Understanding Multimodal Contrastive Learning and Incorporating Unpaired Data

Ryumei Nakada
Rutgers University

Halil Ibrahim Gulluk
Stanford University

Zhun Deng
Columbia University

Wenlong Ji
Stanford University

James Zou
Stanford University

Linjun Zhang
Rutgers University

Abstract

Language-supervised vision models have recently attracted great attention in computer vision. A common approach to build such models is to use contrastive learning on paired data across the two modalities, as exemplified by Contrastive Language-Image Pre-Training (CLIP). In this paper, under linear representation settings, (i) we initiate the investigation of a general class of nonlinear loss functions for multimodal contrastive learning (MMCL) including CLIP loss and show its connection to singular value decomposition (SVD). Namely, we show that each step of loss minimization by gradient descent can be seen as performing SVD on a contrastive cross-covariance matrix. Based on this insight, (ii) we analyze the performance of MMCL. We quantitatively show that the feature learning ability of MMCL can be better than that of unimodal contrastive learning applied to each modality even under the presence of wrongly matched pairs. This characterizes the robustness of MMCL to noisy data. Furthermore, when we have access to additional unpaired data, (iii) we propose a new MMCL loss that incorporates additional unpaired datasets. We show that the algorithm can detect the ground-truth pairs and improve performance by fully exploiting unpaired datasets. The performance of the proposed algorithm was verified by numerical experiments.

Proceedings of the 26[th] International Conference on Artificial Intelligence and Statistics (AISTATS) 2023, Valencia, Spain. PMLR: Volume 206.

1 Introduction

Multimodal learning is a broad class of machine learning algorithms that take advantage of the association of multiple modalities such as text, image, and audio. As the technology of both data collection and sensors advances, we have growing access to data with multiple modes. It has a wide range of applications, including media description, stock return prediction, and drug discovery (Baltrušaitis et al., 2018; Lee and Yoo, 2020; Hu et al., 2021).

Focusing on the research of vision-language models, there have been many breakthroughs in large-scale vision-language pre-training methods (Li et al., 2019; Lu et al., 2019; Tan and Bansal, 2019; Li et al., 2020; Radford et al., 2021; Jia et al., 2021; Li et al., 2022; Yao et al., 2022b; Du et al., 2022). One of the vision-language models is Contrastive Language-Image Pre-training (CLIP) (Radford et al., 2021). Through contrastive loss, CLIP trains dual encoders in the shared representation space by maximizing the similarity of the observed pairs of text and images while minimizing the similarity of the artificially paired data. Through the flexibility of its architecture, CLIP successfully achieves outstanding zero-shot learning performance on ImageNet, outperforming other few-shot linear probes of BiT-M, SimCLRv2, and ResNet50 (Radford et al., 2021). CLIP and its successors are widely used, for example, in semantic segmentation, image generation from captions, and video summarization (Galatolo. et al., 2021; Narasimhan et al., 2021; Li et al., 2021; Xu et al., 2022; Wang et al., 2022).

Despite the great success of multimodal contrastive learning (MMCL), the theoretical understanding of MMCL is still limited. From the perspective of multimodal learning, it has been empirically (Ngiam et al., 2011; Radford et al., 2021) and theoretically (Zadeh et al., 2020; Huang et al., 2021b) shown that the use of multimodal data produces a better representation compared to the use of unimodal data when focused on a single modality. Namely, Huang et al.

(2021b) showed that the difference in downstream task performance of multimodal learning using different subset of modalities depends on the term named latent representation quality. In particular, they showed that multimodal learning using smaller subset of modalities can perform worse than multimodal learning with a full set of modalities under linear representation settings. However, the feature recovery performance of multimodal learning, as well as the comparison with that of unimodal contrastive learning, have not been considered in previous works.

Additionally, a practical issue in multimodal learning is the problem of noisy pairs; the collected raw data may not be correctly aligned due to an error in the data collection procedure. For example, (Radford et al., 2021; Jia et al., 2021) uses a dataset collected from various public sources on the Internet and feeds the collected data directly into the algorithm without cleaning up the noisy association. However, the quality of association between images and word queries used for the search depends on the context of the website, which possibly leads to incorrect alignment of images and words in the collected dataset. To avoid this problem, Li et al. (2020) proposed OSCAR that detects object tags in images and uses them as anchor points for alignment. Although the problem has been recognized in the literature, the analysis of the effect of noisy pairs in MMCL has not been addressed.

Furthermore, multimodal learning requires datasets with specified pair information among modalities. However, data collection procedures can be very expensive in practice. If we can combine abundant unpaired dataset in a semi-supervised manner, we can expect to improve the quality of representation learning with lower cost.

The purpose of this paper is to provide insights on the feature learning ability of MMCL. We summarize **our contributions** as follows. (i) We establish a connection between the general multimodal contrastive loss and the SVD analysis. Namely, assuming that representations are linear, we show that the gradient descent of minimizing multimodal contrastive loss function is equivalent to the gradient ascent of the SVD objective function with contrastive cross-covariance matrix. (ii) We analyze the learning capacity of MMCL under linear loss when there are noisy paired data and the pairs are assumed to be one-to-one. We show that as long as the observed pairs contain an inignorable portion of ground-truth pairs, MMCL can recover the core features with a parametric rate. However, in practice, many-to-many correspondences between modalities are often observed. We showed by real-data analysis that cleaning up pairs by Bipartite Spectral Graph Multi-Partitioning (Dhillon, 2001) improves the performance of learned representations. (iii) We propose a method that incorporates unpaired data and improves the performance of MMCL in a linear representation setting. The theoretical concept was verified by a numerical experiment.

The outline of this paper is as follows. Section 2 establishes the connection between MMCL and SVD. Based on the results of Section 2, in Section 3, we provide a theoretical analysis of MMCL using linear loss on feature learning ability. In Section 4, we discuss the possible improvement of MMCL when additional unpaired data is available. In Section 5, we numerically verify the result of Section 4 that the performance of MMCL improves with additional unpaired data. We also conduct a real-data analysis that deals with many-to-many correspondences of multimodal data. We discuss and conclude our results in Section 6.

1.1 Related Works

Multimodal Learning Due to its applicability and generality, there has been a large amount of literature on multimodal learning since the 1980s (Yuhas et al., 1989). Recently, the development of deep learning brought many advances in multimodal representation learning (Sun et al., 2020). Especially, Ngiam et al. (2011); Srivastava and Salakhutdinov (2012) proposed multimodal learning algorithms to obtain joint representations. Multimodal contrastive representation learning and generative models are also proposed (Shi et al., 2020; Yuan et al., 2021; Radford et al., 2021; Jia et al., 2021). The missing modality problem has been addressed by Ma et al. (2021b,a). From a theoretical point of view, multimodal learning has been shown to outperform unimodal learning focused on one modality (Zadeh et al., 2020; Subramanian et al., 2021; Huang et al., 2021b). For an overview of multimodal learning, see Baltrušaitis et al. (2018); Zhang et al. (2020); Xu et al. (2022); Liang et al. (2022).

Self-Supervised Contrastive Learning Another closely related line of research is unimodal self-supervised contrastive learning (SSCL) for unimodal data. SSCL is a group of self-supervised learning algorithms that learn representations by contrasting two views generated by data augmentation. It has gained popularity in computer vision, natural language processing, and graph learning (Jaiswal et al., 2020; Liu et al., 2021). In particular, Chen et al. (2020b) proposed SimCLR, which uses contrastive loss to train encoders so that they maximize the similarity of similar views generated by data augmentation and minimize the similarity of unrelated views. There have been many works on the theoretical guarantee of SSCL (Saunshi et al., 2019; Wang and Isola, 2020; Ash et al., 2021; Wen and Li, 2021; Huang et al., 2021a; Ji et al., 2021; Tian, 2022; Saunshi et al., 2022). In particular, Wen and Li (2021) proved that contrastive learning using shallow neural networks with appropriate data augmentation can learn the sparse signal despite the presence of spurious noise. Tian (2022) analyzed unimodal contrastive learning and showed that gradient descent applied to nonlinear contrastive loss can be interpreted as gradient ascent of PCA objective function

under game-theoretical formulation. Ji et al. (2021) proved that contrastive learning is equivalent to diagonal-deletion PCA under linear loss settings. They also showed the superiority of contrastive learning over autoencoders under constant signal-to-noise regime. Ko et al. (2022) established a connection between contrastive learning and neighborhood component analysis (Goldberger et al., 2004) which learns Mahalanobis distance metrics.

SVD analysis and CCA The goal of SVD analysis is to find projections that maximize the variance between two projected variables. A closely related method is canonical correlation analysis (CCA) (Harold, 1936; Kettenring, 1971). In CCA, the goal is to find linear projections such that *correlation* between two projected variables is maximized, so that the learned projections fully exploit the association of two datasets. To deal with nonlinear data sets, artificial neural networks were applied to transform data (Lai and Fyfe, 1998, 1999) and kernels were used to allow flexibility in representation space (Akaho, 2006; Hardoon et al., 2004). In particular, Deep CCA (Andrew et al., 2013; Benton et al., 2017; Wang et al., 2016) learns nonlinear embeddings using deep neural networks. Deep CCA was shown to identify latent variables shared between multiple modalities (Lyu and Fu, 2020, 2022). For an overview of CCA-related methods, see Yang et al. (2019).

1.2 Notation

For two sequences of positive numbers $(a_k)_k$ and $(b_k)_k$ indexed by $k \in \mathcal{K}$, we write $a_k \lesssim b_k$ if and only if there exists a constant $C > 0$ independent of the index k such that $\sup_{k\in\mathcal{K}} a_k/b_k < C$ holds. We also write $a_k = O(b_k)$ if $a_k \lesssim b_k$ holds and $a_k = \Omega(b_k)$ if $a_k \gtrsim b_k$ holds. We write $a_k \asymp b_k$ when $a_k \lesssim b_k$ and $a_k \gtrsim b_k$ holds simultaneously. For any matrix A, let $\|A\|$ and $\|A\|_F$ denote the operator norm and Frobenius norm of A, respectively. $\mathbb{O}_{d,r} \triangleq \{O \in \mathbb{R}^{r\times d} : O^\top O = I_r\}$ is a set of orthogonal matrices of order $d \times r$. For any positive integer I, let $[I] = \{1, 2, \cdots, I\}$. We write $a \vee b$ and $a \wedge b$ to denote $\max(a, b)$ and $\min(a, b)$, respectively. For any matrix A, let $P_r(A)$ be the top-r right singular vectors of A. When the right singular vectors are not unique, we choose arbitrary singular vectors. For any matrix A, let $\lambda_j(A)$ be the j-th largest singular value of A. Let $\lambda_{\min}(A)$ and $\lambda_{\max}(A)$ be the minimum and maximum singular values of A, respectively. For any mean zero random variables X and $\tilde{X}$, we define the covariance matrix of X as $\Sigma_X \triangleq \mathbb{E}[XX^\top]$, and the cross-covariance matrix of X and $\tilde{X}$ as $\Sigma_{X,\tilde{X}} \triangleq \mathbb{E}[X\tilde{X}^\top]$. For any square matrix A, define its effective rank $r(A)$ as $r(A) = \mathrm{Tr}(A)/\|A\|$.

2 Multimodal Contrastive Learning and SVD

In this section, we establish the connection between MMCL and SVD. In the following sections, we focus on MMCL with two-modality data.

Suppose that we have n pairs of observations $\{(x_i, \tilde{x}_i)\}_{i=1}^n \subset \mathbb{R}^{d_1+d_2}$, where $x_i \in \mathbb{R}^{d_1}$ and $\tilde{x}_i \in \mathbb{R}^{d_2}$. The multimodal contrastive loss maximizes the similarity of observed pairs, while minimizing the similarity of generated pairs to learn the encoders $g_1 : \mathbb{R}^{d_1} \to \mathbb{R}^r$ and $g_2 : \mathbb{R}^{d_2} \to \mathbb{R}^r$ that share the same representation space. As in the previous literature, we adapt **inner product** of the representation space as a measure of the similarity of two representations for theoretical brevity; Given two encoders g_1 and g_2 for each modality, we measure the similarity of the pair $(x, \tilde{x})$ by $\langle g_1(x), g_2(\tilde{x})\rangle$. This inner product measure has been widely used in the literature (He et al., 2020; Ji et al., 2021; Wang and Liu, 2021; Radford et al., 2021; Jia et al., 2021).

In this paper, we consider **linear representation settings**, that is, $g_1(x) = G_1 x$ and $g_2(\tilde{x}) = G_2\tilde{x}$ for $G_1 \in \mathbb{R}^{r\times d_1}$ and $G_2 \in \mathbb{R}^{r\times d_2}$. The linear representation setting has been widely adapted in the machine learning literature (Jing et al., 2021; Tian et al., 2021; Ji et al., 2021; Wu et al., 2022; Tian, 2022).

2.1 Minimizing Nonlinear Loss via Gradient Descent

Here, we consider a general class of nonlinear loss functions[1], which includes linearized loss, CLIP loss (Radford et al., 2021) or ALIGN loss (Jia et al., 2021). Let $\phi, \psi : \mathbb{R} \to \mathbb{R}$ be differentiable and non-decreasing smooth functions. The non-decreasing property of ϕ and ψ ensures that the loss becomes small when the encoders align only with observed pairs. Define the loss function as follows:

$$\mathcal{L}(G_1, G_2) \triangleq \frac{1}{2C_n}\sum_i \phi\left(\epsilon\psi(0) + \sum_{j:j\neq i}\psi(s_{ij} - s_{ii})\right) + \frac{1}{2C_n}\sum_i \phi\left(\epsilon\psi(0) + \sum_{j:j\neq i}\psi(s_{ji} - s_{ii})\right) + R(G_1, G_2), \tag{2.1}$$

where $s_{ij} \triangleq \langle G_1 x_i, G_2\tilde{x}_j\rangle$, $\epsilon \geq 0$, C_n is a normalizing constant depending only on n, and $R(G_1, G_2)$ is a sufficiently smooth regularization term. We note that regularization techniques have been widely adapted in unimodal SSCL practice (Chen et al., 2020a; He et al., 2020; Grill et al., 2020).

We consider gradient descent as an optimization method under linear representation settings. The following propo-

[1] A similar class of loss functions in SSCL is considered in Tian (2022), where the similarity is measured for augmented views.

sition states that the gradient of loss equation 2.1 with respect to G_k equals to the negative gradient of the SVD objective function minus the penalty term. Thus, if we optimize the loss in equation 2.1 via gradient descent, the search direction of the parameter is exactly the direction that maximizes the SVD objective function with the contrastive cross-covariance matrix[2]. A similar result holds for smooth nonlinear representations, which is deferred to Appendix 8.1.

Proposition 2.1 (Informal)**.** *Let $\beta = \beta(G_1, G_2) \triangleq ((\beta_i)_i, (\beta_{ij})_{i,j})$, where β_i and β_{ij} also depend on the choice of ϕ, ψ, ϵ and ν. Define the contrastive cross-covariance $S(\beta) \triangleq C_n^{-1}\sum_{i=1}^n \beta_i x_i \tilde{x}_i^\top - C_n^{-1}\sum_{i\neq j}\beta_{ij} x_i \tilde{x}_j^\top$. Consider minimizing the nonlinear loss function $\mathcal{L}$ defined above. Then, for $k \in \{1, 2\}$,*

$$\frac{\partial \mathcal{L}}{\partial G_k} = -\left.\frac{\partial \operatorname{tr}\big(G_1 S(\beta) G_2^\top\big)}{\partial G_k}\right|_{\beta=\beta(G_1,G_2)} + \frac{\partial R(G_1, G_2)}{\partial G_k}.$$

The formula of β_i and β_{ij} is available in Appendix 8.1. In the case of (unimodal) SSCL, it has been shown that gradient descent of minimizing the contrastive loss is equivalent to the gradient ascent of the PCA objective function, where the target matrix to apply PCA is given by the contrastive covariance matrix (Tian, 2022). We can consider Proposition 2.1 as an analogy to this result.

Remark 2.1. If $C_n = n(n-1)$ and the loss function is linear, that is, ϕ and ψ are identity functions, then $S = (1/n)\sum_i x_i\tilde{x}_i^\top - 1/(n(n-1))\sum_{i\neq j} x_i\tilde{x}_j^\top = 1/(n-1)\sum_i (x_i - \bar{x})(\tilde{x}_i - \bar{\tilde{x}})^\top$, which is the centered cross-covariance matrix of x and $\tilde{x}$. For InfoNCE loss, $C_n = n$ and ϕ and ψ are set to $\phi(x) = \tau\log(x)$ and $\psi(x) = \exp(x/\tau)$ for some $\epsilon \geq 0$, where $\tau > 0$ is the temperature parameter. Setting $\epsilon = 1$ gives the CLIP and ALIGN loss.

To encourage the encoders to learn diverse features and prevent the collapse of representations, we simultaneously regularize by $\operatorname{tr}\big(G_1 G_1^\top G_2 G_2^\top\big)$. A similar penalty has been considered in unimodal SSCL (Ji et al., 2021). This is especially beneficial when the loss is linear, that is, ϕ and ψ are identity functions, since we can easily make the first two terms in equation 2.1 very small by choosing large G_1 and G_2. For this reason, we consider the regularization term $R(G_1, G_2) = (\rho/2)\|G_1^\top G_2\|_F^2$ for $\rho > 0$. For this regularization, we have the following result, which directly follows from Eckart-Young-Mirsky theorem.

Lemma 2.1. *Fix any $A \in \mathbb{R}^{d_1\times d_2}$ and $\rho > 0$. Let the SVD of A be $\sum_{j=1}^d c_j U_{1,j} U_{2,j}^\top$, where d is the rank of the sum, $c_1 \geq c_2 \geq \cdots \geq c_d > 0$ and $(U_{1,1}, \ldots, U_{1,d}), (U_{2,1}, \ldots, U_{2,d}) \in \mathbb{O}_{r,d}$. Then,*

$$\left\{(G_1, G_2) \in \mathbb{R}^{r\times d_1}\times\mathbb{R}^{r\times d_2} : G_1^\top G_2 = \frac{1}{\rho}\sum_{j=1}^r c_j U_{1,j} U_{2,j}^\top\right\} = \operatorname*{arg\,max}_{G_1\in\mathbb{R}^{r\times d_1}, G_2\in\mathbb{R}^{r\times d_2}} \operatorname{tr}\Big(G_1 A G_2^\top\Big) - (\rho/2)\|G_1^\top G_2\|_F^2. \quad (2.2)$$

In particular, the right singular vectors of G_1 and G_2 are uniquely determined (up to orthogonal transformation) independent of the choice of $\rho > 0$.

Using Lemma 2.1, Proposition 2.1 implies that, at each step of gradient descent during minimization of the of the regularized CLIP loss, the increment of parameter is in the direction of top-r singular vectors of S. Therefore, our result shows the equivalence between the minimization of the loss function 2.1 through gradient descent and top-r SVD with cross-covariance matrix.

3 Robustness of Multimodal Contrastive Learning to Data Noise

In this section, we investigate the robustness of MMCL against noisy pairs. We analyze the following linear loss, which is the loss function in 2.1 with $\phi(x) = x$, $\psi(x) = x$ and $C_n = n(n-1)$.

$$\mathcal{L}(G_1, G_2) = \frac{1}{n(n-1)}\sum_{j\neq i}(s_{ij} - s_{ii}) + R(G_1, G_2). \quad (3.1)$$

Note that this loss function can be rewritten as $\operatorname{tr}\big(G_1\bar{S}G_2\big) + R(G_1, G_2)$, where $\bar{S} \triangleq (n-1)^{-1}\sum_i (x_i - \bar{x})(\tilde{x}_i - \bar{\tilde{x}})^\top$ and thus in this case the minimizer of the loss function is *exactly* the maximizer of the SVD objective function $\operatorname{tr}\big(G_1\bar{S}G_2\big) - R(G_1, G_2)$.

The linear loss function for analyzing representation learning has been used in metric learning (Schroff et al., 2015; He et al., 2018), contrastive learning (Ji et al., 2021) and MMCL (Won et al., 2021; Alsan et al., 2021). Analysis on MMCL using InfoNCE loss is deferred to the Appendix 7.3.

3.1 Data Generating Process

For each modality, we consider the spiked covariance model (Bai and Yao, 2012; Yao et al., 2015; Zhang et al., 2018; Zeng et al., 2019; Ji et al., 2021) as the data generation process. Suppose that we have n observed pairs $\{x_i\}_{i\in[n]}$ and $\{\tilde{x}_i\}_{i\in[n]}$ drawn from the following model:

$$x_i = U_1^* z_i + \xi_i,\ \tilde{x}_i = U_2^*\tilde{z}_i + \tilde{\xi}_i, \quad (3.2)$$
$$z_i = \Sigma_z^{1/2} w_i,\ \ \tilde{z}_i = \Sigma_{\tilde{z}}^{1/2}\tilde{w}_i,\ \ \xi_i = \Sigma_\xi^{1/2}\zeta_i,\ \ \tilde{\xi}_i = \Sigma_{\tilde{\xi}}^{1/2}\tilde{\zeta}_i,$$

where w_i, $\tilde{w}_i$, ζ_i, and $\tilde{\zeta}_i$ have i.i.d. coordinates, each of which follows sub-Gaussian distribution with parameter σ

[2] A closely related notion is the contrastive covariance, which is the covariance matrix of data subracted by the covariance matrix of background noise, introduced in Abid et al. (2017). In the work, authors proposed contrastive principal component analysis, where PCA is applied to contrastive covariance, aiming to eliminate the effect of background noise from the data.

and unit variance. Notice that in model 3.2, U_1^* and z_i are only identifiable up to orthogonal transformation. Thus, we can assume that Σ_z is a diagonal matrix with $(\Sigma_z)_{1,1} \geq (\Sigma_z)_{2,2} \geq \cdots \geq (\Sigma_z)_{r,r}$ without loss of generality. A similar argument holds for U_2^* and $\tilde{z}_i$, and we assume that $\Sigma_{\tilde{z}}$ is also a diagonal matrix with $(\Sigma_{\tilde{z}})_{1,1} \geq (\Sigma_{\tilde{z}})_{2,2} \geq \cdots \geq (\Sigma_{\tilde{z}})_{r,r}$. Furthermore, without loss of generality, we assume $\|\Sigma_z\| = \|\Sigma_{\tilde{z}}\| = 1$.

Since the data are multimodal, we additionally assume the following (noisy) matches between two modalities. Recall that we have n observed pairs $(x_1, \tilde{x}_1), \ldots, (x_n, \tilde{x}_n)$. Define the set of observed indices as $\mathcal{C} \triangleq \{(1,1), \ldots, (n,n)\}$. Let $\mathcal{E} \subset [n] \times [n]$ be the set of n pairs. For the pairs $(i_1, j_1) \in \mathcal{E}$, assume that $w_{i_1} = \tilde{w}_{j_1}$ while ξ_{i_1} and $\tilde{\xi}_{j_1}$ are independent. For pairs $(i_1, j_1) \in [n]^2 \setminus \mathcal{E}$, assume the independence between w_{i_1} and $\tilde{w}_{j_1}$, and between ξ_{i_1} and $\tilde{\xi}_{j_1}$. We note that we can regard the set of pairs as the subset of the edges of the bipartite graph $\{(i,j) : i,j \in [n]\}$. Hereafter, we sometimes call the pairs in $\mathcal{E}$ ground truth edges and the pairs in $\mathcal{C}$ observed edges.

Let $m \triangleq |\mathcal{C} \cap \mathcal{E}| \in \{0, 1, \ldots, n\}$ be the number of observed ground-truth edges, and we define $p_n = 1 - m/n \in [0,1]$ as the *distortion rate* of the bipartite graph. When p_n is small, the information of association in collected data is highly reliable, while when p_n is large, the data contains many noisy pairs, which do not have the valid information between each modality.

We measure the "goodness" of pre-trained encoders by the quality of right-singular vectors of the encoders, since the fine-tuned predictors in downstream tasks only depend on the right-singular vectors in many cases. To see this, we decompose $G_1 \in \mathbb{R}^{r \times d}$ by SVD as $G_1 = VCU^\top$, where $U \in \mathbb{O}_{r,d}$, $V \in \mathbb{O}_{r,r}$ and C is diagonal. Suppose that we have a sample $(y, x) \in \mathbb{R}^{1+d}$ in the downstream task with some metric D. Through fine-tuning, we obtain $f^* = \arg\min_{f \in \mathcal{F}} D(y, f(G_1 x))$. For linear benchmarks, $\mathcal{F} = \{f : f(z) = w^\top z, w \in \mathbb{R}^r\}$ and f^* does not depend on V and C. This also holds for neural networks with a similar argument. To measure the quality of $P_r(G_1)$ (or $P_r(G_2)$, we employ $\sin\Theta$ distance; For two orthogonal matrices $U_1, U_2 \in \mathbb{O}_{d,r}$, the distance is defined as $\|\sin\Theta(U_1, U_2)\|_F \triangleq \|U_{1\perp}^\top U_2\|_F$ for any orthogonal compliment $U_{1\perp}$ of U_1.

3.2 Analysis on Multimodal Contrastive Loss Function

Before formalizing the previous argument, we introduce several assumptions.

Assumption 3.1. Assume that the condition numbers of Σ_z and $\Sigma_{\tilde{z}}$ are bounded; $\|\Sigma_z\|/\lambda_{\min}(\Sigma_z) \leq \kappa_z^2$ and $\|\Sigma_{\tilde{z}}\|/\lambda_{\min}(\Sigma_{\tilde{z}}) \leq \kappa_{\tilde{z}}^2$ for some constants $\kappa_z^2, \kappa_{\tilde{z}}^2 > 0$.

Assumption 3.2. Assume that the signal-to-noise ratio is bounded below; $\|\Sigma_z\|/\|\Sigma_\xi\| \geq s_1^2$ and $\|\Sigma_{\tilde{z}}\|/\|\Sigma_{\tilde{\xi}}\| \geq s_2^2$ for some constants $s_1^2, s_2^2 > 0$.

Assumption 3.1 imposes regularity conditions on covariance matrices, and Assumption 3.2 assumes the signal-to-noise ratio is not too small. Similar assumptions have been commonly used in the machine learning theory literature, e.g., Cai et al. (2019); Yan et al. (2021); Cai et al. (2021); Ji et al. (2021); Wen and Li (2021).

For this setting, we have the following result.

Theorem 3.1. *Suppose that we have a collection of pairs $(x_i, \tilde{x}_i)_{i=1}^n$ generated according to the model 3.2. Suppose Assumptions 3.1 and 3.2 hold. Let G_1 and G_2 be the solution to minimizing the loss 3.1. If $p_n \leq 1 - \eta$ for some constant $\eta > 0$, then, with probability $1 - O(n^{-1})$, we have*

$$\|\sin\Theta(P_r(G_1), U_1^*)\|_F \vee \|\sin\Theta(P_r(G_2), U_2^*)\|_F \lesssim \sqrt{r} \wedge \frac{1}{\eta}\sqrt{\frac{r(r + r(\Sigma_\xi) + r(\Sigma_{\tilde{\xi}}))\log(n + d_1 + d_2)}{n}}.$$

From Theorem 3.1, as long as $1 - p_n$ is strictly bounded away from 0, the feature recovery ability attains square root convergence. In other words, MMCL can learn representations even in the presence of noisy pairs whenever there are inignorable portion of observed ground-truth pairs. The case where $p_n \uparrow 1$ is treated in Section 8.3 in the appendix.

In the following, we compare the performance of MMCL with (unimodal) SSCL applied to each modality separately. SSCL aims to learn representations by contrasting pairs generated by data augmentation (Jaiswal et al., 2020; Liu et al., 2021; Yao et al., 2022a). In particular, we consider SSCL similar to SimCLR (Chen et al., 2020b), where encoders are trained to have similar representations for augmented views from the same sample and to discriminate augmented views from different samples.

Consider minimizing the unimodal linear contrastive loss using the first-mode data $\{x_i\}_{i=1}^n$. Let A be the *random masking augmentation* defined as $A = \mathrm{diag}(a_1, \ldots, a_{d_1})$, where a_i follows i.i.d. $\mathrm{Ber}(1/2)$ distribution. Given A, positive pairs are generated as $(Ax_i, (I - A)x_i)_{i=1}^n$. Let $\mathcal{L}_c(G_1) \triangleq \mathcal{L}(G_1, G_1; (x_i, \tilde{x}_i)_{i \in [n]})$ be the linear loss in 3.1 fed with generated positive pairs. Let G_1^c be the solution to minimizing the expected loss $\mathbb{E}_A[\mathcal{L}_c]$, where expectation is taken with respect to the data augmentation A. For the learned representation, Ji et al. (2021) showed that when Σ_ξ and $\Sigma_{\tilde{\xi}}$ are bounded above, then

$$\mathbb{E}\,\|\sin\Theta(U_1^\star, P_r(G_1^u))\|_F \lesssim \frac{r^{3/2}}{d}\log d + \sqrt{\frac{dr}{n}}.$$

The detailed statement is available in Appendix 7. Note that the assumption that the condition numbers of Σ_ξ and $\Sigma_{\tilde{\xi}}$ are bounded above implies that $r(\Sigma_\xi) \gtrsim d_1$ and $r(\Sigma_{\tilde{\xi}}) \gtrsim d_1$. Ignoring the logarithmic term, and provided that $d_1 \asymp d_2$, we notice that the bound in Theorem 3.1 improves the rate by reducing the bias term $r^{3/2}\log d/d$, while the variance term remains the same. The bias term is

due to the fact that core feature U_1^* loses its information when the random masking data augmentation is applied to the original data. Ji et al. (2021) also provably showed that when the noise covariance shows strong heteroskedasticity, the feature recovery performance of the representations obtained by autoencoders stays constant, while contrastive learning can mitigate the effect of heteroskedasticity. Therefore, under strong heteroskedasticity, MMCL can learn representations better than autoencoders applied to each modality separately.

3.3 Extension to Many-to-Many Correspondence Case

For the analysis in Section 3.2, we assumed one-to-one matches between the modalities and showed that MMCL can learn representations regardless of the distortion rate as long as there is an inignorable portion of ground-truth pairs in observed pairs. However, in practice, the performance of MMCL can be improved by eliminating noisy pairs. In addition, we face a multimodal dataset with many-to-many correspondence. To detect noisy pairs in the many-to-many correspondence case, we employ the Bipartite Spectral Graph Multi-Partitioning algorithm (BSGMP) (Dhillon, 2001). BSGMP is a generalization of the spectral graph partitioning algorithm by which we can detect and eliminate wrongly aligned edges in a bipartite graph that we expect to have a clustered shape.

Consider applying BSGMP first to the dataset generated by matching the MNIST and Fashion-MNIST datasets using labels. Then, we perform MMCL with InfoNCE loss defined as the loss function in equation 2.1 with $\phi(x) = \tau \log(x)$ and $\psi(x) = \exp(x/\tau)$. Details of the algorithm and results are deferred to Section 5.1.

Figure 1 shows that if we apply BSGMP with parameter $k = 10$, the performance of the downstream prediction task improves with moderate distortion rate.

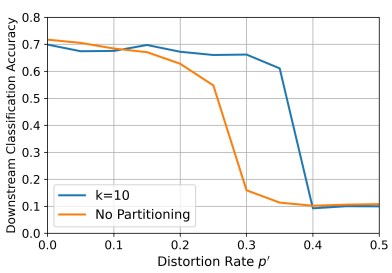

Figure 1: The downstream task performance of MMCL versus the distortion rate p'. The orange curve indicates MMCL without BSGMP, whereas the blue curve indicates MMCL with BSGMP with parameter $k = 10$.

4 Improving Multimodal Learning by Incorporating Unpaired Data

In this section, we propose a modification of CLIP loss to incorporate additional unpaired data and investigate its theoretical property. Due to the abundance of unpaired data in practice, this would greatly improve multimodal learning with the paired training data is limited.

Since MMCL projects the data into the shared representation space, we can explicitly calculate the similarity of any pair using given initial representations. This information, in turn, can be used to test whether a new pair is actually associated or not. More specifically, we consider a setting in which we have access to both the paired dataset $(x_i, \tilde{x}_i)_{i=1}^n$ and the unpaired datasets $(x_i^u)_{i=1}^N$ and $(\tilde{x}_i^u)_{i=1}^N$. Since we can regard the information on association between modalities as labels of pairs, we refer to this setting as a semi-supervised setting and call additional datasets unpaired data. We define the data generation process as follows. Suppose that the paired data $(x_i, \tilde{x}_i)_{i=1}^n$ are generated according to the model equation 3.2 described in Section 2. For the unpaired datasets $(x_i^u)_{i=1}^N$ and $(\tilde{x}_i^u)_{i=1}^N$, we assume the same spiked covariance model. We further assume the matches between two modalities as in Section 3.2. To avoid confusion of notation, let $\mathcal{E}^u \subset [N]^2$ denote the set of N ground-truth pairs for unpaired data. We continue to use the linear representation settings. Define the similarity between x_i^u and $\tilde{x}_j^u$ as $s_{ij}^u = s_{ij}^u(G_1, G_2) \triangleq \langle G_1 x_i^u, G_2 \tilde{x}_j^u \rangle$. We consider the following loss function $\mathcal{L}^u = \mathcal{L}^u(G_1, G_2; \bar{\mathcal{E}}^u)$ with respect to any set of pairs $\bar{\mathcal{E}}^u \subset [N]^2$ to incorporate the unpaired data:

$$\mathcal{L}^u \triangleq -\frac{\nu}{N} \sum_{(i,j)\in\bar{\mathcal{E}}^u} s_{ij}^u + \frac{\tau}{2N} \sum_{i\in[N]} \log \sum_{j\in[N]} e^{s_{ij}^u/\tau} + \frac{\tau}{2N} \sum_{i\in[N]} \log \sum_{j\in[N]} e^{s_{ji}^u/\tau} + R(G_1, G_2), \quad (4.1)$$

where $\nu \geq 1$. Note that this loss is exactly the InfoNCE loss function when $\bar{\mathcal{E}}^u = \{(1,1), \ldots, (N,N)\}$, $\epsilon = 1$ and $\nu = 1$. Setting $\nu > 1$ corresponds to choosing different temperature parameters for the similarity of positive pairs and negative pairs.

Given the loss function, we propose the semi-supervised MMCL in Algorithm 1.

Similar to Proposition 2.1, we can connect the gradient of the loss function and the negative gradient of the SVD objective. Define the contrastive cross-covariance matrix given some set of pairs $\bar{\mathcal{E}}^u$ as $S^u = S^u(\{\beta_{ij}^u\}_{i,j}; \bar{\mathcal{E}}^u) \triangleq \nu N^{-1} \sum_{(i,j)\in\bar{\mathcal{E}}^u} x_i \tilde{x}_j^\top - N^{-1} \sum_{i,j\in[N]} \beta_{ij}^u x_i \tilde{x}_j^\top$, where the formula for β_{ij}^u is available in Appendix 8.2. Then, for $k \in \{1, 2\}$,

$$\frac{\partial \mathcal{L}^u}{\partial G_k} = -\left.\frac{\partial \operatorname{tr}\left(G_1 S^u G_2^\top\right)}{\partial G_k}\right|_{\beta_{ij}^u = \beta_{ij}^u(G_1,G_2)} + \frac{\partial R(G_1, G_2)}{\partial G_k}. \quad (4.3)$$

Observe that each step of the gradient descent corresponds to the gradient ascent of the SVD objective with the negative cross-covariance matrix S^u.

Algorithm 1 Semi-supervised MMCL

Input: Labeled pairs $(x_i, \tilde{x}_i)_{i\in[n]}$, unlabeled data $(x_i^u)_{i\in[N]}$, $(\tilde{x}_i)_{i\in[N]}$, rank $r \geq 1$, parameters $\tau, \nu > 0$.
Obtain the initial representations $G_1^{(0)}$ and $G_2^{(0)}$ from the paired dataset $(x_i, \tilde{x}_i)_{i\in[n]}$ by minimizing InfoNCE loss given by the loss in equation 3.1 with $\phi(x) = \tau\log(1+x)$ and $\psi(x) = e^{x/\tau}$. Calculate the similarity of pairs by $s_{ij}^u = \langle G_1^{(0)} x_i^u, G_2^{(0)} \tilde{x}_j^u \rangle$.
Estimate the set of ground truth pairs $\mathcal{E}^u$ by

$$\hat{\mathcal{E}}^u \triangleq \{(i,j) \in [N]^2 : s_{ij}^u \geq s_{(N)}^u\}, \tag{4.2}$$

where $s_{(N)}^u$ is the N-th largest value of $\{s_{ij}^u : i \in [N], j \in \arg\max_{j'} s_{ij'}^u\} \cup \{(i,j) : j \in [N], i \in \arg\max_{i'} s_{i'j}^u\}$.
Output: G_1 and G_2 obtained by minimizing the loss $\mathcal{L}^u(G_1, G_2; \hat{\mathcal{E}}^u)$.

Motivated by Lemma 2.1, we consider the following two-step procedure to analyze the performance of Algorithm 1.

Step 1. Obtain the initial representations $G_1^{(0)}$ and $G_2^{(0)}$ from the paired dataset $(x_i, \tilde{x}_i)_{i\in[n]}$ by minimizing the linear loss in equation 3.1.

Step 2. Estimate the set of ground truth pairs $\mathcal{E}^u$ by $\hat{\mathcal{E}}^u$ as in Algorithm 1. Solve the following maximization problem, as an approximation to the minimization of the loss in equation 4.1 with $\hat{S}^u \triangleq S^u\big(\{\beta_{ij}^{u(0)}\}_{i,j}; \hat{\mathcal{E}}^u\big)$.

$$\max_{G_1\in\mathbb{R}^{r\times d_1}, G_2\in\mathbb{R}^{r\times d_2}} \operatorname{tr}\Big(G_1 \hat{S}^u G_2^\top\Big) - R(G_1, G_2), \tag{4.4}$$

where $\beta_{ij}^{u(0)} = \beta_{ij}^u(G_1^{(0)}, G_2^{(0)})$ is obtained using initial representations $G_1^{(0)}$ and $G_2^{(0)}$.

Although this is a two-step procedure, we note that an iterative version of this procedure can also be considered. The details and results are deferred to the Appendix 8.8.

To ensure that the obtained initial representations are accurate enough to detect ground truths, we assume the following assumptions.
Assumption 4.1. Suppose that the number of labeled pairs n satisfies

$$n \geq \frac{C}{\rho^2} \frac{(r + r(\Sigma_\xi) + r(\Sigma_{\tilde{\xi}}))^3}{r} \log N \cdot \log(n + d_1 + d_2),$$

where $C > 0$ is some constant depending on $\sigma, s_1, s_2, \kappa_z^2$ and $\kappa_{\tilde{z}}^2$.
Assumption 4.2. Assume that r and n satisfy $\log n \leq cr$, where $c > 0$ is some constant depending only on $\sigma, s_1, s_2, \kappa_z^2$ and $\kappa_{\tilde{z}}^2$.

Assumption 4.1 ensures that $\|G_1^{(0)\top} G_2^{(0)} - \rho^{-1} U_1^* \Sigma_z^{1/2} \Sigma_{\tilde{z}}^{1/2} U_2^{*\top}\|^2 = O((r\log n)^{-1})$ occurs with high probability, allowing one to detect ground truth pairs precisely with high probability. Assumption 4.2 is required to ensure that the similarity of ground truth pairs increases larger than the similarity of uncorrelated pairs. If $r \ll \log n$, it is more likely that two of n independent random vectors in $\mathbb{R}^r$ are close to each other, making the false positive rate in edge detection intolerably high.

For the estimation of ground-truth pairs, we have the following lemma.

Lemma 4.1. *Suppose Assumptions 3.1, 3.2, 4.1, and 4.2 hold. Fix any $\gamma > 0$. Then, with probability $1 - O(N^{-1} \vee n^{-1})$, $\hat{\mathcal{E}}^u = \mathcal{E}^u$ and*

$$\min_{(i,j)\in\hat{\mathcal{E}}^u} \beta_{ij}^{u(0)} = 1 - O\Big(\frac{1}{N^\gamma}\Big), \quad \max_{(i,j)\notin\hat{\mathcal{E}}^u} \beta_i^{u(0)} \lesssim \frac{1}{N^{1+\gamma}}.$$

Lemma 4.1 states that the cross-covariance matrix $\hat{S}^u$ behaves as if $\hat{S}^u \approx (\nu-1)N^{-1}\sum_{(i,j)\in\mathcal{E}^u} x_i \tilde{x}_j^\top$. Thus, when $\nu > 1$, Algorithm 1 has the ability to exploit ground-truth pairs, even if they are not observed. Notice that Assumption 4.1 is mild in the sense that it only requires $\tilde{\Omega}(r^2)$ number of samples when $r(\Sigma_\xi) \vee r(\Sigma_{\tilde{\xi}}) \lesssim r$. Taking advantage of this result, we can improve the performance of feature learning by incorporating unpaired data, as summarized in the next result.

Theorem 4.1. *Suppose Assumptions 3.1, 3.2, 4.1, and 4.2 hold. Fix any $\gamma > 1$ and $\nu > 1$. Choose $\tau \leq C(1+\gamma)^{-1}\sqrt{r/\log n}$, where $C > 0$ is some constant depending on $\sigma, s_1, s_2, \kappa_z^2, \kappa_{\tilde{z}}^2$. Let G_1 and G_2 be the solution to the maximization problem in equation 4.4. Then, with probability $1 - O(N^{-1} \vee n^{-1})$,*

$$\|\sin\Theta(P_r(G_1), U_1^*)\|_F \vee \|\sin\Theta(P_r(G_2), U_2^*)\|_F \lesssim \sqrt{r} \wedge \sqrt{\frac{r(r + r(\Sigma_\xi) + r(\Sigma_{\tilde{\xi}}))\log(N + d_1 + d_2)}{N}}.$$

Theorem 4.1 suggests that our proposed procedure is able to process the unpaired data as if they are paired, greatly improving the multimodal learning performance.

5 Numerical Experiments

In this section, we show that cleaning the noisy pairs using the BSGMP algorithm (Dhillon, 2001) helps MMCL improve the downstream task performance in the many-to-many correspondence case. In addition, we show that we can improve the performance of MMCL by incorporating the unpaired data. [3]

5.1 Eliminating Noisy Pairs

As briefly mentioned in Section 3.3, we use the BSGMP algorithm to eliminate incorrectly aligned edges from the

[3]The code is available at https://github.com/nswa17/MMCL.

training data and compare the performance for different distortion rates.

Algorithm 2 Bipartite Spectral Graph Multi-Partitioning

Input: two-modal dataset $\mathcal{E}'$. The number of clusters k.
Calculate an adjacency matrix A of the bipartite graph.
Let $A_n \triangleq D_1^{-1/2} A D_2^{-1/2}$, where D_1 and D_2 are diagonal matrices with $(D_1)_{ii} = |\{j \in [N] : (x_i, \tilde{x}_j) \in \mathcal{E}'\}|$ and $(D_2)_{ii} = |\{j \in [N] : (x_j, \tilde{x}_i) \in \mathcal{E}'\}|$.
Let $u_2, .., u_{l+1}$ be the left singular vectors and $v_2, .., v_{l+1}$ be the right singular vectors of A_n, where $l \triangleq \log_2 k$.
Define $Z \triangleq \begin{bmatrix} D_1^{-1/2} & U \\ D_2^{-1/2} & V \end{bmatrix}$, where $U \triangleq [u_2, .., u_{l+1}]$ and $V \triangleq [v_2, .., v_{l+1}]$.
Apply k-means algorithm to columns of the matrix Z.
Output: $\mathcal{E}'$ without all intercluster pairs.

In this experiment, we use MNIST and Fashion-MNIST datasets as different modalities. We pair images if they belong to the same class in each modality. For example, all digit-2 images in MNIST and all pullover images in Fashion-MNIST are connected. Note that, apart from the settings in Sections 3.2 and 4, we have a bipartite graph with many-to-many edges. In experiments, the number of training samples is set $n = 500$ for each modality, and the samples are equally distributed among different classes. For the MNIST side, we have fully-connected neural networks, while we use convolutional neural networks for the Fashion-MNIST side. The dimension of the latent space is chosen to be $r = 128$. After creating our dataset, we distort the pairs in the following way. For all $1 \leq i, j \leq n$, if x_i and $\tilde{x}_j$ are paired, we remove this pair with probability p'. Similarly, if x_i and $\tilde{x}_j$ are not paired, we pair them in our dataset with probability p'. Note that due to this distortion, our bipartite graph loses its clustered structure, which we try to regain with Algorithm 2. Although we know the true number of clusters $k = 10$, we treat k as unknown and try a different number of clusters $k = 5, 7, 10, 13, 15$. We also perform MMCL without applying algorithm 2. The performance of the learned representations is measured by a downstream classification accuracy as in Radford et al. (2021) using test data, which is generated in the same way as the training data. That is, for each test datum x on the Fashion-MNIST side, the most similar test datum $\tilde{x}$ on the MNIST side is chosen. We then measure the accuracy by the rate of x and $\tilde{x}$ whose labels are equal. The experiment was performed for different k and p'.

The result is reported in Table 1. When the distortion rate is $p' = 0.1$, the naive use of MMCL ("No Partitioning") performs the best. As p' increases, the downstream task performance decreases if no partitioning is applied. However, applying algorithm 2 with $k = 10$ yields the highest accuracy for $p' = 0.2$ and $p' = 0.3$. Note that this allows one to choose the number of clusters k by cross-validation.

Table 1: Downstream task classification accuracy with different distortion rates p'.

Partitioning	$p' = 0.1$	$p' = 0.2$	$p' = 0.3$
$k = 5$	0.370	0.407	0.353
$k = 7$	0.438	0.482	0.481
$k = 10$	0.675	**0.672**	**0.662**
$k = 13$	0.667	0.669	0.596
$k = 15$	0.650	0.612	0.579
No Partitioning	**0.684**	0.628	0.159

5.2 Incorporating Unpaired Data

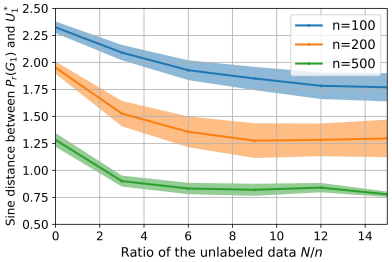

Figure 2: Performance of feature recovery ability measured by $\|\sin\Theta(P_r(G_1), U_1^*)\|_F$ with additional unpaired data with different $n = 100, 200, 500$.

In this experiment, we show that we can improve the performance of MMCL by incorporating unpaired data. Specifically, we generate a synthetic dataset with $d_1 = 40, d_2 = 39, n = 300, r = 10$ according to the model in equation 3.2. $U_1^* \in \mathcal{O}_{d_1,r}$ and $U_2^* \in \mathcal{O}_{d_2,r}$ are random orthonormal matrices, and z_i and $\tilde{z}_i$ are sampled from the standard Gaussian distribution for $i \in [n]$. Similarly, ξ_i and $\tilde{\xi}_i$ are sampled from the zero-mean Gaussian distribution with standard deviation 0.3. The unpaired data are generated in the same way as the paired data with N number of samples for each modalities. We first train the initial linear encoders by minimizing InfoNCE loss with the paired data. Then we train the linear encoders using the procedure described in Section 4 with unpaired data. The performance of the obtained representations is measured by the $\sin\Theta$ distance between U_1^* and $P_r(G_1)$. We consider different ratios N/n with $n = 100, 200, 500$, and the results are summarized in Figure 2.

As the ratio of unpaired data with labeled data N/n increases, we can observe that the $\sin\Theta$ distance decreases, which validates our theory in the sense that if the model is initialized with relatively good accuracy, having more and more unpaired data improves the performance.

5.3 Application to Real Datasets

In this experiment, we show that semi-supervised CLIP improves the performance for real data. The dataset is created

by artificially pairing images from MNIST and Fashion-MNIST datasets based on their labels. Namely, we generate pairs $(x_i, \tilde{x}_i)_{i=1}^n$ so that the digit in MNIST is paired with a random image in Fashion-MNIST with the corresponding digit. Similarly, we generate validation pairs $(x_i^v, \tilde{x}_i^v)_{i=1}^v$ with some v, the details of which are explained later. We use random data from each modality as unlabeled datasets $(x_i^u)_{i=1}^N$ and $(\tilde{x}_i^u)_{i=1}^N$.

In CLIP and ALIGN (Jia et al., 2021; Radford et al., 2021), pre-trained encoders such as Vision Transformers and BERT are used. To speed up the learning process, we first reduce the dimension of data; we train autoencoders consisting of 2-dimensional multilayer convolutional neural networks on datasets $(x_i^u)_{i=1}^N$ and $(\tilde{x}_i^u)_{i=1}^N$ separately. Let the encoders obtained be $E_1 : \mathbb{R}^{28^2} \to \mathbb{R}^{16}$ and $E_2 : \mathbb{R}^{28^2} \to \mathbb{R}^{16}$ for MNIST and Fashion-MNIST, respectively.

We first train the initial representations for the dataset $(E_1(x_i), E_2(\tilde{x}_i))_{i=1}^n$ while fixing the parameters of E_1 and E_2. Then, we train the semi-supervised CLIP with another set of data whose association is unknown. The initial representations consist of two two-layer neural networks.

We call the representations used to estimate the ground-truth pairs *anchor representations*. Given anchor representations G_1^a and G_2^a, we estimate the ground-truth pairs by $\hat{\mathcal{E}}^u$ as in equation 4.2.

Let the initial representations trained on the dataset $(E_1(x_i), E_2(\tilde{x}_i))_{i=1}^n$ be initial anchor representations. Since the performance of semi-supervised CLIP is restricted by the performance of anchor representations, we update anchor representations when the learned representations outperform the current anchor representations by a certain ratio. The performance of representations is measured on validation pairs $(x_i^v, \tilde{x}_i^v)_{i=1}^v$.

We employ the AdamW algorithm and use mini-batch optimization with batch size 64 to reduce the load of calculating the similarity matrix. The number of epochs for initial training and training with labeled data is set 100.

We compare the performance of semi-supervised CLIP with CLIP trained with unknown association. Figure 3 shows the improvement of test accuracy when we additionally train CLIP with unlabeled data. The parameters are set as $N = 59000$, $n = v = 500$, $\tau = 1$ and $\eta = 1.1$. The gray curve indicates the test accuracy of initial representations against the number of epochs. The orange curve indicates the accuracy of representations when we additionally feed the ground-truth pairs of unlabeled data. The blue curve indicates the accuracy of representations with additional unlabeled data. From this result, we can see the improvement in test accuracy with additional unlabeled data.

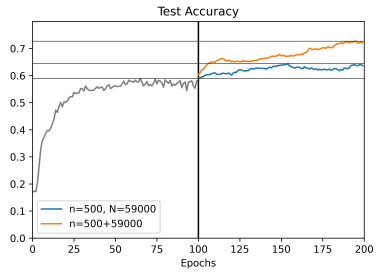

Figure 3: The comparison of the downstream task performance of semi-supervised CLIP and oracle CLIP. The gray curve indicates the performance when training initial representations. The orange curve indicates the performance of semi-supervised CLIP and the blue curve indicates the performance of oracle CLIP.

6 Discussion

In this paper, we provide a theoretical understanding of MMCL in linear representation settings with two-modal data. We showed that for a general class of non-linear MMCL loss performing gradient descent on the loss function is equivalent to gradient ascent of the SVD objective function with contrastive cross-covariance matrix. Using this result, we analyze the feature recovery ability of the approximated algorithm under linear loss in the presence of noisy pairs. We note that the feature recovery ability of MMCL attains a better theoretical bound compared to that attained by SSCL applied separately to each modality. For data with many-to-many correspondence, we numerically show that we can improve the performance of MMCL by eliminating incorrectly paired edges using BS-GMP. We also proposed a semi-supervised framework that incorporates the unpaired dataset to enhance the performance of MMCL. Given a small number of labeled data, it can successfully detect the ground-truth alignment for unpaired data and improve the representations. The improvement in performance with linear encoders is verified by numerical experiment. To the best of our knowledge, this is the first work on the theoretical analysis of MMCL that incorporates the unpaired data. We emphasize that our main contribution is to provide theoretical analysis and insights on MMCL. The analysis of other multimodal pretrain learning algorithms remains a future work. While we verified our theory with proof-of-concept experiments, systematic experiments with non-linear representations on larger datasets is a good direction of future work. It is also possible to extend two-modal contrastive learning to more than three modalities by summing up the loss equation 2.1 for all pairs of modalities. As an analogy to the results of Section 2, we can interpret loss minimization via gradient descent as maximizing the sum of the SVD objective functions (Corollary 7.1.) An analysis of its feature recovery ability remains a future work.

Acknowledgements

The research of Linjun Zhang is partially supported by NSF DMS-2015378. The research of James Zou is partially supported by funding from NSF CAREER and the Sloan Fellowship.

References

Abid, A., Zhang, M. J., Bagaria, V. K., and Zou, J. (2017). Contrastive principal component analysis. *arXiv preprint arXiv:1709.06716*.

Akaho, S. (2006). A kernel method for canonical correlation analysis. *arXiv preprint cs/0609071*.

Alsan, H. F., Yıldız, E., Safdil, E. B., Arslan, F., and Arsan, T. (2021). Multimodal retrieval with contrastive pretraining. In *2021 International Conference on INnovations in Intelligent SysTems and Applications (INISTA)*, pages 1–5. IEEE.

Andrew, G., Arora, R., Bilmes, J., and Livescu, K. (2013). Deep canonical correlation analysis. In *International conference on machine learning*, pages 1247–1255. PMLR.

Ash, J. T., Goel, S., Krishnamurthy, A., and Misra, D. (2021). Investigating the role of negatives in contrastive representation learning. *arXiv preprint arXiv:2106.09943*.

Bai, Z. and Yao, J. (2012). On sample eigenvalues in a generalized spiked population model. *Journal of Multivariate Analysis*, 106:167–177.

Baltrušaitis, T., Ahuja, C., and Morency, L.-P. (2018). Multimodal machine learning: A survey and taxonomy. *IEEE transactions on pattern analysis and machine intelligence*, 41(2):423–443.

Benton, A., Khayrallah, H., Gujral, B., Reisinger, D. A., Zhang, S., and Arora, R. (2017). Deep generalized canonical correlation analysis. *arXiv preprint arXiv:1702.02519*.

Bunea, F. and Xiao, L. (2015). On the sample covariance matrix estimator of reduced effective rank population matrices, with applications to fpca. *Bernoulli*, 21(2):1200–1230.

Cai, T. T., Ma, J., and Zhang, L. (2019). Chime: Clustering of high-dimensional gaussian mixtures with em algorithm and its optimality. *The Annals of Statistics*, 47(3):1234–1267.

Cai, T. T., Wang, Y., and Zhang, L. (2021). The cost of privacy: Optimal rates of convergence for parameter estimation with differential privacy. *The Annals of Statistics*, 49(5):2825–2850.

Chen, T., Kornblith, S., Norouzi, M., and Hinton, G. (2020a). A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR.

Chen, X., Fan, H., Girshick, R., and He, K. (2020b). Improved baselines with momentum contrastive learning. *arXiv preprint arXiv:2003.04297*.

Dhillon, I. S. (2001). Co-clustering documents and words using bipartite spectral graph partitioning. In *Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 269–274.

Du, Y., Liu, Z., Li, J., and Zhao, W. X. (2022). A survey of vision-language pre-trained models. *arXiv preprint arXiv:2202.10936*.

Galatolo., F., Cimino., M., and Vaglini, G. (2021). Generating images from caption and vice versa via clip-guided generative latent space search. *Proceedings of the International Conference on Image Processing and Vision Engineering*.

Goldberger, J., Hinton, G. E., Roweis, S., and Salakhutdinov, R. R. (2004). Neighbourhood components analysis. *Advances in neural information processing systems*, 17.

Golub, G. H. and Van Loan, C. F. (2013). *Matrix computations*. JHU press.

Grill, J.-B., Strub, F., Altché, F., Tallec, C., Richemond, P. H., Buchatskaya, E., Doersch, C., Pires, B. A., Guo, Z. D., Azar, M. G., et al. (2020). Bootstrap your own latent: A new approach to self-supervised learning. *arXiv preprint arXiv:2006.07733*.

Hardoon, D. R., Szedmak, S., and Shawe-Taylor, J. (2004). Canonical correlation analysis: An overview with application to learning methods. *Neural computation*, 16(12):2639–2664.

Harold, H. (1936). Relations between two sets of variates. *Biometrika*, 28(3/4):321.

He, K., Fan, H., Wu, Y., Xie, S., and Girshick, R. (2020). Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9729–9738.

He, X., Zhou, Y., Zhou, Z., Bai, S., and Bai, X. (2018). Triplet-center loss for multi-view 3d object retrieval. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1945–1954.

Hu, P., Huang, Y.-a., Mei, J., Leung, H., Chen, Z.-h., Kuang, Z.-m., You, Z.-h., and Hu, L. (2021). Learning from low-rank multimodal representations for predicting disease-drug associations. *BMC medical informatics and decision making*, 21(1):1–13.

Huang, W., Yi, M., and Zhao, X. (2021a). Towards the generalization of contrastive self-supervised learning. *arXiv preprint arXiv:2111.00743*.

Huang, Y., Du, C., Xue, Z., Chen, X., Zhao, H., and Huang, L. (2021b). What makes multi-modal learning better than single (provably). *Advances in Neural Information Processing Systems*, 34:10944–10956.

Jaiswal, A., Babu, A. R., Zadeh, M. Z., Banerjee, D., and Makedon, F. (2020). A survey on contrastive self-supervised learning. *Technologies*, 9(1):2.

Ji, W., Deng, Z., Nakada, R., Zou, J., and Zhang, L. (2021). The power of contrast for feature learning: A theoretical analysis. *arXiv preprint arXiv:2110.02473*.

Jia, C., Yang, Y., Xia, Y., Chen, Y.-T., Parekh, Z., Pham, H., Le, Q., Sung, Y.-H., Li, Z., and Duerig, T. (2021). Scaling up visual and vision-language representation learning with noisy text supervision. In *International Conference on Machine Learning*, pages 4904–4916. PMLR.

Jing, L., Vincent, P., LeCun, Y., and Tian, Y. (2021). Understanding dimensional collapse in contrastive self-supervised learning. *arXiv preprint arXiv:2110.09348*.

Kettenring, J. R. (1971). Canonical analysis of several sets of variables. *Biometrika*, 58(3):433–451.

Ko, C.-Y., Mohapatra, J., Liu, S., Chen, P.-Y., Daniel, L., and Weng, L. (2022). Revisiting contrastive learning through the lens of neighborhood component analysis: an integrated framework. In *International Conference on Machine Learning*, pages 11387–11412. PMLR.

Lai, P. L. and Fyfe, C. (1998). Canonical correlation analysis using artificial neural networks. In *ESANN*, pages 363–368. Citeseer.

Lai, P. L. and Fyfe, C. (1999). A neural implementation of canonical correlation analysis. *Neural Networks*, 12(10):1391–1397.

Lee, S. I. and Yoo, S. J. (2020). Multimodal deep learning for finance: integrating and forecasting international stock markets. *The Journal of Supercomputing*, 76(10):8294–8312.

Li, F., Zhang, H., Zhang, Y.-F., Liu, S., Guo, J., Ni, L. M., Zhang, P., and Zhang, L. (2022). Vision-language intelligence: Tasks, representation learning, and large models. *arXiv preprint arXiv:2203.01922*.

Li, L. H., Yatskar, M., Yin, D., Hsieh, C.-J., and Chang, K.-W. (2019). Visualbert: A simple and performant baseline for vision and language. *arXiv preprint arXiv:1908.03557*.

Li, X., Yin, X., Li, C., Zhang, P., Hu, X., Zhang, L., Wang, L., Hu, H., Dong, L., Wei, F., et al. (2020). Oscar: Object-semantics aligned pre-training for vision-language tasks. In *European Conference on Computer Vision*, pages 121–137. Springer.

Li, Y., Liang, F., Zhao, L., Cui, Y., Ouyang, W., Shao, J., Yu, F., and Yan, J. (2021). Supervision exists everywhere: A data efficient contrastive language-image pre-training paradigm. *arXiv preprint arXiv:2110.05208*.

Liang, P. P., Zadeh, A., and Morency, L.-P. (2022). Foundations and recent trends in multimodal machine learning: Principles, challenges, and open questions. *arXiv preprint arXiv:2209.03430*.

Liu, X., Zhang, F., Hou, Z., Mian, L., Wang, Z., Zhang, J., and Tang, J. (2021). Self-supervised learning: Generative or contrastive. *IEEE Transactions on Knowledge and Data Engineering*.

Lu, J., Batra, D., Parikh, D., and Lee, S. (2019). Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *Advances in neural information processing systems*, 32.

Lyu, Q. and Fu, X. (2020). Nonlinear multiview analysis: Identifiability and neural network-assisted implementation. *IEEE Transactions on Signal Processing*, 68:2697–2712.

Lyu, Q. and Fu, X. (2022). Finite-sample analysis of deep cca-based unsupervised post-nonlinear multimodal learning. *IEEE Transactions on Neural Networks and Learning Systems*.

Ma, F., Xu, X., Huang, S.-L., and Zhang, L. (2021a). Maximum likelihood estimation for multimodal learning with missing modality. *arXiv preprint arXiv:2108.10513*.

Ma, M., Ren, J., Zhao, L., Tulyakov, S., Wu, C., and Peng, X. (2021b). Smil: Multimodal learning with severely missing modality. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 2302–2310.

Narasimhan, M., Rohrbach, A., and Darrell, T. (2021). Clip-it! language-guided video summarization. *Advances in Neural Information Processing Systems*, 34:13988–14000.

Ngiam, J., Khosla, A., Kim, M., Nam, J., Lee, H., and Ng, A. Y. (2011). Multimodal deep learning. In *ICML*.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al. (2021). Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR.

Saunshi, N., Ash, J., Goel, S., Misra, D., Zhang, C., Arora, S., Kakade, S., and Krishnamurthy, A. (2022). Understanding contrastive learning requires incorporating inductive biases. *arXiv preprint arXiv:2202.14037*.

Saunshi, N., Plevrakis, O., Arora, S., Khodak, M., and Khandeparkar, H. (2019). A theoretical analysis of contrastive unsupervised representation learning. In *International Conference on Machine Learning*, pages 5628–5637. PMLR.

Schroff, F., Kalenichenko, D., and Philbin, J. (2015). Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on*

computer vision and pattern recognition, pages 815–823.

Shi, Y., Paige, B., Torr, P. H., and Siddharth, N. (2020). Relating by contrasting: A data-efficient framework for multimodal generative models. *arXiv preprint arXiv:2007.01179*.

Srivastava, N. and Salakhutdinov, R. R. (2012). Multimodal learning with deep boltzmann machines. *Advances in neural information processing systems*, 25.

Subramanian, V., Syeda-Mahmood, T., and Do, M. N. (2021). Multi-modality fusion using canonical correlation analysis methods: Application in breast cancer survival prediction from histology and genomics. *arXiv preprint arXiv:2111.13987*.

Sun, X., Xu, Y., Cao, P., Kong, Y., Hu, L., Zhang, S., and Wang, Y. (2020). Tcgm: An information-theoretic framework for semi-supervised multi-modality learning. In *European conference on computer vision*, pages 171–188. Springer.

Tan, H. and Bansal, M. (2019). Lxmert: Learning cross-modality encoder representations from transformers. *arXiv preprint arXiv:1908.07490*.

Tian, Y. (2022). Deep contrastive learning is provably (almost) principal component analysis. *arXiv preprint arXiv:2201.12680*.

Tian, Y., Chen, X., and Ganguli, S. (2021). Understanding self-supervised learning dynamics without contrastive pairs. In *International Conference on Machine Learning*, pages 10268–10278. PMLR.

Tropp, J. A. (2012). User-friendly tail bounds for sums of random matrices. *Foundations of computational mathematics*, 12(4):389–434.

Vershynin, R. (2018). *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge university press.

Wainwright, M. J. (2019). *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48. Cambridge University Press.

Wang, F. and Liu, H. (2021). Understanding the behaviour of contrastive loss. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2495–2504.

Wang, T. and Isola, P. (2020). Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In *International Conference on Machine Learning*, pages 9929–9939. PMLR.

Wang, W., Yan, X., Lee, H., and Livescu, K. (2016). Deep variational canonical correlation analysis. *arXiv preprint arXiv:1610.03454*.

Wang, Z., Codella, N., Chen, Y.-C., Zhou, L., Yang, J., Dai, X., Xiao, B., You, H., Chang, S.-F., and Yuan, L. (2022). Clip-td: Clip targeted distillation for vision-language tasks. *arXiv preprint arXiv:2201.05729*.

Wen, Z. and Li, Y. (2021). Toward understanding the feature learning process of self-supervised contrastive learning. In *International Conference on Machine Learning*, pages 11112–11122. PMLR.

Won, M., Oramas, S., Nieto, O., Gouyon, F., and Serra, X. (2021). Multimodal metric learning for tag-based music retrieval. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 591–595. IEEE.

Wu, R., Zhang, L., and Tony Cai, T. (2022). Sparse topic modeling: Computational efficiency, near-optimal algorithms, and statistical inference. *Journal of the American Statistical Association*, pages 1–13.

Xu, P., Zhu, X., and Clifton, D. A. (2022). Multimodal learning with transformers: A survey. *arXiv preprint arXiv:2206.06488*.

Yan, Y., Chen, Y., and Fan, J. (2021). Inference for heteroskedastic pca with missing data. *arXiv preprint arXiv:2107.12365*.

Yang, X., Liu, W., Liu, W., and Tao, D. (2019). A survey on canonical correlation analysis. *IEEE Transactions on Knowledge and Data Engineering*, 33(6):2349–2368.

Yao, H., Wang, Y., Li, S., Zhang, L., Liang, W., Zou, J., and Finn, C. (2022a). Improving out-of-distribution robustness via selective augmentation. In *International Conference on Machine Learning*, pages 25407–25437. PMLR.

Yao, H., Zhang, L., and Finn, C. (2022b). Meta-learning with fewer tasks through task interpolation. In *International Conference on Learning Representations*.

Yao, J., Zheng, S., and Bai, Z. (2015). *Sample covariance matrices and high-dimensional data analysis*. Cambridge University Press Cambridge.

Yu, Y., Wang, T., and Samworth, R. J. (2015). A useful variant of the davis–kahan theorem for statisticians. *Biometrika*, 102(2):315–323.

Yuan, X., Lin, Z., Kuen, J., Zhang, J., Wang, Y., Maire, M., Kale, A., and Faieta, B. (2021). Multimodal contrastive training for visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6995–7004.

Yuhas, B. P., Goldstein, M. H., and Sejnowski, T. J. (1989). Integration of acoustic and visual speech signals using neural networks. *IEEE Communications Magazine*, 27(11):65–71.

Zadeh, A., Liang, P. P., and Morency, L.-P. (2020). Foundations of multimodal co-learning. *Information Fusion*, 64:188–193.

Zeng, X., Xia, Y., and Zhang, L. (2019). Double cross validation for the number of factors in approximate factor models. *arXiv preprint arXiv:1907.01670*.

Zhang, A. R., Cai, T. T., and Wu, Y. (2018). Heteroskedastic pca: Algorithm, optimality, and applications. *arXiv preprint arXiv:1810.08316*.

Zhang, C., Yang, Z., He, X., and Deng, L. (2020). Multimodal intelligence: Representation learning, information fusion, and applications. *IEEE Journal of Selected Topics in Signal Processing*, 14(3):478–493.

Supplementary Materials: Understanding Multimodal Contrastive Learning and Incorporating Unpaired Data

In this appendix, we define the following notations. Let $\mathbf{1}\{A\}$ be an indicator function that takes 1 when A is true, otherwise takes 0. For any square matrix of the same order A and B, we write $A \prec B$ if $u^\top (B-A)u \geq 0$ holds for all unit vector u. Define the set of pairs in $[n]^2 \setminus \mathcal{E}$ containing x_i as $\mathcal{E}^\perp_{i,\cdot} \triangleq \{(i,j) \in \mathcal{E}^\perp : j \in [n]\}$ and similarly the pairs containing $\tilde{x}_j$ as $\mathcal{E}^\perp_{\cdot,j} \triangleq \{(i,j) \in \mathcal{E}^\perp : i \in [n]\}$. For any matrix A, let $\mathrm{SVD}_r(A)$ be the rank-r approximation of A. Let $\mathbb{S}^{d-1} \triangleq \{x \in \mathbb{R}^d : x^\top x = 1\}$ be a sphere on $\mathbb{R}^d$.

7 Omitted Contents

7.1 Numerical Experiments

Here is the algorithm used in the experiment in Section 5.3.

Algorithm 3 Semi-supervised MMCL

Input: Labeled pairs $(x_i, \tilde{x}_i)_{i=1}^n$, validation pairs $(x_i^v, \tilde{x}_i^v)_{i=1}^n$, unlabeled data $(x_i^u)_{i=1}^N, (\tilde{x}_i^u)_{i=1}^N$, temperature $\tau > 0$, update ratio $\eta > 0$.
Obtain the initial representations $G_1^{(0)}$ and $G_2^{(0)}$ from the paired dataset $(E_1(x_i), E_2(\tilde{x}_i))_{i\in[n]}$ by minimizing CLIP loss.
Let $G_1^a = G_1^{(0)}$ and $G_2^a = G_2^{(0)}$.
repeat
  Calculate the similarity of all possible unlabeled pairs by $s_{ij}^u = \langle G_1^{(0)} E_1(x_i^u), G_2^{(0)} E_2(\tilde{x}_j^u)\rangle$ for $i, j \in [N]$.
  Estimate the set of ground truth pairs according to equation 4.2.
  Obtain G_1 and G_2 by minimizing CLIP loss with artificially paired dataset $(E_1(x_i^u), E_2(\tilde{x}_j^u))_{(i,j)\in\hat{\mathcal{E}}^u}$.
  if G_1 and G_2 outperforms G_1^a and G_2^a on validation set $(E_1(x_i^v), E_2(\tilde{x}_i^v))_{i=1}^n$ by η, **then**
    Set $G_1^a = G_1$ and $G_2^a = G_2$.
  end if
until convergence
Output: G_1 and G_2.

7.2 Feature Recovery via SSCL

Define the incoherent constant, which measures the closeness between the standard basis and orthonormal column vectors of a matrix $U \in \mathbb{O}_{d,r}$ as $I(U) \triangleq \max_{i\in[d]} \|e_i^\top U\|^2$. For the learned representation, we invoke the following theorem from Ji et al. (2021).

Lemma 7.1 (Theorem 3.11 from Ji et al. (2021))**.** *Suppose that $n > d \gg r$ and the condition number of Σ_ξ and $\Sigma_{\tilde{\xi}}$ are bounded above, and $I(U^\star) = O(r \log d/d)$. Consider applying random masking augmentation. Then,*

$$\mathbb{E}\left\|\sin\Theta(U_1^\star, P_r(G_1^u))\right\|_F \lesssim \frac{r^{3/2}}{d}\log d + \sqrt{\frac{dr}{n}}.$$

Note that the assumption that the condition numbers of Σ_ξ and $\Sigma_{\tilde{\xi}}$ are bounded above implies that $r(\Sigma_\xi) \gtrsim d_1$ and $r(\Sigma_{\tilde{\xi}}) \gtrsim d_1$. Ignoring the logarithmic term, and provided that $d_1 \asymp d_2$, we notice that the bound in Theorem 3.1 improves the rate in Lemma 7.1 by reducing the bias term $r^{3/2}\log d/d$, while the variance term remains almost the same.

The bias term appearing in the bound in Lemma 7.1 is due to the fact that core feature U_1^* loses its information when the random masking data augmentation is applied to the original data. Also, note that our result 7.1 does not require the incoherent constant assumption, because we can separate core features from noise using the fact that the core features are highly correlated, while noises are not correlated between two modalities. Ji et al. (2021) provably showed that when the noise covariance shows strong heteroskedasticity, the feature recovery performance of representations obtained by autoencoders stays constant, while contrastive learning can mitigate the effect of heteroskedasticity. Therefore, under strong heteroskedasticity, MMCL can learn representations better than autoencoders applied to each modality separately.

7.3 Analysis on Multimodal Contrastive Loss Function with InfoNCE Loss

Before going to the proof, we modify the multimodal contrastive loss with InfoNCE loss with $\epsilon = 1$ as

$$\mathcal{L}(G_1, G_2) \triangleq -\frac{\tau}{2n} \sum_{(i,j)\in\mathcal{E}^u} \log \frac{e^{\nu s_{ii}^u/\tau}}{\sum_{j\in[n]} e^{s_{ij}^u/\tau}} - \frac{\tau}{2n} \sum_{(i,j)\in\mathcal{E}^u} \log \frac{e^{\nu s_{ii}^u/\tau}}{\sum_{i\in[n]} e^{s_{ij}^u/\tau}} + R(G_1, G_2), \tag{7.1}$$

where $\nu \geq 1$. Setting $\nu > 1$ corresponds to choosing different temperature parameters for positive pairs and negative pairs.

Similar to the argument in Proposition 2.1, each step of the gradient descent of minimizing the loss in equation 7.1 corresponds to performing gradient ascent to the objective function $\mathrm{tr}\big(G_1 S G_2^\top\big) - (\rho/2)\|G_1^\top G_2\|_F^2$ with the matrix S given by

$$S = \frac{1}{n}\sum_{i\in[n]} \beta_i x_i \tilde{x}_i^\top - \frac{1}{n}\sum_{i\neq j} \beta_{ij} x_i \tilde{x}_j^\top, \tag{7.2}$$

where

$$\begin{aligned}
\beta_i &= \nu - 1 + \frac{1}{2}\frac{\sum_{j':j'\neq i}\exp(s_{ij'}/\tau)}{\sum_{j':j'\in[n]}\exp(s_{ij'}/\tau)} + \frac{1}{2}\frac{\sum_{j':j'\neq i}\exp(s_{j'i}/\tau)}{\sum_{j':j'\in[n]}\exp(s_{j'i}/\tau)}, \\
\beta_{ij} &= \frac{1}{2}\frac{\exp(s_{ij}/\tau)}{\sum_{j':j'\in[n]}\exp(s_{ij'}/\tau)} + \frac{1}{2}\frac{\exp(s_{ij}/\tau)}{\sum_{i':i'\in[n]}\exp(s_{i'j}/\tau)}.
\end{aligned} \tag{7.3}$$

From this observation and Lemma 2.1, we study the following loss minimization problem as an approximation to MMCL.

Algorithm 4 Approximated Multimodal Contrastive Learning

Input: Data $(x_i)_{i\in[n]}$ and $(\tilde{x}_i)_{i\in[n]}$, rank $r \geq 1$, temperature $\tau > 0$, parameter $\nu \geq 1$, initial representations $G_1^{(0)} \in \mathbb{R}^{r\times d_1}$ and $G_2^{(0)} \in \mathbb{R}^{r\times d_2}$.

Calculate the similarity of pairs by $s_{ij} = \langle G_1^{(0)} x_i, G_2^{(0)} \tilde{x}_j \rangle$. Also calculate β_i and β_{ij} for $i \neq j$ according to equation 7.3. Compute

$$S = \frac{1}{n}\sum_{i\in[n]} \beta_i x_i \tilde{x}_i^\top - \frac{1}{n}\sum_{i\neq j} \beta_{ij} x_i \tilde{x}_j^\top.$$

Perform SVD on S and write $S = \sum_{j=1}^{d_1\wedge d_2} \lambda_j u_{1j} u_{2j}^\top$ so that $\lambda_1 \geq \cdots \geq \lambda_{d_1\wedge d_2}$. Let $G_1 \in \mathbb{R}^{r\times d_1}$ and $G_2 \in \mathbb{R}^{r\times d_2}$ satisfy $G_1^\top G_2 = \sum_{j=1}^r \lambda_j u_{1j} u_{2j}^\top$.

Output: G_1 and G_2.

We introduce an assumption for initial representations.

Assumption 7.1. Assume that there exist a constant $q > 0$ and some small constant $c_q = c_q(\sigma, s_1, s_2, \kappa_z^2, q) > 0$ such that the initial representations $G_1^{(0)}$ and $G_2^{(0)}$ satisfy

$$\|G_1^{(0)\top} G_2^{(0)} - q U_1^* \Sigma_z^{1/2} \Sigma_{\tilde{z}}^{1/2} U_2^{*\top}\|^2 \leq c_q \frac{r}{(r + r(\Sigma_\xi))(r + r(\Sigma_{\tilde{\xi}}))\log n}.$$

The following lemma states that when the initial representations are good enough, then Algorithm 4 can detect the unobserved ground-truth pairs.

Lemma 7.2. *Suppose Assumptions 3.1, 3.2, 4.2, and 7.1 hold. Fix any $\gamma > 0$ and $\nu \geq 1$. Choose $\tau \leq C(1+\gamma)^{-1}\sqrt{r/\log n}$, where $C > 0$ is some constant depending on $\sigma, s_1, s_2, \kappa_z^2, \kappa_{\tilde{z}}^2$. Assume that n satisfies*

$$n \geq \frac{C_q}{c_q} \frac{(r + r(\Sigma_\xi) + r(\Sigma_{\tilde{\xi}}))^{5/2} \log n \sqrt{\log(n + d_1 + d_2)}}{r}, \tag{7.4}$$

where $C_q = C_q(\sigma, s_1, s_2, \kappa_z^2, \kappa_{\tilde{z}}^2, q) > 0$ is some constant. Consider applying Algorithm 4 to the data generated from the model 3.2. Then, with probability $1 - O(n^{-1})$,

$$\min_{(i,j)\in\mathcal{E}\setminus\mathcal{C}} \beta_{ij} = 1 - O\left(\frac{1}{n^\gamma}\right), \quad \max_{(i,j)\notin\mathcal{E}\cup\mathcal{C}} \beta_{ij} \lesssim \frac{1}{n^{1+\gamma}},$$
$$\min_{(i,i)\in\mathcal{E}\cap\mathcal{C}} \beta_i = \nu - 1 + O\left(\frac{1}{n^\gamma}\right), \quad \max_{(i,i)\in\mathcal{C}\setminus\mathcal{E}} \beta_i = \nu - O\left(\frac{1}{n^\gamma}\right).$$

Based on Lemma 7.2, we can show that Algorithm 4 can recover the core features:

Theorem 7.1. *Suppose that Assumptions 3.1, 3.2, 4.2, and 7.1 hold. Suppose that $p_n \leq 1 - \eta$ for some constant $\eta > 0$. Fix any $\gamma > 1$, $\nu \geq 1.1\eta^{-1}$. Choose τ as in Lemma 7.2. Let G_1 and G_2 be the representations obtained from Algorithm 4 applied to the data generated from equation 3.2. Suppose that n satisfies equation 7.4. Then, with probability $1 - O(n^{-1})$,*

$$\|\sin\Theta(P_r(G_1), U_1^*)\|_F \vee \|\sin\Theta(P_r(G_2), U_2^*)\|_F \lesssim \sqrt{r} \wedge \sqrt{\frac{r(r + r(\Sigma_\xi) + r(\Sigma_{\tilde{\xi}}))\log(n + d_1 + d_2)}{n}},$$

and

$$\|G_1^\top G_2 - (\nu - 1 - \nu p_n) U_1^* \Sigma_z^{1/2} \Sigma_{\tilde{z}}^{1/2} U_2^{*\top}\| \leq c_q \frac{r}{(r + r(\Sigma_\xi))(r + r(\Sigma_{\tilde{\xi}}))\log n}. \tag{7.5}$$

It is noted that approximated multimodal contrastive learning can learn representations even in the presence of noisy pairs. equation 7.5 implies that we can further iterate the procedure by obtained representations G_1, G_2 to obtain the same theoretical guarantee.

7.4 Extension to More Than Three Modalities

Here we discuss the extension of MMCL to the case where data have more than three modalities. Specifically, we observe n data $(x_i^\mu)_{i=1}^n \subset \mathbb{R}^{d_\mu}$ from μ-th modality for all $\mu \in [M]$, where M is the number of modalities. As in the main body, we focus on linear representations. We train M linear representations $G_\mu \in \mathbb{R}^{r\times d_\mu}$ for each modality. Since the loss 2.1 is designed to contrast two modalities, one possible extension to multiple modalities is to sum up the contrastive loss for all pairs of modalities; we define the additive multimodal contrastive loss as follows:

$$\mathcal{L}_M(G_1, \ldots, G_M) \triangleq \sum_{1\leq\mu_1<\mu_2\leq M} \mathcal{L}(G_{\mu_1}, G_{\mu_2}), \tag{7.6}$$

where $s_{ij}^{\mu_1,\mu_2} \triangleq \langle G_{\mu_1} x_i^{\mu_1}, G_{\mu_2} x_j^{\mu_2}\rangle$ and

$$\mathcal{L}(G_{\mu_1}, G_{\mu_2}) = \frac{1}{2C_n}\sum_i \phi\left(\epsilon\psi(0) + \sum_{j:j\neq i}\psi(s_{ij}^{\mu_1,\mu_2} - s_{ii}^{\mu_1,\mu_2})\right)$$
$$+ \frac{1}{2C_n}\sum_i \phi\left(\epsilon\psi(0) + \sum_{j:j\neq i}\psi(s_{ji}^{\mu_1,\mu_2} - s_{ii}^{\mu_1,\mu_2})\right) + R(G_{\mu_1}, G_{\mu_2}).$$

Since this is a simple addition of the contrastive loss, its gradient is also a sum of the gradients. We minimize the loss 7.6 via coordinate descent; given the set of $G_1^{(t)}, \ldots, G_M^{(t)}$ at step t, we obtain $G_1^{(t+1)}, \ldots, G_M^{(t+1)}$ by

$$G_\mu^{(t+1)} = G_\mu^{(t)} - \iota \left.\frac{\partial \mathcal{L}_M}{\partial G_\mu}\right|_{G_1=G_1^{(t)},\ldots,G_M=G_M^{(t)}},$$

where $\iota > 0$ is the learning rate.

Then, we have the following result as a corollary from Proposition 8.1.

Corollary 7.1 (Corollary of Proposition 2.1)**.** *Consider minimizing the nonlinear loss function $\mathcal{L}_M$ defined in equation 7.6 by coordinate descent. Suppose that the regularizer R is symmetric, i.e., $R(G_{\mu_1}, G_{\mu_2}) = R(G_{\mu_2}, G_{\mu_1})$ for any $\mu_1 \neq \mu_2$. Then,*

$$\frac{\partial \mathcal{L}_M}{\partial G_\mu} = -\frac{\partial}{\partial G_\mu} \sum_{\mu' \neq \mu} \operatorname{tr}\bigl(G_\mu S_{\mu,\mu'}(\beta) G_{\mu'}^\top\bigr)\Big|_{\beta=\beta(G_1,\dots,G_M)} + \frac{\partial}{\partial G_\mu} \sum_{\mu' \neq \mu} R(G_\mu, G_{\mu'}), \qquad \mu \in [M],$$

where the contrastive cross-covariance $S_{\mu,\mu'}(\beta)$ is given by:

$$S_{\mu,\mu'}(\beta) = \frac{1}{C_n} \sum_{i=1}^n \beta_i^{\mu,\mu'} x_i^\mu x_i^{\mu'} - \frac{1}{C_n} \sum_{i \neq j} \beta_{ij}^{\mu,\mu'} x_i^\mu x_j^{\mu'},$$
$$\beta_{ij}^{\mu,\mu'} \triangleq \frac{\alpha_{ij}^{\mu,\mu'} + \bar{\alpha}_{ji}^{\mu,\mu'}}{2}, \quad \beta_i^{\mu,\mu'} \triangleq \nu \sum_{j \in [n]} \frac{\alpha_{ij}^{\mu,\mu'} + \bar{\alpha}_{ij}^{\mu,\mu'}}{2} - 1,$$

with

$$\alpha_{ij}^{\mu,\mu'} = \epsilon_{ij} \phi'\left(\sum_{j' \in [n]} \epsilon_{ij} \psi(s_{ij'}^{\mu,\mu'} - \nu s_{ii}^{\mu,\mu'}) \right) \psi'(s_{ij}^{\mu,\mu'} - \nu s_{ii}^{\mu,\mu'}),$$
$$\bar{\alpha}_{ij}^{\mu,\mu'} = \epsilon_{ij} \phi'\left(\sum_{j' \in [n]} \epsilon_{ij} \psi(s_{j'i}^{\mu,\mu'} - \nu s_{ii}^{\mu,\mu'}) \right) \psi'(s_{ji}^{\mu,\mu'} - \nu s_{ii}^{\mu,\mu'}),$$

where $\nu \geq 1$ and $\epsilon_{ij} = 1$ for $i \neq j$ and $\epsilon_{ij} = \epsilon$ for $i = j$.

Corollary 7.1 shows that when $R(G_\mu, G_{\mu'}) = \|G_{\mu_1}^\top G_{\mu_2}\|_F^2$, each step of gradient descent in minimizing the additive contrastive loss 7.6 can be interpreted as maximizing the sum of SVD objectives, which is an analogy of the results in Section 2.

8 Proofs

8.1 Proof of Lemma 2.1

Here we prove Lemma 2.1.

Proof. Observe that

$$-2 \operatorname{tr}\bigl(G_1 S G_2^\top\bigr) + \rho \|G_1^\top G_2\|_F^2 = \left\| \rho^{1/2} G_1^\top G_2 - \frac{1}{\rho^{1/2}} S \right\|_F^2 - \frac{1}{\rho} \|S\|_F^2.$$

Eckart-Young-Mirsky theorem (see, for example, Theorem 2.4.8 in Golub and Van Loan (2013)) concludes the proof. □

8.2 Proof of Proposition 2.1

Before going to the proof, we restate Proposition 2.1 in a slightly generalized way. Suppose that we have parameters θ_1 and θ_2 such that $G_1 = G_1(\theta_1)$ and $G_2 = G_2(\theta_2)$ are smooth functions of θ_1 and θ_2, respectively.

Define the loss function $\mathcal{L}'$ as

$$\mathcal{L}'(\theta_1, \theta_2) \triangleq \frac{1}{2C_n} \sum_i \phi\left(\sum_{j \in [n]} \epsilon_{ij} \psi(s_{ij} - \nu s_{ii}) \right) + \frac{1}{2C_n} \sum_i \phi\left(\sum_{j \in [n]} \epsilon_{ij} \psi(s_{ji} - \nu s_{ii}) \right) + R(\theta_1, \theta_2), \tag{8.1}$$

where $\nu \geq 1$ and $\epsilon_{ij} = 1$ for $i \neq j$ and $\epsilon_{ij} = \epsilon$ for $i = j$. Recall that $s_{ij} = \langle G_1(\theta_1) x_i, G_2(\theta_j) \tilde{x}_j \rangle$. The loss in equation 2.1 corresponds to the loss in equation 8.1 with $\nu = 1$ and $G_k = \theta_k \in \mathbb{R}^{r \times d_k}$ for $k = 1, 2$. Setting $\nu > 0$ corresponds to choosing different temperature parameters for positive pairs and negative pairs.

Proposition 8.1 (Restatement of Proposition 2.1)**.** *Consider minimizing the nonlinear loss function $\mathcal{L}'(\theta_1, \theta_2)$ defined in equation 8.1. Then,*

$$\frac{\partial \mathcal{L}'}{\partial \theta_k} = -\left.\frac{\partial \operatorname{tr}\big(G_1 S(\beta) G_2^\top\big)}{\partial \theta_k}\right|_{\beta=\beta(G_1,G_2)} + \frac{\partial R(G_1, G_2)}{\partial \theta_k}, \qquad k \in \{1, 2\},$$

where the contrastive cross-covariance $S(\beta)$ is given by:

$$S(\beta) = \frac{1}{C_n}\sum_{i=1}^n \beta_i x_i \tilde{x}_i^\top - \frac{1}{C_n}\sum_{i\neq j} \beta_{ij} x_i \tilde{x}_j^\top, \;\; \beta_{ij} = \frac{\alpha_{ij} + \bar{\alpha}_{ji}}{2}, \;\; \beta_i = \nu \sum_{j\in[n]} \frac{\alpha_{ij} + \bar{\alpha}_{ij}}{2} - 1,$$

with

$$\alpha_{ij} = \epsilon_{ij}\phi'\left(\sum_{j'\in[n]} \epsilon_{ij}\psi(s_{ij'} - \nu s_{ii})\right)\psi'(s_{ij} - \nu s_{ii}), \;\; \bar{\alpha}_{ij} = \epsilon_{ij}\phi'\left(\sum_{j'\in[n]} \epsilon_{ij}\psi(s_{j'i} - \nu s_{ii})\right)\psi'(s_{ji} - \nu s_{ii}).$$

Proof. Let $\theta_{2,\ell}$ be the k-th component of θ_2. Observe that

$$\begin{aligned}\partial_{\theta_{2,\ell}}\mathcal{L}' = {} & \frac{1}{2C_n}\sum_{i=1}^n \phi'\left(\sum_{j=1}^n \epsilon_{ij}\psi(s_{ij} - \nu s_{ii})\right)\sum_{j=1}^n \epsilon_{ij}\psi'(s_{ij} - \nu s_{ii})(\partial_{\theta_{2,\ell}} s_{ij} - \nu \partial_{\theta_{2,\ell}} s_{ii}) \\ & + \frac{1}{2C_n}\sum_{i=1}^n \phi'\left(\sum_{j=1}^n \epsilon_{ij}\psi(s_{ji} - \nu s_{ii})\right)\sum_{j=1}^n \epsilon_{ij}\psi'(s_{ji} - \nu s_{ii})(\partial_{\theta_{2,\ell}} s_{ji} - \nu \partial_{\theta_{2,\ell}} s_{ii}) + \partial_{\theta_{2,\ell}} R.\end{aligned}$$

Define $\alpha_{ij} \triangleq \epsilon_{ij}\phi'(\sum_{j'\in[n]} \epsilon_{ij'}\psi(s_{ij'} - \nu s_{ii}))\psi'(s_{ij} - \nu s_{ii})$ and $\bar{\alpha}_{ij} \triangleq \epsilon_{ij}\phi'(\sum_{j'\in[n]} \epsilon_{ij'}\psi(s_{j'i} - \nu s_{ii}))\psi'(s_{ji} - \nu s_{ii})$. Then,

$$\partial_{\theta_{2,\ell}}\mathcal{L}' = \frac{1}{2C_n}\sum_{i=1}^n\sum_{j=1}^n \alpha_{ij}(\partial_{\theta_{2,\ell}} s_{ij} - \nu \partial_{\theta_{2,\ell}} s_{ii}) + \frac{1}{2C_n}\sum_{i=1}^n\sum_{j=1}^n \bar{\alpha}_{ij}(\partial_{\theta_{2,\ell}} s_{ji} - \nu \partial_{\theta_{2,\ell}} s_{ii}) + \partial_{\theta_{2,\ell}} R.$$

This further gives

$$\begin{aligned}-\partial_{\theta_{2,\ell}}\mathcal{L}' &= \frac{1}{2C_n}\sum_{i,j\in[n]} \big[\nu(\alpha_{ij} + \bar{\alpha}_{ij})\partial_{\theta_{2,\ell}} s_{ii} - \alpha_{ij}\partial_{\theta_{2,\ell}} s_{ij} - \bar{\alpha}_{ij}\partial_{\theta_{2,\ell}} s_{ji}\big] + \partial_{\theta_{2,\ell}} R \\ &= \frac{1}{C_n}\sum_i \left(\nu \sum_{j\in[n]} \frac{\alpha_{ij} + \bar{\alpha}_{ij}}{2} - 1\right)\partial_{\theta_{2,\ell}} s_{ii} - \frac{1}{C_n}\sum_{i\neq j} \frac{\alpha_{ij} + \bar{\alpha}_{ji}}{2}\partial_{\theta_{2,\ell}} s_{ij} + \partial_{\theta_{2,\ell}} R \\ &= \frac{1}{C_n}\sum_i \beta_i \partial_{\theta_{2,\ell}} s_{ii} - \frac{1}{C_n}\sum_{i\neq j} \beta_{ij}\partial_{\theta_{2,\ell}} s_{ij} + \partial_{\theta_{2,\ell}} R.\end{aligned}$$

Since $\partial_{\theta_{2,\ell}} s_{ij} = \partial_{\theta_{2,\ell}} \operatorname{tr}\big(G_1 x_i \tilde{x}_j^\top G_2^\top\big)$, when β_i and β_{ij} do not depend on θ_1 and $\theta_{2,\ell}$, we obtain

$$-\partial_{\theta_{2,\ell}}\mathcal{L}' = \partial_{\theta_{2,\ell}} \operatorname{tr}\left(G_1\left(\frac{1}{C_n}\sum_i \beta_i x_i \tilde{x}_i^\top\right)G_2^\top\right) - \partial_{\theta_{2,\ell}} \operatorname{tr}\left(G_1\left(\frac{1}{C_n}\sum_{i\neq j} \beta_{ij} x_i \tilde{x}_j^\top\right)G_2^\top\right) + \partial_{\theta_{2,\ell}} R.$$

This yields the claim for $k = 2$ case. By symmetry, we have a similar result for $k = 1$. □

8.3 Proof of Theorem 3.1

We restate Theorem 3.1.

Theorem 8.1 (Restatement of Theorem 3.1)**.** *Suppose that we have a collection of pairs $(x_i, \tilde{x}_i)_{i=1}^n$ generated according to the model 3.2. Suppose Assumptions 3.1 and 3.2 hold. Let G_1 and G_2 be the solution to minimizing the loss 3.1. Then, with probability $1 - O(n^{-1})$, there exists some constant $C = C(\sigma, s_1, s_2, \kappa_z^2, \kappa_{\tilde{z}}^2) > 0$ such that*

$$\left\| G_1^\top G_2 - \frac{1}{\rho} U_1^* \Sigma_z^{1/2} \Sigma_{\tilde{z}}^{1/2} U_2^{*\top} \right\| \le \frac{C}{\rho} \sqrt{\frac{(r + r(\Sigma_\xi) + r(\Sigma_{\tilde{\xi}})) \log(n + d_1 + d_2)}{n}},$$

and

$$\begin{aligned}
&\| \sin\Theta(P_r(G_1), U_1^*) \|_F \vee \| \sin\Theta(P_r(G_2), U_2^*) \|_F \\
&\lesssim \sqrt{r} \wedge \frac{\sqrt{r}}{1 - p_n} \sqrt{\frac{(r + r(\Sigma_\xi) + r(\Sigma_{\tilde{\xi}})) \log(n + d_1 + d_2)}{n}} \left(1 + \frac{1}{1 - p_n} \sqrt{\frac{(r + r(\Sigma_\xi) + r(\Sigma_{\tilde{\xi}})) \log(n + d_1 + d_2)}{n}} \right).
\end{aligned}$$

Let $\varepsilon_n \triangleq \sqrt{(r + r(\Sigma_\xi) + r(\Sigma_{\tilde{\xi}})) \log(n + d_1 + d_2)/n}$. From Theorem 8.1, we can see that the feature recovery bound becomes a trivial bound $\sqrt{r}$ when $1 - p_n \lesssim \varepsilon_n$. Otherwise the feature recovery ability is bounded by $(1 - p_n)^{-1} \sqrt{r} \varepsilon_n$. This implies that even if the portion of noisy pairs grows, MMCL can still recover the core features as $n \to \infty$ as long as the growth is mild.

Corollary 8.1. *Assume the same conditions as in Theorem 8.1. If $p_n \le 1 - \eta$ for some $\eta \in (0, 1]$, then*

$$\| \sin\Theta(P_r(G_1), U_1^*) \|_F \vee \| \sin\Theta(P_r(G_2), U_2^*) \|_F \lesssim \sqrt{r} \wedge \frac{1}{\eta} \sqrt{\frac{r(r + r(\Sigma_\xi) + r(\Sigma_{\tilde{\xi}})) \log(n + d_1 + d_2)}{n}}.$$

Proof. Since the loss function 2.1 can be rewritten as

$$\begin{aligned}
&- \operatorname{tr}\left(G_1 \left(\frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})(\tilde{x}_i - \bar{\tilde{x}})^\top \right) G_2^\top \right) + (\rho^2/2) \| G_1^\top G_2 \|_F^2 \\
&= (\rho^2/2) \left\| G_1^\top G_2 - \frac{1}{\rho^2} \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})(\tilde{x}_i - \bar{\tilde{x}})^\top \right\|_F^2 - \frac{1}{2\rho^2} \left\| \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})(\tilde{x}_i - \bar{\tilde{x}})^\top \right\|_F^2,
\end{aligned}$$

where $\bar{x} = (1/n) \sum_i x_i$ and $\bar{\tilde{x}} = (1/n) \sum_i \tilde{x}_i$. By Eckart-Young-Mirsky theorem (see, for example, Theorem 2.4.8 in Golub and Van Loan (2013)), we have $G_1^\top G_2 = \mathrm{SVD}_r(\rho^{-1}(n-1)^{-1} \sum_{i=1}^n (x_i - \bar{x})(\tilde{x}_i - \bar{\tilde{x}})^\top)$. For notational brevity, define $\Sigma = \Sigma_z^{1/2} \Sigma_{\tilde{z}}^{1/2}$ and $\bar{S} \triangleq (n-1)^{-1} \sum_{i=1}^n (x_i - \bar{x})(\tilde{x}_i - \bar{\tilde{x}})^\top$. Observe that

$$\begin{aligned}
\left\| \mathrm{SVD}_r(\bar{S}) - U_1^* \Sigma U_2^{*\top} \right\| &\le \left\| \mathrm{SVD}_r(\bar{S}) - \bar{S} \right\| + \left\| \bar{S} - U_1^* \Sigma U_2^{*\top} \right\| \\
&= \lambda_{r+1}(\bar{S}) + \left\| \bar{S} - U_1^* \Sigma U_2^{*\top} \right\| \\
&\le \lambda_{r+1}(U_1^* \Sigma U_2^{*\top}) + 2 \left\| \bar{S} - U_1^* \Sigma U_2^{*\top} \right\| \qquad (8.2) \\
&= 2 \left\| \bar{S} - U_1^* \Sigma U_2^{*\top} \right\|. \qquad (8.3)
\end{aligned}$$

Note that

$$\frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})(\tilde{x}_i - \bar{\tilde{x}})^\top = \frac{1}{n-1} \sum_{i=1}^n x_i \tilde{x}_i^\top - \frac{n}{n-1} \bar{x} \bar{\tilde{x}}^\top.$$

Since x_i is independent sub-Gaussian random vectors with parameter $\sqrt{\sigma_z^2 + \sigma_\xi^2}$, from Hoeffding bound (see, for example, Proposition 2.5 in Wainwright (2019)), it holds with probability $1 - O(n^{-1})$ that

$$\| \bar{x} \| \le \sqrt{\frac{2(\sigma_z^2 + \sigma_\xi^2) \log(n d_1)}{n}} \lesssim \sqrt{\frac{\log(n d_1)}{n}}, \qquad (8.4)$$

where the last inequality follows from Assumption 3.2 and $\sigma_z^2 + \sigma_\xi^2 \le \sigma(1 + s_1^{-1})$. A similar bound holds for $\bar{\tilde{x}}$.

For the term $(n-1)^{-1}\sum_i x_i\tilde{x}_i^\top$, observe that

$$\frac{1}{n}\sum_{i\in[n]} x_i\tilde{x}_i^\top = \frac{1}{n}\sum_{(i,i)\in\mathcal{C}\cap\mathcal{E}} x_i\tilde{x}_i^\top + \frac{1}{n}\sum_{(i,i)\in\mathcal{C}\setminus\mathcal{E}} x_i\tilde{x}_i^\top \triangleq T_1 + Q_1.$$

Suppose for a moment that $0 < m < n$.

For the term T_1, note that

$$T_1 = U_1^*\left(\frac{1}{n}\sum_{(i,i)\in\mathcal{C}\cap\mathcal{E}} z_i\tilde{z}_i^\top\right)U_2^{*\top} + \frac{1}{n}\sum_{(i,i)\in\mathcal{C}\cap\mathcal{E}} U_1^* z_i\tilde{\xi}_i^\top + \frac{1}{n}\sum_{(i,i)\in\mathcal{C}\cap\mathcal{E}} \xi_i\tilde{z}_i^\top U_2^{*\top} + \frac{1}{n}\sum_{(i,i)\in\mathcal{C}\cap\mathcal{E}} \xi_i\tilde{\xi}_i^\top.$$

Using Proposition 9.1 and Assumption 3.2, the following bound holds with probability $1-O(n^{-1})$:

$$\begin{aligned}
\Big\|\frac{1}{m}\sum_{(i,i)\in\mathcal{C}\cap\mathcal{E}} z_i\tilde{z}_i^\top - \Sigma_z^{1/2}\Sigma_{\tilde{z}}^{1/2}\Big\| &\lesssim (\|\Sigma_z\|\vee\|\Sigma_{\tilde{z}}\|)\sqrt{\frac{r\log(nr)}{m}},\\
\Big\|\frac{1}{m}\sum_{(i,i)\in\mathcal{C}\cap\mathcal{E}} z_i\tilde{\xi}_i^\top\Big\| &\lesssim \|\Sigma_z\|^{1/2}\|\Sigma_{\tilde{z}}\|^{1/2}\sqrt{\frac{(r+r(\Sigma_{\tilde{\xi}}))\log(nr+nd_2)}{m}},\\
\Big\|\frac{1}{m}\sum_{(i,i)\in\mathcal{C}\cap\mathcal{E}} \xi_i\tilde{z}_i^\top\Big\| &\lesssim \|\Sigma_z\|^{1/2}\|\Sigma_{\tilde{z}}\|^{1/2}\sqrt{\frac{(r+r(\Sigma_{\xi}))\log(nr+nd_1)}{m}},\\
\Big\|\frac{1}{m}\sum_{(i,i)\in\mathcal{C}\cap\mathcal{E}} \xi_i\tilde{z}_i^\top\Big\| &\lesssim \|\Sigma_z\|^{1/2}\|\Sigma_{\tilde{z}}\|^{1/2}\sqrt{\frac{(r(\Sigma_{\xi})+r(\Sigma_{\tilde{\xi}}))\log(nd_1+nd_2)}{m}}.
\end{aligned}$$

Thus, with probability $1-O(n^{-1})$,

$$\|T_1 - (1-p_n)U_1^*\Sigma_z^{1/2}\Sigma_{\tilde{z}}^{1/2}U_2^{*\top}\| \lesssim \sqrt{1-p_n}\sqrt{\frac{(r+r(\Sigma_\xi)+r(\Sigma_{\tilde{\xi}}))\log(nd_1+nd_2)}{n}}. \tag{8.5}$$

Next we bound $\|Q_1\|$. Note that Q_1 is a sum of $n-m$ independent random variables. Using Proposition 9.1,

$$\begin{aligned}
\Big\|\frac{n}{n-m}Q_1\Big\| &\lesssim (\mathrm{tr}(\Sigma_{\tilde{x}})\|\Sigma_x\|\vee\mathrm{tr}(\Sigma_x)\|\Sigma_{\tilde{x}}\|)^{1/2}\sqrt{\frac{\log(nd_1+nd_2)}{n-m}} &(8.6)\\
&\le \Big(\|\Sigma_{\tilde{z}}\|(r+s_2^{-2}r(\Sigma_{\tilde{\xi}}))^{1/2}(1+s_2^{-2})^{1/2}\vee\|\Sigma_z\|(r+s_1^{-2}r(\Sigma_\xi))^{1/2}(1+s_1^{-2})^{1/2}\Big) &(8.7)\\
&\quad\times\sqrt{\frac{\log(nd_1+nd_2)}{n-m}} &(8.8)\\
&\lesssim \sqrt{\frac{(r+r(\Sigma_\xi)+r(\Sigma_{\tilde{\xi}}))\log(nd_1+nd_2)}{n-m}} &(8.9)
\end{aligned}$$

holds with probability at least $1-n^{-1}$, where the last inequality holds from Assumption 3.2.

Therefore, combined with equation 8.4,

$$\begin{aligned}
\big\|\bar{S} - (1-p_n)U_1^*\Sigma U_2^{*\top}\big\| &\lesssim \sqrt{\frac{(r+r(\Sigma_\xi)+r(\Sigma_{\tilde{\xi}}))\log(nd_1+nd_2)}{n}} + \frac{1-p_n}{n-1} + \sqrt{\frac{\log(nd_1)}{n}}\\
&\lesssim \sqrt{\frac{(r+r(\Sigma_\xi)+r(\Sigma_{\tilde{\xi}}))\log(nd_1+nd_2)}{n}}
\end{aligned}$$

holds with probability at least $1-O(n^{-1})$. If $m=0$ or $m=n$, a similar argument gives the same bound with probability $1-O(n^{-1})$. From equation 8.3, we obtain

$$\left\|G_1^\top G_2 - \frac{1}{\rho}U_1^*\Sigma U_2^{*\top}\right\| \lesssim \frac{1}{\rho}\sqrt{\frac{(r+r(\Sigma_\xi)+r(\Sigma_{\tilde{\xi}}))\log(nd_1+nd_2)}{n}}$$

with probability at least $1 - O(n^{-1})$.

Define $\Sigma' \triangleq (1-p_n)U_1^* \Sigma U_2^{*\top}$. Since $\lambda_{\max}(\Sigma) \le (1-p_n)$ and $\lambda_{\min}(\Sigma') \ge (1-p_n)\lambda_{\min}(\Sigma_z^{1/2}\Sigma_{\tilde{z}}^{1/2})$, from Theorem 3 in Yu et al. (2015) with $s \leftarrow r$, $r \leftarrow 1$,

$$\begin{aligned}\|\sin\Theta(P_r(G_1), U_1^*)\|_F \vee \|\sin\Theta(P_r(G_2), U_2^*)\|_F &\le \sqrt{r} \wedge \frac{2(2(1-p_n)\lambda_{\max}(\Sigma_z^{1/2}\Sigma_{\tilde{z}}^{1/2}) + \|\bar{S} - \Sigma'\|)r^{1/2}\|\bar{S} - \Sigma'\|}{(1-p_n)^2\lambda_{\min}^2(\Sigma_z^{1/2}\Sigma_{\tilde{z}}^{1/2})} \\ &\lesssim \sqrt{r} \wedge \sqrt{r}\left(1 + \frac{\|\bar{S} - \Sigma'\|}{1-p_n}\right)\frac{\|\bar{S} - \Sigma'\|}{1-p_n}.\end{aligned}$$

where in the second inequality we used Assumption 3.1. $\log(nd_1 + nd_2) \lesssim \log(n + d_1 + d_2)$ concludes the proof. □

8.4 Proof of Lemma 7.2

Before going to the proof of Lemma 7.2, we introduce the following result.

Lemma 8.1. *Suppose Assumptions 3.1, 3.2, 4.2 and 7.1 hold. Let $(x_i, \tilde{x}_i)_{i=1}^n$ be the paired data generated from the model described in Section 3. Let $\mathcal{E}$ be the set of ground-truth pairs. Let $s_{ij} = \langle G_1^{(0)} x_i, G_2^{(0)} \tilde{x}_i \rangle$ be the similarity evaluated by the initial representations. Then, there exists some constant $C' = C'(\sigma, s_1, s_2, \kappa_z^2, \kappa_{\tilde{z}}^2, q) > 0$ satisfying*

$$\min_{(i_1,j_1)\in\mathcal{E}} \left(s_{i_1 j_1} - \max_{j:(i_1,j)\notin\mathcal{E}} s_{i_1,j} \vee \max_{i:(i,j_1)\notin\mathcal{E}} s_{i,j_1} \right) \ge C'\sqrt{r \log n}.$$

Using Lemma 8.1, we prove Lemma 7.2.

Proof. Fix any $(i_1, j_1) \in \mathcal{E} \setminus \mathcal{C}$. From equation 7.3,

$$\beta_{i_1 j_1} = \frac{1}{2}\frac{\exp(s_{i_1 j_1}/\tau)}{\sum_{j':j'\in[n]} \exp(s_{i_1 j'}/\tau)} + \frac{1}{2}\frac{\exp(s_{i_1 j_1}/\tau)}{\sum_{i':i'\in[n]} \exp(s_{i' j_1}/\tau)}.$$

Taking any $\delta \le C'\sqrt{r \log n}/(\gamma + 1)$ and choosing the temperature parameter as $\tau = \delta/\log n$ gives,

$$\frac{e^{s_{i_1 j_1}/\tau}}{\sum_{j':j'\in[n]} e^{s_{i_1 j'}/\tau}} \ge \frac{n^{s_{i_1 j_1}/\delta}}{n^{\max_{j':j'\neq j_1} s_{i_1 j'}/\delta + 1} + n^{s_{i_1 j_1}/\delta}} = \frac{1}{n^{-(s_{i_1 j_1}/\delta - \max_{j':j'\neq j_1} s_{i_1 j'}/\delta - 1)} + 1}.$$

From the above argument, we can see that

$$s_{i_1 j_1}/\delta - \max_{j':j'\neq i_1} s_{i_1 j'}/\delta \ge \gamma + 1$$

with probability $1 - O(n^{-1})$. Hence

$$\frac{e^{s_{i_1 j_1}/\tau}}{\sum_{j':j'\in[n]} e^{s_{i_1 j'}/\tau}} = 1 - O(n^{-\gamma})$$

holds uniformly over all $(i_1, j_1) \in \mathcal{E} \setminus \mathcal{C}$. Similarly, we have

$$\frac{\exp(s_{i_1 j_1}/\tau)}{\sum_{i':i'\in[n]} \exp(s_{i' j_1}/\tau)} = 1 - O(n^{-\gamma})$$

holds uniformly over all $(i_1, j_1) \in \mathcal{E} \setminus \mathcal{C}$. These give $\min_{(i_1,j_1)\in\mathcal{E}\setminus\mathcal{C}} \beta_{i_1 j_1} = 1 - O(n^{-\gamma})$.

For any $(i_1, j_1) \notin \mathcal{E} \cup \mathcal{C}$, take another node j_2 satisfying $(i_1, j_2) \in \mathcal{E} \setminus \mathcal{C}$. Then, by a similar argument

$$\frac{e^{s_{i_1 j_1}/\tau}}{\sum_{j':j'\in[n]} e^{s_{i_1 j'}/\tau}} \le \frac{n^{s_{i_1 j_1}/\delta + 1} n^{-1}}{n^{s_{i_1 j_2}/\delta}} \le \frac{1}{n^{1+\gamma}}$$

holds with probability $1 - O(n^{-1})$. Similarly,

$$\frac{\exp(s_{i_1 j_1}/\tau)}{\sum_{i':i' \in [n]} \exp(s_{i' j_1}/\tau)} \lesssim n^{-\gamma}.$$

Since this bound is uniform over all $(i_1, j_1) \notin \mathcal{E} \cup \mathcal{C}$, we obtain $\max_{(i_1,j_1)\notin \mathcal{E}\cup\mathcal{C}} \beta_{i_1 j_1} \leq n^{-(1+\gamma)}$.

For β_{i_1}, note that we can rewrite it as

$$\beta_{i_1} = \nu - 1 + \frac{1}{2}\left(\frac{\sum_{j:j\neq i_1} e^{s_{i_1 j}/\tau}}{\sum_{j:j\in[n]} e^{s_{i_1 j}/\tau}} + \frac{\sum_{j:j\neq i_1} e^{s_{j i_1}/\tau}}{\sum_{j:j\in[n]} e^{s_{j i_1}/\tau}} \right).$$

Similar to the above arguments, if $(i_1, i_1) \in \mathcal{E}$,

$$\frac{\sum_{j:j\neq i_1} e^{s_{i_1 j}/\tau}}{\sum_{j:j\in[n]} e^{s_{i_1 j}/\tau}} \lesssim \frac{n^{\max_{j:j\neq i_1} s_{i_1 j}/\delta + 1}}{n^{s_{i_1 i_1}/\delta}} \lesssim n^{-\gamma}.$$

Similarly,

$$\frac{\sum_{j:j\neq i_1} e^{s_{j i_1}/\tau}}{\sum_{j:j\in[n]} e^{s_{j i_1}/\tau}} \lesssim n^{-\gamma}.$$

Thus $\beta_{i1} = \nu - 1 + O(n^{-\gamma})$. If $(i_1, i_1) \notin \mathcal{E}$, there exists some $j_1 \neq i_1$ such that

$$\frac{\sum_{j:j\neq i_1} e^{s_{i_1 j}/\tau}}{\sum_{j:j\in[n]} e^{s_{i_1 j}/\tau}} \geq \frac{e^{s_{i_1 j_1}/\tau}}{e^{s_{i_1 i_1}/\tau} + \sum_{j:j\neq i_1} e^{s_{i_1 j}/\tau}} \geq \frac{n^{s_{i_1 j_1}/\delta}}{n^{s_{i_1 i_1}/\delta} + n^{\max_{j:j\neq i_1} s_{i_1 j}/\delta + 1}} = 1 - O(n^{-\gamma}).$$

By a similar argument, we obtain $\beta_{i_1} = \nu - O(n^{-\gamma})$. □

8.5 Proof of Lemma 8.1

Here we prove Lemma 8.1.

Proof. We first prove

$$\min_{(i_1,j_1)\in\mathcal{E}} \left(s_{i_1 j_1} - \max_{j:(i_1,j)\notin\mathcal{E}} s_{i_1,j} \right) \geq C' \sqrt{r \log n}.$$

Fix any $(i_1, j_1) \in \mathcal{E} \setminus \mathcal{C}$. Define $\Sigma = \Sigma_z^{1/2} \Sigma_{\tilde{z}}^{1/2}$ for notational brevity. Recall that $\|\Sigma_z\| = \|\Sigma_{\tilde{z}}\| = 1$. Since $\tilde{x}_{j_1} = U_2^* \tilde{z}_{j_1} + \tilde{\xi}_{j_1}$ is a sub-Gaussian random vector with parameter $\sqrt{\sigma_{\tilde{z}}^2 + \sigma_{\tilde{\xi}}^2}$, $x_{i_1}^\top U_1^* \Sigma U_2^{*\top} \tilde{x}_{j_1}$ is a sub-Gaussian random variable with parameter $\sqrt{\sigma_{\tilde{z}}^2 + \sigma_{\tilde{\xi}}^2}\|U_1^{*\top} x_{i_1}\|$ conditioned on x_{i_1}. Note that since $(i_1, j_1) \in \mathcal{E} \setminus \mathcal{C}$, $(x_{i_1}, \tilde{x}_{j_1})$ is independent of $\{(x_{i_1}, \tilde{x}_j : (i_1, j) \notin \mathcal{E}\}$. By Lemma 9.2 applied to independent sub-Gaussian random variables $(x_{i_1}^\top U_1^* \Sigma U_2^{*\top} \tilde{x}_j)_{j:(i_1,j)\notin\mathcal{E}}$ conditioned on x_{i_1} and $\tilde{x}_{j_1}$,

$$\begin{aligned}
&\mathbb{P}\left(x_{i_1}^\top U_1^* \Sigma U_2^{*\top} \tilde{x}_{j_1} - \max_{j:(i_1,j)\notin\mathcal{E}} x_{i_1}^\top U_1^* \Sigma U_2^{*\top} \tilde{x}_j \leq t\|U_1^{*\top} x_{i_1}\| \,\middle|\, x_{i_1}, \tilde{x}_{j_1} \right) \\
&= \mathbb{P}\Bigg(\max_{j:(i_1,j)\notin\mathcal{E}} x_{i_1}^\top U_1^* \Sigma U_2^{*\top} \tilde{x}_j - \sqrt{2(\sigma_{\tilde{z}}^2 + \sigma_{\tilde{\xi}}^2)\log(n-1)}\|U_1^{*\top} x_{i_1}\| \\
&\qquad \geq x_{i_1}^\top U_1^* \Sigma U_2^{*\top} \tilde{x}_{j_1} - t\|U_1^{*\top} x_{i_1}\| - \sqrt{2(\sigma_{\tilde{z}}^2 + \sigma_{\tilde{\xi}}^2)\log(n-1)}\|U_1^{*\top} x_{i_1}\| \,\Bigg|\, x_{i_1}, \tilde{x}_{j_1} \Bigg) \\
&\leq \exp\left(-\frac{1}{2(\sigma_{\tilde{z}}^2 + \sigma_{\tilde{\xi}}^2)\|U_1^{*\top} x_{i_1}\|^2} \left(x_{i_1}^\top U_1^* \Sigma U_2^{*\top} \tilde{x}_{j_1} - t\|U_1^{*\top} \tilde{x}_{j_1}\| - \sqrt{2(\sigma_{\tilde{z}}^2 + \sigma_{\tilde{\xi}}^2)\log(n-1)}\|U_1^{*\top} x_{i_1}\| \right)^2 \right) \\
&= \exp\left(-\frac{1}{2(\sigma_{\tilde{z}}^2 + \sigma_{\tilde{\xi}}^2)} \left(\frac{x_{i_1}^\top U_1^* \Sigma U_2^{*\top} \tilde{x}_{j_1}}{\|U_1^{*\top} x_{i_1}\|} - t - \sqrt{2(\sigma_{\tilde{z}}^2 + \sigma_{\tilde{\xi}}^2)\log(n-1)} \right)^2 \right).
\end{aligned} \tag{8.10}$$

Set $t \leftarrow t_0 \triangleq \sqrt{\log n}$. We further bound the far right hand side in equation 8.10. Note that by Assumption 3.2 and $\|\Sigma_z\| = \|\Sigma_{\tilde{z}}\| = 1$,

$$\sqrt{(\sigma_{\tilde{z}}^2 + \sigma_{\tilde{\xi}}^2)\log(n-1)} \le \sigma\sqrt{(1+s_2^{-2})}t_0.$$

From Lemma 9.1, there exists an event E with $\mathbb{P}(E) = 1 - O(n^{-1})$ such that on the event E the following holds uniformly for all $(i_1, j_1) \in \mathcal{E}$: there exists a constant $C_1 = C_1(\sigma, s_1, s_2) > 0$ satisfying

$$\begin{aligned}
&\frac{x_{i_1}^\top U_1^* \Sigma U_2^{*\top} \tilde{x}_{j_1}}{\|U_1^{*\top} x_{i_1}\|} - t_0 - \sqrt{2(\sigma_{\tilde{z}}^2 + \sigma_{\tilde{\xi}}^2)\log(n-1)} \\
&\ge \frac{z_{i_1}^\top \Sigma \tilde{z}_{j_1} - \max_{(i,j)\in\mathcal{E}} |\xi_i^\top U_1^* \Sigma \tilde{z}_j + z_i^\top \Sigma U_2^{*\top} \tilde{\xi}_j + \xi_i^\top U_1^* \Sigma U_2^{*\top} \tilde{\xi}_j|}{\max_{i\in[n]} \|U_1^{*\top} x_i\|} - t_0 - \sqrt{2(\sigma_{\tilde{z}}^2 + \sigma_{\tilde{\xi}}^2)\log(n-1)} \\
&\ge C_1 \left(\frac{\mathrm{tr}\left(\Sigma_z^{1/2} \Sigma \Sigma_{\tilde{z}}^{1/2}\right) - \|\Sigma_z\|^{1/2}\|\Sigma\|\|\Sigma_{\tilde{z}}\|^{1/2}\sqrt{r\log n}}{\sqrt{r}\|\Sigma_z\|^{1/2}} - \frac{\sqrt{r\log n}\|\Sigma_z\|^{1/2}\|\Sigma\|\|\Sigma_{\tilde{z}}\|^{1/2}}{\sqrt{r}\|\Sigma_z\|^{1/2}} - \sqrt{\log n} \right) \\
&\ge C_1(\sqrt{r}/(\kappa_z^2\kappa_{\tilde{z}}^2) - 3\sqrt{\log n}) \\
&\ge C_1\sqrt{r}\left(\frac{1}{\kappa_z^2\kappa_{\tilde{z}}^2} - 3\sqrt{c}\right),
\end{aligned}$$

where in the third inequality, we used

$$\mathrm{tr}\left(\Sigma_z^{1/2}\Sigma\Sigma_{\tilde{z}}^{1/2}\right) \ge r\lambda_{\min}(\Sigma_z)\lambda_{\min}(\Sigma_{\tilde{z}}) \ge r\|\Sigma_z\|\|\Sigma_{\tilde{z}}\|/(\kappa_z^2\kappa_{\tilde{z}}^2),$$

which holds by Assumption 3.1. Retaking $c \leftarrow c \wedge 2^{-1}(3\kappa_z^2\kappa_{\tilde{z}}^2)^2$, we obtain

$$\frac{x_{i_1}^\top U_1^* \Sigma U_2^{*\top} \tilde{x}_{j_1}}{\|U_1^{*\top} x_{i_1}\|} - t_0 - \sqrt{2(\sigma_{\tilde{z}}^2 + \sigma_{\tilde{\xi}}^2)\log(n-1)} \ge \frac{C_2\sqrt{r}}{2} \tag{8.11}$$

for some constant $C_2 = C_2(c, \sigma, s_1, s_2, \kappa_z^2, \kappa_{\tilde{z}}^2) > 0$.

Since

$$\sigma_{\tilde{z}}^2 + \sigma_{\tilde{\xi}}^2 \le \sigma^2\|\Sigma_{\tilde{z}}\|(1+s_2^{-2}),$$

equation 8.10 becomes

$$\exp\left(-\frac{1}{2(\sigma_{\tilde{z}}^2 + \sigma_{\tilde{\xi}}^2)}\left(\frac{x_{i_1}^\top U_1^* \Sigma U_2^{*\top} \tilde{x}_{j_1}}{\|U_1^{*\top} x_{i_1}\|} - t_0 - \sqrt{2(\sigma_{\tilde{z}}^2 + \sigma_{\tilde{\xi}}^2)\log(n-1)}\right)^2\right) \le \exp\left(-\frac{C_2^2 r}{8\sigma^2(1+s_2^{-2})}\right).$$

Retaking $c \leftarrow c \wedge 2C_2^2/(8\sigma^2(1+s_2^{-2}))$, we have

$$\exp\left(-\frac{C_2^2 r}{8\sigma^2(1+s_2^{-2})}\right) \le \exp(-2\log n) = \frac{1}{n^2}.$$

Therefore, by Lemma 9.1,

$$x_{i_1}^\top U_1^* \Sigma U_2^{*\top} \tilde{x}_{j_1} - \max_{j:(i_1,j)\notin\mathcal{E}} x_{i_1}^\top U_1^* \Sigma U_2^{*\top} \tilde{x}_j \ge t_0\|U_1^{*\top} x_{i_1}\| \ge \sqrt{\frac{r\log n}{2\kappa_z^{-2}}}$$

holds uniformly for all $(i_1, j_1) \in \mathcal{E}$ with probability $1 - O(n^{-1})$.

Furthermore,

$$\begin{aligned}
&x_{i_1}^\top G_1^\top G_2 \tilde{x}_{j_1} - \max_{j:(i_1,j)\notin\mathcal{E}} x_{i_1}^\top G_1^\top G_2 \tilde{x}_j \\
&= x_{i_1}^\top (G_1^\top G_2 - qU_1^* \Sigma U_2^{*\top})\tilde{x}_{j_1} + q x_{i_1}^\top U_1^* U_2^{*\top} \tilde{x}_{j_1} \\
&\quad - \max_{j:(i_1,j)\notin\mathcal{E}} [x_{i_1}^\top (G_1^\top G_2 - qU_1^* \Sigma U_2^{*\top})\tilde{x}_j + q x_{i_1}^\top U_1^* \Sigma U_2^{*\top} \tilde{x}_j] \\
&\ge -2\max_{i,j\in[n]} R_{ij} + q(x_{i_1}^\top U_1^* \Sigma U_2^{*\top} \tilde{x}_{j_1} - \max_{j:(i_1,j)\notin\mathcal{E}} x_{i_1}^\top U_1^* \Sigma U_2^{*\top} \tilde{x}_j),
\end{aligned}$$

where $R_{ij} \triangleq x_i^\top (G_1^\top G_2 - qU_1^* \Sigma U_2^{*\top}) \tilde{x}_j$. Note that

$$\max_{i,j\in[n]} R_{ij} \le \|G_1^\top G_2 - qU_1^* \Sigma U_2^{*\top}\| \max_{i\in[n]} \|x_i\| \max_{i\in[n]} \|\tilde{x}_i\|.$$

From Lemma 9.1 and by assumption, there exists a constant $C_3 = C_3(\sigma, s_1, s_2)$ satisfying

$$\max_{i,j\in[n]} R_{ij} \le C_3 \|G_1^\top G_2 - qU_1^* \Sigma U_2^{*\top}\| \|\Sigma_z\|^{1/2} \|\Sigma_{\tilde{z}}\|^{1/2} (r^{1/2} + r^{1/2}(\Sigma_\xi))(r^{1/2} + r^{1/2}(\Sigma_{\tilde{\xi}})) \log n \le C_3 c_q \sqrt{r \log n} \tag{8.12}$$

on the event E, where the last inequality follows from the assumption on $\|G_1^\top G_2 - qU_1^* \Sigma U_2^{*\top}\|$. We can take c_q small enough so that $C_3 c_q \le q\kappa_z/(4\sqrt{2})$. Then,

$$x_{i_1}^\top G_1^\top G_2 \tilde{x}_{j_1} - \max_{j:(i_1,j)\notin\mathcal{E}} x_{i_1}^\top G_1^\top G_2 \tilde{x}_j \ge q\kappa_z \sqrt{\frac{r\log n}{2}} - q\kappa_z \frac{1}{2}\sqrt{\frac{r\log n}{2}} = q\kappa_z \frac{1}{2}\sqrt{\frac{r\log n}{2}}.$$

Fix $\bar{C} > 0$ and let $H_{i_1,j_1}(\bar{C})$ be the event defined as

$$H_{i_1,j_1}(\bar{C}) \triangleq \left\{ x_{i_1}^\top G_1^\top G_2 \tilde{x}_{j_1} - \max_{j:(i_1,j)\notin\mathcal{E}} x_{i_1}^\top G_1^\top G_2 \tilde{x}_j \le \bar{C}\sqrt{r\log n} \right\}.$$

Then, from above arguments, there exists a constant $C' = C'(\sigma, s_1, s_2, \kappa_z^2, \kappa_{\tilde{z}}^2, q) > 0$ and a universal constant $c' > 0$ such that

$$\max_{(i_1,j_1)\in\mathcal{E}} \mathbb{P}(H_{i_1,j_1}(C') | x_{i_1}, \tilde{x}_{j_1}) \le cn^{-2}$$

holds in the event E. Observe

$$\begin{aligned}
\mathbb{P}\left(\bigcup_{(i_1,j_1)\in\mathcal{E}} H_{i_1,j_1}(C') \right) &= \mathbb{P}\left(E \cap \bigcup_{(i_1,j_1)\in\mathcal{E}} H_{i_1,j_1}(C') \right) + \mathbb{P}\left(E^c \cap \bigcup_{(i_1,j_1)\in\mathcal{E}} H_{i_1,j_1}(C') \right) \\
&\le \sum_{(i_1,j_1)\in\mathcal{E}} \mathbb{P}(H_{i_1,j_1}(C') \cap E) + \mathbb{P}(E^c).
\end{aligned}$$

Note that

$$\begin{aligned}
\mathbb{P}(H_{i_1,j_1}(C') \cap E) &\le \mathbb{P}(H_{i_1,j_1}(C') \cap \{\mathbb{P}(H_{i_1,j_1}(C') | x_{i_1}\tilde{x}_{j_1}) \le c' n^{-2}\}) \\
&= \mathbb{E}[\mathbb{E}[\mathbf{1}_{H_{i_1,j_1}(C')} \mathbf{1}\{\mathbb{P}(H_{i_1,j_1}(C') | x_{i_1}, \tilde{x}_{j_1}) \le c' n^{-2}\} | x_{i_1}, \tilde{x}_{j_1}]] \\
&= \mathbb{E}[\mathbf{1}\{\mathbb{P}(H_{i_1,j_1}(C') | x_{i_1}, \tilde{x}_{j_1}) \le c' n^{-2}\} \mathbb{P}(H_{i_1,j_1}(C') | x_{i_1}, \tilde{x}_{j_1})] \\
&\le c' n^{-2}.
\end{aligned}$$

Therefore $\mathbb{P}(\bigcup_{i_1,j_1} H_{i_1,j_i}(C')) \lesssim n^{-1}$ and thus

$$\min_{(i_1,j_1)\in\mathcal{E}} \left(s_{i_1 j_1} - \max_{j:(i_1,j)\notin\mathcal{E}} s_{i_1,j} \right) = \min_{(i_1,j_1)\in\mathcal{E}} \left(x_{i_1}^\top G_1^\top G_2 \tilde{x}_{j_1} - \max_{j:(i_1,j)\notin\mathcal{E}} x_{i_1}^\top G_1^\top G_2 \tilde{x}_j \right) \ge C'\sqrt{r\log n}$$

holds with probability $1 - O(n^{-1})$. A similar argument gives $\min_{(i_1,j_1)\in\mathcal{E}} \left(s_{i_1 j_1} - \max_{i:(i,j_1)\notin\mathcal{E}} s_{i,j_1} \right) \ge C'\sqrt{r\log n}$. □

8.6 Proof of Theorem 7.1

Here we prove Theorem 7.1.

Proof. By Lemma 7.2, we can rewrite S in Algorithm 4 as

$$\begin{aligned}
S &= \frac{1}{n}\sum_{(i,i)\in\mathcal{C}\cap\mathcal{E}}\beta_i x_i\tilde{x}_i^\top + \frac{1}{n}\sum_{(i,i)\in\mathcal{C}\setminus\mathcal{E}}\beta_i x_i\tilde{x}_i^\top - \frac{1}{n}\sum_{(i,j)\in\mathcal{E}\setminus\mathcal{C}}\beta_{ij} x_i\tilde{x}_j^\top - \frac{1}{n}\sum_{(i,j)\notin\mathcal{E}\cup\mathcal{C}}\beta_{ij} x_i\tilde{x}_j^\top \\
&= \frac{\nu-1}{n}\sum_{(i,i)\in\mathcal{C}\cap\mathcal{E}} x_i\tilde{x}_i^\top + \frac{\nu}{n}\sum_{(i,i)\in\mathcal{C}\setminus\mathcal{E}} x_i\tilde{x}_i^\top - \frac{1}{n}\sum_{(i,j)\in\mathcal{E}\setminus\mathcal{C}} x_i\tilde{x}_j^\top + \frac{1}{n}\sum_{(i,j)\in\mathcal{E}\setminus\mathcal{C}}(1-\beta_{ij})x_i\tilde{x}_j^\top \\
&\quad - \frac{1}{n}\sum_{(i,i)\in\mathcal{C}\cap\mathcal{E}}(\nu-1-\beta_i)x_i\tilde{x}_i^\top - \frac{1}{n}\sum_{(i,i)\in\mathcal{C}\setminus\mathcal{E}}(\nu-\beta_i)x_i\tilde{x}_i^\top - \frac{1}{n}\sum_{(i,j)\notin\mathcal{E}\cup\mathcal{C}}\beta_{ij}x_i\tilde{x}_j^\top \\
&\triangleq T_1 + Q_1 - T_2 + R_1 - R_2 - R_3 - R_4.
\end{aligned}$$

We first bound R_1, R_2, R_3 and R_4. For the term R_1, from Lemma 9.1 and Lemma 7.2,

$$\begin{aligned}
\|R_1\| &\le \frac{n-m}{n}\max_{(i,j)\in\mathcal{E}\setminus\mathcal{C}}(1-\beta_{ij})\max_{i\in[n]}\|x_i\|\max_{i\in[n]}\|\tilde{x}_i\| \\
&\lesssim \frac{n}{n^{1+\gamma}}\|\Sigma_z\|^{1/2}\|\Sigma_{\tilde{z}}\|^{1/2}(r+r(\Sigma_\xi))^{1/2}(r+r(\Sigma_{\tilde{\xi}}))^{1/2}\log n \\
&\lesssim \frac{1}{n^\gamma}(\|\Sigma_z\|\vee\|\Sigma_{\tilde{z}}\|)(r+r(\Sigma_\xi)+r(\Sigma_{\tilde{\xi}}))\log n
\end{aligned}$$

holds with probability at least $1-O(n^{-1})$. Similary,

$$\begin{aligned}
\|R_2\| &\le \frac{m}{n}\max_{(i,i)\in\mathcal{C}\cap\mathcal{E}}(\nu-1-\beta_i)\max_{i\in[n]}\|x_i\|\max_{i\in[n]}\|\tilde{x}_i\| \\
&\lesssim \frac{1}{n^\gamma}\|\Sigma_z\|^{1/2}\|\Sigma_{\tilde{z}}\|^{1/2}(r+r(\Sigma_\xi))^{1/2}(r+r(\Sigma_{\tilde{\xi}}))^{1/2}\log n \\
&\lesssim \frac{1}{n^\gamma}(\|\Sigma_z\|\vee\|\Sigma_{\tilde{z}}\|)(r+r(\Sigma_\xi)+r(\Sigma_{\tilde{\xi}}))\log n, \\
\|R_3\| &\le \frac{n-m}{n}\max_{(i,i)\in\mathcal{C}\setminus\mathcal{E}}(\nu-\beta_i)\max_{i\in[n]}\|x_i\|\max_{i\in[n]}\|\tilde{x}_i\| \\
&\lesssim \frac{1}{n^\gamma}\|\Sigma_z\|^{1/2}\|\Sigma_{\tilde{z}}\|^{1/2}(r+r(\Sigma_\xi))^{1/2}(r+r(\Sigma_{\tilde{\xi}}))^{1/2}\log n \\
&\lesssim \frac{1}{n^\gamma}(\|\Sigma_z\|\vee\|\Sigma_{\tilde{z}}\|)(r+r(\Sigma_\xi)+r(\Sigma_{\tilde{\xi}}))\log n, \\
\|R_4\| &\le \frac{n^2-2n+m}{n}\max_{(i,j)\notin\mathcal{E}\cup\mathcal{C}}\beta_{ij}\max_{i\in[n]}\|x_i\|\max_{i\in[n]}\|\tilde{x}_i\| \\
&\lesssim \frac{1}{n^\gamma}(\|\Sigma_z\|\vee\|\Sigma_{\tilde{z}}\|)(r+r(\Sigma_\xi)+r(\Sigma_{\tilde{\xi}}))\log n
\end{aligned}$$

holds with probability at least $1-O(n^{-1})$. We can bound the terms T_1 and Q_1 as in equation 8.5 and equation 8.9. Similar to the argument in equation 8.9, with probability $1-O(n^{-1})$,

$$\left\|T_2 - \frac{n-m}{n}U_1^*\Sigma_z^{1/2}\Sigma_{\tilde{z}}^{1/2}U_2^{*\top}\right\| \lesssim \frac{n-m}{n}(\|\Sigma_z\|\vee\|\Sigma_{\tilde{z}}\|)\sqrt{\frac{(r+r(\Sigma_\xi)+r(\Sigma_{\tilde{\xi}}))\log(nd_1+nd_2)}{n-m}}.$$

Therefore,

$$\left\|S - \Big(\nu\frac{m}{n} - 1\Big)U_1^*\Sigma_z^{1/2}\Sigma_{\tilde{z}}^{1/2}U_2^{*\top}\right\| \tag{8.13}$$

$$\le \left\|T_1 - (\nu-1)\frac{m}{n}U_1^*\Sigma_z^{1/2}\Sigma_{\tilde{z}}^{1/2}U_2^{*\top}\right\| + \left\|T_2 - \frac{n-m}{n}U_1^*\Sigma_z^{1/2}\Sigma_{\tilde{z}}^{1/2}U_2^{*\top}\right\| + \|Q_1\| + \|R_1\| + \|R_2\| + \|R_3\| + \|R_4\| \tag{8.14}$$

$$\lesssim \frac{(r + r(\Sigma_\xi) + r(\Sigma_{\tilde{\xi}}))\log n}{n^\gamma} \tag{8.15}$$

$$+ \left[\Big(1 - \frac{m}{n}\Big)^{1/2} + \Big(\frac{m}{n}\Big)^{1/2} + \Big(1 - \frac{m}{n}\Big)^{1/2}\right]\sqrt{\frac{(r + r(\Sigma_\xi) + r(\Sigma_{\tilde{\xi}}))\log(nd_1 + nd_2)}{n}}$$

$$\lesssim \frac{(r + r(\Sigma_\xi) + r(\Sigma_{\tilde{\xi}}))\log n}{n^\gamma} + \sqrt{\frac{(r + r(\Sigma_\xi) + r(\Sigma_{\tilde{\xi}}))\log(nd_1 + nd_2)}{n}}. \tag{8.16}$$

Let $\Sigma' \triangleq (\nu m/n - 1)U_1^*\Sigma_z^{1/2}\Sigma_{\tilde{z}}^{1/2}U_2^{*\top} = (\nu - 1 - \nu p_n)U_1^*\Sigma_z^{1/2}\Sigma_{\tilde{z}}^{1/2}U_2^{*\top}$. Then $\lambda_r(\Sigma') = |\nu - 1 - \nu p_n|\lambda_{\min}(\Sigma_z^{1/2}\Sigma_{\tilde{z}}^{1/2})$ and $\lambda_{\max}(\Sigma') = |\nu - 1 - \nu p_n|\lambda_{\max}(\Sigma_z^{1/2}\Sigma_{\tilde{z}}^{1/2})$. From Theorem 3 in Yu et al. (2015) with $s \leftarrow r$, $r \leftarrow 1$,

$$\begin{aligned}
&\|\sin\Theta(P_r(G_1), U_1^*)\|_F \vee \|\sin\Theta(P_r(G_2), U_2^*)\|_F \\
&\quad\le \sqrt{r} \wedge \frac{2(2|\nu - 1 - \nu p_n|\lambda_{\max}(\Sigma_z^{1/2}\Sigma_{\tilde{z}}^{1/2}) + \|S - \Sigma'\|)r^{1/2}\|S - \Sigma'\|}{(\nu - 1 - \nu p_n)^2\lambda_{\min}^2(\Sigma_z^{1/2}\Sigma_{\tilde{z}}^{1/2})} \\
&\quad\lesssim \sqrt{r} \wedge \sqrt{r}\left[\frac{(r + r(\Sigma_\xi) + r(\Sigma_{\tilde{\xi}}))\log n}{n^\gamma} + \sqrt{\frac{(r + r(\Sigma_\xi) + r(\Sigma_{\tilde{\xi}}))\log(nd_1 + nd_2)}{n}}\right] \\
&\quad\lesssim \sqrt{r} \wedge \sqrt{\frac{r(r + r(\Sigma_\xi) + r(\Sigma_{\tilde{\xi}}))\log(nd_1 + nd_2)}{n}},
\end{aligned} \tag{8.17}$$

where in the second inequality we used Assumption 3.1 and $\nu - 1 - \nu p_n \ge \nu\eta - 1 \ge 0.1$.

Furthermore, from equation 8.16 and since $G_1^\top G_2 = \mathrm{SVD}_r(S)$, there exists a constant $C_q = C_q(\sigma, s_1, s_2, \kappa_z^2, \kappa_{\tilde{z}}^2, q) > 0$ satisfying

$$\begin{aligned}
\|G_1^\top G_2 - (\nu - 1 - \nu p_n)\Sigma\| &= \|\mathrm{SVD}_r(S) - \Sigma'\| \\
&\le \lambda_{r+1}(S) + \|S - \Sigma'\| \\
&\le \lambda_{r+1}(\Sigma') + 2\|S - \Sigma'\| \\
&\le C_q\left(\sqrt{r} \wedge \sqrt{\frac{(r + r(\Sigma_\xi) + r(\Sigma_{\tilde{\xi}}))\log(nd_1 + nd_2)}{n}}\right).
\end{aligned}$$

Thus, the condition in equation 7.4 implies that $\|G_1^\top G_2 - (\nu - 1 - \nu p_n)\Sigma\| \le c_q r/((r + r(\Sigma_\xi))(r + r(\Sigma_{\tilde{\xi}}))\log n)$.

□

8.7 Proof of equation 4.3

Here we restate the result of 4.3, which follows by a similar argument in the proof of Proposition 2.1.

Proposition 8.2. *Consider minimizing the nonlinear loss function $\mathcal{L}^u$ defined in equation 4.1. Then,*

$$\frac{\partial\mathcal{L}}{\partial G_k} = -\left.\frac{\partial\,\mathrm{tr}\big(G_1 S(\beta) G_2^\top\big)}{\partial G_k}\right|_{\beta^u = \beta^u(G_1, G_2)} + \frac{\partial R(G_1, G_2)}{\partial G_k}, \qquad k \in \{1, 2\},$$

where the contrastive cross-covariance $S(\beta^u)$ is given by:

$$S(\beta^u) = \frac{1}{N}\sum_{(i,j)\in\bar{\mathcal{E}}^u}\beta_i^u x_i\tilde{x}_j^\top - \frac{1}{N}\sum_{(i,j)\notin\bar{\mathcal{E}}^u}\beta_{ij}^u x_i\tilde{x}_j^\top,$$

with

$$\beta_i^u = \nu - 1 + \frac{1}{2}\frac{e^{s_{ij}^u/\tau}}{\sum_{j'\in[N]} e^{s_{ij'}^u/\tau}} + \frac{1}{2}\frac{e^{s_{ij}^u/\tau}}{\sum_{i'\in[N]} e^{s_{i'j}^u/\tau}},$$
$$\beta_{ij}^u = \frac{1}{2}\frac{e^{s_{ij}^u/\tau}}{\sum_{j'\in[N]} e^{s_{ij'}^u/\tau}} + \frac{1}{2}\frac{e^{s_{ij}^u/\tau}}{\sum_{i'\in[N]} e^{s_{i'j}^u/\tau}}.$$

8.8 Proof of Lemma 4.1

Here we consider Algorithm 1. We restate Lemma 4.1.

Lemma 8.2 (Restatement of Lemma 4.1)**.** *Suppose Assumptions 3.1, 3.2, 4.1, and 4.2 hold. Fix any $\gamma > 0$ and $\nu \geq 1$. Choose $\tau \leq C(1+\gamma)^{-1}\sqrt{r/\log N}$, where $C > 0$ is some constant depending on $\sigma, s_1, s_2, \kappa_z^2, \kappa_{\tilde{z}}^2$. Consider applying Algorithm 1 to the data generated from the model 3.2. Then, with probability $1 - O(N^{-1} \vee n^{-1})$, $\hat{\mathcal{E}}^u = \mathcal{E}^u$ and*

$$\min_{(i,j)\in\mathcal{E}^u\setminus\bar{\mathcal{E}}^u} \beta_{ij}^{u(0)} = 1 - O\left(\frac{1}{N^\gamma}\right), \quad \max_{(i,j)\notin\mathcal{E}^u\cup\bar{\mathcal{E}}^u} \beta_{ij}^{u(0)} \lesssim \frac{1}{N^{1+\gamma}},$$
$$\min_{(i,i)\in\mathcal{E}^u\cap\bar{\mathcal{E}}^u} \beta_i^{u(0)} = \nu - 1 + O\left(\frac{1}{N^\gamma}\right), \quad \max_{(i,i)\in\bar{\mathcal{E}}^u\setminus\mathcal{E}^u} \beta_i^{u(0)} = \nu - O\left(\frac{1}{N^\gamma}\right).$$

Proof. Since the initial representations $G_1^{(0)}$ and $G_2^{(0)}$ are the solution to the minimization of the loss 3.1 with the dataset $(x_i, \tilde{x}_i)_{i=1}^n$, Theorem 8.1 and Assumption 4.1 give

$$\left\| G_1^{(0)\top} G_2^{(0)} - \frac{1}{\rho} U_1^* \Sigma_z^{1/2} \Sigma_{\tilde{z}}^{1/2} U_2^{*\top} \right\| \leq c_q \frac{r}{(r + r(\Sigma_\xi))(r + r(\Sigma_{\tilde{\xi}}))\log N} \tag{8.18}$$

with probability $1 - O(n^{-1})$.

From Lemma 8.1, with probability $1 - O(N^{-1} \vee n^{-1})$,

$$\min_{(i_1,j_1)\in\mathcal{E}^u} \left(s_{i_1 j_1} - \max_{j:(i_1,j)\notin\mathcal{E}^u} s_{i_1,j} \vee \max_{i:(i,j_1)\notin\mathcal{E}^u} s_{i,j_1} \right) \geq C'\sqrt{r\log N}.$$

This implies that $\{(i,j) \in [N]^2 : i \in [N], j \in \arg\max_{j'} s_{ij'}^u\} = \{(i,j) \in [N]^2 : j \in [N], i \in \arg\max_{i'} s_{i'j}^u\} = \mathcal{E}^u$ with high probability.

The conclusion directly follows by Lemma 7.2 with substitution $\mathcal{C} \leftarrow \bar{\mathcal{E}}^u$ and $\mathcal{E} \leftarrow \mathcal{E}^u$, since $(x_i, \tilde{x}_i)_{i=1}^n$ and $(x_i^u, \tilde{x}_i^u)_{i=1}^N$ are independent. □

8.9 Proof of Theorem 4.1

We restate Theorem 4.1.

Theorem 8.2 (Restatement of Theorem 4.1)**.** *Suppose Assumptions 3.1, 3.2, 4.1 and 4.2 hold. Fix any $\gamma > 2$ and $\nu \geq 1.1$. Choose τ as in Lemma 8.2. Consider applying Algorithm 1 to the data $(x_i^u, \tilde{x}_i^u)_{i=1}^N$ generated from equation 3.2, whose association is unknown. Then, with probability $1 - O(N^{-1} \vee n^{-1})$,*

$$\|\sin\Theta(P_r(G_1), U_1^*)\|_F \vee \|\sin\Theta(P_r(G_2), U_2^*)\|_F \lesssim \sqrt{r} \wedge \sqrt{\frac{r(r + r(\Sigma_\xi) + r(\Sigma_{\tilde{\xi}}))\log(N + d_1 + d_2)}{N}}.$$

Proof. From Lemma 8.2, we have $\hat{\mathcal{E}}^u = \mathcal{E}^u$ with high probability. Thus we treat $\mathcal{E}^u$ as known for brevity and let $\mathcal{E}^u = \{(1,1), \ldots, (N,N)\}$ without loss of generality. Since the loss function in 4.1 is exactly the same as the loss function in 7.1 when $\mathcal{E}^u = \{(1,1), \ldots, (N,N)\}$, the conclusion follows by Theorem 7.1 applied to $(x_i^u, \tilde{x}_i^u)_{i=1}^N$ with $p_n \equiv 0$. □

9 Auxiliary Results

Here, we list the auxiliary results that are used in the proofs.

Lemma 9.1. *Suppose Assumptions 3.1 and 3.2 hold. Fix any $\Sigma \in \mathbb{R}^{r\times r}$. There exists some constant $c = c(\sigma, s_1, \kappa_z^2) \in (0, 1]$ such that if $\log n \le cr$, the following inequalities hold with probability $1 - O(n^{-1})$:*

$$\begin{aligned}
&\max_{i\in[n]} \|x_i\| \le C_1\|\Sigma_z\|^{1/2}(r^{1/2} + r^{1/2}(\Sigma_\xi))\sqrt{\log n},\\
&\max_{i\in[n]} \|\tilde{x}_i\| \le C_2\|\Sigma_{\tilde{z}}\|^{1/2}(r^{1/2} + r^{1/2}(\Sigma_{\tilde{\xi}}))\sqrt{\log n},\\
&\max_{(i,j)\in\mathcal{E}} |\xi_i^\top U_1^*\Sigma\tilde{z}_j + z_i^\top\Sigma U_2^{*\top}\tilde{\xi}_j + \xi_i^\top U_1^*\Sigma U_2^{*\top}\tilde{\xi}_j| \le C_3\sqrt{r\log n}\|\Sigma_z\|^{1/2}\|\Sigma\|\|\Sigma_{\tilde{z}}\|^{1/2},\\
&\max_{(i,j)\in\mathcal{E}} \left|z_i^\top\Sigma\tilde{z}_j - \mathrm{tr}\left(\Sigma_z^{1/2}\Sigma\Sigma_{\tilde{z}}^{1/2}\right)\right| \le C_4\|\Sigma_z\|^{1/2}\|\Sigma\|\|\Sigma_{\tilde{z}}\|^{1/2}\sqrt{r\log n},\\
&\max_{i\in[n]} \|U_1^{*\top}x_i\| \le C_5\sqrt{r\|\Sigma_z\|},\\
&\min_{i\in[n]} \|U_1^{*\top}x_i\| \ge \sqrt{\frac{r\|\Sigma_z\|}{2\kappa_z^{-2}}},
\end{aligned}$$

where $C_1 = C_1(\sigma, s_1), C_2 = C_2(\sigma, s_2), C_3 = C_3(\sigma, s_1, s_2), C_4 = C_4(\sigma)$ and $C_5 = C_5(\sigma, s_1) > 0$ are some constants.

Proof of Lemma 9.1. Let

$$c = 2^{-1}(2C'''(\sigma)(1 \vee s_1^{-1})\kappa_z^2)^{-2} \wedge 1.$$

From Corollary 9.1, Assumption 3.2, and by the union bound argument,

$$\begin{aligned}
\max_i \|x_i\| &\le \max_i \|z_i\| + \max_i \|\xi_i\|\\
&\le \mathrm{tr}^{1/2}(\Sigma_z) + \mathrm{tr}^{1/2}(\Sigma_\xi) + C(\sigma)(\|\Sigma_z\|^{1/2} + \|\Sigma_\xi\|^{1/2})\sqrt{2\log n}\\
&\le \|\Sigma_z\|^{1/2}(r^{1/2} + s_1^{-1}r^{1/2}(\Sigma_\xi))(1 + C(\sigma)\sqrt{2\log n})\\
&\le C_1(\sigma, s_1)\|\Sigma_z\|^{1/2}(r^{1/2} + r^{1/2}(\Sigma_\xi))\sqrt{\log n}
\end{aligned}$$

holds with probability at least $1 - 2n^{-1}$, where $C_1(\sigma, s_1) \triangleq (1 \vee s_1^{-1})(1 \vee C(\sigma))$. Similarly,

$$\max_i \|\tilde{x}_i\| \le C_2(\sigma, s_2)\|\Sigma_{\tilde{z}}\|^{1/2}(r^{1/2} + r^{1/2}(\Sigma_{\tilde{\xi}}))\sqrt{\log n}$$

holds with probability at least $1 - 2n^{-1}$, where $C_2(\sigma, s_2) \triangleq (1 \vee s_2^{-1})(1 \vee C(\sigma))$.

By Lemma 9.3 and the union bound argument, there exists some constant $C'(\sigma) > 0$ such that

$$\begin{aligned}
\max_{(i,j)\in\mathcal{E}} |\xi_i^\top U_1^*\tilde{\Sigma}z_j| &\le C'(\sigma)(\|\Sigma_\xi^{1/2}U_1^*\Sigma\Sigma_{\tilde{z}}^{1/2}\|_F\sqrt{2\log n} \vee 2\|\Sigma_\xi^{1/2}U_1^*\Sigma\Sigma_{\tilde{z}}^{1/2}\|\log n)\\
&\le C'(\sigma)(\mathrm{tr}^{1/2}(U_1^{*\top}\Sigma_\xi U_1^*\Sigma\Sigma_{\tilde{z}})\sqrt{2\log n} \vee 2\|\Sigma_\xi\|^{1/2}\|\Sigma\|\|\Sigma_{\tilde{z}}\|^{1/2}\log n)\\
&\le C'(\sigma)\|\Sigma_\xi\|^{1/2}\|\Sigma\|\|\Sigma_{\tilde{z}}\|^{1/2}(\sqrt{2r\log n} \vee 2\log n)
\end{aligned}$$

holds with probability at least $1 - n^{-1}$. Since $\log n \le \sqrt{r\log n}$, the far right-hand side can be further bounded by $2C'(\sigma)s_1^{-1}\|\Sigma_z\|^{1/2}\|\Sigma\|\|\Sigma_{\tilde{z}}\|^{1/2}\sqrt{r\log n}$. By a similar argument combined with Assumption 3.2, there exists some constant $C_3(\sigma, s_1, s_2) > 0$ such that

$$\begin{aligned}
&\max_{(i,j)\in\mathcal{E}} |\xi_i^\top U_1^*\Sigma\tilde{z}_j + z_i^\top\Sigma U_2^{*\top}\tilde{\xi}_j + \xi_i^\top U_1^*\Sigma U_2^{*\top}\tilde{\xi}_j|\\
&\quad\le \max_{(i,j)\in\mathcal{E}} |\xi_i^\top U_1^*\Sigma\tilde{z}_j| + \max_{(i,j)\in\mathcal{E}} |z_i^\top\Sigma U_2^{*\top}\tilde{\xi}_j| + \max_{(i,j)\in\mathcal{E}} |\xi_i^\top U_1^*\Sigma U_2^{*\top}\tilde{\xi}_j|\\
&\quad\le C_3(c, \sigma, s_1, s_2)\|\Sigma_z\|^{1/2}\|\Sigma\|\|\Sigma_{\tilde{z}}\|^{1/2}\sqrt{r\log n}
\end{aligned}$$

holds with probability at least $1 - 3n^{-1}$, where we used Cauchy-Schwarz inequality in the last inequality.

Fix any $(i, j) \in \mathcal{E}$. Since $\Sigma_z^{-1/2} z_i = \Sigma_{\tilde{z}}^{-1/2} \tilde{z}_j$, applying Lemma 9.4 with $X \leftarrow \Sigma_z^{-1/2} z_i$, $A \leftarrow \Sigma_z^{1/2} \Sigma \Sigma_{\tilde{z}}^{1/2}$, and $t \leftarrow \log n^2$ yields the following. There exists a constant $C''(\sigma) > 0$ such that

$$\begin{aligned}\left| z_i^\top \Sigma \tilde{z}_j - \operatorname{tr}\left(\Sigma_z^{1/2} \Sigma \Sigma_{\tilde{z}}^{1/2}\right) \right| &\le C''(\sigma) \left(\|\Sigma_z^{1/2} \Sigma \Sigma_{\tilde{z}}^{1/2}\|_F \sqrt{\log n^2} \vee \|\Sigma_z^{1/2} \Sigma \Sigma_{\tilde{z}}^{1/2}\| \log n^2 \right) \\ &\le C_4(\sigma) \|\Sigma_z^{1/2} \Sigma \Sigma_{\tilde{z}}^{1/2}\| \sqrt{r \log n}\end{aligned}$$

holds with probability at least $1 - n^{-2}$, where the last inequality is again from $\log n \le \sqrt{r \log n}$. By the union bound argument, we obtain

$$\max_{(i,j) \in \mathcal{E}} \left| z_i^\top \Sigma \tilde{z}_j - \operatorname{tr}\left(\Sigma_z^{1/2} \Sigma \Sigma_{\tilde{z}}^{1/2}\right) \right| \le C_4(\sigma) \|\Sigma_z^{1/2}\| \|\Sigma\| \|\Sigma_{\tilde{z}}^{1/2}\| \sqrt{r \log n}$$

with probability at least $1 - n^{-1}$.

From Corollary 9.1, Assumption 3.2 and by the union bound argument,

$$\begin{aligned}\max_i \|U_1^{*\top} x_i\| &\le \max_i \|z_i\| + \max_i \|U_1^{*\top} \xi_i\| \\ &\le \operatorname{tr}^{1/2}(\Sigma_z) + \operatorname{tr}^{1/2}(U_1^{*\top} \Sigma_\xi U_1^*) + C(\sigma)(\|\Sigma_z\|^{1/2} + \|\Sigma_\xi\|^{1/2}) \sqrt{2 \log n} \\ &\le (\sqrt{r} + C(\sigma) \sqrt{2 \log n}) \|\Sigma_z\|^{1/2} (1 + s_1^{-1}) \\ &\le C_5(\sigma, s_1) \sqrt{r} \|\Sigma_z\|^{1/2}\end{aligned}$$

holds with probability at least $1 - 2n^{-1}$, where $C_5 > 0$ is some constant.

Since $\|z_i\|^2 = \|U_1^{*\top} x_i\|^2 - 2 z_i^\top U_1^{*\top} \xi_i - \|U_1^{*\top} \xi_i\|^2 \le \|U_1^{*\top} x_i\|^2 - 2 z_i^\top U_1^{*\top} \xi_i$, there exists some constant $C'''(\sigma) > 0$ such that

$$\begin{aligned}\min_i \|U_1^{*\top} x_i\|^2 &\ge \min_i \|z_i\|^2 - 2 \max_i |z_i^\top U_1^{*\top} \xi_i| \\ &\ge \operatorname{tr}(\Sigma_z) - C'''(\sigma) (\|\Sigma_z\|^{1/2} \operatorname{tr}^{1/2}(\Sigma_z) \sqrt{2 \log n} \vee 2 \|\Sigma_z\| \log n) \\ &\quad - C'''(\sigma) \|\Sigma_\xi\|^{1/2} \|\Sigma_z\|^{1/2} (\sqrt{2r \log n} \vee 2 \log n) \\ &\ge r \|\Sigma_z\| \frac{\lambda_{\min}(\Sigma_z)}{\|\Sigma_z\|} - 2C'''(\sigma)(1 \vee s_1^{-1}) \|\Sigma_z\| \sqrt{r \log n} \\ &= r \|\Sigma_z\| \frac{\lambda_{\min}(\Sigma_z)}{\|\Sigma_z\|} \left(1 - 2C'''(\sigma)(1 \vee s_1^{-1}) \frac{\|\Sigma_z\|}{\lambda_{\min}(\Sigma_z)} \sqrt{\frac{\log n}{r}} \right)\end{aligned}$$

holds with probability at least $1 - 2n^{-1}$, where the second inequality follows from Lemma 9.4 and Assumption 3.2. From Assumption 3.1 and the definition of c,

$$\min_i \|U_1^{*\top} x_i\|^2 \ge r \|\Sigma_z\| \kappa_z^{-2} \big(1 - 2C'''(\sigma)(1 \vee s_1^{-1}) \kappa_z^2 \sqrt{c}\big) \ge (1/2) r \|\Sigma_z\| \kappa_z^{-2}.$$

□

Lemma 9.2. *Suppose $X_1, \ldots, X_n$ are i.i.d. sub-Gaussian random variables with parameter σ. Then,*

$$\mathbb{P}(\max_i X_i - \sqrt{2\sigma^2 \log n} \ge t) \le \exp\left(-\frac{t^2}{2\sigma^2} \right)$$

holds for all $t \ge 0$.

Proof. Observe that

$$\begin{aligned}\mathbb{P}(\max_i X_i - \sqrt{2\sigma^2 \log n} \geq t) &= \mathbb{P}\bigg(\bigcup_i \{X_i \geq t + \sqrt{2\sigma^2 \log n}\}\bigg) \\ &\leq n\mathbb{P}(X_1 \geq t + \sqrt{2\sigma^2 \log n}) \\ &\leq n \exp\bigg(-\frac{(t + \sqrt{2\sigma^2 \log n})^2}{2\sigma^2}\bigg) \\ &\leq \exp\bigg(-\frac{t^2}{2\sigma^2}\bigg).\end{aligned}$$

□

Lemma 9.3. *Let $X = (X_1, \ldots, X_{d_1})$ and $\tilde{X} = (\tilde{X}_1, \ldots, \tilde{X}_{d_2})$ be mean zero random vectors taking values in $\mathbb{R}^d$. Let $A \in \mathbb{R}^{d_1 \times d_2}$ be a non-random matrix. Suppose $\Sigma_X^{-1/2} X$ and $\Sigma_{\tilde{X}}^{-1/2} \tilde{X}$ are independent and have i.i.d. sub-Gaussian coordinates with parameter σ. Then, there exists a constant $C = C(\sigma) > 0$ such that with probability at least $1 - e^{-t}$,*

$$|X^\top A \tilde{X}| \leq C\Big(\|\Sigma_X^{1/2} A \Sigma_{\tilde{X}}^{1/2}\|_F \sqrt{t} \vee \|\Sigma_X^{1/2} A \Sigma_{\tilde{X}}^{1/2}\| t\Big).$$

holds for all $t > 0$.

Proof of Lemma 9.3. The proof follows by a similar argument as in the proof of Theorem 6.2.1, Lemma 6.2.2 and Lemma 6.2.3 in Vershynin (2018). □

We also use the following Hanson-Wright inequality. See, for example, Theorem 6.2.1 in Vershynin (2018).

Lemma 9.4. *Let $X = (X_1, \ldots, X_{d_1})$ be mean zero random vectors taking values in $\mathbb{R}^d$. Let $A \in \mathbb{R}^{d_1 \times d_1}$ be a non-random matrix. Suppose $\Sigma_X^{-1/2} X$ have i.i.d. sub-Gaussian coordinates with parameter σ. Then, there exists a constant $C = C(\sigma) > 0$ such that with probability at least $1 - e^{-t}$,*

$$|X^\top A X - \mathrm{tr}(A\Sigma_X)| \leq C\Big(\|\Sigma_X^{1/2} A \Sigma_X^{1/2}\|_F \sqrt{t} \vee \|\Sigma_X^{1/2} A \Sigma_X^{1/2}\| t\Big).$$

holds for all $t > 0$.

The following corollary is adapted from the proof of Theorem 6.3.2 in Vershynin (2018).

Corollary 9.1. *Let X be a random vector in $\mathbb{R}^d$. Suppose $\Sigma_X^{-1/2} X$ has i.i.d. sub-Gaussian coordinates with parameter σ. Let $A \in \mathbb{R}^{r \times d}$ be any non-random matrix. Then, there exists a constant $C = C(\sigma) > 0$ such that*

$$|\|AX\| - \mathrm{tr}^{1/2}(A\Sigma_X A^\top)| \leq C(\sigma)\|A\Sigma_X A^\top\|^{1/2} \sqrt{t}$$

holds with probability at least $1 - e^{-t}$ for all $t > 0$.

Assumption 9.1. Let X and $\tilde{X}$ be mean zero random vectors taking values in $\mathbb{R}^{d_1}$ and $\mathbb{R}^{d_2}$, respectively. Assume the following

- $\mathbb{E}[(u^\top X)^2] \geq c_1 \|u^\top X\|_{\psi_2}^2$ holds for any $u \in \mathbb{R}^{d_1}$,
- $\mathbb{E}[(v^\top \tilde{X})^2] \geq c_2 \|v^\top \tilde{X}\|_{\psi_2}^2$ holds for any $v \in \mathbb{R}^{d_2}$.

Proposition 9.1. *Let X and $\tilde{X}$ be mean zero random vectors taking values in $\mathbb{R}^{d_1}$ and $\mathbb{R}^{d_2}$, respectively. Suppose X and $\tilde{X}$ satisfy Assumption 9.1. Let $(a_i)_i$ be a bounded sequence of positive numbers such that $\max_i a_i \leq a$. Let $\{(X_i, \tilde{X}_i)\}_i$ be independent copies of $(X, \tilde{X})$ and $\hat{\Sigma}^a_{X,\tilde{X}} \triangleq (1/n)\sum_{i=1}^n a_i X_i \tilde{X}_i^\top$. Let $\Sigma^a_{X,\tilde{X}} = (1/n)(\sum_{i=1}^n a_i)\mathbb{E}[X\tilde{X}^\top]$. Then, there exists a constant $C = C(c_1, c_2) > 0$ such that with probability at least $1 - e^{-t}$,*

$$\begin{aligned}&\|\hat{\Sigma}^a_{X,\tilde{X}} - \Sigma^a_{X,\tilde{X}}\| \\ &\leq Ca\bigg[(\mathrm{tr}(\Sigma_{\tilde{X}})\|\Sigma_X\| \vee \mathrm{tr}(\Sigma_X)\|\Sigma_{\tilde{X}}\|)^{1/2}\sqrt{\frac{t + \log(d_1 + d_2)}{n}} \vee (\mathrm{tr}(\Sigma_X)\,\mathrm{tr}(\Sigma_{\tilde{X}}))^{1/2}\frac{t + \log(d_1 + d_2)}{n}\bigg].\end{aligned}$$

holds for all $t > 0$.

Notice that when $X = \tilde{X}$, we recover the bound given in Theorem 2.2 of Bunea and Xiao (2015).

Proof of Proposition 9.1. Let $B_i \triangleq X_i\tilde{X}_i^\top - \Sigma_{X,\tilde{X}}$. Define symmetric matrices A_i of order $d_1 + d_2$ as

$$A_i \triangleq \begin{pmatrix} O & B_i \\ B_i^\top & O \end{pmatrix}.$$

Since $A_1^{2k} = \mathrm{diag}((B_1B_1^\top)^k, (B_1^\top B_1)^k)$ for $k \geq 2$, $\|\mathbb{E}[A_1^{2k}]\| \leq \|\mathbb{E}[(B_1B_1^\top)^k]\| \vee \|\mathbb{E}[(B_1^\top B_1)^k]\|$ and A_1^{2k} is positive semi-definite. From Lemma 9.5, for $k \geq 1$,

$$\|\mathbb{E}[A_1^{2k}]\| \leq \frac{(2k)!}{2} R^{2k-2}\sigma^2,$$

where σ^2 and R are defined in Lemma 9.5. Fix any $u \in \mathbb{S}^{d_1+d_2-1}$. By Cauchy-Schwarz inequality and Jensen's inequality, for $k \geq 2$,

$$\mathbb{E}[u^\top A_1^{2k-1}u] \leq \mathbb{E}[\sqrt{u^\top A_1^{2k-2}uu^\top A_1^{2k}u}] \leq \sqrt{u^\top \mathbb{E}[A_1^{2k-2}]uu^\top \mathbb{E}[A_1^{2k}]u} \leq \frac{\sqrt{(2k)!(2k-2)!}}{2} R^{2k-3}\sigma^2.$$

Observe that

$$\frac{\sqrt{(2k)!(2k-2)!}}{(2k-1)!} = \sqrt{\frac{2k}{2k-1}} \leq \frac{2}{\sqrt{3}}.$$

Therefore, substituting $\sigma^2 \leftarrow (4/3)^{1/2}\sigma^2$, we obtain the bound $\|\mathbb{E}[A_1^k]\| \leq (k!/2)R^{k-2}\sigma^2$ for all $k \geq 2$. Applying Theorem 6.2 in Tropp (2012) to $(1/n)\sum_{i=1}^n a_iA_i$, and using $\|B_i\| = \|(I_{d_1}O_{d_1\times d_2})A_i(O_{d_2\times d_1}I_{d_2})^\top\| \leq \|A_i\|$ and $\|E(a_iA_i)^k\| \leq (k)!(aR)^{k-2}(a\sigma)^2/2$, we obtain the following bound: there exists a constant $C > 0$ only depending on c_1 and c_2 such that

$$\begin{aligned}
&\|\hat{\Sigma}^a_{X,\tilde{X}} - \Sigma^a_{X,\tilde{X}}\| \\
&\leq Ca\left[(\mathrm{tr}(\Sigma_{\tilde{X}})\|\Sigma_X\| \vee \mathrm{tr}(\Sigma_X)\|\Sigma_{\tilde{X}}\|)^{1/2}\sqrt{\frac{t+\log(d_1+d_2)}{n}} \vee (\mathrm{tr}(\Sigma_X)\,\mathrm{tr}(\Sigma_{\tilde{X}}))^{1/2}\frac{t+\log(d_1+d_2)}{n}\right].
\end{aligned}$$

holds with probability at least $1 - e^{-t}$ for all $t > 0$. □

Lemma 9.5. *Let X and $\tilde{X}$ be mean zero random vectors taking values in $\mathbb{R}^{d_1}$ and $\mathbb{R}^{d_2}$, respectively. Suppose X and $\tilde{X}$ satisfy Assumption 9.1. Then, for $k \geq 1$,*

$$\|\mathbb{E}[((X\tilde{X}^\top - \Sigma_{X,\tilde{X}})(X\tilde{X}^\top - \Sigma_{X,\tilde{X}})^\top)^k]\| \vee \|\mathbb{E}[((\tilde{X}X^\top - \Sigma_{\tilde{X},X})(\tilde{X}X^\top - \Sigma_{\tilde{X},X})^\top)^k]\| \leq \frac{(2k)!}{2}L^{2k-2}\sigma^2,$$

where

$$\begin{aligned}
\sigma^2 &\triangleq \frac{65\cdot 16e^2}{c_1c_2 \wedge 1}(\mathrm{tr}(\Sigma_{\tilde{X}})\|\Sigma_X\| \vee \mathrm{tr}(\Sigma_X)\|\Sigma_{\tilde{X}}\|) \\
L &\triangleq \frac{8e}{(c_1c_2 \wedge 1)^{1/2}}(\mathrm{tr}(\Sigma_X)\,\mathrm{tr}(\Sigma_{\tilde{X}}))^{1/2}.
\end{aligned}$$

Proof of Lemma 9.5. Define $B \triangleq X\tilde{X}^\top - \Sigma_{X,\tilde{X}}$. To use the matrix bernstein inequality, we bound $\|\mathbb{E}[(BB^\top)^k]\|$. Fix any $u \in \mathbb{S}^{d_1-1}$. Then

$$u^\top(BB^\top)^k u \leq \|B\|^{2k-2}u^\top BB^\top u.$$

Since for any matrices C and D, $2CC^\top + 2DD^\top - (C-D)(C-D)^\top = (C+D)(C+D)^\top$, we obtain $0 \preceq (C-D)(C-D)^\top \preceq 2CC^\top + 2DD^\top \preceq 2CC^\top + 2\|D\|^2 I$. Also, $\|C-D\|^{2k-2} \leq 2^{2k-3}(\|C\|^{2k-2} + \|D\|^{2k-2})$. These results give

$$\begin{aligned}
u^\top(BB^\top)^k u &\leq \|B\|^{2k-2}u^\top BB^\top u \\
&\leq 2^{2k-1}(\|X\tilde{X}^\top\|^{2k-2} + \|\Sigma_{X,\tilde{X}}\|^{2k-2})(u^\top X\tilde{X}^\top\tilde{X}X^\top u + \|\Sigma_{X,\tilde{X}}\|^2) \\
&= 2^{2k-1}(\|X\tilde{X}^\top\|^{2k-2}u^\top X\tilde{X}^\top\tilde{X}X^\top u + \|X\tilde{X}^\top\|^{2k-2}\|\Sigma_{X,\tilde{X}}\|^2) \\
&\quad + 2^{2k-1}(\|\Sigma_{X,\tilde{X}}\|^{2k-1}u^\top X\tilde{X}^\top\tilde{X}X^\top u + \|\Sigma_{X,\tilde{X}}\|^{2k}).
\end{aligned}$$

$$\mathbb{E}[u^\top (BB^\top)^k u] \le 2^{2k-1}(\mathbb{E}[\|X\tilde{X}^\top\|^{2k-2}u^\top X\tilde{X}^\top\tilde{X}X^\top u] + \mathbb{E}[\|X\tilde{X}^\top\|^{2k-2}]\|\Sigma_{X,\tilde{X}}\|^2) + 2^{2k-1}(\|\Sigma_{X,\tilde{X}}\|^{2k-2}\mathbb{E}[u^\top X\tilde{X}^\top\tilde{X}X^\top u] + \|\Sigma_{X,\tilde{X}}\|^{2k}).$$

Notice that

$$\mathbb{E}[\|X\tilde{X}^\top\|^{2k-2}u^\top X\tilde{X}^\top\tilde{X}X^\top u] \le \sqrt{\mathbb{E}[\|X\tilde{X}^\top\|^{2(2k-2)}]\mathbb{E}[(u^\top X\tilde{X}^\top\tilde{X}X^\top u)^2]},$$

where the first inequality follows from Cauchy-Schwarz inequality. By Lemma A.2 from Bunea and Xiao (2015) and Assumption 9.1,

$$\mathbb{E}[\|X\tilde{X}^\top\|^{2(2k-2)}] \le \sqrt{\mathbb{E}[\|X\|^{4(2k-2)}]\mathbb{E}[\|\tilde{X}\|^{4(2k-2)}]} \le \frac{(4(2k-2))^{2(2k-2)}}{c_1^{2k-2}c_2^{2k-2}}(\operatorname{tr}(\Sigma_X))^{2k-2}(\operatorname{tr}(\Sigma_{\tilde{X}}))^{2k-2},$$

$$\mathbb{E}[(u^\top X\tilde{X}^\top\tilde{X}X^\top u)^2] = \mathbb{E}[(u^\top X)^4\|\tilde{X}\|^4] \le \sqrt{\mathbb{E}[(u^\top X)^8]\mathbb{E}[\|\tilde{X}\|^8]} \le \left(\frac{64}{c_1c_2}\operatorname{tr}(\Sigma_{\tilde{X}})\|\Sigma_X\|\right)^2.$$

Thus

$$\mathbb{E}[\|X\tilde{X}^\top\|^{2k-2}u^\top X\tilde{X}^\top\tilde{X}X^\top u] \le \frac{64(4(2k-2))^{2k-2}}{c_1^kc_2^k}(\operatorname{tr}(\Sigma_X)\operatorname{tr}(\Sigma_{\tilde{X}}))^{(2k-2)/2}\operatorname{tr}(\Sigma_{\tilde{X}})\|\Sigma_X\|.$$

Similarly,

$$\mathbb{E}[\|X\tilde{X}^\top\|^{2k-2}] \le \sqrt{\mathbb{E}[\|X\|^{2(2k-2)}]\mathbb{E}[\|\tilde{X}\|^{2(2k-2)}]} \le \frac{2(2k-2)^{2k-2}}{c_1^{k-1}c_2^{k-1}}(\operatorname{tr}(\Sigma_X)\operatorname{tr}(\Sigma_{\tilde{X}}))^{(2k-2)/2},$$

$$\mathbb{E}[u^\top X\tilde{X}^\top\tilde{X}X^\top u] \le \sqrt{\mathbb{E}[(u^\top X)^4]\mathbb{E}[\|\tilde{X}\|^4]} \le \frac{16}{c_1c_2}\operatorname{tr}(\Sigma_{\tilde{X}})\|\Sigma_X\|.$$

Also, for any $(u,v) \in \mathbb{S}^{d_1-1}\times\mathbb{S}^{d_2-1}$,

$$u^\top\Sigma_{X,\tilde{X}}v = \mathbb{E}[u^\top X\tilde{X}^\top v] \le \sqrt{\mathbb{E}[(u^\top X)^2]\mathbb{E}[(v^\top\tilde{X})^2]} \le \sqrt{\|\Sigma_X\|\|\Sigma_{\tilde{X}}\|} \le \sqrt{\operatorname{tr}(\Sigma_{\tilde{X}})\|\Sigma_X\|}.$$

This gives $\|\Sigma_{X,\tilde{X}}\|^2 \le \operatorname{tr}(\Sigma_{\tilde{X}})\|\Sigma_X\| \le \operatorname{tr}(\Sigma_{\tilde{X}})\operatorname{tr}(\Sigma_X)$. Therefore,

$$\begin{aligned}
\mathbb{E}[u^\top (BB^\top)^k u] &\le 2^{2k-1}\frac{64(4(2k-2))^{2k-2}}{c_1^kc_2^k}(\operatorname{tr}(\Sigma_X)\operatorname{tr}(\Sigma_{\tilde{X}}))^{(2k-2)/2}\operatorname{tr}(\Sigma_{\tilde{X}})\|\Sigma_X\| \\
&\quad + 2^{2k-1}\frac{(2(2k-2))^{2k-2}}{c_1^{k-1}c_2^{k-1}}(\operatorname{tr}(\Sigma_X)\operatorname{tr}(\Sigma_{\tilde{X}}))^{(2k-2)/2}\|\Sigma_{X,\tilde{X}}\|^2 \\
&\quad + 2^{2k-1}\|\Sigma_{X,\tilde{X}}\|^{2k-2}\frac{16}{c_1c_2}\operatorname{tr}(\Sigma_{\tilde{X}})\|\Sigma_X\| + 2^{2k-1}\|\Sigma_{X,\tilde{X}}\|^{2k} \\
&\le 2^{2k-1}\left(\frac{(4(2k-1))^{2k-1}}{c_1^{k-1}c_2^{k-1}}(\operatorname{tr}(\Sigma_X)\operatorname{tr}(\Sigma_{\tilde{X}}))^{(2k-2)/2} + \|\Sigma_{X,\tilde{X}}\|^{2k-2}\right) \\
&\quad \times \operatorname{tr}(\Sigma_{\tilde{X}})\|\Sigma_X\|\left(\frac{64}{c_1c_2}+1\right) \\
&\le 2^{2k}\frac{(4(2k-1))^{2k-1}}{c_1^{k-1}c_2^{k-1}\wedge 1}(\operatorname{tr}(\Sigma_X)\operatorname{tr}(\Sigma_{\tilde{X}}))^{(2k-2)/2}\operatorname{tr}(\Sigma_{\tilde{X}})\|\Sigma_X\|\frac{65}{c_1c_2\wedge 1}.
\end{aligned}$$

Hence

$$\|\mathbb{E}[(BB^\top)^k]\| \le \frac{(2k)!}{2}\frac{2}{(2k)!}2^{2k}\frac{(4(2k-1))^{2k-1}}{c_1^{k-1}c_2^{k-1}\wedge 1}(\operatorname{tr}(\Sigma_X)\operatorname{tr}(\Sigma_{\tilde{X}}))^{(2k-2)/2}\operatorname{tr}(\Sigma_{\tilde{X}})\|\Sigma_X\|\frac{65}{c_1c_2\wedge 1}.$$

Using the Stirling's formula $k! \geq \sqrt{2\rho} k^{k+1/2} e^{-k} e^{1/(12k+1)} \geq 2k^{k+1/2} e^{-k}$,

$$
\begin{aligned}
&\|\mathbb{E}[(BB^\top)^k]\| \\
&\leq \frac{(2k)!}{2} \frac{e^{2k}}{(2k)^{2k+1/2}} 2^{2k} \frac{(4(2k-1))^{2k-1}}{c_1^{k-1} c_2^{k-1} \wedge 1} (\operatorname{tr}(\Sigma_X) \operatorname{tr}(\Sigma_{\tilde{X}}))^{(2k-2)/2} \operatorname{tr}(\Sigma_{\tilde{X}}) \|\Sigma_X\| \frac{65}{c_1 c_2 \wedge 1} \\
&\leq \frac{(2k)!}{2} \left(65 \cdot 16 e^2 \frac{1}{c_1 c_2 \wedge 1} \operatorname{tr}(\Sigma_{\tilde{X}}) \|\Sigma_X\| \right) \left(8e (\operatorname{tr}(\Sigma_X) \operatorname{tr}(\Sigma_{\tilde{X}}))^{1/2} \frac{1}{c_1^{1/2} c_2^{1/2} \wedge 1} \right)^{2k-2}.
\end{aligned}
$$

By symmetry of $X, \tilde{X}$, this concludes the proof. □