Finite-Sample Analysis of Deep CCA-Based Unsupervised Post-Nonlinear Multimodal Learning

Qi Lyu and Xiao Fu

***Abstract*—Canonical correlation analysis (CCA) has been essential in *unsupervised* multimodal/multiview latent representation leaning and data fusion. Classic CCA extracts shared information from multiple modalities of data using *linear* transformations. In recent years, deep neural networks-based *nonlinear* feature extractors were combined with CCA to come up with new variants, namely, the "DeepCCA" line of work. These approaches were shown to have enhanced performance in many applications. However, theoretical supports of DeepCCA are often lacking. To address this challenge, the authors' recent work in [1] showed that, under a reasonable post-nonlinear generative model, a carefully designed DeepCCA criterion provably removes unknown distortions in data generation and identifies the shared information across modalities. Nonetheless, a critical assumption used in [1] for identifiability analysis was that unlimited data is available—which is unrealistic. This brief paper puts forth a finite-sample analysis of the DeepCCA method in [1]. The main result is that the finite-sample version of the method can still estimate the shared information with a guaranteed accuracy—when the number of samples is sufficiently large. Our analytical approach is a nontrivial integration of statistical learning, numerical differentiation, and robust system identification—which may be of interest beyond the scope of DeepCCA and benefit other unsupervised learning paradigms.**

***Index Terms*—Unsupervised multimodal analysis, post-nonlinear mixture model, sample complexity, identifiability**

I. Introduction

Data often comes with multiple modalities (e.g., audio and video describing the same events). *Canonical correlation analysis* (CCA) [2], [3] has been one of the most prominent unsupervised multimodal learning tools. CCA seeks linear transformations to "project" the high-dimensional multimodal data to a low-dimensional space, so that the low-dimensional embedding represents the *shared* information across the modalities. Both empirical and theoretical evidence shows that CCA admits a series of appealing features. In particular, CCA is robust to strong unknown modality-specific interference and admits identifiability of the shared latent components, under reasonable linear mixture-type generative models [4], [5].

In recent years, nonlinear function approximators (e.g., kernel functions and deep neural networks) are combined with CCA to come up with nonlinear variants. The deep learning-based CCA (DeepCCA) learning paradigm is particularly attractive due to its balance between efficiency and effectiveness [6], [7]. However, despite of empirical successes, understanding to DeepCCA has been largely elusive.

Recently, the authors made important advances towards understanding DeepCCA under a reasonable multimodal *post-nonlinear mixture* (PNM) model [1]. PNM models are widely used in machine learning for modeling complex data generating processes where unknown nonlinear distortions arise; see applications in source separation [8], brain signal embedding [9], and causality discovery [10]. Under PNM, the work in [1] showed that a carefully designed DeepCCA learning criterion guarantees to extract modality-shared latent information, even if strong private interferences and unknown nonlinearities are present. To our best knowledge, this is the first identifiability analysis of nonlinear multimodal models under DeepCCA [1]. Nonetheless, a critical gap is yet to fill. Specifically, the proof in [1] used a key assumption that unlimited data are available—which is far from realistic. Extending the result to the finite-sample regime is nontrivial, since the major analytical steps in [1] rely on differentiability of certain functions that hinges on the unlimited data assumption.

Contributions. This brief paper offers a finite-sample analysis of the DeepCCA method in [1] under the same multimodal PNM model. We show that the finite-sample version of the learning criterion in [1] still provably extracts the shared information across modalities, given a sufficiently large sample size. The idea is to convert the model identification problem to a special regression problem, and use statistical error analysis and *numerical differentiation* techniques to quantify the nonlinearity removal effectiveness. Then, the shared information identification accuracy can be quantified by robustness analysis of system identification. The main result presents finite-sample guarantees under the settings of [1]. We should mention that finite-sample analysis is rare in latent component analysis, perhaps due to the challenging nature of the unsupervised settings. Therefore, the analytical approach may be of interest beyond DeepCCA and benefit other unsupervised learning paradigms, e.g., nonlinear independent component analysis (nICA)—whose analyses still largely rely on unlimited data assumptions; see, e.g., [8], [11]–[14].

II. Background

A. Generative Model, CCA, and Deep CCA

Consider two modalities/views of data, i.e., $\boldsymbol{x}_\ell^{(q)} \in \mathbb{R}^{M_q}$ for $q = 1, 2$. The first view $\boldsymbol{x}_\ell^{(1)}$ (e.g., video) and second view $\boldsymbol{x}_\ell^{(2)}$ (e.g., audio) are both representations of the same entity

This work is supported in part by the National Science Foundation (NSF) under Project NSF ECCS-1808159 and in part by the Army Research Office (ARO) under Project ARO W911NF-21-1-0227.

Q. Lyu and X. Fu are with the School of Electrical Engineering and Computer Science, Oregon State University, Corvallis, OR 97331, United States. Email: {lyuqi, xiao.fu}@oregonstate.edu

or event (e.g., 'cat' or a 'car collision'). CCA extracts low-dimensional representations from the two views by finding shared information in $\boldsymbol{x}_\ell^{(q)}$ for $q = 1, 2$. This is often more effective than single view methods, e.g., principal component analysis (PCA) and nonnegative matrix factorization (NMF) [15], [16], as noticed in [4], [5].

To understand the effectiveness of the classic CCA, the work in [4], [5] took an unsupervised generative model learning viewpoint. For example, the recent work in [5] uses the following model

$$\boldsymbol{x}_\ell^{(q)} = \boldsymbol{A}^{(q)}\boldsymbol{z}_\ell^{(q)},\ \boldsymbol{z}_\ell^{(q)} = [\boldsymbol{s}_\ell^\top, (\boldsymbol{c}_\ell^{(q)})^\top]^\top, \tag{1}$$

where $q = 1, 2$, $\boldsymbol{s}_\ell \in \mathcal{S} \subseteq \mathbb{R}^D$ is the shared latent component across views and $\boldsymbol{c}_\ell^{(q)} \in \mathcal{C}_q \subseteq \mathbb{R}^{D_q}$ is the private component in view q (which could be interference), and $\mathrm{range}(\boldsymbol{A}^{(q)})$ is where view q resides. The goal amounts to extracting the shared information represented by $\boldsymbol{s}_\ell$ from the data; see similar modeling and estimation goals in [4], where the private information was modeled as colored Gaussian noise.

The authors' work in [1] extended the above perspective to the nonlinear regime, i.e.,

$$\boldsymbol{x}_\ell^{(q)} = \boldsymbol{g}^{(q)}\left(\boldsymbol{A}^{(q)}\boldsymbol{z}_\ell^{(q)}\right), \tag{2}$$

where $\boldsymbol{g}^{(q)}(\cdot) = [g_1^{(q)}(\cdot), \ldots, g_{M_q}^{(q)}(\cdot)]$ represent *unknown* nonlinear distortions in the data generating/acquisition process. The model is reminiscent of the PNM model in nonlinear blind source separation [8], [17]. PNM makes much sense in many data acquisition processes, e.g., brain signal (EEG/MEG) sensing [9], hyperspectral analysis [18], [19], and image data generation [1]. It also finds applications in causality discovery [10]. Given (2), the authors proposed a criterion to find the shared information $\boldsymbol{s}_\ell$ [1]—which can be recast as the following nonlinear CCA formulation:

$$\min_{\boldsymbol{B}^{(q)},\ \boldsymbol{f}^{(q)}} \ \mathbb{E}\left[\left\|\boldsymbol{B}^{(1)}\boldsymbol{f}^{(1)}\left(\boldsymbol{x}^{(1)}\right) - \boldsymbol{B}^{(2)}\boldsymbol{f}^{(2)}\left(\boldsymbol{x}^{(2)}\right)\right\|_2^2\right] \tag{3a}$$

$$\text{s.t. } \mathbb{E}\left[\boldsymbol{B}^{(q)}\boldsymbol{f}^{(q)}\left(\boldsymbol{x}^{(q)}\right)\left(\boldsymbol{B}^{(q)}\boldsymbol{f}^{(q)}\left(\boldsymbol{x}^{(q)}\right)\right)^\top\right] = \boldsymbol{I}, \tag{3b}$$

$$\mathbb{E}\left[\boldsymbol{f}^{(q)}(\boldsymbol{x}^{(q)})\right] = \boldsymbol{0},\ \boldsymbol{f}^{(q)} : \text{invertible}. \tag{3c}$$

where the expectation is taken over $(\boldsymbol{x}^{(1)}, \boldsymbol{x}^{(2)}) \sim \mathcal{D}$ and $\mathcal{D}$ has a positive probability density over $\mathcal{X}_1 \times \mathcal{X}_2$ in which $\mathcal{X}_q$ is the domain where $\boldsymbol{x}^{(q)}$ is defined over. Note that $\boldsymbol{f}^{(q)} = [f_1^{(q)}, \ldots, f_{M_q}^{(q)}]$ and $f_m^{(q)}(\cdot) : \mathbb{R} \to \mathbb{R}$ is a scalar function. If one uses a DNN to represent f_m, then the above becomes a `DeepCCA` formulation[1]. In practice, the invertibility and zero mean constraint on $f_m^{(q)}(x_m)$ is to prevent $f_m^{(q)}(x_m)$ from outputting trivial solutions, e.g., constants. The invertibility can be promoted by using a data reconstruction regularization [1] or using special function approximators, e.g., normalizing flows [20].

The idea is to use $f_m^{(q)}$ to cancel $g_m^{(q)}$ and to make the problem essentially a linear CCA problem. If $\boldsymbol{f}^{(q)}$'s and (3c) are not used and $\mathbb{E}[\boldsymbol{x}^{(q)}] = \boldsymbol{0}$, the formulation in (3) becomes the classic linear CCA problem.

[1]A remark is that there are other `DeepCCA` formulations, e.g., those in [6], [7], but not designed for the generative model in (2). These versions of `DeepCCA` are out of the scope of this brief paper.

B. Identifiability Theory and Critical Gap

The takeaway in [5] is that classic CCA provably extracts $\mathrm{range}(\boldsymbol{S}^\top)$ where $\boldsymbol{S} = [\boldsymbol{s}_1, \ldots, \boldsymbol{s}_N]^\top$, i.e., the subspace contains shared components under (1). The authors showed that similar results also hold under (2), if one use neural networks (or other universal function learners) to replace the linear projectors in CCA. To be precise, the authors had following assumption and theorem in [1]:

Assumption 1 Under (2), assume that $M_q \geq D + D_q$ and that the mixing matrices $\boldsymbol{A}^{(q)}$ are drawn from any absolutely continuous distributions. Assume that the components in $[s_{d,\ell}, (\boldsymbol{c}_\ell^{(1)})^\top, (\boldsymbol{c}_\ell^{(2)})^\top]^\top$ are *ubiquitously unanchored* (see definition in [1]) for any d, and also $D_q(D_q+1)/2 \geq M_q$.

The *ubiquitously unanchored* (UU) condition means that when fixing one variable, the others can still take any possible values within a certain local region, which further implies that for any x, y satisfy such condition, $\frac{\partial x}{\partial y} = 0$ always holds. In other words, it means that the ubiquitously unanchored components are "locally free". This is much less stringent than some commonly used conditions in the nonlinear learning literature, e.g., statistical independence. As shown in [1], two UU variables can be strongly dependent. Under Assumption 1, the authors showed that

Theorem 1 *[1] Under Assumption 1, suppose that $(\boldsymbol{B}^{(q)}, \boldsymbol{f}^{(q)})$ for $q = 1, 2$ are solutions of* (3) *with $\|\boldsymbol{B}^{(q)}\|_0 = DM_q$ and $f_m^{(q)}$ being a universal function approximator. If* (3) *is solved over the entire continuous domain $\mathcal{X}_1 \times \mathcal{X}_2$, then, the composition $f_i^{(q)} \circ g_i^{(q)}(x)$ for all i, q are affine functions with probability one. In addition, $\boldsymbol{B}^{(q)}\boldsymbol{f}^{(q)}(\boldsymbol{x}^{(q)}) = \boldsymbol{\Theta}\boldsymbol{s}$ for any $\boldsymbol{x}^{(q)} \in \mathcal{X}_q$, where $\boldsymbol{\Theta} \in \mathbb{R}^{D\times D}$ is any nonsingular matrix.*

Theorem 1 asserts that in the *population* case, the `DeepCCA` criterion extracts the shared information up to a nonsingular linear transformation. Notably, the criterion achieves provable shared information extraction even if the private components/interference $\boldsymbol{c}_\ell^{(q)}$ have overwhelming energy over the shared components. This property is inherited from classic CCA [5]. However, single modal tools such as PCA always extracts high energy components first. This articulates the advantages of using multiple modalities.

One critical gap left unfilled in [1] is as follows. The main identifiability theorem for using neural CCA under (2) is derived under the premises that unlimited data is available and that the function approximator is universal. In particular, unlimited/infinite sample may never be available in practice—

but the method in [19] works well with the finite sample version of (3), i.e.,

$$\min_{\boldsymbol{B}^{(q)},\ \boldsymbol{f}^{(q)}} \ \frac{1}{N}\sum_{\ell=1}^{N}\left\|\boldsymbol{B}^{(1)}\boldsymbol{f}^{(1)}\left(\boldsymbol{x}_\ell^{(1)}\right)-\boldsymbol{B}^{(2)}\boldsymbol{f}^{(2)}\left(\boldsymbol{x}_\ell^{(2)}\right)\right\|_2^2 \tag{4a}$$

$$\text{s.t. } \frac{1}{N}\sum_{\ell=1}^{N}\boldsymbol{B}^{(q)}\boldsymbol{f}^{(q)}\left(\boldsymbol{x}_\ell^{(q)}\right)\left(\boldsymbol{B}^{(q)}\boldsymbol{f}^{(q)}\left(\boldsymbol{x}_\ell^{(q)}\right)\right)^\top=\boldsymbol{I}, \tag{4b}$$

$$\frac{1}{N}\sum_{\ell=1}^{N}\boldsymbol{f}^{(q)}(\boldsymbol{x}_\ell^{(q)})=\boldsymbol{0},\ \ \boldsymbol{f}^{(q)}:\text{invertible}. \tag{4c}$$

Characterizing the performance of (4) gives rise to an important and challenging research question. Note that since the problem is unsupervised, sample complexity analysis tools developed for supervised learning, e.g., [21], are not directly applicable. More importantly, since the success of unsupervised learning is not by measuring the cost function value as in supervised learning, the performance metric design and its analytical underpinning are challenging questions. In this work, we provide an answer to this inquiry, which is an integration of three major analytical steps. First, we analyze the statistical error of the fitting problem. This analysis is reminiscent of generalization analysis in supervised learning. Second, we use numerical differentiation techniques to characterize the nonlinearity removal performance over unseen samples, which is reminiscent of the authors' recent work in [19] that shows finite sample identifiability of a special single view model. Third, we leverage the first step to recast the `DeepCCA` formulation as a noisy least squares problem, and characterize its solution—which serves as our problem success metric.

III. Main Result

Our main result in this work is as follows. For notational simplicity we assume that $M=M_1=M_2$ and $D_1=D_2$. First, we have the following assumptions:

Assumption 2 In the generative model, $g_m^{(q)}\in\mathcal{G}$, where the function class $\mathcal{G}$ is 4th-order differentiable and bounded. In addition, $c_j^{(q)}\in[-C_p,C_p]$ with $0<C_p<\infty$ for $j\in[D_q]$.

Assumption 3 The learning function f_m is taken from $\mathcal{F}$, where the function class $\mathcal{F}$ is 4th-order differentiable and bounded. In addition, the function class $\mathcal{F}$ has bounded Rademacher complexity $\mathfrak{R}_N$ under N samples from $\mathcal{D}$. Besides, assume that $\left|B_{i,j}^{(q)}\right|\le C_b$.

Assumption 4 Assume that there exists $\boldsymbol{B}^{(q)},\boldsymbol{f}^{(q)}$ such that $\frac{1}{N}\sum_{\ell=1}^{N}\left\|\boldsymbol{B}^{(1)}\boldsymbol{f}^{(1)}(\boldsymbol{x}_\ell^{(1)})-\boldsymbol{B}^{(2)}\boldsymbol{f}^{(2)}(\boldsymbol{x}_\ell^{(2)})\right\|_2^2\le\nu$ under $f_m^{(q)}\in\mathcal{F}$ for all m.

Assumption 5 The absolute value of 4th-order derivative of

$$\left(\boldsymbol{b}_d^{(1)}\right)^\top\boldsymbol{f}^{(1)}\left(\boldsymbol{x}_\ell^{(1)}\right)-\left(\boldsymbol{b}_d^{(2)}\right)^\top\boldsymbol{f}^{(2)}\left(\boldsymbol{x}_\ell^{(2)}\right)$$

is bounded by C_ϕ for $d\in[D]$, where $(\boldsymbol{f}^{(q)},\boldsymbol{B}^{(q)})$ is any solution of (4) and $(\boldsymbol{b}_d^{(q)})^\top$ is the dth row of $\boldsymbol{B}^{(q)}$.

Note that Assumption 4 can always be fulfilled using powerful function approximators, e.g., deep neural networks. The boundedness assumptions hold if the learned functions are bounded, which is easy to check or regularize. Under these assumptions, we show that:

Theorem 2 *Under* (2)*, the assumptions in Theorem 1, and Assumptions 2-5 (see in the next section), assume that* $(\boldsymbol{x}_\ell^{(1)},\boldsymbol{x}_\ell^{(2)})$ *for* $\ell=1,\dots,N$ *are i.i.d. samples of* $(\boldsymbol{x}^{(1)},\boldsymbol{x}^{(2)})\sim\mathcal{D}$*. Denote* $(\boldsymbol{f}^{(q)},\boldsymbol{B}^{q)})$ *as any optimal solution of* (4) *and* $\mathfrak{R}_N$ *the Rademacher complexity of* $\mathcal{F}$*. Then, the following holds with probability of at least* $1-\alpha-\delta$*:*

$$\boldsymbol{B}^{(q)}\left[\boldsymbol{f}^{(q)}(\boldsymbol{x}_1^{(q)}),\cdots,\boldsymbol{f}^{(q)}(\boldsymbol{x}_N^{(q)})\right]=(\boldsymbol{\Theta}^{(q)})^\top\boldsymbol{S}+(\boldsymbol{\Omega}^{(q)})^\top\boldsymbol{C}^{(q)},$$

where $\boldsymbol{S}=[\boldsymbol{s}_1,\dots,\boldsymbol{s}_N]$ *and* $\boldsymbol{C}^{(q)}=[\boldsymbol{c}_1^{(q)},\dots,\boldsymbol{c}_N^{(q)}]$*, and*

$$\left\|\boldsymbol{\Omega}^{(q)}\right\|_F=O\left(\frac{D(D+D_1)}{\alpha}\left(M\mathfrak{R}_N+\sqrt{\log(1/\delta)/N}\right)^{1/4}+\sqrt{\nu}\right),$$

if $-C_p+\kappa_i\le c_i^{(q)}\le C_p-\kappa_i$ *for all* $i\in[D_q]$*, where* $\kappa_i=\Omega(\left(MC_b\mathfrak{R}_N+\sqrt{\log(1/\delta)/N}\right)^{1/8}/C_\phi^{1/4})$.

Theorem 2 means that if N is sufficiently large, then the solution $(\boldsymbol{B}^{(q)},\ \boldsymbol{f}^{(q)})$ approximately extracts the subspace of $\boldsymbol{S}$, i.e., $\text{range}(\boldsymbol{S}^\top)$. The Rademacher complexity $\mathfrak{R}_N$ is determined by the selected function class $\mathcal{F}$ and the sample size N. For instance, if $\mathcal{F}$ is modeled with one-hidden-layer neural network with R neurons and ReLU activation, then $\mathfrak{R}_N=O(\sqrt{R/N})$. In practice, one may use an expressive network (e.g., by increasing R) in order to reduce ν. But under a fixed N, one may also hope to avoid using an overly large R that makes $\mathfrak{R}_N$ dominant—which may hurt the performance.

IV. Proof of Main Theorem

We start with the following lemma:

Lemma 1 *Consider the function class*

$$\mathcal{H}=\left\{l\left(\boldsymbol{x}_\ell^{(1)},\boldsymbol{x}_\ell^{(2)}\right)\middle|\,l\left(\boldsymbol{x}_\ell^{(1)},\boldsymbol{x}_\ell^{(2)}\right)=\left\|\boldsymbol{B}^{(1)}\boldsymbol{f}^{(1)}\left(\boldsymbol{x}_\ell^{(1)}\right)-\boldsymbol{B}^{(2)}\boldsymbol{f}^{(2)}\left(\boldsymbol{x}_\ell^{(2)}\right)\right\|_2^2\right\},$$

where each $f_m^{(q)}(\cdot):\mathbb{R}\to\mathbb{R}\in\mathcal{F}$ *as defined in Assumption 3. Assume that* $(\boldsymbol{x}_\ell^{(1)},\boldsymbol{x}_\ell^{(2)})$ *for* $\ell\in[N]$ *are i.i.d. samples drawn from* $\mathcal{D}$*. Then, the Rademacher complexity of class* $\mathcal{H}$ *is bounded by*

$$\mathfrak{R}_N(\mathcal{H})\le 8DMC_b\mathfrak{R}_N,$$

where $\mathfrak{R}_N$ *and* C_b *are defined in Assumption 3.*

Proof: Note that the cost function can be rewritten as

$$\sum_{d=1}^{D}\left(\sum_{i=1}^{M}B_{d,i}^{(1)}f_i^{(1)}\left(x_i^{(1)}\right)-\sum_{j=1}^{M}B_{d,j}^{(2)}f_j^{(2)}\left(x_j^{(2)}\right)\right)^2.$$

Then according to the property of Rademacher complexity, the complexity of function class of $\sum_{m=1}^{M} B_{d,m}^{(q)} f_m^{(q)}(\cdot)$ is upper bounded by $MC_b\mathfrak{R}_N$.

Since we have orthogonality constraint on $\boldsymbol{B}^{(q)}\boldsymbol{f}^{(q)}(\boldsymbol{x}_\ell^{(q)})$, which indicates that $|\sum_{m=1}^{M} B_{d,m}^{(q)} f_m^{(q)}(\cdot)| \leq 1$. Therefore, the function $|\sum_{i=1}^{M} B_{d,i}^{(1)} f_i^{(1)}(x_i^{(1)}) - \sum_{j=1}^{M} B_{d,j}^{(2)} f_j^{(2)}(x_j^{(2)})|$ is bounded within $[0, 2]$, with its Rademacher complexity bounded by $2MC_b\mathfrak{R}_N$. By the composition property of Rademacher complexity, we have

$$\mathfrak{R}_N(\mathcal{H}) \leq 8DMC_b\mathfrak{R}_N,$$

which completes the proof. ∎

By applying Lemma 1 and [21, Theorem 26.5], we have the following hold with probability at least $1-\delta$:

$$v_\infty(\boldsymbol{f}, \boldsymbol{B}) \leq v_N(\boldsymbol{f}, \boldsymbol{B}) + 2\mathfrak{R}_N(\mathcal{H}) + 16D\sqrt{\frac{2\log(4/\delta)}{N}},$$

where we use $v_N(\boldsymbol{f}, \boldsymbol{B})$ and $v_\infty(\boldsymbol{f}, \boldsymbol{B})$ to denote the cost values of (4) and (3), respectively, under the solution $\boldsymbol{f} = (\boldsymbol{f}^{(1)}, \boldsymbol{f}^{(2)})$ and $\boldsymbol{B} = (\boldsymbol{B}^{(1)}, \boldsymbol{B}^{(2)})$. Therefore, we have $v_\infty(\boldsymbol{f}, \boldsymbol{B}) \leq \varepsilon$, where $\varepsilon := \nu + 16DMC_b\mathfrak{R}_N + 16D\sqrt{2\log(4/\delta)/N}$. Let us denote $\|\boldsymbol{B}^{(1)}\boldsymbol{f}^{(1)}(\boldsymbol{x}_\ell^{(1)}) - \boldsymbol{B}^{(2)}\boldsymbol{f}^{(2)}(\boldsymbol{x}_\ell^{(2)})\|_2^2 = \varepsilon_\ell$, where $\mathbb{E}[\varepsilon_\ell] \leq \varepsilon$. Next, define

$$\varepsilon_{\ell,d} := \left(\left(\boldsymbol{b}_d^{(1)}\right)^\top \boldsymbol{f}^{(1)}\left(\boldsymbol{x}_\ell^{(1)}\right) - \left(\boldsymbol{b}_d^{(2)}\right)^\top \boldsymbol{f}^{(2)}\left(\boldsymbol{x}_\ell^{(2)}\right)\right)^2,$$

for $d = 1, \cdots, D$ with $\varepsilon_\ell = \sum_{d=1}^{D} \varepsilon_{\ell,d}$. Obviously, $\mathbb{E}[\varepsilon_{\ell,d}] \leq \varepsilon/D$. Also denote

$$\phi\left(\boldsymbol{s}_\ell, \boldsymbol{c}_\ell^{(1)}, \boldsymbol{c}_\ell^{(2)}\right) = \left(\boldsymbol{b}_d^{(1)}\right)^\top \boldsymbol{h}^{(1)}\left(\boldsymbol{A}^{(1)}\begin{bmatrix}\boldsymbol{s}_\ell \\ \boldsymbol{c}_\ell^{(1)}\end{bmatrix}\right) - \left(\boldsymbol{b}_d^{(2)}\right)^\top \boldsymbol{h}^{(2)}\left(\boldsymbol{A}^{(2)}\begin{bmatrix}\boldsymbol{s}_\ell \\ \boldsymbol{c}_\ell^{(2)}\end{bmatrix}\right) = \pm\sqrt{\varepsilon_{\ell,d}}.$$

Next, we will estimate the second-order (cross-)derivatives of ϕ. We consider $q = 1$ here, and the same proof applies to $q = 2$. Here, we consider $\phi(\boldsymbol{s}_\ell, \boldsymbol{c}_\ell^{(1)}, \boldsymbol{c}_\ell^{(2)})$ as a function of $\boldsymbol{c}_\ell^{(1)}$ with fixed $\boldsymbol{s}_\ell$ and $\boldsymbol{c}_\ell^{(2)}$, thus we denote it as $\phi(\boldsymbol{c}_\ell^{(1)})$ for conciseness.

A. Estimating $\partial^2\phi\left(\boldsymbol{c}_\ell^{(1)}\right)/\partial\left([\boldsymbol{c}_\ell^{(1)}]_i\right)^2$

For any continuous function $\omega(z)$ that admits non-vanishing 4th order derivatives, the second order derivative at z can be estimated as follows:

$$\omega''(z) = \frac{\omega(z+\Delta z) - 2\omega(z) + \omega(z-\Delta z)}{\Delta z^2} - \frac{\Delta z^2}{12}\omega^{(4)}(\xi), \tag{5}$$

where $\xi \in (z-\Delta z, z+\Delta z)$.

Let $\Delta\boldsymbol{c}_i^{(1)} = [0, \ldots, \Delta, \ldots, 0]^\top$ with Δ at position i, and define another two points as

$$\left(\boldsymbol{b}_d^{(1)}\right)^\top \boldsymbol{h}^{(1)}\left(\boldsymbol{A}^{(1)}\begin{bmatrix}\boldsymbol{s}_\ell \\ \boldsymbol{c}_\ell^{(1)} + \Delta\boldsymbol{c}_i^{(1)}\end{bmatrix}\right) - \left(\boldsymbol{b}_d^{(2)}\right)^\top \boldsymbol{h}^{(2)}\left(\boldsymbol{A}^{(2)}\begin{bmatrix}\boldsymbol{s}_\ell \\ \boldsymbol{c}_\ell^{(2)}\end{bmatrix}\right) = \pm\sqrt{\varepsilon_{\widetilde{\ell},d}},$$

$$\left(\boldsymbol{b}_d^{(1)}\right)^\top \boldsymbol{h}^{(1)}\left(\boldsymbol{A}^{(1)}\begin{bmatrix}\boldsymbol{s}_\ell \\ \boldsymbol{c}_\ell^{(1)} - \Delta\boldsymbol{c}_i^{(1)}\end{bmatrix}\right) - \left(\boldsymbol{b}_d^{(2)}\right)^\top \boldsymbol{h}^{(2)}\left(\boldsymbol{A}^{(2)}\begin{bmatrix}\boldsymbol{s}_\ell \\ \boldsymbol{c}_\ell^{(2)}\end{bmatrix}\right) = \pm\sqrt{\varepsilon_{\widehat{\ell},d}}.$$

Since s_i, $\boldsymbol{c}^{(1)}$ and $\boldsymbol{c}^{(2)}$ satisfy the ubiquitously unanchored condition [1], the point $\boldsymbol{z}_\ell^{(q)} + \Delta\boldsymbol{c}_i^{(q)}$ exists if $|\Delta|$ is sufficiently small. We also denote $\varepsilon_{\widetilde{\ell}} = \sum_{d=1}^{D} \varepsilon_{\widetilde{\ell},d}$ and $\varepsilon_{\widehat{\ell}} = \sum_{d=1}^{D} \varepsilon_{\widehat{\ell},d}$. By (5), we have

$$\frac{\partial^2\phi\left(\boldsymbol{c}_\ell^{(1)}\right)}{\partial\left([\boldsymbol{c}_\ell^{(1)}]_i\right)^2} = \frac{\pm\sqrt{\varepsilon_{\widetilde{\ell},d}} \mp 2\sqrt{\varepsilon_{\ell,d}} \pm \sqrt{\varepsilon_{\widehat{\ell},d}}}{\Delta^2} - \frac{\Delta^2}{12}\phi^{(4)}(\boldsymbol{\xi}_i),$$

where $\boldsymbol{\xi}_i \in \left(\boldsymbol{c}_\ell^{(1)} - \Delta\boldsymbol{c}_i^{(1)}, \boldsymbol{c}_\ell^{(1)} + \Delta\boldsymbol{c}_i^{(1)}\right)$, i.e.,

$$\begin{aligned}
&\left|\sum_{m=1}^{M} (a_{m,D+i}^{(1)})^2 (h_m^{(1)})''\left(\boldsymbol{A}^{(1)}[\boldsymbol{s}_\ell^\top, (\boldsymbol{c}_\ell^{(1)})^\top]^\top\right)\right| \\
&= \left|\frac{\pm\sqrt{\varepsilon_{\widetilde{\ell},d}} \mp 2\sqrt{\varepsilon_{\ell,d}} \pm \sqrt{\varepsilon_{\widehat{\ell},d}}}{\Delta^2} - \frac{\Delta^2}{12}\phi^{(4)}(\boldsymbol{\xi}_i)\right| \\
&\leq \frac{\sqrt{\varepsilon_{\widetilde{\ell},d}} + 2\sqrt{\varepsilon_{\ell,d}} + \sqrt{\varepsilon_{\widehat{\ell},d}}}{\Delta^2} + \frac{\Delta^2}{12}\left|\phi^{(4)}(\boldsymbol{\xi}_i)\right|.
\end{aligned}$$

By taking expectations, the following holds with probability of at least $1-\delta$:

$$\begin{aligned}
&\mathbb{E}\left[\left|\sum_{m=1}^{M} (a_{m,D+i}^{(1)})^2 (h_m^{(1)})''\left(\boldsymbol{A}^{(1)}[\boldsymbol{s}_\ell^\top, (\boldsymbol{c}_\ell^{(1)})^\top]^\top\right)\right|\right] \\
&\leq \frac{\mathbb{E}\left[\sqrt{\varepsilon_{\widetilde{\ell},d}}\right] + 2\mathbb{E}\left[\sqrt{\varepsilon_{\ell,d}}\right] + \mathbb{E}\left[\sqrt{\varepsilon_{\widehat{\ell},d}}\right]}{\Delta^2} + \frac{\Delta^2}{12}\left|\phi^{(4)}(\boldsymbol{\xi}_i)\right| \\
&\leq \frac{4\sqrt{\varepsilon/D}}{\Delta^2} + \frac{\Delta^2}{12}\left|\phi^{(4)}(\boldsymbol{\xi}_i)\right|,
\end{aligned}$$

where the second inequality is by the Jensen's inequality

$$\mathbb{E}[\sqrt{\varepsilon_{\ell,d}}] \leq \sqrt{\mathbb{E}[\varepsilon_{\ell,d}]} \leq \sqrt{\varepsilon/D},$$

which holds by the concavity of $\sqrt{x}$.

We are interested in finding the minimum upper bound of

$$\inf_{0<\Delta<\min\left\{C_p+c_{\ell,i}^{(1)}, C_p-c_{\ell,i}^{(1)}\right\}} \frac{4\sqrt{\varepsilon/D}}{\Delta^2} + \frac{\Delta^2}{12}\left|\phi^{(4)}(\boldsymbol{\xi}_i)\right|. \tag{6}$$

Note that the function in (6) is convex in Δ and is smooth. The minimizer is as follows:

$$\Delta^* \in \left\{\left(48\sqrt{\varepsilon/D}/|\phi^{(4)}(\boldsymbol{\xi}_i)|\right)^{1/4}, \min\left\{C_p + c_{\ell,i}^{(1)}, C_p - c_{\ell,i}^{(1)}\right\}\right\},$$

which gives us the minimum upper bound

$$\inf_{\Delta} \frac{4\sqrt{\varepsilon/D}}{\Delta^2} + \frac{\Delta^2}{12}\left|\phi^{(4)}(\boldsymbol{\xi}_i)\right| \leq \min\left\{\frac{2\sqrt{3\left|\phi^{(4)}(\boldsymbol{\xi}_i)\right|}(\varepsilon/D)^{1/4}}{3}, \frac{4\sqrt{\varepsilon/D}}{\kappa_i^2} + \frac{\left|\phi^{(4)}(\boldsymbol{\xi}_i)\right|}{12}\kappa_i^2\right\},$$

where $\kappa_i = \min\{C_p + c_{\ell,i}^{(1)}, C_p - c_{\ell,i}^{(1)}\}$.

If $\kappa_i \geq \left(\sqrt[4]{8\varepsilon/D}/|\phi^{(4)}(\boldsymbol{\xi}_i)|\right)^{1/4}$, then we can bound

$$\mathbb{E}\left[\left|\sum_{m=1}^{M}(a_{m,D+i}^{(1)})^2(h_m^{(1)})''\left(\boldsymbol{A}^{(1)}[\boldsymbol{s}_\ell^\top, (\boldsymbol{c}_\ell^{(1)})^\top]^\top\right)\right|\right] \leq 2\sqrt{3\left|\phi^{(4)}(\boldsymbol{\xi}_i)\right|}(\varepsilon/D)^{1/4}/3.$$

With fixed N and $\varepsilon = \nu + 16DMC_b\mathfrak{R}_N + 16D\sqrt{2\log(4/\delta)/N}$, which gives the following bound

$$\mathbb{E}\left[\left|\sum_{m=1}^{M}(a_{m,D+i}^{(1)})^2(h_m^{(1)})''\left(\boldsymbol{A}^{(1)}[\boldsymbol{s}_\ell^\top, (\boldsymbol{c}_\ell^{(1)})^\top]^\top\right)\right|\right] \leq \frac{4\sqrt{3C_\phi}}{3}\left(\nu + MC_b\mathfrak{R}_N + \sqrt{2\log(4/\delta)/N}\right)^{1/4}, \tag{7}$$

if $\kappa_i \geq 2\sqrt[4]{12}\left(\nu + MC_b\mathfrak{R}_N + \sqrt{2\log(4/\delta)/N}\right)^{1/8}/C_\phi^{1/4}$.

B. Estimating 2nd-Order Cross-Derivatives

By using similar techniques, we can estimate the cross-derivatives $\partial^2\phi\left(\boldsymbol{c}_\ell^{(1)}\right)/\partial\left([\boldsymbol{c}_\ell]_i^{(1)}\right)\partial\left([\boldsymbol{c}_\ell]_j^{(1)}\right)$. As a result, the following holds with probability at least $1-\delta$

$$\mathbb{E}\left[\left|\sum_{m=1}^{M}a_{m,D+i}^{(1)}a_{m,D+j}^{(1)}(h_m^{(1)})''\left(\boldsymbol{A}^{(1)}[\boldsymbol{s}_\ell^\top, (\boldsymbol{c}_\ell^{(1)})^\top]^\top\right)\right|\right] \leq \frac{\sqrt{3C_\phi}}{6}\left(\nu + MC_b\mathfrak{R}_N + \sqrt{2\log(4/\delta)/N}\right)^{1/4} \tag{8}$$

if $\kappa \geq \left(\frac{3}{C_\phi}\right)^{1/4}\left(\nu + 16MC_b\mathfrak{R}_N + 16\sqrt{2\log(4/\delta)/N}\right)^{1/8}$, where $\kappa = \min\{C_p + c_{\ell,i}^{(1)}, C_p - c_{\ell,i}^{(1)}, C_p + c_{\ell,j}^{(1)}, C_p - c_{\ell,j}^{(1)}\}$.

C. Bounding $(h_m^{(q)})''$

By aggregating results (7) and (8), we have the following bound since $\|\cdot\|_2$ is upper bounded by $\|\cdot\|_1$:

$$\mathbb{E}\left[\left\|\boldsymbol{G}^{(1)}(\boldsymbol{h}^{(1)})''\left(\boldsymbol{A}^{(1)}\boldsymbol{z}_\ell^{(1)}\right)\right\|_2^2\right] = O\left(C_\phi\left(\nu + MC_b\mathfrak{R}_N + \sqrt{2\log(4/\delta)/N}\right)^{1/2}\right),$$

where $\boldsymbol{G}^{(1)} \in \mathbb{R}^{\frac{D_1(D_1+1)}{2}\times M}$ is defined as follows

$$\boldsymbol{G}^{(1)} := \begin{bmatrix} \left(\boldsymbol{a}_{D+1}^{(1)} \circledast \boldsymbol{a}_{D+1}^{(1)}\right)^\top \\ \vdots \\ \left(\boldsymbol{a}_{D+D_1}^{(1)} \circledast \boldsymbol{a}_{D+D_1}^{(1)}\right)^\top \\ \left(\boldsymbol{a}_{D+1}^{(1)} \circledast \boldsymbol{a}_{D+2}^{(1)}\right)^\top \\ \vdots \\ \left(\boldsymbol{a}_{D+D_1-1}^{(1)} \circledast \boldsymbol{a}_{D+D_1}^{(1)}\right)^\top \end{bmatrix},$$

and $\circledast$ denotes Hadamard product. Then, one can express

$$\mathbb{E}\left[\left\|(\boldsymbol{h}^{(1)})''\left(\boldsymbol{A}^{(1)}\boldsymbol{z}_\ell^{(1)}\right)\right\|_2^2\right] = O\left(\frac{C_\phi}{\sigma_{\min}^2(\boldsymbol{G}^{(1)})}\left(\nu + MC_b\mathfrak{R}_N + \sqrt{2\log(4/\delta)/N}\right)^{1/2}\right),$$

where the above is because $\boldsymbol{G}^{(1)}$ is not a function of data.

By Chebyshev's inequality, we have the following

$$\Pr\left[\left|\left\|(\boldsymbol{h}^{(1)})''\right\|_2 - \mathbb{E}\left[\left\|(\boldsymbol{h}^{(1)})''\right\|_2\right]\right| < \frac{\sigma}{\alpha}\right] \geq 1-\alpha$$

which means with probability at least $1-\alpha-\delta$

$$\left\|(\boldsymbol{h}^{(1)})''\right\|_2 < \mathbb{E}\left[\left\|(\boldsymbol{h}^{(1)})''\right\|_2\right] + \frac{\sigma}{\alpha} \leq \mathbb{E}\left[\left\|(\boldsymbol{h}^{(1)})''\right\|_1\right] + O\left(\frac{1}{\alpha\sigma_{\min}(\boldsymbol{G}^{(1)})}\left(\nu + MC_b\mathfrak{R}_N + \sqrt{2\log(4/\delta)/N}\right)^{1/4}\right), \tag{9}$$

where the second inequality is because ℓ_1 norm is an upper bound for ℓ_2 norm and by definition of standard deviation

$$\sigma^2 = \mathbb{E}\left[\left\|(\boldsymbol{h}^{(1)})''\right\|_2^2\right] - \mathbb{E}\left[\left\|(\boldsymbol{h}^{(1)})''\right\|_2\right]^2 \leq \mathbb{E}\left[\left\|(\boldsymbol{h}^{(1)})''\right\|_2^2\right] = O\left(\frac{1}{\sigma_{\min}^2(\boldsymbol{G}^{(1)})}\left(\nu + MC_b\mathfrak{R}_N + \sqrt{2\log(4/\delta)/N}\right)^{1/2}\right).$$

D. Final Bound

By denoting $\bar{h}_m^{(q)} = h_m^{(q)}\left((\boldsymbol{a}_m^{(q)})^\top\boldsymbol{0}\right) = h_m^{(q)}(0)$ for short and with Taylor expansion at point $\boldsymbol{z}^{(q)} = \boldsymbol{0}$ with the Lagrange remainder form, we have the following

$$\left[\boldsymbol{f}^{(q)}(\boldsymbol{x}_\ell^{(q)})\right]_m = h_m^{(q)}\left((\boldsymbol{a}_m^{(q)})^\top\boldsymbol{z}_\ell^{(q)}\right) = \bar{h}_m^{(q)} + (\bar{h}_m^{(q)})'t_{m,\ell}^{(q)} + (h_m^{(q)})''(\omega_{m,\ell}^{(q)})\left(t_{m,\ell}^{(q)}\right)^2,$$

where $\omega_{m,\ell}^{(q)} \in \left(-|t_{m,\ell}^{(q)}|, |t_{m,\ell}^{(q)}|\right)$, $t_{m,\ell}^{(q)} = \sum_{j=1}^{D+D_q}a_{m,j}^{(q)}[\boldsymbol{z}_\ell^{(q)}]_j$.

Since $(\boldsymbol{f}, \boldsymbol{B})$ is an optimal solution, we must have

$$1/\sqrt{N}\boldsymbol{B}^{(1)}\boldsymbol{E}^{(1)} = 1/\sqrt{N}\boldsymbol{B}^{(2)}\boldsymbol{E}^{(2)} + \boldsymbol{Q}$$

where $\boldsymbol{Q} \in \mathbb{R}^{D\times N}$ is an error term with $\|\boldsymbol{Q}\|_F^2 \leq \nu$ (by Assumption 4), and $E_{i,j}^{(q)} = \bar{h}_i^{(q)} + (\bar{h}_i^{(q)})'t_{i,j}^{(q)} + (h_i^{(q)})''(t_{i,j}^{(q)})^2$ (with $(h_m^{(q)})'' = (h_m^{(q)})''(\omega_{m,\ell}^{(q)})$ for short). The constant terms $\bar{h}_m^{(q)}$'s should be 0, since we have zero-mean constraint on *seen* samples $\boldsymbol{B}^{(q)}\boldsymbol{f}^{(q)}(\boldsymbol{x}_\ell^{(q)})$ and each dimension of $\boldsymbol{z}_\ell^{(q)}$ is also zero-mean. By rearranging terms, we have

$$\frac{1}{\sqrt{N}}\left(\boldsymbol{B}^{(1)}\boldsymbol{F}^{(1)} - \boldsymbol{B}^{(2)}\boldsymbol{F}^{(2)}\right) = \frac{1}{\sqrt{N}}\left(\boldsymbol{B}^{(2)}\boldsymbol{\Xi}^{(2)} - \boldsymbol{B}^{(1)}\boldsymbol{\Xi}^{(1)}\right) + \boldsymbol{Q},$$

where $\boldsymbol{F}^{(q)} \in \mathbb{R}^{M_q\times N}$, $\boldsymbol{\Xi}^{(q)} \in \mathbb{R}^{M_q\times N}$, $F_{ij}^{(q)} = (\bar{h}_i^{(q)})'t_{i,j}^{(q)}$ and $\Xi_{ij}^{(q)} = (h_i^{(q)})''\left(t_{i,j}^{(q)}\right)^2$, which leads to

$$\frac{1}{\sqrt{N}}(\boldsymbol{B}^{(1)}\underbrace{\boldsymbol{D}_1^{(1)}\boldsymbol{A}^{(1)}}_{\widehat{\boldsymbol{A}}^{(1)}}\boldsymbol{Z}^{(1)} - \boldsymbol{B}^{(2)}\underbrace{\boldsymbol{D}_1^{(2)}\boldsymbol{A}^{(2)}}_{\widehat{\boldsymbol{A}}^{(2)}}\boldsymbol{Z}^{(2)}) = \frac{1}{\sqrt{N}}\boldsymbol{\Gamma} + \boldsymbol{Q}, \tag{10}$$

where we define

$$\boldsymbol{\Gamma} = \boldsymbol{B}^{(2)}\boldsymbol{D}_2^{(2)} \circledast \left(\boldsymbol{A}^{(2)}\boldsymbol{Z}^{(2)}\right)^{\circledast 2} - \boldsymbol{B}^{(1)}\boldsymbol{D}_2^{(1)} \circledast \left(\boldsymbol{A}^{(1)}\boldsymbol{Z}^{(1)}\right)^{\circledast 2},$$

with $\boldsymbol{X}^{\circledast 2}$ denoting element-wise squaring.

Its transposed version is

$$(\widetilde{\boldsymbol{Z}}^{(1)})^\top(\widehat{\boldsymbol{A}}^{(1)})^\top(\boldsymbol{B}^{(1)})^\top - (\widetilde{\boldsymbol{Z}}^{(2)})^\top(\widehat{\boldsymbol{A}}^{(2)})^\top(\boldsymbol{B}^{(2)})^\top = \widetilde{\boldsymbol{\Gamma}}^\top + \boldsymbol{Q}^\top$$

where $\widetilde{\boldsymbol{Z}}^{(q)}, \widetilde{\boldsymbol{\Gamma}}$ are the variables scaled by $\frac{1}{\sqrt{N}}$.

Without loss of generality, we may set $(\boldsymbol{B}^{(q)})^\top = \widehat{\boldsymbol{A}}^{(q)}\left((\widehat{\boldsymbol{A}}^{(q)})^\top\widehat{\boldsymbol{A}}^{(q)}\right)^{-1}\boldsymbol{R}^{(q)}$ where $\boldsymbol{R}^{(q)} \in \mathbb{R}^{(D+D_q)\times D}$. This is because $(\boldsymbol{B}^{(q)})^\top$ can always be decomposed into a component in the subspace spanned by $\widehat{\boldsymbol{A}}^{(q)}$ and one is orthogonal and the latter is always canceled by multiplying with $(\widehat{\boldsymbol{A}}^{(q)})^\top$.

Then, we have the following

$$\boldsymbol{T}\boldsymbol{R} = \widetilde{\boldsymbol{\Gamma}}^\top + \boldsymbol{Q}^\top,$$

where we define

$$\boldsymbol{T} = \frac{1}{\sqrt{N}}[\boldsymbol{S}^\top, (\boldsymbol{C}^{(1)})^\top, (\boldsymbol{C}^{(2)})^\top] \in \mathbb{R}^{N\times(D+D_1+D_2)}$$

$$\boldsymbol{R} = \begin{bmatrix} \boldsymbol{R}^{(1)}(1:D,:) - \boldsymbol{R}^{(2)}(1:D,:) \\ \boldsymbol{R}^{(1)}(D+1:\text{end},:) \\ -\boldsymbol{R}^{(2)}(D+1:\text{end},:) \end{bmatrix} \in \mathbb{R}^{(D+D_1+D_2)\times D}$$

where $\boldsymbol{R}^{(q)}(1:D,:)$ denotes rows 1 to D of $\boldsymbol{R}^{(q)}$.

If matrix $\boldsymbol{T}$ is full column rank, then we have $\boldsymbol{R} = \boldsymbol{T}^\dagger(\widetilde{\boldsymbol{\Gamma}}^\top + \boldsymbol{Q}^\top)$. Thus the ℓ_2-norm of $\boldsymbol{R}$ is bounded as

$$\begin{aligned}\|\boldsymbol{R}\|_2 &\le \|\boldsymbol{T}^\dagger\|_2(\|\widetilde{\boldsymbol{\Gamma}}^\top\|_2 + \|\boldsymbol{Q}^\top\|_2) \\ &\le \|\boldsymbol{T}^\dagger\|_2\left(\sum_{q=1}^{2}\left\|\boldsymbol{D}_2^{(q)} \circledast \left(\boldsymbol{A}^{(q)}\widetilde{\boldsymbol{Z}}^{(q)}\right)^{\circledast 2}\right\|_2 \left\|\boldsymbol{B}^{(q)}\right\|_2 + \|\boldsymbol{Q}\|_2\right).\end{aligned}$$

Note that for Hadamard product, we have the following

$$\|\boldsymbol{A}\circledast\boldsymbol{B}\|_F = \sqrt{\sum_{i,j}a_{ij}^2b_{ij}^2} \le \max_{i,j}(|a_{ij}|)\sqrt{\sum_{i,j}b_{ij}^2} = \|\boldsymbol{A}\|_{\max}\|\boldsymbol{B}\|_F.$$

Thus, by defining $\boldsymbol{\Psi}^{(q)} = \left(\boldsymbol{A}^{(q)}\widetilde{\boldsymbol{Z}}^{(q)}\right)^{\circledast 2}$ we have

$$\begin{aligned}\|\boldsymbol{R}\|_2 &\le \|\boldsymbol{T}^\dagger\|_2\left(\sum_{q=1}^{2}\left\|\boldsymbol{D}_2^{(q)}\right\|_{\max}\left\|\boldsymbol{\Psi}^{(q)}\right\|_F \left\|\boldsymbol{B}^{(q)}\right\|_2 + \|\boldsymbol{Q}\|_2\right) \\ &\le \|\boldsymbol{T}^\dagger\|_2\left(\sum_{q=1}^{2}\left\|\boldsymbol{D}_2^{(q)}\right\|_{\max}\|\boldsymbol{A}^{(q)}\boldsymbol{Z}^{(q)}\|_{\max}(D+D_q)\right. \\ &\qquad \left.\|\boldsymbol{A}^{(q)}\|_2\|\widetilde{\boldsymbol{Z}}^{(q)}\|_2\left\|\boldsymbol{B}^{(q)}\right\|_2 + \|\boldsymbol{Q}\|_2\right),\end{aligned}$$

where

$$\begin{aligned}\left\|\boldsymbol{\Psi}^{(q)}\right\|_F &\le \|\boldsymbol{A}^{(q)}\widetilde{\boldsymbol{Z}}^{(q)}\|_{\max}\|\boldsymbol{A}^{(q)}\|_F\|\widetilde{\boldsymbol{Z}}^{(q)}\|_F, \\ \|\boldsymbol{A}^{(q)}\|_F &\le \sqrt{D+D_q}\|\boldsymbol{A}^{(q)}\|_2, \\ \|\widetilde{\boldsymbol{Z}}^{(q)}\|_F &\le \sqrt{D+D_q}\|\widetilde{\boldsymbol{Z}}^{(q)}\|_2, \\ \|\boldsymbol{Q}\|_2 &\le \|\boldsymbol{Q}\|_F \le \sqrt{\nu},\end{aligned}$$

with the rank of $\boldsymbol{A}^{(q)}$ and $\boldsymbol{Z}^{(q)}$ assumed $D+D_q$ and we assume $\|\boldsymbol{T}^\dagger\|_2 \le C_T, \|\widetilde{\boldsymbol{Z}}^{(q)}\|_2 \le C_Z$ since $\mathcal{S}, \mathcal{C}_q$ are bounded sets, $\|\boldsymbol{A}^{(q)}\|_2 \le C_A$ since $\boldsymbol{A}^{(q)}$ is fixed and $\|\boldsymbol{B}^{(q)}\|_2 \le C_B$ because we have orthogonality constraint (3b).

Combining with (9), we have the following bound with probability at least $1-\alpha-\delta$:

$$\|\boldsymbol{R}\|_2 \le O\left(\frac{D+D_1}{\alpha}\left(\nu + M\mathfrak{R}_N + \sqrt{2\log(4/\delta)/N}\right)^{1/4} + \sqrt{\nu}\right),$$

which implies that if N is sufficiently large and ν is small, then $\boldsymbol{R}$ is close to $\boldsymbol{0}$. If we define

$$\boldsymbol{R}^{(q)}(1:D,:) = \boldsymbol{\Theta}^{(q)},\ \boldsymbol{R}^{(q)}(D+1:\text{end},:) = \boldsymbol{\Omega}^{(q)},$$

the above means

$$\left\|\begin{bmatrix}\boldsymbol{\Theta}^{(1)} - \boldsymbol{\Theta}^{(2)} \\ \boldsymbol{\Omega}^{(1)} \\ -\boldsymbol{\Omega}^{(2)}\end{bmatrix}\right\|_F = O\left(\frac{D(D+D_1)}{\alpha}\left(\nu + M\mathfrak{R}_N + \sqrt{2\log(4/\delta)/N}\right)^{1/4} + \sqrt{\nu}\right),$$

when the rank of $\boldsymbol{R}$ is D. Then we have the following for $q = 1, 2$

$$\left\|\boldsymbol{\Omega}^{(q)}\right\|_F = O\left(\frac{D(D+D_1)}{\alpha}\left(M\mathfrak{R}_N + \sqrt{2\log(4/\delta)/N}\right)^{1/4} + \sqrt{\nu}\right),$$

which completes the proof.

V. Experimental Results

In this section, we present both synthetic and real data experiments. We should mention that the work [1] has extensive simulations and real data experiments to showcase the effectiveness of the learning criterion of interest—and the readers are referred to [1] for details. We will only focus on validating the insights revealed by Theorem 2. We use the optimization algorithm for tackling (4) from [1], which is available on the authors' websites.

By Theorem 2, the composition $\widehat{f}_m^{(q)} \circ g_m^{(q)}$ should be approximately affine for all m and q, if the neural network structure is appropriately chosen under a fixed N. To be more precise, if the neural networks representing the learning functions have one hidden layer, then the number of neurons R has to strike a fine balance. Specifically, when N is fixed, increasing R will help reduce the representation error ν, which improves the performance. However, increasing R also enlarges the Rademacher complexity $\mathfrak{R}_N = O(\sqrt{R/N})$, which may make the performance worse. Hence, under a finite N, one hopes to use expressive enough neural networks with a sufficiently large R, but not an overly complex neural network with an exceedingly large R. In this section, we will validate this claim.

A. Synthetic Data

In this subsection, we validate the theorem using synthetic data.

1) Data Generation: We use a similar data generation model as in [1]. Specifically, we set the shared components to be $\boldsymbol{S} = [\boldsymbol{s}_1, \ldots, \boldsymbol{s}_N] \in \mathbb{R}^{2\times N}$ are sampled from a parabola [i.e., $s_{2,\ell} = (s_{1,\ell})^2$, where $s_{1,\ell} \in (-1, 1)$]. The view-specific components $\boldsymbol{c}_\ell^{(q)} \in \mathbb{R}^3$ for $q = 1, 2$ are set to be i.i.d. Gaussian components with mean$= -0.5$/variance$= 1$ and mean$=0.8$/variance$=1.5^2$, respectively. We set $M_1 = M_2 = 5$.

TABLE I
GENERATIVE FUNCTIONS USED FOR THE SYNTHETIC DATA SIMULATION.

	Generative function
First view	$g_1^{(1)}(x) = 3\text{sigmoid}(x) + 0.1x$
	$g_2^{(1)}(x) = 5\text{sigmoid}(x) + 0.2x$
	$g_3^{(1)}(x) = 0.2\exp(x)$
	$g_4^{(1)}(x) = -4\text{sigmoid}(x) - 0.3x$
	$g_5^{(1)}(x) = -3\text{sigmoid}(x) - 0.2x$
Second view	$g_1^{(2)}(x) = 5\tanh(x) + 0.2x$
	$g_2^{(2)}(x) = 2\tanh(x) + 0.1x$
	$g_3^{(2)}(x) = 0.1x^3 + x$
	$g_4^{(2)}(x) = -5\tanh(x) - 0.4x$
	$g_5^{(2)}(x) = -6\tanh(x) - 0.3x$

The mixing matrices $\boldsymbol{A}^{(1)}, \boldsymbol{A}^{(2)} \in \mathbb{R}^{5\times 5}$ follow the zero-mean unit-variance i.i.d. Gaussian distribution. The nonlinear generative functions are listed in Tab. I.

2) Performance Metric: By Theorem 2, the nonlinear multiview analysis method recovers the range space of $\boldsymbol{S}^\top$ plus an additional noise term determined by $\boldsymbol{\Omega}^{(q)}$ and $\boldsymbol{C}^{(q)}$. To quantitatively evaluate the performance of recovery of the shared components, we employ the *subspace distance* measure as in [1]:

$$\text{dist}(\mathcal{S}, \widehat{\mathcal{S}}) = \|\boldsymbol{P}_{\boldsymbol{s}}^{\perp} \boldsymbol{Q}_{\widehat{\boldsymbol{s}}}^{\top}\|_2$$

as the performance metric, where $\mathcal{S} = \text{range}(\boldsymbol{S}^\top)$ and $\widehat{\mathcal{S}} = \text{range}(\widehat{\boldsymbol{S}}^\top)$, $\boldsymbol{P}_{\boldsymbol{s}}^{\perp}$ is defined as $\boldsymbol{P}_{\boldsymbol{s}}^{\perp} = \boldsymbol{I} - \boldsymbol{S}^\top(\boldsymbol{S}\boldsymbol{S}^\top)^{-1}\boldsymbol{S}$, and $\boldsymbol{Q}_{\widehat{\boldsymbol{s}}}$ is the orthogonal basis of $\widehat{\boldsymbol{S}}$. This metric is in between 0 and 1. Ideally, if the energy of $\boldsymbol{\Omega}^{(q)}$ in Theorem 2 is small enough, then the noise term is negligible, which means that the subspace of $\boldsymbol{S}^\top$ is well recovered. We should see $\text{dist}(\mathcal{S}, \widehat{\mathcal{S}}) \approx 0$ in this case. This provides a measure of the "size" of the residual $\boldsymbol{\Omega}^{(q)}$.

3) Results: Fig. 1 shows the learned composition functions $\widehat{f}_m^{(q)} \circ g_m^{(q)}$. Here, we use a one-hidden-layer neural network with R neurons to model each $f_m^{(q)}$. We first observe how the performance changes when the neural network's complexity changes by fixing the sample size to be $N = 2,000$ and varying $R \in \{8, 16, 32, 64, 128, 1024\}$. One can see that the composition function $\widehat{f}_1^{(1)} \circ g_1^{(1)}$ becomes increasingly closer to an affine function from $R = 8$ to $R = 128$, with $R = 128$ giving the best result. However, the result of $R = 1024$ is obviously worse than that of $R = 128$. This validates Theorem 2: although increasing the complexity of the function class will decrease the realization gap ν, a too large R will lead to performance degradation under fixed sample size due to the increase of $\mathfrak{R}_N$.

Fig. 2 gets a closer look at the relationship between R and the nonlinearity removal performance under different N's. Specifically, using the learned functions under the settings of the previous simulations, we plot the subspace distance computed over a test set of 1,000 samples. The results are averaged over 10 different random trials. It is clear that the subspace distance measure decreases when N gets larger over all different R. More importantly, given any fixed N, the performance is a trade-off between ν and R. When R is too small, the function is not powerful enough to model

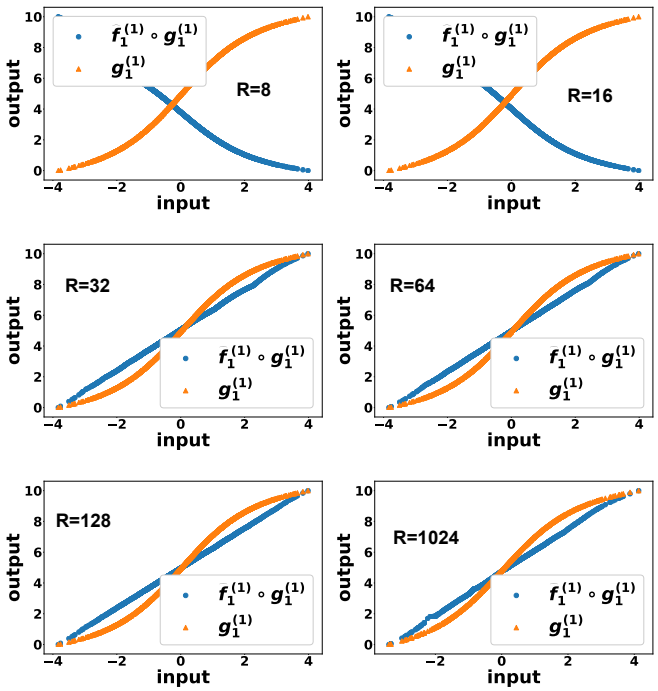

Fig. 1. Learned composition functions by method in [1] with varying R and fixed $N = 2000$.

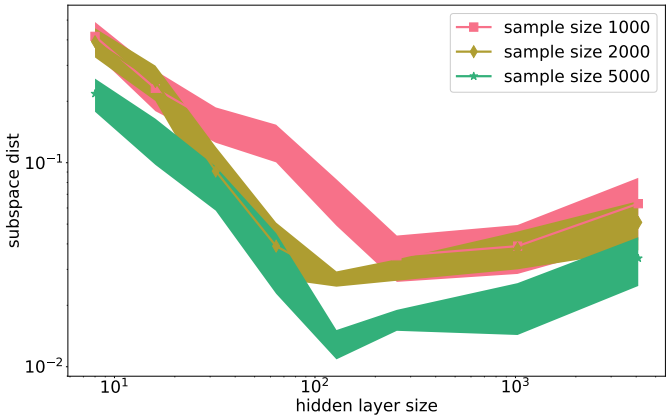

Fig. 2. Subspace distances under different network structures and various sample size for method proposed in [1]

the nonlinear transformation, which leads to a large ν in Theorem 2. Thus, the performance is far from satisfactory. Increasing the hidden size R improves the performance. For example, when $N = 1,000$, the subspace distance is 0.23 when $R = 16$. It is reduced to 0.04 when $R = 256$. However, after a certain point, the performance starts to deteriorate again. This is because that the complexity of the function class, i.e., $\mathfrak{R}_N = O(\sqrt{R/N})$, becomes dominant in this case. Hence, an appropriate function class should be expressive enough but not overly complex. This observation is consistent with what we proved in Theorem 2.

B. Real Data

Following the real data experiment in [1], we use the *Multiview Digit Dataset* [22] which consists of low-level

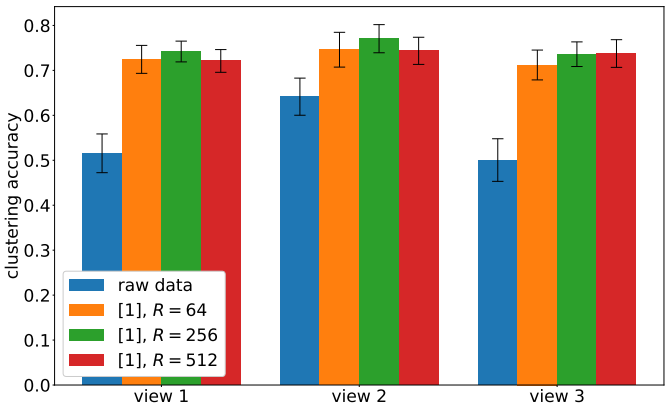

Fig. 3. Clustering accuracy for each view on the Multiview Digit Dataset [22], with different hidden size R by method in [1] and on the raw data.

features of handwritten digits 0 to 9. The three views include 76 Fourier coefficients of the character shapes, 64 Karhunen-Loève coefficients, and 47 Zernike moments, respectively. The formulation proposed in [1] is able to handle multiple modalities of inputs, sames as those in [23]–[25], by introducing a slack variable. We split the dataset into 1,200/400/400 for training/validation/testing, respectively. To measure the performance, we use spectral clustering [26] to perform a downstream clustering task. In particular, after learning the $\boldsymbol{f}^{(q)}$ for each view, we do clustering on $\boldsymbol{f}^{(q)}(\boldsymbol{x}_\ell^{(q)})$ for all $q \in \{1,2,3\}$ and ℓ over all the data samples and then report the accuracy on the test set. The results are averaged over 5 random trials. The parameter settings follow that in [1].

The clustering accuracy results on the testing set are plotted in Fig. 3. Note that by clustering on the raw features, the performance is around 50%. By applying the multiview approach in [1], the accuracy increases substantially. Nonetheless, we also observe similar trade-offs as in the simulations. In particular, the accuracy is highest when $R = 256$. In comparison, when $R = 64$ and $R = 512$ the accuracy both decrease to certain extents (except for view 3, which essentially outputs the same accuracy under $R = 128$ and $R = 512$). This is probably because that $R = 64$ is not powerful enough while $R = 512$ starts to be overly complex. This again corroborates our claim in Theorem 2.

VI. Conclusions

In this work, we presented finite-sample analysis of a `DeepCCA` formulation for identifying the post-nonlinear multimodal model. Our result filled a critical gap between the previous model identification theorem that relies on unlimited data and the practical cases where only finite samples are available. Our analytical approach is an integration of statistical learning, numerical differentiation, and robust system identification. This framework is bound to finding applications in a wider spectrum of sample complexity problems associated with nonlinear unsupervised learning. The synthetic and real data experiment results corroborate our finite-sample analysis.

References

[1] Q. Lyu and X. Fu, "Nonlinear multiview analysis: Identifiability and neural network-assisted implementation," *IEEE Trans. Signal Process.*, vol. 68, pp. 2697–2712, 2020.

[2] H. Hotelling, "Relations between two sets of variates," *Biometrika*, vol. 28, no. 3/4, pp. 321–377, 1936.

[3] D. R. Hardoon, S. Szedmak, and J. Shawe-Taylor, "Canonical correlation analysis: An overview with application to learning methods," *Neural Computation*, vol. 16, no. 12, pp. 2639–2664, 2004.

[4] F. R. Bach and M. I. Jordan, "A probabilistic interpretation of canonical correlation analysis," 2005.

[5] M. S. Ibrahim and N. D. Sidiropoulos, "Reliable detection of unknown cell-edge users via canonical correlation analysis," *IEEE Trans. Wireless Commun.*, vol. 19, no. 6, pp. 4170–4182, 2020.

[6] G. Andrew, R. Arora, J. Bilmes, and K. Livescu, "Deep canonical correlation analysis," in *International conference on machine learning*, vol. 28, no. 3, 17–19 Jun 2013, pp. 1247–1255.

[7] W. Wang, R. Arora, K. Livescu, and J. Bilmes, "On deep multi-view representation learning," in *Proceedings of ICML*, 2015, pp. 1083–1092.

[8] A. Taleb and C. Jutten, "Source separation in post-nonlinear mixtures," *IEEE Trans. Signal Process.*, vol. 47, no. 10, pp. 2807–2820, 1999.

[9] F. Oveisi, "EEG signal classification using nonlinear Independent Component Analysis," in *Proc. IEEE ICASSP 2009*, April 2009, pp. 361–364.

[10] K. Zhang and A. Hyvarinen, "On the identifiability of the post-nonlinear causal model," *arXiv preprint arXiv:1205.2599*, 2012.

[11] S. Achard and C. Jutten, "Identifiability of post-nonlinear mixtures," *IEEE Signal Process. Lett.*, vol. 12, no. 5, pp. 423–426, 2005.

[12] A. Hyvarinen, H. Sasaki, and R. Turner, "Nonlinear ICA using auxiliary variables and generalized contrastive learning," in *International Conference on Artificial Intelligence and Statistics*, 2019, pp. 859–868.

[13] I. Khemakhem, D. Kingma, R. Monti, and A. Hyvarinen, "Variational autoencoders and nonlinear ICA: A unifying framework," in *International Conference on Artificial Intelligence and Statistics*. PMLR, 2020, pp. 2207–2217.

[14] L. Gresele, P. K. Rubenstein, A. Mehrjou, F. Locatello, and B. Schölkopf, "The incomplete Rosetta stone problem: Identifiability results for multi-view nonlinear ICA," in *Proceedings of Uncertainty in Artificial Intelligence*, 2020, pp. 217–227.

[15] X. Fu, K. Huang, N. D. Sidiropoulos, and W.-K. Ma, "Nonnegative matrix factorization for signal and data analytics: Identifiability, Algorithms, and Applications," *IEEE Signal Process. Mag.*, vol. 36, no. 2, pp. 59–80, March 2019.

[16] N. Gillis, *Nonnegative Matrix Factorization*. SIAM, 2020.

[17] E. Oja, "The nonlinear PCA learning rule in Independent Component Analysis," *Neurocomputing*, vol. 17, no. 1, pp. 25–45, 1997.

[18] Y. Altmann, A. Halimi, N. Dobigeon, and J.-Y. Tourneret, "A post nonlinear mixing model for hyperspectral images unmixing," in *2011 IEEE International Geoscience and Remote Sensing Symposium*, 2011, pp. 1882–1885.

[19] Q. Lyu and X. Fu, "Identifiability-guaranteed simplex-structured post-nonlinear mixture learning via autoencoder," *IEEE Trans. Signal Process.*, 2021.

[20] D. Rezende and S. Mohamed, "Variational inference with normalizing flows," in *International conference on machine learning*. PMLR, 2015, pp. 1530–1538.

[21] S. Shalev-Shwartz and S. Ben-David, *Understanding machine learning: From theory to algorithms*. Cambridge university press, 2014.

[22] M. van Breukelen, R. P. Duin, D. M. Tax, and J. Den Hartog, "Handwritten digit recognition by combined classifiers," *Kybernetika*, vol. 34, no. 4, pp. 381–386, 1998.

[23] A. Benton, H. Khayrallah, B. Gujral, D. Reisinger, S. Zhang, and R. Arora, "Deep generalized canonical correlation analysis," in *RepL4NLP at ACL*, 2017.

[24] C. Tang, X. Zhu, X. Liu, M. Li, P. Wang, C. Zhang, and L. Wang, "Learning a joint affinity graph for multiview subspace clustering," *IEEE Transactions on Multimedia*, vol. 21, no. 7, pp. 1724–1736, 2018.

[25] C. Tang, X. Zheng, X. Liu, W. Zhang, J. Zhang, J. Xiong, and L. Wang, "Cross-view locality preserved diversity and consensus learning for multi-view unsupervised feature selection," *IEEE Transactions on Knowledge and Data Engineering*, 2021.

[26] A. Y. Ng, M. I. Jordan, and Y. Weiss, "On spectral clustering: Analysis and an algorithm," in *Proceedings of NIPS 2002*, 2002, pp. 849–856.