be the ϵ-covering number of Φ. Let D be a joint distribution on $\mathcal{W}\times\mathbb{R}^k$ such that $\|z - \phi^\star(w)\|_2 \le U$ and $\mathbb{E}[z|w] = \phi^\star(w)$. Define $R(\phi) := \mathbb{E}_{w\sim D}[\|\phi(w) - \phi^\star(w)\|_2^2]$, $\hat{L}_n(\phi) := \frac{1}{n}\sum_{i=1}^{n}\|\phi(w_i) - z_i\|_2^2$, and $\hat{R}_n(\phi) := \hat{L}_n(\phi) - \hat{L}_n(\phi^\star)$. Then for any $\delta \in (0,1)$, with probability at least $1-\delta$,*

- *any $\phi \in \Phi$ satisfy:*

$$R(\phi) \le \frac{a}{2}R(\phi) + \frac{(9/a+5)U^2\log(4\mathcal{N}(\Phi, U/n, \|\cdot\|_\infty)/\delta)}{n}, \quad \forall a > 0. \tag{34}$$

- *all empirical risk minimizers $\hat{\phi} \in \arg\min_{\phi\in\Phi}\hat{L}_n(\phi)$ satisfy:*

$$R(\hat{\phi}) \le 2\inf_{\phi\in\Phi}\mathbb{E}_D[(\phi(w) - \phi^\star(w))^2] + \frac{92U^2}{n}\left(\log 2\mathcal{N}(\Phi, U/n, \|\cdot\|_\infty) + \log\frac{2}{\delta}\right). \tag{35}$$

Proof of Lemma D.19. For each $\phi \in \Phi$, first define the loss function $\ell_\phi : \mathcal{W}\times\mathbb{R}^k \to \mathbb{R}$ as

$$\ell_\phi(w, z) := \|\phi(w) - z\|_2^2.$$

Also, define the excess loss function $\Delta\ell_\phi : \mathcal{W}\times\mathbb{R}^k \to \mathbb{R}$ as

$$\Delta\ell_\phi(w,z) := \ell_\phi(w,z) - \ell_{\phi^\star}(w,z) = \|\phi(w) - z\|_2^2 - \|\phi^\star(w) - z\|_2^2.$$

We then bound the excess loss function as follows. For any $\phi \in \Phi$ and $(w,z) \in \mathcal{W}\times\mathbb{R}^k$, we have:

$$\begin{aligned}
|\Delta\ell_\phi(w,z)| &= \left|\|\phi(w) - z\|_2^2 - \|\phi^\star(w) - z\|_2^2\right| \\
&= \left|\|\phi(w) - \phi^\star(w) + \phi^\star(w) - z\|_2^2 - \|\phi^\star(w) - z\|_2^2\right| \\
&= \left|\|\phi(w) - \phi^\star(w)\|_2^2 + 2\langle\phi(w) - \phi^\star(w), \phi^\star(w) - z\rangle\right| \\
&\le \|\phi(w) - \phi^\star(w)\|_2^2 + 2\|\phi(w) - \phi^\star(w)\|_2\|\phi^\star(w) - z\|_2 \\
&\le U^2 + 2U^2 = 3U^2.
\end{aligned}$$

Here, the first inequality follows from the Cauchy-Schwarz inequality, and the second inequality uses the assumptions that $\sup_{w\in\mathcal{W}}\sup_{\phi\in\Phi}\|\phi(w)-\phi^\star(w)\|_2 \le U$ and $\|z-\phi^\star(w)\|_2 \le U$.

Next, we construct an ϵ-cover of Φ. Let $\epsilon > 0$ be fixed. By the definition of covering numbers, there exists an ϵ-cover Φ_ϵ of Φ under the $\|\cdot\|_\infty$ norm, such that $|\Phi_\epsilon| = \mathcal{N}(\Phi, \epsilon, \|\cdot\|_\infty)$. This means that for every $\phi \in \Phi$, there exists a $\phi_\epsilon \in \Phi_\epsilon$ such that $\|\phi - \phi_\epsilon\|_\infty \le \epsilon$. We further bound the difference between the true and empirical excess risks for functions in the cover. For each $\phi_\epsilon \in \Phi_\epsilon$, define $Z_i(\phi_\epsilon) := \Delta\ell_{\phi_\epsilon}(w_i, z_i)$. Then we have: (a) $\mathbb{E}[Z_i(\phi_\epsilon)] = \mathbb{E}_{(w,z)\sim D}[\Delta\ell_{\phi_\epsilon}(w,z)] =: R(\phi_\epsilon)$, (b) $|Z_i(\phi_\epsilon)| \le 3U^2$ almost surely, and (c) $\mathbb{E}[Z_i(\phi_\epsilon)^2] \le 9U^2 R(\phi_\epsilon)$.

By Bernstein's inequality and a union bound over Φ_ϵ, we have with probability at least $1-\delta/2$, for all $\phi_\epsilon \in \Phi_\epsilon$:

$$\begin{aligned}\left|R(\phi_\epsilon) - \frac{1}{n}\sum_{i=1}^n Z_i(\phi_\epsilon)\right| &\le \sqrt{\frac{2\cdot 9U^2 R(\phi_\epsilon)\log(4|\Phi_\epsilon|/\delta)}{n}} + \frac{U^2\log(4|\Phi_\epsilon|/\delta)}{n} \\ &= \sqrt{\frac{18U^2 R(\phi_\epsilon)\log(4\mathcal{N}(\Phi,\epsilon,\|\cdot\|_\infty)/\delta)}{n}} + \frac{U^2\log(4\mathcal{N}(\Phi,\epsilon,\|\cdot\|_\infty)/\delta)}{n}.\end{aligned}$$

We then extend the bound to all functions in Φ. For any $\phi \in \Phi$, let $\phi_\epsilon \in \Phi_\epsilon$ be such that $\|\phi - \phi_\epsilon\|_\infty \le \epsilon$. Then, we have:

$$\begin{aligned}&\left|R(\phi) - \frac{1}{n}\sum_{i=1}^n \Delta\ell_\phi(w_i, z_i)\right| \\ &\le |R(\phi) - R(\phi_\epsilon)| + \left|R(\phi_\epsilon) - \frac{1}{n}\sum_{i=1}^n Z_i(\phi_\epsilon)\right| + \left|\frac{1}{n}\sum_{i=1}^n Z_i(\phi_\epsilon) - \frac{1}{n}\sum_{i=1}^n \Delta\ell_\phi(w_i, z_i)\right| \\ &\le 2U\epsilon + \sqrt{\frac{18U^2 R(\phi_\epsilon)\log(4\mathcal{N}(\Phi,\epsilon,\|\cdot\|_\infty)/\delta)}{n}} + \frac{U^2\log(4\mathcal{N}(\Phi,\epsilon,\|\cdot\|_\infty)/\delta)}{n} + 2U\epsilon \\ &\le \sqrt{\frac{18U^2 R(\phi)\log(4\mathcal{N}(\Phi,\epsilon,\|\cdot\|_\infty)/\delta)}{n}} + \frac{U^2\log(4\mathcal{N}(\Phi,\epsilon,\|\cdot\|_\infty)/\delta)}{n} + 4U\epsilon.\end{aligned}$$

We will choose ϵ and simplify the bound. Set $\epsilon = U/n$. Then, with probability at least $1-\delta/2$, for all $\phi \in \Phi$:

$$\begin{aligned}&\left|R(\phi) - \frac{1}{n}\sum_{i=1}^n \Delta\ell_\phi(w_i, z_i)\right| \\ &\le \sqrt{\frac{18U^2 R(\phi)\log(4\mathcal{N}(\Phi,U/n,\|\cdot\|_\infty)/\delta)}{n}} + \frac{U^2\log(4\mathcal{N}(\Phi,U/n,\|\cdot\|_\infty)/\delta)}{n} + \frac{4U^2}{n} \\ &\le \sqrt{\frac{18U^2 R(\phi)\log(4\mathcal{N}(\Phi,U/n,\|\cdot\|_\infty)/\delta)}{n}} + \frac{5U^2\log(4\mathcal{N}(\Phi,U/n,\|\cdot\|_\infty)/\delta)}{n} \\ &\le \frac{a}{2}R(\phi) + \frac{(9/a+5)U^2\log(4\mathcal{N}(\Phi,U/n,\|\cdot\|_\infty)/\delta)}{n}\end{aligned}$$

for any $a > 0$.

Next, we can bound the risk of the ERM solution. Let $\tilde{\phi} \in \arg\min_{\phi\in\Phi} R(\phi)$ be a minimizer of the true risk. Then, on the event of the previous step, we have:

$$\begin{aligned}R(\hat{\phi}) - R(\tilde{\phi}) &= R(\hat{\phi}) - \hat{R}_n(\hat{\phi}) + \hat{R}_n(\hat{\phi}) - \hat{R}_n(\tilde{\phi}) + \hat{R}_n(\tilde{\phi}) - R(\tilde{\phi}) \\ &\le (R(\hat{\phi}) - \hat{R}_n(\hat{\phi})) + (\hat{R}_n(\tilde{\phi}) - R(\tilde{\phi})) \\ &\le \frac{a}{2}(R(\tilde{\phi}) + R(\hat{\phi})) + 2\cdot\frac{(9/a+5)U^2\log(4\mathcal{N}(\Phi,U/n,\|\cdot\|_\infty)/\delta)}{n} \\ &\le aR(\hat{\phi}) + 2\cdot\frac{(9/a+5)U^2\log(4\mathcal{N}(\Phi,U/n,\|\cdot\|_\infty)/\delta)}{n}.\end{aligned}$$

Here, $\hat{R}_n(\phi) := \frac{1}{n}\sum_{i=1}^n \Delta\ell_\phi(w_i, z_i)$ is the empirical excess risk, and the first inequality uses the fact that $\hat{R}_n(\hat{\phi}) - \hat{R}_n(\tilde{\phi}) \leq 0$ since $\hat{\phi}$ minimizes the empirical risk. Choosing $a = 1/2$, we get:

$$R(\hat{\phi}) - R(\tilde{\phi}) \leq \frac{1}{2}R(\hat{\phi}) + \frac{46U^2 \log(4\mathcal{N}(\Phi, U/n, \|\cdot\|_\infty)/\delta)}{n}.$$

Rearranging the terms, we obtain:

$$\begin{aligned} R(\hat{\phi}) &\leq 2R(\tilde{\phi}) + \frac{92U^2 \log(4\mathcal{N}(\Phi, M/n, \|\cdot\|_\infty)/\delta)}{n} \\ &\leq 2\inf_{\phi\in\Phi} \mathbb{E}_D[(\phi(w) - \phi^\star(w))^2] + \frac{92U^2 \log(4\mathcal{N}(\Phi, M/n, \|\cdot\|_\infty)/\delta)}{n}. \end{aligned}$$

□

Lemma D.20. *Let $\mathcal{F}$ and $\mathcal{G}$ be two classes of embedding functions, where $\underline{\mathbf{f}} \in \mathcal{F}$ maps from an arbitrary domain $\mathcal{X}$ to a Hilbert t-Module $\mathcal{M}$ and $\underline{\mathbf{g}} \in \mathcal{G}$ maps from an arbitrary domain $\mathcal{Y}$ to $\mathcal{M}$. Let $\mathcal{H}$ be a class of functions of the form $\underline{h}(x, y) = \langle \underline{\mathbf{f}}(x), \underline{\mathbf{g}}(y)\rangle_\mathcal{M}$, where $\underline{\mathbf{f}} \in \mathcal{F}$ and $\underline{\mathbf{g}} \in \mathcal{G}$. Let $\mathcal{N}_\mathcal{F}(\epsilon)$ denote the ϵ-covering number of $\mathcal{F}$, $\mathcal{N}_\mathcal{G}(\epsilon)$ denote the ϵ-covering number of $\mathcal{G}$, and $\mathcal{N}_\mathcal{H}(\epsilon)$ denote the ϵ-covering number of $\mathcal{H}$. If there exists a constant B such that for all $\underline{\mathbf{f}} \in \mathcal{F}$, $\underline{\mathbf{g}} \in \mathcal{G}$, $x \in \mathcal{X}$, and $y \in \mathcal{Y}$, we have $\|\underline{\mathbf{f}}(x)\|_\mathcal{M} \leq B$ and $\|\underline{\mathbf{g}}(y)\|_\mathcal{M} \leq B$, then for any $\epsilon > 0$, the following inequality holds:*

$$\mathcal{N}_\mathcal{H}(\epsilon) \leq \mathcal{N}_\mathcal{F}\left(\frac{\epsilon}{2B}\right) \cdot \mathcal{N}_\mathcal{G}\left(\frac{\epsilon}{2B}\right).$$

Proof of Lemma D.20. Let $\{\underline{\mathbf{f}}_1, \ldots, \underline{\mathbf{f}}_{N_\mathcal{F}}\}$ be an $\frac{\epsilon}{2B}$-cover of $\mathcal{F}$ and $\{\underline{\mathbf{g}}_1, \ldots, \underline{\mathbf{g}}_{N_\mathcal{G}}\}$ be an $\frac{\epsilon}{2B}$-cover of $\mathcal{G}$, where $N_\mathcal{F} = \mathcal{N}_\mathcal{F}\left(\frac{\epsilon}{2B}\right)$ and $N_\mathcal{G} = \mathcal{N}_\mathcal{G}\left(\frac{\epsilon}{2B}\right)$.

For any $\underline{h} \in \mathcal{H}$, there exist $\underline{\mathbf{f}} \in \mathcal{F}$ and $\underline{\mathbf{g}} \in \mathcal{G}$ such that $\underline{h}(x, y) = \langle \underline{\mathbf{f}}(x), \underline{\mathbf{g}}(y)\rangle_\mathcal{M}$. By the definition of the covering sets, there exist $\underline{\mathbf{f}}_i$ and $\underline{\mathbf{g}}_j$ such that for all $x \in \mathcal{X}$ and $y \in \mathcal{Y}$, we have $\|\underline{\mathbf{f}}(x) - \underline{\mathbf{f}}_i(x)\|_\mathcal{M} \leq \frac{\epsilon}{2B}$ and $\|\underline{\mathbf{g}}(y) - \underline{\mathbf{g}}_j(y)\|_\mathcal{M} \leq \frac{\epsilon}{2B}$.

Now, consider the function $\underline{h}_{ij}(x, y) = \langle \underline{\mathbf{f}}_i(x), \underline{\mathbf{g}}_j(y)\rangle_\mathcal{M}$. We have:

$$\begin{aligned} \|\underline{h}(x, y) - \underline{h}_{ij}(x, y)\| &= \|\langle \underline{\mathbf{f}}(x), \underline{\mathbf{g}}(y)\rangle_\mathcal{M} - \langle \underline{\mathbf{f}}_i(x), \underline{\mathbf{g}}_j(y)\rangle_\mathcal{M}\| \\ &= \|\langle \underline{\mathbf{f}}(x) - \underline{\mathbf{f}}_i(x), \underline{\mathbf{g}}(y)\rangle_\mathcal{M} + \langle \underline{\mathbf{f}}_i(x), \underline{\mathbf{g}}(y) - \underline{\mathbf{g}}_j(y)\rangle_\mathcal{M}\| \\ &\leq \|\langle \underline{\mathbf{f}}(x) - \underline{\mathbf{f}}_i(x), \underline{\mathbf{g}}(y)\rangle_\mathcal{M}\| + \|\langle \underline{\mathbf{f}}_i(x), \underline{\mathbf{g}}(y) - \underline{\mathbf{g}}_j(y)\rangle_\mathcal{M}\| \\ &\leq \|\underline{\mathbf{f}}(x) - \underline{\mathbf{f}}_i(x)\|_\mathcal{M}\|\underline{\mathbf{g}}(y)\|_\mathcal{M} + \|\underline{\mathbf{f}}_i(x)\|_\mathcal{M} \cdot \|\underline{\mathbf{g}}(y) - \underline{\mathbf{g}}_j(y)\|_\mathcal{M} \\ &\leq \frac{\epsilon}{2B} \cdot B + B \cdot \frac{\epsilon}{2B} = \epsilon. \end{aligned}$$

This shows that the set $\{\underline{h}_{ij} : 1 \leq i \leq N_\mathcal{F}, 1 \leq j \leq N_\mathcal{G}\}$ forms an ϵ-cover of $\mathcal{H}$. The cardinality of this set is $N_\mathcal{F} \cdot N_\mathcal{G} = \mathcal{N}_\mathcal{F}\left(\frac{\epsilon}{2B}\right) \cdot \mathcal{N}_\mathcal{G}\left(\frac{\epsilon}{2B}\right)$, which implies that $\mathcal{N}_\mathcal{H}(\epsilon) \leq \mathcal{N}_\mathcal{F}\left(\frac{\epsilon}{2B}\right) \cdot \mathcal{N}_\mathcal{G}\left(\frac{\epsilon}{2B}\right)$. □

Lemma D.21. *Let $\mathcal{F}$ and $\mathcal{G}$ be two classes of vector-valued functions, where $f \in \mathcal{F}$ maps from an arbitrary domain $\mathcal{X}$ to $\mathbb{R}^r$ and $g \in \mathcal{G}$ maps from an arbitrary domain $\mathcal{Y}$ to $\mathbb{R}^r$. Let $\mathcal{H}$ be a class of bivariate functions of the form $h(x, y) = \langle f(x), g(y)\rangle$, where $f \in \mathcal{F}$ and $g \in \mathcal{G}$. Let $\mathcal{N}_\mathcal{F}(\epsilon)$ denote the ϵ-covering number of $\mathcal{F}$, $\mathcal{N}_\mathcal{G}(\epsilon)$ denote the ϵ-covering number of $\mathcal{G}$, and $\mathcal{N}_\mathcal{H}(\epsilon)$ denote the ϵ-covering number of $\mathcal{H}$. If there exists a constant B such that for all $f \in \mathcal{F}$, $g \in \mathcal{G}$, $x \in \mathcal{X}$, and $y \in \mathcal{Y}$, we have $\|f(x)\| \leq B$ and $\|g(y)\| \leq B$, then for any $\epsilon > 0$, the following inequality holds:*

$$\mathcal{N}_\mathcal{H}(\epsilon) \leq \mathcal{N}_\mathcal{F}\left(\frac{\epsilon}{2B}\right) \cdot \mathcal{N}_\mathcal{G}\left(\frac{\epsilon}{2B}\right).$$

Proof of Lemma D.21. Let $\{f_1, \ldots, f_{N_\mathcal{F}}\}$ be an $\frac{\epsilon}{2B}$-cover of $\mathcal{F}$ and $\{g_1, \ldots, g_{N_\mathcal{G}}\}$ be an $\frac{\epsilon}{2B}$-cover of $\mathcal{G}$, where $N_\mathcal{F} = \mathcal{N}_\mathcal{F}\left(\frac{\epsilon}{2B}\right)$ and $N_\mathcal{G} = \mathcal{N}_\mathcal{G}\left(\frac{\epsilon}{2B}\right)$.

For any $h \in \mathcal{H}$, there exist $f \in \mathcal{F}$ and $g \in \mathcal{G}$ such that $h(x, y) = \langle f(x), g(y) \rangle$. By the definition of the covering sets, there exist f_i and g_j such that for all $x \in \mathcal{X}$ and $y \in \mathcal{Y}$, we have $\|f(x) - f_i(x)\| \leq \frac{\epsilon}{2B}$ and $\|g(y) - g_j(y)\| \leq \frac{\epsilon}{2B}$.

Now, consider the function $h_{ij}(x, y) = \langle f_i(x), g_j(y) \rangle$. We have:

$$\begin{aligned}
|h(x, y) - h_{ij}(x, y)| &= |\langle f(x), g(y) \rangle - \langle f_i(x), g_j(y) \rangle| \\
&= |\langle f(x) - f_i(x), g(y) \rangle + \langle f_i(x), g(y) - g_j(y) \rangle| \\
&\leq |\langle f(x) - f_i(x), g(y) \rangle| + |\langle f_i(x), g(y) - g_j(y) \rangle| \\
&\leq \|f(x) - f_i(x)\| \cdot \|g(y)\| + \|f_i(x)\| \cdot \|g(y) - g_j(y)\| \\
&\leq \frac{\epsilon}{2B} \cdot B + B \cdot \frac{\epsilon}{2B} = \epsilon.
\end{aligned}$$

This shows that the set $\{h_{ij} : 1 \leq i \leq N_{\mathcal{F}}, 1 \leq j \leq N_{\mathcal{G}}\}$ forms an ϵ-cover of $\mathcal{H}$. The cardinality of this set is $N_{\mathcal{F}} \cdot N_{\mathcal{G}} = \mathcal{N}_{\mathcal{F}}\left(\frac{\epsilon}{2B}\right) \cdot \mathcal{N}_{\mathcal{G}}\left(\frac{\epsilon}{2B}\right)$, which implies that $\mathcal{N}_{\mathcal{H}}(\epsilon) \leq \mathcal{N}_{\mathcal{F}}\left(\frac{\epsilon}{2B}\right) \cdot \mathcal{N}_{\mathcal{G}}\left(\frac{\epsilon}{2B}\right)$. □

Proof of Lemma D.15. Let $\Phi := \{(x, y) \mapsto \langle f(x), g(y) \rangle, (f(x), g(y)) \in \breve{\mathcal{F}}_{r_i}^{(i)} \times \breve{\mathcal{G}}_{r_i}^{(i)}\}$ be a class of bivariate functions for each $i \in [K]$. We define the following target functions:

$$\begin{aligned}
\breve{\phi}_{3,\star}^{(i)}(x, y) &:= \langle \breve{\underline{\mathbf{f}}}^{\star,(i)}(x), \breve{\underline{\mathbf{g}}}^{\star,(i)}(y) \rangle, \\
\breve{\phi}_{4,\star}^{(i)}(x, y) &:= \langle \breve{\tilde{\underline{\mathbf{f}}}}^{(i)}(x), \breve{\hat{\mathbf{Q}}}_{r_i}^{(i)} \breve{\tilde{\underline{\mathbf{g}}}}^{(i)}(y) \rangle,
\end{aligned}$$

for all $i \in [K]$. Let $\mathcal{D}_3$ be the distribution of $(x, y, z) \sim \mathcal{D}_{\text{train}}$, and $\mathcal{D}_4$ be the distribution of (x', y', z'), where $(x', y') \sim \mathcal{D}_{1\otimes 1}$ and $z' = \breve{\phi}_{4,\star}^{(i)}(x', y')$. We compute the following bounds:

$$\begin{aligned}
\sup_{x,y} \max_{\phi \in \Phi} \|\phi(x, y) - \breve{\phi}_{3,\star}^{(i)}(x, y)\| &\leq 2B^2, \\
\sup_{x,y} \max_{\phi \in \Phi} \|\phi(x, y) - \breve{\phi}_{4,\star}^{(i)}(x, y)\| &\leq (1 + \sqrt{2n_1})B^2.
\end{aligned}$$

According to Lemma D.21, the covering number of Φ satisfies

$$\mathcal{N}(\Phi, U/n, \|\cdot\|_\infty) \leq \mathcal{N}(\breve{\mathcal{F}}_{r_i}^{(i)}, U/(2Bn), \|\cdot\|_\infty) \mathcal{N}(\breve{\mathcal{G}}_{r_i}^{(i)}, U/(2Bn), \|\cdot\|_\infty).$$

We define the following covering numbers:

$$\begin{aligned}
\mathscr{N}_{r_i,n_3} &:= \mathcal{N}(\breve{\mathcal{F}}_{r_i}^{(i)}, B/n_3, \|\cdot\|_\infty) \cdot \mathcal{N}(\breve{\mathcal{G}}_{r_i}^{(i)}, B/n_3, \|\cdot\|_\infty), \\
\mathscr{N}_{r_i,n_4} &:= \mathcal{N}(\breve{\mathcal{F}}_{r_i}^{(i)}, (1 + \sqrt{2n_1})B/n_4, \|\cdot\|_\infty) \cdot \mathcal{N}(\breve{\mathcal{G}}_{r_i}^{(i)}, (1 + \sqrt{2n_1})B/n_4, \|\cdot\|_\infty).
\end{aligned}$$

For $k \in \{3, 4\}$, let R_k and $\hat{L}_{k,n_k}, \hat{R}_{k,n_k}$ denote the corresponding population and empirical excess risks as in Lemma D.19. By applying Lemma D.19 with $\alpha = 1/2$ and a union bound over all $i \in [K]$, we have with probability at least $1 - \delta/(3K)$, for all $\phi \in \Phi$:

$$\begin{aligned}
|R_3(\phi) - \hat{R}_{3,n_3}(\phi)| &\leq \frac{1}{4} R_3(\phi) + (9 \cdot 2 + 5) \frac{(2B^2)^2 (\mathscr{N}_{r_i,n_3} + \log(12K/\delta))}{n_3}, \\
|R_4(\phi) - \hat{R}_{4,n_4}(\phi)| &\leq \frac{1}{4} R_4(\phi) + (9 \cdot 2 + 5) \frac{((1 + \sqrt{2n_1})B^2)^2 (\mathscr{N}_{r_i,n_4} + \log(12K/\delta))}{n_4}.
\end{aligned}$$

Now, define the combined population risk $R_{\breve{\nu}^{(i)}}(\phi) := R_3(\phi) + \breve{\nu}^{(i)} R_4(\phi)$. Let

$$\begin{aligned}
\hat{\phi} &\in \operatorname*{argmin}_{\phi \in \Phi} \hat{L}_{3,n_3}(\phi) + \breve{\nu}^{(i)} \hat{L}_{4,n_4}(\phi) \\
&= \operatorname*{argmin}_{\phi \in \Phi} \hat{R}_{3,n_3}(\phi) + \breve{\nu}^{(i)} \hat{R}_{4,n_4}(\phi)
\end{aligned}$$

be an empirical risk minimizer, and let $\tilde{\phi} \in \operatorname{argmin}_{\phi \in \Phi} R_{\breve{\nu}^{(i)}}(\phi)$ be a population risk minimizer. Then, we have

$$\begin{aligned}
&R_{\breve{\nu}^{(i)}}(\hat{\phi}) - R_{\breve{\nu}^{(i)}}(\tilde{\phi}) \\
&\le \frac{1}{4}(R_{\breve{\nu}^{(i)}}(\hat{\phi}) + R_{\breve{\nu}^{(i)}}(\tilde{\phi})) + 2(9 \cdot 2 + 5)\frac{4B^4(\mathscr{N}_{r_i,n_3} + \log(12K/\delta))}{n_3} \\
&\quad + 2\breve{\nu}^{(i)}(9 \cdot 2 + 5)\frac{(2 + 4n_1)B^4(\mathscr{N}_{r_i,n_4} + \log(12K/\delta))}{n_4} \\
&\le \frac{1}{2}R_{\breve{\nu}^{(i)}}(\hat{\phi}) + 184\left(1 + \frac{2\breve{\nu}^{(i)}n_1 n_3}{n_4}\right)\frac{B^4(\mathscr{N}_{r_i,n_4} + \log(12K/\delta))}{n_3},
\end{aligned}$$

where the last inequality holds because $n_4 \ge n_1 n_3 \max\{2\breve{\nu}^{(i)}, 3\},\ \forall i \in [K]$ implies $\mathscr{N}_{r_i,n_4} \ge \mathscr{N}_{r_i,n_3}$.

Rearranging terms, we get

$$R_{\breve{\nu}^{(i)}}(\hat{\phi}) \le 2R_{\breve{\nu}^{(i)}}(\tilde{\phi}) + 368\left(1 + \frac{2\breve{\nu}^{(i)}n_1 n_3}{n_4}\right)\frac{B^4(\mathscr{N}_{r_i,n_4} + \log(12K/\delta))}{n_3}.$$

To conclude the proof, we bound the terms $R_{\breve{\nu}^{(i)}}(\hat{\phi})$ and $R_{\breve{\nu}^{(i)}}(\tilde{\phi})$. First, for the population risk minimizer $\tilde{\phi}$, we have

$$\begin{aligned}
R_{\breve{\nu}^{(i)}}(\tilde{\phi}) &= \inf_{\phi \in \Phi} R_{\breve{\nu}^{(i)}}(\phi) \\
&= \inf_{(f,g) \in \breve{\mathscr{F}}^{(i)}_{r_i} \times \breve{\mathscr{G}}^{(i)}_{r_i}} \breve{\mathcal{R}}^{(i)}(f, g; \mathcal{D}_{\mathrm{train}}) + \breve{\nu}^{(i)}\mathbb{E}_{\mathcal{D}_{1\otimes 1}}[(\langle f, g\rangle - \langle \breve{\tilde{\underline{\mathbf{f}}}}^{(i)}, \breve{\hat{\mathbf{Q}}}^{(i)} \cdot \breve{\tilde{\underline{\mathbf{g}}}}^{(i)}\rangle)^2] \\
&\le \breve{\mathcal{R}}^{(i)}(\breve{\underline{\mathbf{f}}}^{\star,(i)}_r, \breve{\underline{\mathbf{g}}}^{\star,(i)}_r; \mathcal{D}_{\mathrm{train}}) + \breve{\nu}^{(i)}\mathbb{E}_{\mathcal{D}_{1\otimes 1}}[(\langle \breve{\underline{\mathbf{f}}}^{\star,(i)}_r, \breve{\underline{\mathbf{g}}}^{\star,(i)}_r\rangle - \langle \breve{\tilde{\underline{\mathbf{f}}}}^{(i)}, \breve{\hat{\mathbf{Q}}}^{(i)} \cdot \breve{\tilde{\underline{\mathbf{g}}}}^{(i)}\rangle)^2] \\
&\le \kappa_{\mathrm{apx}}\breve{\mathcal{R}}^{(i)}(\breve{\underline{\mathbf{f}}}^{\star,(i)}_{r_i}, \breve{\underline{\mathbf{g}}}^{\star,(i)}_{r_i}; \mathcal{D}_{1\otimes 1}) + \breve{\nu}^{(i)}\mathbb{E}_{\mathcal{D}_{1\otimes 1}}[(\langle \breve{\underline{\mathbf{f}}}^{\star,(i)}_{r_i}, \breve{\underline{\mathbf{g}}}^{\star,(i)}_{r_i}\rangle - \langle \breve{\tilde{\underline{\mathbf{f}}}}^{(i)}, \breve{\hat{\mathbf{Q}}}^{(i)} \cdot \breve{\tilde{\underline{\mathbf{g}}}}^{(i)}\rangle)^2] \\
&= \kappa_{\mathrm{apx}}\mathsf{\mathbf{tail}}^{(i)\star}_2(r_i) + \breve{\nu}^{(i)}\breve{\mathcal{R}}^{(i)}_{[r_i]}(\breve{\tilde{\underline{\mathbf{f}}}}^{(i)}, \breve{\hat{\mathbf{Q}}}^{(i)}_{r_i} \cdot \breve{\tilde{\underline{\mathbf{g}}}}^{(i)}; \mathcal{D}_{1\otimes 1}).
\end{aligned}$$

Second, for the empirical risk minimizer $\hat{\phi}$, we have

$$\begin{aligned}
&\breve{\mathcal{R}}^{(i)}(\breve{\hat{\underline{\mathbf{f}}}}^{(i)}, \breve{\hat{\underline{\mathbf{g}}}}^{(i)}; \mathcal{D}_{\mathrm{train}}) + \frac{\breve{\nu}^{(i)}}{2}\breve{\mathcal{R}}^{(i)}_{[r_i]}(\breve{\hat{\underline{\mathbf{f}}}}^{(i)}, \breve{\hat{\underline{\mathbf{g}}}}^{(i)}; \mathcal{D}_{1\otimes 1}) \\
&= \breve{\mathcal{R}}^{(i)}(\hat{\underline{\mathbf{f}}}, \hat{\underline{\mathbf{g}}}; \mathcal{D}_{\mathrm{train}}) + \frac{\breve{\nu}^{(i)}}{2}\mathbb{E}_{\mathcal{D}_{1\otimes 1}}[(\langle \breve{\hat{\underline{\mathbf{f}}}}^{(i)}, \breve{\hat{\underline{\mathbf{g}}}}^{(i)}\rangle - \langle \breve{\underline{\mathbf{f}}}^{\star,(i)}_{r_i}, \breve{\underline{\mathbf{g}}}^{\star,(i)}_{r_i}\rangle)^2] \\
&\le \breve{\mathcal{R}}^{(i)}(\breve{\hat{\underline{\mathbf{f}}}}^{(i)}, \breve{\hat{\underline{\mathbf{g}}}}^{(i)}; \mathcal{D}_{\mathrm{train}}) + \breve{\nu}^{(i)}\mathbb{E}_{\mathcal{D}_{1\otimes 1}}[(\langle \breve{\tilde{\underline{\mathbf{f}}}}^{(i)}, \breve{\hat{\mathbf{Q}}}^{(i)}_{r_i} \cdot \breve{\tilde{\underline{\mathbf{g}}}}^{(i)}\rangle - \langle \breve{\hat{\underline{\mathbf{f}}}}^{(i)}, \breve{\hat{\underline{\mathbf{g}}}}^{(i)}\rangle)^2] \\
&\quad + \breve{\nu}^{(i)}\mathbb{E}_{\mathcal{D}_{1\otimes 1}}[(\langle \breve{\tilde{\underline{\mathbf{f}}}}^{(i)}, \breve{\hat{\mathbf{Q}}}_{r_i} \cdot \breve{\tilde{\underline{\mathbf{g}}}}^{(i)}\rangle - \langle \breve{\underline{\mathbf{f}}}^{\star,(i)}_{r_i}, \breve{\underline{\mathbf{g}}}^{\star,(i)}_{r_i}\rangle)^2] \\
&= R_3(\hat{\phi}) + \breve{\nu}^{(i)}R_4(\hat{\phi}) + \breve{\nu}^{(i)}\breve{\mathcal{R}}^{(i)}_{[r_i]}(\breve{\tilde{\underline{\mathbf{f}}}}^{(i)}, \breve{\hat{\mathbf{Q}}}_{r_i} \cdot \breve{\tilde{\underline{\mathbf{g}}}}^{(i)}; \mathcal{D}_{1\otimes 1}) \\
&= R_{\breve{\nu}^{(i)}}(\hat{\phi}) + \breve{\nu}^{(i)}\breve{\mathcal{R}}^{(i)}_{[r_i]}(\breve{\tilde{\underline{\mathbf{f}}}}^{(i)}, \breve{\hat{\mathbf{Q}}}_{r_i} \cdot \breve{\tilde{\underline{\mathbf{g}}}}^{(i)}; \mathcal{D}_{1\otimes 1}).
\end{aligned}$$

Combining the bounds for $R_{\breve{\nu}^{(i)}}(\hat{\phi})$ and $R_{\breve{\nu}^{(i)}}(\tilde{\phi})$, we conclude

$$\begin{aligned}
&\breve{\mathcal{R}}^{(i)}(\breve{\hat{\underline{\mathbf{f}}}}^{(i)}, \breve{\hat{\underline{\mathbf{g}}}}^{(i)}; \mathcal{D}_{\mathrm{train}}) + \frac{\breve{\nu}^{(i)}}{2}\breve{\mathcal{R}}^{(i)}_{[r_i]}(\breve{\hat{\underline{\mathbf{f}}}}^{(i)}, \breve{\hat{\underline{\mathbf{g}}}}^{(i)}; \mathcal{D}_{1\otimes 1}) \\
&\quad \le 2\kappa_{\mathrm{apx}}\mathsf{\mathbf{tail}}^{(i)\star}_2(r_i) + 3\breve{\nu}^{(i)}\breve{\mathcal{R}}^{(i)}_{[r_i]}(\breve{\tilde{\underline{\mathbf{f}}}}^{(i)}, \breve{\hat{\mathbf{Q}}}_{r_i} \cdot \breve{\tilde{\underline{\mathbf{g}}}}^{(i)}; \mathcal{D}_{1\otimes 1}) \\
&\qquad + 368\left(1 + \frac{2\breve{\nu}^{(i)}n_1 n_3}{n_4}\right)\frac{B^4(\mathscr{N}_{r_i,n_4} + \log(12K/\delta))}{n_3},
\end{aligned}$$

which completes the proof. □

D.3.2 A Lemma on the Risk Bound for t-Bilinear Combinatorial Extrapolation

We now present a lemma on the risk bound for t-bilinear combinatorial extrapolation. This bound depends on the upper bound ϵ_{trn} on the risk of the learned embedding $(\hat{\underline{\mathbf{f}}}, \hat{\underline{\mathbf{g}}})$ for the training distribution $\mathcal{D}_{\text{train}}$, the upper bound $\epsilon_{1\otimes 1}$ on the risk for the top-block distribution $\mathcal{D}_{1\otimes 1}$, and $\underline{\sigma}_r(\hat{\underline{\mathbf{f}}}, \hat{\underline{\mathbf{g}}})$.

Lemma D.22. *Given vectors* $\mathbf{r} = (r_i)_{i=1}^K \in \mathbb{N}^K$, $\boldsymbol{\alpha} = (\alpha_i)_{i=1}^K \in \mathbb{R}^K$, $\boldsymbol{\epsilon}_{\text{trn}} = (\breve{\epsilon}_{\text{trn}}^{(i)})_{i=1}^K \in \mathbb{R}^K$, *and* $\boldsymbol{\epsilon}_{\mathcal{D}_{1\otimes 1}} = (\breve{\epsilon}_{\mathcal{D}_{1\otimes 1}}^{(i)})_{i=1}^K \in \mathbb{R}^K$, *where* $\alpha_i \geq 1$, $\breve{\epsilon}_{\text{trn}}^{(i)}$, $\breve{\epsilon}_{\mathcal{D}_{1\otimes 1}}^{(i)} \geq 0$ *for all* $i \in [K]$, *suppose* $(\hat{\underline{\mathbf{f}}}, \hat{\underline{\mathbf{g}}})$ *are* $\boldsymbol{\alpha}$*-conditioned and* $(\boldsymbol{\epsilon}_{\text{trn}}, \boldsymbol{\epsilon}_{\mathcal{D}_{1\otimes 1}})$*-accurate* $\mathbb{R}^{\|\mathbf{r}\|_\infty \times 1 \times K}$*-embeddings of tensor-multl-rank* $\mathbf{r}$, *where* $r_i \leq \breve{\sigma}_1^{\star,(i)}/(40\breve{\epsilon}_{\mathcal{D}_{1\otimes 1}}^{(i)})$. *If* $\breve{\sigma}_{r_i}^{\star,(i)} > 0$, *then*

$$\begin{aligned}&\breve{\mathcal{R}}^{(i)}(\breve{\hat{\underline{\mathbf{f}}}}^{(i)}, \breve{\hat{\underline{\mathbf{g}}}}^{(i)}; \mathcal{D}_{\text{test}}) \\ &\lesssim_\star \left(r_i^4(\breve{\epsilon}_{\mathcal{D}_{1\otimes 1}}^{(i)})^2 + \alpha_i r_i^2 (\breve{\sigma}_{r_i+1}^{\star,(i)})^2 + \textsf{\textbf{tail}}_1^{(i)\star}(r_i)^2\right) + \alpha_i \left(\frac{r_i^6(\breve{\epsilon}_{\mathcal{D}_{1\otimes 1}}^{(i)})^4 + (\breve{\epsilon}_{\text{trn}}^{(i)})^4 + \textsf{\textbf{tail}}_2^{(i)\star}(r_i)^2}{(\sigma_{r_i}^\star)^2}\right).\end{aligned}$$

Lemma D.22 provides an upper bound on the excess risk $\breve{\mathcal{R}}^{(i)}(\breve{\hat{\underline{\mathbf{f}}}}^{(i)}, \breve{\hat{\underline{\mathbf{g}}}}^{(i)}; \mathcal{D}_{\text{test}})$ for each i-th frequency component of the learned embeddings on $\mathcal{D}_{\text{test}}$. This bound depends on the embedding rank r_i, accuracy parameters $\breve{\epsilon}_{\text{trn}}^{(i)}$ and $\breve{\epsilon}_{\mathcal{D}_{1\otimes 1}}^{(i)}$, conditioning parameter α_i, and the singular values and tail sums of the true embeddings. The embeddings must be $\boldsymbol{\alpha}$-conditioned and $(\boldsymbol{\epsilon}_{\text{trn}}, \boldsymbol{\epsilon}_{\mathcal{D}_{1\otimes 1}})$-accurate, with r_i limited relative to $\breve{\sigma}_1^{\star,(i)}/\breve{\epsilon}_{\mathcal{D}_{1\otimes 1}}^{(i)}$.

Proof of Lemma D.22. The proof closely follows the arguments in Theorem 8 of [46], extending them to the multi-output setting using the tensor algebraic tools developed in our work.

Let $\mathbf{s} = (s_1, \cdots, s_K)^\top \in \mathbb{N}^K$ and $\boldsymbol{\epsilon} = (\breve{\epsilon}_i, \cdots, \breve{\epsilon}_K)$ be such that

$$\breve{\epsilon}_i^2 \geq \inf_{s' \geq s_i - 1} \breve{\mathcal{R}}_{[s']}^{(i)}(\breve{\hat{\underline{\mathbf{f}}}}^{(i)}, \breve{\hat{\underline{\mathbf{g}}}}^{(i)}; \mathcal{D}_{1\otimes 1}) := \mathbb{E}_{\mathcal{D}_{1\otimes 1}}[(\langle \breve{\hat{\underline{\mathbf{f}}}}^{(i)}, \breve{\hat{\underline{\mathbf{g}}}}^{(i)} \rangle - \langle \mathbf{P}_{s_i}^\star \breve{\underline{\mathbf{f}}}^{\star,(i)}, \mathbf{P}_{s_i}^\star \breve{\underline{\mathbf{g}}}^{\star,(i)} \rangle)^2], \tag{36}$$

$$s_i < \frac{\|M(\underline{\boldsymbol{\Sigma}}_{1\otimes 1}^\star)^{(i)}\|_{\text{op}}}{40\breve{\epsilon}_i}, \tag{37}$$

where $\mathbf{P}_{s_i}^\star$ is the projection operator onto the top-s_i eigenspaces of $M(\underline{\boldsymbol{\Sigma}}_{1\otimes 1}^\star)^{(i)}$.

Suppose $(\hat{\underline{\mathbf{f}}}, \hat{\underline{\mathbf{g}}})$ are *full-multi-rank-*$\mathbf{r}$. Then there exists an index vector $\mathbf{k} = (k_1, \cdots, k_K)^\top \in \mathbb{N}^K$ with $k_i \in [\min\{r_i, s_i - 1\}]$ for all $i \in [K]$, and functions $\underline{\mathbf{f}} : \mathcal{X} \to \mathcal{M}$ and $\underline{\mathbf{g}} : \mathcal{Y} \to \mathcal{M}$ such that $(\underline{\mathbf{f}}, \underline{\mathbf{g}})$ are *aligned* $\mathbf{k}$*-proxies* of $(\hat{\underline{\mathbf{f}}}, \hat{\underline{\mathbf{g}}})$. The construction of $(\underline{\mathbf{f}}, \underline{\mathbf{g}})$ follows the same steps as in the proof of Theorem 8(a) in [46], applied to each component $(\breve{\hat{\underline{\mathbf{f}}}}^{(i)}, \breve{\hat{\underline{\mathbf{g}}}}^{(i)})$ separately. Specifically, we set

$$\breve{\underline{\mathbf{f}}}^{(i)} = (\iota_{\mathbf{r}_i} \circ (\breve{\underline{\mathbf{T}}}^{(i)})^{-1}) \breve{\hat{\underline{\mathbf{f}}}}^{(i)}, \quad \breve{\underline{\mathbf{g}}}^{(i)} = (\iota_{\mathbf{r}_i} \circ \breve{\underline{\mathbf{T}}}^{(i)}) \breve{\hat{\underline{\mathbf{g}}}}^{(i)},$$

where $\iota_{\mathbf{r}_i} : \mathbb{R}^{r_i} \to \mathcal{M}_i$ is a t-isometric inclusion, $\mathcal{M}_i$ is the i-th component Hilbert space of $\mathcal{M}$, and $\breve{\underline{\mathbf{T}}}^{(i)}$ is the balancing operator for $(\breve{\hat{\underline{\mathbf{f}}}}^{(i)}, \breve{\hat{\underline{\mathbf{g}}}}^{(i)})$ as defined in Lemma 15. The index k_i is chosen as the largest integer in $[\min\{r_i, s_i - 1\}]$ such that $\breve{\sigma}_{k_i}^{\star,(i)} - \breve{\sigma}_{k_i+1}^{\star,(i)} \geq \frac{\breve{\sigma}_{k_i}^{\star,(i)}}{\min\{r_i, s_i - 1\}}$. The existence of such a k_i is guaranteed by the gap condition (36) and the fact that $(\breve{\hat{\underline{\mathbf{f}}}}^{(i)}, \breve{\hat{\underline{\mathbf{g}}}}^{(i)})$ are full-rank, as shown in [46].

With the aligned proxies $(\underline{\mathbf{f}}, \underline{\mathbf{g}})$ constructed, we can bound their error terms using the same arguments as in the proof of Theorem $\bar{8}$(b) in [46], applied componentwise. This yields the bounds

$$\begin{aligned}\Delta_0(\breve{\underline{\mathbf{f}}}^{(i)}, \breve{\underline{\mathbf{g}}}^{(i)}, k_i) + \textsf{\textbf{tail}}_2^{(i)\star}(k_i) &\lesssim \textsf{\textbf{tail}}_2^{(i)\star}(s_i) + s_i^3 \breve{\epsilon}_i^2 + s_i (\breve{\sigma}_{s_i}^{\star,(i)})^2, \\ \Delta_1(\breve{\underline{\mathbf{f}}}^{(i)}, \breve{\underline{\mathbf{g}}}^{(i)}, k_i) &\lesssim \textsf{\textbf{tail}}_1^{(i)\star}(s_i) + (\sqrt{r_i} + s_i^2)\breve{\epsilon}_i + s_i \breve{\sigma}_{s_i}^{\star,(i)}.\end{aligned}$$

Finally, suppose $\breve{\epsilon}_i^2 \leq (1-\alpha_i^{-1})(\breve{\sigma}_{r_i}^{\star,(i)})^2$ for some $\alpha_i \geq 1$. Then, as shown in the proof of Theorem 8(c) in [46], we have $\sigma_{r_i}(\breve{\hat{\underline{\mathbf{f}}}}^{(i)}, \breve{\hat{\underline{\mathbf{g}}}}^{(i)})^2 \geq (\breve{\sigma}_{r_i}^{\star,(i)})^2/\alpha_i$, and hence $(\hat{\underline{\mathbf{f}}}, \hat{\underline{\mathbf{g}}})$ are necessarily full-multi-rank-$\mathbf{r}$.

Combining these bounds and applying a refined error decomposition in each i-th frequency component induced by the transform $M(\cdot)$ in Eq. (1), we arrive at the risk bound

$$
\begin{aligned}
&\breve{\mathcal{R}}^{(i)}(\breve{\hat{\underline{\mathbf{f}}}}^{(i)}, \breve{\hat{\underline{\mathbf{g}}}}^{(i)}; \mathcal{D}_{\mathrm{test}}) \\
&\lesssim \kappa_{\mathrm{tst}}\kappa_{\mathrm{cov}}^2 \left(\Delta_1(\breve{\underline{\mathbf{f}}}^{(i)}, \breve{\underline{\mathbf{g}}}^{(i)}, k_i)^2 + \frac{1}{\sigma_{r_i}(\breve{\hat{\underline{\mathbf{f}}}}^{(i)}, \breve{\hat{\underline{\mathbf{g}}}}^{(i)})^2} \Big(\mathsf{tail}_2^{(i)\star}(k_i) + \Delta_0(\breve{\underline{\mathbf{f}}}^{(i)}, \breve{\underline{\mathbf{g}}}^{(i)}, k_i) + \kappa_{\mathrm{cov}}\kappa_{\mathrm{trn}}\breve{\mathcal{R}}^{(i)}(\breve{\underline{\mathbf{f}}}^{(i)}, \breve{\underline{\mathbf{g}}}^{(i)}; \mathcal{D}_{\mathrm{train}}) \Big)^2 \right) \\
&\lesssim_\star s_i^4 \breve{\epsilon}_i^2 + \alpha_i s_i^2 (\breve{\sigma}_{s_i+1}^{\star,(i)})^2 + \mathsf{tail}_1^{(i)\star}(s_i)^2 + \alpha_i \left(\frac{s_i^6 (\breve{\epsilon}_i)^4 + (\breve{\epsilon}_{\mathrm{trn}}^{(i)})^4 + \mathsf{tail}_2^{(i)\star}(s_i)^2}{(\breve{\sigma}_{s_i}^{\star,(i)})^2} \right) \\
&\lesssim_\star r_i^4 (\breve{\epsilon}_{\mathcal{D}_{1\otimes 1}}^{(i)})^2 + \alpha_i r_i^2 (\breve{\sigma}_{r_i+1}^{\star,(i)})^2 + \mathsf{tail}_1^{(i)\star}(r_i)^2 + \alpha_i \left(\frac{r_i^6 (\breve{\epsilon}_{\mathcal{D}_{1\otimes 1}}^{(i)})^4 + (\breve{\epsilon}_{\mathrm{trn}}^{(i)})^4 + \mathsf{tail}_2^{(i)\star}(r_i)^2}{(\sigma_{r_i}^{\star})^2} \right).
\end{aligned}
$$

Here, in the last step, we used the fact that $s_i \leq r_i$ and $\breve{\epsilon}_i \leq \breve{\epsilon}_{\mathcal{D}_{1\otimes 1}}^{(i)}$, and we absorbed the constant factors into the $\lesssim_\star$ notation. This completes the proof of the lemma. □

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The paper introduces a tensor spectral theory approach to address multi-output regression challenges under combinatorial distribution shift (CDS), marking a pioneering theoretical exploration in this area.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: It acknowledges several limitations: the potential inadequacy of the spectral methods against real-world data complexity and the uncertain robustness of results from controlled experiments. We suggest future research to refine these methods fostering the development of more effective solutions.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory Assumptions and Proofs**

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [Yes]

Justification: The paper provide the full set of assumptions and a complete (and correct) proof for the theroetical results.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental Result Reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer:[Yes]

Justification: This paper presents a pioneering theoretical exploration of multi-output regression challenges under CDS using a generalized tensor spectral theory. Since no existing algorithms address these specific models, our experiments use numerical simulations with synthetic data to validate preliminary results. Details are provided in the appendix.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).

(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: We provide the code for our conceptually validation.

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (https://nips.cc/public/guides/CodeSubmissionPolicy) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (https://nips.cc/public/guides/CodeSubmissionPolicy) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental Setting/Details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: We show the experiment details in the appendix.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment Statistical Significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: The paper reports error bars and provides appropriate information about the statistical significance of the experiments, with results obtained from multiple repeated trials.

   Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments Compute Resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: For the experiment, the paper provides sufficient information on the computer resources needed to reproduce the experiments.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code Of Ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics https://neurips.cc/public/EthicsGuidelines?

   Answer: [Yes]

   Justification: This paper conforms in every respect with the NeurIPS Code of Ethics.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader Impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [NA]

Justification: The paper provides a new tensor spectral perspective for the multi-output regression problem due to combinatorial distribution shift. It focuses solely on these technical aspects and does not have potential societal impacts.

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

Answer: [NA]

Justification: This paper does not describe safeguards for the responsible release of data or models because it focuses exclusively on the technical aspects of addressing combinatorial distribution shift in multi-output regression. Therefore, this question is not applicable (N/A).

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [NA]

Justification: This paper does not use any external assets such as code, data, or models. Therefore, the question regarding the crediting of creators or original owners, as well as the respect for licenses and terms of use, is not applicable (N/A).

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, paperswithcode.com/datasets has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer:[NA]

    Justification: This paper does not introduce any new assets. Therefore, the question regarding the documentation of new assets is not applicable (N/A).

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: This paper does not involve crowdsourcing experiments or research with human subjects. Therefore, the question regarding instructions, screenshots, and compensation details is not applicable (N/A).

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [NA]

Justification: This paper does not involve study participants or human subjects. Therefore, the question regarding potential risks, disclosures to subjects, and IRB approvals is not applicable (N/A).

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.