Sharper Generalization Bounds for Clustering

Shaojie Li [1 2] **Yong Liu** [1 2]

Abstract

Existing generalization analysis of clustering mainly focuses on specific instantiations, such as (kernel) k-means, and a unified framework for studying clustering performance is still lacking. Besides, the existing excess clustering risk bounds are mostly of order $\mathcal{O}(K/\sqrt{n})$ provided that the underlying distribution has bounded support, where n is the sample size and K is the cluster numbers, or of order $\mathcal{O}(K^2/n)$ under strong assumptions on the underlying distribution, where these assumptions are hard to be verified in general. In this paper, we propose a unified clustering learning framework and investigate its excess risk bounds, obtaining state-of-the-art upper bounds under mild assumptions. Specifically, we derive sharper bounds of order $\mathcal{O}(K^2/n)$ under mild assumptions on the covering number of the hypothesis spaces, where these assumptions are easy to be verified. Moreover, for the hard clustering scheme, such as (kernel) k-means, if just assume the hypothesis functions to be bounded, we improve the upper bounds from the order $\mathcal{O}(K/\sqrt{n})$ to $\mathcal{O}(\sqrt{K}/\sqrt{n})$. Furthermore, state-of-the-art bounds of faster order $\mathcal{O}(K/n)$ are obtained with the covering number assumptions.

1. Introduction

Clustering is one of the fundamental issues in unsupervised learning and has been used in various applications (Xu & Wunsch, 2005; Von Luxburg, 2007; Jain, 2010; Shaham et al., 2018; Liu et al., 2018). In a clustering scheme, datasets are divided into several subgroups, such that data points in the same subgroup are more similar to each other than to those in other subgroups. Although clustering has been studied for decades, by contrast to the thriving of clustering algorithm design and application, the statistical theory of clustering may appear to be not sufficient. The existing excess risk bounds are mostly derived for different specific instantiations of clustering learning problems, such as k-means (Thorpe et al., 2015; Tang & Monteleoni, 2016; Levrard et al., 2013; Antos, 2005), kernel k-means (Biau et al., 2008; Antos et al., 2005; Levrard et al., 2015; Levrard, 2018) or spectral clustering (Terada & Yamamoto, 2019), and a unified framework to study generalization performance for clustering learning is still lacking. Moreover, the existing excess clustering risk bounds either have a slow convergence rate or require pretty strong assumptions on the underlying distribution to get the faster convergence rate, however, whose assumptions are hard to be verified in general. Specifically, if the distribution has bounded support, the excess risk upper bounds are mostly of order $\mathcal{O}(K/\sqrt{n})$ (Linder, 2000; Biau et al., 2008; Maurer & Pontil, 2010). If the distribution further satisfies a strong assumption called margin condition, the faster convergence rate of order $\mathcal{O}(K^2/n)$ appears to be obtained (Chou, 1994; Antos et al., 2005; Levrard et al., 2015).

Motivated by these problems, we first propose a clustering learning framework, which is suitable for (kernel) k-means, soft k-means, spectral clustering, neural network clustering scheme, etc, and then investigate its excess risk bounds. We start our analysis by proposed clustering Rademacher complexity and show that the upper bound obtained by it is just of order $\mathcal{O}(K/\sqrt{n})$ under the bounded hypothesis functions assumption. Since the Rademacher complexity (Bartlett & Mendelson, 2002) considers the worst case, we further define the more reasonable local clustering Rademacher complexity (Bartlett et al., 2005) and use it to get a basic excess risk bound with the fixed point under the same assumptions. After that, by using some assumptions of the covering number on the hypothesis function classes, we obtain suitable fixed points and get sharper excess risk upper bounds, whose convergence rates are of order $\mathcal{O}(K^2/n)$. Note that our assumptions of covering number on hypothesis classes are mild and easy to be verified compared with the margin condition (Pollard et al., 1982; Chou, 1994; Antos et al., 2005; Levrard et al., 2015; Terada & Yamamoto, 2019). Furthermore, different from bounding the clustering Rademacher complexity by the maximum Rademacher complexity of the restrictions of the function class along

[1]Gaoling School of Artificial Intelligence, Renmin University of China, Beijing, China [2]Beijing Key Laboratory of Big Data Management and Analysis Methods, Beijing, China. Correspondence to: Yong Liu <liuyonggsai@ruc.edu.cn>.

Proceedings of the 38^{th} International Conference on Machine Learning, PMLR 139, 2021.

each coordinate with timing a factor of $\mathcal{O}(K)$, we provide a new bound for clustering Rademacher complexity, reducing the order of K to $\sqrt{K}$. The excess clustering risk bounds are therefore improved to $\mathcal{O}(L\sqrt{K}/\sqrt{n})$ and $\mathcal{O}(L^2K/n)$ dependent on the clustering Rademacher complexity and the local clustering Rademacher complexity respectively, where L denotes a Lipschitz parameter. Particularly for the hard clustering scheme, $L = 1$, hence, our excess clustering risk bounds have convergence rates of order $\mathcal{O}(\sqrt{K}/\sqrt{n})$ and $\mathcal{O}(K/n)$, respectively, which are state-of-the-art.

2. Related Work

In this section, we introduce the related work on excess risk bound analysis of clustering learning and the local Rademacher complexity.

2.1. Excess Risk Bounds

The existing excess risk bounds of clustering learning are mostly derived for (kernel) k-means.

VC-dimension Bounds. VC-dimension technique is commonly used for finite-dimensional problem of k-means. (Linder et al., 1994; Bartlett et al., 1998; Linder, 2000) show that the excess clustering risk bounds obtained by VC-dimension technique are of order $\mathcal{O}(\sqrt{K}/\sqrt{n})$ provided that the underlying distribution has bounded support. In addition, the problem of quantifying how good empirically designed minimizers are, compared to the truly optimal ones, has also been extensively studied in (Linder, 2002).

Rademacher Complexity Bounds. An emerging problem in finite dimension is that the upper bound usually relevant to the dimension d, while the hypothesis space of kernel k-means is typically an infinite-dimensional Hilbert space. (Biau et al., 2008; Maurer & Pontil, 2010; Canas et al., 2012; Fefferman et al., 2016; Calandriello & Rosasco, 2018; Fischer, 2010) use the Rademacher complexity technique to extend the previous results (Linder et al., 1994; Bartlett et al., 1998; Linder, 2000) and provide dimension-independent bounds for kernel k-means, whose bounds are mostly of order $\mathcal{O}(K/\sqrt{n})$.

U-Process Bounds. (Cao et al., 2016) study the excess risk of clustering learning from the perspective of similarity learning, not only focusing on k-means. (Clémençcon, 2011) study the hard clustering scheme where a sample is assigned to each subgroup with probability 0 or 1. They all model the clustering learning as pairwise learning problems and use the tool of U-process to analyze the excess clustering risk bound. They use symmetry of U-statistics to control supremum of a U-process in generalization analysis by the supremum of a Rademacher process, which can be bounded by standard techniques in the i.i.d context, providing upper bounds of order $\mathcal{O}(K/\sqrt{n})$.

Margin Condition Bounds. Several results (Pollard et al., 1982; Chou, 1994; Antos et al., 2005) show that the convergence rate can be improved to $\mathcal{O}(C_K/n)$ under different sets of assumption on the distribution in the finite-dimensional problem of k-means, where C_K denotes a parameter with respect to K. Note that the relationship in their results between K and the upper bound is not clear. (Levrard et al., 2013) show that these different sets of assumptions turn out to be equivalent in the continuous density case to a technical condition, and also provide an upper bound of order $\mathcal{O}(C_K/n)$. In a context of a separable Hilbert space, (Levrard et al., 2015) extend the above results and propose an assumption on the underlying distribution, called margin condition, to satisfy the technical condition. They establish a fast convergence rate of order $\mathcal{O}(K^2/n)$. Under the framework of margin condition, (Levrard, 2018) refine the results in (Levrard et al., 2015) and give an excess clustering risk upper bound of order $\mathcal{O}(K/n)$ under another pretty strong assumption. (Terada & Yamamoto, 2019) provide the convergence rate of order $\mathcal{O}(K^2/n)$ for normalized cut (Von Luxburg, 2007) based on margin condition assumption (Levrard et al., 2015), which is a related work on spectral clustering. We will show the margin condition and explain that it is hard to be verified in general in Section 5.2.

2.2. Local Rademacher Complexity

Local Rademacher complexity is an important tool in statistical learning theory, and has been used to obtain better generalization error bounds for many important supervised learning problems (Bartlett et al., 2005; Koltchinskii et al., 2006; Liu & Liao, 2015; Liu et al., 2017b;a; Xu et al., 2016; Yousefi et al., 2018; Li et al., 2018). In the generalization analysis of clustering learning, (Levrard et al., 2015; Levrard, 2018) also apply it to obtain the sharper excess risk bounds, however, their results are built on the margin condition, leaving challenges in the clustering learning since their assumptions are hard to be verified in general. In this paper, we use assumptions of covering number instead of margin condition to derive sharper excess risk bounds together with the local Rademacher complexity technique. Since the covering number of many popular function classes are known, thus our assumptions are milder and easier to check.

3. A Clustering Learning Framework

In this section, after a brief description of notations, we first introduce the clustering learning framework and then present some examples to explain it.

3.1. Notations

Assume that μ is an underlying probability distribution on the feature space $\mathcal{X}$, and $S = \{\mathbf{x}_i\}_{i=1}^n \in \mathcal{X}^n$ is a set of samples drawn independent and identically distributed (i.i.d) from μ, where n is the sample size. The empirical distribution of S is defined as $\mu_n(S) = \frac{1}{n}\sum_{i=1}^n \mathbb{I}\{\mathbf{x}_i \in S\}$, where $\mathbb{I}$ denotes the indictor function. The samples are vectors in euclidian space $\mathbb{R}^m$ typically such that $m \gg 1$. The l_p norm on $\mathbb{R}^m$ is defined by $\|\mathbf{x}\|_p = (\sum_{i=1}^m |x_i|^p)^{1/p}$ when $1 \leq p < \infty$ and by $\|\mathbf{x}\|_\infty = \max_{1\leq i\leq m} |x_i|$ in the case $p = \infty$. The expectation and the variance of a random variable X are denoted as $\mathbb{E}(X)$ and $V(X)$ respectively. $\widetilde{\mathcal{O}}$ hides logarithmic terms.

3.2. Framework

Clustering aims to divide the samples into several clusters such that samples lying in the same cluster have more similarities than those in other clusters. Assume that a clustering task wants to partition the samples into K clusters, $W : \mathcal{X}^2 \to \mathbb{R}_+$ is a pairwise distance-based function used to measure the dissimilarity between pair observations, and $Z = [Z_1, ..., Z_K]$ is a collection of K partition functions $Z_k : \mathcal{X}^2 \to \mathbb{R}_+$ for $k = 1, ..., K$, which is together with the function W and used to divide the given dataset S into K disjoint clusters. The clustering framework can be cast as the problem of minimizing the following criterion:

$$\hat{L}_n(W, Z) = \frac{1}{n(n-1)} \sum_{k=1}^{K} \sum_{i,j=1, i\neq j}^{n} W(\mathbf{x}_i, \mathbf{x}_j) Z_k(\mathbf{x}_i, \mathbf{x}_j), \tag{1}$$

over all possible functions Z_k for $k = 1, ..., K$ and W. Assume that W and Z_k for $k = 1, .., K$ are symmetry, that is for all $(x, x') \in \mathcal{X}^2$, there have $D(x, x') = D(x', x)$ and $Z_k(x, x') = Z_k(x', x)$ for $k = 1, ..., K$.

To make the notations concise, let $f_{W,Z_k}(X, X') = W(X, X')Z_k(X, X')$ and $f_{W,Z} = (f_{W,Z_1}, ..., f_{W,Z_K})$ be a vector-valued function of the collection of $f_{W,Z_1}, ..., f_{W,Z_K}$, then $\hat{L}_n(W, Z)$ can be written as

$$\hat{L}_n(f_{W,Z}) = \frac{1}{n(n-1)} \sum_{k=1}^{K} \sum_{i,j=1, i\neq j}^{n} f_{W,Z_k}(\mathbf{x}_i, \mathbf{x}_j), \tag{2}$$

The expectation of the Eq. (2) is defined as:

$$L(f_{W,Z}) = \mathbb{E} \sum_{k=1}^{K} f_{W,Z_k}(X, X'). \tag{3}$$

where (X, X') is a pair of i.i.d random variables drawn from the distribution μ.

Let $\mathcal{F}$ be a family of vector-valued functions $f_{W,Z}$:

$$\mathcal{F} := \left\{ f_{W,Z} | f_{W,Z}(X, X') \in \mathbb{R}_+^K, \forall X, X' \in \mathcal{X} \right\}$$

and $\mathcal{F}_k$ be a function class of the output coordinate k of $\mathcal{F}$. Usually, $\hat{L}_n(f_{W,Z})$ is called empirical clustering risk, and $L(f_{W,Z})$ called (expected) clustering risk. Since the underlying distribution μ is unknown, a clustering algorithm always minimizes the empirical clustering risk $\hat{L}_n$ to obtain the final partitions, formalized as:

$$\hat{f}^*_{W,Z} = \arg\min_{f_{W,Z}\in\mathcal{F}} \hat{L}_n(f_{W,Z}),$$

over all possible vector-valued function $f_{W,Z}$ in function class $\mathcal{F}$. The optimal risk of the feature space $\mathcal{X}$ is the infimum of the clustering risk $L(f_{W,Z})$:

$$L^* = \inf_{f_{W,Z}\in\mathcal{F}} L(f_{W,Z}).$$

We now define the excess function class used in Section 4,

$$\mathcal{F}_{exc} := \left\{ f_{W,Z} - f^*_{W,Z} | f_{W,Z}, f^*_{W,Z} \in \mathcal{F} \right\},$$

where $f^*_{W,Z} = \arg\min_{f_{W,Z}\in\mathcal{F}} L(f_{W,Z})$.

By definition, the empirical excess clustering risk is $L(\hat{f}^*_{W,Z}) - L^*$, and the (expected) excess clustering risk is $\mathbb{E}[L(\hat{f}^*_{W,Z})] - L^*$, which are typical research objectives in learning theory (Li et al., 2019b;a; Yin et al., 2020; Kang et al., 2021), and are the main objectives this paper focuses on.

3.3. Examples

We now briefly explain our proposed clustering framework in the following examples.

Example 1: k-Means and Kernel k-Means. In k-means, we can set $W(x, x') = \|x - x'\|_2^2$, and set $Z_k(x, x') = \mathbb{I}\{(x, x') \in C_k^2\}$, where $k = 1, ..., K$ and where $C_1, ..., C_K$ are partitions of the feature space $\mathcal{X}$. In kernel k-means, we can set $W(x, x') = \phi(\|\psi(x) - \psi(x')\|_p)$, where $p \geq 1$, $\psi : \mathcal{X} \to \mathcal{H}$ is the feature map which maps $\mathcal{X}$ into a reproducing kernel Hilbert space (RKHS) $\mathcal{H}$ and $\phi : \mathbb{R}_+ \to \mathbb{R}_+$ is a monotone function such that $\phi(0) = 0$ and $\phi(t) > 0$ for all $t > 0$, and set $Z_k(x, x') = \mathbb{I}\{(\psi(x), \psi(x')) \in C_k^2\}$, where $k = 1, ..., K$ and where $C_1, ..., C_K$ are partitions of $\mathcal{H}$. In particular, k-means and kernel k-means belong to the hard clustering scheme. In the hard clustering scheme, a pair of observations can at most correspond to one cluster, which means that Z_k is valued either 0 or 1 for a pair of observations where $k = 1, ..., K$, and at most one valued 1. Thus the corresponding formula of the hard clustering scheme can be easily written.

Example 2: Soft k-Means. In soft k-means, or in the broader soft clustering scheme, the cluster 'center', such as the mean, median, or other metric measuring the 'center' of a cluster, is contributed by all the samples, different from the hard clustering scheme. In this case, the pairwise distance-based function W can also be chosen as $W(x, x') = \phi(\|x-$

$x'\|_p)$, where $p \geq 1$, and $\phi(x)$ is same as the hard clustering scheme; For any $k = 1, ..., K$, if $Z_k(x, x')$ is restricted to the samples, it should satisfy the following constraints:

$$Z_k(\mathbf{x}_i, \mathbf{x}_j) \in [0, 1], \quad \sum_{k=1}^{K} Z_k(\mathbf{x}_i, \mathbf{x}_j) = 1 \quad \forall i \neq j,$$

$$\text{and} \sum_{i,j=1, i\neq j}^{n} Z_k(\mathbf{x}_i, \mathbf{x}_j) \leq n.$$

Example 3: Spectral Clustering Scheme. Assume that the samples are constructed as the fully connected graph, W in this case can be chosen as the Gaussian Kernel function $W(x, x') = \exp(-\|x - x'\|_2^2/(2\sigma^2))$, where σ^2 is the variance determining the connectivity length scale (Von Luxburg, 2007); Since eigenvectors can be seen as empirical versions of underlying eigenfunctions (Rosasco et al., 2010), thus $Z_k(\mathbf{x}_i, \mathbf{x}_j) = (f_k(\mathbf{x}_i) - f_k(\mathbf{x}_j))^2$ if $Z_k(\cdot, \cdot)$ is restricted to samples $\mathbf{x}_i, \mathbf{x}_j$, where $f_k(\cdot)$ is the k-th eigenfunction, and $f_k(\mathbf{x}_i)$ and $f_k(\mathbf{x}_j)$ are elements of the corresponding k-th eigenvector. This definition of Z_k and W is suitable for NCut and RatioCut spectral clustering problem (Von Luxburg, 2007). Of course, if the samples are constructed as other types of graph, pairwise distance function $W(x, x')$ can be written as corresponding formulas.

Example 4: Neural Network Clustering Scheme. Assume that the feature map encoded by a neural network model after many non-linear layers is $\psi(x) : \mathcal{X} \to \mathcal{R}$, where $\mathcal{R}$ is a non-linear high dimensional function space. In this case, $W(x, x') = \phi(\|\psi(x) - \psi(x')\|_p)$, where $p \geq 1$, and $\phi(x)$ is same as the hard clustering scheme. If we use the hard clustering scheme, $Z_k(x, x') = \mathbb{I}\{(\psi(x), \psi(x') \in C_k^2\}$, where $k = 1, ..., K$ and where $C_1, ..., C_K$ are partitions of $\mathcal{R}$.

The above examples suggest that our proposed clustering framework is generalized well and suitable for a lot of clustering algorithms.

4. Sharper Excess Clustering Risk Bounds

In this section, we first introduce the assumptions used. Then, we prove that the clustering Rademacher complexity technique can just obtain an excess risk bound of slow convergence rate. Under the same condition of bounded hypothesis functions, we then derive a basic excess risk bound by the local clustering Rademacher complexity. This basic bound can be further together with some mild assumptions of the covering number and used to derive sharper bounds.

4.1. Assumptions

Assumption 1. *Assume that the hypothesis functions $f_{W,Z_k}(\cdot, \cdot) \in [0, M]$ for $k = 1, ..., K$ where $M > 0$ is a constant, and that $\sum_{k=1}^{K} f_{W,Z_k}(\cdot, \cdot) \in [0, E]$ where $E > 0$ is a constant.*

Assumption 1 is a very mild assumption. In examples 1, 2 and 4, it can be easily fulfilled if the pairwise distance function W is bounded or normalized because the partition functions $Z_1, ..., Z_K$ are obviously bounded. In example 3, if the function W is bounded, the corresponding integral operator related to W is a bounded operator and the eigenfunctions in RKHS are thus continuous and bounded (Rosasco et al., 2010), which means f_{W,Z_k} for $k = 1, ..., K$ are bounded. Obviously, $\sum_{k=1}^{K} f_{W,Z_k}$ is bounded if $f_{W,Z_1}, ..., f_{W,Z_K}$ are bounded. Besides, since $Z_1, ..., Z_K$ are usually indicator functions, which means $E = M$, so it is unreasonable to assume that $E \leq KM$.

Definition 1 (covering number (Zhou, 2002))**.** *For any $\epsilon > 0$ and a function class $\mathcal{H}$, the L_2 covering number $\mathcal{N}(\epsilon, \mathcal{H}, \|\cdot\|_{\ell_2(\mu_n(S))})$ is the supremum over samples $S = \{\mathbf{x}_1, ..., \mathbf{x}_n\}$ of the size of a minimal cover $\mathcal{C}_\epsilon$ such that $\forall f \in \mathcal{H}$, $\exists f_\epsilon \in \mathcal{C}_\epsilon$ s.t. $\sqrt{\frac{1}{n}\sum_{i=1}^{n}(f(\mathbf{x}_i) - f_\epsilon(\mathbf{x}_i))^2} \leq \epsilon$. Furthermore, the following covering number is introduced:*

$$\mathcal{N}(\epsilon, \mathcal{H}, \|\cdot\|_2) := \sup_n \sup_{\mu_n} \mathcal{N}\left(\epsilon, \mathcal{H}, \|\cdot\|_{\ell_2(\mu_n)}\right).$$

Assumption 2 (logarithmic covering number)**.** *Assume that there exist three positive constants γ, d and p satisfying $\log \mathcal{N}(\epsilon, \mathcal{F}_k, \|\cdot\|_2) \leq d \log^p(\gamma/\epsilon)$ for any $0 < \epsilon \leq \gamma$ and $k = 1, ..., K$.*

Many popular function classes satisfy Assumption 2 when the function classes $\mathcal{F}_k$ for $k = 1, ..., K$ are bounded:

1.) Any function space with finite VC-dimension (Vaart & Wellner, 1997), including linear functions and univariate polynomials of degree e (for which $d = e + 1$ and $p = 1$) as special cases; In particular, the corresponding function class of k-means clustering is a VC major class with finite VC dimension, see section 19.1 in (Devroye et al., 2013) for details.

2.) Any unit Euclidean ball $\mathcal{B}_2 \subset \mathbb{R}^d$ with fixed $\epsilon \in (0, 1)$ (Rigollet & Hütter, 2015).

3.) Any RKHS based on a kernel with rank d (Carl & Triebel, 1980).

Assumption 3 (polynomial covering number)**.** *Assume that there exist two constants $\gamma > 0$ and $p > 0$ satisfying $\log \mathcal{N}(\epsilon, \mathcal{F}_k, \|\cdot\|_2) \leq \gamma\epsilon^{-p}$ for any $k = 1, ..., K$.*

Classes that fulfill Assumption 3 are known as satisfying the uniform entropy condition (Wellner et al., 2013). If the function classes $\mathcal{F}_k$ for $k = 1, ..., K$ are bounded, this type of covering number is satisfied by many Sobolev/Besov classes (Gu, 2013). For instance, if the kernel eigenvalues decay at a rate of $t^{-\frac{2}{p}}$, where t denotes a sequence notation, then the RKHS satisfies this assumption of covering number (Carl & Triebel, 1980). The popular RKHSs of Gaussian, polynomial and finite rank kernels satisfy this assumption.

Assumption 4. *Assume that there exist two constants $\gamma > 0$ and $p > 0$ satisfying $\log \mathcal{N}(\epsilon, \mathcal{F}_k, \|\cdot\|_2) \le \gamma \epsilon^{-p} \log^2 \frac{2}{\epsilon}$ for any $k = 1, ..., K$.*

Assumption 4 combines the logarithmic and the polynomial covering number.

4.2. A Basic Excess Clustering Risk Bound

Rademacher complexity is widely used to measure the complexity of the function class and has been used for classification and regression problems (Bartlett & Mendelson, 2002; Koltchinskii et al., 2002). However, the traditional Rademacher complexity definition is not suitable for clustering learning. The summation over all pairs of observations in Eq. (1) makes its study more difficult, rendering standard techniques in the i.i.d case not applicable in this context. Inspired by (Clémençon et al., 2008; 2005), by using the permutations in U-process, we convert the non-sum-of-i.i.d pairwise function to a sum-of-i.i.d form. The empirical clustering Rademacher complexity thus defined on the $\mathcal{F}$ function space as follows:

Definition 2. *The empirical clustering Rademacher complexity of $\mathcal{F}$ is:*

$$R_n(\mathcal{F}) = \mathbb{E}_\sigma \left[\sup_{f_{W,Z} \in \mathcal{F}} \left| \frac{2}{\lfloor \frac{n}{2} \rfloor} \sum_{i=1}^{\lfloor \frac{n}{2} \rfloor} \sum_{k=1}^{K} \sigma_i f_{W,Z_k} \left(\mathbf{x}_i, \mathbf{x}_{i+\lfloor \frac{n}{2} \rfloor} \right) \right| \right],$$

where $\sigma_1, ..., \sigma_{\lfloor n/2 \rfloor}$ are i.i.d Rademacher variables, taking values in $\{-1, 1\}$ with equal probability. These Rademacher variables are independent of the samples $S = \{\mathbf{x}_1, ..., \mathbf{x}_n\}$. The expected clustering Rademacher complexity of $\mathcal{F}$ is defined as follows:

$$R(\mathcal{F}) = \mathbb{E}_S R_n(\mathcal{F}).$$

By the proposed clustering Rademacher complexity, we have the following Theorem.

Theorem 1. *Under Assumption 1, with probability $1 - \delta$, we have*

$$L(\hat{f}^*_{W,Z}) - L^* \le \frac{8KM}{\sqrt{n}} + E \sqrt{\frac{8 \log \frac{1}{\delta}}{n}}.$$

Remark 1. Theorem 1 suggests that the (empirical) excess clustering risk upper bound obtained by Rademacher complexity has a convergence rate of order $\mathcal{O}(K/\sqrt{n})$ provided that the hypothesis functions have bounded support.

It is worth noticing that Rademacher complexity considers the worst-case of the element in function space, neglecting that the algorithm will likely pick functions that have a small error. As a result, the best error rate that can be obtained via the Rademacher complexity is at least of order $\mathcal{O}(1/\sqrt{n})$. Indeed, the type of algorithms we consider here is known in the statistical literature as M-estimators. They minimize the empirical loss in a fixed class of functions. (Bartlett et al., 2005; Liu et al., 2019) demonstrate that the local Rademacher complexity is more reasonable to be served as a complexity measure. Since the local Rademacher complexity cannot be defined on the non-sum-of-i.i.d pairwise objective of Eq. (1), combined with *Definition 2*, we define the expectation local Rademacher complexity for clustering learning as follows.

Definition 3. *For any $r > 0$, the expectation local Rademacher complexity of the function class $\mathcal{F}_{exc}$ for clustering learning is defined as:*

$$\begin{aligned} R(\mathcal{F}^r_{exc}) := R\big(\{\alpha(f_{W,Z} - f^*_{W,Z}) \,|\alpha \in [0,1], \\ f_{W,Z} - f^*_{W,Z} \in \mathcal{F}_{exc}, \eta \le r\}\big), \end{aligned}$$

where $\eta := L[(\sum_{k=1}^K \alpha(f_{W,Z_k} - f^*_{W,Z_k}))^2]$, that is $\eta = \mathbb{E}[(\sum_{k=1}^K \alpha(f_{W,Z_k} - f^*_{W,Z_k}))^2]$.

If we pick a variance as small as possible while requiring that the $f_{W,Z} - f^*_{W,Z}$ is still in $\mathcal{F}^r_{exc}$, we can choose a much smaller class $\mathcal{F}^r_{exc} \subseteq \mathcal{F}_{exc}$ and obtain sharper excess clustering risk upper bounds. With the local clustering Rademacher complexity, we obtain a basic excess clustering risk upper bound.

Theorem 2. *Under Assumption 1, and let r^* be the fixed point of $R(\mathcal{F}^r_{exc})$, that is r^* is the solution of $R(\mathcal{F}^r_{exc}) = r$ with respect to r. Then, $\forall h > \max\left(1, \frac{\sqrt{2}}{4E}\right)$, with probability $1 - \delta$:*

$$L(\hat{f}^*_{W,Z}) - L^* \le c_{h,E} r^* + \frac{c_{h,\delta}}{n-1}, \tag{4}$$

where $c_{h,E}$ and $c_{h,\delta}$ are constants dependent on h, E and h, δ respectively.

A basic excess clustering risk upper bound is given by Theorem 2. We can choose suitable fixed points r^* to derive sharper bounds, which will be introduced in Section 4.3.

Remark 2. [Proof Techniques] [1] The generalization error bounds in (Bartlett et al., 2005; Koltchinskii et al., 2006) are complex, we therefore provide excess risk bound here, which need fine-grained generalization analysis (see proof of Theorem 2 in Appendix) and define the local Rademacher complexity on the excess function class $\mathcal{F}_{exc}$. [2] The Eq. (2) involves with pairwise observations, hinders the standard i.i.d technique of local Rademacher complexity (Bartlett et al., 2005) to apply to. To overcome this difficulty, we consider the U-process (Clémençon et al., 2005; 2008) to convert the Eq. (2) to a sum-of-i.i.d form (see Lemma 4 in

Appendix); [3] To present this basic excess clustering risk bound, we must use the Bennett concentration inequality (Bousquet, 2002a) to bound the empirical uniform deviation to get a term of order $\mathcal{O}(1/n)$, which can be together with the fixed point r^* to derive result (4) (see Lemma 4 and Lemma 5 in Appendix).

4.3. Sharper Excess Clustering Risk Bounds

In this paper, we use the covering number to bound the expected local clustering Rademacher complexity $R(\mathcal{F}^r_{exc})$ to obtain the fixed point r^* and derive the sharper excess clustering risk bounds. Covering number (Kolmogorov & Tikhomirov, 1959; Zhou, 2002) is also a tool to measure the complexity of the function class, which implies the 'richness' of a function class and is widely used in statistical learning theory. By the covering number, we can derive the following sharper excess risk bounds for clustering learning.

Theorem 3. *With different assumptions, we have the following different results:*

1.) Under Assumptions 1 and 2, with probability $1-\delta$*, we have*

$$L(\hat{f}^*_{W,Z}) - L^* \leq c_1 K^2 \frac{\log^p(n)}{n} + \frac{c_2}{n-1},$$

where c_1 *and* c_2 *are constants dependent on* γ, d, p, h, M, E *and* h, δ *respectively.*

2.) Under Assumptions 1 and 3, with probability $1-\delta$*, we have*

$$L(\hat{f}^*_{W,Z}) - L^* \leq c_1 K^2 \frac{1}{n^{\frac{2}{p+2}}} + \frac{c_2}{n-1},$$

where c_1 *and* c_2 *are constants dependent on* γ, p, h, M, E *and* h, δ *respectively.*

3.) Under Assumptions 1 and 4, with probability $1-\delta$*, we have*

$$L(\hat{f}^*_{W,Z}) - L^* \leq c_1 K^2 \frac{(\log n)^{\frac{2-p}{p+2}}}{n^{\frac{2}{p+2}}} \log \frac{n}{(\log n)^{\frac{2}{p+2}}} + \frac{c_2}{n-1},$$

where c_1 *and* c_2 *are constants dependent on* γ, p, h, M, E *and* h, δ *respectively.*

For result (1) in Theorem 3, the empirical excess clustering risk $L(\hat{f}^*_{W,Z}) - L^*$ has a convergence rate of order $\widetilde{\mathcal{O}}(K^2/n)$, which is much faster than $\mathcal{O}(K/\sqrt{n})$ because usually $K \ll n$. For results (2) and (3) in Theorem 3, the empirical excess clustering risk $L(\hat{f}^*_{W,Z}) - L^*$ have convergence rates of order $\mathcal{O}(K^2/n^{\frac{2}{p+2}})$ and $\widetilde{\mathcal{O}}(K^2/n^{\frac{2}{p+2}})$ respectively, which are all faster than $\mathcal{O}(K/\sqrt{n})$ when $0 < p < 2$. It is easy to verify that the upper bound of the expected excess clustering risk $\mathbb{E}[L(\hat{f}^*_{W,Z})] - L^*$ has the same order as $L(\hat{f}^*_{W,Z}) - L^*$.

According to (Hanneke, 2016; Zhivotovskiy & Hanneke, 2018; Ehrenfeucht et al., 1989), if a function class is a VC major class with finite VC dimension, which is a special case of Assumption 2, its excess risk lower bound is of order $\mathcal{O}(d/n)$. Since the upper bound of excess clustering risk in result (1) is also of order $\widetilde{\mathcal{O}}(d/n)$ (refer to Remark 4), thus the convergence rate obtained here is optimal in a minimax sense. Besides, (Rakhlin et al., 2017; Rakhlin & Sridharan, 2015) show that if there have the bounded loss function assumption and the polynomial covering number assumption on the hypothesis function space, which corresponds to Assumption 1 and Assumption 3, the lower bound obtained is of order $\mathcal{O}(n^{-\frac{2}{p+2}})$ for supervised learning problems. Since we obtain the upper bound of order $\mathcal{O}(n^{-\frac{2}{p+2}})$ for the more complex pairwise clustering learning problem under the same assumptions, thus this convergence rate in result (2) may be optimal rate as well.

Remark 3. [Proof Techniques] To present Theorem 3, we must establish the relationship between the expected local clustering Rademacher complexity and the covering number, however, the conventional Dudley's covering number bounds (Dudley, 1978; Bousquet, 2002b; Bartlett et al., 2005; Srebro & Sridharan, 2010) are built on the empirical constraints. In Definition 3, our local clustering Rademacher complexity has a constraint: $L[(\sum_{k=1}^K (f_{W,Z_k} - f^*_{W,Z_k}))^2] \leq r$ which is constructed on the expectation. To overcome this difficulty, we extend Theorem 2 in (Lei et al., 2016) to our pairwise clustering learning problem (see Lemma 6 in Appendix). Subsequently, by using the commonly used mild assumptions of covering number, the suitable fixed points r^* can be derived so that it can be substituted into the basic excess clustering risk bound of Theorem 2.

Remark 4. The factor K^2 appears because of the local Rademacher complexity technique. For example in result (1) of Theorem 3, we need to solve the following equality:

$$r^* = cK\left(x + \sqrt{r^* x}\right),$$

where $x = \frac{d \log^p(2\gamma n^{1/2})}{n}$ and c is a constant. Since usually $K \ll n$, so the upper bound in Theorem 3 is obviously faster than that in Theorem 1. For the extreme case of focusing on K, we further improve the order of K in Section 5.

5. Improve the Order of K

The results in Section 4 and the existing studies on data-dependent excess risk bounds for clustering (Levrard et al., 2013; Biau et al., 2008; Levrard et al., 2015; Calandriello & Rosasco, 2018; Clémençon, 2011) usually build on the following result for the Rademacher complexity:

$$R(\mathcal{F}) \leq K \max_k R(\mathcal{F}_k),$$

where $K \max_k R(\mathcal{F}_k)$ means the maximum Rademacher complexity of the restrictions of the function class along each coordinate with timing a factor of $\mathcal{O}(K)$. Therefore, the existing excess risk bounds are linearly dependent on K obtained by the clustering Rademacher complexity or K^2 obtained by the local clustering Rademacher complexity. However, for fine-grained analysis in the social network or recommendation systems, the number of clusters K may be very large. In this section, we reduce the order of K to improve the results of Section 4. Moreover, we will compare the results this paper obtained with the related work.

5.1. Improved Excess Clustering Risk Bounds

We first show that the clustering Rademacher complexity can be bounded by the maximum Rademacher complexity of the restrictions of the function class along each coordinate with timing a factor of $\mathcal{O}(\sqrt{K})$ by refining the Theorem 1 in (Foster & Rakhlin, 2019). Under the new clustering Rademacher complexity bound, we improve the convergence rates of the excess clustering risk of Section 4.

Assumption 5. *Assume the function $\sum_{k=1}^K f_{W,Z_k}$ to be L-lipschitz continuous with respect to the L_∞ norm, that is: $\forall f_{W,Z}, f'_{W,Z} \in \mathcal{F}$,*

$$\left\| \sum_{k=1}^K f_{W,Z_k} - \sum_{k=1}^K f'_{W,Z_k} \right\|_\infty \leq L \left\| f_{W,Z} - f'_{W,Z} \right\|_\infty .$$

Assumption 5 is a very mild assumption. In fact, the function $\sum_{k=1}^K f_{W,Z_k}$ in our proposed clustering framework is K-Lipschitz w.r.t. the L_∞ norm in the worst case (see Lemma 7 in the supplementary material for details). In particular, for the hard clustering scheme, such as k-means or kernel k-means, the function $\sum_{k=1}^K f_{W,Z_k}$ is 1-Lipschitz with respect to the L_∞ norm (see Lemma 7 in the supplementary material), which allows us to derive the state-of-the-art convergence rates for the hard clustering scheme.

Theorem 4. *Under Assumption 1 and 5, for any $\eta > 0$ and $S = \mathbf{x}_{i=1}^n \in \mathcal{X}^n$, there exists a constant $C > 0$ such that*

$$R_n(\mathcal{F}) \leq CL\sqrt{K} \max_k \widetilde{R}_n(\mathcal{F}_k) \log^{\frac{3}{2}+\eta}(\sqrt{n}),$$

where $\widetilde{R}_n(\mathcal{F}_k) = \sup_{\mathcal{S}\in\mathcal{X}^n} R_n(\mathcal{F}_k)$.

Remark 5. From Theorem 4, one can see that the clustering Rademacher complexity can be bounded by the maximum Rademacher complexity of the restrictions of the function class along each coordinate with timing a factor of $\mathcal{O}(\sqrt{K})$. It allows us to reduce the order of K of results of Section 4.

Remark 6. [Proof Techniques] The Theorem 1 in (Foster & Rakhlin, 2019) has a term $\max_i \widetilde{R}_n(\mathcal{F}_i)$ in the denominator, which hinders the construction of the relationship between the expected local clustering Rademacher complexity and the covering number. To present the following Theorems 5 and 6, we must prove the lower bound of the term $\max_i \widetilde{R}_n(\mathcal{F}_i)$ to overcome this difficulty, where the proof refers to using the Khintchine inequality (Haagerup, 1981) and the U-process technique (see the proof of Theorem 4 in the supplementary material for details).

Based on Theorem 4, we can derive new excess clustering risk upper bounds with a lower order of K.

Theorem 5. *Under Assumptions 1 and 5, for any $\eta > 0$, there exist a constant $C > 0$ such that with probability $1-\delta$,*

$$L(\hat{f}^*_{W,Z}) - L^* \leq 8MCL\sqrt{K}\frac{\log^{\frac{3}{2}+\eta}(\sqrt{n})}{\sqrt{n}} + E\sqrt{\frac{8\log\frac{1}{\delta}}{n}}.$$

From Theorem 5, one can see that the upper bound of the (empirical) excess clustering risk has a convergence rate of order $\widetilde{\mathcal{O}}(L\sqrt{K}/\sqrt{n})$. For the hard clustering scheme whose $L = 1$, the convergence rate is of order $\widetilde{\mathcal{O}}(\sqrt{K}/\sqrt{n})$. In other words, if there just have assumption 1, we improve the existing excess clustering risk upper bound to $\widetilde{\mathcal{O}}(\sqrt{K}/\sqrt{n})$ for the hard clustering scheme.

Theorem 6. *With different assumptions, we have the following different results:*

1.) *Under Assumptions 1, 2 and 5, for any $\eta > 0$, with probability $1-\delta$, we have*

$$L(\hat{f}^*_{W,Z}) - L^* \leq \mathcal{O}\left(L^2K\frac{\log^{3+p+2\eta}(\sqrt{n})}{n}\right).$$

2.) *Under Assumptions 1, 3 and 5, for any $\eta > 0$, with probability $1-\delta$, we have*

$$L(\hat{f}^*_{W,Z}) - L^* \leq \mathcal{O}\left(L^2K\frac{\log^{3+2\eta}(\sqrt{n})}{n^{\frac{2}{p+2}}}\right).$$

3.) *Under Assumptions 1, 4 and 5, for any $\eta > 0$, with probability $1-\delta$, we have*

$$L(\hat{f}^*_{W,Z}) - L^* \leq \mathcal{O}\left(L^2K\frac{(\log n)^r}{n^{\frac{2}{p+2}}}\log\frac{n}{(\log n)^{\frac{2}{p+2}}}\right),$$

where $r = \frac{2-p}{p+2} + 3 + 2\eta$.

Theorem 6 reduces the order of K for the upper bounds of the (empirical) excess clustering risk, from K^2 in Theorem 3 to K. For the hard clustering scheme whose $L = 1$, the convergence rate of excess clustering risk in result (1) of Theorem 6 is of order $\widetilde{\mathcal{O}}(K/n)$ and in results (2) and (3) of Theorem 6 are of order $\widetilde{\mathcal{O}}(K/n^{\frac{2}{p+2}})$, which are all faster than the results in Theorem 3. From Theorem 6, one can see that the best convergence rate obtained for (empirical) excess clustering risk is of order $\widetilde{\mathcal{O}}(K/n)$, which is the state-of-the-art convergence rate under mild assumptions.

5.2. Comparison with Related Work

Bounded Support. Suppose the distribution μ has bounded support, the excess clustering bounds are mostly of order $\mathcal{O}\big(\sqrt{\frac{K}{n}}\big)$ for finite-dimensional problems(Linder et al., 1994; Bartlett et al., 1998; Linder, 2000; 2002), or of order $\mathcal{O}\big(\frac{K}{\sqrt{n}}\big)$ (Biau et al., 2008; Maurer & Pontil, 2010; Canas et al., 2012; Fefferman et al., 2016; Calandriello & Rosasco, 2018; Fischer, 2010) for infinite-dimensional problems. For example, it is shown in (Bartlett et al., 1998) that if μ has bounded support, then

$$\mathbb{E}[L(\hat{f}^*_{W,Z})] - L^* \leq C \min\left(\frac{\sqrt{Km}}{\sqrt{n}}, \frac{\sqrt{K^{1-\frac{2}{m}} m \log n}}{\sqrt{n}}\right),$$

where m is the dimension of the sample $x \in \mathbb{R}^m$. Or, it is shown in (Biau et al., 2008) that if μ has bounded support R, then

$$\mathbb{E}[L(\hat{f}^*_{W,Z})] - L^* \leq 12KR^2/\sqrt{n}.$$

While supposing the hypothesis function has a bounded support, (Clémençcon, 2011; Cao et al., 2016) obtain the convergence rate of order $\mathcal{O}(\frac{K}{\sqrt{n}})$. For example, it is shown in (Clémençcon, 2011) that if the loss function has bounded support B, then with probability at least $1-\delta$,

$$L(\hat{f}^*_{W,Z}) - L^* \leq C_{\delta,B} K/\sqrt{n},$$

where $C_{\delta,B}$ is a constant dependent on B and δ.

From Theorem 5, one can see that our obtained bound for hard clustering scheme is of order $\mathcal{O}(\sqrt{K}/\sqrt{n})$ provided that there just have a bounded hypothesis function assumption. Note that this upper bound is suitable for infinite-dimensional problem, thus it is faster than $\mathcal{O}(K/\sqrt{n})$ that is provided for infinite dimension problem in (Biau et al., 2008; Maurer & Pontil, 2010; Calandriello & Rosasco, 2018; Clémençcon, 2011).

Margin Condition. As shown in Section 2, (Levrard et al., 2015) propose an assumption on μ, called margin condition, to satisfy the technical condition (Pollard et al., 1982; Chou, 1994; Antos et al., 2005; Levrard et al., 2013) and establish a fast convergence rate of order $\mathcal{O}(K^2/n)$. Let $\mathcal{M}$ be the set of all $c^* = \{c_1^*, ..., c_K^*\}$ where c^* is the set of optimal cluster centers constructed on the underlying distribution μ for (kernel) k-means algorithm. For $t \geq 0$, we define $p(t) := \sup_{c^* \in \mathcal{M}} P(f(c^*)^t)$, where, for any set $A \subset \mathcal{E}$, the term A^t stands for the t-neighborhood of A in $\mathcal{E}$ and where $f(c^*)$ denotes the frontier of the Voronoi diagram generated by c^*. Specifically, $f(c) := \bigcup_{i \neq j} U_i(c) \cap U_j(c)$ where $U_j(c) := \{x \in \mathcal{E} : \forall i \in \{1, ..., K\}, d(x, c_j) \leq d(x, c_i)\}$ and $d(\cdot,\cdot)$ is a distance metric defined on the metric space $\mathcal{E}$. The margin condition consists of: (1) for any $x \in \mathcal{E}$, $P(x : |x| \leq R) = 1$ for some $R > 0$; (2) suppose there exists $r_o > 0$ such that for all $0 < t < r_o$, $p(t) \leq \frac{Bp_{min}}{128R^2} t$, where $B := \inf_{c^* \in \mathcal{M}, i \neq j} |c_i^* - c_j^*|$ and $p_{min} := \inf_{c^* \in \mathcal{M}, 1 \leq j \leq K} P(U_j(c^*))$. Since $P(f(c^*)^t)$ corresponds to the probability mass of the frontier of the associated Voronoi diagram of c^* inflated by t, the margin condition suggests that if $p(t)$ does not increase too rapidly with t, the faster rate $\mathcal{O}(K^2/n)$ appears to be obtained. From the introduction of the margin condition, one can see that it is a strong assumption and difficult to be verified in general. Assume that the underlying distribution μ satisfies the margin condition with radius r_o, and let δ denote the quantity $\frac{p_{min} B^2 r_o^2}{64R^2} \wedge \epsilon$ where $\epsilon > 0$ is a constant derived by the margin condition (Levrard et al., 2015), (Levrard, 2018) refine the results in (Levrard et al., 2015):

$$\mathbb{E}[L(\hat{f}^*_{W,Z})] - L^* \leq \frac{C(K + \log(|\mathcal{M}|))R^2}{np_{min}} + \left[\frac{20KR^2}{\sqrt{n}} - \delta\right] \mathbb{I}_{\delta < \gamma} + \frac{R^2 e^{-\frac{n}{2R^4}\left((\delta - \frac{12KR^2}{\sqrt{n}})\right)^2}}{\sqrt{n}} \mathbb{I}_{\delta \geq \gamma},$$

where $\gamma = \frac{12KR^2}{\sqrt{n}}$. They show that if $\mathbb{E}[L(\hat{f}^*_{W,Z})] - L^* \leq \delta$, a faster convergence rate of order $\mathcal{O}(K/n)$ appears to be obtained. Their result is built on stronger assumptions than (Levrard et al., 2015) and also difficult to be verified in general. Moreover, (Terada & Yamamoto, 2019) give the excess risk bound of order $\mathcal{O}(K^2/n)$ for NCut (Von Luxburg, 2007) under the framework of margin condition.

Compared with these margin condition-based bounds, our excess clustering risk bounds are first suitable for many clustering algorithms, not just k-means. Besides, except for the bounded function assumption, we obtain the convergence rate of fast order $\mathcal{O}(K^2/n)$ just assuming mild assumptions of covering number on the function classes $\mathcal{F}_k$ for $k = 1, ..., K$. Since the covering numbers of many popular function classes are already known, thus our assumption is quite mild and easy to check. In addition, we obtain the state-of-the-art convergence rate of order $\mathcal{O}(K/n)$ for the hard clustering scheme under the same mild assumptions.

6. Conclusion

In this paper, we propose a clustering learning framework and analyze its excess risk. We all obtain sharper excess clustering upper bounds under two sets of mild assumptions: bounded support of hypothesis functions and assumptions of covering number on hypothesis function spaces, respectively. Besides, the state-of-the-art upper bound of order $\mathcal{O}(K/n)$ is derived for the hard clustering scheme under mild assumptions. We believe our work will provide a new studying perspective for clustering learning.

In future work, we will investigate the lower bound and the optimal rate for our proposed clustering framework.

7. Acknowledgment

This work was supported in part by the National Natural Science Foundation of China NO. 62076234, the Beijing Outstanding Young Scientist Program NO. BJJWZYJH012019100020098, China Unicom Innovation Ecological Cooperation Plan, the Public Computing Cloud of Renmin University of China, and the Major Innovation & Planning Interdisciplinary Platform for the "Double-First Class" Initiative, Renmin University of China.

References

Antos, A. Improved minimax bounds on the test and training distortion of empirically designed vector quantizers. *IEEE Transactions on Information Theory*, 51(11):4022–4032, 2005.

Antos, A., Gyorfi, L., and Gyorgy, A. Individual convergence rates in empirical vector quantizer design. *IEEE Transactions on Information Theory*, 51(11):4013–4022, 2005.

Bartlett, P. L. and Mendelson, S. Rademacher and gaussian complexities: risk bounds and structural results. *Journal of Machine Learning Research*, 3(Nov):463–482, 2002.

Bartlett, P. L., Linder, T., and Lugosi, G. The minimax distortion redundancy in empirical quantizer design. *IEEE Transactions on Information theory*, 44(5):1802–1813, 1998.

Bartlett, P. L., Bousquet, O., Mendelson, S., et al. Local rademacher complexities. *The Annals of Statistics*, 33(4): 1497–1537, 2005.

Biau, G., Devroye, L., and Lugosi, G. On the performance of clustering in hilbert spaces. *IEEE Transactions on Information Theory*, 54(2):781–790, 2008.

Bousquet, O. A bennett concentration inequality and its application to suprema of empirical processes. *Comptes Rendus Mathematique*, 334(6):495–500, 2002a.

Bousquet, O. *Concentration inequalities and empirical processes theory applied to the analysis of learning algorithms*. PhD thesis, Department of Applied Mathematics, École Polytechnique, 2002b.

Calandriello, D. and Rosasco, L. Statistical and computational trade-offs in kernel k-means. *Advances in Neural Information Processing Systems (NeurIPS 2018)*, 31: 9357–9367, 2018.

Canas, G., Poggio, T., and Rosasco, L. Learning manifolds with k-means and k-flats. *Advances in Neural Information Processing Systems (NIPS 2012)*, 25:2465–2473, 2012.

Cao, Q., Guo, Z.-C., and Ying, Y. Generalization bounds for metric and similarity learning. *Machine Learning*, 102 (1):115–132, 2016.

Carl, B. and Triebel, H. Inequalities between eigenvalues, entropy numbers, and related quantities of compact operators in banach spaces. *Mathematische Annalen*, 251(2): 129–133, 1980.

Chou, P. A. The distortion of vector quantizers trained on n vectors decreases to the optimum as $Op(1/n)$. In *Proceedings of 1994 IEEE International Symposium on Information Theory (ISIT 1994)*, pp. 457. IEEE, 1994.

Clémençcon, S. On U-processes and clustering performance. In *Advances in Neural Information Processing Systems (NIPS 2011)*, pp. 37–45, 2011.

Clémençon, S., Lugosi, G., and Vayatis, N. Ranking and scoring using empirical risk minimization. In *International Conference on Computational Learning Theory (COLT 2005)*, pp. 1–15. Springer, 2005.

Clémençon, S., Lugosi, G., Vayatis, N., et al. Ranking and empirical minimization of U-statistics. *The Annals of Statistics*, 36(2):844–874, 2008.

Devroye, L., Györfi, L., and Lugosi, G. *A probabilistic theory of pattern recognition*, volume 31. Springer Science & Business Media, 2013.

Dudley, R. M. Central limit theorems for empirical measures. *The Annals of Probability*, pp. 899–929, 1978.

Ehrenfeucht, A., Haussler, D., Kearns, M., and Valiant, L. A general lower bound on the number of examples needed for learning. *Information and Computation*, 82(3):247–261, 1989.

Fefferman, C., Mitter, S., and Narayanan, H. Testing the manifold hypothesis. *Journal of the American Mathematical Society*, 29(4):983–1049, 2016.

Fischer, A. Quantization and clustering with bregman divergences. *Journal of Multivariate Analysis*, 101(9):2207–2221, 2010.

Foster, D. J. and Rakhlin, A. l_∞ vector contraction for rademacher complexity. *arXiv preprint arXiv:1911.06468*, 2019.

Gu, C. *Smoothing spline ANOVA models*, volume 297. Springer Science & Business Media, 2013.

Haagerup, U. The best constants in the khintchine inequality. *Studia Mathematica*, 70:231–283, 1981.

Hanneke, S. Refined error bounds for several learning algorithms. *The Journal of Machine Learning Research*, 17 (1):4667–4721, 2016.

Jain, A. K. Data clustering: 50 years beyond k-means. *Pattern recognition letters*, 31(8):651–666, 2010.

Kang, Y., Liu, Y., Niu, B., and Wang, W. Weighted distributed differential privacy erm: Convex and non-convex. *Computers & Security*, 106:102275, 2021.

Kolmogorov, A. N. and Tikhomirov, V. M. ε-entropy and ε-capacity of sets in function spaces. *Uspekhi Matematicheskikh Nauk*, 14(2):3–86, 1959.

Koltchinskii, V., Panchenko, D., et al. Empirical margin distributions and bounding the generalization error of combined classifiers. *The Annals of Statistics*, 30(1): 1–50, 2002.

Koltchinskii, V. et al. Local rademacher complexities and oracle inequalities in risk minimization. *The Annals of Statistics*, 34(6):2593–2656, 2006.

Lei, Y., Ding, L., and Bi, Y. Local rademacher complexity bounds based on covering numbers. *Neurocomputing*, 218:320–330, 2016.

Levrard, C. Quantization/clustering: when and why does k-means work? *arXiv preprint arXiv:1801.03742*, 2018.

Levrard, C. et al. Fast rates for empirical vector quantization. *Electronic Journal of Statistics*, 7:1716–1746, 2013.

Levrard, C. et al. Nonasymptotic bounds for vector quantization in hilbert spaces. *The Annals of Statistics*, 43(2): 592–619, 2015.

Li, J., Liu, Y., Yin, R., Zhang, H., Ding, L., and Wang, W. Multi-class learning: From theory to algorithm. In *Advances In Neural Information Processing Systems (NeurIPS 2018)*, pp. 1586–1595, 2018.

Li, J., Liu, Y., Yin, R., and Wang, W. Approximate manifold regularization: Scalable algorithm and generalization analysis. In *International Joint Conference on Artificial Intelligence (IJCAI 2019)*, pp. 2887–2893, 2019a.

Li, J., Liu, Y., Yin, R., and Wang, W. Multi-class learning using unlabeled samples: Theory and algorithm. In *International Joint Conference on Artificial Intelligence (IJCAI 2019)*, pp. 2880–2886, 2019b.

Linder, T. On the training distortion of vector quantizers. *IEEE Transactions on Information Theory*, 46(4):1617–1623, 2000.

Linder, T. Learning-theoretic methods in vector quantization. In *Principles of nonparametric learning*, pp. 163–210. Springer, 2002.

Linder, T., Lugosi, G., and Zeger, K. Rates of convergence in the source coding theorem, in empirical quantizer design, and in universal lossy source coding. *IEEE Transactions on Information Theory*, 40(6):1728–1740, 1994.

Liu, F., Choi, D., Xie, L., and Roeder, K. Global spectral clustering in dynamic networks. *Proceedings of the National Academy of Sciences*, 115(5):927–932, 2018.

Liu, Y. and Liao, S. Eigenvalues ratio for kernel selection of kernel methods. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI 2015)*, number 1, pp. 2814–2820, 2015.

Liu, Y., Liao, S., Lin, H., Yue, Y., and Wang, W. Generalization analysis for ranking using integral operator. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI 2017)*, pp. 2273–2279, 2017a.

Liu, Y., Liao, S., Lin, H., Yue, Y., and Wang, W. Infinite kernel learning: generalization bounds and algorithms. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI 2017)*, pp. 2280–2286, 2017b.

Liu, Y., Liao, S., Jiang, S., Ding, L., Lin, H., and Wang, W. Fast cross-validation for kernel-based algorithms. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42(5):1083–1096, 2019.

Maurer, A. and Pontil, M. K-dimensional coding schemes in hilbert spaces. *IEEE Transactions on Information Theory*, 56(11):5839–5846, 2010.

Pollard, D. et al. A central limit theorem for k-means clustering. *The Annals of Probability*, 10(4):919–926, 1982.

Rakhlin, A. and Sridharan, K. Online nonparametric regression with general loss functions. *arXiv preprint arXiv:1501.06598*, 2015.

Rakhlin, A., Sridharan, K., Tsybakov, A. B., et al. Empirical entropy, minimax regret and minimax risk. *Bernoulli*, 23 (2):789–824, 2017.

Rigollet, P. and Hütter, J.-C. High dimensional statistics. *Lecture notes for course 18S997*, 2015.

Rosasco, L., Belkin, M., and De Vito, E. On learning with integral operators. *Journal of Machine Learning Research*, 11(1):905–934, 2010.

Shaham, U., Stanton, K., Li, H., Nadler, B., Basri, R., and Kluger, Y. Spectralnet: Spectral clustering using deep neural networks. *arXiv preprint arXiv:1801.01587*, 2018.

Srebro, N. and Sridharan, K. Note on refined dudley integral covering number bound. *Unpublished results. http://ttic. uchicago. edu/karthik/dudley. pdf*, 2010.

Tang, C. and Monteleoni, C. On lloyd's algorithm: New theoretical insights for clustering in practice. In *Artificial Intelligence and Statistics (AISTATS 2016)*, pp. 1280–1289, 2016.

Terada, Y. and Yamamoto, M. Kernel normalized cut: a theoretical revisit. In *International Conference on Machine Learning (ICML 2019)*, pp. 6206–6214, 2019.

Thorpe, M., Theil, F., Johansen, A. M., and Cade, N. Convergence of the k-means minimization problem using γ-convergence. *SIAM Journal on Applied Mathematics*, 75(6):2444–2474, 2015.

Vaart, A. v. d. and Wellner, J. A. Weak convergence and empirical processes with applications to statistics. *Journal of the Royal Statistical Society-Series A Statistics in Society*, 160(3):596–608, 1997.

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

Wellner, J. et al. *Weak convergence and empirical processes: with applications to statistics*. Springer Science & Business Media, 2013.

Xu, C., Liu, T., Tao, D., and Xu, C. Local rademacher complexity for multi-label learning. *IEEE Transactions on Image Processing*, 25(3):1495–1507, 2016.

Xu, R. and Wunsch, D. Survey of clustering algorithms. *IEEE Transactions on neural networks*, 16(3):645–678, 2005.

Yin, R., Liu, Y., Lu, L., Wang, W., and Meng, D. Divide-and-conquer learning with nyström: Optimal rate and algorithm. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI 2020)*, pp. 6696–6703, 2020.

Yousefi, N., Lei, Y., Kloft, M., Mollaghasemi, M., and Anagnostopoulos, G. C. Local rademacher complexity-based learning guarantees for multi-task learning. *The Journal of Machine Learning Research*, 19(1):1385–1431, 2018.

Zhivotovskiy, N. and Hanneke, S. Localization of vc classes: Beyond local rademacher complexities. *Theoretical Computer Science*, 742:27–49, 2018.

Zhou, D.-X. The covering number in learning theory. *Journal of Complexity*, 18(3):739–767, 2002.