DISCRIMINATIVE SIMILARITY FOR DATA CLUSTERING

Anonymous authors
Paper under double-blind review

ABSTRACT

Similarity-based clustering methods separate data into clusters according to the pairwise similarity between the data, and the pairwise similarity is crucial for their performance. In this paper, we propose *Clustering by Discriminative Similarity (CDS)*, a novel method which learns discriminative similarity for data clustering. CDS learns an unsupervised similarity-based classifier from each data partition, and searches for the optimal partition of the data by minimizing the generalization error of the learnt classifiers associated with the data partitions. By generalization analysis via Rademacher complexity, the generalization error bound for the unsupervised similarity-based classifier is expressed as the sum of discriminative similarity between the data from different classes. It is proved that the derived discriminative similarity can also be induced by the integrated squared error bound for kernel density classification. In order to evaluate the performance of the proposed discriminative similarity, we propose a new clustering method using a kernel as the similarity function, CDS via unsupervised kernel classification (CDSK), with its effectiveness demonstrated by experimental results.

1 INTRODUCTION

Similarity-based clustering methods segment the data based on the similarity measure between the data points, such as spectral clustering (Ng et al., 2001), pairwise clustering method (Shental et al., 2003), K-means (Hartigan & Wong, 1979), and kernel K-means (Schölkopf et al., 1998). The success of similarity-based clustering highly depends on the underlying pairwise similarity over the data, which in most cases are constructed empirically, e.g. by Gaussian kernel or the K-Nearest-Neighbor (KNN) graph. In this paper, we model data clustering as a multiclass classification problem and seek for the data partition where the associated classifier, trained on cluster labels, can have low generalization error. Therefore, it is natural to formulate data clustering problem as a problem of training unsupervised classifiers: a classifier can be trained upon each candidate partition of the data, and the quality of the data partition can be evaluated by the performance of the trained classifier. Such classifier trained on a hypothetical labeling associated with a data partition is termed an unsupervised classifier.

We present *Clustering by Discriminative Similarity (CDS)*, wherein discriminative similarity is derived by the generalization error bound for an unsupervised similarity-based classifier. CDS is based on a novel framework of discriminative clustering by unsupervised classification wherein an unsupervised classifier is learnt from unlabeled data and the preferred hypothetical labeling should minimize the generalization error bound for the learnt classifier. When the popular Support Vector Machines (SVMs) is used in this framework, unsupervised SVM (Xu et al., 2004) can be deduced. In this paper, a similarity-based classifier motivated by similarity learning (Balcan et al., 2008; Cortes et al., 2013), is used as the unsupervised classifier. By generalization analysis via Rademacher complexity, the generalization error bound for the unsupervised similarity-based classifier is expressed as sum of pairwise similarity between the data from different classes. Such pairwise similarity, parameterized by the weights of the unsupervised similarity-based classifier, serves as the discriminative similarity. The term "discriminative similarity" emphasizes the fact that the similarity is learnt so as to improve the discriminative capability of a certain classifier such as the aforementioned unsupervised similarity-based classifier.

1.1 Contributions and Main Results

Firstly, we present Clustering by Discriminative Similarity (CDS) where discriminative similarity is induced by the generalization error bound for unsupervised similarity-based classifier on unlabeled data. The generalization bound for such similarity-based classifier is of independent interest, which is among the few results of generalization bounds for classification using general similarity functions (Section B.1 of the appdendix). When the general similarity function is set to a Positive Semi-Definite (PSD) kernel, the derived discriminative similarity between two data points $\mathbf{x}_i$,$\mathbf{x}_j$ is $S_{ij}^K = 2(\boldsymbol{\alpha}_i + \boldsymbol{\alpha}_j - \lambda\boldsymbol{\alpha}_i\boldsymbol{\alpha}_j)K(\mathbf{x}_i - \mathbf{x}_j)$, where K can be an arbitrary PSD kernel and $\boldsymbol{\alpha}_i$ is the kernel weight associated with $\mathbf{x}_i$. With theoretical and empirical study, we argue that S_{ij}^K should be used for data clustering instead of the conventional kernel similarity corresponding to uniform kernel weights. In the case of binary classification, we prove that the derived discriminative similarity S_{ij}^K has the same form as the similarity induced by the integrated squared error bound for kernel density classification (Section A of the appendix). Such connection suggests that there exists information-theoretic measure which is implicitly equivalent to our CDS framework for unsupervised learning, and our CDS framework is well grounded for learning similarity from unlabeled data.

Secondly, based on our CDS model, we develop a clustering algorithm termed Clustering by Discriminative Similarity via unsupervised Kernel classification (CDSK) in Section 4. CDSK uses a PSD kernel as the similarity function, and outperforms competing clustering algorithms, including nonparametric discriminative similarity based clustering methods and similarity graph based clustering methods, demonstrating the effectiveness of CDSK. When the kernel weights $\{\boldsymbol{\alpha}_i\}$ are uniform, CDSK is equivalent to kernel K-Means (Schölkopf et al., 1998). CDSK is more flexible by learning adaptive kernel weights associated with different data points.

1.2 Connection to Related Works

Our CDS model is related to a class of discriminative clustering methods which classify unlabeled data by various measures on discriminative unsupervised classifiers, and the measures include generalization error (Xu et al., 2004) or the entropy of the posterior distribution of the label (Gomes et al., 2010). Discriminative clustering methods (Xu et al., 2004) predict the labels of unlabeled data by minimizing the generalization error bound for the unsupervised classifier with respect to the hypothetical labeling. Unsupervised SVM is proposed in Xu et al. (2004) which learns a binary classifier to partition unlabeled data with the maximum margin between different clusters. The theoretical properties of unsupervised SVM are further analyzed in Karnin et al. (2012). Kernel logistic regression classifier is employed in Gomes et al. (2010), and it uses the entropy of the posterior distribution of the class label by the classifier to measure the quality of the hypothetical labeling. CDS model performs discriminative clustering based on a novel unsupervised classification framework by considering similarity-based or kernel classifiers which are important classification methods in the supervised learning literature. In contrasts with kernel similarity with uniform weights, the induced discriminative similarity with learnable weights enhances its capability to represent complex interconnection between data. The generalization analysis for CDS is primarily based on distribution free Rademacher complexity. While Yang et al. (2014a) propose nonparametric discriminative similarity for clustering, the nonparametric similarity requires probability density estimation which is difficult for high-dimensional data, and the fixed nonparametric similarity is not adaptive to complicated data distribution.

The paper is organized as follows. We introduce the problem setup of Clustering by Discriminative Similarity in Section 2. We then derive the generalization error bound for the unsupervised similarity-based classifier for CDS in Section 3 where the proposed discriminative similarity is induced by the error bound. The application of CDS to data clustering is shown in Section 4. Throughout this paper, the term kernel stands for the PSD kernel if no special notes are made.

2 Problem Setup

We introduce the problem setup of the formulation of clustering by unsupervised classification. Given unlabeled data $\{\mathbf{x}_l\}_{l=1}^n \subset \mathbb{R}^d$, clustering is equivalent to searching for the hypothetical labeling which is optimal in some sense. Each hypothetical labeling corresponds to a candidate data

partition. Figure 1 illustrates four binary hypothetical labelings which correspond to four partitions of the data, and the data is divided into two clusters by each hypothetical labeling.

The discriminative clustering literature (Xu et al., 2004; Gomes et al., 2010) has demonstrated the potential of multi-class classification for clustering problem. Inspired by the natural connection between clustering and classification, we proposes the framework of Clustering by Unsupervised Classification which models clustering problem as a multi-class classification problem. A classifier is learnt from unlabeled data with a hypothetical labeling, which is associated with a candidate partition of the unlabeled data. The optimal hypothetical labeling is supposed to be the one such that its associated classifier has the minimum generalization error bound. To study the generalization bound for the classifier learnt from hypothetical labeling, the concept of classification model is needed. Given unlabeled data $\{\mathbf{x}_l\}_{l=1}^n$, a classification model $M_{\mathcal{Y}}$ is constructed for any hypothetical labeling $\mathcal{Y} = \{y_l\}_{l=1}^n$ as follows.

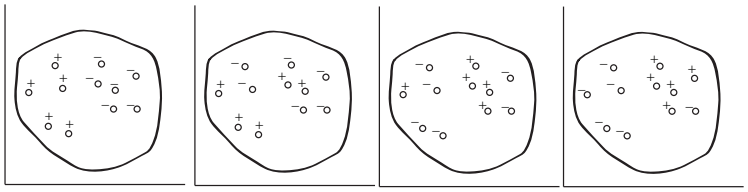


Figure 1: Illustration of binary hypothetical labelings

Definition 2.1. The classification model corresponding to the hypothetical labeling $\mathcal{Y} = \{y_l\}_{l=1}^n$ is defined as $M_{\mathcal{Y}} = (\mathcal{S}, F)$. $\mathcal{S} = \{\mathbf{x}_l, y_l\}_{l=1}^n$ are the labeled data by the hypothetical labeling $\mathcal{Y}$, and $\mathcal{S}$ are assumed to be i.i.d. samples drawn from the some unknown joint distribution P_{XY}, where (X, Y) is a random couple, $X \in \mathcal{X} \subseteq \mathbb{R}^d$ represents the data in some domain $\mathcal{X}$, and $Y \in \{1, 2, ..., c\}$ is the class label of X, c is the number of classes. F is a classifier trained on $\mathcal{S}$. The generalization error of the classification model $M_{\mathcal{Y}}$ is defined as the generalization error of the classifier F in $M_{\mathcal{Y}}$.

The basic assumption of CDS is that the optimal hypothetical labeling minimizes the generalization error bound for the classification model. With f being different classifiers, different discriminative clustering models can be derived. When SVMs is used as the classifier F in the above discriminative model, unsupervised SVM (Xu et al., 2004) is obtained.

In Balcan et al. (2008), the authors proposes a classification method using general similarity functions. The classification rule measures the similarity of the test data to each class, and then assigns the test data to the class such that the weighed average of the similarity between the test data and the training data belonging to that class is maximized over all the classes. Inspired by this classification method, we now consider using a general symmetric and continuous function $S\colon \mathcal{X} \times \mathcal{X} \to [-1, 1]$ as the similarity function in our CDS model. We propose the following hypothesis,

$$h_S(\mathbf{x}, y) = \sum_{i\colon y_i = y} \boldsymbol{\alpha}_i S(\mathbf{x}, \mathbf{x}_i). \tag{1}$$

In the next section, we derive generalization bound for the unsupervised similarity-based classifier based on the above hypothesis, and such generalization bound leads to discriminative similarities for data clustering. When S is a PSD kernel, minimizing the generalization error bound amounts to minimization of a new form of kernel similarity between data from different clusters, which lays the foundation of a new clustering algorithm presented in Section 4.

3 Generalization Bound for Similarity-based Classifier

In this section, the generalization error bound for the classification model in Definition 2.1 with the unsupervised similarity-based classifier is derived as a sum of discriminative similarity between the data from different classes.

3.1 GENERALIZATION BOUND

The following notations are introduced before our analysis. Let $\boldsymbol{\alpha} = [\boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_n]^\top$ be the nonzero weights that sum up to 1, $\boldsymbol{\alpha}^{(y)}$ be a $n \times 1$ column vector representing the weights belonging to class y such that $\boldsymbol{\alpha}_i^{(y)}$ is $\boldsymbol{\alpha}_i$ if $y = y_i$, and 0 otherwise. The margin of the labeled sample $(\mathbf{x}, y)$ is defined as $m_{h_S}(\mathbf{x}, y) = h_S(\mathbf{x}, y) - \text{argmax}_{y' \neq y} h_S(\mathbf{x}, y')$, the sample $(\mathbf{x}, y)$ is classified correctly if $m_{h_S}(\mathbf{x}, y) \geq 0$.

The general similarity-based classifier f_S predicts the label of the input $\mathbf{x}$ by $f_S(\mathbf{x}) = \text{argmax}_{y \in \{1,\ldots,c\}} h_S(\mathbf{x}, y)$. We then begin to derive the generalization error bound for f_S using the Rademacher complexity of the function class comprised of all the possible margin functions m_{h_S}. The Rademacher complexity (Bartlett & Mendelson, 2003; Koltchinskii, 2001) of a function class is defined below:

Definition 3.1. Let $\{\sigma_i\}_{i=1}^n$ be n i.i.d. random variables such that $\Pr[\sigma_i = 1] = \Pr[\sigma_i = -1] = \frac{1}{2}$. The Rademacher complexity of a function class $\mathcal{A}$ is defined as

$$\mathfrak{R}(\mathcal{A}) = \mathbb{E}_{\{\sigma_i\},\{\mathbf{x}_i\}} \left[\sup_{h \in \mathcal{A}} \left| \frac{1}{n} \sum_{i=1}^n \sigma_i h(\mathbf{x}_i) \right| \right]. \tag{2}$$

In order to analyze the generalization property of the classification rule using the general similarity function, we first investigate the properties of general similarity function and its relationship to PSD kernels in terms of eigenvalues and eigenfunctions of the associated integral operator. The integral operator $(L_S f)(\mathbf{x}) = \int S(\mathbf{x}, \mathbf{t}) f(\mathbf{t}) d\mathbf{t}$ is well defined. It can be verified that L_S is a compact operator since S is continuous. According to the spectral theorem in operator theory, there exists an orthonormal basis $\{\phi_1, \phi_2, \ldots\}$ of $\mathcal{L}^2$ which is comprised of the eigenfunctions of L_S, where $\mathcal{L}^2$ is the space of measurable functions which are defined over $\mathcal{X}$ and square Lebesgue integrable. ϕ_k is the eigenfunction of L_S with eigenvalue λ_k if $L_S \phi_k = \lambda_k \phi_k$. The following lemma shows that under certain assumption on the eigenvalues and eigenfunctions of L_S, a general symmetric and continuous similarity can be decomposed into two PSD kernels.

Lemma 3.1. Suppose $S \colon \mathcal{X} \times \mathcal{X} \to [-1, 1]$ is a symmetric continuous function, and $\{\lambda_k\}$ and $\{\phi_k\}$ are the eigenvalues and eigenfunctions of L_S respectively. Suppose $\sum_{k \geq 1} \lambda_k |\phi_k(\mathbf{x})|^2 < C$ for some constant $C > 0$. Then $S(\mathbf{x}, \mathbf{t}) = \sum_{k \geq 1} \lambda_k \phi_k(\mathbf{x}) \phi_k(\mathbf{t})$ for any $\mathbf{x}, \mathbf{t} \in \mathcal{X}$, and it can be decomposed as the difference between two positive semi-definite kernels, that is, $S(\mathbf{x}, \mathbf{t}) = S^+(\mathbf{x}, \mathbf{t}) - S^-(\mathbf{x}, \mathbf{t})$, with

$$S^+(\mathbf{x}, \mathbf{t}) = \sum_{k \colon \lambda_k \geq 0} \lambda_k \phi_k(\mathbf{x}) \phi_k(\mathbf{t}), \quad S^-(\mathbf{x}, \mathbf{t}) = \sum_{k : \lambda_k < 0} |\lambda_k| \phi_k(\mathbf{x}) \phi_k(\mathbf{t}). \tag{3}$$

We now use a regularization term to bound the Rademacher complexity for the classification rule using the general similarity function. Let $\Omega^+(\boldsymbol{\alpha}) = \sum_{y=1}^c {\boldsymbol{\alpha}^{(y)}}^\top \mathbf{S}^+ \boldsymbol{\alpha}^{(y)}$ and $\Omega^-(\boldsymbol{\alpha}) = \sum_{y=1}^c {\boldsymbol{\alpha}^{(y)}}^\top \mathbf{S}^- \boldsymbol{\alpha}^{(y)}$ with $[\mathbf{S}^+]_{ij} = S^+(\mathbf{x}_i, \mathbf{x}_j)$ and $[\mathbf{S}^-]_{ij} = S^-(\mathbf{x}_i, \mathbf{x}_j)$. The space $\mathcal{H}_y$ of all the hypothesis $h_S(\cdot, y)$ associated with label y is defined as

$$\mathcal{H}_{S,y} = \{(\mathbf{x}, y) \to \sum_{i \colon y_i = y} \boldsymbol{\alpha}_i S(\mathbf{x}, \mathbf{x}_i) \colon \boldsymbol{\alpha} \geq \mathbf{0}, \mathbf{1}^\top \boldsymbol{\alpha} = 1, \Omega^+(\boldsymbol{\alpha}) \leq {B^+}^2, \Omega^-(\boldsymbol{\alpha}) \leq {B^-}^2\}, 1 \leq y \leq c,$$

with positive number B^+ and B^- which bound Ω^+ and Ω^- respectively. Let the hypothesis space comprising all possible margin functions be $\mathcal{H}_S = \{(\mathbf{x}, y) \to m_{h_S}(\mathbf{x}, y) \colon h_S(\mathbf{x}, y) \in \mathcal{H}_{S,y}\}$. We then present the main result in this section about the generalization error of unsupervised similarity-based classifier f_S.

Theorem 3.2. Given the discriminative model $M_{\mathcal{Y}} = (\mathcal{S}, f_S)$, suppose $\Omega^+(\boldsymbol{\alpha}) \leq {B^+}^2$, $\Omega^-(\boldsymbol{\alpha}) \leq {B^-}^2$, $\sup_{\mathbf{x} \in \mathcal{X}} |S^+(\mathbf{x}, \mathbf{x})| \leq R^2$, $\sup_{\mathbf{x} \in \mathcal{X}} |S^-(\mathbf{x}, \mathbf{x})| \leq R^2$ for positive constants B^+, B^- and R.

Then with probability $1-\delta$ over the labeled data $\mathcal{S}$ with respect to any distribution in P_{XY}, under the assumptions of Lemma 3.1, the generalization error of the general classifier f_S satisfies

$$\begin{aligned} R(f_S) =& \Pr\left[Y \neq f_S(X)\right] \\ \leq& \widehat{R}_n(f_S) + \frac{8R(2c-1)c(B^+ + B^-)}{\gamma\sqrt{n}} + \left(\frac{16c(2c-1)(B^+ + B^-)R^2}{\gamma} + 1\right)\sqrt{\frac{\log\frac{4}{\delta}}{2n}}, \end{aligned} \tag{4}$$

where $\widehat{R}_n(f_S) = \frac{1}{n}\sum_{i=1}^{n}\Phi\Big(\frac{h_S(\mathbf{x}_i,y_i) - \sum_{y\neq y_i} h_S(\mathbf{x}_i,y)}{\gamma}\Big)$ is the empirical loss of f_S on the labeled data, $\gamma > 0$ is a constant and Φ is defined as $\Phi(x) = \min\{1, \max\{0, 1-x\}\}$. Moreover, if $\gamma \geq c$, the empirical loss $\widehat{R}_n(f_S)$ is

$$\widehat{R}_n(f_S) = 1 - \frac{1}{n\gamma}\sum_{i,j=1}^{n}\frac{\boldsymbol{\alpha}_i + \boldsymbol{\alpha}_j}{2}S(\mathbf{x}_i,\mathbf{x}_j) + \frac{1}{n\gamma}\sum_{1\leq i<j\leq n}^{n} 2(\boldsymbol{\alpha}_i + \boldsymbol{\alpha}_j)S(\mathbf{x}_i,\mathbf{x}_j)\mathbb{I}_{y_i\neq y_j}. \tag{5}$$

Remark 3.3. Lemma E.3 in the Appendix shows that the Rademacher complexity of $\mathcal{H}_S$ is bounded in terms of B^+ and B^-, and that is why these two quantities appear on the RHS of (4). In addition, when S is a Positive Semi-Definite (PSD) kernel K, it can be verified that $S^- \equiv 0$, $S = S^+$.

Remark 3.4. When the decomposition $S = S^+ - S^-$ exists and S^+, S^- are PSD kernels, S is the kernel of some Reproducing Kernel Kreĭn Space (RKKS) (Mary, 2003). Ong et al. (2004) and Loosli et al. (2016) analyzed the problem of learning SVM-style classifiers with indefinite kernels from the Kreĭn space. However, their work does not show when and how an indefinite and general similarity function can have PSD decomposition, as well as the generalization analysis for the similarity-based classifier using such general indefinite function as similarity measure. Our analysis deals with these problems in Lemma 3.1 and Theorem 3.2. It should be emphasized that our generalization bound is of independent interest in supervised learning, because it is among the few results of generalization bounds using general similarity-based classifier. Section B.1 shows that the our bound is a principled result with strong connection to established generalization error bound for Support Vector Machines (SVMs) or Kernel Machines.

3.2 Clustering by Discriminative Similarity

We let

$$S_{ij}^{\text{sim}} = 2(\boldsymbol{\alpha}_i + \boldsymbol{\alpha}_j)S(\mathbf{x}_i,\mathbf{x}_j) - 2\lambda\boldsymbol{\alpha}_i\boldsymbol{\alpha}_j S^+(\mathbf{x}_i,\mathbf{x}_j) - 2\lambda\boldsymbol{\alpha}_i\boldsymbol{\alpha}_j S^-(\mathbf{x}_i,\mathbf{x}_j) \tag{6}$$

be the discriminative similarity between data from different classes, which is induced by the generalization error bound (4) for the unsupervised general similarity-based classifier f_S. Minimizing the bound (4) motivates us to consider the following problem that minimizes $\widehat{R}_n(f_S) + \lambda\big(\Omega^+(\boldsymbol{\alpha}) + \Omega^-(\boldsymbol{\alpha})\big)$, i.e.,

$$\begin{aligned} &\min_{\boldsymbol{\alpha},\mathcal{Y}} \sum_{1\leq i<j\leq n}^{n} S_{ij}^{\text{sim}}\mathbb{I}_{y_i\neq y_j} - \sum_{i,j=1}^{n}\frac{\boldsymbol{\alpha}_i + \boldsymbol{\alpha}_j}{2}S(\mathbf{x}_i,\mathbf{x}_j) + \lambda(\boldsymbol{\alpha}^\top\mathbf{S}^+\boldsymbol{\alpha} + \boldsymbol{\alpha}^\top\mathbf{S}^-\boldsymbol{\alpha}) \\ &s.t.\ \boldsymbol{\alpha} \geq \mathbf{0}, \mathbf{1}^\top\boldsymbol{\alpha} = 1, \mathcal{Y} = \{y_i\}_{i=1}^n, \end{aligned} \tag{7}$$

where $\lambda > 0$ is the weighting parameter for the regularization term $\Omega^+(\boldsymbol{\alpha}) + \Omega^-(\boldsymbol{\alpha})$. Note that we do not set λ to $\frac{16c(2c-1)R^2 + 8\sqrt{2}R(2c-1)c}{\sqrt{2}\gamma}$ exactly matching the RHS of (4), because λ controls the weight of the regularization term which bounds the unknown complexity of the function class $\mathcal{H}_S$. Note that (7) encourages the discriminative similarity S_{ij}^{sim} between the data from different classes small. The optimization problem (7) forms the formulation of Clustering by Discriminative Similarity (CDS).

By Remark 3.3, when S is a PSD kernel K, $S^- \equiv 0$, $S = S^+$, S_{ij}^{sim} reduces to the following discriminative similarity for PSD kernels:

$$S_{ij}^K = 2(\boldsymbol{\alpha}_i + \boldsymbol{\alpha}_j - \lambda\boldsymbol{\alpha}_i\boldsymbol{\alpha}_j)K(\mathbf{x}_i - \mathbf{x}_j), 1 \leq i,j \leq n, \tag{8}$$

and S_{ij}^K is the similarity induced by the unsupervised kernel classifier by the kernel K.

Without loss of generality, we set $K = K_\tau(\mathbf{x}) = \exp(-\frac{\|\mathbf{x}\|_2^2}{2\tau^2})$ which is the isotropic Gaussian kernel with kernel bandwidth $\tau > 0$, and we omit the constant that makes integral of K unit.

When setting the general similarity function to kernel K_τ, CDS aims to minimize the error bound for the corresponding unsupervised kernel classifier, which amounts to minimizing the following objective function

$$\min_{\boldsymbol{\alpha}\in\Lambda, \mathcal{Y}=\{y_i\}_{i=1}^n} \sum_{1\le i<j\le n}^{n} S_{ij}^K \mathbb{1}_{y_i\neq y_j} - \sum_{i,j=1}^{n} \frac{\boldsymbol{\alpha}_i + \boldsymbol{\alpha}_j}{2} K_\tau(\mathbf{x}_i - \mathbf{x}_j) + \lambda \boldsymbol{\alpha}^\top \mathbf{K}\boldsymbol{\alpha}, \tag{9}$$

where S_{ij}^K is defined in (8) with $K = K_\tau$. $\mathbf{K} \in \mathbb{R}^{n\times n}$ and $\mathbf{K}_{ij} = K_\tau(\mathbf{x}_i - \mathbf{x}_j)$. λ is tuned such that $S_{ij}^K \ge 0$, e.g. $\lambda \le 2$. In Section A, it is shown that the discriminative similarity (8) can also be induced from the perspective of kernel density classification by kernel density estimators with nonuniform weights. It supports the theoretical justification for the induced discriminative similarity in this section.

4 Application to Data Clustering

In this section, we propose a novel data clustering method termed Clustering by Discriminative Similarity via unsupervised Kernel classification (CDSK) which is an empirical method inspired by our CDS model when the similarity function is a PSD kernel $K = K_\tau$. In accordance with the CDS model in Section 3.2, CDSK aims to minimize (9). However, problem (9) involves minimization with respect to discrete cluster labels $\mathcal{Y} = \{y_i\}$ which is NP-hard. In addition, it potentially results in a trivial solution which puts all the data in a single cluster due to the lack of constraints on the cluster balance. When $\mathbf{Y}$ is a binary matrix where each column is a membership vector for a particular cluster, $\sum_{1\le i<j\le n}^{n} S_{ij}^K \mathbb{1}_{y_i\neq y_j} = \frac{1}{2}\mathrm{Tr}(\mathbf{Y}^\top \mathbf{L}^K \mathbf{Y})$. Therefore, (9) is relaxed in the proposed optimization problem for CDSK below:

$$\min_{\boldsymbol{\alpha}\in\Lambda, \mathbf{Y}\in\mathbb{R}^{n\times c}} \frac{1}{2}\mathrm{Tr}(\mathbf{Y}^\top \mathbf{L}^K \mathbf{Y}) - \sum_{i,j=1}^{n} \frac{\boldsymbol{\alpha}_i + \boldsymbol{\alpha}_j}{2} K_\tau(\mathbf{x}_i - \mathbf{x}_j) + \lambda \boldsymbol{\alpha}^\top \mathbf{K}\boldsymbol{\alpha} \quad s.t.\ \mathbf{Y}^\top \mathbf{D}^K \mathbf{Y} = \mathbf{I}_c, \tag{10}$$

where $\Lambda = \{\boldsymbol{\alpha} \colon \boldsymbol{\alpha} \ge \mathbf{0}, \mathbf{1}^\top \boldsymbol{\alpha} = 1\}$, $\mathbf{S}_{ij}^K = S_{ij}^K$, $\mathbf{L}^K = \mathbf{D}^K - \mathbf{S}^K$ is the graph Laplacian computed with $\mathbf{S}^K$, $\mathbf{D}^K$ is a diagonal matrix with each diagonal element being the sum of the corresponding row of $\mathbf{S}^K$: $[\mathbf{D}^K]_{ii} = \sum_{j=1}^{n} \mathbf{S}_{ij}^K$, $\mathbf{I}_c$ is a $c \times c$ identity matrix, c is the number of clusters. The constraint in (10) is used to balance the cluster size. This is because minimizing (9) without any constraint on the cluster size results in a trivial solution where all data points form a single cluster. Inspired by spectral clustering (Ng et al., 2001), the constraint $\mathbf{Y}^\top \mathbf{D}^K \mathbf{Y} = \mathbf{I}_c$ used in CDSK prevents imbalanced data clusters.

Problem (10) is optimized by coordinate descent. In each iteration of coordinate descent, optimization with respect to $\mathbf{Y}$ is performed with fixed $\boldsymbol{\alpha}$, which is exactly the same problem as that of spectral clustering with a solution formed by the smallest c eigenvectors of the normalized graph Laplacian $(\mathbf{D}^K)^{-1/2}\mathbf{L}^K(\mathbf{D}^K)^{-1/2}$; then the optimization with respect to $\boldsymbol{\alpha}$ is performed with fixed $\mathbf{Y}$, which is a standard constrained quadratic programming problem. The iteration of coordinate descent proceeds until convergence or the maximum iteration number M is achieved. Each iteration solves two subproblems, (11) and (12). The algorithm of CDSK is described in Algorithm 1.
The first subproblem (11) takes $\mathcal{O}(n^2c)$ steps using truncated Singular Value Decomposition (SVD) by Krylov subspace iterative method. We adopt Sequential Minimal Optimization (SMO) (Platt, 1998) to solve the second subproblem (12), which takes roughly $\mathcal{O}(n^{2.1})$ steps as reported in Platt (1998). Therefore, the overall time complexity of CDSK is $\mathcal{O}(Mcn^2 + Mn^{2.1})$. Furthermore, Section C in the appendix explains the theoretical properties of the coordinate descent algorithm for problem (10).

Algorithm 1 Clustering by Discriminative Similarity via unsupervised Kernel classification (CDSK)

Input: Unlabeled dataset $\{\mathbf{x}_l\}_{l=1}^n$, parameter λ, maximum iteration number M.
for $t \leftarrow 1$ to M **do**
With fixed $\boldsymbol{\alpha}$, solve

$$\min_{\mathbf{Y}\in\mathbb{R}^{n\times c}} \mathrm{Tr}(\mathbf{Y}^\top \mathbf{L}^K \mathbf{Y}) \quad s.t.\ \mathbf{Y}^\top \mathbf{D}^K \mathbf{Y} = \mathbf{I}_c, \tag{11}$$

With fixed $\mathbf{Y}$, solve

$$\min_{\boldsymbol{\alpha}\in\Lambda,} \mathrm{Tr}(\mathbf{Y}^\top \mathbf{L}^K \mathbf{Y}) - \sum_{i,j=1}^{n} \frac{\boldsymbol{\alpha}_i + \boldsymbol{\alpha}_j}{2} K_\tau(\mathbf{x}_i - \mathbf{x}_j) + \lambda \boldsymbol{\alpha}^\top \mathbf{K} \boldsymbol{\alpha}$$
$$s.t.\ \mathbf{Y}^\top \mathbf{D}^K \mathbf{Y} = \mathbf{I}_c, \tag{12}$$

end for
Perform K-Means Clustering on rows of $\mathbf{Y}$ to obtain the clustering result.

The baseline named SC-NS performs spectral clustering on the nonparametric similarity proposed in Yang et al. (2014a). The baseline named SC-MS first constructs a similarity matrix between data denoted by $\mathbf{W}$, where $\mathbf{W}_{ij} = K_\tau(\mathbf{x}_i - \mathbf{x}_j)$, then optimize the kernel bandwidth h by minimizing $\sum_i \|\mathbf{x}_i - \frac{1}{d_i}\sum_j \mathbf{W}_{ij}\mathbf{x}_j\|_2$ where $d_i = \sum_j \mathbf{W}_{ij}$. SC-MS then performs spectral clustering on $\mathbf{W}$ with the kernel bandwidth h obtained from the optimization.

To demonstrate the advantage of the proposed parametric discriminative similarity, we compare CDSK to various baseline clustering methods. SC stands for Spectral Clustering, which is the best performer among spectral clustering with similarity matrix set by Gaussian kernel (SCK), spectral clustering with similarity matrix set by a manifold-based similarity learning method (SC-MS) (Karasuyama & Mamitsuka, 2013), and spectral clustering with similarity matrix set by the nonparametric discriminative similarity (SC-NS) in (Yang et al., 2014a). In SC-MS, Gaussian kernel is used as data similarity, and the parameters of the diagonal covariance matrix is optimized so as to minimize the data reconstruction error term. SC-NS minimizes nonparametric kernel similarity between data across different clusters, which is the same objective as that of kernel K-Means (Schölkopf et al., 1998), so its performance is the same as kernel K-Means. Please refer to Section 5 for discussion about other baselines.

Datasets. We conduct experiments on the Yale face dataset, UCI Ionosphere dataset, the MNIST handwritten digits dataset and the Georgia Face dataset. The Yale face dataset has face images of 15 people with 11 images for each person. The Ionosphere data contains 351 points of dimensionality 34. The Georgia Face dataset contains images of 50 people, and each person is represented by 15 color images with cluttered background. COIL-20 dataset has 1440 images of size 32×32 for 20 objects with background removed in all images. The COIL-100 dataset contains 100 objects with 72 images of size 32×32 for each object. CMU PIE face data contains 11554 cropped face images of size 32×32 for 68 persons, and there are around 170 facial images for each person under different illumination and expressions. The UMIST face dataset is comprised of 575 images of size 112×92 for 20 people. CMU Multi-PIE (MPIE) data (Gross et al., 2010) contains 8916 facial images captured in four sessions. The MNIST handwritten digits database has a total number of 70000 samples of dimensionality 1024 for digits from 0 to 9. The digits are normalized and centered in a fixed-size image. The Extended Yale Face Database B (Yale-B) dataset contains face images for 38 subjects with 64 frontal face images taken under different illuminations for each subject. CIFAR-10 dataset consists of 50000 training images and 10000 testing images in 10 classes, and each image is a color one of size 32×32, and we perform data clustering using all the training and testing images. We also use the miniImageNet dataset used in (Vinyals et al., 2016) to evaluate the potential of clustering methods. MiniImageNet consists of $60,000$ color images of size 84×84 with 100 classes, and each class has 600 images. MiniImageNet is known to be more complex than the CIFAR-10 dataset, and we perform clustering on the 64 classes in miniImageNet which are used for few-shot learning, so $38,400$ images are used for clustering. For every clustering method involving randomness such as K-Means clustering, we report the average performance of running it for 10 times.

Performance Measures and Tuning λ by Cross Validation. We use Accuracy (AC) and Normalized Mutual Information (NMI) (Zheng et al., 2004) as the performance measures. The results of different clustering methods are shown in Table 1 in the format of AC(NMI). Except for SC-MS, the kernel bandwidth in all methods is set as the variance of the pairwise Euclidean distance between the data. λ is the weight for the regularization term in our derived generalization bound. As explained in Section B.1 of the appendix, λ plays the same role as the weight in the regularization term of SVMs or Kernel Machines. Following the common practice in the literature of SVM or Kernel Machines, λ can be tuned by Cross-Validation (CV). While this is an unsupervised learning task and these is no labeled data for CV, we still developed a well-motivated CV procedure. Following the practice in (Mairal et al., 2012), we randomly sampled 10% of the given data as the validation data, then perform CDSK on the validation data. The best λ is chosen among the discrete values between $[0.05, 05]$ with a step of 0.05 which minimizes the average entropy of the obtained embedding matrix $\mathbf{Y} \in \mathbb{R}^{n\times c}$ by Algorithm 1, where the average entropy is compute as $\frac{1}{n}\sum_{i=1}^{n} \mathrm{entropy}(\mathrm{softmax}(\mathbf{Y}^i))$. This is because we would like to choose λ which renders the most confident clustering embedding. We perform CDSK on all the datasets using this tuning strategy and observe improved performance as shown in the above two tables.

Computational Complexity. Suppose the optimization of CDSK comprises M iterations of coordinate descent, then the overall time complexity of CDSK is $\mathcal{O}(Mcn^2 + Mn^{2.1})$. M is set to 20 throughout all the experiments. For clustering methods involving random operations, the average performance over 10 runs is reported.

Table 1: Clustering results on CMU PIE, CMU Multi-PIE which contains the facial images captured in four sessions (S1 to S4) and UMIST face dataset. c in the left column is the cluster number, i.e. the first c clusters of the entire data are used for clustering.

Methods \ Dataset	Yale-B	Ionosphere	Georgia Face	COIL-20	COIL-100	CMU PIE	UMIST Face
K-Means	0.09(0.13)	0.71(0.13)	0.50(0.69)	0.65(0.76)	0.49(0.75)	0.08(0.19)	0.42(0.64)
SC	0.11(0.15)	0.74(0.22)	0.52(0.70)	0.43(0.62)	0.28(0.59)	0.07(0.18)	0.42(0.61)
ℓ^1-Graph (Elhamifar & Vidal, 2013)	0.79(0.78)	0.51(0.12)	0.54(0.70)	0.79(0.91)	0.53(0.80)	0.23(0.34)	0.44(0.65)
SMCE (Elhamifar & Vidal, 2011)	0.34(0.39)	0.68(0.09)	**0.60(0.74)**	0.88(0.88)	0.56(0.81)	0.16(0.34)	0.45(0.66)
Lap-ℓ^1-Graph (Yang et al., 2014b)	0.79(0.78)	0.50(0.09)	0.58(0.73)	0.79(0.91)	0.56(0.81)	0.30(**0.51**)	0.50(0.69)
RAG (Zhu et al., 2014)	0.13(0.19)	0.70(0.11)	0.17(0.38)	0.50(0.64)	0.58(0.81)	0.14(0.34)	0.26(0.28)
MMC (Xu et al., 2004)	0.71(0.69)	0.75(0.21)	0.42(0.58)	0.80(0.89)	0.61(0.63)	0.22(0.30)	0.51(0.56)
BMMC (Chen et al., 2014)	0.65(0.63)	0.70(0.15)	0.34(0.41)	0.82(0.93)	0.64(0.69)	0.18(0.23)	0.55(0.61)
RIM (Gomes et al., 2010)	0.62(0.74)	0.59(0.08)	0.39(0.56)	0.77(0.82)	0.71(0.79)	0.26(0.34)	0.40(0.53)
RatioRF (Bicego et al., 2021)	0.39(0.53)	0.62(0.05)	0.18(0.40)	0.65(0.75)	0.36(0.64)	0.15(0.36)	0.29(0.34)
CDSK (**Ours**)	**0.83(0.86)**	**0.76(0.25)**	**0.60(0.74)**	**0.93(0.97)**	**0.78(0.92)**	**0.32**(0.50)	**0.67(0.80)**

Table 2: Clustering results on CMU PIE, CMU Multi-PIE which contains the facial images captured in four sessions (S1 to S4) and UMIST face dataset. c in the left column is the cluster number, i.e. the first c clusters of the entire data are used for clustering.

Methods \ Dataset	MPIE S1	MPIE S2	MPIE S3	MPIE S4	MNIST	CIFAR-10	Mini-ImageNet
KM	0.12(0.50)	0.13(0.48)	0.13(0.48)	0.13(0.49)	0.52(0.47)	0.19(0.06)	0.27(0.33)
SC	0.13(0.53)	0.14(0.51)	0.14(0.52)	0.15(0.53)	0.38(0.36)	0.21(0.04)	0.29(0.35)
ℓ^1-Graph (Elhamifar & Vidal, 2013)	0.59(0.77)	0.70(0.81)	0.63(0.79)	0.68(0.81)	0.57(0.61)	0.28(0.24)	0.28(0.37)
SMCE (Elhamifar & Vidal, 2011)	0.17(0.55)	0.19(0.53)	0.19(0.52)	0.18(0.53)	0.65(0.67)	0.31(0.30)	0.29(0.37)
Lap-ℓ^1-Graph (Yang et al., 2014b)	0.59(0.77)	0.70(0.81)	0.63(0.79)	0.68(0.81)	0.56(0.60)	0.29(0.30)	0.29(0.37)
RAG (Zhu et al., 2014)	0.34(0.75)	0.30(0.69)	0.31(0.68)	0.29(0.70)	0.59(0.51)	0.22(0.10)	0.18(0.33)
MMC (Xu et al., 2004)	0.49(0.58)	0.51(0.60)	0.53(0.65)	0.50(0.61)	0.64(0.60)	0.31(0.28)	0.19(0.34)
BMMC (Chen et al., 2014)	0.40(0.51)	0.44(0.59)	0.45(0.61)	0.49(0.66)	0.66(0.69)	0.29(0.26)	0.16(0.32)
RIM (Gomes et al., 2010)	0.50(0.63)	0.52(0.68)	0.55(0.71)	0.51(0.67)	0.54(0.62)	0.20(0.25)	0.17(0.38)
RatioRF (Bicego et al., 2021)	0.54(**0.85**)	0.55(0.86)	0.64(0.86)	0.62(0.86)	0.48(0.39)	0.20(0.09)	0.26(0.38)
CDSK (**Ours**)	**0.66(0.85)**	**0.72(0.88)**	**0.68(0.87)**	**0.73(0.89)**	**0.76(0.75)**	**0.46(0.39)**	**0.31(0.41)**

5 Detailed Discussion about Significance of CDSK over Existing Discriminative and Similarity-Based Clustering Methods

Effective data similarity highly depends on the underlying probabilistic distribution and geometric structure of the data, and these two characteristics leads to "data-driven" similarity, such as (Zhu et al., 2014; Bicego et al., 2021; Ng et al., 2001; Shental et al., 2003; Hartigan & Wong, 1979;

Schölkopf et al., 1998) and similarity based on geometric structure of the data, such as the subspace structure (Sparse Subspace Clustering, or SSC in (Elhamifar & Vidal, 2013)). Note that the sparse graph method, ℓ^1-Graph (Yan & Wang, 2009), has the same formulation as SSC. Most existing clustering methods based on data-driven or geometric structure-driven similarity suffer from a common deficiency, that is, the similarity is not explicitly optimized for the purpose of separating underlying clusters. In particular, the Random Forest-based similarity (Zhu et al., 2014; Bicego et al., 2021) is extracted from features in decision trees. Previous works about subspace-based similarity (Yan & Wang, 2009; Elhamifar & Vidal, 2013) try to make sure that only data points lying on or close to the same subspace have nonzero similarity, so that data points from the same subspace can form a cluster. However, it is not guaranteed that features in the decision trees are discriminative enough to separate clusters, because the candidate data partition (or candidate cluster labels) do not participate in the feature or similarity extraction process. Note that synthetically generated negative class are suggested in (Zhu et al., 2014; Bicego et al., 2021) to train unsupervised random forest, however, the synthetic labels are not for the original data. Moreover, it is well known that the existing subspace learning methods only obtain reliable subspace-based similarity with restrictive geometric assumptions on the data and the underlying subspaces, such as large principal angle between intersecting subspaces (Soltanolkotabi & Candes, 2012; Elhamifar & Vidal, 2013).

Therefore, it is particularly important to derive similarity for clustering which meets two requirements: (1) discriminative measure with information such as cluster partition is used to derive such similarity so as to achieve compelling clustering performance; (2) it requires less restrictive assumptions on the geometric structure of the data than current geometric structure-based similarity learning methods, such as subspace clustering (Yan & Wang, 2009; Elhamifar & Vidal, 2013).

Significance. The proposed discriminative similarity of this paper meets these two requirements. First, the discriminative similarity is derived by the generalization error bound associated with candidate cluster labeling, and minimizing the objective function (10) renders a joint optimization of discriminative similarity and candidate cluster labeling in a way such that the similarity-based classifier has small generalization error bound. Second, our framework only assumes a mild classification model in Definition 2.1, which only requires an unknown joint distribution over data and its labels. In this way, the restrictive geometric assumptions are avoided in our method. Compared to the existing discriminative clustering methods, such as MMC (Xu et al., 2004), BMMC (Chen et al., 2014), RIM (Gomes et al., 2010), and the other discriminative clustering methods such as (Huang et al., 2015; Nguyen et al., 2017), the optimization problem of CDSK with discriminative similarity-based formulation is much easier to solve and it enjoys convexity and efficiency in each iteration of coordinate descent described in Algorithm 1. In particular, as mentioned in Section D of the appendix, the first step (11) of each iteration can be solved by efficient SVD or other randomized large-scale SVD methods, and the second step (12) of each iteration can be solved by efficient SMO (Platt, 1998). Moreover, the optimization problems in these two steps are either convex or having closed-form solution. In contrast, MMC requires expensive semidefinite programming. RIM has to solve a nonconvex optimization problem and its formulation does not guarantee that the trained multi-class kernelized logistic regression has low classification error on candidate labeling, which explains why it has inferior performance compared to our method. The discriminative Extreme Learning Machine (Huang et al., 2015) trains ELM using labels produced by a simple clustering method such as K-means, and the potentially poor cluster labels by the simple clustering method can easily result in unsatisfactory performance of this method. The discriminative Bayesian nonparametric clustering (Nguyen et al., 2017) and BMMC (Chen et al., 2014) require extra efforts of sampling hidden variables and tuning hyperparameters to generate the desirable number of clusters (or model selection), which could reduce the effect of discriminative measures used in these Bayesian nonparametric methods.

6 Conclusion

We propose a new clustering framework termed Clustering by Discriminative Similarity (CDS). CDS searches for the optimal partition of the data where the associated unsupervised classifier has minimum generalization error bound. Under this framework, discriminative similarity is induced by the generalization error bound for unsupervised similarity-based classifier, and CDS minimizes discriminative similarity between different clusters. It is also proved that the discriminative similarity can be induced from kernel density classification. Based on CDS, we propose a new clustering method, CDS via unsupervised kernel classification (CDSK), and demonstrate its effectiveness in data clustering.

REFERENCES

Maria-Florina Balcan, Avrim Blum, and Nathan Srebro. A theory of learning with similarity functions. *Mach. Learn.*, 72(1-2):89–112, 2008.

Peter L. Bartlett and Shahar Mendelson. Rademacher and gaussian complexities: Risk bounds and structural results. *J. Mach. Learn. Res.*, 3:463–482, March 2003.

Manuele Bicego, Ferdinando Cicalese, and Antonella Mensi. Ratiorf: a novel measure for random forest clustering based on the tversky's ratio model. *IEEE Transactions on Knowledge and Data Engineering*, pp. 1–1, 2021.

Changyou Chen, Jun Zhu, and Xinhua Zhang. Robust bayesian max-margin clustering. In Zoubin Ghahramani, Max Welling, Corinna Cortes, Neil D. Lawrence, and Kilian Q. Weinberger (eds.), *Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, December 8-13 2014, Montreal, Quebec, Canada*, pp. 532–540, 2014.

Corinna Cortes, Mehryar Mohri, and Afshin Rostamizadeh. Multi-class classification with maximum margin multiple kernel. In *Proceedings of the 30th International Conference on Machine Learning (ICML)*, pp. 46–54, Atlanta, GA, 2013.

Ehsan Elhamifar and René Vidal. Sparse manifold clustering and embedding. In *Advances in Neural Information Processing Systems (NIPS)*, pp. 55–63, Granada, Spain, 2011.

Ehsan Elhamifar and René Vidal. Sparse subspace clustering: Algorithm, theory, and applications. *IEEE Trans. Pattern Anal. Mach. Intell.*, 35(11):2765–2781, 2013.

Evarist Gine and Joel Zinn. Some limit theorems for empirical processes. *Ann. Probab.*, 12(4): 929–989, 11 1984.

Mark Girolami and Chao He. Probability density estimation from optimally condensed data samples. *IEEE Trans. Pattern Anal. Mach. Intell.*, 25(10):1253–1264, 2003.

Ryan Gomes, Andreas Krause, and Pietro Perona. Discriminative clustering by regularized information maximization. In *Advances in Neural Information Processing Systems (NIPS)*, pp. 775–783, Vancouver, Canada, 2010.

Ralph Gross, Iain A. Matthews, Jeffrey F. Cohn, Takeo Kanade, and Simon Baker. Multi-pie. *Image Vis. Comput.*, 28(5):807–813, 2010.

J. A. Hartigan and M. A. Wong. A K-means clustering algorithm. *Applied Statistics*, 28:100–108, 1979.

Gao Huang, Tianchi Liu, Yan Yang, Zhiping Lin, Shiji Song, and Cheng Wu. Discriminative clustering via extreme learning machine. *Neural Networks*, 70:1–8, 2015.

Masayuki Karasuyama and Hiroshi Mamitsuka. Manifold-based similarity adaptation for label propagation. In *Advances in Neural Information Processing Systems (NIPS)*, pp. 1547–1555, Lake Tahoe, NV, 2013.

Zohar Shay Karnin, Edo Liberty, Shachar Lovett, Roy Schwartz, and Omri Weinstein. Unsupervised SVMs: On the complexity of the furthest hyperplane problem. In *Proceedings of the 25th Annual Conference on Learning Theory (COLT)*, pp. 2.1–2.17, Edinburgh, Scotland, 2012.

JooSeuk Kim and Clayton D. Scott. Performance analysis for l_2 kernel classification. In *Advances in Neural Information Processing Systems (NIPS)*, pp. 833–840, Vancouver, Canada, 2008.

JooSeuk Kim and Clayton D. Scott. Robust kernel density estimation. *J. Mach. Learn. Res.*, 13: 2529–2565, 2012.

V. Koltchinskii and D. Panchenko. Empirical margin distributions and bounding the generalization error of combined classifiers. *Ann. Statist.*, 30(1):1–50, 02 2002. doi: 10.1214/aos/1015362183.

Vladimir Koltchinskii. Rademacher penalties and structural risk minimization. *IEEE Trans. Inf. Theory*, 47(5):1902–1914, 2001.

James R. Lee, Shayan Oveis Gharan, and Luca Trevisan. Multi-way spectral partitioning and higher-order cheeger inequalities. In Howard J. Karloff and Toniann Pitassi (eds.), *Proceedings of the 44th Symposium on Theory of Computing Conference, STOC 2012, New York, NY, USA, May 19 - 22, 2012*, pp. 1117–1130. ACM, 2012.

Gaëlle Loosli, Stéphane Canu, and Cheng Soon Ong. Learning SVM in kreĭn spaces. *IEEE Trans. Pattern Anal. Mach. Intell.*, 38(6):1204–1216, 2016.

Ravi Sastry Ganti Mahapatruni and Alexander G. Gray. CAKE: convex adaptive kernel density estimation. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics (AISTATS)*, pp. 498–506, Fort Lauderdale, FL, 2011.

Julien Mairal, Francis Bach, and Jean Ponce. Task-driven dictionary learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(4):791–804, 2012.

Xavier Mary. Hilbertian subspaces, subdualities and applications. *Ph.D. Dissertation, Institut National des Sciences Appliquees Rouen*, 2003.

Andrew Y. Ng, Michael I. Jordan, and Yair Weiss. On spectral clustering: Analysis and an algorithm. In *Advances in Neural Information Processing Systems (NIPS)*, pp. 849–856, Vancouver, Canada], 2001.

Vu Nguyen, Dinh Q. Phung, Trung Le, and Hung Bui. Discriminative bayesian nonparametric clustering. In Carles Sierra (ed.), *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI 2017, Melbourne, Australia, August 19-25, 2017*, pp. 2550–2556. ijcai.org, 2017.

Cheng Soon Ong, Xavier Mary, Stéphane Canu, and Alexander J. Smola. Learning with non-positive kernels. In *Proceedings of the Twenty-first International Conference on Machine Learning (ICML)*, Banff, Alberta, Canada, 2004.

John Platt. Sequential minimal optimization: A fast algorithm for training support vector machines. Technical report, 1998.

Bernhard Schölkopf, Alexander Smola, and Klaus-Robert Müller. Nonlinear component analysis as a kernel eigenvalue problem. *Neural Comput.*, 10(5):1299–1319, July 1998.

Noam Shental, Assaf Zomet, Tomer Hertz, and Yair Weiss. Pairwise clustering and graphical models. In *Advances in Neural Information Processing Systems (NIPS)*, pp. 185–192, Vancouver and Whistler, Canada, 2003.

Mahdi Soltanolkotabi and Emmanuel J. Candes. A geometric analysis of subspace clustering with outliers. *Ann. Statist.*, 40(4):2195–2238, 08 2012.

Oriol Vinyals, Charles Blundell, Tim Lillicrap, Koray Kavukcuoglu, and Daan Wierstra. Matching networks for one shot learning. In Daniel D. Lee, Masashi Sugiyama, Ulrike von Luxburg, Isabelle Guyon, and Roman Garnett (eds.), *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain*, pp. 3630–3638, 2016.

Linli Xu, James Neufeld, Bryce Larson, and Dale Schuurmans. Maximum margin clustering. In *Advances in Neural Information Processing Systems (NIPS)*, pp. 1537–1544, Vancouver, Canada], 2004.

Shuicheng Yan and Huan Wang. Semi-supervised learning by sparse representation. In *Proceedings of the SIAM International Conference on Data Mining (SDM)*, pp. 792–801, Sparks, NV, 2009.

Yingzhen Yang, Feng Liang, Shuicheng Yan, Zhangyang Wang, and Thomas S. Huang. On a theory of nonparametric pairwise similarity for clustering: Connecting clustering to classification. In *Advances in Neural Information Processing Systems (NIPS)*, pp. 145–153, Montreal, Canada, 2014a.

Yingzhen Yang, Zhangyang Wang, Jianchao Yang, Jiangping Wang, Shiyu Chang, and Thomas S. Huang. Data clustering by laplacian regularized l1-graph. In *Proceedings of the Twenty-Eighth AAAI Conference on Artificial Intelligence (AAAI)*, pp. 3148–3149, Québec City, Canada, 2014b.

Xin Zheng, Deng Cai, Xiaofei He, Wei-Ying Ma, and Xueyin Lin. Locality preserving clustering for image database. In *Proceedings of the 12th ACM International Conference on Multimedia (MM)*, pp. 885–891, New York, NY, 2004.

Xiatian Zhu, Chen Change Loy, and Shaogang Gong. Constructing robust affinity graphs for spectral clustering. In *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1450–1457, 2014.

A Connection to Kernel Density Classification

In this section, we show that the discriminative similarity (8) can also be induced from kernel density classification with varying weights on the data, and binary classification is considered in this section. For any classification model $M_{\mathcal{Y}} = (\mathcal{S}, f)$ with hypothetical labeling $\mathcal{Y}$ and the labeled data $\mathcal{S} = \{\mathbf{x}_i, y_i\}_{i=1}^n$, suppose the joint distribution P_{XY} over $\mathcal{X} \times \{1, 2\}$ has probabilistic density function $p(\mathbf{x}, y)$. Let P_X be the induced marginal distribution over the data with probabilistic density function $p(\mathbf{x})$. Robust kernel density estimation methods (Girolami & He, 2003; Kim & Scott, 2008; Mahapatruni & Gray, 2011; Kim & Scott, 2012) suggest the following kernel density estimator where the kernel contributions of different data points are reflected by different nonnegative weights that sum up to 1:

$$\widehat{p}(\mathbf{x}) = \tau_0 \sum_{i=1}^{n} \boldsymbol{\alpha}_i K_\tau(\mathbf{x} - \mathbf{x}_i), \mathbf{1}^\top \boldsymbol{\alpha} = 1, \boldsymbol{\alpha} \geq 0, \tag{13}$$

where $\tau_0 = \frac{1}{(2\pi)^{d/2} h^d}$. Based on (13), it is straightforward to obtain the following kernel density estimator of the density function $p(\mathbf{x}, y)$:

$$\widehat{p}(\mathbf{x}, y) = \tau_0 \sum_{i: y_i = y} \boldsymbol{\alpha}_i K_\tau(\mathbf{x} - \mathbf{x}_i). \tag{14}$$

Kernel density classifier is learnt from the labeled data $\mathcal{S}$ and constructed by kernel density estimators (14). Kernel density classifier resembles the Bayes classifier, and it classifies the test data $\mathbf{x}$ based on the conditional label distribution $P(Y|X = \mathbf{x})$, or equivalently, $\mathbf{x}$ is assigned to class 1 if $\widehat{p}(\mathbf{x}, 1) - \widehat{p}(\mathbf{x}, 2) \geq 0$, otherwise it is assigned to class 2. Intuitively, it is preferred that the decision function $\widehat{r}(\mathbf{x}, \boldsymbol{\alpha}) = \widehat{p}(\mathbf{x}, 1) - \widehat{p}(\mathbf{x}, 2)$ is close to the true Bayes decision function $r = p(\mathbf{x}, 1) - p(\mathbf{x}, 2)$. Girolami & He (2003); Kim & Scott (2008) propose to use Integrated Squared Error (ISE) as the metric to measure the distance between the kernel density estimators and their true counterparts, and the oracle inequality is obtained that relates the performance of the L_2 classifier in Kim & Scott (2008) to the best possible performance of kernel density classifier in the same category. ISE is adopted in our analysis of kernel density classification, and the ISE between the decision function $\widehat{r}$ and the true Bayes decision function r is defined as

$$\mathrm{ISE}(\widehat{r}, r) = \|\widehat{r} - r\|_{\mathcal{L}_2}^2 = \int_{\mathbb{R}^d} (\widehat{r} - r)^2 dx. \tag{15}$$

The upper bound for the ISE $\mathrm{ISE}(r, \widehat{r})$ also induces discriminative similarity between the data from different classes, which is presented in the following theorem.

Theorem A.1. Let $n_1 = \sum_{i=1} \mathbb{I}_{y_i=1}$ and $n_2 = \sum_{i=1} \mathbb{I}_{y_i=2}$. With probability at least $1 - 2n_2 \exp\big(-2(n-1)\varepsilon^2\big) - 2n \exp\big(-2n\varepsilon^2\big)$ over the labeled data $\mathcal{S}$, the ISE between the decision function $\widehat{r}(\mathbf{x}, \boldsymbol{\alpha})$ and the true Bayes decision function $r(\mathbf{x})$ satisfies

$$\mathrm{ISE}(\widehat{r}, r) \leq \frac{\tau_0}{n} \widehat{\mathrm{ISE}}(\widehat{r}, r) + \tau_1 K(\boldsymbol{\alpha}) + 2\tau_0 \Big(\frac{1}{n-1} + \varepsilon \Big), \tag{16}$$

where

$$\widehat{\mathrm{ISE}}(\widehat{r}, r) = 4 \sum_{1 \leq i < j \leq n} (\boldsymbol{\alpha}_i + \boldsymbol{\alpha}_j) K_\tau(\mathbf{x}_i - \mathbf{x}_j) \mathbb{I}_{y_i \neq y_j} - \sum_{i,j=1}^{n} (\boldsymbol{\alpha}_i + \boldsymbol{\alpha}_j) K_\tau(\mathbf{x}_i - \mathbf{x}_j), \tag{17}$$

$$K(\boldsymbol{\alpha}) = \boldsymbol{\alpha}^\top(\mathbf{K}_{\sqrt{2}h})\boldsymbol{\alpha} - 4 \sum_{1\le i<j\le n} \boldsymbol{\alpha}_i\boldsymbol{\alpha}_j K_{\sqrt{2}h}(\mathbf{x}_i - \mathbf{x}_j)\mathbb{1}_{y_i\neq y_j}, \tag{18}$$

and $\mathbf{K}_{\sqrt{2}h}$ is the gram matrix evaluated on the data $\{\mathbf{x}_i\}_{i=1}^n$ with the kernel $K_{\sqrt{2}h}$.

Let $\lambda_1 > 0$ be a weighting parameter, then the cost function $\widehat{\mathrm{ISE}} + \lambda_1 K(\boldsymbol{\alpha})$, designed according to the empirical term $\mathrm{ISE}(\widehat{r}, r)$ and the regularization term $K(\boldsymbol{\alpha})$ in the ISE error bound (16), can be expressed as

$$\widehat{\mathrm{ISE}} + \lambda_1 K(\boldsymbol{\alpha}) \le \sum_{1\le i<j\le n} S_{ij}^{\mathrm{ise}}\mathbb{1}_{y_i\neq y_j} - \sum_{i,j=1}^{n}(\boldsymbol{\alpha}_i + \boldsymbol{\alpha}_j)K_\tau(\mathbf{x}_i - \mathbf{x}_j) + \lambda_1\boldsymbol{\alpha}^\top\mathbf{K}_{\sqrt{2}h}\boldsymbol{\alpha},$$

where the first term is comprised of sum of similarity between data from different classes with similarity $S_{ij}^{\mathrm{ise}} = 4(\boldsymbol{\alpha}_i + \boldsymbol{\alpha}_j - \lambda_1\boldsymbol{\alpha}_i\boldsymbol{\alpha}_j)K_\tau(\mathbf{x}_i - \mathbf{x}_j)$, and S_{ij}^{ise} is the discriminative similarity induced by the ISE bound for kernel density classification. Note that S_{ij}^{ise} has the same form as the discriminative similarity S_{ij}^K (8) induced by our CDS model, up to a scaling constant and the choice of the balancing parameter λ. The proof of Theorem A.1 is deferred to Section E.3.

B More Explanation about the Theoretical Results in Section 3.1

B.1 Theoretical Significance of the Bound (4)

To the best of our knowledge, our generalization error bound (4) is the first principled result about generalization error bound for general similarity-based classifier with strong connection to the established generalization error bound for Support Vector Machines (SVMs) or Kernel Machines.

We now explain this claim. When the similarity function S is a PSD kernel function, we have $S^- \equiv 0, S = S^+$ as explained in Remark 3.3. As a reminder, S is the general similarity function used in the similarity-based classification, and S^+, S^- are PSD kernel functions, and it is proved that S can be decomposed by $S = S^+ - S^-$ under the mild conditions of Lemma 3.1. It follows that we can set $\Omega^-(\boldsymbol{\alpha}) = 0$ and $B^- = 0$. Plugging $B^- = 0$ in the derived generalization error bound for the general similarity-based classification (4), we have

$$\mathrm{Prob}\left[Y \neq f_S(X)\right] \le \widehat{R}_n(f_S) + \frac{8R(2c-1)cB^+}{\gamma\sqrt{n}} + \left(\frac{16c(2c-1)B^+R^2}{\gamma} + 1\right)\sqrt{\frac{\log\frac{4}{\delta}}{2n}}. \tag{19}$$

According to its definition, $\Omega^+(\boldsymbol{\alpha}) = \sum_{y=1}^{c} {\boldsymbol{\alpha}^{(y)}}^\top\mathbf{S}\boldsymbol{\alpha}^{(y)}$ because $\mathbf{S} = \mathbf{S}^+$. We define $B := B^+$. Because $\Omega^+(\boldsymbol{\alpha}) \le {B^+}^2$ as mentioned in Theorem 3.2, B satisfies $\sum_{y=1}^{c} {\boldsymbol{\alpha}^{(y)}}^\top\mathbf{S}\boldsymbol{\alpha}^{(y)} \le B^2$. As a result, when S is a PSD kernel function, inequality (19) becomes

$$\mathrm{Prob}\left[Y \neq f_S(X)\right] \le \widehat{R}_n(f_S) + \frac{8R(2c-1)cB}{\gamma\sqrt{n}} + \left(\frac{16c(2c-1)R^2B}{\gamma} + 1\right)\sqrt{\frac{\log\frac{4}{\delta}}{2n}}, \tag{20}$$

with $\sum_{y=1}^{c} {\boldsymbol{\alpha}^{(y)}}^\top\mathbf{S}\boldsymbol{\alpha}^{(y)} \le B^2$.

Note that the bound (20) is in fact the generalization error bound for supervised learning when using S as the similarity function in the similarity-based classification. At the end of this subsection, we provide a lemma proving that $\sum_{y=1}^{c} {\boldsymbol{\alpha}^{(y)}}^\top\mathbf{S}\boldsymbol{\alpha}^{(y)} \le B^2 \Rightarrow \boldsymbol{\alpha}^\top\mathbf{S}\boldsymbol{\alpha} \le cB^2$, where c is the number of classes.

Now we compare the generalization error bound (20) to the established generalization error bound for Kernel Machines in (Bartlett & Mendelson, 2003, Theorem 21) for the case that $c = 2$ with

notations adapted to our analysis. The bound in (Bartlett & Mendelson, 2003, Theorem 21) is for binary classification, which is presented as follows:

$$\text{Prob}\left[Y \neq f_S(X)\right] \leq \widehat{R}_n(f_S) + \frac{4\sqrt{R}B}{\gamma\sqrt{n}} + \left(\frac{8}{\gamma}+1\right)\sqrt{\frac{\log 4/\delta}{2n}}, \boldsymbol{\alpha}^\top \mathbf{S}\boldsymbol{\alpha} \leq B^2. \tag{21}$$

Comparing our generalization error bound (2) with $c = 2$ to the max-margin generalization error bound (21), it can be easily seen that the two bounds are equivalent up to a constant scaling factor. In fact, our bound (2) is more general which handles multi-class classification.

Lemma B.1. When S is a PSD kernel, then $\sum_{y=1}^{c} {\boldsymbol{\alpha}^{(y)}}^\top \mathbf{S}\boldsymbol{\alpha}^{(y)} \leq B^2 \Rightarrow \boldsymbol{\alpha}^\top \mathbf{S}\boldsymbol{\alpha} \leq cB^2$.

Proof. Let $\mathcal{H}$ be the Reproducing Kernel Hilbert Space associated with the PSD kernel function S, and $\mathcal{H}$ is also called the feature space associated with $\mathbf{S}$. We use $\langle \cdot, \cdot \rangle_{\mathcal{H}}$ to denote the inner product in the feature space $\mathcal{H}$. Then we have $S(\mathbf{x}_i, \mathbf{x}_j) = \mathbf{S}_{ij} = \langle \Phi(\mathbf{x}_i), \Phi(\mathbf{x}_j) \rangle_{\mathcal{H}}$ where Φ is the feature mapping associated with $\mathcal{H}$. Because $\boldsymbol{\alpha} = \sum_{y=1}^{c} \boldsymbol{\alpha}^y$, it can be verified that

$\boldsymbol{\alpha}^\top \mathbf{S}\boldsymbol{\alpha} = \sum_{i=1}^{n}\sum_{j=1}^{n} \boldsymbol{\alpha}_i \boldsymbol{\alpha}_j \mathbf{S}_{ij} = \langle \sum_{i=1}^{n} \boldsymbol{\alpha}_i \Phi(\mathbf{x}_i), \sum_{j=1}^{n} \boldsymbol{\alpha}_j \Phi(\mathbf{x}_j) \rangle \leq c \sum_{y=1}^{c} \langle e_y, e_y \rangle_{\mathcal{H}}$, where $e_y = \sum_{i:\, y_i = y} \boldsymbol{\alpha}_i^y \Phi(\mathbf{x}_i)$. It follows that $\boldsymbol{\alpha}^\top \mathbf{S}\boldsymbol{\alpha} \leq c \sum_{y=1}^{c} {\boldsymbol{\alpha}^{(y)}}^\top \mathbf{S}\boldsymbol{\alpha}^{(y)} \leq cB^2$. □

B.2 Tightness of the Bound

Note that the generalization error $\Pr\left[Y \neq f(X)\right]$ is bounded by $\widehat{R}_n(f) = \frac{1}{n}\sum_{i=1}^{n} \Phi\left(\frac{h_S(\mathbf{x}_i, y_i) - \sum_{y \neq y_i} h_S(\mathbf{x}_i, y)}{\gamma}\right)$ in Theorem 3.2. The underlying principle behind this bound and all such bounds in the statistical machine learning literature such as Bartlett & Mendelson (2003) is the following property about empirical process (adjusted using our notations):

$$\mathfrak{R}(\mathcal{H}) \leq \mathbb{E}_{\mathbf{x},y} \sup_{h \in \mathcal{H}} |\mathbb{E}_{\mathbf{x},y} \widehat{R}_n(f) - R_n(f)| \leq 2\mathfrak{R}(\mathcal{H}), \tag{22}$$

where $\mathbb{E}_{\mathbf{x},y}$ indicates expectation with respect to random couple $(\mathbf{x}, y) \sim P_{XY}$, and P_{XY} is a joint distribution in a discriminative model $M_{\mathcal{Y}} = (\mathcal{S}, f)$. (22) is introduced in the classical properties of empirical process in Gine & Zinn (1984). By Lemma E.2 of this paper, with probability at least $1 - \delta$ over the data $\{\mathbf{x}_i\}_{i=1}^n \overset{\text{i.i.d.}}{\sim} P_{XY}$,

$$\mathfrak{R}(\mathcal{H}) \leq \frac{(2c-1)c}{\sqrt{n}} B + \sqrt{2}Bc(2c-1)\sqrt{\frac{\ln \frac{2}{\delta}}{2n}} = \mathcal{O}(\frac{B}{\sqrt{n}}) \tag{23}$$

for some constant B. It follows from (22), (23), and concentration inequality (such as McDiarmid's inequality) that for each sufficiently large n, with large probability, $\sup_{h \in \mathcal{H}} |\mathbb{E}_{\mathbf{x},y} \widehat{R}_n(f) - R_n(f)|$ is less than $\mathcal{O}(\frac{B}{\sqrt{n}})$. Therefore, we can bound the expectation of the empirical loss, i.e. $\mathbb{E}_{\mathbf{x},y} \widehat{R}_n(f)$, tightly using the empirical loss $R_n(f)$ uniformly over the function space $\mathcal{H}$.

C Theoretical Properties of the Coordinate Descent Algorithm in Section 4

In this subsection, we give a detailed explanation about the theoretical properties of the coordinate descent algorithm presented in Section 4. We first explain how the objective function of CDSK (10) is connected to the objective function (9) developed in our theoretical analysis. It should be emphasized that (9) cannot be directly used for data clustering since it cannot avoids the trivial solution where all the data are in a single cluster. We adopt the broadly used formulation of normalized

cut and use $\sum_{k=1}^{c} \frac{\mathrm{cut}(\mathbf{A}_k, \bar{\mathbf{A}}_k)}{\mathrm{vol}(\mathbf{A}_k)}$ to replace $\sum_{i<j} S_{ij}^K \mathbb{1}_{y_i \neq y_j}$ in (9), leading to the following optimization problem:

$$\min_{\boldsymbol{\alpha} \in \Lambda, \mathcal{Y} = \{y_i\}_{i=1}^n} \sum_{k=1}^{c} \frac{\mathrm{cut}(\mathbf{A}_k, \bar{\mathbf{A}}_k)}{\mathrm{vol}(\mathbf{A}_k)} - \sum_{i,j=1}^{n} \frac{\boldsymbol{\alpha}_i + \boldsymbol{\alpha}_j}{2} K_\tau(\mathbf{x}_i - \mathbf{x}_j) + \lambda \boldsymbol{\alpha}^\top \mathbf{K} \boldsymbol{\alpha} \triangleq \bar{Q}(\boldsymbol{\alpha}, \mathcal{Y}), \quad (9')$$

where $\{\mathbf{A}\}_{k=1}^c$ are c data clusters according to the cluster labels $\{y_i\}$, $\bar{\mathbf{A}}_k$ is the complement of $\mathbf{A}_k$, $\mathrm{cut}(\mathbf{A}, \mathbf{B}) = \sum_{\mathbf{x}_i \in \mathbf{A}, \mathbf{x}_j \in \mathbf{B}} S_{ij}^K$, $\mathrm{vol}(\mathbf{A}) = \sum_{\mathbf{x}_i \in \mathbf{A}, 1 \leq j \leq n} S_{ij}^K$. We have the following theorem, which can be derived based on Theorem 4.1 in the work of multi-way Cheeger inequalities (Lee et al., 2012).

Theorem C.1. $\min_{\boldsymbol{\alpha} \in \Lambda, \mathcal{Y}} \sum_{k=1}^{c} \frac{\mathrm{cut}(\mathbf{A}_k, \bar{\mathbf{A}}_k)}{\mathrm{vol}(\mathbf{A}_k)} \lesssim \sum_{t=1}^{c} \sigma_t(\mathbf{L}^{\mathrm{nor}})$, where $\mathbf{L}^{\mathrm{nor}}$ is the normalized graph Laplacian $\mathbf{L}^{\mathrm{nor}} = (\mathbf{D}^K)^{-1/2} \mathbf{L}^K (\mathbf{D}^K)^{-1/2}$, $a \lesssim b$ indicates $a < Cb$ for some constant C and $\sigma_t(\cdot)$ indicates the t-th smallest singular value of a matrix.

Based on Theorem C.1, we resort to solve the following more tractable problem, that is,

$$\min_{\boldsymbol{\alpha} \in \Lambda} \sum_{t=1}^{k} \sigma_t(\mathbf{L}^{\mathrm{nor}}) - \sum_{i,j=1}^{n} \frac{\boldsymbol{\alpha}_i + \boldsymbol{\alpha}_j}{2} K_\tau(\mathbf{x}_i - \mathbf{x}_j) + \lambda \boldsymbol{\alpha}^\top \mathbf{K} \boldsymbol{\alpha} \triangleq Q(\boldsymbol{\alpha}), \quad (9'')$$

because Q (9") is an upper bound for $\bar{Q}$ (9') up to a constant scaling factor. It can be verified that problem (10) is equivalent to (9"), and (9") is the underlying optimization problem for data clustering in Section 4. The following proposition shows that the iterative coordinate descent algorithm in Section 4 reduces the value of Q at each iteration.

Proposition C.2. The coordinate descent algorithm for problem (10) reduces the value of the objective function $Q(\boldsymbol{\alpha})$ at each iteration.

Proof. Let $Q'(\boldsymbol{\alpha}, \mathbf{Y}) \triangleq \mathrm{Tr}(\mathbf{Y}^\top \mathbf{L}^K \mathbf{Y}) - \sum_{i,j=1}^{n} \frac{\boldsymbol{\alpha}_i + \boldsymbol{\alpha}_j}{2} K_\tau(\mathbf{x}_i - \mathbf{x}_j) + \lambda \boldsymbol{\alpha}^\top \mathbf{K} \boldsymbol{\alpha}$, and we use superscript to denote the iteration number of coordinate descent. At iteration m, after solving the subproblems (11) and (12), we have $Q'(\boldsymbol{\alpha}^{(m)}, \mathbf{Y}^{(m)}) = Q(\boldsymbol{\alpha}^{(m)})$. At iteration $m+1$, by solving the subproblems (4) and (3) in order again, we have $Q'(\boldsymbol{\alpha}^{(m+1)}, \mathbf{Y}^{(m+1)}) = Q(\boldsymbol{\alpha}^{(m+1)})$. Because of the nature of coordinate descent, $Q'(\boldsymbol{\alpha}^{(m+1)}, \mathbf{Y}^{(m+1)}) \leq Q'(\boldsymbol{\alpha}^{(m)}, \mathbf{Y}^{(m)})$, it follows that $Q(\boldsymbol{\alpha}^{(m+1)}) \leq Q(\boldsymbol{\alpha}^{(m)})$. □

Based on Proposition C.2, the iterations of coordinate descent are similar to that of EM algorithms and they reduce the value of Q, where $\mathbf{Y}$ plays the role of latent variable for EM algorithms.

D More Details about the Experiments

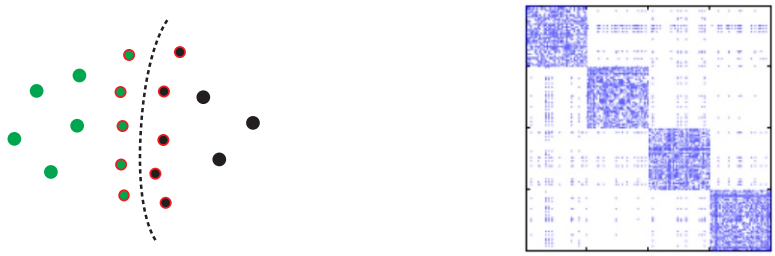
Figure 2 illustrates the importance of the adaptive weights $\boldsymbol{\alpha}$ in the discriminative similarity, as well as the illustration of the learned discriminative similarity.

As illustrated in Figure 2(a), the adaptive weights $\boldsymbol{\alpha}$ help reduce similarity between data close to cluster boundary, which is difficult for kernel similarity with uniform weights such as that for kernel K-means. Figure 2(b) illustrates the desirable block-diagonal structure of the proposed discriminative similarity on a dataset used for experiments.

The optimization of CDSK comprises M iterations of coordinate descent, wherein each iteration solves the following two subproblems.

1) With constant $\boldsymbol{\alpha}$,

$$\min_{\mathbf{Y} \in \mathbb{R}^{n \times c}} \mathrm{Tr}(\mathbf{Y}^\top \mathbf{L}^K \mathbf{Y}) \quad s.t.\ \mathbf{Y}^\top \mathbf{D}^K \mathbf{Y} = \mathbf{I}_c, \quad (24)$$



(a) (b)

Figure 2: (a) Illustration of the importance of the adaptive weights $\boldsymbol{\alpha}$ in the discriminative similarity. The green dots and black dots are data from two clusters separated by the dashed curve. Kernel similarity with uniform weights, such as that for kernel K-means, would incur large similarity between data (marked with red circles) close to the cluster boundary, although they come from different clusters. The misleading large similarity potentially hurts the clustering performance. The learnable and adaptive weights α help reduce the similarity between data close to cluster boundary by a data-level weighting on the kernel similarity. Bluer pixels represent larger discriminative similarity. (b) Illustration of the discriminative similarity generated by CDSK between facial images of 4 random people in MPIE Session 4 (MPIE S4), a dataset used for experiments.

2) With constant $\mathbf{Y}$,

$$\begin{aligned} \min_{\boldsymbol{\alpha}\in\Lambda,} \ & \mathrm{Tr}(\mathbf{Y}^\top \mathbf{L}^K \mathbf{Y}) - \sum_{i,j=1}^{n} \frac{\boldsymbol{\alpha}_i + \boldsymbol{\alpha}_j}{2} K_\tau(\mathbf{x}_i - \mathbf{x}_j) + \lambda \boldsymbol{\alpha}^\top \mathbf{K} \boldsymbol{\alpha} \\ s.t. \ & \mathbf{Y}^\top \mathbf{D}^K \mathbf{Y} = \mathbf{I}_c, \end{aligned} \tag{25}$$

The first subproblem (24) takes $\mathcal{O}(n^2 c)$ steps using truncated Singular Value Decomposition (SVD) by Krylov subspace iterative method. We adopt Sequential Minimal Optimization (SMO) (Platt, 1998) to solve the second subproblem (25), which takes roughly $\mathcal{O}(n^{2.1})$ steps as reported in (Platt, 1998). Therefore, the overall time complexity of CDSK is $\mathcal{O}(Mcn^2 + Mn^{2.1})$.

E PROOFS

E.1 PROOF OF LEMMA 3.1

Before stating the proof of Lemma 3.1, we introduce the famous spectral theorem in operator theory below.

Theorem E.1. (Spectral Theorem) Let L be a compact linear operator on a Hilbert space $\mathcal{H}$. Then there exists in H an orthonormal basis $\{\phi_1, \phi_2, \ldots\}$ consisting of eigenvectors of L. If λ_k is the eigenvalue corresponding to ϕ_k, then the set $\{\lambda_k\}$ is either finite or $\lambda_k \to 0$ when $k \to \infty$. In addition, the eigenvalues are real if L is self-adjoint.

Recall that the integral operator by S is defined as

$$(L_S f)(\mathbf{x}) = \int S(\mathbf{x}, \mathbf{t}) f(\mathbf{t}) d\mathbf{t},$$

and we are ready to prove Lemma 3.1.

Proof of Lemma 3.1. It can be verified that L_S is a compact operator. Therefore, according to Theorem E.1, $\{\phi_k\}$ is an orthogonal basis of $\mathcal{L}^2$. Note that ϕ_k is the eigenfunction of L_S with eigenvalue λ_k if $L_S \phi_k = \lambda_k \phi_k$.

With fixed $\mathbf{x} \in \mathcal{X}$, we then have

$$\begin{aligned}
&\Big|\sum_{k=m}^{m+\ell} \lambda_k \phi_k(\mathbf{x})\phi_k(\mathbf{t})\Big| \\
\leq &\big(\sum_{k=m}^{m+\ell} |\lambda_k||\phi_k(\mathbf{x})|^2\big)^{\frac{1}{2}} \cdot \big(\sum_{k=m}^{m+\ell} |\lambda_k||\phi_k(\mathbf{t})|^2\big)^{\frac{1}{2}} \\
\leq &\sqrt{C}\big(\sum_{k=m}^{m+\ell} |\lambda_k||\phi_k(\mathbf{x})|^2\big)^{\frac{1}{2}}.
\end{aligned}$$

It follows that the series $\sum_{k\geq 1} \lambda_k \phi_k(\mathbf{x})\phi_k(\mathbf{t})$ converges to a continuous function $e_{\mathbf{x}}$ uniformly on $\mathbf{t}$. This is because $\phi_k = \frac{L_S \phi_k}{\lambda_k}$ is continuous for nonzero λ_k.

On the other hand, for fixed $\mathbf{x} \in \mathcal{X}$, as a function in $\mathcal{L}^2$,

$$S(\mathbf{x}, \cdot) = \sum_{k\geq 1} \langle S(\mathbf{x}, \cdot), \phi_k\rangle \phi_k = \sum_{k\geq 1} \lambda_k \phi_k(\mathbf{x})\phi_k(\cdot).$$

Therefore, for fixed $\mathbf{x} \in \mathcal{X}$, $S(\mathbf{x}, \cdot) = \sum_{k\geq 1} \lambda_k \phi_k(\mathbf{x})\phi_k(\cdot) = e_{\mathbf{x}}(\cdot)$ almost surely w.r.t the Lebesgue measure. Since both are continuous functions, we must have $S(\mathbf{x}, \mathbf{t}) = \sum_{k\geq 1} \lambda_k \phi_k(\mathbf{x})\phi_k(\mathbf{t})$ for any $\mathbf{t} \in \mathcal{X}$. It follows that $S(\mathbf{x}, \mathbf{t}) = \sum_{k\geq 1} \lambda_k \phi_k(\mathbf{x})\phi_k(\mathbf{t})$ for any $\mathbf{x}, \mathbf{t} \in \mathcal{X}$.

We now consider two series which correspond to the positive eigenvalues and negative eigenvalues of L_S, namely $\sum_{k:\, \lambda_k \geq 0} \lambda_k \phi_k(\mathbf{x})\phi_k(\cdot)$ and $\sum_{k:\lambda_k<0} |\lambda_k|\phi_k(\mathbf{x})\phi_k(\cdot)$. Using similar argument, for fixed $\mathbf{x}$, both series converge to a continuous function, and we let

$$\begin{aligned}
S^+(\mathbf{x}, \mathbf{t}) &= \sum_{k:\, \lambda_k \geq 0} \lambda_k \phi_k(\mathbf{x})\phi_k(\mathbf{t}), \\
S^-(\mathbf{x}, \mathbf{t}) &= \sum_{k:\lambda_k<0} |\lambda_k|\phi_k(\mathbf{x})\phi_k(\mathbf{t}).
\end{aligned}$$

$S^+(\mathbf{x}, \mathbf{t})$ and $S^-(\mathbf{x}, \mathbf{t})$ are continuous function in $\mathbf{x}$ and $\mathbf{t}$. All the eigenvalues of S^+ and S^- are nonnegative, and it can be verified that both are PSD kernels since

$$\begin{aligned}
\sum_{i,j=1}^{n} c_i c_j S^+(\mathbf{x}_i, \mathbf{x}_j) &= \sum_{i,j=1}^{n} c_i c_j \sum_{k:\, \lambda_k \geq 0} \lambda_k \phi_k(\mathbf{x}_i)\phi_k(\mathbf{x}_j) \\
&= \sum_{k:\, \lambda_k \geq 0} \lambda_k \sum_{i,j=1}^{n} c_i c_j \phi_k(\mathbf{x}_i)\phi_k(\mathbf{x}_j) \\
&= \sum_{k:\, \lambda_k \geq 0} \lambda_k (\sum_{i=1}^{n} c_i \phi(\mathbf{x}_i))^2 \geq 0.
\end{aligned}$$

Similarly argument applies to S^-. Therefore, S is decomposed as $S(\mathbf{x}, \mathbf{t}) = S^+(\mathbf{x}, \mathbf{t}) - S^-(\mathbf{x}, \mathbf{t})$.

□

E.2 Proof of Theorem 3.2

Lemma E.3 will be used in the Proof of Theorem 3.2. The following lemma is introduced for the proof of Lemma E.3, whose proof appears in the end of this subsection.

Lemma E.2. The Rademacher complexity of the class $\mathcal{H}_S$ satisfies

$$\mathfrak{R}(\mathcal{H}_S) \leq (2c-1) \sum_{y=1}^{c} \mathfrak{R}(\mathcal{H}_{S,y}). \tag{26}$$

Lemma E.3. Define $\Omega^+(\boldsymbol{\alpha}) = \sum_{y=1}^{c} \boldsymbol{\alpha}^{(y)\top}\mathbf{S}^+\boldsymbol{\alpha}^{(y)}$ and $\Omega^-(\boldsymbol{\alpha}) = \sum_{y=1}^{c} \boldsymbol{\alpha}^{(y)\top}\mathbf{S}^-\boldsymbol{\alpha}^{(y)}$. When $\Omega^+(\boldsymbol{\alpha}) \leq {B^+}^2$,$\Omega^-(\boldsymbol{\alpha}) \leq {B^-}^2$ for positive constant B^+ and B^-, $\sup_{\mathbf{x}\in\mathcal{X}} |S^+(\mathbf{x},\mathbf{x})| \leq R^2$, $\sup_{\mathbf{x}\in\mathcal{X}} |S^-(\mathbf{x},\mathbf{x})| \leq R^2$ for some $R > 0$, then with probability at least $1-\delta$ over the data $\{\mathbf{x}_i\}_{i=1}^n$, the Rademacher complexity of the class $\mathcal{H}_S$ satisfies

$$\mathfrak{R}(\mathcal{H}_S) \leq \frac{R(2c-1)c(B^+ + B^-)}{\sqrt{n}} + 2c(2c-1)(B^+ + B^-)R^2\sqrt{\frac{\ln\frac{2}{\delta}}{2n}}. \tag{27}$$

Proof of Lemma E.3 . According to Lemma 3.1, S is decomposed into two PSD kernels as $S = S^+ - S^-$. Therefore, the are two Reproducing Kernel Hilbert Spaces $\mathcal{H}_S^+$ and $\mathcal{H}_S^-$ that are associated with S^+ and S^- respectively, and the canonical feature mappings in $\mathcal{H}_S^+$ and $\mathcal{H}_S^-$ are ϕ^+ and ϕ^-, with $S^+(\mathbf{x},\mathbf{t}) = \langle\phi^+(\mathbf{x}),\phi^+(\mathbf{t})\rangle_{H_K^+}$ and $S^-(\mathbf{x},\mathbf{t}) = \langle\phi^-(\mathbf{x}),\phi^-(\mathbf{t})\rangle_{H_K^-}$. In the following text, we will omit the subscripts H_K^+ and H_K^- without confusion.

For any $1 \leq y \leq c$,

$$h_S(\mathbf{x},y) = \sum_{i:\, y_i=y} \boldsymbol{\alpha}_i S(\mathbf{x},\mathbf{x}_i) = \langle\mathbf{w}^+,\phi^+(\mathbf{x})\rangle - \langle\mathbf{w}^-,\phi^-(\mathbf{x})\rangle$$

with $\|\mathbf{w}^+\|^2 = \boldsymbol{\alpha}^{(y)\top}\mathbf{S}^+\boldsymbol{\alpha}^{(y)} \leq {B^+}^2$ and $\|\mathbf{w}^-\|^2 = \boldsymbol{\alpha}^{(y)\top}\mathbf{S}^-\boldsymbol{\alpha}^{(y)} \leq {B^-}^2$. Therefore,

$$\begin{aligned}
&\mathcal{H}_{S,y} \subseteq \tilde{\mathcal{H}}_{S,y} = \{(\mathbf{x},y) \to \langle\mathbf{w}^+,\phi^+(\mathbf{x})\rangle - \langle\mathbf{w}^-,\phi^-(\mathbf{x})\rangle,\\
&\|\mathbf{w}^+\|^2 \leq {B^+}^2, \|\mathbf{w}^-\|^2 \leq {B^-}^2\}, 1 \leq y \leq c,
\end{aligned}$$

and $\mathfrak{R}(\mathcal{H}_{S,y}) \subseteq \mathfrak{R}(\tilde{\mathcal{H}}_{S,y})$. Since we are deriving upper bound for $\mathfrak{R}(\mathcal{H}_{S,y})$, we slightly abuse the notation and let $\mathcal{H}_{S,y}$ represent $\tilde{\mathcal{H}}_{S,y}$ in the remaining part of this proof.

For $\mathbf{x},\mathbf{t} \in \mathbb{R}^d$ and any $h_S \in \mathcal{H}_{S,y}$, we have

$$\begin{aligned}
|h_S(\mathbf{x}) - h_S(\mathbf{t})| &= |\langle\mathbf{w}^+,\phi^+(\mathbf{x})\rangle - \langle\mathbf{w}^-,\phi^-(\mathbf{x})\rangle - \langle\mathbf{w}^+,\phi^+(\mathbf{t})\rangle + \langle\mathbf{w}^-,\phi^-(\mathbf{t})\rangle|\\
&= |\langle\mathbf{w}^+,\phi^+(\mathbf{x}) - \phi^+(\mathbf{t})\rangle + \langle\mathbf{w}^-,\phi^-(\mathbf{t}) - \phi^-(\mathbf{x})\rangle|\\
&\leq B^+\|\phi^+(\mathbf{x}) - \phi^+(\mathbf{t})\| + B^-\|\phi^-(\mathbf{x}) - \phi^-(\mathbf{t})\|)\\
&\leq (B^+ + B^-)\sqrt{S^+(\mathbf{x},\mathbf{x}) + S^+(\mathbf{t},\mathbf{t}) + 2\sqrt{S^+(\mathbf{x},\mathbf{x})S^+(\mathbf{t},\mathbf{t})}}\\
&\leq 2R^2(B^+ + B^-).
\end{aligned}$$

We now approximate the Rademacher complexity of the function class $\mathcal{H}_{S,y}$ with its empirical version $\widehat{\mathfrak{R}}(\mathcal{H}_{S,y})$ using the sample $\{\mathbf{x}_i\}$. For each $1 \leq y \leq c$, Define $E_{\{\mathbf{x}_i\}}^{(y)} = \widehat{\mathfrak{R}}(\mathcal{H}_{S,y}) =$

$\mathbb{E}_{\{\sigma_i\}}\Big[\sup_{h_S(\cdot,y)\in\mathcal{H}_{S,y}}\Big|\frac{1}{n}\sum_{i=1}^{n}\sigma_i h_S(\mathbf{x}_i,y)\Big|\Big]$, then $\sum_{y=1}^{c}\mathfrak{R}(\mathcal{H}_{S,y})=\mathbb{E}_{\{\mathbf{x}_i\}}\Big[\sum_{y=1}^{c}E^{(y)}_{\{\mathbf{x}_i\}}\Big]$, and

$$
\begin{aligned}
&\sup_{\mathbf{x}_1,\ldots,\mathbf{x}_n,\mathbf{x}_t^{'}}\Big|E^{(y)}_{\mathbf{x}_1,\ldots,\mathbf{x}_{t-1},\mathbf{x}_t,\mathbf{x}_{t+1},\ldots,\mathbf{x}_n}-E^{(y)}_{\mathbf{x}_1,\ldots,\mathbf{x}_{t-1},\mathbf{x}_t^{'},\mathbf{x}_{t+1},\ldots,\mathbf{x}_n}\Big|\\
&=\sup_{\mathbf{x}_1,\ldots,\mathbf{x}_n,\mathbf{x}_t^{'}}\Bigg|\mathbb{E}_{\{\sigma_i\}}\Bigg[\sup_{h_S(\cdot,y)\in\mathcal{H}_{S,y}}\Big|\frac{1}{n}\sum_{i=1}^{n}\sigma_i h_S(\mathbf{x}_i,y)\Big|-\sup_{h_S(\cdot,y)\in\mathcal{H}_{S,y}}\Big|\frac{1}{n}\sum_{i\neq t}\sigma_i h_S(\mathbf{x}_i,y)+\frac{h_S(\mathbf{x}_t^{'},y)}{n}\Big|\Bigg]\Bigg|\\
&\leq\sup_{\mathbf{x}_1,\ldots,\mathbf{x}_n,\mathbf{x}_t^{'}}\mathbb{E}_{\{\sigma_i\}}\Bigg[\Bigg|\sup_{h_S(\cdot,y)\in\mathcal{H}_{S,y}}\Big|\frac{1}{n}\sum_{i=1}^{n}\sigma_i h_S(\mathbf{x}_i,y)\Big|-\sup_{h_S(\cdot,y)\in\mathcal{H}_{S,y}}\Big|\frac{1}{n}\sum_{i\neq t}\sigma_i h_S(\mathbf{x}_i,y)+\frac{h_S(\mathbf{x}_t^{'},y)}{n}\Big|\Bigg|\Bigg]\\
&\leq\sup_{\mathbf{x}_1,\ldots,\mathbf{x}_n,\mathbf{x}_t^{'}}\mathbb{E}_{\{\sigma_i\}}\Bigg[\sup_{h_S(\cdot,y)\in\mathcal{H}_{S,y}}\Big|\Big|\frac{1}{n}\sum_{i=1}^{n}\sigma_i h_S(\mathbf{x}_i,y)\Big|-\Big|\frac{1}{n}\sum_{i\neq t}\sigma_i h_S(\mathbf{x}_i,y)+\frac{h_S(\mathbf{x}_t^{'},y)}{n}\Big|\Big|\Bigg]\\
&\leq\sup_{\mathbf{x}_1,\ldots,\mathbf{x}_n,\mathbf{x}_t^{'}}\mathbb{E}_{\{\sigma_i\}}\Bigg[\sup_{h_S(\cdot,y)\in\mathcal{H}_{S,y}}\Big|\frac{1}{n}\sum_{i=1}^{n}\sigma_i h_S(\mathbf{x}_i,y)-\Big(\frac{1}{n}\sum_{i\neq t}\sigma_i h_S(\mathbf{x}_i,y)+\frac{h_S(\mathbf{x}_t^{'},y)}{n}\Big)\Big|\Bigg]\\
&=\sup_{\mathbf{x}_t,\mathbf{x}_t^{'}}\mathbb{E}_{\{\sigma_i\}}\Bigg[\sup_{h_S(\cdot,y)\in\mathcal{H}_{S,y}}\Big|\frac{h_S(\mathbf{x}_t,y)}{n}-\frac{h_S(\mathbf{x}_t^{'},y)}{n}\Big|\Bigg]\\
&\leq\frac{2R^2(B^{+}+B^{-})}{n}.
\end{aligned}
$$

It follows that $\Big|\sum_{y=1}^{c}E^{(y)}_{\mathbf{x}_1,\ldots,\mathbf{x}_{t-1},\mathbf{x}_t,\mathbf{x}_{t+1},\ldots,\mathbf{x}_n}-\sum_{y=1}^{c}E^{(y)}_{\mathbf{x}_1,\ldots,\mathbf{x}_{t-1},\mathbf{x}_t^{'},\mathbf{x}_{t+1},\ldots,\mathbf{x}_n}\Big|\leq\frac{2R^2(B^{+}+B^{-})c}{n}$. According to the McDiarmid's Inequality,

$$
\Pr\Big[\Big|\sum_{y=1}^{c}\widehat{\mathfrak{R}}(\mathcal{H}_{S,y})-\sum_{y=1}^{c}\mathfrak{R}(\mathcal{H}_{S,y})\Big|\geq\varepsilon\Big]\leq 2\exp\big(-\frac{n\varepsilon^2}{2(B^{+}+B^{-})^2R^4c^2}\big). \tag{28}
$$

Now we derive the upper bound for the empirical Rademacher complexity:

$$
\begin{aligned}
&\sum_{y=1}^{c}\widehat{\mathfrak{R}}(\mathcal{H}_{S,y})=\sum_{y=1}^{c}\mathbb{E}_{\{\sigma_i\}}\Bigg[\sup_{h_S\in\mathcal{H}_{S,y}}\Big|\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{n}\sigma_i h_S(\mathbf{x}_i)\Big|\Bigg] \qquad (29)\\
&\leq\frac{1}{n}\sum_{y=1}^{c}\mathbb{E}_{\{\sigma_i\}}\Bigg[\sup_{\|\mathbf{w}^{+}\|\leq B^{+},\|\mathbf{w}^{-}\|\leq B^{-}}\Big|\sum_{i=1}^{n}\sigma_i\big(\langle\mathbf{w}^{+},\phi^{+}(\mathbf{x}_i)\rangle-\langle\mathbf{w}^{-},\phi^{-}(\mathbf{x}_i)\rangle\big)\Big|\Bigg]\\
&\leq\frac{1}{n}\sum_{y=1}^{c}\mathbb{E}_{\{\sigma_i\}}\Bigg[B^{+}\|\sum_{i=1}^{n}\sigma_i\phi^{+}(\mathbf{x}_i)\|+B^{-}\|\sum_{i=1}^{n}\sigma_i\phi^{-}(\mathbf{x}_i)\|\Bigg]\\
&=\frac{B^{+}c}{n}\mathbb{E}_{\{\sigma_i\}}\Bigg[\|\sum_{i=1}^{n}\sigma_i\phi^{+}(\mathbf{x}_i)\|\Bigg]+\frac{B^{-}c}{n}\mathbb{E}_{\{\sigma_i\}}\Bigg[\|\sum_{i=1}^{n}\sigma_i\phi^{-}(\mathbf{x}_i)\|\Bigg]\\
&\leq\frac{B^{+}c}{n}\sqrt{\mathbb{E}_{\{\sigma_i\}}\Bigg[\|\sum_{i=1}^{n}\sigma_i\phi^{+}(\mathbf{x}_i)\|^2\Bigg]}+\frac{B^{-}c}{n}\sqrt{\mathbb{E}_{\{\sigma_i\}}\Bigg[\|\sum_{i=1}^{n}\sigma_i\phi^{-}(\mathbf{x}_i)\|^2\Bigg]}\\
&\leq\frac{B^{+}c}{n}\sqrt{\sum_{i}^{n}s^{+}(\mathbf{x}_i,\mathbf{x}_i)}+\frac{B^{-}c}{n}\sqrt{\sum_{i}^{n}s^{-}(\mathbf{x}_i,\mathbf{x}_i)}\leq\frac{Rc}{\sqrt{n}}(B^{+}+B^{-}).
\end{aligned}
$$

By Lemma E.2, (28) and (29), with probability at least $1-\delta$, we have

$$
\mathfrak{R}(\mathcal{H}_S)\leq(2c-1)\sum_{y=1}^{c}\mathfrak{R}(\mathcal{H}_{S,y})\leq\frac{R(2c-1)c(B^{+}+B^{-})}{\sqrt{n}}+2c(2c-1)(B^{+}+B^{-})R^2\sqrt{\frac{\ln\frac{2}{\delta}}{2n}}. \tag{30}
$$

□

Proof of Theorem 3.2 . According to Theorem 2 in Koltchinskii & Panchenko (2002), with probability $1-\delta$ over the labeled data $\mathcal{S}$ with respect to any distribution in $\mathcal{P}$, the generalization error of the kernel classifier f_S satisfies

$$R(f_S) \leq \widehat{R}_n(f_S) + \frac{8}{\gamma}\mathfrak{R}(\mathcal{H}_S) + \sqrt{\frac{\ln 2/\delta}{2n}}, \tag{31}$$

where $\widehat{R}_n(f_S) = \frac{1}{n}\sum_{i=1}^{n}\Phi(\frac{m_{h_S}(\mathbf{x}_i,y_i)}{\gamma})$ is empirical error of the classifier for $\gamma > 0$. Note that $\Phi(\frac{m_{h_S}(\mathbf{x}_i,y_i)}{\gamma}) \leq \Phi\Big(\frac{h_S(\mathbf{x}_i,y_i)-\sum_{y\neq y_i} h_S(\mathbf{x}_i,y)}{\gamma}\Big)$, applying Lemma E.3, (4) holds with probability $1-\delta$. When $\gamma \geq c-1$, it can be verified that $\Big|h_S(\mathbf{x}_i,y_i) - \sum_{y\neq y_i} h_S(\mathbf{x}_i,y)\Big| \leq c-1$ for all $(\mathbf{x}_i,y_i)$, so that

$$\Phi\Big(\frac{h_S(\mathbf{x}_i,y_i) - \sum_{y\neq y_i} h_S(\mathbf{x}_i,y)}{\gamma}\Big) = 1 - \frac{h(\mathbf{x}_i,y_i) - \sum_{y\neq y_i} h(\mathbf{x}_i,y)}{\gamma},$$

and (5) is obtained.

□

Proof of Lemma E.2. Inspired by Koltchinskii & Panchenko (2002), we first prove that the Rademacher complexity of the function class formed by the maximum of several hypotheses is bounded by two times the sum of the Rademacher complexity of the function classes that these hypothesis belong to. That is,

$$\mathfrak{R}(\mathcal{H}_{\max}) \leq 2\sum_{y=1}^{k}\mathfrak{R}(\mathcal{H}_{S,y}), \tag{32}$$

where $\mathcal{H}_{\max} = \{\max\{h_1,\ldots,h_k\}\colon h_y \in \mathcal{H}_{S,y}, 1\leq y\leq k\}$ for $1\leq k\leq c-1$.

If no confusion arises, the notations $(\{\sigma_i\},\{\mathbf{x}_i,y_i\})$ are omitted in the subscript of the expectation operator in the following text, i.e. $\mathbb{E}_{\{\sigma_i\},\{\mathbf{x}_i,y_i\}}$ is abbreviated to $\mathbb{E}$. According to Theorem 11 of Koltchinskii & Panchenko (2002), it can be verified that

$$\begin{aligned}
&\mathbb{E}_{\{\sigma_i\},\{\mathbf{x}_i,y_i\}}\left[\left(\sup_{h\in\mathcal{H}_{\max}}\left|\frac{1}{n}\sum_{i=1}^{n}\sigma_i h_S(\mathbf{x}_i)\right|\right)^+\right] \\
&\leq \sum_{y=1}^{k}\mathbb{E}_{\{\sigma_i\},\{\mathbf{x}_i,y_i\}}\left[\left(\sup_{h\in\mathcal{H}_{S,y}}\left|\frac{1}{n}\sum_{i=1}^{n}\sigma_i h_S(\mathbf{x}_i)\right|\right)^+\right].
\end{aligned}$$

Therefore,

$$\begin{aligned}
\mathfrak{R}(\mathcal{H}_{\max}) &= \mathbb{E}_{\{\sigma_i\},\{\mathbf{x}_i,y_i\}}\left[\sup_{h\in\mathcal{H}_{\max}}\left|\frac{1}{n}\sum_{i=1}^{n}\sigma_i h_S(\mathbf{x}_i)\right|\right] \\
&\leq \mathbb{E}_{\{\sigma_i\},\{\mathbf{x}_i,y_i\}}\left[\Big(\sup_{h\in\mathcal{H}_{\max}}\frac{1}{n}\sum_{i=1}^{n}\sigma_i h_S(\mathbf{x}_i)\Big)^+\right] \\
&\quad+\mathbb{E}_{\{\sigma_i\},\{\mathbf{x}_i,y_i\}}\left[\Big(\sup_{h\in\mathcal{H}_{\max}}-\frac{1}{n}\sum_{i=1}^{n}\sigma_i h_S(\mathbf{x}_i)\Big)^+\right] \\
&= 2\mathbb{E}_{\{\sigma_i\},\{\mathbf{x}_i,y_i\}}\left[\Big(\sup_{h\in\mathcal{H}_{\max}}\frac{1}{n}\sum_{i=1}^{n}\sigma_i h_S(\mathbf{x}_i)\Big)^+\right] \\
&\leq 2\sum_{y=1}^{k}\mathbb{E}_{\{\sigma_i\},\{\mathbf{x}_i,y_i\}}\left[\Big(\sup_{h\in\mathcal{H}_{S,y}}\frac{1}{n}\sum_{i=1}^{n}\sigma_i h_S(\mathbf{x}_i)\Big)^+\right]
\end{aligned}$$

$$\leq 2\sum_{y=1}^{k}\mathbb{E}_{\{\sigma_i\},\{\mathbf{x}_i,y_i\}}\left[\sup_{h\in\mathcal{H}_{S,y}}\left|\frac{1}{n}\sum_{i=1}^{n}\sigma_i h_S(\mathbf{x}_i)\right|\right] = 2\sum_{y=1}^{k}\mathfrak{R}(\mathcal{H}_{S,y}). \tag{33}$$

The equality in the third line of (33) is due to the fact that $-\sigma_i$ has the same distribution as σ_i. Using this fact again, (32), we have

$$\begin{aligned}
\mathfrak{R}(\mathcal{H}_S) &= \mathbb{E}_{\{\sigma_i\},\{\mathbf{x}_i,y_i\}}\left[\sup_{m_{h_S}\in\mathcal{H}_S}\left|\frac{1}{n}\sum_{i=1}^{n}\sigma_i m_{h_S}(\mathbf{x}_i,y_i)\right|\right] \\
&= \mathbb{E}_{\{\sigma_i\},\{\mathbf{x}_i,y_i\}}\left[\sup_{m_{h_S}\in\mathcal{H}_S}\left|\frac{1}{n}\sum_{i=1}^{n}\sigma_i\sum_{y=1}^{c}m_{h_S}(\mathbf{x}_i,y)\mathbb{1}_{y=y_i}\right|\right] \\
&\leq \sum_{y=1}^{c}\mathbb{E}_{\{\sigma_i\},\{\mathbf{x}_i,y_i\}}\left[\sup_{m_{h_S}\in\mathcal{H}_S}\left|\frac{1}{n}\sum_{i=1}^{n}\sigma_i m_{h_S}(\mathbf{x}_i,y)\mathbb{1}_{y=y_i}\right|\right] \\
&\leq \frac{1}{2n}\sum_{y=1}^{c}\mathbb{E}_{\{\sigma_i\},\{\mathbf{x}_i,y_i\}}\left[\sup_{m_{h_S}\in\mathcal{H}_S}\left|\sum_{i=1}^{n}\sigma_i m_{h_S}(\mathbf{x}_i,y)(2\mathbb{1}_{y=y_i}-1)\right|\right] \\
&\quad+\frac{1}{2n}\sum_{y=1}^{c}\mathbb{E}_{\{\sigma_i\},\{\mathbf{x}_i\}}\left[\sup_{m_{h_S}\in\mathcal{H}_S}\left|\sum_{i=1}^{n}\sigma_i m_{h_S}(\mathbf{x}_i,y)\right|\right] \\
&= \frac{1}{n}\sum_{y=1}^{c}\mathbb{E}_{\{\sigma_i\},\{\mathbf{x}_i\}}\left[\sup_{m_{h_S}\in\mathcal{H}_S}\left|\sum_{i=1}^{n}\sigma_i m_{h_S}(\mathbf{x}_i,y)\right|\right].
\end{aligned} \tag{34}$$

Also, for any given $1\leq y\leq c$,

$$\begin{aligned}
&\frac{1}{n}\mathbb{E}_{\{\sigma_i\},\{\mathbf{x}_i\}}\left[\sup_{m_{h_S}\in\mathcal{H}_S}\left|\sum_{i=1}^{n}\sigma_i m_{h_S}(\mathbf{x}_i,y)\right|\right] \\
&= \frac{1}{n}\mathbb{E}_{\{\sigma_i\},\{\mathbf{x}_i\}}\left[\sup_{h_S(\cdot,y)\in\mathcal{H}_{S,y},y=1\ldots c}\left|\sum_{i=1}^{n}\sigma_i h_S(\mathbf{x}_i,y)-\sigma_i\text{argmax}_{y'\neq y}h_S(\mathbf{x}_i,y')\right|\right] \\
&\leq \frac{1}{n}\mathbb{E}_{\{\sigma_i\},\{\mathbf{x}_i\}}\left[\sup_{h_S(\cdot,y)\in\mathcal{H}_{S,y}}\left|\sum_{i=1}^{n}\sigma_i h_S(\mathbf{x}_i,y)\right|\right] \\
&\quad+\frac{1}{n}\mathbb{E}_{\{\sigma_i\},\{\mathbf{x}_i\}}\left[\sup_{h_S(\cdot,y')\in\mathcal{H}'_{S,y},y'\neq y}\left|\sum_{i=1}^{n}\sigma_i\text{argmax}_{y'\neq y}h_S(\mathbf{x}_i,y')\right|\right] \\
&\leq \frac{1}{n}\mathbb{E}_{\{\sigma_i\},\{\mathbf{x}_i\}}\left[\sup_{h_S(\cdot,y)\in\mathcal{H}_{S,y}}\left|\sum_{i=1}^{n}\sigma_i h_S(\mathbf{x}_i,y)\right|\right]+\frac{2}{n}\sum_{y'\neq y}\mathbb{E}_{\{\sigma_i\},\{\mathbf{x}_i\}}\left[\sup_{h_S(\cdot,y')\in\mathcal{H}'_{S,y}}\left|\sum_{i=1}^{n}\sigma_i h_S(\mathbf{x}_i,y')\right|\right].
\end{aligned} \tag{35}$$

Combining (34) and (35),

$$\begin{aligned}
\mathfrak{R}(\mathcal{H}_S) &\leq \sum_{y=1}^{c}\frac{1}{n}\mathbb{E}_{\{\sigma_i\},\{\mathbf{x}_i\}}\left[\sup_{h_S(\cdot,y)\in\mathcal{H}_{S,y}}\left|\sum_{i=1}^{n}\sigma_i h_S(\mathbf{x}_i,y)\right|\right] \\
&\quad+\sum_{y=1}^{c}\sum_{y=1}^{c}\frac{2}{n}\sum_{y'\neq y}\mathbb{E}_{\{\sigma_i\},\{\mathbf{x}_i\}}\left[\sup_{h_S(\cdot,y')\in\mathcal{H}'_{S,y}}\left|\sum_{i=1}^{n}\sigma_i h_S(\mathbf{x}_i,y')\right|\right] \\
&= (2c-1)\sum_{y=1}^{c}\mathbb{E}_{\{\sigma_i\},\{\mathbf{x}_i\}}\left[\sup_{h_S(\cdot,y)\in\mathcal{H}_{S,y}}\left|\frac{1}{n}\sum_{i=1}^{n}\sigma_i h_S(\mathbf{x}_i,y)\right|\right] \\
&= (2c-1)\sum_{y=1}^{c}\mathfrak{R}(\mathcal{H}_{S,y}).
\end{aligned} \tag{36}$$

□

E.3 PROOF OF THEOREM A.1

Proof. According to definition of ISE,

$$\mathrm{ISE}(\widehat{r}, r) = \int_{\mathbb{R}^d} (\widehat{r} - r)^2 dx = \int_{\mathbb{R}^d} \widehat{r}(\mathbf{x}, \boldsymbol{\alpha})^2 dx - 2\int_{\mathbb{R}^d} \widehat{r}(\mathbf{x}, \boldsymbol{\alpha}) r(\mathbf{x}) dx + \int_{\mathbb{R}^d} r(\mathbf{x})^2 dx. \tag{37}$$

For a given distribution, $\int_{\mathbb{R}^d} r(\mathbf{x})^2 dx$ is a constant. By Gaussian convolution theorem,

$$\int_{\mathbb{R}^d} \widehat{r}(\mathbf{x}, \boldsymbol{\alpha})^2 dx = \tau_1 \sum_{y=1}^{2} \boldsymbol{\alpha}^{(y)\top} (\mathbf{K}_{\sqrt{2}h}) \boldsymbol{\alpha}^{(y)} - \tau_1 \sum_{1 \le i < j \le n} 2\boldsymbol{\alpha}_i \boldsymbol{\alpha}_j \mathbf{K}_{\sqrt{2}h}(\mathbf{x}_i - \mathbf{x}_j) \mathbb{1}_{y_i \neq y_j}, \tag{38}$$

where $\tau_1 = \frac{1}{(2\pi)^{d/2}(\sqrt{2}h)^d}$. Moreover,

$$\begin{aligned} &\int_{\mathbb{R}^d} \widehat{r}(\mathbf{x}, \boldsymbol{\alpha}) r(\mathbf{x}) dx \\ &= \int_{\mathbb{R}^d} \widehat{p}(\mathbf{x}, 1) p(\mathbf{x}, 1) dx + \int_{\mathbb{R}^d} \widehat{p}(\mathbf{x}, 2) p(\mathbf{x}, 2) dx - \int_{\mathbb{R}^d} \widehat{p}(\mathbf{x}, 1) p(\mathbf{x}, 2) dx - \int_{\mathbb{R}^d} \widehat{p}(\mathbf{x}, 2) p(\mathbf{x}, 1) dx. \end{aligned} \tag{39}$$

Note that

$$\frac{1}{\tau_0} \int_{\mathbb{R}^d} \widehat{p}(\mathbf{x}, 1) p(\mathbf{x}, 1) dx = \sum_{j:\, y_j = 1} \int_{\mathbb{R}^d} \boldsymbol{\alpha}_j K_\tau(\mathbf{x} - \mathbf{x}_j) p(\mathbf{x}, 1) dx,$$

we then use the empirical term $\frac{\sum_{i:\, i \neq j} \boldsymbol{\alpha}_j K_\tau(\mathbf{x}_i - \mathbf{x}_j) \mathbb{1}_{y_i = 1}}{n-1}$ to approximate the integral $\int_{\mathbb{R}^d} \boldsymbol{\alpha}_j K_\tau(\mathbf{x} - \mathbf{x}_j) p(\mathbf{x}, 1) dx$. Since $\mathbb{E}_{\{\mathbf{x}_i, y_i\}_{i \neq j}} \left[\frac{\sum_{i:\, i \neq j} \boldsymbol{\alpha}_j K_\tau(\mathbf{x}_i - \mathbf{x}_j) \mathbb{1}_{y_i = 1}}{n-1} \right] = \int_{\mathbb{R}^d} \boldsymbol{\alpha}_j K_\tau(\mathbf{x} - \mathbf{x}_j) p(\mathbf{x}, 1) dx$, and bounded difference holds for $\frac{\sum_{i:\, i \neq j} \boldsymbol{\alpha}_j K_\tau(\mathbf{x}_i - \mathbf{x}_j) \mathbb{1}_{y_i = 1}}{n-1}$, therefore,

$$\Pr \left[\left| \frac{\sum_{i:\, i \neq j} \boldsymbol{\alpha}_j K_\tau(\mathbf{x}_i - \mathbf{x}_j) \mathbb{1}_{y_i = 1}}{n-1} - \int_{\mathbb{R}^d} \boldsymbol{\alpha}_j K_\tau(\mathbf{x} - \mathbf{x}_j) p(\mathbf{x}, 1) dx \right| \ge \boldsymbol{\alpha}_j \varepsilon \right] \le 2 \exp\big(-2(n-1)\varepsilon^2 \big).$$

It follows that with probability at least $1 - 2n_1 \exp\big(-2(n-1)\varepsilon^2 \big)$, where n_i is the number of data points with label i,

$$\left| \frac{\sum_{i,j:\, i \neq j, y_i = y_j = 1} \boldsymbol{\alpha}_j K_\tau(\mathbf{x}_i - \mathbf{x}_j)}{n-1} - \frac{1}{\tau_0} \int_{\mathbb{R}^d} \widehat{p}(\mathbf{x}, 1) p(\mathbf{x}, 1) dx \right| \le \sum_{j:\, y_j = 1} \boldsymbol{\alpha}_j \varepsilon. \tag{40}$$

Similarly, with probability at least $1 - 2n_2 \exp\big(-2(n-1)\varepsilon^2 \big)$,

$$\left| \frac{\sum_{i,j:\, i \neq j, y_i = y_j = 2} \boldsymbol{\alpha}_j K_\tau(\mathbf{x}_i - \mathbf{x}_j)}{n-1} - \frac{1}{\tau_0} \int_{\mathbb{R}^d} \widehat{p}(\mathbf{x}, 2) p(\mathbf{x}, 2) dx \right| \le \sum_{j:\, y_j = 2} \boldsymbol{\alpha}_j \varepsilon. \tag{41}$$

It follows from (40) and (41) that with probability at least $1 - 2n \exp\big(-2(n-1)\varepsilon^2 \big)$,

$$\left| \frac{\sum_{i,j:\, i \neq j, y_i = y_j} \boldsymbol{\alpha}_j K_\tau(\mathbf{x}_i - \mathbf{x}_j)}{n-1} - \frac{1}{\tau_0} \int_{\mathbb{R}^d} \big(\widehat{p}(\mathbf{x}, 1) p(\mathbf{x}, 1) + \widehat{p}(\mathbf{x}, 2) p(\mathbf{x}, 2) \big) dx \right| \le \varepsilon. \tag{42}$$

In the same way, with probability at least $1 - 2n \exp\big(-2n\varepsilon^2 \big)$,

$$\left| \frac{\sum_{i,j:\, y_i \neq y_j} \boldsymbol{\alpha}_j K_\tau(\mathbf{x}_i - \mathbf{x}_j)}{n} - \frac{1}{\tau_0} \int_{\mathbb{R}^d} \big(\widehat{p}(\mathbf{x}, 1) p(\mathbf{x}, 2) + \widehat{p}(\mathbf{x}, 2) p(\mathbf{x}, 1) \big) dx \right| \le \varepsilon. \tag{43}$$

Based on (42) and (43), with probability at least $1 - 2n_2 \exp\big(-2(n-1)\varepsilon^2\big) - 2n\exp\big(-2n\varepsilon^2\big)$,

$$
\begin{aligned}
\mathrm{ISE}(\widehat{r}, r) \leq\ & 2\tau_0 \frac{\sum\limits_{i,j\,:\,y_i \neq y_j} \boldsymbol{\alpha}_j K_\tau(\mathbf{x}_i - \mathbf{x}_j)}{n} - 2\tau_0 \frac{\sum\limits_{i,j\,:\,i \neq j, y_i = y_j} \boldsymbol{\alpha}_j K_\tau(\mathbf{x}_i - \mathbf{x}_j)}{n-1} \\
& + \tau_1 \sum_{y=1}^{2} \boldsymbol{\alpha}^{(y)\top} (\mathbf{K}_{\sqrt{2}h}) \boldsymbol{\alpha}^{(y)} - \tau_1 \sum_{1 \leq i < j \leq n} 2\boldsymbol{\alpha}_i \boldsymbol{\alpha}_j K_{\sqrt{2}h}(\mathbf{x}_i - \mathbf{x}_j) \mathbb{1}_{y_i \neq y_j} + 2\tau_0 \varepsilon \\
\leq\ & 4\tau_0 \frac{\sum\limits_{1 \leq i < j \leq n} (\boldsymbol{\alpha}_i + \boldsymbol{\alpha}_j) K_\tau(\mathbf{x}_i - \mathbf{x}_j) \mathbb{1}_{y_i \neq y_j}}{n} - \tau_0 \frac{\sum\limits_{i,j=1}^{n} (\boldsymbol{\alpha}_i + \boldsymbol{\alpha}_j) K_\tau(\mathbf{x}_i - \mathbf{x}_j)}{n} \\
& + \tau_1 \sum_{y=1}^{2} \boldsymbol{\alpha}^{(y)\top} (\mathbf{K}_{\sqrt{2}h}) \boldsymbol{\alpha}^{(y)} - \tau_1 \sum_{1 \leq i < j \leq n} 2\boldsymbol{\alpha}_i \boldsymbol{\alpha}_j K_{\sqrt{2}h}(\mathbf{x}_i - \mathbf{x}_j) \mathbb{1}_{y_i \neq y_j} + 2\tau_0 \Big(\frac{1}{n-1} + \varepsilon\Big).
\end{aligned}
\tag{44}
$$

The conclusion of this theorem can be obtained from (44). □