When and How Does Known Class Help Discover Unknown Ones? Provable Understanding Through Spectral Analysis

Yiyou Sun [1] **Zhenmei Shi** [1] **Yingyu Liang** [1] **Yixuan Li** [1]

Abstract

Novel Class Discovery (NCD) aims at inferring novel classes in an unlabeled set by leveraging prior knowledge from a labeled set with known classes. Despite its importance, there is a lack of theoretical foundations for NCD. This paper bridges the gap by providing an analytical framework to formalize and investigate *when and how known classes can help discover novel classes*. Tailored to the NCD problem, we introduce a graph-theoretic representation that can be learned by a novel NCD Spectral Contrastive Loss (NSCL). Minimizing this objective is equivalent to factorizing the graph's adjacency matrix, which allows us to derive a provable error bound and provide the sufficient and necessary condition for NCD. Empirically, NSCL can match or outperform several strong baselines on common benchmark datasets, which is appealing for practical usage while enjoying theoretical guarantees. Code is available at: `https://github.com/deeplearning-wisc/NSCL.git`.

1. Introduction

Though modern machine learning methods have achieved remarkable success (He et al., 2016; Chen et al., 2020; Song et al., 2020; Wang et al., 2022), the vast majority of learning algorithms have been driven by the closed-world setting, where the classes are assumed stationary and unchanged between training and testing. However, machine learning models in the open world will inevitably encounter novel classes that are outside the existing known categories (Sun et al., 2021; 2022; Ming et al., 2022; 2023). Novel Class Discovery (NCD) (Han et al., 2019) has emerged as an important problem, which aims to cluster similar samples in an unlabeled dataset (of novel classes) by way of utilizing knowledge from the labeled data (of known classes). Key to NCD is harnessing the power of labeled data for possible knowledge sharing and transfer to the unlabeled data (Hsu et al., 2018; Han et al., 2019; Hsu et al., 2019; Zhong et al., 2021b; Han et al., 2020a; Yang et al., 2022; Sun & Li, 2023).

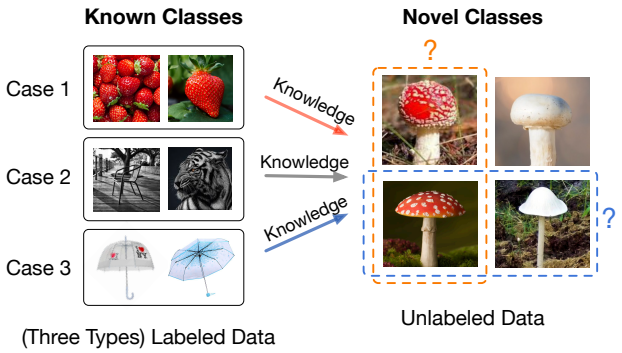

Figure 1. **Novel Class Discovery (NCD)** aims to cluster similar samples in unlabeled data (right), by way of utilizing knowledge from the labeled data (left). We illustrate scenarios where different known classes could result in different novel clusters (e.g., red mushrooms or mushrooms with umbrella shapes). This paper aims to provide a formal understanding.

One promising approach for NCD is to learn feature representation jointly from both labeled and unlabeled data, so that meaningful cluster structures emerge as novel classes. We argue that interesting intricacies can arise in this learning process—the resulting novel clusters may be very different, depending on the type of known class provided. We exemplify the nuances in Figure 1. In one scenario, the novel class "red mushroom" can be discovered, provided with the known class "strawberry" of a shared color feature. Alternatively, a different novel class can also emerge by grouping the bottom two images together (as "mushroom with umbrella shape" class), if the umbrella-shape images are given as a known class to the learner. We argue—perhaps obviously—that a formalized understanding of the intricate phenomenon is needed. This motivates our research:

When and how does the known class help discover novel classes?

[1]Department of Computer Sciences, University of Wisconsin - Madison. Correspondence to: Yiyou Sun, Yixuan Li <sunyiyou, sharonli@cs.wisc.edu>.

Proceedings of the 40th International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023. Copyright 2023 by the author(s).

Despite the empirical successes in recent years, there is a limited theoretical understanding and formalization for novel class discovery. To the best of our knowledge, there is no prior work that investigated this research question from a rigorous theoretical standpoint or provided provable error bound. Our work thus complements the existing works by filling in the critical blank.

In this paper, we start by formalizing a new learning algorithm that facilitates the understanding of NCD from a spectral analysis perspective. Our theoretical framework first introduces a graph-theoretic representation tailored for NCD, where the vertices are all the labeled and unlabeled data points, and classes form connected sub-graphs (Section 4.1). Based on this graph representation, we then introduce a new loss called NCD Spectral Contrastive Loss (NSCL) and show that minimizing our loss is equivalent to performing spectral decomposition on the graph (Section 4.2). Such equivalence allows us to derive the formal error bound for NCD based on the properties of the graph, which directly encodes the relations between known and novel classes.

We analyze the NCD quality by the linear probing performance on novel data, which is the least error of all possible linear classifiers with the learned representation. Our main result (Theorem 5.5) suggests that the linear probing error can be significantly reduced (even to 0) when the linear span of known samples' feature covers the "ignorance space" of unlabeled data in discovering novel classes. Lastly, we verify that our theoretical guarantees can translate into empirical effectiveness. In particular, NSCL establishes competitive performance on common NCD benchmarks, outperforming the best baseline by **10.6**% on the CIFAR-100-50 dataset (with 50 novel classes).

Our **main contributions** are:

1. We provide the first provable framework for the NCD problem, formalizing it by spectral decomposition of the graph containing both known and novel data. Our framework allows the research community to gain insights from a graph-theoretic perspective.

2. We propose a new loss called NCD Spectral Contrastive Loss (NSCL) and show that minimizing our loss is equivalent to performing singular decomposition on the graph. The loss leads to strong empirical performance while enjoying theoretical guarantees.

3. We provide theoretical insight by formally defining the semantic relationship between known and novel classes. Based on that, we derive an error bound of novel class discovery and investigate the sufficient and necessary conditions for the perfect discovery results.

2. Related Work

Novel class discovery. Early works tackled novel category discovery (NCD) as a transfer learning problem, such as DTC (Han et al., 2019), KCL (Hsu et al., 2018), MCL (Hsu et al., 2019). Many subsequent works incorporate representation learning for NCD, including RankStats (Han et al., 2020a), NCL (Zhong et al., 2021a) and UNO (Fini et al., 2021). CompEx (Yang et al., 2022) further uses a novelty detection module to better separate novel and known. However, none of the previous works theoretically analyzed the key question: *when and how do known classes help?* Li et al. (2022) try to answer this question from an empirical perspective by comparing labeled datasets from different levels of semantic similarity. Chi et al. (2021) directly define a solvable condition for the NCD problem but do not investigate the semantic relationship between known and novel classes. Our paper is the first work that systematically investigates the "when and how" questions by modeling the sample relevance from a graph-theoretic perspective and providing a provable error bound for the NCD problem.

Spectral graph theory. Spectral graph theory is a classic research problem (Chung, 1997; Cheeger, 2015; Kannan et al., 2004; Lee et al., 2014; McSherry, 2001), which aims to partition the graph by studying the eigenspace of the adjacency matrix. The spectral graph theory is also widely applied in machine learning (Ng et al., 2001; Shi & Malik, 2000; Blum, 2001; Zhu et al., 2003; Argyriou et al., 2005; Shaham et al., 2018). Recently, HaoChen et al. (2021) derive a spectral contrastive loss from the factorization of the graph's adjacency matrix which facilitates theoretical study in unsupervised domain adaptation (Shen et al., 2022; HaoChen et al., 2022). The graph definition in existing works is purely formed by the unlabeled data, whereas our graph and adjacency matrix is uniquely tailored for the NCD problem setting and consists of both labeled data from known classes and unlabeled data from novel classes. We offer new theoretical guarantees and insights based on the relations between known and novel classes, which has not been explored in the previous literature.

Theoretical analysis on contrastive learning. Recent works have advanced contrastive learning with empirical success (Chen et al., 2020; Khosla et al., 2020; Zhang et al., 2021; Wang et al., 2022), which necessitates a theoretical foundation. Arora et al. (2019); Lee et al. (2021); Tosh et al. (2021a;b); Balestriero & LeCun (2022); Shi et al. (2023) provided provable guarantees on the representations learned by contrastive learning for linear probing. Shen et al. (2022); HaoChen et al. (2021; 2022) further modeled the pairwise relation from the graphic view and provided error analysis of the downstream tasks. However, the existing body of work has mostly focused on *unsupervised learning*. There is no prior theoretical work considering the NCD problem

where both labeled and unlabeled data are presented. In this paper, we systematically investigate how the label information can change the representation manifold and affect the downstream novel class discovery task.

3. Setup

Formally, we describe the data setup and learning goal for novel class discovery (NCD).

Data setup. We consider the empirical training set $\mathcal{D}_l \cup \mathcal{D}_u$ as a union of labeled and unlabeled data. The labeled dataset is given by $\mathcal{D}_l = \{(\bar{x}_1, y_1), \ldots, (\bar{x}_i, y_i), \ldots\}$, where y_i belongs to known class space $\mathcal{Y}_l$; and the unlabeled dataset is $\mathcal{D}_u = \{\bar{x}_1, \ldots, \bar{x}_j, \ldots\}$. We assume that each unlabeled sample $\bar{x} \in \mathcal{D}_u$ belongs to one of the **novel** classes, *which do not overlap with the known classes* $\mathcal{Y}_l$. We use $\mathcal{P}_l$ and $\mathcal{P}_u$ to denote the marginal distributions of labeled and unlabeled data in the input space. Further, we let $\mathcal{P}_{l_i}$ denote the distribution of labeled samples with class label $i \in \mathcal{Y}_l$.

Learning goal. We assume that there exists an underlying class space $\mathcal{Y}_u = \{1, ..., |\mathcal{Y}_u|\}$ for unlabeled data $\mathcal{X}_u$, which is not revealed to the learner. The goal of novel class discovery is to learn a clustering for the novel data, which can be mapped to $\mathcal{Y}_u$ with low error.

4. Spectral Contrastive Learning for Novel Class Discovery

In this section, we introduce a new learning algorithm for NCD, from a graph-theoretic perspective. NCD is inherently a clustering problem—grouping similar points in unlabeled data $\mathcal{D}_u$ into the same cluster, by way of possibly utilizing helpful information from the labeled data $\mathcal{D}_l$. This clustering process can be fundamentally modeled by a graph, where the vertices are all the data points and classes form connected sub-graphs. Our novel framework first introduces a graph-theoretic representation for NCD, where edges connect similar data points (Section 4.1). We then propose a new loss that performs spectral decomposition on the similarity graph and can be written as a contrastive learning objective on neural net representations (Section 4.2).

4.1. Graph-Theoretic Representation for NCD

We start by formally defining the augmentation graph and adjacency matrix. For notation clarity, we use $\bar{x}$ to indicate the natural sample (raw inputs without augmentation). Given an $\bar{x}$, we use $\mathcal{T}(x|\bar{x})$ to denote the probability of x being augmented from $\bar{x}$. For instance, when $\bar{x}$ represents an image, $\mathcal{T}(\cdot|\bar{x})$ can be the distribution of common augmentations such as Gaussian blur, color distortion, and random cropping. The augmentation allows us to define a general population space $\mathcal{X}$, which contains all the original images along with their augmentations. In our case, $\mathcal{X}$ ($|\mathcal{X}| = N$) is composed of two parts $\mathcal{X}_l$ ($|\mathcal{X}_l| = N_l$), $\mathcal{X}_u$ ($|\mathcal{X}_u| = N_u$) which represents the division into labeled data with known classes and unlabeled data with novel classes respectively. Unlike unsupervised learning (Chen et al., 2020), NCD has access to both labeled and unlabeled data. This leads to two cases where two samples x and x^+ form a **positive pair** if:

(a) x and x^+ are augmented from the same unlabeled image $\bar{x}_u \sim \mathcal{P}_u$.

(b) x and x^+ are augmented from two labeled samples $\bar{x}_l$ and $\bar{x}'_l$ *with the same known class* i. In other words, both $\bar{x}_l$ and $\bar{x}'_l$ are drawn independently from $\mathcal{P}_{l_i}$.

We define the graph $G(\mathcal{X}, w)$ with vertex set $\mathcal{X}$ and edge weights w. For any two augmented data $x, x' \in \mathcal{X}$, $w_{xx'}$ is the marginal probability of generating the pair (x, x'):

$$\begin{aligned} w_{xx'} \triangleq \alpha \sum_{i \in \mathcal{Y}_l} \mathbb{E}_{\bar{x}_l \sim \mathcal{P}_{l_i}} \mathbb{E}_{\bar{x}'_l \sim \mathcal{P}_{l_i}} \underbrace{\mathcal{T}(x|\bar{x}_l)\mathcal{T}(x'|\bar{x}'_l)}_{\text{case (b)}} \\ + \beta \mathbb{E}_{\bar{x}_u \sim \mathcal{P}_u} \underbrace{\mathcal{T}(x|\bar{x}_u)\mathcal{T}(x'|\bar{x}_u)}_{\text{case (a)}}, \end{aligned} \tag{1}$$

where α, β modulates the importance between unlabeled and labeled data. The magnitude of $w_{xx'}$ indicates the "positiveness" or similarity between x and x'. We then use $w_x = \sum_{x' \in \mathcal{X}} w_{xx'}$ to denote the total edge weights connected to vertex x.

As a standard technique in graph theory (Chung, 1997), we use the *normalized adjacency matrix*:

$$\dot{A} \triangleq D^{-1/2} A D^{-1/2}, \tag{2}$$

where $A \in \mathbb{R}^{N \times N}$ is adjacency matrix with entries $A_{xx'} = w_{xx'}$ and $D \in \mathbb{R}^{N \times N}$ is a diagonal matrix with $D_{xx} = w_x$. The normalization balances the degree of each node, reducing the influence of vertices with very large degrees. The adjacency matrix defines the probability of x and x' being considered as the positive pair from the perspective of augmentation, which helps derive the NCD Spectral Contrastive Loss as we show next.

4.2. NCD Spectral Contrastive Learning

In this subsection, we propose a formal definition of NCD Spectral Contrastive Loss, which can be derived from a spectral decomposition of $\dot{A}$. The derivation of the loss is inspired by (HaoChen et al., 2021), and allows us to theoretically show the equivalence between learning feature embeddings and the projection on the top-k SVD components of $\dot{A}$. Importantly, such equivalence facilitates the theoretical understanding based on the semantic relation between known and novel classes encoded in $\dot{A}$.

Specifically, we consider low-rank matrix approximation:

$$\min_{F\in\mathbb{R}^{N\times k}} \mathcal{L}_{\mathrm{mf}}(F,A) \triangleq \left\|\dot{A} - FF^\top\right\|_F^2 \tag{3}$$

According to the Eckart–Young–Mirsky theorem (Eckart & Young, 1936), the minimizer of this loss function is $F^* \in \mathbb{R}^{N\times k}$ such that $F^*F^{*\top}$ contains the top-k components of $\dot{A}$'s SVD decomposition.

Now, if we view each row $\mathbf{f}_x^\top$ of F as a learned feature embedding $f : \mathcal{X} \mapsto \mathbb{R}^k$, the $\mathcal{L}_{\mathrm{mf}}(F,A)$ can be written as a form of the contrastive learning objective. We formalize this connection in Theorem 4.1 below.

Theorem 4.1. *We define $\mathbf{f}_x = \sqrt{w_x} f(x)$ for some function f. Recall α, β are hyper-parameters defined in Eq. (1). Then minimizing the loss function $\mathcal{L}_{\mathrm{mf}}(F,A)$ is equivalent to minimizing the following loss function for f, which we term **NCD Spectral Contrastive Loss (NSCL)**:*

$$\begin{aligned}\mathcal{L}_{nscl}(f) \triangleq & -2\alpha\mathcal{L}_1(f) - 2\beta\mathcal{L}_2(f) \\ & + \alpha^2\mathcal{L}_3(f) + 2\alpha\beta\mathcal{L}_4(f) + \beta^2\mathcal{L}_5(f),\end{aligned} \tag{4}$$

where

$$\mathcal{L}_1(f) = \sum_{i\in\mathcal{Y}_l} \mathop{\mathbb{E}}_{\substack{\bar{x}_l\sim\mathcal{P}_{l_i}, \bar{x}_l'\sim\mathcal{P}_{l_i}, \\ x\sim\mathcal{T}(\cdot|\bar{x}_l), x^+\sim\mathcal{T}(\cdot|\bar{x}_l')}} \left[f(x)^\top f\left(x^+\right)\right],$$

$$\mathcal{L}_2(f) = \mathop{\mathbb{E}}_{\substack{\bar{x}_u\sim\mathcal{P}_u, \\ x\sim\mathcal{T}(\cdot|\bar{x}_u), x^+\sim\mathcal{T}(\cdot|\bar{x}_u)}} \left[f(x)^\top f\left(x^+\right)\right],$$

$$\mathcal{L}_3(f) = \sum_{i\in\mathcal{Y}_l}\sum_{j\in\mathcal{Y}_l} \mathop{\mathbb{E}}_{\substack{\bar{x}_l\sim\mathcal{P}_{l_i}, \bar{x}_l'\sim\mathcal{P}_{l_j}, \\ x\sim\mathcal{T}(\cdot|\bar{x}_l), x^-\sim\mathcal{T}(\cdot|\bar{x}_l')}} \left[\left(f(x)^\top f\left(x^-\right)\right)^2\right],$$

$$\mathcal{L}_4(f) = \sum_{i\in\mathcal{Y}_l} \mathop{\mathbb{E}}_{\substack{\bar{x}_l\sim\mathcal{P}_{l_i}, \bar{x}_u\sim\mathcal{P}_u, \\ x\sim\mathcal{T}(\cdot|\bar{x}_l), x^-\sim\mathcal{T}(\cdot|\bar{x}_u)}} \left[\left(f(x)^\top f\left(x^-\right)\right)^2\right],$$

$$\mathcal{L}_5(f) = \mathop{\mathbb{E}}_{\substack{\bar{x}_u\sim\mathcal{P}_u, \bar{x}_u'\sim\mathcal{P}_u, \\ x\sim\mathcal{T}(\cdot|\bar{x}_u), x^-\sim\mathcal{T}(\cdot|\bar{x}_u')}} \left[\left(f(x)^\top f\left(x^-\right)\right)^2\right].$$

Proof. (*sketch*) We can expand $\mathcal{L}_{\mathrm{mf}}(F,A)$ and obtain

$$\mathcal{L}_{\mathrm{mf}}(F,A) = \sum_{x,x'\in\mathcal{X}} \left(\frac{w_{xx'}}{\sqrt{w_x w_{x'}}} - \mathbf{f}_x^\top \mathbf{f}_{x'}\right)^2 = const + \sum_{x,x'\in\mathcal{X}} \left(-2w_{xx'} f(x)^\top f\left(x'\right) + w_x w_{x'}\left(f(x)^\top f\left(x'\right)\right)^2\right)$$

The form of $\mathcal{L}_{nscl}(f)$ is derived from plugging $w_{xx'}$ (defined in Eq. (1)) and w_x. We include the details in Appendix A.2. □

Interpretation of $\mathcal{L}_{nscl}(f)$. At a high level, $\mathcal{L}_1$ and $\mathcal{L}_2$ push the embeddings of **positive pairs** to be closer while $\mathcal{L}_3$, $\mathcal{L}_4$ and $\mathcal{L}_5$ pull away the embeddings of **negative pairs**. In particular, $\mathcal{L}_1$ samples two random augmentation views of two images from labeled data with the **same** class label, and $\mathcal{L}_2$ samples two views from the same image in $\mathcal{X}_u$. For negative pairs, $\mathcal{L}_3$ uses two augmentation views from two samples in $\mathcal{X}_l$ with **any** class label. $\mathcal{L}_4$ uses two views of one sample in $\mathcal{X}_l$ and another one in $\mathcal{X}_u$. $\mathcal{L}_5$ uses two views from two random samples in $\mathcal{X}_u$.

5. Theoretical Analysis

So far we have presented a spectral approach for NCD based on the augmentation graph. Under this formulation, we now formally investigate and analyze: ***when and how does the known class help discover novel class?*** We start by showing that analyzing the linear probing performance is equivalent to analyzing the regression residual using singular vectors of $\dot{A}$ in Sec. 3. We then construct a toy example to illustrate and verify the key insight in Sec. 5.2. We finally provide a formal theory for the general case in Sec. 5.3.

5.1. Theoretical Setup

Representation for unlabeled data. We apply NCD spectral learning objective $\mathcal{L}_{nscl}(f)$ in Equation 4 and assume the optimizer is capable to obtain the representation that minimizes the loss. We can then obtain the F^* s.t. $F^*F^{*\top}$ are the top-k components of $\dot{A}$'s SVD decomposition. To ease the analysis, we will focus on the top-k singular vectors $V^* \in \mathbb{R}^{N\times k}$ of $\dot{A}$ such that $F^* = V^*\sqrt{\Sigma_k}$, where Σ_k is the diagonal matrix with top-k singular values $(\sigma_1, ..., \sigma_k)$.

Since we are primarily interested in the unlabeled data, we split V^* into two parts: $U^* \in \mathbb{R}^{N_u\times k}$ for unlabeled data and $L^* \in \mathbb{R}^{N_l\times k}$ for labeled data, respectively. Assuming the first N_l rows/columns in $\dot{A}$ corresponds to the labeled data, we can conveniently rewrite V^* as:

$$V^* = \begin{bmatrix} L^*(\text{labeled part}) \\ U^*(\text{unlabeled part}) \end{bmatrix} \tag{5}$$

Linear probing evaluation. With the learned representation for the unlabeled data, we can evaluate NCD quality by the linear probing performance. The strategy is commonly used in self-supervised learning (Chen et al., 2020). Specifically, the weight of a linear classifier is denoted as $\mathbf{M} \in \mathbb{R}^{k\times|\mathcal{Y}_u|}$. The class prediction is given by $h(x; f, \mathbf{M}) = \mathrm{argmax}_{i\in\mathcal{Y}_u}(f(x)^\top\mathbf{M})_i$. The linear probing performance is given by the least error of all possible linear classifiers:

$$\mathcal{E}(f) \triangleq \min_{\mathbf{M}\in\mathbb{R}^{k\times|\mathcal{Y}_u|}} \sum_{x\in\mathcal{X}_u} \mathbb{1}\left[y(x) \neq h(x; f, \mathbf{M})\right], \tag{6}$$

where $y(x)$ indicates the ground-truth class of x.

Residual analysis. With defined U^*, we can bound the linear probing error $\mathcal{E}(f)$ by the residual of the regression

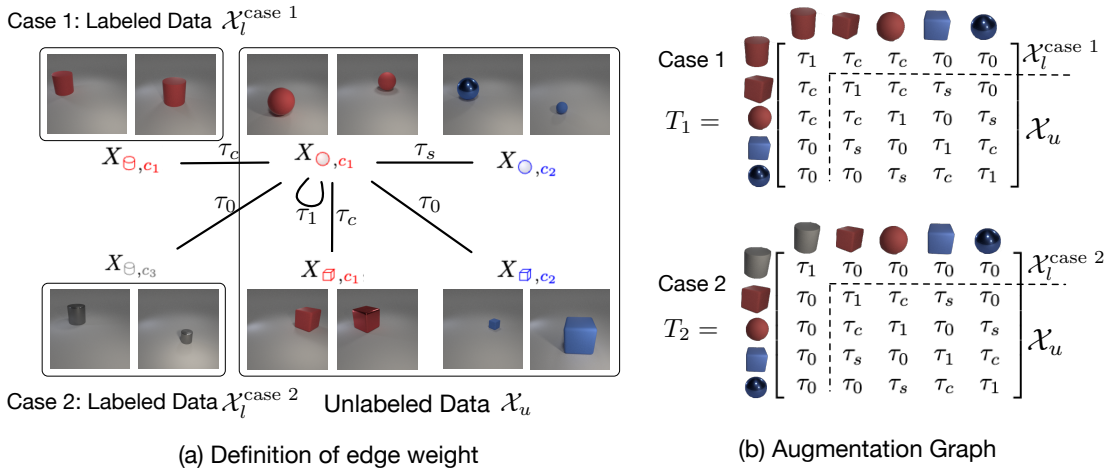

Figure 2. An illustrative example for theoretical analysis. **(a)** The unlabeled data $\mathcal{X}_u$ consists of 3D objects of sphere/cube with red/blue colors. We consider two cases of labeled data: (1) Case 1 uses a red cylinder $X_{\ominus,c_1}$ which is correlated with the target novel class (red). (2) Case 2 uses gray cylinder $X_{\ominus,c_3}$ which has no correlation with $\mathcal{X}_u$. **(b)** The augmentation matrices for case 1 and case 2 respectively. See definition in Eq. (7). Best viewed in color.

error $\mathcal{R}(U^*)$ as we show in Lemma 5.1 with proof in Appendix A.1.

Lemma 5.1. *Denote the* $\mathbf{y}(x) \in \mathbb{R}^{|\mathcal{Y}_u|}$ *as a one-hot vector whose* $y(x)$*-th position is 1 and 0 elsewhere. Let* $\mathbf{Y} \in \mathbb{R}^{N_u \times |\mathcal{Y}_u|}$ *as a binary mask whose rows are stacked by* $\mathbf{y}(x)$*. We have:*

$$\mathcal{R}(U^*) \triangleq \min_{\mathbf{M} \in \mathbb{R}^{k \times |\mathcal{Y}_u|}} \|\mathbf{Y} - U^* \mathbf{M}\|_F^2 \geq \frac{1}{2}\mathcal{E}(f).$$

Note that we can rewrite $\mathcal{R}(U^*)$ as the summation of individual residual terms $\mathcal{R}(U^*, \vec{y}_i)$: $\mathcal{R}(U^*) = \sum_{i \in \mathcal{Y}_u} \mathcal{R}(U^*, \vec{y}_i)$, where

$$\mathcal{R}(U^*, \vec{y}_i) \triangleq \min_{\vec{\mu}_i \in \mathbb{R}^k} \|\vec{y}_i - U^* \vec{\mu}_i\|_2^2,$$

and $\vec{y}_i \in \mathbb{R}^{N_u}$ is the i-th column of $\mathbf{Y}$ and $\vec{\mu}_i \in \mathbb{R}^k$ is the i-th column of $\mathbf{M}$. Without losing the generality, our analysis will revolve around the residual term $\mathcal{R}(U^*, \vec{y}_i)$ for specific class i. It is clear that if learned representation U^* encodes more information of the label vector $\vec{y}_i$, the residual $\mathcal{R}(U^*, \vec{y}_i)$ becomes smaller[1]. Such insight can be used to investigate which type of known class is more helpful for learning the representation of novel classes.

5.2. An Illustrative Example

We consider a toy example that helps illustrate the core idea of our theoretical findings. Specifically, the example aims to cluster 3D objects of different colors and shapes, as shown in Figure 2 (a). These images are generated by a 3D rendering software (Johnson et al., 2017) with user-defined properties including colors, shape, size, position, etc.

In what follows, we define two data configurations and corresponding graphs, where the labeled data is correlated with the attribute of unlabeled data (**case 1**) vs. not (**case 2**). We are interested in contrasting the representations (in form of singular vectors) and residuals derived from both scenarios. The proof of all theorems in this section is provided in Appendix B.

Motivation and data design. For simplicity, we focus on two main properties: color and shape. Formally, the images with shape s and color c are sampled from a generation procedure $\mathcal{G}$:

$$X_{s,c} \sim \mathcal{G}(s, c),$$

where $s \in \{\square\,(\text{cube}), \bigcirc\,(\text{sphere}), \ominus\,(\text{cylinder})\}$ and $c \in \{c_1(\text{red}), c_2(\text{blue}), c_3(\text{gray})\}$. We then construct our unlabeled dataset containing red/blue cubes/spheres as:

$$\mathcal{X}_u \triangleq \{X_{\square,c_1}, X_{\bigcirc,c_1}, X_{\square,c_2}, X_{\bigcirc,c_2}\}.$$

For simplicity, we assume each element in $\mathcal{X}_u$ is a single example. W.o.l.g, we also assume the red cube and red sphere form the target novel class. Then the corresponding labeling vector on $\mathcal{X}_u$ is defined by:

$$\vec{y} \triangleq \{1, 1, 0, 0\}.$$

To answer *"when and how does the known class help discover novel class?"*, we construct two separate scenarios: one helps and the other one does not. Specifically, in the first

[1] In an extreme case, if the first column of U^* is exactly the same as $\vec{y}_i$, one can set $\vec{\mu}_i = [1, 0, 0, ...]^\top$ to make residual zero.

case, we let the labeled data $\mathcal{X}_l^{\text{case 1}}$ be strongly correlated with the target class (red color) in unlabeled data:

$$\mathcal{X}_l^{\text{case 1}} \triangleq \{X_{\ominus,c_1}\}(\text{red cylinder}).$$

In the second case, we construct the labeled data that has no correlation with any novel classes. We use gray cylinders which have no overlap in either shape and color:

$$\mathcal{X}_l^{\text{case 2}} \triangleq \{X_{\ominus,c_3}\}(\text{gray cylinder}).$$

Putting it together, our entire training dataset is $\mathcal{X}^{\text{case 1}} = \mathcal{X}_l^{\text{case 1}} \cup \mathcal{X}_u$ or $\mathcal{X}^{\text{case 2}} = \mathcal{X}_l^{\text{case 2}} \cup \mathcal{X}_u$. We aim to verify the hypothesis that: *the representation learned by $\mathcal{X}^{\text{case 1}}$ provides a much smaller regression residual to $\vec{y}$ than $\mathcal{X}^{\text{case 2}}$ for color class.*

Augmentation graph. Based on the data, we now define the probability of augmenting an image $X_{s,c}$ to another $X'_{s',c'}$:

$$\mathcal{T}\left(X'_{s',c'} \mid X_{s,c}\right) = \begin{cases} \tau_1 & \text{if} \quad s = s', c = c', \\ \tau_s & \text{if} \quad s = s', c \neq c', \\ \tau_c & \text{if} \quad s \neq s', c = c', \\ \tau_0 & \text{if} \quad s \neq s', c \neq c', \end{cases} \tag{7}$$

It is natural to assume the magnitude order that follows $\tau_1 \gg \max(\tau_s, \tau_c)$ and $\min(\tau_s, \tau_c) \gg \tau_0$. In two data settings $\mathcal{X}^{\text{case 1}}$ and $\mathcal{X}^{\text{case 2}}$, the corresponding augmentation matrices T_1, T_2 formed by $\mathcal{T}(\cdot|\cdot)$ are presented in Fig. 2 (b). According to Eq. (1), it can be verified that the adjacency matrices are $A_1 = T_1^2$ and $A_2 = T_2^2$ respectively.

Main analysis. We are primarily interested in analyzing the difference of the representation space derived from A_1 vs. A_2. Since $\tau_1 \gg \max(\tau_s, \tau_c)$, one can show that A_1 and A_2 are positive-definite. The singular vector is thus equivalent to the eigenvector. Also note that A_1 and their square root T_1 have the same eigenvectors and order. It is thus equivalent to analyzing the eigenvectors of T_1. Same with A_2 and T_2. In this toy example, we consider the eigenvalue problem of the unnormalized adjacency matrix[2] for simplicity.

We put analysis on the top-2 eigenvectors $V_1^*, V_2^* \in \mathbb{R}^{5\times 2}$ for A_1/A_2 —- as we will see later, the top-1 eigenvector of T_1/T_2 usually functions at distinguishing known vs novel data, while the 2nd eigenvector functions at distinguishing color or shape.

We let $U_1^* \in \mathbb{R}^{4\times 2}$ contains the last 4 rows of V_1^*, and corresponds to the "representation" for the unlabeled data only. U_2^* is defined in the same way *w.r.t.* A_2. We have the following theorem:

Theorem 5.2. *Assume $\tau_1 = 1$, $\tau_0 = 0$, $\tau_s < 1.5\tau_c$. We have*

$$U_1^* = \begin{bmatrix} a_1 & a_1 & b_1 & b_1 \\ a_2 & a_2 & b_2 & b_2 \end{bmatrix}^\top,$$

where a_1, b_1 are some positive real numbers, and a_2, b_2 has different signs.

$$U_2^* = \begin{cases} \frac{1}{2}\begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 & -1 & -1 \end{bmatrix}^\top, & \text{if } \tau_s < \tau_c, \\ \frac{1}{2}\begin{bmatrix} 1 & 1 & 1 & 1 \\ -1 & 1 & -1 & 1 \end{bmatrix}^\top, & \text{if } \tau_s > \tau_c, \end{cases}$$

With label vector $\vec{y} = \{1, 1, 0, 0\}$, we have

$$\mathcal{R}(U_1^*, \vec{y}) = 0, \mathcal{R}(U_2^*, \vec{y}) = \begin{cases} 0, & \text{if } \tau_s < \tau_c \\ 1, & \text{if } \tau_s > \tau_c. \end{cases} \tag{8}$$

Interpretation of Theorem 5.2: The discussion can be divided into two cases: (1) $\tau_s < \tau_c$. (2) $\tau_s > \tau_c$. In the **first case** $\tau_s < \tau_c$, the connection between the same-color data pair is already stronger than the same-shape data pair. Thus the eigenvector corresponding to color information ($\frac{1}{2}[1, 1, -1, -1]^\top$) will be more prominent (and ranked higher in U_2^*) than "shape eigenvector" ($\frac{1}{2}[-1, 1, -1, 1]^\top$). Since the feature U_2^* already encodes sufficient information (color) of the labeling vector $\vec{y}$, fitting $\vec{y}$ becomes easy and the residual $\mathcal{R}(U_2^*, \vec{y})$ becomes 0.

In NCD, **we are more interested in the second case** ($\tau_s > \tau_c$), where unlabeled data indeed need some help from labeled data for better clustering. Such help comes from the semantic connection between labeled data and unlabeled data. In our toy example, the semantic connection comes from the first row/column of T_1 and T_2. However, the first row/column of T_2 is $[1, 0, 0, 0, 0]$, which means there is no extra information offered from $\mathcal{X}_l^{\text{case 2}}$. It is because $\mathcal{X}_l^{\text{case 2}}$ contains gray cylinders which have neither colors nor shapes connection to unlabeled data $\mathcal{X}_u$. Contrarily, $\mathcal{X}_l^{\text{case 1}}$ with red cylinder provides strong color prior. This allows the "color eigenvector" ($[a_2, a_2, -b_2, -b_2]$) to become a main component in U_1^*, making the residual $\mathcal{R}(U_1^*, \vec{y}) = 0$ even when $\tau_s > \tau_c$.

Main takeaway. In Theorem 5.2, we have verified the hypothesis that incorporating labeled data $\mathcal{X}_l^{\text{case 1}}$ (red cylinder) can reduce the residual $\mathcal{R}(U_1^*, \vec{y})$ more than using $\mathcal{X}_l^{\text{case 2}}$, especially when color is a weaker signal than shape in unlabeled data.

Extension: A more general result. Note that T_1 and T_2 are special cases of the following $T(t)$ with $t \in [\tau_0, \tau_c]$:

$$T(t) = \begin{bmatrix} \tau_1 & t & t & \tau_0 & \tau_0 \\ t & \tau_1 & \tau_c & \tau_s & \tau_0 \\ t & \tau_c & \tau_1 & \tau_0 & \tau_s \\ \tau_0 & \tau_s & \tau_0 & \tau_1 & \tau_c \\ \tau_0 & \tau_0 & \tau_s & \tau_c & \tau_1 \end{bmatrix},$$

where t indicates the strength of the connection between labeled data and a novel class in unlabeled data. Let U_t^*

[2]The normalized/unnormalized adjacency matrix corresponds to the NCut/RatioCut problem respectively (Von Luxburg, 2007).

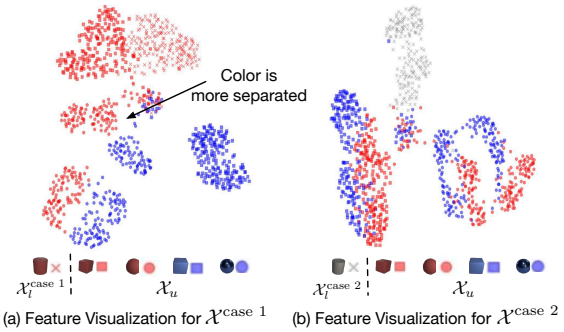

Figure 3. UMAP (McInnes et al., 2018) visualization of the feature embedding learned from $\mathcal{X}^{\text{case 1}}$ and $\mathcal{X}^{\text{case 2}}$ respectively. The model is trained with NCD Spectral Contrastive Loss.

be the representation for unlabeled data derived from $T(t)$. The following theorem indicates that the residual decreases when t increases and the residual becomes 0 when t is larger than a threshold $\bar{t}$ depending on the gap between τ_s and τ_c.

Theorem 5.3. *Assume $\tau_1 = 1$, $\tau_0 = 0$, $1.5\tau_c > \tau_s > \tau_c$. Let $\bar{t} = \sqrt{\frac{2(\tau_s-\tau_c)^2\tau_c}{2\tau_c-\tau_s}}$, $r : \mathbb{R} \mapsto (0,1)$ be a real value function, we have*

$$\mathcal{R}(U_t^*, \vec{y}) = \begin{cases} 0, & \text{if } t \in (\bar{t}, \tau_s), \\ r(t), & \text{if } t \in (0, \bar{t}), \\ 1, & \text{if } t = 0. \end{cases} \tag{9}$$

Can adding labeled data be harmful? We exemplify the scenario in Figure 1, where the umbrella images are given as a known class, undesirably causing the "mushroom with umbrella shape" to be grouped together. To formally analyze this case, we construct **case 3**:

$$\mathcal{X}_l^{\text{case 3}} \triangleq \{X_{\text{☂},c_3}\}(\text{gray cube}).$$

In this case, we have the following Lemma 5.4.

Lemma 5.4. *If $\frac{\tau_c}{\tau_s} \in (1, 1.5)$, $\mathcal{R}(U_3^*, \vec{y}) - \mathcal{R}(U_2^*, \vec{y}) = 1$.*

The residual in case 3 is now larger than in case 2, since the shape is treated as a more important feature than the color feature (which relates to the target class). The main takeaway of this lemma is that the labeled data can be harmful when its connection with unlabeled data is undesirably stronger in the spurious feature dimension.

Qualitative results. The theoretical results can be verified in our empirical results by visualization in Fig. 3. Due to the space limitation, we include experimental details in Appendix D.2. As seen in Fig. 3 (a), the features of unlabeled data $\mathcal{X}_u$ jointly learned with red cylinder $\mathcal{X}_l^{\text{case 1}}$ are more distinguishable by color attribute, as opposed to Fig. 3 (b).

5.3. Main Theory

The toy example offers an important insight that using the labeled data help reduce the residual when it provides the missing information of unlabeled data. In this section, we will formalize this insight by extending the toy example to a more general setting with N samples. We start with the definition of notations.

Notations. Recall that $V^* \in \mathbb{R}^{N\times k}$ is defined as the top-k singular vectors of $\dot{A}$, which is further split into two parts $L^* = [l_1, l_2, \cdots, l_k] \in \mathbb{R}^{N_l\times k}$, $U^* = [u_1, u_2, \cdots, u_k] \in \mathbb{R}^{N_u\times k}$, for labeled and unlabeled samples respectively. Then we let $V^\flat \in \mathbb{R}^{N\times(N-k)}$ be the remaining singular vectors of $\dot{A}$ except top-k. Similarly, we split $V^\flat$ into two parts ($L^\flat = [l_{k+1}, l_{k+2}, \cdots, l_N] \in \mathbb{R}^{N_l\times(N-k)}$, $U^\flat = [u_{k+1}, u_{k+2}, \cdots, u_N] \in \mathbb{R}^{N_u\times(N-k)}$).

We now present our first main result in Theorem 5.5.

Theorem 5.5. *Denote the projection matrix $\mathsf{P}_{L^\flat} = L^{\flat\top}(L^\flat L^{\flat\top})^\dagger L^\flat$, where † denotes the Moore-Penrose inverse. For any labeling vector $\vec{y} \in \{0,1\}^{N_u}$, we have*

$$\mathcal{R}(U^*, \vec{y}) \leq \|(I - \mathsf{P}_{L^\flat})U^{\flat\top}\vec{y}\|_2^2. \tag{10}$$

Interpretation of Theorem 5.5. The bound of residual in Ineq. (10) is composed of two projections: $U^{\flat\top}$ and $(I - \mathsf{P}_{L^\flat})$. We first consider the ignorance space formed by the first projection:

$$\textbf{ignorance space} \triangleq U^{\flat\top}\vec{y},$$

which contains the information of the labeling vector $\vec{y}$ that is not encoded in the learned representation U^* of the unlabeled data. Intuitively, when $\mathcal{R}(U^*, \vec{y}) > 0$, the labeling vector $\vec{y}$ does not lie in the span of the existing representation U^*. On the other hand, $\mathcal{R}\left([\ U^* \quad U^\flat\ ], \vec{y}\right) = 0$ since U^* together with $U^\flat$ forms a full rank space. We also define a measure of the ignorance degree of the current feature space: **ignorance degree** $\triangleq \mathfrak{T}(\vec{y}) = \frac{\|U^{\flat\top}\vec{y}\|_2}{\|\vec{y}\|_2}$.

The second projection matrix $(I - \mathsf{P}_{L^\flat})$ is composed of $L^\flat$, which we deem as the extra knowledge from known classes:

$$\textbf{extra knowledge} \triangleq L^\flat.$$

Multiplying the second projection matrix $(I - \mathsf{P}_{L^\flat})$ further reduces the norm of the ignorance space by considering the extra knowledge from labeled data, since $\mathsf{P}_{L^\flat}$ is a projection matrix that projects a vector to the linear span of $L^\flat$. In the extreme case, when $U^{\flat\top}\vec{y}$ fully lies in the linear span of $L^\flat$, the residual $\mathcal{R}(U^*, \vec{y})$ goes 0.

Next, we present another main theorem that bounds the linear probing error $\mathcal{E}(f)$ based on the relations between the known and novel classes. See Appendix C.3 for a detailed discussion and assumption.

Theorem 5.6. *Let* $[\ A_{ul} \in \mathbb{R}^{N_u \times N_l}, A_{uu} \in \mathbb{R}^{N_u \times N_u}\ ]$ *be the sub-matrix of the last* N_u *rows of* $\dot{A}$*, and* q_i *be the* i*-th eigenvector of* A_{uu}*. The linear probing error can be bounded as follows:*

$$\mathcal{E}(f) \lesssim \frac{2N_u}{|\mathcal{Y}_u|}\left(\sum_i^{|\mathcal{Y}_u|} \underbrace{\mathfrak{T}(\vec{y}_i)}_{\text{ignorance degree}} (1 - \underbrace{\kappa(\vec{y}_i)^2}_{\text{knowledge coverage}}) + \frac{\|\dot{A} - \bar{A}\|_2}{\sigma_k - \sigma_{k+1}}\right),$$

where

$$\kappa(\vec{y}) = \cos(\bar{U}^{\flat\top}\vec{y}, \bar{\mathfrak{l}}^{\flat}) \gtrsim \min_{i>k, j>k} \frac{2\sqrt{\frac{\vec{y}^\top q_i}{\vec{\eta}_u^\top q_i}\frac{\vec{y}^\top q_j}{\vec{\eta}_u^\top q_j}}}{\frac{\vec{y}^\top q_i}{\vec{\eta}_u^\top q_i} + \frac{\vec{y}^\top q_j}{\vec{\eta}_u^\top q_j}},$$

and $\bar{A}$ is the approximation of $\dot{A}$ by taking the expectation in the rows/columns of labeled samples (Appendix C.2) with a similar motivation as the SBM model (Holland et al., 1983). In such condition, $\bar{U}^{\flat\top}$, $\bar{\mathfrak{l}}^{\flat}$ and η_u is the approximation to $U^{\flat\top}$, $L^{\flat}$ and A_{ul} accordingly.

Interpretation of $\kappa(\vec{y})$. We provide the detailed derivation of $\kappa(\vec{y})$ in Lemma C.10. Intuitively, $\kappa(\vec{y})$ measures the usefulness and relevance of knowledge from known classes for NCD. We formally call it coverage, which measures the cosine distance between the ignorance space and the extra knowledge:

$$\textbf{coverage} \triangleq \kappa(\vec{y}) = \cos(\underbrace{\bar{U}^{\flat\top}\vec{y}}_{\text{ignorance space}}, \underbrace{\bar{\mathfrak{l}}^{\flat}}_{\text{extra knowledge}}).$$

Our Theorem 5.6 thus meaningfully shows that the linear probing error can be bounded more tightly as $\kappa(\vec{y})$ increases (i.e., when labeled data provides more useful information for the unlabeled data).

Implication of Theorem 5.6. Our theorem allows us to formalize answers to the "When and How" question. Firstly, the Theorem answers "*how the labeled data helps*"—because the knowledge from the known classes changes the representation of unlabeled data and reduces the ignorance space for novel class discovery. Secondly, the Theorem answers "*when the labeled data helps*". Specifically, labeled data helps when the coverage between ignorance space and extra knowledge is nonzero. In the extreme case, if the extra knowledge fully covers the ignorance space, we get the perfect performance (0 linear probing error).

6. Experiments on Common Benchmarks

Beyond theoretical insights, we show empirically that our proposed NCD spectral loss is effective on common benchmark datasets CIFAR-10 and CIFAR-100 (Krizhevsky et al., 2009). Following the well-established NCD benchmarks (Han et al., 2019; 2020b; Fini et al., 2021), each dataset is divided into two subsets, the labeled set that contains labeled images belonging to a set of known classes, and an unlabeled set with novel classes. Our comparison is on three benchmarks: `C10-5` means CIFAR-10 datasets split with 5 known classes and 5 novel classes and `C100-80` means CIFAR-100 datasets split with 80 known classes while `C100-50` has 50 known classes. The division is consistent with Fini et al. (2021). We train the model by the proposed NSCL algorithm with details in Appendix D.1 and measure performance on the features in the penultimate layer of ResNet-18.

Table 1. Main Results. Results are reported in clustering accuracy (%) on the *training* split of the novel set. With the learned feature, we perform a K-Means clustering with the default setting in Python's `sklearn` package. The accuracy of the novel classes is measured by solving an optimal assignment problem using the Hungarian algorithm (Kuhn, 1955). "C" is short for CIFAR. SCL denotes training with Spectral Contrastive Loss purely on $\mathcal{D}_u$ while SCL‡ is trained on $\mathcal{D}_u \cup \mathcal{D}_l$ unsupervisedly.

Method	C10-5	C100-80	C100-50
KCL (Hsu et al., 2018)	72.3	42.1	-
MCL (Hsu et al., 2019)	70.9	21.5	-
DTC (Han et al., 2019)	88.7	67.3	35.9
RS+ (Han et al., 2020a)	91.7	75.2	44.1
DualRank (Zhao & Han, 2021)	91.6	75.3	-
Joint (Jia et al., 2021)	93.4	76.4	-
UNO (Fini et al., 2021)	92.6	85.0	52.9
ComEx (Yang et al., 2022)	93.6	85.7	53.4
SCL (HaoChen et al., 2021)	92.4	72.7	51.8
SCL‡ (HaoChen et al., 2021)	93.7	68.9	53.3
NSCL (Ours)	**97.5**	**85.9**	**64.0**

NSCL is competitive in discovering novel classes. Our proposed loss NSCL is amenable to the theoretical understanding of NCD, which is our primary goal of this work. Beyond theory, we show that NSCL is equally desirable in empirical performance. In particular, NSCL outperforms its rivals by a significant margin, as evidenced in Table 1. Our comparison covers an extensive collection of common NCD algorithms and baselines. In particular, on C100-50, we improve upon the best baseline ComEx by **10.6**%. This finding further validates that putting analysis on NSCL is appealing for both theoretical and empirical reasons.

Ablation study on the unsupervised counterpart. To verify whether the known classes indeed help discover new classes, we compare NSCL with the unsupervised counterpart (dubbed SCL) that is purely trained on the unlabeled data $\mathcal{D}_u$. Results show that the labeled data offers tremendous help and improves **13.2**% in novel class accuracy.

Supervision signals are important in the labeled data. We also analyze how much the supervision signals in labeled

Table 2. Comparison of results reported in overall/novel/known accuracy (%) on the *test* split of CIFAR. The three metrics are calculated as follows. (1) **Known accuracy**: For the features from the labeled data, we train an additional linear head by linear probing and then measure classification accuracy based on the prediction $\vec{h}_l$; (2) **Novel accuracy**: For features from the unlabeled data, we perform a K-Means clustering with the default setting in Python's `sklearn` package, which produces the clustering prediction $\vec{h}_u$. The clustering accuracy is further measured by solving an optimal assignment problem using the Hungarian algorithm (Kuhn, 1955); (3) **Overall accuracy**. The overall accuracy is measured by concatenating the prediction $\vec{h}_l$ and $\vec{h}_u$ and then solving the assignment problem.

Method	C10-5			C100-50		
	All	Novel	Known	All	Novel	Known
DTC (Han et al., 2019)	68.7	78.6	58.7	32.5	34.7	30.2
RankStats (Han et al., 2020a)	89.7	88.8	90.6	55.3	40.9	69.7
UNO (Fini et al., 2021)	**95.8**	95.1	96.6	65.4	52.0	78.8
ComEx (Yang et al., 2022)	95.0	93.2	**96.7**	67.2	54.5	**80.1**
NSCL (Ours)	95.5	**96.7**	94.2	**67.4**	**57.1**	77.4

data help. To investigate it, we compare our method NSCL with SCL trained on $\mathcal{D}_u \cup \mathcal{D}_l$ in a purely unsupervised manner. The difference is that SCL does not utilize the label information in $\mathcal{D}_l$. We denote this setting as SCL‡ in Table 1. Results show that NSCL provides stronger performance than SCL‡. The ablation suggests that relevant knowledge of known classes indeed provides meaningful help in novel class discovery.

NSCL is competitive in the inductive setting. We report performance comparison in Table 2, comprehensively measuring three accuracy metrics—for all/novel/known classes respectively. Different from Table 1 which reports clustering results in a transductive manner, the performance in Table 2 is reported on the test split. For evaluation, we first collect the feature representations and then report overall/novel/known accuracy with inference details provided in the caption of Table 2. We see that NSCL establishes comparable performance with baselines on the labeled data from known classes and superior performance on novel class discovery. Notably, NSCL outperforms UNO (Fini et al., 2021) on `C10-5` by 1.6% and outperforms ComEx (Yang et al., 2022) by 2.6% on `C100-50` in terms of novel accuracy.

7. Conclusion

In this paper, we present a theoretical framework of novel class discovery and provide new insight on the research question: "*when and how does the known class help discover novel classes?*". Specifically, we propose a graph-theoretic representation that can be learned through a new NCD Spectral Contrastive Loss (NSCL). Minimizing this objective is equivalent to factoring the graph's adjacency matrix, which allows us to analyze the NCD quality by measuring the linear probing error on novel samples' features. Our main result (Theorem 5.5) suggests such error can be significantly reduced (even to 0) when the linear span of known samples' feature covers the "ignorance space" of unlabeled data in discovering novel classes. Our framework is also empirically appealing to use since it can achieve similar or better performance than existing methods on benchmark datasets.

Broader impacts. Our new framework opens a new door to the NCD community in the following way:

- NSCL provides a framework to answer the fundamental question that is shared across all NCD methods. At a high level, NSCL analyzes how the new knowledge changes the representation space that leads to different discovery outcomes. This finding can be generalizable to other NCD methods which may differ in the way of incorporating new knowledge.
- NSCL can be compatible with prior NCD methods. Note that NSCL is a representation learning method. With that being said, one can possibly "plug" NSCL into existing learning objectives for NCD. Take the two most popular prior works in NCD as an example. For example, we can use the encoder learned by NSCL in RS+ (Han et al., 2020a) and UNO (Fini et al., 2021).

To summarize, NSCL is an important building block in the NCD research area and have broader impacts both theoretically and empirically.

Acknowledgement

Li is supported in part by the AFOSR Young Investigator Award under No. FA9550-23-1-0184; Philanthropic Fund from SFF; and faculty research awards/gifts from Google, Meta, and Amazon. Liang is partially supported by Air Force Grant FA9550-18-1-0166, the National Science Foundation (NSF) Grants 2008559-IIS and CCF-2046710. Any opinions, findings, conclusions, or recommendations expressed in this material are those of the authors and do not necessarily reflect the views, policies, or endorsements either expressed or implied, of the sponsors. The authors would also like to thank ICML reviewers for their helpful suggestions and feedback.

References

Argyriou, A., Herbster, M., and Pontil, M. Combining graph laplacians for semi–supervised learning. *Advances in Neural Information Processing Systems*, 18, 2005.

Arora, S., Khandeparkar, H., Khodak, M., Plevrakis, O., and Saunshi, N. A theoretical analysis of contrastive unsupervised representation learning. *International Conference on Machine Learning*, 2019.

Balestriero, R. and LeCun, Y. Contrastive and non-contrastive self-supervised learning recover global and local spectral embedding methods. *Advances in Neural Information Processing Systems*, 2022.

Blum, A. Learning form labeled and unlabeled data using graph mincuts. In *18th International Conference on Machine Learning*, 2001.

Cheeger, J. A lower bound for the smallest eigenvalue of the laplacian. In *Problems in analysis*, pp. 195–200. Princeton University Press, 2015.

Chen, T., Kornblith, S., Norouzi, M., and Hinton, G. A simple framework for contrastive learning of visual representations. In *Proceedings of the international conference on machine learning*, pp. 1597–1607. PMLR, 2020.

Chen, X. and He, K. Exploring simple siamese representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15750–15758, 2021.

Chi, H., Liu, F., Yang, W., Lan, L., Liu, T., Han, B., Niu, G., Zhou, M., and Sugiyama, M. Meta discovery: Learning to discover novel classes given very limited data. In *International Conference on Learning Representations*, 2021.

Chung, F. R. *Spectral graph theory*, volume 92. American Mathematical Soc., 1997.

Eckart, C. and Young, G. The approximation of one matrix by another of lower rank. *Psychometrika*, 1(3):211–218, 1936.

Fini, E., Sangineto, E., Lathuilière, S., Zhong, Z., Nabi, M., and Ricci, E. A unified objective for novel class discovery. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 9284–9292, 2021.

Han, K., Vedaldi, A., and Zisserman, A. Learning to discover novel visual categories via deep transfer clustering. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019.

Han, K., Rebuffi, S., Ehrhardt, S., Vedaldi, A., and Zisserman, A. Automatically discovering and learning new visual categories with ranking statistics. In *Proceedings of the 8th Intennational Conference on Learning Representations*. Schloss Dagstuhl-Leibniz-Zentrum für Informatik, 2020a.

Han, K., Rebuffi, S.-A., Ehrhardt, S., Vedaldi, A., and Zisserman, A. Automatically discovering and learning new visual categories with ranking statistics. *International Conference on Learning Representations*, 2020b.

HaoChen, J. Z., Wei, C., Gaidon, A., and Ma, T. Provable guarantees for self-supervised deep learning with spectral contrastive loss. *Advances in Neural Information Processing Systems*, 34:5000–5011, 2021.

HaoChen, J. Z., Wei, C., Kumar, A., and Ma, T. Beyond separability: Analyzing the linear transferability of contrastive representations to related subpopulations. *Advances in neural information processing systems*, 2022.

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Holland, P. W., Laskey, K. B., and Leinhardt, S. Stochastic blockmodels: First steps. *Social networks*, 5(2):109–137, 1983.

Hsu, Y.-C., Lv, Z., and Kira, Z. Learning to cluster in order to transfer across domains and tasks. *Proceedings of the International Conference on Learning Representations*, 2018.

Hsu, Y.-C., Lv, Z., Schlosser, J., Odom, P., and Kira, Z. Multi-class classification without multi-class labels. *Proceedings of the International Conference on Learning Representations*, 2019.

Jia, X., Han, K., Zhu, Y., and Green, B. Joint representation learning and novel category discovery on single-and multi-modal data. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 610–619, 2021.

Johnson, J., Hariharan, B., van der Maaten, L., Fei-Fei, L., Zitnick, C. L., and Girshick, R. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In *CVPR*, 2017.

Kannan, R., Vempala, S., and Vetta, A. On clusterings: Good, bad and spectral. *Journal of the ACM (JACM)*, 51 (3):497–515, 2004.

Khosla, P., Teterwak, P., Wang, C., Sarna, A., Tian, Y., Isola, P., Maschinot, A., Liu, C., and Krishnan, D. Supervised

contrastive learning. *Advances in Neural Information Processing Systems*, 33:18661–18673, 2020.

Krizhevsky, A., Hinton, G., et al. Learning multiple layers of features from tiny images. 2009.

Kuhn, H. W. The hungarian method for the assignment problem. *Naval research logistics quarterly*, 2(1-2):83–97, 1955.

Lee, J. D., Lei, Q., Saunshi, N., and Zhuo, J. Predicting what you already know helps: Provable self-supervised learning. *Advances in Neural Information Processing Systems*, 34:309–323, 2021.

Lee, J. R., Gharan, S. O., and Trevisan, L. Multiway spectral partitioning and higher-order cheeger inequalities. *Journal of the ACM (JACM)*, 61(6):1–30, 2014.

Li, Z., Otholt, J., Dai, B., Meinel, C., Yang, H., et al. A closer look at novel class discovery from the labeled set. *arXiv preprint arXiv:2209.09120*, 2022.

McInnes, L., Healy, J., Saul, N., and Grossberger, L. Umap: Uniform manifold approximation and projection. *The Journal of Open Source Software*, 3(29):861, 2018.

McSherry, F. Spectral partitioning of random graphs. In *Proceedings 42nd IEEE Symposium on Foundations of Computer Science*, pp. 529–537. IEEE, 2001.

Ming, Y., Cai, Z., Gu, J., Sun, Y., Li, W., and Li, Y. Delving into out-of-distribution detection with vision-language representations. *Proceedings of the Advances in Neural Information Processing Systems*, 2022.

Ming, Y., Sun, Y., Dia, O., and Li, Y. How to exploit hyperspherical embeddings for out-of-distribution detection? In *The Eleventh International Conference on Learning Representations*, 2023.

Ng, A., Jordan, M., and Weiss, Y. On spectral clustering: Analysis and an algorithm. *Advances in neural information processing systems*, 14, 2001.

Shaham, U., Stanton, K., Li, H., Nadler, B., Basri, R., and Kluger, Y. Spectralnet: Spectral clustering using deep neural networks. *International Conference on Learning Representations*, 2018.

Shen, K., Jones, R. M., Kumar, A., Xie, S. M., HaoChen, J. Z., Ma, T., and Liang, P. Connect, not collapse: Explaining contrastive learning for unsupervised domain adaptation. In *International Conference on Machine Learning*, pp. 19847–19878. PMLR, 2022.

Shi, J. and Malik, J. Normalized cuts and image segmentation. *IEEE Transactions on pattern analysis and machine intelligence*, 22(8):888–905, 2000.

Shi, Z., Chen, J., Li, K., Raghuram, J., Wu, X., Liang, Y., and Jha, S. The trade-off between universality and label efficiency of representations from contrastive learning. In *International Conference on Learning Representations*, 2023.

Song, Y., Sohl-Dickstein, J., Kingma, D. P., Kumar, A., Ermon, S., and Poole, B. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations*, 2020.

Sun, Y. and Li, Y. Opencon: Open-world contrastive learning. *Transactions of Machine Learning Research*, 2023.

Sun, Y., Su, T., and Tu, Z. Faster r-cnn based autonomous navigation for vehicles in warehouse. In *2017 IEEE International Conference on Advanced Intelligent Mechatronics (AIM)*, pp. 1639–1644. IEEE, 2017.

Sun, Y., Ravi, S. N., and Singh, V. Adaptive activation thresholding: Dynamic routing type behavior for interpretability in convolutional neural networks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4938–4947, 2019.

Sun, Y., Guo, C., and Li, Y. React: Out-of-distribution detection with rectified activations. In *Proceedings of the Advances in Neural Information Processing Systems*, 2021.

Sun, Y., Ming, Y., Zhu, X., and Li, Y. Out-of-distribution detection with deep nearest neighbors. In *International Conference on Machine Learning*, pp. 20827–20840. PMLR, 2022.

Tosh, C., Krishnamurthy, A., and Hsu, D. Contrastive estimation reveals topic posterior information to linear models. *J. Mach. Learn. Res.*, 22:281–1, 2021a.

Tosh, C., Krishnamurthy, A., and Hsu, D. Contrastive learning, multi-view redundancy, and linear models. In *Algorithmic Learning Theory*, pp. 1179–1206. PMLR, 2021b.

Von Luxburg, U. A tutorial on spectral clustering. *Statistics and computing*, 17(4):395–416, 2007.

Wang, H., Xiao, R., Li, Y., Feng, L., Niu, G., Chen, G., and Zhao, J. Pico: Contrastive label disambiguation for partial label learning. *Proceedings of the International Conference on Learning Representations*, 2022.

Yang, M., Zhu, Y., Yu, J., Wu, A., and Deng, C. Divide and conquer: Compositional experts for generalized novel class discovery. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14268–14277, 2022.

Zhang, D., Nan, F., Wei, X., Li, S., Zhu, H., McKeown, K., Nallapati, R., Arnold, A., and Xiang, B. Supporting clustering with contrastive learning. *Proceedings of the North American Chapter of the Association for Computational Linguistics*, 2021.

Zhao, B. and Han, K. Novel visual category discovery with dual ranking statistics and mutual knowledge distillation. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2021.

Zhong, Z., Fini, E., Roy, S., Luo, Z., Ricci, E., and Sebe, N. Neighborhood contrastive learning for novel class discovery. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 10867–10875, 2021a.

Zhong, Z., Zhu, L., Luo, Z., Li, S., Yang, Y., and Sebe, N. Openmix: Reviving known knowledge for discovering novel visual categories in an open world. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9462–9470, 2021b.

Zhu, X., Ghahramani, Z., and Lafferty, J. D. Semi-supervised learning using gaussian fields and harmonic functions. In *Proceedings of the 20th International conference on Machine learning (ICML-03)*, pp. 912–919, 2003.

Appendix

A. Proof Details for Section 4

A.1. Bound Linear Probing Error by Regression Residual

Lemma A.1. *(Recap of Lemma 5.1) Denote by $\mathbf{y}(x) \in \mathbb{R}^{C_u}$ a one-hot vector, whose $y(x)$-th position is 1 and 0 elsewhere. Let $\mathbf{Y} \in \mathbb{R}^{N_u \times C_u}$ be a matrix whose rows are stacked by $\mathbf{y}(x)$. We have:*

$$\mathcal{R}(U^*) \triangleq \min_{\mathbf{M} \in \mathbb{R}^{k \times C_u}} \|\mathbf{Y} - U^* \mathbf{M}\|_F^2 \geq \frac{1}{2}\mathcal{E}(f)$$

Proof. Suppose $\tilde{f}(x) = \sqrt{w_x} f(x)$, we first show that

$$\|\mathbf{y}(x) - \tilde{f}(x)^\top \mathbf{M}\|^2 \geq \frac{1}{2}\mathbb{1}\left[y(x) \neq h(x; \tilde{f}, M)\right]$$

If $y(x) = h(x; \tilde{f}, M)$, it is clear that $\|\mathbf{y}(x) - \tilde{f}(x)^\top \mathbf{M}\|^2 \geq 0$. If $y(x) \neq h(x; \tilde{f}, M)$, then there exists another index $y' \neq y(x)$ so that $\tilde{f}(x)^\top \vec{\mu}_{y'} \geq \tilde{f}(x)^\top \vec{\mu}_{y(x)}$. Then,

$$\begin{aligned}
\|\mathbf{y}(x) - \tilde{f}(x)^\top \mathbf{M}\|_2^2 &\geq (1 - \tilde{f}(x)^\top \vec{\mu}_{y(x)})^2 + (\tilde{f}(x)^\top \vec{\mu}_{y'})^2 \\
&\geq \frac{1}{2}(1 - \tilde{f}(x)^\top \vec{\mu}_{y(x)} + \tilde{f}(x)^\top \vec{\mu}_{y'})^2 \\
&\geq \frac{1}{2},
\end{aligned}$$

where the first inequality is by only keeping y'-th and $y(x)$-th terms in the l_2 norm. We can then prove the lemma by:

$$\begin{aligned}
\mathcal{R}(U^*) &= \min_{\mathbf{M} \in \mathbb{R}^{k \times C_u}} \|\mathbf{Y} - U^* \mathbf{M}\|_F^2 \\
&= \min_{\mathbf{M} \in \mathbb{R}^{k \times C_u}} \sum_{x \in \mathcal{X}_u} \|\mathbf{y}(x) - \sqrt{w_x} f(x)^\top \Sigma_k^{-\frac{1}{2}} \mathbf{M}\|^2 \\
&= \min_{\mathbf{M} \in \mathbb{R}^{k \times C_u}} \sum_{x \in \mathcal{X}_u} \|\mathbf{y}(x) - \sqrt{w_x} f(x)^\top \mathbf{M}\|^2 \\
&\geq \frac{1}{2} \min_{\mathbf{M} \in \mathbb{R}^{k \times C_u}} \sum_{x \in \mathcal{X}_u} \mathbb{1}\left[y(x) \neq h(x; \tilde{f}, M)\right] \\
&= \frac{1}{2}\mathcal{E}(f),
\end{aligned}$$

where the second equation is given by $F^* \Sigma_k^{-\frac{1}{2}} = V_k$, and U^* is the last N_u rows of V_k, and the last equation is based on the fact that multiplying a scalar value on the output does not change the prediction result ($h(x; f, \mathbf{M}) = h(x; \tilde{f}, \mathbf{M})$). □

A.2. Spectral Contrastive Loss

Theorem A.2. *(Recap of Theorem 4.1) We define $\mathbf{f}_x = \sqrt{w_x} f(x)$ for some function f. Recall α, β is a hyper-parameter defined in Eq. (1). Then minimizing the loss function $\mathcal{L}_{\mathrm{mf}}(F, A)$ is equivalent to minimizing the following loss function for f, which we term **NCD Spectral Contrastive Loss (NSCL)**:*

$$\begin{aligned}
\mathcal{L}_{nscl}(f) \triangleq &-2\alpha\mathcal{L}_1(f) - 2\beta\mathcal{L}_2(f) \\
&+ \alpha^2\mathcal{L}_3(f) + 2\alpha\beta\mathcal{L}_4(f) + \beta^2\mathcal{L}_5(f),
\end{aligned} \tag{11}$$

where

$$\mathcal{L}_1(f) = \sum_{i\in\mathcal{Y}_l} \mathop{\mathbb{E}}_{\substack{\bar{x}_l\sim\mathcal{P}_{l_i},\bar{x}'_l\sim\mathcal{P}_{l_i},\\ x\sim\mathcal{T}(\cdot|\bar{x}_l),x^+\sim\mathcal{T}(\cdot|\bar{x}'_l)}} \left[f(x)^\top f\left(x^+\right)\right], \mathcal{L}_2(f) = \mathop{\mathbb{E}}_{\substack{\bar{x}_u\sim\mathcal{P}_u,\\ x\sim\mathcal{T}(\cdot|\bar{x}_u),x^+\sim\mathcal{T}(\cdot|\bar{x}_u)}} \left[f(x)^\top f\left(x^+\right)\right],$$

$$\mathcal{L}_3(f) = \sum_{i\in\mathcal{Y}_l}\sum_{j\in\mathcal{Y}_l} \mathop{\mathbb{E}}_{\substack{\bar{x}_l\sim\mathcal{P}_{l_i},\bar{x}'_l\sim\mathcal{P}_{l_j},\\ x\sim\mathcal{T}(\cdot|\bar{x}_l),x^-\sim\mathcal{T}(\cdot|\bar{x}'_l)}} \left[\left(f(x)^\top f\left(x^-\right)\right)^2\right], \mathcal{L}_4(f) = \sum_{i\in\mathcal{Y}_l} \mathop{\mathbb{E}}_{\substack{\bar{x}_l\sim\mathcal{P}_{l_i},\bar{x}_u\sim\mathcal{P}_u,\\ x\sim\mathcal{T}(\cdot|\bar{x}_l),x^-\sim\mathcal{T}(\cdot|\bar{x}_u)}} \left[\left(f(x)^\top f\left(x^-\right)\right)^2\right],$$

$$\mathcal{L}_5(f) = \mathop{\mathbb{E}}_{\substack{\bar{x}_u\sim\mathcal{P}_u,\bar{x}'_u\sim\mathcal{P}_u,\\ x\sim\mathcal{T}(\cdot|\bar{x}_u),x^-\sim\mathcal{T}(\cdot|\bar{x}'_u)}} \left[\left(f(x)^\top f\left(x^-\right)\right)^2\right].$$

Proof. We can expand $\mathcal{L}_{\mathrm{mf}}(F, A)$ and obtain

$$\mathcal{L}_{\mathrm{mf}}(F, A) = \sum_{x,x'\in\mathcal{X}} \left(\frac{w_{xx'}}{\sqrt{w_x w_{x'}}} - \mathbf{f}_x^\top \mathbf{f}_{x'}\right)^2 = \text{const} + \sum_{x,x'\in\mathcal{X}} \left(-2w_{xx'} f(x)^\top f\left(x'\right) + w_x w_{x'} \left(f(x)^\top f\left(x'\right)\right)^2\right),$$

where $\mathbf{f}_x = \sqrt{w_x} f(x)$ is a re-scaled version of $f(x)$. At a high level we follow the proof in (HaoChen et al., 2021), while the specific form of loss varies with the different definitions of positive/negative pairs. The form of $\mathcal{L}_{nscl}(f)$ is derived from plugging $w_{xx'}$ and w_x.

Recall that $w_{xx'}$ is defined by

$$w_{xx'} = \alpha \sum_{i\in\mathcal{Y}_l} \mathbb{E}_{\bar{x}_l\sim\mathcal{P}_{l_i}} \mathbb{E}_{\bar{x}'_l\sim\mathcal{P}_{l_i}} \mathcal{T}(x|\bar{x}_l)\mathcal{T}\left(x'|\bar{x}'_l\right) + \beta \mathbb{E}_{\bar{x}_u\sim\mathcal{P}_u} \mathcal{T}(x|\bar{x}_u)\mathcal{T}\left(x'|\bar{x}_u\right),$$

and w_x is given by

$$\begin{aligned} w_x &= \sum_{x'} w_{xx'} \\ &= \alpha \sum_{i\in\mathcal{Y}_l} \mathbb{E}_{\bar{x}_l\sim\mathcal{P}_{l_i}} \mathbb{E}_{\bar{x}'_l\sim\mathcal{P}_{l_i}} \mathcal{T}(x|\bar{x}_l) \sum_{x'} \mathcal{T}\left(x'|\bar{x}'_l\right) + \beta \mathbb{E}_{\bar{x}_u\sim\mathcal{P}_u} \mathcal{T}(x|\bar{x}_u) \sum_{x'} \mathcal{T}\left(x'|\bar{x}_u\right) \\ &= \alpha \sum_{i\in\mathcal{Y}_l} \mathbb{E}_{\bar{x}_l\sim\mathcal{P}_{l_i}} \mathcal{T}(x|\bar{x}_l) + \beta \mathbb{E}_{\bar{x}_u\sim\mathcal{P}_u} \mathcal{T}(x|\bar{x}_u). \end{aligned}$$

Plugging $w_{xx'}$ we have,

$$\begin{aligned} &-2 \sum_{x,x'\in\mathcal{X}} w_{xx'} f(x)^\top f\left(x'\right) = -2 \sum_{x,x^+\in\mathcal{X}} w_{xx^+} f(x)^\top f\left(x^+\right) \\ &= -2\alpha \sum_{i\in\mathcal{Y}_l} \mathbb{E}_{\bar{x}_l\sim\mathcal{P}_{l_i}} \mathbb{E}_{\bar{x}'_l\sim\mathcal{P}_{l_i}} \sum_{x,x'\in\mathcal{X}} \mathcal{T}(x|\bar{x}_l)\mathcal{T}\left(x'|\bar{x}'_l\right) f(x)^\top f\left(x'\right) - 2\beta \mathbb{E}_{\bar{x}_u\sim\mathcal{P}_u} \sum_{x,x'} \mathcal{T}(x|\bar{x}_u)\mathcal{T}\left(x'|\bar{x}_u\right) f(x)^\top f\left(x'\right) \\ &= -2\alpha \sum_{i\in\mathcal{Y}_l} \mathop{\mathbb{E}}_{\substack{\bar{x}_l\sim\mathcal{P}_{l_i},\bar{x}'_l\sim\mathcal{P}_{l_i},\\ x\sim\mathcal{T}(\cdot|\bar{x}_l),x^+\sim\mathcal{T}(\cdot|\bar{x}'_l)}} \left[f(x)^\top f\left(x^+\right)\right] - 2\beta \mathop{\mathbb{E}}_{\substack{\bar{x}_u\sim\mathcal{P}_u,\\ x\sim\mathcal{T}(\cdot|\bar{x}_u),x^+\sim\mathcal{T}(\cdot|\bar{x}_u)}} \left[f(x)^\top f\left(x^+\right)\right] = -2\alpha\mathcal{L}_1(f) - 2\beta\mathcal{L}_2(f) \end{aligned}$$

Plugging w_x and $w_{x'}$ we have,

$$
\begin{aligned}
&\sum_{x,x'\in\mathcal{X}} w_x w_{x'} \left(f(x)^\top f\left(x'\right)\right)^2 = \sum_{x,x^-\in\mathcal{X}} w_x w_{x^-} \left(f(x)^\top f\left(x^-\right)\right)^2 \\
&= \sum_{x,x'\in\mathcal{X}} \left(\alpha \sum_{i\in\mathcal{Y}_l} \mathbb{E}_{\bar{x}_l\sim\mathcal{P}_{l_i}} \mathcal{T}(x|\bar{x}_l) + \beta \mathbb{E}_{\bar{x}_u\sim\mathcal{P}_u} \mathcal{T}(x|\bar{x}_u)\right) \cdot \\
&\qquad\qquad \left(\alpha \sum_{j\in\mathcal{Y}_l} \mathbb{E}_{\bar{x}'_l\sim\mathcal{P}_{l_j}} \mathcal{T}(x^-|\bar{x}'_l) + \beta \mathbb{E}_{\bar{x}'_u\sim\mathcal{P}_u} \mathcal{T}(x^-|\bar{x}'_u)\right) \left(f(x)^\top f\left(x^-\right)\right)^2 \\
&= \alpha^2 \sum_{x,x^-\in\mathcal{X}} \sum_{i\in\mathcal{Y}_l} \mathbb{E}_{\bar{x}_l\sim\mathcal{P}_{l_i}} \mathcal{T}(x|\bar{x}_l) \sum_{j\in\mathcal{Y}_l} \mathbb{E}_{\bar{x}'_l\sim\mathcal{P}_{l_j}} \mathcal{T}(x^-|\bar{x}'_l) \left(f(x)^\top f\left(x^-\right)\right)^2 \\
&+ 2\alpha\beta \sum_{x,x^-\in\mathcal{X}} \sum_{i\in\mathcal{Y}_l} \mathbb{E}_{\bar{x}_l\sim\mathcal{P}_{l_i}} \mathcal{T}(x|\bar{x}_l) \mathbb{E}_{\bar{x}_u\sim\mathcal{P}_u} \mathcal{T}(x^-|\bar{x}_u) \left(f(x)^\top f\left(x^-\right)\right)^2 \\
&+ \beta^2 \sum_{x,x^-\in\mathcal{X}} \mathbb{E}_{\bar{x}_u\sim\mathcal{P}_u} \mathcal{T}(x|\bar{x}_u) \mathbb{E}_{\bar{x}'_u\sim\mathcal{P}_u} \mathcal{T}(x^-|\bar{x}'_u) \left(f(x)^\top f\left(x^-\right)\right)^2 \\
&= \alpha^2 \sum_{i\in\mathcal{Y}_l} \sum_{j\in\mathcal{Y}_l} \mathop{\mathbb{E}}_{\substack{\bar{x}_l\sim\mathcal{P}_{l_i},\bar{x}'_l\sim\mathcal{P}_{l_j},\\ x\sim\mathcal{T}(\cdot|\bar{x}_l),x^-\sim\mathcal{T}(\cdot|\bar{x}'_l)}} \left[\left(f(x)^\top f\left(x^-\right)\right)^2\right] + 2\alpha\beta \sum_{i\in\mathcal{Y}_l} \mathop{\mathbb{E}}_{\substack{\bar{x}_l\sim\mathcal{P}_{l_i},\bar{x}_u\sim\mathcal{P}_u,\\ x\sim\mathcal{T}(\cdot|\bar{x}_l),x^-\sim\mathcal{T}(\cdot|\bar{x}_u)}} \left[\left(f(x)^\top f\left(x^-\right)\right)^2\right] \\
&+ \beta^2 \mathop{\mathbb{E}}_{\substack{\bar{x}_u\sim\mathcal{P}_u,\bar{x}'_u\sim\mathcal{P}_u,\\ x\sim\mathcal{T}(\cdot|\bar{x}_u),x^-\sim\mathcal{T}(\cdot|\bar{x}'_u)}} \left[\left(f(x)^\top f\left(x^-\right)\right)^2\right] \\
&= \alpha^2 \mathcal{L}_3(f) + 2\alpha\beta \mathcal{L}_4(f) + \beta^2 \mathcal{L}_5(f).
\end{aligned}
$$

□

B. Proof for Eigenvalue in Toy Example

Before we present the proof of Theorem 5.2, Theorem5.3 and Lemma 5.4, we first present the following lemma B.1 which extensively explore the order and the form of eigenvectors of the general form $T(t)$. Note that T_1 and T_2 are special cases of the following $T(t)$ with $t \in [\tau_0, \tau_c]$:

$$T(t) = \begin{bmatrix} \tau_1 & t & t & \tau_0 & \tau_0 \\ t & \tau_1 & \tau_c & \tau_s & \tau_0 \\ t & \tau_c & \tau_1 & \tau_0 & \tau_s \\ \tau_0 & \tau_s & \tau_0 & \tau_1 & \tau_c \\ \tau_0 & \tau_0 & \tau_s & \tau_c & \tau_1 \end{bmatrix},$$

where t indicates the strength of the connection between labeled data and a novel class in unlabeled data.

Lemma B.1. *Assume $\tau_1 = 1$, $\tau_0 = 0$, $\tau_c < \tau_s < 1.5\tau_c$, $\bar{t} = \sqrt{\frac{2(\tau_s-\tau_c)^2\tau_c}{2\tau_c-\tau_s}}$, let $a(\lambda) = \frac{\lambda-1}{2t}$ and $b(\lambda) = \frac{\tau_s(\lambda-1)}{2(\lambda-1-\tau_c)t}$ are real value functions, the matrix $T(t)$'s eigenvectors (not necessarily l_2-normalized) and its eigenvalues are the following:*

(Case 1): If $t \in (\bar{t}, \tau_c]$,

$$\begin{array}{ll} v_1 = [1, a(\lambda_1), a(\lambda_1), b(\lambda_1), b(\lambda_1)]^\top, & \lambda_1 > 1 + \tau_s + \tau_c, \\ v_2 = [1, a(\lambda_2), a(\lambda_2), b(\lambda_2), b(\lambda_2)]^\top, & \lambda_2 \in [1 + \tau_s - \tau_c, 1 + \tau_c) \\ v_3 = [0, -1, 1, -1, 1]^\top, & \lambda_3 = 1 + \tau_s - \tau_c, \\ v_4 = [1, a(\lambda_4), a(\lambda_4), b(\lambda_4), b(\lambda_4)]^\top, & \lambda_4 \in (1 - \tau_s - \tau_c, 1) \\ v_5 = [0, 1, -1, -1, 1]^\top, & \lambda_5 = 1 - \tau_s - \tau_c, \end{array}$$

(Case 2): If $t \in (0, \bar{t})$,

$$\begin{array}{ll} v_1 = [1, a(\lambda_1), a(\lambda_1), b(\lambda_1), b(\lambda_1)]^\top, & \lambda_1 > 1 + \tau_s + \tau_c, \\ v_2 = [0, -1, 1, -1, 1]^\top, & \lambda_2 = 1 + \tau_s - \tau_c, \\ v_3 = [1, a(\lambda_3), a(\lambda_3), b(\lambda_3), b(\lambda_3)]^\top, & \lambda_3 \in [1, 1 + \tau_s - \tau_c) \\ v_4 = [1, a(\lambda_4), a(\lambda_4), b(\lambda_4), b(\lambda_4)]^\top, & \lambda_4 \in (1 - \tau_s - \tau_c, 1) \\ v_5 = [0, 1, -1, -1, 1]^\top, & \lambda_5 = 1 - \tau_s - \tau_c, \end{array}$$

(Case 3): If $t = 0$,

$$\begin{array}{ll} v_1 = [0, 1, 1, 1, 1]^\top, & \lambda_1 = 1 + \tau_s + \tau_c, \\ v_2 = [0, -1, 1, -1, 1]^\top, & \lambda_2 = 1 + \tau_s - \tau_c, \\ v_3 = [1, 0, 0, 0, 0]^\top, & \lambda_3 = 1 \\ v_4 = [0, 1, 1, -1, -1]^\top, & \lambda_4 = 1 - \tau_s + \tau_c \\ v_5 = [0, 1, -1, -1, 1]^\top, & \lambda_5 = 1 - \tau_s - \tau_c, \end{array}$$

Proof. For $t = 0$, Case 3, we can verify by direct calculation.

Now for Case 1 and Case 2, we consider $t \in (0, \tau_c)$. For any $i \in [5]$, denote $\hat{\lambda}_i$ as unordered eigenvalue and $\hat{v}_i$ is its corresponding eigenvector. We can direct verify that

$$\hat{\lambda}_1 = 1 + \tau_s - \tau_c \tag{12}$$

$$\hat{\lambda}_2 = 1 - \tau_s - \tau_c, \tag{13}$$

are two eigenvalues of $\tilde{A}_t$ and

$$\hat{v}_1 = [0, -1, 1, -1, 1]^\top \tag{14}$$

$$\hat{v}_2 = [0, 1, -1, -1, 1]^\top, \tag{15}$$

are two corresponding eigenvectors. Now, we prove for $i \in \{3,4,5\}$, $\hat{v}_i = [1, a(\hat{\lambda}_i), a(\hat{\lambda}_i), b(\hat{\lambda}_i), b(\hat{\lambda}_i)]^\top$ are eigenvector for $\hat{\lambda}_i$. For $i \in \{3,4,5\}$ we only need to show

$$\begin{cases} 1 + 2ta(\hat{\lambda}_i) & = \hat{\lambda}_i \\ t + (1+\tau_c)a(\hat{\lambda}_i) + \tau_s b(\hat{\lambda}_i) & = \hat{\lambda}_i a(\hat{\lambda}_i) \\ \tau_s a(\hat{\lambda}_i) + (1+\tau_c) b(\hat{\lambda}_i) & = \hat{\lambda}_i b(\hat{\lambda}_i). \end{cases} \tag{16}$$

Equivalently to

$$\begin{cases} 1 + 2ta(\hat{\lambda}_i) - \hat{\lambda}_i & = 0 \\ t + (1+\tau_c+\tau_s-\hat{\lambda}_i)(a(\hat{\lambda}_i)d + b(\hat{\lambda}_i)) & = 0 \\ t + (1+\tau_c-\tau_s-\hat{\lambda}_i)(a(\hat{\lambda}_i) - b(\hat{\lambda}_i)) & = 0. \end{cases} \tag{17}$$

Let $z_i = \hat{\lambda}_i - 1$. Equivalently to

$$\begin{cases} 1 + 2ta(\hat{\lambda}_i) - \hat{\lambda}_i & = 0 \\ (\hat{\lambda}_i - 1 - \tau_c) b(\hat{\lambda}_i) - \tau_s a(\hat{\lambda}_i) & = 0 \\ z_i^3 - 2\tau_c z_i^2 + (\tau_c^2 - \tau_s^2 - 2t^2) z_i + 2\tau_c t^2 & = 0. \end{cases} \tag{18}$$

Let $g(z) = z^3 - 2\tau_c z^2 + (\tau_c^2 - \tau_s^2 - 2t^2)z + 2\tau_c t^2$, we can verify that $g(-\infty) < 0,\ \ g(-\tau_c - \tau_s) = -4\tau_c(\tau_c+\tau_s)^2 + 4t^2\tau_c + 2t^2\tau_s < 0,\ \ g(0) = 2\tau_c t^2 > 0,\ \ g(\tau_c) = -\tau_s^2\tau_c < 0,\ \ g(\tau_c+\tau_s) = -2\tau_s t^2 < 0,\ \ g(+\infty) > 0$. Thus, we have three solutions and satisfying $1 - \tau_c - \tau_s < \hat{\lambda}_5 < 1 < \hat{\lambda}_4 < 1 + \tau_c < 1 + \tau_c + \tau_s < \hat{\lambda}_3$. As $\hat{\lambda}_i \neq 1 + \tau_c$ for $i \in \{3,4,5\}$, thus, equivalently to

$$\begin{cases} a(\hat{\lambda}_i) & = \frac{\hat{\lambda}_i - 1}{2t} \\ b(\hat{\lambda}_i) & = \frac{\tau_s(\hat{\lambda}_i - 1)}{2(\hat{\lambda}_i - 1 - \tau_c)t} \\ (\hat{\lambda}_i - 1)^3 - 2\tau_c(\hat{\lambda}_i - 1)^2 + (\tau_c^2 - \tau_s^2 - 2t^2)(\hat{\lambda}_i - 1) + 2\tau_c t^2 & = 0. \end{cases} \tag{19}$$

When $t > \bar{t}$, we have $g(\tau_s - \tau_c) > 0$. Thus, we have $1 - \tau_c - \tau_s < \hat{\lambda}_5 < 1 + \tau_s - \tau_c < \hat{\lambda}_4 < 1 + \tau_c + \tau_s < \hat{\lambda}_3$. By reorder, we finish Case 1.

When $t < \bar{t}$, we have $g(\tau_s - \tau_c) < 0$. Thus, we have $1 - \tau_c - \tau_s < \hat{\lambda}_5 < 1 < \hat{\lambda}_4 < 1 + \tau_s - \tau_c < 1 + \tau_c + \tau_s < \hat{\lambda}_3$. By reorder the eigenvectors w.r.t the size of eigenvalues, we finish Case 2. □

Theorem B.2. *(Recap of Theorem 5.2) Assume $\tau_1 = 1$, $\tau_0 = 0$, $\tau_s < 1.5\tau_c$. We have*

$$U_1^* = \begin{bmatrix} a_1 & a_1 & b_1 & b_1 \\ a_2 & a_2 & b_2 & b_2 \end{bmatrix}^\top,$$

where a_1, b_1 are some positive real numbers, and a_2, b_2 has different signs.

$$U_2^* = \begin{cases} \frac{1}{2}\begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 & -1 & -1 \end{bmatrix}^\top, & \text{if } \tau_s < \tau_c, \\ \frac{1}{2}\begin{bmatrix} 1 & 1 & 1 & 1 \\ -1 & 1 & -1 & 1 \end{bmatrix}^\top, & \text{if } \tau_s > \tau_c, \end{cases}$$

With label vector $\vec{y} = \{1,1,0,0\}$, we have

$$\mathcal{R}(U_1^*, \vec{y}) = 0, \mathcal{R}(U_2^*, \vec{y}) = \begin{cases} 0, & \text{if } \tau_s < \tau_c \\ 1, & \text{if } \tau_s > \tau_c. \end{cases} \tag{20}$$

Proof. In the Case 1 and Case 3 of Lemma B.1, we have shown the U_1^* and U_2^* case when $\tau_s > \tau_c$ respectively. In this proof, we just need to show the case when $\tau_s < \tau_c$. For U_2^* and $\tau_s < \tau_c$, since $t = 0$, we can directly prove by giving the eigenvectors with order:

$$\begin{aligned}
v_1 &= [0,1,1,1,1]^\top, & \lambda_1 &= 1+\tau_s+\tau_c,\\
v_2 &= [0,1,1,-1,-1]^\top, & \lambda_2 &= 1-\tau_s+\tau_c\\
v_3 &= [1,0,0,0,0]^\top, & \lambda_3 &= 1\\
v_4 &= [0,-1,1,-1,1]^\top, & \lambda_4 &= 1+\tau_s-\tau_c,\\
v_5 &= [0,1,-1,-1,1]^\top, & \lambda_5 &= 1-\tau_s-\tau_c,
\end{aligned}$$

For U_1^*, one can see that in the Case 1 of Lemma B.1, we still have $\lambda_2 > \lambda_3$ since $\tau_s < 1.5\tau_c < 2\tau_c$ holds. Therefore the order of v_2 and v_3 does not change. Then U_1^* is the concatenation of the last four dimensions of v_2 and v_1.

Now we would like to show that a_1, b_1 are positive and a_2, b_2 have different signs. We have shown in Lemma B.1 that $a(\lambda) = \frac{\lambda-1}{2t}$ and $b(\lambda) = \frac{\tau_s(\lambda-1)}{2(\lambda-1-\tau_c)t}$. Since $a_1 = a(\lambda_1)$ and $b_1 = b(\lambda_1)$, one can show that $a_1 > 0, b_1 > 0$ since $\lambda_1 > 1+\tau_s+\tau_c$. For $\lambda_2 \in [1+\tau_s-\tau_c, 1+\tau_c)$, it is clear that $a_2 = a(\lambda_2) > 0 > b(\lambda_2) = b_2$ when $\tau_s > \tau_c$, and conversely we have $a_2 = a(\lambda_2) < 0 < b(\lambda_2) = b_2$ when $\tau_s < \tau_c$. So a_2 and b_2 have different signs in both cases.

Recall $\mathcal{R}(U^*, \vec{y})$ is defined as:

$$\mathcal{R}(U^*, \vec{y}) = \min_{\vec{\mu}\in\mathbb{R}^k} \|\vec{y} - U^*\vec{\mu}\|_2^2,$$

Let $\vec{\mu} = [\frac{b_2}{a_1b_2-a_2b_1}, \frac{-b_1}{a_1b_2-a_2b_1}]^\top$, $\mathcal{R}(U_1^*, \vec{y}) = 0$. If $\tau_s < \tau_c$, let $\vec{\mu} = [1,1]^\top$, then $\mathcal{R}(U_2^*, \vec{y}) = 0$. If $\tau_s > \tau_c$, $\vec{\mu}^* = U_2^{*\top}\vec{y} = [1,0]^\top$ is the minimizer and we have $\mathcal{R}(U_2^*, \vec{y}) = 1$.

□

Theorem B.3. *(Recap of Theorem 5.3) Assume $\tau_1 = 1$, $\tau_0 = 0$, $1.5\tau_c > \tau_s > \tau_c$. Let $\bar{t} = \sqrt{\frac{2(\tau_s-\tau_c)^2\tau_c}{2\tau_c-\tau_s}}$, $r : \mathbb{R} \mapsto (0,1)$ as a real value function, we have*

$$\mathcal{R}(U_t^*, \vec{y}) = \begin{cases} 0, & \text{if } t \in (\bar{t}, \tau_s),\\ r(t), & \text{if } t \in (0, \bar{t})\\ 1, & \text{if } t = 0. \end{cases} \tag{21}$$

Proof. According to Lemma B.1, if $t \in (\bar{t}, \tau_s)$,

$$U_t^* = \begin{bmatrix} a_1 & a_1 & b_1 & b_1\\ a_2 & a_2 & b_2 & b_2 \end{bmatrix}^\top,$$

where a_1, b_1 are some positive real numbers, and a_2, b_2 has different signs. Let $\vec{\mu} = [\frac{b_2}{a_1b_2-a_2b_1}, \frac{-b_1}{a_1b_2-a_2b_1}]^\top$, $\mathcal{R}(U_t^*, \vec{y}) = 0$. If $t = 0$, $\mathcal{R}(U_t^*, \vec{y}) = 0$, which is proved in Theorem B.2 when $\tau_s > \tau_c$. If $t \in (0, \bar{t})$, as shown in Lemma B.1, we have

$$U_t^* = \begin{bmatrix} \frac{\lambda_1-1}{2t} & \frac{\lambda_1-1}{2t} & \frac{\tau_s(\lambda_1-1)}{2(\lambda_1-1-\tau_c)t} & \frac{\tau_s(\lambda_1-1)}{2(\lambda_1-1-\tau_c)t}\\ -1 & 1 & -1 & 1 \end{bmatrix}^\top,$$

where $\lambda_1 > 0$. $\vec{\mu}_* = (U_t^{*\top}U_t^*)^\dagger U_t^{*\top}\vec{y} = [\frac{\frac{\lambda_1-1}{2t}}{(\frac{\lambda_1-1}{2t})^2+(\frac{\tau_s(\lambda_1-1)}{2(\lambda_1-1-\tau_c)t})^2}, 0]^\top$, then:

$$\mathcal{R}(U_t^*, \vec{y}) = \frac{2\tau_s^2}{(\lambda_1-1-\tau_c)^2+\tau_s^2} = r(\lambda_1) \in (0,1).$$

Note that λ_1 is a value dependent on t, therefore $r(\lambda_1)$ can be represented as $r(t)$.

□

Lemma B.4. *(Recap of Lemma 5.4) If $\tau_s < \tau_c < 1.5\tau_s$, $\mathcal{R}(U_3^*, \vec{y}) = 1, \mathcal{R}(U_2^*, \vec{y}) = 0$.*

Proof. When $\mathcal{X}_l^{\text{case 3}} \triangleq \{X_{\square,c_3}\}$(gray cube), we have

$$T_3 = \begin{bmatrix} \tau_1 & \tau_s & \tau_0 & \tau_s & \tau_0 \\ \tau_s & \tau_1 & \tau_c & \tau_s & \tau_0 \\ \tau_0 & \tau_c & \tau_1 & \tau_0 & \tau_s \\ \tau_s & \tau_s & \tau_0 & \tau_1 & \tau_c \\ \tau_0 & \tau_0 & \tau_s & \tau_c & \tau_1 \end{bmatrix},$$

Follow the same proof in Lemma B.1, one can show that

$$U_3^* = \begin{bmatrix} a_1 & b_1 & a_1 & b_1 \\ a_2 & b_2 & a_2 & b_2 \end{bmatrix}^\top,$$

where a_1, b_1 are some positive real numbers, and a_2, b_2 has different signs. Note that U_3^* forms the same linear span as

$$\frac{1}{2}\begin{bmatrix} 1 & 1 & 1 & 1 \\ -1 & 1 & -1 & 1 \end{bmatrix}^\top.$$

Therefore, we have $\mathcal{R}(U_3^*, \vec{y}) = 1$ as proved in Theorem B.2. □

C. Additional Details for Section 5.3

This section acts as an expanded version of Section 5.3. We will first show in Section C.1 with the background and proof for Theorem 5.5 with the original adjacency matrix $\dot{A}$. Then we present the analysis based on the approximation matrix $\bar{A}$ in Section C.2. Finally, we show the formal proof of our main Theorem 5.6 in Section C.3. The proof of Theorem 5.6 requires two important ingredients (Lemma C.6 and Lemma C.10) with proof deferred in Section C.4 and Section C.5 respectively.

C.1. Sufficient and Necessary Condition for Perfect Residual

We first present the formal analysis in Theorem C.1 which is an extended version of Theorem 5.5 without approximation and we start with the recap of definitions.

Notations. Recall that $V^* \in \mathbb{R}^{N\times k}$ is defined as the top-k singular vectors of $\dot{A}$ and we split the eigen-matrix into two parts for labeled and unlabeled samples respectively:

$$V^* = \begin{bmatrix} L^* \in \mathbb{R}^{N_l\times k} \\ U^* \in \mathbb{R}^{N_u\times k} \end{bmatrix} = \begin{bmatrix} l_1 & l_2 & \cdots & l_k \\ u_1 & u_2 & \cdots & u_k \end{bmatrix}$$

for labeled and unlabeled samples respectively. Then we let $V^\flat \in \mathbb{R}^{N\times(N-k)}$ be the remaining singular vectors of $\dot{A}$ except top-k. Similarly, we split $V^\flat$ into two parts:

$$V^\flat = \begin{bmatrix} L^\flat \in \mathbb{R}^{N_l\times(N-k)} \\ U^\flat \in \mathbb{R}^{N_u\times(N-k)} \end{bmatrix} = \begin{bmatrix} l_{k+1} & l_{k+2} & \cdots & l_N \\ u_{k+1} & u_{k+2} & \cdots & u_N \end{bmatrix}.$$

We can also split the matrix $\dot{A}$ at the N_l-th row and the N_l-th column and we obtain $A_{ll} \in \mathbb{R}^{N_l\times N_l}, A_{ul} \in \mathbb{R}^{N_u\times N_l}, A_{uu} \in \mathbb{R}^{N_u\times N_u}$ with

$$\dot{A} = \begin{bmatrix} A_{ll} & A_{ul}^\top \\ A_{ul} & A_{uu} \end{bmatrix}.$$

Theorem C.1. ***(No approximation)*** *Denote the projection matrix $\mathsf{P}_{L^\flat} = L^{\flat\top}(L^\flat L^{\flat\top})^\dagger L^\flat$, where † denotes the Moore-Penrose inverse. For any labeling vector $\vec{y} \in \{0,1\}^{N_u}$, we have*

$$\mathcal{R}(U^*, \vec{y}) \le \|(I - \mathsf{P}_{L^\flat})U^{\flat\top}\vec{y}\|_2^2. \tag{22}$$

The sufficient and necessary condition for $\mathcal{R}(U^*, \vec{y}) = 0$ is $\vec{\omega} \in \mathbb{R}^{N_l}$ such that

$$\forall i = k+1, \ldots, N, \langle \vec{y}^\top(\sigma_i I - A_{uu})^\dagger A_{ul}, l_i\rangle = \langle \vec{\omega}, l_i\rangle \tag{23}$$

where σ_i is the i-th largest eigenvalue of $\dot{A}$.

Proof. Define $\vec{y}' = [\vec{\zeta}^\top, \vec{y}^\top]^\top$ as an extended labeling vector, where $\vec{\zeta} \in \mathbb{R}^{N_l}$ can be a "placeholder" vector with any values. We have

$$\begin{aligned}
\mathcal{R}\left(U^*, \vec{y}\right) &= \min_{\vec{\mu}\in\mathbb{R}^k} \|\vec{y} - U^*\vec{\mu}\|_2^2 \\
&= \min_{\vec{\mu}\in\mathbb{R}^k, \vec{\zeta}\in\mathbb{R}^{N_l}} \|\vec{y}' - V^*\vec{\mu}\|_2^2 \\
&= \min_{\vec{\zeta}\in\mathbb{R}^{N_l}} \|\vec{y}' - V^*V^{*\top}\vec{y}'\|_2^2 \\
&= \min_{\vec{\zeta}\in\mathbb{R}^{N_l}} \|V^{\flat\top}\vec{y}'\|_2^2 \\
&= \min_{\vec{\zeta}\in\mathbb{R}^{N_l}} \|L^{\flat\top}\vec{\zeta} + U^{\flat\top}\vec{y}\|_2^2 \\
&= \|(I - L^{\flat\top}(L^\flat L^{\flat\top})^\dagger L^\flat)U^{\flat\top}\vec{y}\|_2^2.
\end{aligned}$$

The sufficient and necessary condition for $\mathcal{R}(U^*, \vec{y}) = 0$ is:

$$\exists \vec{\omega} \in \mathbb{R}^{N_l}, \forall i = k+1, \ldots, N, u_i^\top \vec{y} = l_i^\top \vec{\omega}.$$

We then look into the relationship between l_i and u_i. Since

$$\begin{bmatrix} A_{ll} & A_{ul}^\top \\ A_{ul} & A_{uu} \end{bmatrix} \begin{bmatrix} l_i \\ u_i \end{bmatrix} = \sigma_i \begin{bmatrix} l_i \\ u_i \end{bmatrix},$$

we have the following results:

$$u_i = (\sigma_i I - A_{uu})^\dagger A_{ul} l_i.$$

So the sufficient and necessary condition becomes: there exists $\vec{\omega} \in \mathbb{R}^{N_l}$ such that

$$\forall i = k+1, \ldots, N, \langle \vec{y}^\top (\sigma_i I - A_{uu})^\dagger A_{ul}, l_i \rangle = \langle \vec{\omega}, l_i \rangle, \tag{24}$$

where σ_i is the i-th largest singular value of $\dot{A}$. □

Interpretation of Theorem C.1. The bound of residual in Ineq. (10) composed of two projections: $U^{\flat\top}$ and $(I - \mathsf{P}_{L^\flat})$. If we only consider $\|U^{\flat\top}\vec{y}\|_2^2$, it is equivalent to $\vec{y}^\top(I - U^*U^{*\top})\vec{y}$ which indicates the information in $\vec{y}$ that is not covered by the learned representation U^*. Then multiplying the second projection matrix $(I - \mathsf{P}_{L^\flat})$ further reduces the residual by considering the information from labeled data, since $\mathsf{P}_{L^\flat}$ is a projection matrix that projects a vector to the linear span of $L^\flat$. In the extreme case, when $U^{\flat\top}\vec{y}$ fully lies in the linear span of $L^\flat$, the residual $\mathcal{R}(U^*, \vec{y})$ becomes 0. To provide further insights about Eq. (23), we analyze in a simplified setting by approximating $\dot{A}$ in the next section.

C.2. Analysis with Approximation

In Theorem C.1, we put an analysis on how $L^\flat$ can influence the residual function. However, $L^\flat$ is a matrix with N_l rows, so it is hard to quantitatively understand the effect of N_l labeled samples individually. We resort to viewing the labeled samples as a whole. Our idea is motivated by the Stochastic Block Model (SBM) (Holland et al., 1983) model, which analyzes the probability between different communities instead of individual values. In our case, we aim to analyze the probability vector $\eta_u \in \mathbb{R}^{N_u}$ denoting the chance of each unlabeled data point having the same augmentation view as one of the samples from the known class. The relationship between η_u and A_{uu} is then of our interest. Specifically, we define $\bar{A}$ with values at (i, j) be the following:

$$\bar{A}_{x_i x_j} = \begin{cases} \dot{A}_{x_i x_j} & \text{if } x_i \in \mathcal{X}_u, x_j \in \mathcal{X}_u, \\ \mathbb{E}_{x'\in\mathcal{X}_l} \dot{A}_{x_i x'} & \text{if } x_i \in \mathcal{X}_u, x_j \in \mathcal{X}_l, \\ \mathbb{E}_{x'\in\mathcal{X}_l} \dot{A}_{x' x_j} & \text{if } x_i \in \mathcal{X}_l, x_j \in \mathcal{X}_u, \\ \mathbb{E}_{x',x''\in\mathcal{X}_l} \dot{A}_{x' x''} & \text{if } x_i \in \mathcal{X}_l, x_j \in \mathcal{X}_l. \end{cases} \tag{25}$$

The probability is estimated by taking the average. It is equivalent to multiplying matrix P and $P^\top$ on left and right side, where $P \in \mathbb{R}^{N\times N}$ is given by:

$$P = \begin{bmatrix} \frac{1}{N_l}\mathbf{1}_{N_l\times N_l} & \mathbf{0}_{N_l\times N_u} \\ \mathbf{0}_{N_u\times N_l} & I_{N_u} \end{bmatrix},$$

where $\mathbf{1}_{n\times m}$ and $\mathbf{0}_{n\times m}$ represent matrix filled with 1 and 0 respectively with shape $n \times m$. Then we can write $\bar{A} \in \mathbb{R}^{N\times N}$, the approximated version of A, as follows:

$$\bar{A} = PAP^\top = \begin{bmatrix} \eta_l \mathbf{1}_{N_l\times N_l} & \mathbf{1}_{N_l\times 1}\vec{\eta}_u^\top \\ \vec{\eta}_u \mathbf{1}_{1\times N_l} & A_{uu}, \end{bmatrix},$$

where $\eta_l \in \mathbb{R}$ and $\vec{\eta}_u \in \mathbb{R}^{N_u\times 1}$. Our analysis can then focus on how η_u influences the representation space learned by A_{uu}. Similar to Section C.1, we define the top-k and the remainder singular vectors with corresponding splits as :

$$\bar{V}^* = \begin{bmatrix} \bar{L}^* \\ \bar{U}^* \end{bmatrix} = \begin{bmatrix} \bar{l}_1 & \bar{l}_2 & \cdots & \bar{l}_k \\ \bar{u}_1 & \bar{u}_2 & \cdots & \bar{u}_k \end{bmatrix},$$

$$\bar{V}^\flat = \begin{bmatrix} \bar{L}^\flat \\ \bar{U}^\flat \end{bmatrix} = \begin{bmatrix} \bar{l}_{k+1} & \bar{l}_{k+2} & \cdots & \bar{l}_N \\ \bar{u}_{k+1} & \bar{u}_{k+2} & \cdots & \bar{u}_N \end{bmatrix}.$$

Note that due to the special structure of $\bar{A}$ with N_l duplicated rows and columns, the eigenvector $\bar{V}$ has a special structure as we demonstrate in the next Lemma C.2. We defer the proof to Section C.2.1.

Lemma C.2. *Since A_{uu} is symmetric and has large diagonal values, we assume A_{uu} is a positive semi-definite matrix. $\bar{L}^*$ is stacked by the same row such that $\bar{L}^* = \frac{\mathbf{1}_{N_l\times 1}}{N_l}\bar{\mathfrak{l}}^{*\top}$, where $\bar{\mathfrak{l}}^* \in \mathbb{R}^k$ and that $\bar{L}^\flat$ has the following form:*

$$\bar{L}^\flat = \begin{bmatrix} \frac{\mathbf{1}_{N_l\times 1}}{N_l}\bar{\mathfrak{l}}'^\top & \bar{l}_{N-\Theta+1} & \dots & \bar{l}_N \end{bmatrix},$$

where Θ is the rank of the null space for $A_{uu} - \frac{\eta_u\eta_u^\top}{\eta_l}$, $\bar{\mathfrak{l}}' \in \mathcal{R}^{N-k-\Theta}$ with non-zero values, and $\bar{l}_{N-\Theta+1}, ..., \bar{l}_N$ are all perpendicular to $\mathbf{1}_{N_l}$.

By property in Lemma C.2, we define:

$$\bar{\mathfrak{l}}^\flat \triangleq \bar{L}^{\flat\top}\mathbf{1}_{N_l\times 1} = \begin{bmatrix} \bar{\mathfrak{l}}'^\top & 0 & \dots & 0 \end{bmatrix}^\top \in \mathbb{R}^{N-k}. \tag{26}$$

Definition C.3. To ease the notation, we let $\mathcal{I} \triangleq \{k+1, k+2, ..., N-\Theta\}$ and we mainly discuss $i \in \mathcal{I}$.

These definitions facilitate the presentation of the following Theorem C.4.

Theorem C.4. ***(With approximation)*** *Denote $\mathfrak{T}(\vec{y}) = \frac{\|\bar{U}^{\flat\top}\vec{y}\|_2}{\|\vec{y}\|_2}$ and $\kappa(\vec{y}) = \cos(\bar{U}^{\flat\top}\vec{y}, \bar{\mathfrak{l}}^\flat)$, where $\cos$ measures the cosine distance between two vectors. Let σ_i as the i-th largest eigenvalue of $\dot{A}$ and $\bar{\sigma}_i$ is for $\bar{A}$. For a labeling vector $\vec{y} \in \{0,1\}^{N_u}$, we have*

$$\mathcal{R}(\bar{U}^*, \vec{y}) = \frac{N_u}{|\mathcal{Y}_u|}(1-\kappa(\vec{y})^2)\mathfrak{T}(\vec{y})^2. \tag{27}$$

If the ignorance degree $\mathfrak{T}(\vec{y})$ is non-zero, the sufficient and necessary condition for $\mathcal{R}(\bar{U}^*, \vec{y}) = 0$: there exists $\omega \in \mathbb{R}$ such that

$$\forall i \in \mathcal{I}, \vec{y}^\top(\bar{\sigma}_i I - A_{uu})^\dagger \vec{\eta}_u = \omega. \tag{28}$$

Proof. Define $\vec{y}' = [\zeta \mathbf{1}_{1\times N_l}, \vec{y}^\top]^\top$ as an extended labeling vector where ζ is any real number. We have

$$\begin{aligned}
\mathcal{R}\left(\bar{U}^*, \vec{y}\right) &= \min_{\vec{\mu}\in\mathbb{R}^k} \|\vec{y} - \bar{U}^*\vec{\mu}\|_2^2 \\
&= \min_{\vec{\mu}\in\mathbb{R}^k, \zeta\in\mathbb{R}} \{\|\vec{y} - \bar{U}^*\vec{\mu}\|_2^2 + \|(\zeta - \bar{\mathfrak{l}}^{*\top}\vec{\mu})\mathbf{1}_{1\times N_l}\|_2^2\} \\
&= \min_{\vec{\mu}\in\mathbb{R}^k, \zeta\in\mathbb{R}} \|\vec{y}' - \bar{V}^*\vec{\mu}\|_2^2 \\
&= \min_{\zeta\in\mathbb{R}} \|\vec{y}' - \bar{V}^*\bar{V}^{*\top}\vec{y}'\|_2^2 \\
&= \min_{\zeta\in\mathbb{R}} \|\bar{V}^{\flat\top}\vec{y}'\|_2^2 \\
&= \min_{\zeta\in\mathbb{R}} \|\zeta\bar{L}^{\flat\top}\mathbf{1}_{N_l\times 1} + \bar{U}^{\flat\top}\vec{y}\|_2^2 \\
&= \min_{\zeta\in\mathbb{R}} \|\zeta\bar{\mathfrak{l}}^\flat + \bar{U}^{\flat\top}\vec{y}\|_2^2 \\
&= \|(I - \frac{\bar{\mathfrak{l}}^\flat\bar{\mathfrak{l}}^{\flat\top}}{\|\bar{\mathfrak{l}}^\flat\|_2^2})\bar{U}^{\flat\top}\vec{y}\|_2^2 \\
&= (1-\kappa(\vec{y})^2)\|\bar{U}^{\flat\top}\vec{y}\|_2^2 \\
&= \frac{N_u}{|\mathcal{Y}_u|}(1-\kappa(\vec{y})^2)\mathfrak{T}(\vec{y})^2.
\end{aligned}$$

We then look into the components of $\bar{\mathfrak{l}}^\flat$ and $\bar{U}^\flat$. According to Lemma C.2, when $i > N - \Theta$, we have:

$$\bar{\mathfrak{l}}^\flat = \begin{bmatrix} \bar{\mathfrak{l}}'^\top & 0 & \dots & 0 \end{bmatrix}^\top = \begin{bmatrix} (\bar{\mathfrak{l}}^\flat)_{k+1} & (\bar{\mathfrak{l}}^\flat)_{k+2} & \cdots & (\bar{\mathfrak{l}}^\flat)_{N-\Theta} & 0\cdots & 0 \end{bmatrix}. \tag{29}$$

And the sufficient and necessary condition for $\mathcal{R}(\bar{U}^*, \vec{y})$ to be minimized by $\bar{\mathfrak{l}}^\flat$ is:

$$\exists\omega\in\mathbb{R}, \forall i\in\mathcal{I}, \bar{u}_i^{\flat\top}\vec{y} = \omega(\bar{\mathfrak{l}}^\flat)_i. \tag{30}$$

Note that for $i \in \mathcal{I}$,

$$\begin{bmatrix} \eta_l\mathbf{1}_{N_l\times N_l} & \mathbf{1}_{N_l\times 1}\vec{\eta}_u^\top \\ \vec{\eta}_u\mathbf{1}_{1\times N_l} & A_{uu} \end{bmatrix} \begin{bmatrix} \bar{l}_i \\ \bar{u}_i \end{bmatrix} = \bar{\sigma}_i \begin{bmatrix} \bar{l}_i \\ \bar{u}_i \end{bmatrix}.$$

Also since $(\bar{\mathfrak{l}}^\flat)_i = \mathbf{1}_{1\times N_l}\bar{l}_i \in \mathbb{R}$, we have the following results:

$$\bar{u}_i = (\bar{\sigma}_i I - A_{uu})^\dagger\vec{\eta}_u(\bar{\mathfrak{l}}^\flat)_i.$$

Thus, the sufficient and necessary condition (30) becomes: there exists $\omega \in \mathbb{R}$ such that

$$\forall i\in\mathcal{I}, \vec{y}^\top(\bar{\sigma}_i I - A_{uu})^\dagger\vec{\eta}_u = \omega. \tag{31}$$

□

C.2.1. Proof of Lemma C.2

Proof. To understand the structure of $\bar{U}$ and $\bar{L}$, we consider the eigenvalue problem:

$$\begin{bmatrix} \eta_l\mathbf{1}_{N_l\times N_l} & \mathbf{1}_{N_l\times 1}\vec{\eta}_u^\top \\ \vec{\eta}_u\mathbf{1}_{1\times N_l} & A_{uu} \end{bmatrix} \begin{bmatrix} \bar{l}_i \\ \bar{u}_i \end{bmatrix} = \bar{\sigma}_i \begin{bmatrix} \bar{l}_i \\ \bar{u}_i \end{bmatrix}.$$

In the non-trivial case, $\eta_l \neq 0, \vec{\eta}_u \neq \mathbf{0}_{N_l}$, we have the following two equations:

$$\begin{aligned}
\eta_l\mathbf{1}_{N_l\times 1}\mathbf{1}_{1\times N_l}\bar{l}_i + \mathbf{1}_{N_l\times 1}\vec{\eta}_u^\top\bar{u}_i &= \bar{\sigma}_i\bar{l}_i \\
(\bar{\sigma}_i I - A_{uu})\bar{u}_i &= \vec{\eta}_u\mathbf{1}_{1\times N_l}\bar{l}_i.
\end{aligned}$$

(Case 1) When $\bar{\sigma}_i \neq 0$, then $\bar{l}_i$ has N_l duplicated scalar values $\frac{\vec{\eta}_u^\top \bar{u}_i}{\bar{\sigma}_i - N_l \eta_l}$ for the first equation to satisfy.

(Case 2) When $\bar{\sigma}_i = 0$, then by combing the two equations, we have:

$$A_{uu}\bar{u}_i = \frac{\vec{\eta}_u \vec{\eta}_u^\top}{\eta_l}\bar{u}_i.$$

If $A_{uu} - \frac{\vec{\eta}_u \vec{\eta}_u^\top}{\eta_l}$ is a full rank matrix, then $\bar{u}_i = \mathbf{0}_{N_u}$, and by the first equation $\mathbf{1}_{1\times N_l}\bar{l}_i = 0$. If $A_{uu} - \frac{\vec{\eta}_u \vec{\eta}_u^\top}{\eta_l}$ is a deficiency matrix and $\text{rank}(A_{uu} - \frac{\vec{\eta}_u \vec{\eta}_u^\top}{\eta_l}) \geq \text{rank}(A_{uu})$[3], then $\bar{u}_i$ lies in the null space formed by $\vec{\eta}_u$ and A_{uu} jointly, then $\vec{\eta}_u^\top \bar{u}_i = 0$, we still have $\mathbf{1}_{1\times N_l}\bar{l}_i = 0$.

Therefore when $i \in \{1, \ldots, k\}$, $\bar{\sigma}$ is non-zero values, so that $\bar{L}^*$ is stacked by the same row such that $\bar{L}^* = \frac{\mathbf{1}_{N_l\times 1}}{N_l}\bar{\mathfrak{l}}^{*\top}$, where $\bar{\mathfrak{l}}^* \in \mathbb{R}^k$. For $i \in \{k+1, \ldots, N\}$, $\bar{L}^\flat$ has the following form:

$$\bar{L}^\flat = \begin{bmatrix} \frac{\mathbf{1}_{N_l\times 1}}{N_l}\bar{\mathfrak{l}}'^\top & \bar{l}_{N-\Theta+1} & \ldots & \bar{l}_N \end{bmatrix},$$

where Θ is the rank of the null space for $A_{uu} - \frac{\eta_u \eta_u^\top}{\eta_l}$, $\bar{\mathfrak{l}}' \in \mathcal{R}^{N-k-\Theta}$, and $\bar{l}_{N-\Theta+1}, \ldots, \bar{l}_N$ are all perpendicular to $\mathbf{1}_{N_l}$. □

C.3. Proof for the Main Theorem 5.6

In this section, we provide the main proof of Theorem 5.6. For reader's convenience, we provide the recap version in Theorem C.5 by omitting the definition claim, where the detailed definition of $A_{ul}, A_{ll}, q_i, \bar{U}^{\flat\top}, \bar{\mathfrak{l}}^\flat, \vec{\eta}_u$ is in Section C.2.

The proof of Theorem 5.6 consists of four steps. Firstly, $\mathcal{E}(f)$ is bounded by $\mathcal{R}(U^*)$ as we show in Lemma 5.1. Secondly, the residual $\mathcal{R}(U^*, \vec{y})$ of the original representation can be approximated by the residual $\mathcal{R}\left(\bar{U}^*, \vec{y}\right)$ analyzed in Section C.2. Thirdly, the approximation error bound is in the order of $\frac{\|\dot{A}-\bar{A}\|_2}{\sigma_k - \sigma_{k+1}}$ as shown in Section C.4. Finally, we show that the coverage measurement $\kappa(\vec{y})$ can be lower bounded in Section C.5.

Theorem C.5. *(Recap of Theorem 5.6) Based on the assumptions made in Lemma C.6, Lemma C.9 and Lemma C.10. The linear probing error is bounded by:*

$$\mathcal{E}(f) \lesssim \frac{2N_u}{|\mathcal{Y}_u|}\left(\sum_i^{|\mathcal{Y}_u|} \mathfrak{T}(\vec{y}_i)(1-\kappa(\vec{y}_i)^2) + \frac{\|\dot{A}-\bar{A}\|_2}{\sigma_k - \sigma_{k+1}}\right), \tag{32}$$

where for single labeling vector $\vec{y}$,

$$\kappa(\vec{y}) = \cos(\bar{U}^{\flat\top}\vec{y}, \bar{\mathfrak{l}}^\flat) \gtrsim \min_{i>k, j>k} \frac{2\sqrt{\frac{\vec{y}^\top q_i}{\vec{\eta}_u^\top q_i}\frac{\vec{y}^\top q_j}{\vec{\eta}_u^\top q_j}}}{\frac{\vec{y}^\top q_i}{\vec{\eta}_u^\top q_i} + \frac{\vec{y}^\top q_j}{\vec{\eta}_u^\top q_j}}.$$

Proof. According to Lemma 5.1, we have

$$\mathcal{E}(f) \leq 2\mathcal{R}(U^*) = 2\sum_{i\in\mathcal{Y}_u}\mathcal{R}(U^*, \vec{y}_i),$$

where we can view each $\vec{y}_i$ separately. For simplicity, we use $\vec{y}$ in the following proof. As show in Section C.2, $\mathcal{R}(U^*, \vec{y})$ can be approximately estimated by $\mathcal{R}(\bar{U}^*, \vec{y}_i) = (1-\kappa(\vec{y})^2)\|\bar{U}^{\flat\top}\vec{y}_i\|_2^2 = \mathfrak{T}(\vec{y}_i)(1-\kappa(\vec{y})^2)\|\vec{y}_i\|_2^2$. Such approximation bound is given by

$$\mathcal{R}(U^*, \vec{y}) \lesssim \mathcal{R}(\bar{U}^*, \vec{y}) + \frac{2\|\dot{A}-\bar{A}\|_2}{\sigma_k - \sigma_{k+1}}\|\vec{y}\|_2^2,$$

[3]When $\text{rank}(A_{uu} - \frac{\vec{\eta}_u \vec{\eta}_u^\top}{\eta_l}) < \text{rank}(A_{uu})$, it means that η_u happens to cancel out one of the direction in A_{uu}. Such an event has zero probability almost sure in reality. We do not consider this case in our proof.

as shown in Lemma C.6 in Section C.4. Putting things together, we have

$$\mathcal{E}(f) \lesssim 2\sum_{i}^{|\mathcal{Y}_u|} \mathfrak{T}(\vec{y}_i)(1-\kappa(\vec{y})^2)\|\vec{y}_i\|_2^2 + \frac{2\|\dot{A}-\bar{A}\|_2}{\sigma_k-\sigma_{k+1}}\|\vec{y}_i\|_2^2.$$

If the sample size in the novel class is balanced, we have $\|\vec{y}\|_2^2 = \frac{N_u}{|\mathcal{Y}_u|}$, we have:

$$\mathcal{E}(f) \lesssim \frac{2N_u}{|\mathcal{Y}_u|}\left(\sum_{i}^{|\mathcal{Y}_u|} \mathfrak{T}(\vec{y}_i)(1-\kappa(\vec{y})^2) + \frac{\|\dot{A}-\bar{A}\|_2}{\sigma_k-\sigma_{k+1}}\right),$$

Finally, the lower bound of κ is given by Lemma C.10 and proved in Section C.5. □

C.4. Error Bound by Approximation

We see in Section C.2 that we use the approximated version $\bar{U}^*$ instead of the actual feature representation U^*, which creates a gap. In this section, we will present a formal analysis on the gap between the induced residuals $\mathcal{R}(U^*, \vec{y})$ and $\mathcal{R}(\bar{U}^*, \vec{y})$.

Lemma C.6. *When $\|\dot{A}-\bar{A}\|_2 < \frac{1}{2}(\sigma_k-\sigma_{k+1})$ and $c_u \triangleq \mathbb{E}_{i\in\mathcal{I}}(1-\|\bar{u}_i\|_2^2)$ is a non-zero value[4], we have*

$$\mathcal{R}(U^*, \vec{y}) \lesssim \mathcal{R}(\bar{U}^*, \vec{y}) + 2\frac{\|\dot{A}-\bar{A}\|_2}{\sigma_k-\sigma_{k+1}}\|\vec{y}\|_2^2.$$

Proof. Recall that $\vec{y}' = [\zeta\mathbf{1}_{1\times N_l}, \vec{y}^\top]^\top$ is an extended labeling vector where ζ is any real number defined in the proof of Theorem C.4. We let $\zeta^* = \arg\min_{\zeta\in\mathbb{R}} \|\bar{V}^{\flat\top}\vec{y}'\|_2^2$ so that $\bar{y}^* = [\zeta^*\mathbf{1}_{1\times N_l}, \vec{y}^\top]$. We then define $\delta \triangleq \min\{\sigma_k - \bar{\sigma}_{k+1}, \bar{\sigma}_k - \sigma_{k+1}\}$,

$$\begin{aligned}
\mathcal{R}(U^*, \vec{y}) =& \min_{\zeta\in\mathbb{R}} \|V^{\flat\top}\vec{y}'\|_2^2 \\
=& \min_{\zeta\in\mathbb{R}} \vec{y}'^\top V^\flat V^{\flat\top}\vec{y}' \\
=& \min_{\zeta\in\mathbb{R}} (\vec{y}'^\top \bar{V}^\flat\bar{V}^{\flat\top}\vec{y}' + \vec{y}'^\top V^\flat V^{\flat\top}\vec{y}' - \vec{y}'^\top \bar{V}^\flat\bar{V}^{\flat\top}\vec{y}') \\
\leq& \mathcal{R}(\bar{U}^*, \vec{y}) + |\bar{y}^{*\top}(V^\flat V^{\flat\top} - \bar{V}^\flat\bar{V}^{\flat\top})\bar{y}^*| \\
\leq& \mathcal{R}(\bar{U}^*, \vec{y}) + \|V^\flat V^{\flat\top} - \bar{V}^\flat\bar{V}^{\flat\top}\|\|\bar{y}^*\|_2^2 \\
=& \mathcal{R}(\bar{U}^*, \vec{y}) + \|V^{\flat\top}\bar{V}^*\|\|\bar{y}^*\|_2^2 \\
\leq& \mathcal{R}(\bar{U}^*, \vec{y}) + \frac{\|\dot{A}-\bar{A}\|_2}{\delta}\|\bar{y}^*\|_2^2 \\
\leq& \mathcal{R}(\bar{U}^*, \vec{y}) + \frac{2\|\dot{A}-\bar{A}\|_2}{\sigma_k-\sigma_{k+1}}\|\bar{y}^*\|_2^2,
\end{aligned}$$

where the second last inequality is from Davis-Kahan theorem on subspace distance $\|V^\flat V^{\flat\top} - \bar{V}^\flat\bar{V}^{\flat\top}\| = \|V^{\flat\top}\bar{V}^*\| = \|\bar{V}^{\flat\top}V^*\|$, and the last inequality is from Weyl's inequality so that $\delta \geq (\sigma_k - \sigma_{k+1}) - \|\dot{A}-\bar{A}\|_2 \geq \frac{1}{2}(\sigma_k - \sigma_{k+1})$.

We then investigate the magnitude order of $\|\bar{y}^*\|_2^2$. Note that $\|\bar{y}^*\|_2^2 = \|\vec{y}\|_2^2 + N_l(\zeta^*)^2$ and $\zeta^* = \frac{\bar{\mathbf{l}}^{\flat\top}\bar{U}^{\flat\top}\vec{y}}{\|\bar{\mathbf{l}}^\flat\|_2^2}$ according to the

[4]Note that $c_u = 0$ happens in an extreme case that $\forall i \in \mathcal{I}, \|\bar{l}_i\|_2^2 = 0$ which means the extra knowledge is purely irrelevant to the feature representation. Specifically, this could happen when A_{ul} (defined in Section C.1) is a zero matrix.

proof of Theorem C.4. Then,

$$\begin{aligned}\|\bar{y}^*\|_2^2 &= \|\vec{y}\|_2^2 + \frac{N_l(\bar{\mathfrak{l}}^{\flat\top}\bar{U}^{\flat\top}\vec{y})^2}{\|\bar{\mathfrak{l}}^{\flat}\|_2^4} \\ &= \|\vec{y}\|_2^2 + \frac{N_l\kappa(\vec{y})^2\|\bar{U}^{\flat\top}\vec{y}\|_2^2}{\|\bar{\mathfrak{l}}^{\flat}\|_2^2} \\ &= \|\vec{y}\|_2^2\left(1+\frac{N_l\kappa(\vec{y})^2\mathfrak{T}(\vec{y})^2}{\|\bar{\mathfrak{l}}^{\flat}\|_2^2}\right) \\ &= \|\vec{y}\|_2^2\left(1+\frac{\kappa(\vec{y})^2\mathfrak{T}(\vec{y})^2}{\sum_{i=k+1}^{N-\Theta}(1-\|\bar{u}_i\|_2^2)}\right),\end{aligned}$$

where the last equation is given by Lemma C.2 when $i > N-\Theta$, $(\bar{\mathfrak{l}}^{\flat})_i = 0$ and also by the fact that when $i \in \mathcal{I}$, $1-\|\bar{u}_i\|_2^2 = \|\bar{l}_i\|_2^2 = N_l(\frac{(\bar{\mathfrak{l}}^{\flat})_i}{N_l})^2 = (\bar{\mathfrak{l}}^{\flat})_i^2/N_l$. Then by the assumption that c_u is non-zero, we have

$$\|\bar{y}^*\|_2^2 = \|\vec{y}\|_2^2(1+\frac{\kappa(\vec{y})^2\mathfrak{T}(\vec{y})^2}{(N-\Theta-k)c_u}) \lesssim \|\vec{y}\|_2^2(1+O(\frac{1}{N})).$$

By plugging back $\|\bar{y}^*\|_2^2$, we have

$$\mathcal{R}(U^*,\vec{y}) \lesssim \mathcal{R}(\bar{U}^*,\vec{y}) + \frac{2\|\dot{A}-\bar{A}\|_2}{\sigma_k-\sigma_{k+1}}\|\vec{y}\|_2^2.$$

□

C.5. Analysis on the Coverage Measurement $\kappa(\vec{y})$

So far we have shown in Theorem C.4 that the sufficient and necessary condition for a zero residual is when the coverage measurement $\kappa(\vec{y}) = \cos(\bar{U}^{\flat\top}\vec{y}, \bar{\mathfrak{l}}^{\flat})$ equals to one. In this section, we provide a deeper analysis on $\kappa(\vec{y})$ in a less restrictive case.

Recall that we have proved in Theorem C.4 that the sufficient and necessary condition for $\kappa(\vec{y}) = 1$ is:

$$\exists\omega \in \mathbb{R}, \forall i \in \mathcal{I}, \vec{y}^\top(\bar{\sigma}_i I - A_{uu})^\dagger\vec{\eta}_u = \omega. \tag{33}$$

In a general case, we consider ω_i which is variant on i:

$$\omega_i \triangleq \vec{y}^\top(\bar{\sigma}_i I - A_{uu})^\dagger\vec{\eta}_u.$$

Our discussion on $\kappa(\vec{y})$ is based on the following definitions:

Definition C.7. Let q_j and d_j as the j-th eigenvector/eigenvalue of A_{uu}. Then we define $\tilde{\mathbf{y}}_j \triangleq \vec{y}^\top q_j$ and $\tilde{\boldsymbol{\eta}}_j \triangleq \vec{\eta}_u^\top q_j$.

Before showing the bound on $\kappa(\vec{y})$, we first show the following Lemma C.8 and Lemma C.9 which is the important ingredient needed to derive the lower bound of $\kappa(\vec{y})$. We defer the proof to Section C.5.1 and Section C.5.2 respectively.

Lemma C.8. *Let $\Omega \in \mathbb{R}^{(N-\Theta-k)\times(N-\Theta-k)}$ be the diagonal matrix with $\Omega_{i'i'} = \omega_i$ ($i' = i-k$ to be aligned with the indexing of ω_i). For any vector $\mathfrak{l} \in \mathbb{R}^{N-\Theta-k}$, we have the following inequality:*

$$1 \geq \frac{\mathfrak{l}^\top\Omega\mathfrak{l}}{\|\Omega\mathfrak{l}\|_2\|\mathfrak{l}\|_2} \geq \min_{i,j\in\mathcal{I}}\frac{2\sqrt{\omega_i\omega_j}}{\sqrt{\omega_j}+\sqrt{\omega_i}},$$

A sufficient and necessary condition for $\frac{\mathfrak{l}^\top\Omega\mathfrak{l}}{\|\Omega\mathfrak{l}\|_2\|\mathfrak{l}\|_2}$ being 1 for all $\mathfrak{l}$ is to let ω_i be the same for all $i \in \mathcal{I}$.

Lemma C.9. *Assume η_u is upper bounded by a small value $\frac{1}{M}$: $\max_{j=1\ldots N_u}(\vec{\eta}_u)_j = \frac{1}{M}$.[5] For each indexing pair $i \in \mathcal{I}$ and $i' \in \mathcal{I}$ with order $\omega_i < \omega_{i'}$, we have*

$$\frac{\omega_i}{\omega_{i'}} \gtrsim \frac{\vec{y}^\top q_i}{\vec{\eta}_u^\top q_i}\Big/\frac{\vec{y}^\top q_{i'}}{\vec{\eta}_u^\top q_{i'}}.$$

[5] Such assumption is used to align the magnitude later in the proof between $\vec{y} \in [0,1]$ and $\vec{\eta}_u \in [0, \frac{1}{M}]$ for the value range.

Putting the ingredients together, we can finally derive an analytical lower bound of $\kappa(\vec{y})$ in Lemma C.10 based on the angle of $\vec{y}$ / $\vec{\eta}_u$ to each eigenvector of A_{uu}.

Lemma C.10. *W.o.l.g, we let $\omega > 0$ and assume that $\omega_i > 0, \forall i \in \mathcal{I}$ so that perturbation of ω_i to ω to be not significant enough to change the sign of ω. we have:*

$$\kappa(\vec{y}) = \cos(\bar{U}^{\flat\top}\vec{y}, \bar{\mathfrak{l}}^{\flat}) \gtrsim \min_{i>k, j>k} \frac{2\sqrt{\frac{\vec{y}^\top q_i}{\vec{\eta}_u^\top q_i}\frac{\vec{y}^\top q_j}{\vec{\eta}_u^\top q_j}}}{\frac{\vec{y}^\top q_i}{\vec{\eta}_u^\top q_i} + \frac{\vec{y}^\top q_j}{\vec{\eta}_u^\top q_j}},$$

Proof. Recall that

$$\bar{u}_i = (\bar{\sigma}_i I - A_{uu})^\dagger \vec{\eta}_u (\bar{\mathfrak{l}}^{\flat})_i,$$

we consider the specific form of $\kappa(\vec{y})$,

$$\begin{aligned}
\kappa(\vec{y}) &= \cos\left(\bar{U}^{\flat\top}\vec{y}, \bar{\mathfrak{l}}^{\flat}\right) \\
&= \frac{\sum_{i=k+1}^{N} \omega_i (\bar{\mathfrak{l}}^{\flat})_i^2}{\sqrt{\sum_{i=k+1}^{N} \omega_i^2 (\bar{\mathfrak{l}}^{\flat})_i^2}\sqrt{\sum_{i=k+1}^{N} (\bar{\mathfrak{l}}^{\flat})_i^2}} \\
&= \frac{\sum_{i\in\mathcal{I}} \omega_i (\bar{\mathfrak{l}}^{\flat})_i^2}{\sqrt{\sum_{i\in\mathcal{I}} \omega_i^2 (\bar{\mathfrak{l}}^{\flat})_i^2}\sqrt{\sum_{i\in\mathcal{I}} (\bar{\mathfrak{l}}^{\flat})_i^2}} \\
&= \frac{\bar{\mathfrak{l}}'^\top \Omega \bar{\mathfrak{l}}'}{\|\Omega\bar{\mathfrak{l}}'\|_2 \|\bar{\mathfrak{l}}'\|_2},
\end{aligned}$$

where $\Omega \in \mathbb{R}^{N_u - k - \Theta}$ is a diagonal matrix defined in Lemma C.8, and $\bar{\mathfrak{l}}'$ is defined in Eq. (29). According to Lemma C.8, we have

$$\begin{aligned}
\kappa(\vec{y}) &= \frac{\bar{\mathfrak{l}}'^\top \Omega \bar{\mathfrak{l}}'}{\|\Omega\bar{\mathfrak{l}}'\|_2 \|\bar{\mathfrak{l}}'\|_2} \\
&\geq \min_{i,j\in\mathcal{I}} \frac{2\sqrt{\omega_i\omega_j}}{\sqrt{\omega_j} + \sqrt{\omega_i}} \\
&= \min_{i,j\in\mathcal{I}} \frac{2}{\sqrt{\frac{\omega_j}{\omega_i}} + \sqrt{\frac{\omega_i}{\omega_j}}},
\end{aligned}$$

Then by Lemma C.9 and by the fact that $\frac{2}{t+\frac{1}{t}}$ is a monotonically increasing function when $t \in (0, 1)$:

$$\begin{aligned}
\kappa(\vec{y}) &\geq \min_{i,j\in\mathcal{I}} \frac{2}{\sqrt{\frac{\omega_j}{\omega_i}} + \sqrt{\frac{\omega_i}{\omega_j}}} \\
&\gtrsim \min_{i,j\in\mathcal{I}} \frac{2}{\sqrt{\frac{\vec{y}^\top q_i}{\vec{\eta}_u^\top q_i} / \frac{\vec{y}^\top q_j}{\vec{\eta}_u^\top q_j}} + \sqrt{\frac{\vec{y}^\top q_j}{\vec{\eta}_u^\top q_j} / \frac{\vec{y}^\top q_i}{\vec{\eta}_u^\top q_i}}} \\
&> \min_{i>k, j>k} \frac{2\sqrt{\frac{\vec{y}^\top q_i}{\vec{\eta}_u^\top q_i}\frac{\vec{y}^\top q_j}{\vec{\eta}_u^\top q_j}}}{\frac{\vec{y}^\top q_i}{\vec{\eta}_u^\top q_i} + \frac{\vec{y}^\top q_j}{\vec{\eta}_u^\top q_j}}.
\end{aligned}$$

□

C.5.1. PROOF FOR LEMMA C.8

Proof. Consider the function $g(\mathfrak{l}) = \frac{\mathfrak{l}^\top \Omega \mathfrak{l}}{\|\Omega\mathfrak{l}\|_2 \|\mathfrak{l}\|_2}$, the directional derivative $\partial g(\mathfrak{l})/\partial \mathfrak{l}$ is given by:

$$\frac{\partial g(\mathfrak{l})}{\partial \mathfrak{l}} = \frac{2\Omega\mathfrak{l}\|\Omega\mathfrak{l}\|_2\|\mathfrak{l}\|_2 - \Omega^2\mathfrak{l}\frac{\|\mathfrak{l}\|_2}{\|\Omega\mathfrak{l}\|_2}\mathfrak{l}^\top\Omega\mathfrak{l} - \mathfrak{l}\frac{\|\Omega\mathfrak{l}\|_2}{\|\mathfrak{l}\|_2}\mathfrak{l}^\top\Omega\mathfrak{l}}{\|\Omega\mathfrak{l}\|_2^2\|\mathfrak{l}\|_2^2}.$$

The condition for $\partial g(\mathfrak{l})/\partial \mathfrak{l} = 0$ is

$$2\Omega\mathfrak{l} = \Omega^2\mathfrak{l}\frac{\mathfrak{l}^\top\Omega\mathfrak{l}}{\|\Omega\mathfrak{l}\|_2^2} + \mathfrak{l}\frac{\mathfrak{l}^\top\Omega\mathfrak{l}}{\|\mathfrak{l}\|_2^2}.$$

Note that the first condition to satisfy this equation is to let $\mathfrak{l}$ as the eigenvectors of $2\Omega - \Omega^2\frac{\mathfrak{l}^\top\Omega\mathfrak{l}}{\|\Omega\mathfrak{l}\|_2^2}$ which is a diagonal matrix. Then one of the solutions sets is $\mathfrak{l} = c\mathbf{e}_j$ where c is any non-zero scalar value and $\mathbf{e}_j$ is the unit vector with j-th value 1 and 0 elsewhere. Note that this solution set corresponds to the maximum value of $g(\mathfrak{l})$ which is 1. We are then looking into the local minimum value of $g(\mathfrak{l})$ by another solution set. We consider another solution set by considering the following matrix as deficiency:

$$\Gamma \triangleq 2\Omega - \Omega^2\frac{\mathfrak{l}^\top\Omega\mathfrak{l}}{\|\Omega\mathfrak{l}\|_2^2} - \frac{\mathfrak{l}^\top\Omega\mathfrak{l}}{\|\mathfrak{l}\|_2^2}I,$$

where $\mathfrak{l}$ lies in the null space of this matrix. If we let $\varrho = \frac{\|\mathfrak{l}\|_2}{\|\Omega\mathfrak{l}\|_2}$, we have:

$$\Gamma = 2\Omega - \varrho g(\hat{\mathfrak{l}})\Omega^2 - \varrho^{-1}g(\hat{\mathfrak{l}})I$$

and

$$\Gamma_{i'i'} = 2\omega_i - \varrho g(\hat{\mathfrak{l}})\omega_i^2 - \varrho^{-1}g(\hat{\mathfrak{l}}),$$

where i' is indexed starting from 1 and i is indexed starting from k. Note that $\Gamma_{i'i'}$ only has two zero roots. If we consider all ω_i(s) in Ω to be different, Γ can have at most two zero values in the diagonal. Let ω_a, ω_b as two roots of $2\omega - \varrho g(\hat{\mathfrak{l}})\omega^2 - \varrho^{-1}g(\hat{\mathfrak{l}})$, we have:

$$\varrho\omega_a + (\varrho\omega_a)^{-1} = \varrho\omega_b + (\varrho\omega_b)^{-1} = \frac{2}{g(\hat{\mathfrak{l}})}$$

$$\varrho = \frac{\sqrt{\omega_b}}{\sqrt{\omega_a}}, g(\hat{\mathfrak{l}}) = \frac{2}{\sqrt{\frac{\omega_b}{\omega_a}} + \sqrt{\frac{\omega_a}{\omega_b}}},$$

which corresponds to one local minimal with the indexing pair (a, b). By enumerating all the indexing pairs, we have the global minimum of $g(\mathfrak{l})$:

$$g(\mathfrak{l}^*) = \min_{i,j\in\mathcal{I}} \frac{2\sqrt{\omega_i\omega_j}}{\sqrt{\omega_j} + \sqrt{\omega_i}}.$$

Note that when some ω_i, ω_j are identical, this is a special case where the local minimum is equal to the maximum 1. Therefore a sufficient and necessary condition for $g(\mathfrak{l}) = 1$ is to let ω_i be the same for all $i \in \mathcal{I}$.

□

C.5.2. PROOF FOR LEMMA C.9

Proof. We can write ω_i by $\tilde{\mathbf{y}}$ and $\tilde{\boldsymbol{\eta}}$ in Definition C.7:

$$\begin{aligned}
\omega_i &= \vec{y}^\top(\bar{\sigma}_i I - A_{uu})^\dagger \vec{\eta}_u \\
&= \sum_{j\in\mathcal{I}} \frac{(\vec{y}^\top q_j)(\vec{\eta}_u^\top q_j)}{\bar{\sigma}_i - d_j} + \sum_{j=N-\Theta+1}^{N_u} \frac{(\vec{y}^\top q_j)(\vec{\eta}_u^\top q_j)}{\bar{\sigma}_i} \\
&= \sum_{j\in\mathcal{I}} \frac{\tilde{\mathbf{y}}_j\tilde{\boldsymbol{\eta}}_j}{\bar{\sigma}_i - d_j} + \frac{1}{\bar{\sigma}_i}\sum_{j=N-\Theta+1}^{N_u} \tilde{\mathbf{y}}_j\tilde{\boldsymbol{\eta}}_j.
\end{aligned}$$

We then look into the value of $\bar{\sigma}_i$ by solving the eigenvalue problem:

$$
\begin{aligned}
\begin{bmatrix} \eta_l \mathbf{1}_{N_l \times N_l} & \mathbf{1}_{N_l \times 1} \vec{\eta}_u^\top \\ \vec{\eta}_u \mathbf{1}_{1 \times N_l} & A_{uu} \end{bmatrix} \begin{bmatrix} \bar{l}_i \\ \bar{u}_i \end{bmatrix} &= \bar{\sigma}_i \begin{bmatrix} \bar{l}_i \\ \bar{u}_i \end{bmatrix} \\
\iff \eta_l \mathbf{1}_{N_l \times N_l} \bar{l}_i + \mathbf{1}_{N_l \times 1} \vec{\eta}_u^\top \bar{u}_i &= \bar{\sigma}_i \bar{l}_i \\
\iff \mathbf{1}_{N_l \times 1} \eta_l (\vec{\mathfrak{l}}^\flat)_i + \mathbf{1}_{N_l \times 1} \vec{\eta}_u^\top \bar{u}_i &= \mathbf{1}_{N_l \times 1} \bar{\sigma}_i \frac{1}{N_l} (\vec{\mathfrak{l}}^\flat)_i \\
\iff \eta_l (\vec{\mathfrak{l}}^\flat)_i + \vec{\eta}_u^\top \bar{u}_i &= \bar{\sigma}_i \frac{1}{N_l} (\vec{\mathfrak{l}}^\flat)_i \\
\iff \eta_l (\vec{\mathfrak{l}}^\flat)_i + \vec{\eta}_u^\top (\bar{\sigma}_i I - A_{uu})^\dagger \vec{\eta}_u (\vec{\mathfrak{l}}^\flat)_i &= \bar{\sigma}_i \frac{1}{N_l} (\vec{\mathfrak{l}}^\flat)_i \\
\iff \eta_l + \vec{\eta}_u^\top (\bar{\sigma}_i I - A_{uu})^\dagger \vec{\eta}_u &= \frac{\bar{\sigma}_i}{N_l} \\
\iff \eta_l + \sum_{j \in \mathcal{I}} \frac{\tilde{\boldsymbol{\eta}}_j^2}{\bar{\sigma}_i - d_j} &= \frac{\bar{\sigma}_i}{N_l}
\end{aligned}
$$

Note that we get a $(|\mathcal{I}|+1)$-th degree polynomials of $\bar{\sigma}_i$ with $(|\mathcal{I}|+1)$ roots. By observation, we see that there is one root significantly large ($\approx N_l \eta_l$) since N_l and other $|\mathcal{I}|$ roots are very close to each d_j. Based on this intuition, we approximately view it as a unary quadratic equation:

$$
\eta_l + \phi_i + \frac{\tilde{\boldsymbol{\eta}}_i^2}{\bar{\sigma}_i - d_i} = \frac{\bar{\sigma}_i}{N_l},
$$

where we let $\phi_i \triangleq \sum_{j \in \mathcal{I}, j \neq i} \frac{\tilde{\boldsymbol{\eta}}_j^2}{\bar{\sigma}_i - d_j}$. We then proceed by solving this unary quadratic equation by viewing ϕ_i as a variable.

$$
\begin{aligned}
\bar{\sigma}_i(\bar{\sigma}_i - d_i) &= N_l \eta_l (\bar{\sigma}_i - d_i) + N_l \phi_i (\bar{\sigma}_i - d_i) + N_l \tilde{\boldsymbol{\eta}}_i^2 \\
\iff \bar{\sigma}_i^2 &= (d_i + N_l(\eta_l + \phi_i))\bar{\sigma}_i + N_l(\tilde{\boldsymbol{\eta}}_i^2 - (\eta_l + \phi_i) d_i) \\
\iff \bar{\sigma}_i &= \frac{d_i + N_l(\eta_l + \phi_i)}{2} \pm \sqrt{\frac{(d_i + N_l(\eta_l + \phi_i))^2}{4} + N_l(\tilde{\boldsymbol{\eta}}_i^2 - (\eta_l + \phi_i) d_i)} \\
\iff \bar{\sigma}_i &= \frac{d_i + N_l(\eta_l + \phi_i)}{2} \pm \sqrt{\frac{(N_l(\eta_l + \phi_i) - d_i)^2}{4} + N_l \tilde{\boldsymbol{\eta}}_i^2} \\
\iff \bar{\sigma}_i &= \frac{d_i + N_l(\eta_l + \phi_i)}{2} \pm \left(\frac{N_l(\eta_l + \phi_i) - d_i}{2} + \frac{N_l \tilde{\boldsymbol{\eta}}_i^2}{\frac{N_l(\eta_l + \phi_i) - d_i}{2} + \sqrt{\frac{(N_l(\eta_l + \phi_i) - d_i)^2}{4} + N_l \tilde{\boldsymbol{\eta}}_i^2}} \right) \\
\iff \bar{\sigma}_i &= \frac{d_i + N_l(\eta_l + \phi_i)}{2} \pm \left(\frac{N_l(\eta_l + \phi_i) - d_i}{2} + \frac{1}{\frac{\eta_l + \phi_i - \frac{d_i}{N_l}}{2\tilde{\boldsymbol{\eta}}_i^2} + \sqrt{(\frac{\eta_l + \phi_i - \frac{d_i}{N_l}}{2\tilde{\boldsymbol{\eta}}_i^2})^2 + 1}} \right) \\
\iff \bar{\sigma}_i &= \frac{d_i + N_l(\eta_l + \phi_i)}{2} \pm \left(\frac{N_l(\eta_l + \phi_i) - d_i}{2} + \frac{\tilde{\boldsymbol{\eta}}_i^2}{\eta_l + \phi_i - \frac{d_i}{N_l}} - O((\frac{\tilde{\boldsymbol{\eta}}_i^2}{\eta_l + \phi_i})^2) \right)
\end{aligned}
$$

Here we see that $\bar{\sigma}_i$ has two approximated solutions: in the first case, when $\pm$ becomes $+$, $\bar{\sigma}_i \approx N_l \eta_l$ which is the unique very large solution as we mentioned. Another solution is by picking $\pm$ as $-$, we then have $\bar{\sigma}_i \approx d_i - \frac{\tilde{\boldsymbol{\eta}}_i^2}{\eta_l + \phi_i - \frac{d_i}{N_l}}$. The second case is what we are using in this proof since we are looking at the indexing of ω_i with $i \in \mathcal{I}$, which is beyond top-k.

For each indexing pair i and i' with order $\omega_i < \omega_{i'}$, we plug in the solution of $\bar{\sigma}_i$ and $\bar{\sigma}'_i$ respectively:

$$
\begin{aligned}
\frac{\omega_i}{\omega_{i'}} &= \frac{\sum_{j\in\mathcal{I}} \frac{\tilde{\mathbf{y}}_j\tilde{\boldsymbol{\eta}}_j}{d_j-\bar{\sigma}_i} + \frac{1}{\bar{\sigma}'_i}\sum_{j=N-\Theta+1}^{N_u}\tilde{\mathbf{y}}_j\tilde{\boldsymbol{\eta}}_j}{\sum_{j\in\mathcal{I}} \frac{\tilde{\mathbf{y}}_j\tilde{\boldsymbol{\eta}}_j}{d_j-\bar{\sigma}_{i'}} + \frac{1}{\bar{\sigma}''_i}\sum_{j=N-\Theta+1}^{N_u}\tilde{\mathbf{y}}_j\tilde{\boldsymbol{\eta}}_j} \\
&= \frac{\frac{\tilde{\mathbf{y}}_i\tilde{\boldsymbol{\eta}}_i}{d_i-\bar{\sigma}_i} + \sum_{j\in\mathcal{I},j\neq i} \frac{\tilde{\mathbf{y}}_j\tilde{\boldsymbol{\eta}}_j}{d_j-\bar{\sigma}_i} + \frac{1}{\bar{\sigma}'_i}\sum_{j=N-\Theta+1}^{N_u}\tilde{\mathbf{y}}_j\tilde{\boldsymbol{\eta}}_j}{\frac{\tilde{\mathbf{y}}_{i'}\tilde{\boldsymbol{\eta}}_{i'}}{d_{i'}-\bar{\sigma}_{i'}} + \sum_{j\in\mathcal{I},j\neq i'} \frac{\tilde{\mathbf{y}}_j\tilde{\boldsymbol{\eta}}_j}{d_j-\bar{\sigma}_{i'}} + \frac{1}{\bar{\sigma}''_i}\sum_{j=N-\Theta+1}^{N_u}\tilde{\mathbf{y}}_j\tilde{\boldsymbol{\eta}}_j} \\
&= \frac{\frac{\tilde{\mathbf{y}}_i}{\tilde{\boldsymbol{\eta}}_i}(\eta_l+\phi_i) + \tilde{\mathbf{y}}_i\tilde{\boldsymbol{\eta}}_i(O((\frac{\tilde{\boldsymbol{\eta}}_i^2}{\eta_l+\phi_i})^2) - O(\frac{1}{N_l})) + \sum_{j\in\mathcal{I},j\neq i} \frac{\tilde{\mathbf{y}}_j\tilde{\boldsymbol{\eta}}_j}{d_j-\bar{\sigma}_i} + \frac{1}{\bar{\sigma}'_i}\sum_{j=N-\Theta+1}^{N_u}\tilde{\mathbf{y}}_j\tilde{\boldsymbol{\eta}}_j}{\frac{\tilde{\mathbf{y}}_{i'}}{\tilde{\boldsymbol{\eta}}_{i'}}(\eta_l+\phi_{i'}) + \tilde{\mathbf{y}}_{i'}\tilde{\boldsymbol{\eta}}_{i'}(O((\frac{\tilde{\boldsymbol{\eta}}_{i'}^2}{\eta_l+\phi'_i})^2) - O(\frac{1}{N_l})) + \sum_{j\in\mathcal{I},j\neq i'} \frac{\tilde{\mathbf{y}}_j\tilde{\boldsymbol{\eta}}_j}{d_j-\bar{\sigma}_{i'}} + \frac{1}{\bar{\sigma}''_i}\sum_{j=N-\Theta+1}^{N_u}\tilde{\mathbf{y}}_j\tilde{\boldsymbol{\eta}}_j} \\
&= \frac{\frac{\tilde{\mathbf{y}}_i}{\tilde{\boldsymbol{\eta}}_i}\eta_l + \tilde{\mathbf{y}}_i\tilde{\boldsymbol{\eta}}_i(O((\frac{\tilde{\boldsymbol{\eta}}_i^2}{\eta_l+\phi_i})^2) - O(\frac{1}{N_l})) + \sum_{j\in\mathcal{I},j\neq i} \frac{1}{d_j-\bar{\sigma}_i}\tilde{\boldsymbol{\eta}}_j(\tilde{\mathbf{y}}_j + \tilde{\mathbf{y}}_i\frac{\tilde{\boldsymbol{\eta}}_j}{\tilde{\boldsymbol{\eta}}_i}) + \frac{1}{\bar{\sigma}'_i}\sum_{j=N-\Theta+1}^{N_u}\tilde{\mathbf{y}}_j\tilde{\boldsymbol{\eta}}_j}{\frac{\tilde{\mathbf{y}}_{i'}}{\tilde{\boldsymbol{\eta}}_{i'}}\eta_l + \tilde{\mathbf{y}}_{i'}\tilde{\boldsymbol{\eta}}_{i'}(O((\frac{\tilde{\boldsymbol{\eta}}_{i'}^2}{\eta_l+\phi'_i})^2) - O(\frac{1}{N_l})) + \sum_{j\in\mathcal{I},j\neq i'} \frac{1}{d_j-\bar{\sigma}_{i'}}\tilde{\boldsymbol{\eta}}_j(\tilde{\mathbf{y}}_j + \tilde{\mathbf{y}}_{i'}\frac{\tilde{\boldsymbol{\eta}}_j}{\tilde{\boldsymbol{\eta}}_{i'}}) + \frac{1}{\bar{\sigma}''_i}\sum_{j=N-\Theta+1}^{N_u}\tilde{\mathbf{y}}_j\tilde{\boldsymbol{\eta}}_j}.
\end{aligned}
$$

According to assumption that η_u is bounded by $\frac{1}{M}$, we align the magnitude between $\vec{y}$ and $\vec{\eta}_u$ by defining $\vec{\eta}'_u = M\vec{\eta}_u$ which is now also in the range of $[0,1]$ like $\vec{y}$. Then we also scale the following terms: $\tilde{\boldsymbol{\eta}}' = M\tilde{\boldsymbol{\eta}}$. Therefore we can simplify the equation to be:

$$
\begin{aligned}
\frac{\omega_i}{\omega_{i'}} &= \frac{M\frac{\tilde{\mathbf{y}}_i}{\tilde{\boldsymbol{\eta}}'_i}\eta_l + \frac{1}{M}\tilde{\mathbf{y}}_i\tilde{\boldsymbol{\eta}}'_i(O(\frac{1}{M^4}(\frac{\tilde{\boldsymbol{\eta}}_i'^2}{\eta_l+\phi_i})^2) - O(\frac{1}{N_l})) + \frac{1}{M}\sum_{j\in\mathcal{I},j\neq i} \frac{1}{d_j-\bar{\sigma}_i}\tilde{\boldsymbol{\eta}}'_j(\tilde{\mathbf{y}}_j + \tilde{\mathbf{y}}_i\frac{\tilde{\boldsymbol{\eta}}'_j}{\tilde{\boldsymbol{\eta}}'_i}) + \frac{1}{M\bar{\sigma}'_i}\sum_{j=N-\Theta+1}^{N_u}\tilde{\mathbf{y}}_j\tilde{\boldsymbol{\eta}}'_j}{M\frac{\tilde{\mathbf{y}}_{i'}}{\tilde{\boldsymbol{\eta}}'_{i'}}\eta_l + \frac{1}{M}\tilde{\mathbf{y}}_{i'}\tilde{\boldsymbol{\eta}}'_{i'}(O(\frac{1}{M^4}(\frac{\tilde{\boldsymbol{\eta}}_{i'}'^2}{\eta_l+\phi'_i})^2) - O(\frac{1}{N_l})) + \frac{1}{M}\sum_{j\in\mathcal{I},j\neq i'} \frac{1}{d_j-\bar{\sigma}_{i'}}\tilde{\boldsymbol{\eta}}'_j(\tilde{\mathbf{y}}_j + \tilde{\mathbf{y}}_{i'}\frac{\tilde{\boldsymbol{\eta}}'_j}{\tilde{\boldsymbol{\eta}}'_{i'}}) + \frac{1}{M\bar{\sigma}''_i}\sum_{j=N-\Theta+1}^{N_u}\tilde{\mathbf{y}}_j\tilde{\boldsymbol{\eta}}'_j} \\
&= \frac{\frac{\tilde{\mathbf{y}}_i}{\tilde{\boldsymbol{\eta}}'_i}\eta_l + \tilde{\mathbf{y}}_i\tilde{\boldsymbol{\eta}}'_i(O(\frac{1}{M^6}) - O(\frac{1}{M^2N_l})) + \frac{1}{M^2}\sum_{j\in\mathcal{I},j\neq i} \frac{1}{d_j-d_i+O(\frac{1}{M^2})}\tilde{\boldsymbol{\eta}}'_j(\tilde{\mathbf{y}}_j + \tilde{\mathbf{y}}_i\frac{\tilde{\boldsymbol{\eta}}'_j}{\tilde{\boldsymbol{\eta}}'_i}) + \tilde{\mathbf{y}}_i\frac{\tilde{\boldsymbol{\eta}}'_j}{\tilde{\boldsymbol{\eta}}'_i}) + \frac{1}{M^2\bar{\sigma}'_i}\sum_{j=N-\Theta+1}^{N_u}\tilde{\mathbf{y}}_j\tilde{\boldsymbol{\eta}}'_j}{\frac{\tilde{\mathbf{y}}_{i'}}{\tilde{\boldsymbol{\eta}}'_{i'}}\eta_l + \tilde{\mathbf{y}}_{i'}\tilde{\boldsymbol{\eta}}'_{i'}(O(\frac{1}{M^6}) - O(\frac{1}{M^2N_l})) + \frac{1}{M^2}\sum_{j\in\mathcal{I},j\neq i'} \frac{1}{d_j-d_{i'}+O(\frac{1}{M^2})}\tilde{\boldsymbol{\eta}}'_j(\tilde{\mathbf{y}}_j + \tilde{\mathbf{y}}_{i'}\frac{\tilde{\boldsymbol{\eta}}'_j}{\tilde{\boldsymbol{\eta}}'_{i'}}) + \frac{1}{M^2\bar{\sigma}''_i}\sum_{j=N-\Theta+1}^{N_u}\tilde{\mathbf{y}}_j\tilde{\boldsymbol{\eta}}'_j} \\
&= \frac{\frac{\tilde{\mathbf{y}}_i}{\tilde{\boldsymbol{\eta}}'_i}\eta_l + O(\frac{1}{M^2})}{\frac{\tilde{\mathbf{y}}_{i'}}{\tilde{\boldsymbol{\eta}}'_{i'}}\eta_l + O(\frac{1}{M^2})},
\end{aligned}
$$

where we simply regard the remaining term with a magnitude much smaller than M. Note that M can be viewed as the magnitude gap of $\frac{\max_i(\vec{y})_i}{\max_i(\vec{\eta}_u)_i}$. In our case, $\max_i(\vec{y})_i$ is set to 1. However, one can always multiply $\vec{y}$ with a large constant to make M significantly large without changing the residual analysis in the main theorem. In summary, we have

$$
\frac{\omega_i}{\omega_{i'}} \gtrsim \frac{\frac{\tilde{\mathbf{y}}_i}{\tilde{\boldsymbol{\eta}}'_i}\eta_l}{\frac{\tilde{\mathbf{y}}_{i'}}{\tilde{\boldsymbol{\eta}}'_{i'}}\eta_l} = \frac{\tilde{\mathbf{y}}_i}{\tilde{\boldsymbol{\eta}}_i} / \frac{\tilde{\mathbf{y}}_{i'}}{\tilde{\boldsymbol{\eta}}_{i'}} = \frac{\vec{y}^\top q_i}{\vec{\eta}_u^\top q_i} / \frac{\vec{y}^\top q_{i'}}{\vec{\eta}_u^\top q_{i'}}.
$$

□

D. Experimental Details

D.1. Details of Training Configurations

For a fair comparison, we use ResNet-18 (He et al., 2016) as the backbone for all methods. We add a trainable two-layer MLP projection head that projects the feature from the penultimate layer to an embedding space $\mathbb{R}^k$ ($k = 1000$). We use the same data augmentation strategies as SimSiam (Chen & He, 2021; HaoChen et al., 2021). We train our model $f(\cdot)$ for 1200 epochs by NCD Spectral Contrastive Loss defined in Eq. (4). We set $\alpha = 0.0225$ and $\beta = 2$. We use SGD with momentum 0.95 as an optimizer with cosine annealing (lr=0.03), weight decay 5e-4, and batch size 512. We also conduct a sensitivity analysis of the hyper-parameters in Figure 4. The performance comparison for each hyper-parameter is reported by fixing other hyper-parameters. The results suggest that the novel class discovery performance of NSCL is stable when α, β in a reasonable range and with different learning rates.

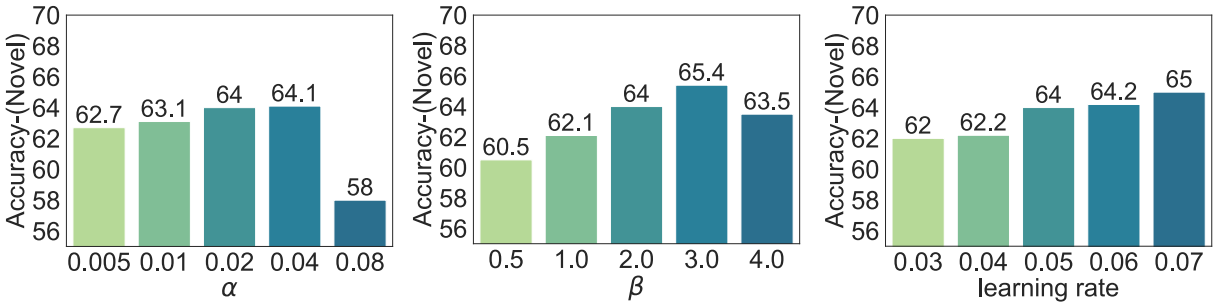

Figure 4. Sensitivity analysis of hyper-parameters α, β, and learning rate. We use the training split of CIFAR-100-50/50, and report the novel class accuracy.

D.2. Experimental Details of Toy Example

Recap of set up. In Section 5.1 we consider a toy example that helps illustrate the core idea of our theoretical findings. Specifically, the example aims to cluster 3D objects of different colors and shapes, generated by a 3D rendering software (Johnson et al., 2017) with user-defined properties including colors, shape, size, position, etc.

In what follows, we define two data configurations and corresponding graphs, where the labeled data is correlated with the attribute of unlabeled data (**case 1**) vs. not (**case 2**). For both cases, we have an unlabeled dataset containing red/blue cubes/spheres as:

$$\mathcal{X}_u \triangleq \{X_{\square,c_1}, X_{\bigcirc,c_1}, X_{\square,c_2}, X_{\bigcirc,c_2}\}.$$

In the first case, we let the labeled data $\mathcal{X}_l^{\text{case 1}}$ be strongly correlated with the target class (red color) in unlabeled data:

$$\mathcal{X}_l^{\text{case 1}} \triangleq \{X_{\ominus,c_1}\}(\text{red cylinder}).$$

In the second case, we use gray cylinders which have no overlap in either shape and color:

$$\mathcal{X}_l^{\text{case 2}} \triangleq \{X_{\ominus,c_3}\}(\text{gray cylinder}).$$

Putting it together, our entire training dataset is $\mathcal{X}^{\text{case 1}} = \mathcal{X}_l^{\text{case 1}} \cup \mathcal{X}_u$ or $\mathcal{X}^{\text{case 2}} = \mathcal{X}_l^{\text{case 2}} \cup \mathcal{X}_u$.

Experimental details for Figure 3. For training, we rendered 2500 samples for each type of data (4 types in $\mathcal{X}_u$ and 1 type in $\mathcal{X}_l$). In total, we have 12500 samples for both $\mathcal{X}^{\text{case 1}}$ and $\mathcal{X}^{\text{case 2}}$. For training, we use the same data augmentation strategy as in SimSiam (Chen & He, 2021). We use ResNet18 and train the model for 40 epochs (sufficient for convergence) with a fixed learning rate of 0.005, using NSCL defined in Eq. (4). We set $\alpha = 0.04$ and $\beta = 1$, respectively. Our visualization is by PyTorch implementation of UMAP (McInnes et al., 2018), with parameters (`n_neighbors=30, min_dist=1.5, spread=2, metric=euclidean`).