Refined Learning Bounds for Kernel and Approximate k-Means

Anonymous Author(s)
Affiliation
Address
email

Abstract

Kernel k-means is one of the most popular approaches to clustering and its theoretical properties have been investigated for decades. However, the existing state-of-the-art risk bounds are of order $\mathcal{O}(k/\sqrt{n})$, which do not match with the stated lower bound $\Omega(\sqrt{k/n})$ in terms of k. In this paper, we study the statistical properties of kernel k-means and Nyström-based kernel k-means, and obtain optimal clustering risk bounds, which improve the existing risk bounds. Particularly, based on a refined upper bound of the clustering Rademacher complexity, we first derive an optimal risk bound of rate $\mathcal{O}(\sqrt{k/n})$ for empirical risk minimizer (ERM), and further extend it to general cases beyond ERM. Then, we analyze the statistical effect of computational approximations of Nyström kernel k-means, and prove that it achieves the same statistical accuracy as the original kernel k-means considering only $\Omega(\sqrt{nk})$ Nyström landmark points. We further relax the restriction of landmark points from $\Omega(\sqrt{nk})$ to $\Omega(\sqrt{n})$ under a mild condition. Finally, we validate the theoretical findings via numerical experiments.

1 Introduction

Clustering, a fundamental data mining task, is used in numerous applications including web search, medical imaging, gene expression analysis, social network analysis and recommendation systems [50, 49, 23, 36]. k-means is arguably one of the most popular approaches to clustering, producing clusters with piece-wise linear boundaries. Its kernel version, which employs a nonlinear distance function, has the ability to find clusters of varying densities and distributions, greatly improving the flexibility of the approach [18, 47].

To understand (kernel) k-means and guide the development of new clustering algorithms, researchers have investigated its theoretical properties for decades. The consistency of the empirical minimizer was demonstrated by [39, 41, 1]. Rates of convergence and non-asymptotic performance bounds were considered by [40, 13, 32, 7, 31, 14, 20]. Most of the proposed risk bounds are dependent upon the dimension of the hypothesis space. For example, Bartlett et al. [7] provided, under certain mild assumptions, a clustering risk bound of order $\mathcal{O}(\sqrt{kd/n})$, where d is the dimension of the hypothesis space and n is the size of the training set. However, the hypothesis space of kernel k-means is typically an infinite-dimensional Hilbert space, such as the reproducing kernel Hilbert space (RKHS) associated with Gaussian kernels [44]. Thus, the existing theoretical analysis of k-means are not usually suitable for explaining its kernel version. Recently, [20, 16, 10, 35, 3, 27, 24, 9] extended the previous results, and provided dimension-independent bounds for kernel k-means. As shown in [16], if the feature map associated with the kernel function satisfies $\|\Phi\| \leq 1$, then the clustering risk bounds are of order $\mathcal{O}(k/\sqrt{n})$. These clustering risk bounds for kernel k-means are usually linearly dependent on the number of clusters k. However, the number of clusters k may be very large in some domains, such as social networks and recommendation systems. Thus, from a theoretical perspective, these existing bounds of $\mathcal{O}(k/\sqrt{n})$ do not match with the stated lower bound $\Omega(\sqrt{k/n})$ in k [7].

Although kernel k-means is one of the most popular clustering methods, it requires the computation of a $n \times n$ kernel matrix. As for other kernel methods, this becomes unfeasible for large-scale problems, and thus deriving approximate computations, such as partial decompositions [6, 28], random projection [16, 15], Nyström approximations [19, 11, 9, 37, 47, 52, 51], and random feature approximations [42, 12, 5, 43, 33, 30], has become the subject of numerous recent works. However, few of these optimization-based methods focused on the underlying excess risk problem. To the best of our knowledge, the only two results providing excess risk guarantees for approximate kernel k-means are [16] and [9]. In [16], Devroye and Lugosi considered the excess clustering risk when the approximate Hilbert space is obtained using Gaussian projections. In [9], Calandriello and Rosasco showed that, when sampling $\Omega(\sqrt{n})$ Nyström landmarks, the excess risk bound can reach $\mathcal{O}(k/\sqrt{n})$. The excess risk bounds of [9] and [16] are both linearly dependent on k and thus do not match with the theoretical lower bound [7].

In this paper, we study the kernel k-means in terms of both statistical and computational requirements. Our major contributions include two parts:

1) A (nearly) optimal excess clustering risk bound of rate $\tilde{\mathcal{O}}(\sqrt{k/n})$[1] is proposed for empirical risk minimization (ERM) (see Theorem 1). To the best of our knowledge, this is the first (nearly) optimal excess risk bound for kernel k-means in terms of both k and n. Beyond ERM, we further extend the result of Theorem 1 to general cases (see Theorem 2 and Theorem 3).

2) A (nearly) optimal excess risk bound for Nyström kernel k-means is also obtained when sampling $\Omega(\sqrt{nk})$ points (see Theorem 4). We further relax the restriction of landmark points from $\Omega(\sqrt{nk})$ to $\Omega(\sqrt{n})$ (see Theorem 5) and extend it to general cases (see Theorem 6 and Theorem 7). This result shows that we can use the Nyström method to improve the effectiveness of kernel k-means, while guaranteeing the optimal generalization performance.

The rest of the paper is organized as follows. In Section 2, we introduce some notations and provide an overview of kernel k-means. In Section 3, we provide nearly optimal excess risk bounds. In Section 4, we quantify the statistical effect of computational approximations of the Nyström-based kernel k-means. In Section 5, we validate our theoretical findings by performing experiments on simulated data. We end in Section 6 with conclusion. All the detailed proofs are deferred to the Appendix.

2 Background

In this section, we will introduce some notations and provide a brief introduction of kernel k-means. Please refer to [18, 9] for more details.

2.1 Notations

Assume $\mathbb{P}$ is a (unknown) distribution on $\mathcal{X}$, and $\mathcal{S} = \{\mathbf{x}_i\}_{i=1}^n \in \mathcal{X}$ is a set of n samples drawn i.i.d. from $\mathbb{P}$. We denote $\mathbb{P}_n(\mathcal{S}) = \frac{1}{n}\sum_{i=1}^n \mathbf{1}\{\mathbf{x}_i \in \mathcal{S}\}$ as the *empirical* distribution. Let $\kappa : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be a mercer kernel [45], and $\mathcal{H}$ be its associated RKHS [46], which is the completion of the linear span of the set of functions: $\mathcal{H} = \overline{\mathrm{span}\{\kappa(\mathbf{x}, \cdot), \mathbf{x} \in \mathcal{X}\}}$. We denote the Cartesian product of $\mathcal{H}$ by $\mathcal{H}^k = \otimes_{i=1}^k \mathcal{H}$. We use the *feature map* $\psi : \mathcal{X} \to \mathcal{H}$ to map $\mathcal{X}$ into the Hilbert space $\mathcal{H}$, and assume that $\mathcal{H}$ is separable, such that for any $\mathbf{x} \in \mathcal{X}$, we have $\Phi_\mathbf{x} = \psi(\mathbf{x})$. Intuitively, in the rest of the paper, the reader can assume that $\Phi_\mathbf{x} \in \mathbb{R}^d$ with $d \gg n$ or even infinite. From here on, we will denote the inner product of $\mathcal{H}$ by $\langle \cdot, \cdot \rangle$, and the associated norm by $\|\cdot\|$, and assume that $\|\Phi_\mathbf{x}\| \le 1$ for any $\mathbf{x} \in \mathcal{X}$. We let $\mathcal{D} = \{\Phi_i = \psi(\mathbf{x}_i)\}_{i=1}^n$, and denote $[\mathbf{K}]_{i,j} = \kappa(\mathbf{x}_i, \mathbf{x}_j) = \langle \Phi_i, \Phi_j \rangle$ as the kernel matrix.

The notations $\mu = \mathcal{O}(\nu)$ and $\mu = \Omega(\nu)$ mean that there exist constants c, c_1, c_2 such that $\mu \le c\nu$ and $c_1\nu \le \mu \le c_2\nu$, respectively. We use $\tilde{\mathcal{O}}$ and $\tilde{\Omega}$ to hide logarithmic terms.

[1] $\tilde{\mathcal{O}}$ hides logarithmic terms.

2.2 Kernel k-Means

In this paper, we aim at partitioning the given dataset into k disjoint $clusters$, each characterized by its $centroid$ $\mathbf{c}_j$. The Voronoi cell associated with a centroid $\mathbf{c}_j$ is defined as [9]

$$\mathcal{C}_j := \left\{ i : j = \arg\min_{s=1,\ldots,k} \|\Phi_i - \mathbf{c}_s\|^2 \right\}.$$

Let $\mathbf{C} = [\mathbf{c}_1, \ldots, \mathbf{c}_k]$ be a collection of k centroids from $\mathcal{H}^k$. In this paper, we focus on the so-called $kernel$ k-$means$ clustering, by minimizing the $empirical\ squared\ norm\ criterion$

$$\mathcal{W}(\mathbf{C}, \mathbb{P}_n) := \frac{1}{n} \sum_{i=1}^{n} \min_{j=1,\ldots,k} \|\Phi_i - \mathbf{c}_j\|^2 \tag{1}$$

over all possible choices of cluster centers $\mathbf{C} \in \mathcal{H}^k$. From [18, 9], we know that $\mathcal{W}(\mathbf{C}, \mathbb{P}_n)$ can be written as

$$\begin{aligned}
\mathcal{W}(\mathbf{C}, \mathbb{P}_n) :&= \frac{1}{n} \min_{\mathbf{C} \in \mathcal{H}^k} \sum_{j=1}^{k} \sum_{i \in \mathcal{C}_j} \left\| \Phi_i - \frac{1}{|\mathcal{C}_j|} \sum_{t \in \mathcal{C}_j} \Phi_t \right\|^2 \\
&= \frac{1}{n} \min_{\mathbf{C} \in \mathcal{H}^k} \sum_{j=1}^{k} \sum_{i \in \mathcal{C}_j} \left(\kappa(\mathbf{x}_i, \mathbf{x}_i) - \frac{2}{|\mathcal{C}_j|} \sum_{t \in \mathcal{C}_j} \kappa(\mathbf{x}_i, \mathbf{x}_t) + \frac{1}{|\mathcal{C}_j|^2} \sum_{t, t' \in \mathcal{C}_j} \kappa(\mathbf{x}_t, \mathbf{x}_{t'}) \right).
\end{aligned}$$

The $empirical\ risk\ minimizer\ (ERM)$ is defined as

$$\mathbf{C}_n := \arg\min_{\mathbf{C} \in \mathcal{H}^k} \mathcal{W}(\mathbf{C}, \mathbb{P}_n). \tag{2}$$

The performance of a clustering scheme given by the collection $\mathbf{C} = [\mathbf{c}_1, \ldots, \mathbf{c}_k] \in \mathcal{H}^k$ of cluster centers is usually measured by the $expected\ squared\ norm\ criterion$ or $expected\ clustering\ risk$

$$\mathcal{W}(\mathbf{C}, \mathbb{P}) := \int \min_{j=1,\ldots,k} \|\Phi_{\mathbf{x}} - \mathbf{c}_j\|^2 d\mathbb{P}(\mathbf{x}).$$

Given a $\mathbf{C} \in \mathcal{H}^k$, let $f_{\mathbf{C}} = (f_{\mathbf{c}_1}, \ldots, f_{\mathbf{c}_k})$ be a k-$valued$ function of the collection $\mathbf{C}$ with $f_{\mathbf{c}_j}(\mathbf{x}) = \|\Phi_{\mathbf{x}} - \mathbf{c}_j\|^2$ and $\mathcal{F}_{\mathbf{C}}$ be a family of k-$valued$ functions with

$$\mathcal{F}_{\mathbf{C}} := \Big\{ f_{\mathbf{C}} = (f_{\mathbf{c}_1}, \ldots, f_{\mathbf{c}_k}) : \mathbf{C} \in \mathcal{H}^k \Big\}. \tag{3}$$

Let $\varphi : \mathbb{R}^k \to \mathbb{R}$ be a minimum function:

$$\forall \boldsymbol{\alpha} \in \mathbb{R}^k, \varphi(\boldsymbol{\alpha}) = \min_{i=1,\ldots,k} \alpha_i \tag{4}$$

and $\mathcal{G}_{\mathbf{C}}$ be a “minimum" family of the functions $\mathcal{F}_{\mathbf{C}}$,

$$\mathcal{G}_{\mathbf{C}} : = \Big\{ g_{\mathbf{C}} = \varphi \circ f_{\mathbf{C}} \;\Big|\; f_{\mathbf{C}} \in \mathcal{F}_{\mathbf{C}}, g_{\mathbf{C}}(\mathbf{x}) = \varphi(f_{\mathbf{C}}(\mathbf{x})) \Big\}. \tag{5}$$

From the definition of $\varphi(f_{\mathbf{C}}(\mathbf{x})) = \min(f_{\mathbf{c}_1}(\mathbf{x}), \ldots, f_{\mathbf{c}_k}(\mathbf{x}))$, one can see that the empirical and expected squared norm criteria can be respectively written as

$$\mathcal{W}(\mathbf{C}, \mathbb{P}_n) := \frac{1}{n} \sum_{i=1}^{n} \varphi(f_{\mathbf{C}}(\mathbf{x}_i)) \text{ and } \mathcal{W}(\mathbf{C}, \mathbb{P}) := \int \varphi(f_{\mathbf{C}}(\mathbf{x})) d\mathbb{P}(\mathbf{x}).$$

In this paper, we consider bounding the $excess\ clustering\ risk$ $\mathcal{E}(\mathbf{C}_n)$ of the empirical risk minimizer [16]:

$$\mathcal{E}(\mathbf{C}_n) := \mathbb{E}_{\mathcal{D}}[\mathcal{W}(\mathbf{C}_n, \mathbb{P})] - \mathcal{W}^*(\mathbb{P}),$$

where $\mathcal{W}^*(\mathbb{P}) = \inf_{\mathbf{C} \in \mathcal{H}^k} \mathcal{W}(\mathbf{C}, \mathbb{P})$ is the $optimal$ clustering risk. In the following, we will ignore the subscript $\mathcal{D}$ if the input dataset $\mathcal{D}$ is clear.

2.3 The Existing Excess Clustering Risk Bounds

According to [7], we know that there exists a collection of centroids $\mathbf{C} \in \mathcal{H}^k$, a constant c, and a distribution $\mathbb{P}$ with $\|\Phi_{\mathbf{x}}\| \leq 1$ for any $\mathbf{x} \in \mathcal{X}$, such that

$$\mathbb{E}[\mathcal{W}(\mathbf{C}, \mathbb{P})] - \mathcal{W}^*(\mathbb{P}) \geq c\sqrt{\frac{k^{1-4/d}}{n}}.$$

Note that d is the dimension of $\Phi_{\mathbf{x}}$, which is usually very large or even infinite. Thus, the lower bound of kernel k-means is $\Omega\big(\sqrt{k/n}\big)$. However, most of the existing risk bounds proposed for kernel k-means are $\mathcal{O}(k/\sqrt{n})$ [16, 10, 35, 20, 9]:

Lemma 1 ([16], Theorem 2.1)**.** *If $\|\Phi_{\mathbf{x}}\| \leq 1$ for any $\mathbf{x} \in \mathcal{X}$, then there exists a constant c such that*

$$\mathbb{E}[\mathcal{W}(\mathbf{C}_n, \mathbb{P})] - \mathcal{W}^*(\mathbb{P}) \leq c\frac{k}{\sqrt{n}},$$

where $\mathbf{C}_n$ is the ERM of $\mathcal{W}(\mathbf{C}, \mathbb{P}_n)$ defined in (2)*.*

Note that the number of clusters k may be very large for fine-grained analyses in social networks or recommendation systems. This leaves us with the question: is it possible to prove a bound of rate $\sqrt{k/n}$, which is (nearly) optimal in terms of both k and n? In this paper, we attempt to answer this.

3 Main Results

In this section, we will provide nearly optimal excess risk bounds for kernel k-means. There are very few works focus on the underlying excess risk problem for kernel k-means. To the best of our knowledge, there are only two results [16, 9] providing excess risk bounds for kernel k-means or approximate kernel k-means. However, these bounds of [16, 9] are all linearly dependent on k. Based on a refined upper bound of clustering Rademacher complexity (see Lemma 3 for detail), we derive a (nearly) optimal excess risk bound of linearly dependent on $\sqrt{k}$.

Theorem 1. *If $\forall \mathbf{x} \in \mathcal{X}, \|\Phi_{\mathbf{x}}\| \leq 1$, then for any $\delta \in (0,1)$, there exists a constant c, and with probability at least $1-\delta$, we have,*

$$\mathbb{E}[\mathcal{W}(\mathbf{C}_n, \mathbb{P})] - \mathcal{W}^*(\mathbb{P}) \leq c\left(\sqrt{\frac{k}{n}}\log^2\left(\sqrt{n}\right) + \sqrt{\frac{\log\frac{1}{\delta}}{n}}\right).$$

From Theorem 1, we know that

$$\mathbb{E}[\mathcal{W}(\mathbf{C}_n, \mathbb{P})] - \mathcal{W}^*(\mathbb{P}) \leq \tilde{\mathcal{O}}\left(\sqrt{\frac{k}{n}}\right),$$

which matches the theoretical lower bound $\Omega\big(\sqrt{k/n}\big)$ when d is large [7]. Thus, our proposed bound is **(nearly) optimal**.

Remark (Fast Rates). Some results suggest that the learning rate of kernel k-means can reach $\mathcal{O}(k/n)$ under certain assumptions on the distribution. Chou [13] pointed out that, if continuous densities of distribution satisfy certain regularity properties, the expected excess risk is of rate $\mathcal{O}(k/n)$. An improved result was obtained by [3], who proved that the learning rate can reach $\mathcal{O}(k/n)$ for any distribution supported on a finite set. Levrard [27] further showed that, if the distribution satisfies a margin condition, the learning rate can also reach $\mathcal{O}(k/n)$. Based on the notion of local Rademacher complexity, the expected excess risk has a rate faster than $\mathcal{O}(k/\sqrt{n})$ given in [24, 29]. However, as pointed out, these conditions are difficult to verify in general. Moreover, these expected excess risk bounds are linearly dependent on k. In the future, we will consider studying whether it is possible to prove a bound of $\mathcal{O}(\sqrt{k}/n)$ under certain strict assumptions.

3.1 Further Results: Beyond ERM

So far we have provided guarantees for $\mathbf{C}_n$, that is, the optimal ERM in $\mathcal{H}^k$. Note that obtaining the optimal ERM $\mathbf{C}_n$ is a NP-hard problem in general [2]. In the following, we will consider the risk bound for a general $\tilde{\mathbf{C}}_n$, which only requires that its empirical squared norm criterion is not far from that of $\mathbf{C}_n$.

Theorem 2. *If* $\forall \mathbf{x} \in \mathcal{X}, \|\Phi_{\mathbf{x}}\| \leq 1$ *and*

$$\mathbb{E}\big[\mathcal{W}(\tilde{\mathbf{C}}_n, \mathbb{P}_n) - \mathcal{W}(\mathbf{C}_n, \mathbb{P}_n)\big] \leq \zeta,$$

then for any $\delta \in (0,1)$*, there exists a constant c and, with probability at least* $1-\delta$*, we have*

$$\mathbb{E}[\mathcal{W}(\tilde{\mathbf{C}}_n, \mathbb{P})] - \mathcal{W}^*(\mathbb{P}) \leq c\sqrt{\frac{k}{n}} \log^2\left(\sqrt{n}\right) + c\sqrt{\frac{\log \frac{1}{\delta}}{n}} + \zeta.$$

From the above theorem, one can see that if the discrepancy between the empirical squared norm criterion of $\tilde{\mathbf{C}}_n$ and $\mathbf{C}_n$ is small, that is $\zeta \leq \mathcal{O}(\sqrt{k/n})$, the risk bound of $\tilde{\mathbf{C}}_n$ is (nearly) optimal.

3.2 Further Results: k-means++

Lloyd's algorithm [34] is the most popular k-means algorithm and when coupled with a careful k-means++ seeding [4], a good approximate solution $\tilde{\mathbf{C}}_n$ can be obtained. Recently, based on a simple combination of k-means++ sampling and a local search strategy, an improved k-means++ algorithm was proposed [25]. It was shown that the empirical squared norm criterion of $\tilde{\mathbf{C}}_n$ can be up to a constant factor from the optimal empirical solution:

Lemma 2 ([25])**.** *If* $\mathbf{C}_n^{\mathcal{A}}$ *is returned by the improved* k*-means++ algorithm with a local search strategy [25], then*

$$\mathbb{E}_{\mathcal{A}}[\mathcal{W}(\mathbf{C}_n^{\mathcal{A}}, \mathbb{P}_n)] \leq \beta \cdot \mathcal{W}(\mathbf{C}_n, \mathbb{P}_n),$$

where β *is a constant and* $\mathcal{A}$ *is the randomness derived from the* k*-means++ initialization.*

Please refer to [25] for details. Note that we can use the algorithm from [25] for kernel k-means by replacing the Euclidean distance $\|\mathbf{x}_i - \mathbf{x}_j\|^2$ with $\|\Phi_i - \Phi_j\|_{\mathcal{H}}^2 = \kappa(\mathbf{x}_i, \mathbf{x}_i) - 2\kappa(\mathbf{x}_i, \mathbf{x}_j) + \kappa(\mathbf{x}_j, \mathbf{x}_j)$. In the following, we derive a risk bound for $\mathbf{C}_n^{\mathcal{A}}$.

Theorem 3. *If* $\forall \mathbf{x} \in \mathcal{X}, \|\Phi_{\mathbf{x}}\| \leq 1$*, and* $\mathbf{C}_n^{\mathcal{A}}$ *is returned by the improved* k*-means++ algorithm with a local search strategy [25], then for any* $\delta \in (0,1)$*, with a probability at least* $1-\delta$*, we have*

$$\mathbb{E}_{\mathcal{D}}\left[\mathbb{E}_{\mathcal{A}}[\mathcal{W}(\mathbf{C}_n^{\mathcal{A}}, \mathbb{P})]\right] \leq \tilde{\mathcal{O}}\left(\sqrt{\frac{k}{n}} + \mathcal{W}^*(\mathbb{P})\right).$$

The above result implies that if the optimal clustering risk $\mathcal{W}^*(\mathbb{P})$ is small, the risk of $\mathcal{W}(\mathbf{C}_n^{\mathcal{A}}, \mathbb{P})$ can reach $\tilde{\mathcal{O}}(\sqrt{k/n})$.

4 Risk Analysis of Nyström Kernel k-Means

Kernel k-means is one of the most popular clustering methods. However, it requires the computation of a $n \times n$ kernel matrix. This renders it non-scalable to large datasets that contain more than a few tens of thousands of points. In particular, simply constructing and storing the kernel matrix $\mathbf{K}$ takes $O(n^2)$ time and space.

The Nyström method [19] is a popular method for approximating the kernel matrix. The properties of Nyström approximations for kernel k-means have recently been studied in [11, 15, 37, 9, 47, 53]. However, most of these works focus on the computation area. To the best of our knowledge, the only study providing excess risk guarantees for the Nyström kernel k-means is [9]. However, its excess risk bound is linearly dependent on k. In the following, we will improve it from k to $\sqrt{k}$.

4.1 Nyström Kernel k-Means

To derive the excess risk bound of Nyström kernel k-means, we first briefly introduce some notations. Given a dataset $\mathcal{D} = \{\Phi_i\}_{i=1}^n$, we use

$$\mathcal{I} = \{\Phi_i\}_{i=1}^m \subseteq \mathcal{D}$$

as a collection of landmark points to replace $\mathcal{D}$. Let $\mathcal{H}_m$ be a linear span of $\mathcal{I} = \{\Phi_i\}_{i=1}^m$,

$$\mathcal{H}_m = \operatorname{span}\left\{\sum_{i=1}^m \alpha_i \Phi_i, \alpha_i \in \mathbb{R}, \Phi_i \in \mathcal{I}\right\},$$

and $\mathcal{H}_m^k = \otimes_{i=1}^k \mathcal{H}_m$ be its Cartesian product. The Nyström kernel k-means, i.e., the approximate kernel k-means over $\mathcal{H}_m^k$, can be written as [9]:

$$\mathbf{C}_{n,m} = \underset{\mathbf{C} \in \mathcal{H}_m^k}{\arg\min} \frac{1}{n} \sum_{i=1}^n \min_{j=1,\ldots,k} \|\Phi_i - \mathbf{c}_j\|^2 .$$

The Nyström kernel k-means can be done in $\mathcal{O}(nm)$ space and $\mathcal{O}(nmkt + nm^2)$ time using t steps of Lloyd's algorithm for k clusters [34]. Please refer to [9] for more details.

4.2 Excess Risk Bound of Nyström Kernel k-Means

Denote with $\Xi = \operatorname{Tr}(\mathbf{K}^{\mathrm{T}}(\mathbf{K}+\mathbf{I})^{-1})$ the so-called effective dimension of $\mathbf{K}$ [43, 9]. Note that

$$\operatorname{Tr}\left(\mathbf{K}^{\mathrm{T}}(\mathbf{K}+\mathbf{I})^{-1}\right) \leq \operatorname{Tr}\left(\mathbf{K}^{\mathrm{T}}(\mathbf{K})^{+}\right),$$

so we can obtain that $\Xi \leq \operatorname{Rank}(\mathbf{K})$. Thus, the effective dimension Ξ can be seen as a soft version of the rank.

Theorem 4. *If $\forall \mathbf{x} \in \mathcal{X}, \|\Phi_{\mathbf{x}}\| \leq 1$, and the size of a uniform sampling is*

$$m \geq \Omega\left(\frac{\sqrt{n}\log(1/\delta)\min(k,\Xi)}{\sqrt{k}}\right),$$

then, with probability at least $1-\delta$, we have

$$\mathbb{E}[\mathcal{W}(\mathbf{C}_{n,m}, \mathbb{P})] - \mathcal{W}^*(\mathbb{P}) \leq \mathcal{O}\left(\sqrt{\frac{k}{n}}\log\left(\frac{n}{\delta}\right)\right).$$

Note that

$$\frac{\sqrt{n}\min(k,\Xi)}{\sqrt{k}} \leq \sqrt{nk}.$$

Thus, from a statistical point, Theorem 4 shows that when sampling $\tilde{\Omega}(\sqrt{nk})$ points, the Nyström kernel k-means achieves the same excess risk as the exact one does. This result demonstrates that we can improve the computational aspect of kernel k-means using Nyström embedding, while maintaining **optimal** generalization guarantees.

Remark. Calandriello and Rosasco [9] have reported that if $m \geq \tilde{\Omega}(\sqrt{n})$, an excess risk bound of rate $\tilde{\mathcal{O}}(k/\sqrt{n})$ for Nyström kernel k-means can be obtained, which seems to be better than our $\tilde{\Omega}(\sqrt{nk})$. However, it should be noted that the risk bound in [9] is linearly dependent on k, while ours is linearly dependent on $\sqrt{k}$. From the proof of Lemma 10, if we want to obtain a risk of linear dependence on k, we only need

$$m \geq \Omega\left(\frac{\sqrt{n}\log(1/\delta)\min(k,\Xi)}{k}\right) = \tilde{\Omega}(\sqrt{n}),$$

which is the same as [9]. In the following, we will show that we can relax the restriction of landmark points under a mild condition.

4.3 Further Results: Reducing the Sampling Points

From Theorem 4, we know that we need $\tilde{\Omega}(\sqrt{nk})$ sampling points to guarantee the nearly optimal rate for approximating kernel k-means. In the following, we show how to reduce the sampling points from $\tilde{\Omega}(\sqrt{nk})$ to $\tilde{\Omega}(\sqrt{n})$ under a basic assumption on the eigenvalues of the kernel matrix.

Theorem 5. *Let λ_i be the i-th eigenvalue of the kernel matrix $\mathbf{K}$, $i = 1, \ldots, n$, and $\lambda_{i+1} \leq \lambda_i$. If $\forall \mathbf{x} \in \mathcal{X}, \|\Phi_\mathbf{x}\| \leq 1$, the eigenvalues satisfy the assumption*

$$\exists \alpha > 1, c > 0 : \lambda_i \leq c i^{-\alpha},$$

and the size of an uniform sampling is

$$m \geq \Omega\left(\sqrt{n} \log(1/\delta)\right).$$

then, with probability at least $1 - \delta$, we have

$$\mathbb{E}[\mathcal{W}(\mathbf{C}_{n,m}, \mathbb{P})] - \mathcal{W}^*(\mathbb{P}) \leq \mathcal{O}\left(\sqrt{\frac{k}{n}} \log\left(\frac{n}{\delta}\right)\right).$$

The assumption of algebraically decreasing eigenvalues of the kernel matrix is a common assumption, and met by the popular finite rank kernels and shift invariant kernel [48], for example. The above results show that we can guarantee the optimal generalization performance when only sampling $\tilde{\Omega}(\sqrt{n})$ points, which is much better than $\tilde{\Omega}(\sqrt{nk})$ when k is large.

4.4 Further Results: Beyond ERM

In the following, we show that our result can be extended to general cases beyond ERM.

Theorem 6. *Under the same assumptions as Theorem 5, if*

$$\mathbb{E}\big[\mathcal{W}(\tilde{\mathbf{C}}_{n,m}, \mathbb{P}_n) - \mathcal{W}(\mathbf{C}_{n,m}, \mathbb{P}_n)\big] \leq \zeta,$$

and the size of an uniform sampling is

$$m \geq \Omega\left(\sqrt{n} \log(1/\delta)\right),$$

then, with probability at least $1 - \delta$, we have

$$\mathbb{E}[\mathcal{W}(\tilde{\mathbf{C}}_{n,m}, \mathbb{P})] - \mathcal{W}^*(\mathbb{P}) \leq \tilde{\mathcal{O}}\left(\sqrt{\frac{k}{n}} + \zeta\right).$$

The above result demonstrates that the risk bound of $\tilde{\mathbf{C}}_{n,m}$ is optimal when $\mathbb{E}\big[\mathcal{W}(\tilde{\mathbf{C}}_{n,m}, \mathbb{P}_n) - \mathcal{W}(\mathbf{C}_{n,m}, \mathbb{P}_n)\big]$ is small.

4.5 Further Results: k-means ++

If adopting the improved k-kernel means++ sampling with a local search strategy [25] for Nyström kernel k-means, we can obtain the following results:

Theorem 7. *Under the same assumptions as Theorem 5, $\mathbf{C}^{\mathcal{A}}_{n,m}$ is returned by the improved k-means++ algorithm with a local search strategy [25], if the size of an uniform sampling is*

$$m \geq \Omega\left(\sqrt{n} \log(1/\delta)\right),$$

then with probability at least $1 - \delta$, we have

$$\mathbb{E}_{\mathcal{D}}\left[\mathbb{E}_{\mathcal{A}}[\mathcal{W}(\mathbf{C}^{\mathcal{A}}_{n,m}, \mathbb{P})]\right] \leq \tilde{\mathcal{O}}\left(\sqrt{\frac{k}{n}} + \mathcal{W}^*(\mathbb{P})\right),$$

where $\mathcal{A}$ is the randomness derived from the k-means++ initialization.

The above result implies that if the optimal clustering risk $\mathcal{W}^*(\mathbb{P})$ is small, i.e. $\mathcal{W}^*(\mathbb{P}) \leq \tilde{\mathcal{O}}(\sqrt{k/n})$, the risk of $\mathcal{W}(\mathbf{C}^{\mathcal{A}}_{n,m}, \mathbb{P})$ can reach $\tilde{\mathcal{O}}(\sqrt{k/n})$.

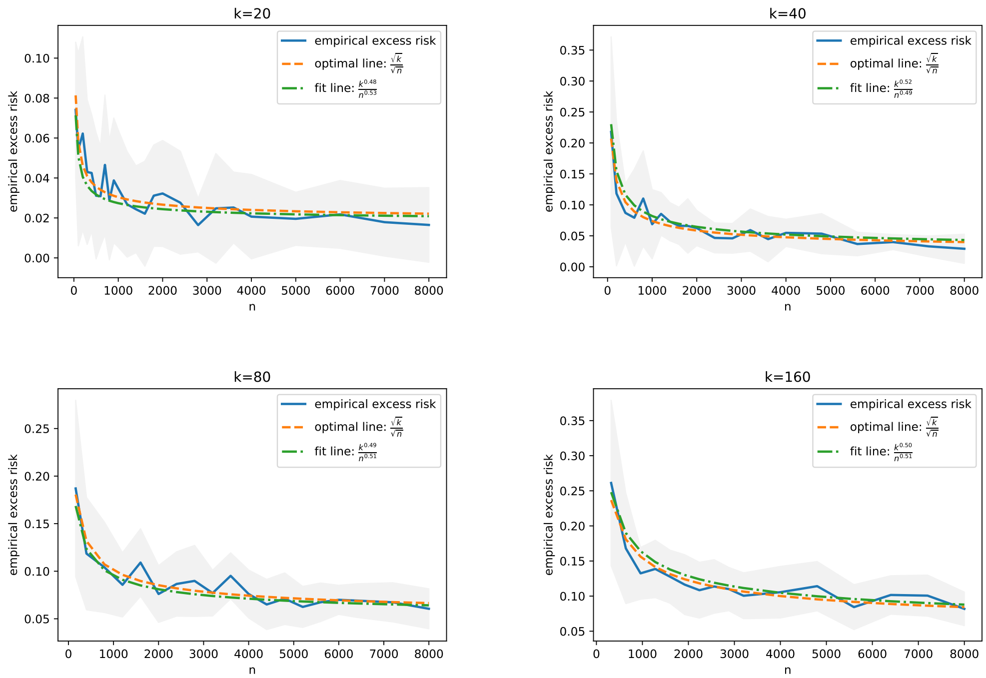

Figure 1: The empirical excess error of kernel k-means on the test set with different sizes of training data n and the number of clustering k. The blue line means the empirical excess error on the test set with different sizes of training data. The dotted orange line means the optimal rate of theoretical findings. The dotted green line means the fit curve of the empirical excess error.

5 Numerical Experiments

In this section, we will validate our theoretical findings by performing experiments on simulated data for kernel k-means and approximate k-means.

Let $\mathbf{c}_i^* \in \mathbb{R}^{10}$, $i = 1, \ldots, k$, be the clustering centers, where the values of the 10 dimensions are 1 or -1 with equal probability. We generate the ith clustering samples $\mathcal{C}_i$ from the normal distribution with mean $\mathbf{c}_i^*$ and variance 2, $|\mathcal{C}_1| = \ldots = |\mathcal{C}_k|$. In the experiments, we consider the popular Gaussian kernel

$$\kappa(\mathbf{x}, \mathbf{x}') = \exp\left(-\frac{\|\mathbf{x} - \mathbf{x}'\|^2}{10}\right).$$

Based on the above construction method, it is easy to verify that the optimal clustering risk is

$$\mathcal{W}^*(\mathbb{P}) = \int \min_{j=1,\ldots,k} \|\Phi_{\mathbf{x}} - \Phi_{\mathbf{c}_j^*}\|^2 d\mathbb{P}(\mathbf{x}) = \int \min_{j=1,\ldots,k} 2(1 - \kappa(\mathbf{x}, \mathbf{c}_j^*)) d\mathbb{P}(\mathbf{x}).$$

5.1 Kernel k-Means

In the first experiment, we validate our theoretical findings of kernel k-means. We generate $\sum_{i=1}^k |\mathcal{C}_i|$ samples of k clustering centers for training and 10,000 samples for testing. The empirical excess risk of kernel k-means on the test set can be written as

$$\frac{\sum_{\mathbf{x}_i \in \mathcal{D}_t} \min_{j=1,\ldots,k} \|\Phi_{\mathbf{x}_i} - \Phi_{\mathbf{c}_j}\|^2 - \min_{j=1,\ldots,k} \|\Phi_{\mathbf{x}_i} - \Phi_{\mathbf{c}_j^*}\|^2}{|\mathcal{D}_t|},$$

where $\mathbf{C}_n = [\mathbf{c}_1, \ldots, \mathbf{c}_k]$ is the solution returned by the kernel k-means using Lloyd's algorithm [34], and $\mathcal{D}_t$ is the test set.

The empirical excess errors of kernel k-means on the test set with different sizes of training data and numbers of k are given in Figure 1. We can see that the line of best fit for empirical excess risks is

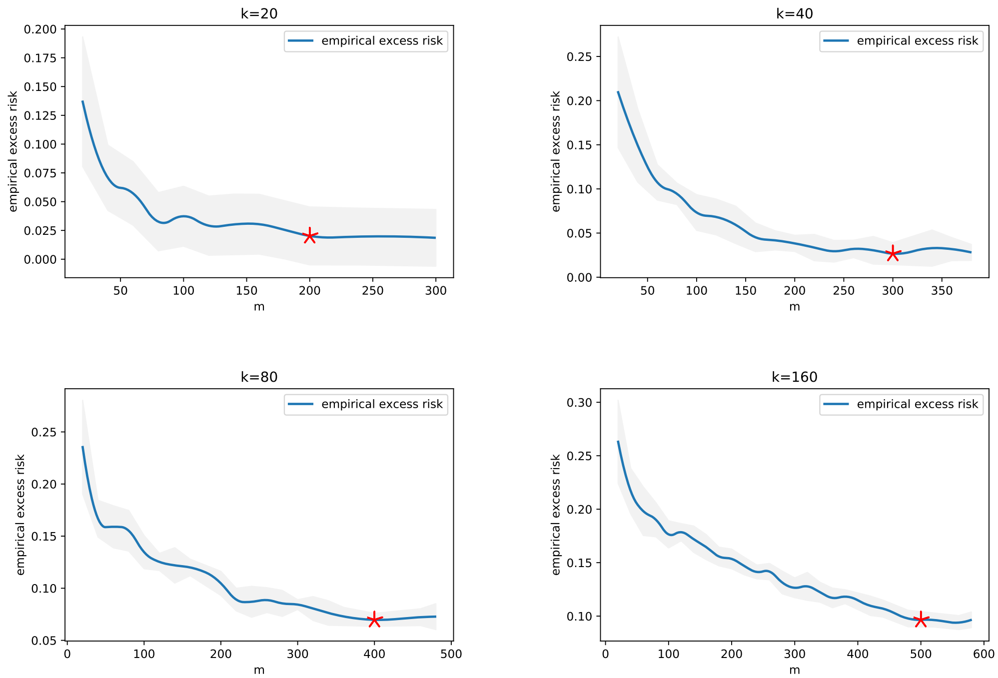

Figure 2: The empirical excess error of the approximate kernel k-means on the test set with different uniform samplings m. The red star is the lower bound of the sampling landmarks, which, when increased, does not decrease the error.

$\frac{k^{0.48}}{n^{0.53}}$ for $k = 20$, $\frac{k^{0.52}}{n^{0.49}}$ for $k = 40$, $\frac{k^{0.49}}{n^{0.51}}$ for $k = 80$, and $\frac{k^{0.50}}{n^{0.51}}$ for $k = 160$, achieving the predicted rate $\frac{k^{0.5}}{n^{0.5}}$ (from Theorem 1), which verifies our theoretical findings.

5.2 Approximate Kernel k-Means

In the second experiment, we validate our theoretical findings of approximate kernel k-means on simulated data.

The data generation rule is the same as that in the kernel k-means. We generate 10,000 samples ($|\mathcal{C}_i| = 10000/k$) for training and 10,000 samples for testing. The empirical excess errors of the approximate kernel k-means on the test set with different uniform samplings m are given in Figure 2, which can be summarized as follows: 1) There exists a lower bound of the sampling landmarks l which does not decrease the error when increase its value. This verifies the theoretical statement in Theorem 4. 2) The lower bound of l increases with the number of the clusters k. This result confirms Theorem 4 once again.

6 Conclusion

In this paper, we derive nearly optimal risk bounds for both kernel k-means and Nyström kernel k-means of learning rate of $\mathcal{O}(\sqrt{k/n})$, which fills the gap ignoring the optimal risk bounds for (approximate) kernel k-means. Furthermore, we extend these results to general cases beyond ERM and k-means++. Our result may provide a new perspective to study the optimal statistical properties of unsupervised learning.

In the future, we will consider studying whether it is possible to prove a bound of $\mathcal{O}(\sqrt{k}/n)$ under certain strict assumptions.

References

[1] E. Abaya and G. Wise. Convergence of vector quantizers with applications to optimal quantization. *SIAM Journal on Applied Mathematics*, 44(1):183–189, 1984.

[2] D. Aloise, A. Deshpande, P. Hansen, and P. Popat. Np-hardness of euclidean sum-of-squares clustering. *Machine learning*, 75(2):245–248, 2009.

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

[4] D. Arthur and S. Vassilvitskii. K-means++: The advantages of careful seeding. In *Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 1027–1035, 2007.

[5] F. Bach. On the equivalence between quadrature rules and random features. *arXiv preprint arXiv:1502.06800*, 2015.

[6] F. Bach and M. Jordan. Predictive low-rank decomposition for kernel methods. In *Proceedings of the 22nd International Conference on Machine Learning (ICML)*, pages 33–40, 2005.

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

[8] P. Bartlett and S. Mendelson. Rademacher and Gaussian complexities: risk bounds and structural results. *Journal of Machine Learning Research*, 3:463–482, 2002.

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

[10] G. Canas, T. Poggio, and L. Rosasco. Learning manifolds with k-means and k-flats. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 2465–2473, 2012.

[11] R. Chitta, R. Jin, T. Havens, and A. Jain. Approximate kernel k-means: Solution to large scale kernel clustering. In *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, pages 895–903, 2011.

[12] R. Chitta, R. Jin, and A. Jain. Efficient kernel clustering using random fourier features. In *Proceeding of 12th International Conference on Data Mining (ICDM)*, pages 161–170, 2012.

[13] P. Chou. The distortion of vector quantizers trained on n vectors decreases to the optimum as $o_p(1/n)$. In *Proceedings of 1994 IEEE International Symposium on Information Theory*, page 457, 1994.

[14] S. Clémençcon. On u-processes and clustering performance. *Advances in Neural Information Processing Systems (NeurIPS)*, 24:37–45, 2011.

[15] M. Cohen, S. Elder, C. Musco, C. Musco, and M. Persu. Dimensionality reduction for k-means clustering and low rank approximation. In *Proceedings of the Forty-Seventh Annual ACM Symposium on Theory of Computing*, pages 163–172, 2015.

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

[17] L. Devroye, L. Györfi, and G.Lugosi. *A Probabilistic Theory of Pattern Recognition*. New York: Springer-Verlag, 1996.

[18] I. Dhillon, Y. Guan, and B. Kulis. Kernel k-means: spectral clustering and normalized cuts. In *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, pages 551–556, 2004.

[19] P. Drineas and M. W. Mahoney. On the Nyström method for approximating a Gram matrix for improved kernel-based learning. *Journal of Machine Learning Research*, 6:2153–2175, 2005.

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

[21] D. Foster and A. Rakhlin. ℓ_∞ vector contraction for Rademacher complexity. *arXiv preprint arXiv:1911.06468*, 2019.

[22] U. Haagerup. The best constants in the khintchine inequality. *Studia Mathematica*, 70(3):231–283, 1981.

[23] A. Jain. Data clustering: 50 years beyond k-means. *Pattern Recognition Letters*, 31(8):651–666, 2010.

[24] V. Koltchinskii. Local Rademacher complexities and oracle inequalities in risk minimization. *The Annals of Statistics*, 34(6):2593–2656, 2006.

[25] S. Lattanzi and C. Sohler. A better k-means++ algorithm via local search. In *Proceedings of 36th International Conference on Machine Learning (ICML)*, pages 3662–3671, 2019.

[26] Y. Lei, Ürün Dogan, D.-X. Zhou, and M. Kloft. Data-dependent generalization bounds for multi-class classification. *IEEE Transactions on Information Theory*, 65(5):2995–3021, 2019.

[27] C. Levrard. Nonasymptotic bounds for vector quantization in Hilbert spaces. *The Annals of Statistics*, 43(2):592–619, 2015.

[28] J. Li, Y. Liu, and W. Wang. Distributed learning with random features. *arXiv preprint arXiv:1906.03155*, 2019.

[29] S. Li and Y. Liu. Sharper generalization bounds for clustering. In *Proceedings of 38th International Conference on Machine Learning (ICML)*, 2021.

[30] Z. Li, J.-F. Ton, D. Oglic, and D. Sejdinovic. Towards a unified analysis of random fourier features. In *Proceedings of the 36th International Conference on Machine Learning (ICML)*, pages 3905–3914, 2019.

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

[32] T. Linder, G. Lugosi, and K. Zeger. 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.

[33] F. Liu, X. Huang, Y. Chen, and J. A. Suykens. Random features for kernel approximation: A survey in algorithms, theory, and beyond. *arXiv preprint arXiv:2004.11154*, 2020.

[34] S. Lloyd. Least squares quantization in pcm. *IEEE Transactions on Information Theory*, 28(2):129–137, 1982.

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

[36] B. Mirkin. *Mathematical classification and clustering*, volume 11. Springer Science & Business Media, 2013.

[37] D. Oglic and T. Gärtner. Nyström method with kernel k-means++ samples as landmarks. In *Proceedings of the 34th International Conference on Machine Learning (ICML)*, pages 2652–2660. JMLR. org, 2017.

[38] L. Oneto, A. Ghio, S. Ridella, and D. Anguita. Local Rademacher complexity: Sharper risk bounds with and without unlabeled samples. *Neural Networks*, 65:115–125, 2015.

[39] D. Pollard. Strong consistency of k-means clustering. *The Annals of Statistics*, pages 135–140, 1981.

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

[41] D. Pollard. Quantization and the method of k-means. *IEEE Transactions on Information Theory*, 28(2):199–205, 1982.

[42] A. Rahimi and B. Recht. Random features for large-scale kernel machines. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 1177–1184, 2007.

[43] A. Rudi and L. Rosasco. Generalization properties of learning with random features. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 3215–3225, 2017.

[44] B. Schölkopf and A. Smola. *Learning with kernels: Support vector machines, regularization, optimization, and beyond*. MIT Press, Cambridge, MA, USA, 2002.

[45] I. Steinwart and A. Christmann. *Support Vector Machines*. Springer Verlag, New York, 2008.

[46] V. Vapnik. *The Nature of Statistical Learning Theory*. Springer Verlag, 2000.

[47] S. Wang, A. Gittens, and M. W. Mahoney. Scalable kernel k-means clustering with nyström approximation: relative-error bounds. *The Journal of Machine Learning Research*, 20(1):431–479, 2019.

[48] R. Williamson, A. Smola, and B. Scholkopf. Generalization performance of regularization networks and support vector machines via entropy numbers of compact operators. *IEEE transactions on Information Theory*, 47(6):2516–2532, 2001.

[49] R. Xu and D. Wunsch. Survey of clustering algorithms. *IEEE Transactions on Neural Networks*, 16(3):645–678, 2005.

[50] R. Xu and D. Wunsch. *Clustering*, volume 10. John Wiley & Sons, 2008.

[51] T. Yang, Y.-F. Li, M. Mahdavi, R. Jin, and Z.-H. Zhou. Nyström method vs random Fourier features: A theoretical and empirical comparison. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 476–484, 2012.

[52] R. Yin, Y. Liu, L. Lu, W. Wang, and D. Meng. Divide-and-conquer learning with Nyström: Optimal rate and algorithm. In *Proceedings of the 34th AAAI Conference on Artificial Intelligence (AAAI)*, pages 6696–6703, 2020.

[53] R. Yin, Y. Liu, W. Wang, and D. Meng. Distributed Nyström kernel learning with communications. In *Proceedings of 28th International Conference on Machine Learning (ICML)*, 2021.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes]
   (c) Did you discuss any potential negative societal impacts of your work? [N/A]
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes]
   (b) Did you include complete proofs of all theoretical results? [Yes]
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [N/A]
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes]
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes]
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [N/A]

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [N/A]
   (b) Did you mention the license of the assets? [N/A]
   (c) Did you include any new assets either in the supplemental material or as a URL? [N/A]
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]

Appendix: Upper Bound for the Clustering Rademacher Complexity

Definition 1 (Clustering Rademacher Complexity)**.** *Let $\mathcal{G}_{\mathbf{C}}$ be a family of functions defined in* (5)*, $\mathcal{S} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)$ be a fixed sample of size n with elements in $\mathcal{X}$, and $\mathcal{D} = \{\Phi_i = \psi(\mathbf{x}_i)\}_{i=1}^n$. Then, the clustering empirical Rademacher complexity of $\mathcal{G}_{\mathbf{C}}$ with respect to $\mathcal{D}$ is defined by*

$$\mathcal{R}_n(\mathcal{G}_{\mathbf{C}}) = \mathbb{E}_{\boldsymbol{\sigma}} \left[\sup_{g_{\mathbf{C}} \in \mathcal{G}_{\mathbf{C}}} \left| \sum_{i=1}^n \sigma_i g_{\mathbf{C}}(\mathbf{x}_i) \right| \right],$$

where $\sigma_1, \ldots, \sigma_n$ are independent random variables with equal probability of taking values $+1$ or -1. Its expectation is $\mathcal{R}(\mathcal{G}_{\mathbf{C}}) = \mathbb{E}\left[\mathcal{R}_n(\mathcal{G}_{\mathbf{C}})\right]$.

Based on the recently improvement of the upper bound of Rademacher complexity of L-Lipschitz with respect to the L_∞ norm [21], we provide a refined bound of clustering Rademacher complexity:

Lemma 3. *If $\forall \mathbf{x} \in \mathcal{X}, \|\Phi_{\mathbf{x}}\| \leq 1$, then, for any $\mathcal{S} = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\} \in \mathcal{X}^n$, there exists a constant $c > 0$ such that*

$$\mathcal{R}_n(\mathcal{G}_{\mathbf{C}}) \leq c\sqrt{k} \max_i \tilde{\mathcal{R}}_n(\mathcal{F}_{\mathbf{C}_i}) \log^2(\sqrt{n}),$$

where $\mathcal{G}_{\mathbf{C}}$ is a family of clustering functions defined in (5)*, $\mathcal{F}_{\mathbf{C}}$ is a family of k-valued functions associate with the clustering center $\mathbf{C} = [\mathbf{c}_1, \ldots, \mathbf{c}_k]$ defined in* (3)*, $\mathcal{F}_{\mathbf{C}_i}$ is a family of the output coordinate i of $\mathcal{F}_{\mathbf{C}}$, and $\tilde{\mathcal{R}}_n(\mathcal{F}_{\mathbf{C}_i}) = \sup_{\mathcal{S} \in \mathcal{X}^n} \mathcal{R}_n(\mathcal{F}_{\mathbf{C}_i})$.*

The above result shows that the upper bound of the clustering Rademacher complexity is linearly dependent on $\sqrt{k}$, which substantially improves the existing bounds linearly dependent on k.

Remark. The upper bound of the clustering Rademacher complexity involves a constant c and a logarithmic term $\log(n)$. Thus, if one requires its absolute value to be smaller than the existing bounds defined, there may exist some cases which acquire a large k. However, from a statistical perspective, our bound with linear dependence on $\sqrt{k}$ substantially improves the existing ones with linear dependence on k.

In the following, we will show that Lemma 3 cannot be improved from a statistical view when ignoring the logarithmic terms.

Lemma 4. *There exists a set $\mathbf{C} \in \mathcal{H}^k$ and data sequence $\mathcal{D} = \{\Phi_1, \ldots, \Phi_n\}$ such that*

$$\mathcal{R}_n(\mathcal{G}_{\mathbf{C}}) \geq \frac{\sqrt{k}}{3\sqrt{2}} \cdot \max_i \tilde{\mathcal{R}}_n(\mathcal{F}_{\mathbf{C}_i}).$$

Lemma 4 shows that the lower bound of $\mathcal{R}_n(\mathcal{G}_{\mathbf{C}})$ is $\Omega\big(\sqrt{k} \max_i \tilde{\mathcal{R}}_n(\mathcal{F}_{\mathbf{C}_i})\big)$, which implies that the upper bound of order $\tilde{\mathcal{O}}\big(\sqrt{k} \max_i \tilde{\mathcal{R}}_n(\mathcal{F}_{\mathbf{C}_i})\big)$ in Lemma 3 is **(nearly) optimal** when ignoring the logarithmic terms

Remark. A lower bound linearly dependent on k for a k-valued function class $\mathcal{F} \subseteq \{f : \mathcal{X} \to \mathbb{R}^k\}$ has been given in [21],

$$\mathcal{R}_n(\phi \circ \mathcal{F}) \geq \frac{k}{2\sqrt{2}} \cdot \max_i \tilde{\mathcal{R}}_n(\phi \circ \mathcal{F}_i),$$

which does not match the upper bound of $\sqrt{k}$. However our bound in Lemma 4 does match.

Appendix: Proof of Lemma 3

To prove Lemma 3, we first give the following two lemmas:

Lemma 5 (L_∞ Contraction Inequality, Theorem 1 in [21])**.** *Let $\mathcal{F} \subseteq \{f : \mathcal{X} \to \mathbb{R}^k\}$, and let $\phi : \mathbb{R}^k \to \mathbb{R}$ be L-Lipschitz with respect to the L_∞ norm, that is $\|\phi(\mathbf{v}) - \phi(\mathbf{v}')\|_\infty \leq L \cdot \|\mathbf{v} - \mathbf{v}'\|_\infty$, $\forall \mathbf{v}, \mathbf{v}' \in \mathbb{R}^k$. For any $a > 0$, there exists a constant $C > 0$ such that if $\max\{|\phi(f(\mathbf{x}))|, \|f(\mathbf{x})\|_\infty\} \leq \rho$, then*

$$\mathcal{R}_n(\phi \circ \mathcal{F}) \leq C \cdot L\sqrt{k} \max_i \tilde{\mathcal{R}}_n(\mathcal{F}_i) \log^{\frac{3}{2}+a}\left(\frac{\rho n}{\max_i \tilde{\mathcal{R}}_n(\mathcal{F}_i)}\right),$$

where $\mathcal{R}_n(\phi \circ \mathcal{F}) = \mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{f \in \mathcal{F}} \left|\sum_{i=1}^n \sigma_i \phi(f(\mathbf{x}_i))\right|\right]$, $\tilde{\mathcal{R}}_n(\mathcal{F}_i) = \sup_{\mathcal{S} \in \mathcal{X}^n} \mathcal{R}_n(\mathcal{F}_i)$.

Lemma 6 (Lemma 24(a) in [26] with $p = 2$)**.** *Let $\eta_1, \ldots, \eta_n \in \mathcal{H}$, where $\mathcal{H}$ is a Hilbert space with $\|\cdot\|$ being the associated norm. Let $\sigma_1, \ldots, \sigma_n$ be a sequence of independent Rademacher variables. Then, we have*

$$\mathbb{E}_{\boldsymbol{\sigma}} \left\| \sum_{i=1}^{n} \sigma_i \eta_i \right\|^2 \leq \sum_{i=1}^{n} \|\eta_i\|^2 \tag{6}$$

and

$$\mathbb{E} \left\| \sum_{i=1}^{n} \sigma_i \eta_i \right\| \geq \frac{\sqrt{2}}{2} \sqrt{\sum_{i=1}^{n} \|\eta_i\|^2}. \tag{7}$$

Proof of Lemma 3. We first show that the minimum function

$$\varphi(\boldsymbol{\nu}) = \min(\nu_1, \ldots, \nu_k)$$

defined in (4) is 1-Lipschitz continuous with respect to the L_∞-norm, that is

$$\forall \boldsymbol{\nu}, \boldsymbol{\nu}' \in \mathbb{R}^k, |\varphi(\boldsymbol{\nu}) - \varphi(\boldsymbol{\nu}')| \leq \|\boldsymbol{\nu} - \boldsymbol{\nu}'\|_\infty. \tag{8}$$

Without loss of generality, we assume that $\varphi(\boldsymbol{\nu}) \geq \varphi(\boldsymbol{\nu}')$. Let

$$j = \underset{i=1,\ldots,k}{\arg\min}\, \nu'_i,$$

then from the definition of φ, we know that $\varphi(\boldsymbol{\nu}') = \nu'_j$. Thus, we can obtain that

$$\begin{aligned}
|\varphi(\boldsymbol{\nu}) - \varphi(\boldsymbol{\nu}')| &= \varphi(\boldsymbol{\nu}) - \nu'_j \\
&\leq \nu_j - \nu'_j \qquad \text{(by the fact that } \varphi(\boldsymbol{\nu}) \leq \nu_j) \\
&\leq \|\boldsymbol{\nu} - \boldsymbol{\nu}'\|_\infty.
\end{aligned}$$

We then show that $\max\{|\varphi(f_{\mathbf{C}}(\mathbf{x}))|, \|f_{\mathbf{C}}(\mathbf{x})\|_\infty\}$ is bounded by a constant. From the definition of $f_{\mathbf{C}}$ (see Eq.(3)), we know that

$$f_{\mathbf{C}}(\mathbf{x}) = (f_{\mathbf{c}_1}(\mathbf{x}) \ldots, f_{\mathbf{c}_k}(\mathbf{x})) \text{ and } f_{\mathbf{c}_j}(\mathbf{x}) = \|\Phi_{\mathbf{x}} - \mathbf{c}_j\|^2.$$

Note that $\|\Phi_{\mathbf{x}}\| \leq 1$ and $\mathbf{c}_j \in \mathcal{H}$, so we have

$$\|\mathbf{c}_j\| \leq 1 \text{ and } f_{\mathbf{c}_j}(\mathbf{x}) \leq 2\|\Phi_{\mathbf{x}}\| + 2\|\mathbf{c}_j\| \leq 4, \forall \mathbf{x} \in \mathcal{X}. \tag{9}$$

Thus, one can see that

$$\|f_{\mathbf{C}}(\mathbf{x})\|_\infty = \max_j |f_{\mathbf{c}_j}(\mathbf{x})| \leq 4 \text{ and } |\varphi(f_{\mathbf{C}}(\mathbf{x}))| = |\min_{j=1,\ldots,k} f_{\mathbf{c}_j}(\mathbf{x})| \leq 4.$$

From the above analysis, we know that $\varphi(\boldsymbol{\nu})$ is 1-continuous with respect to the L_∞-norm, and $\max\{|\varphi(f_{\mathbf{C}}(\mathbf{x}))|, \|f_{\mathbf{C}}(\mathbf{x})\|_\infty\} \leq 4$. Thus, using Lemma 5 with $L = 1$, $\rho = 4$ and $a = 1/2$, we have

$$\mathcal{R}_n(\mathcal{G}_{\mathbf{C}}) \leq C\sqrt{k} \max_i \tilde{\mathcal{R}}_n(\mathcal{F}_{\mathbf{C}_i}) \log^2 \left(\frac{4n}{\max_i \tilde{\mathcal{R}}_n(\mathcal{F}_{\mathbf{C}_i})} \right). \tag{10}$$

Let

$$c_i := \sup_{\mathbf{x} \in \mathcal{X}} \sup_{f_{\mathbf{c}} \in \mathcal{F}_{\mathbf{C}_i}} |f_{\mathbf{c}}(\mathbf{x})| \text{ and } c = \max\{c_i, i = 1, \ldots, k\}. \tag{11}$$

From (9), we know that c is a constant and $c \leq 4$. By definition of $\tilde{\mathcal{R}}_n(\mathcal{F}_{\mathbf{C}_i})$, we can obtain that

$$\begin{aligned}
\forall j, \tilde{\mathcal{R}}_n(\mathcal{F}_{\mathbf{C}_j}) &= \sup_{\mathcal{S} \in \mathcal{X}^n} \mathbb{E}_{\boldsymbol{\sigma}} \left[\sup_{f_{\mathbf{c}} \in \mathcal{F}_{\mathbf{C}_j}} \left| \sum_{i=1}^{n} \sigma_i f_{\mathbf{c}}(\mathbf{x}_i) \right| \right] \\
&\geq \sup_{\mathbf{x} \in \mathcal{X}} \mathbb{E}_{\boldsymbol{\sigma}} \left[\sup_{f_{\mathbf{c}} \in \mathcal{F}_{\mathbf{C}_j}} \left| \sum_{i=1}^{n} \sigma_i f_{\mathbf{c}}(\mathbf{x}) \right| \right] \\
&\geq \sup_{\mathbf{x} \in \mathcal{X}, f_{\mathbf{c}} \in \mathcal{F}_{\mathbf{C}_j}} \mathbb{E}_{\boldsymbol{\sigma}} \left| \sum_{i=1}^{n} \sigma_i f_{\mathbf{c}}(\mathbf{x}) \right| \quad \text{(by Jensen's inequality)} \\
&\geq \frac{\sqrt{2n}}{2} \sup_{\mathbf{x} \in \mathcal{X}, f_{\mathbf{c}} \in \mathcal{F}_{\mathbf{C}_j}} \sqrt{|f_{\mathbf{c}}(\mathbf{x})|} \quad \text{(by Eq.(7) of Lemma 6)} \\
&= \frac{\sqrt{2nc_j}}{2} \quad \text{(by Eq.(11))}.
\end{aligned} \tag{12}$$

Thus, one can see that $\max_i \tilde{\mathcal{R}}_n(\mathcal{F}_{\mathbf{C}_i}) \geq \frac{\sqrt{2cn}}{2}$, where $c = \max\{c_i, i = 1, \ldots, k\}$. So, we have $\frac{n}{\max_i \tilde{\mathcal{R}}_n(\mathcal{F}_{\mathbf{C}_i})} \leq \sqrt{\frac{2n}{c}}$. Plugging this into (10) proves the result. □

Appendix: Proof of Theorem 1

To prove Theorem 1, we first give the following two lemmas:

Lemma 7. *If* $\forall \mathbf{x} \in \mathcal{X}, \|\Phi_{\mathbf{x}}\| \leq 1$, *then for all* $\mathcal{S} \in \mathcal{X}^n$ *and* $\mathbf{C} \in \mathcal{H}^k$, *we have*

$$\max_i \tilde{\mathcal{R}}_n(\mathcal{F}_{\mathbf{C}_i}) \leq 3\sqrt{n}.$$

Proof. $\forall \mathcal{S} \in \mathcal{X}^n$, $\mathbf{C} \in \mathcal{H}^k$ and $i \in \{1, \ldots, k\}$, we have

$$\begin{aligned}
\mathcal{R}_n(\mathcal{F}_{\mathbf{C}_i}) =& \mathbb{E}_{\boldsymbol{\sigma}} \sup_{f_{\mathbf{c}} \in \mathcal{F}_{\mathbf{C}_i}} \left| \sum_{j=1}^n \sigma_j f_{\mathbf{c}}(\mathbf{x}_j) \right| \\
=& \mathbb{E}_{\boldsymbol{\sigma}} \sup_{\mathbf{c} \in \mathcal{H}} \left| \sum_{j=1}^n \sigma_j \|\Phi_j - \mathbf{c}\|^2 \right| \\
=& \mathbb{E}_{\boldsymbol{\sigma}} \sup_{\mathbf{c} \in \mathcal{H}} \left| \sum_{j=1}^n \sigma_j \left[-2\langle \Phi_j, \mathbf{c} \rangle + \|\mathbf{c}\|^2 + \|\Phi_j\|^2 \right] \right| \\
=& \mathbb{E}_{\boldsymbol{\sigma}} \sup_{\mathbf{c} \in \mathcal{H}} \left| \sum_{j=1}^n \sigma_j \left[-2\langle \Phi_j, \mathbf{c} \rangle + \|\mathbf{c}\|^2 \right] \right| \\
\leq& 2\mathbb{E}_{\boldsymbol{\sigma}} \sup_{\mathbf{c} \in \mathcal{H}} \left| \sum_{j=1}^n \sigma_j \langle \Phi_j, \mathbf{c} \rangle \right| + \mathbb{E}_{\boldsymbol{\sigma}} \sup_{\mathbf{c} \in \mathcal{H}} \left| \sum_{j=1}^n \sigma_j \|\mathbf{c}\|^2 \right|.
\end{aligned} \tag{13}$$

One can see that

$$\begin{aligned}
\mathbb{E}_{\boldsymbol{\sigma}} \sup_{\mathbf{c} \in \mathcal{H}} \left| \sum_{j=1}^n \sigma_j \|\mathbf{c}\|^2 \right| &\leq \mathbb{E}_{\boldsymbol{\sigma}} \left| \sum_{j=1}^n \sigma_j \right| \quad (\text{since } \|\mathbf{c}\| \leq 1) \\
&\leq \sqrt{\mathbb{E}_{\boldsymbol{\sigma}} \left| \sum_{j=1}^n \sigma_j \right|^2} \leq \sqrt{n} \quad (\text{by Eq.(6) of Lemma 6}),
\end{aligned} \tag{14}$$

and

$$\begin{aligned}
\mathbb{E}_{\boldsymbol{\sigma}} \sup_{\mathbf{c} \in \mathcal{H}} \left| \sum_{j=1}^n \sigma_j \langle \Phi_j, \mathbf{c} \rangle \right| =& \mathbb{E}_{\boldsymbol{\sigma}} \sup_{\mathbf{c} \in \mathcal{H}} \left| \left\langle \sum_{j=1}^n \sigma_j \Phi_j, \mathbf{c} \right\rangle \right| \\
\leq& \mathbb{E}_{\boldsymbol{\sigma}} \left\| \sum_{j=1}^n \sigma_j \Phi_j \right\| \quad (\text{by } \|\mathbf{c}\| \leq 1) \\
\leq& \sqrt{\mathbb{E}_{\boldsymbol{\sigma}} \left\| \sum_{j=1}^n \sigma_j \Phi_j \right\|^2} \leq \sqrt{\sum_{i=1}^n \|\Phi_i\|^2} \ (\text{by Eq.(6) of Lemma 6}) \\
\leq& \sqrt{n} \quad (\text{since } \|\Phi_i\| \leq 1).
\end{aligned} \tag{15}$$

Substituting (14) and (15) into (13), we can prove the result. □

To prove Theorem 1, we first propose the following lemma:

Lemma 8. *For any $\delta \in (0,1)$, with probability $1-\delta$, there exists a constant $c > 0$, such that*

$$\mathcal{R}(\mathcal{G}_{\mathbf{C}}) \le c\sqrt{kn}\log^2\left(\sqrt{n}\right) + \sqrt{2n\log\left(\frac{1}{\delta}\right)}.$$

Proof. From [38] or [8], with probability $1-\delta$, we have

$$\mathcal{R}(\mathcal{G}_{\mathbf{C}}) \le \mathcal{R}_n(\mathcal{G}_{\mathbf{C}}) + \sqrt{2n\log\left(\frac{1}{\delta}\right)}. \tag{16}$$

Thus, we have

$$\begin{aligned}
&\mathcal{R}(\mathcal{G}_{\mathbf{C}})\\
\le&\mathcal{R}_n(\mathcal{G}_{\mathbf{C}}) + \sqrt{2n\log\left(\frac{1}{\delta}\right)}\\
\le&c\sqrt{k}\max_i \tilde{\mathcal{R}}_n(\mathcal{F}_{\mathbf{C}_i})\log^2\left(\sqrt{n}\right) + \sqrt{2n\log\left(\frac{1}{\delta}\right)} && \text{(by Lemma 3)}\\
\le&3c\sqrt{kn}\log^2\left(\sqrt{n}\right) + \sqrt{2n\log\left(\frac{1}{\delta}\right)}. && \text{(by Lemma 7)}
\end{aligned}$$

□

Proof of Theorem 1. The starting point of our analysis is the following elementary inequality (see Ch.8 in [17] or page 2 in [16]):

$$\begin{aligned}
&\mathbb{E}[\mathcal{W}(\mathbf{C}_n,\mathbb{P})] - \mathcal{W}^*(\mathbb{P})\\
=&\mathbb{E}\left[\mathcal{W}(\mathbf{C}_n,\mathbb{P}) - \mathcal{W}(\mathbf{C}_n,\mathbb{P}_n)\right] + \mathbb{E}\left[\mathcal{W}(\mathbf{C}_n,\mathbb{P}_n)\right] - \mathcal{W}^*(\mathbb{P})\\
\le&\mathbb{E}\left[\mathcal{W}(\mathbf{C}_n,\mathbb{P}) - \mathcal{W}(\mathbf{C}_n,\mathbb{P}_n)\right] + \mathbb{E}\left[\mathcal{W}(\mathbf{C}^*,\mathbb{P}_n)\right] - \mathcal{W}^*(\mathbb{P})\\
&(\mathcal{W}(\mathbf{C}_n,\mathbb{P}_n) \le \mathcal{W}(\mathbf{C}^*,\mathbb{P}_n) \text{ as } \mathbf{C}_n \text{ is optimal w.r.t. } \mathcal{W}(\cdot,\mathbb{P}_n))\\
\le&\mathbb{E}\sup_{\mathbf{C}\in\mathcal{H}^k}\big(\mathcal{W}(\mathbf{C},\mathbb{P}) - \mathcal{W}(\mathbf{C},\mathbb{P}_n)\big) + \sup_{\mathbf{C}\in\mathcal{H}^k}\mathbb{E}\left[\mathcal{W}(\mathbf{C},\mathbb{P}_n) - \mathcal{W}(\mathbf{C},\mathbb{P})\right]\\
\le&2\mathbb{E}\sup_{\mathbf{C}\in\mathcal{H}^k}\big|\mathcal{W}(\mathbf{C},\mathbb{P}_n) - \mathcal{W}(\mathbf{C},\mathbb{P})\big|.
\end{aligned} \tag{17}$$

Let $\mathbf{x}'_1,\ldots,\mathbf{x}'_n$ be a copy of $\mathbf{x}_1,\ldots,\mathbf{x}_n$, independent of the σ_i's. Then, by a standard symmetrization argument [8] (can also be seen in the proof of Lemma 4.3 of [16]), we can write

$$\begin{aligned}
\mathbb{E}\sup_{\mathbf{C}\in\mathcal{H}^k}\big|\mathcal{W}(\mathbf{C},\mathbb{P}_n) - \mathcal{W}(\mathbf{C},\mathbb{P})\big| \le&\mathbb{E}\sup_{g_{\mathbf{C}}\in\mathcal{G}_{\mathbf{C}}}\left|\frac{1}{n}\sum_{i=1}^n \sigma_i\left[g_{\mathbf{C}}(\mathbf{x}) - g_{\mathbf{C}}(\mathbf{x}')\right]\right|\\
\le&2\mathbb{E}\sup_{g_{\mathbf{C}}\in\mathcal{G}_{\mathbf{C}}}\left|\frac{1}{n}\sum_{i=1}^n \sigma_i g_{\mathbf{C}}(\mathbf{x})\right| = \frac{2}{n}\mathcal{R}(\mathcal{G}_{\mathbf{C}}).
\end{aligned} \tag{18}$$

Thus, we can obtain that

$$\begin{aligned}
\mathbb{E}\left[\mathcal{W}(\mathbf{C}_n,\mathbb{P})\right] - \mathcal{W}^*(\mathbb{P}) \le&\frac{4}{n}\mathcal{R}(\mathcal{G}_{\mathbf{C}}) \text{ (by Eq.(17) and Eq.(18))}\\
\le&4c\sqrt{\frac{k}{n}}\log^2\left(\sqrt{n}\right) + 4\sqrt{\frac{2\log\frac{1}{\delta}}{n}} && \text{(by Lemma 8).}
\end{aligned}$$

This proves the result. □

Appendix: Proof of Theorem 2

Proof. Note that

$$\begin{aligned}
&\mathbb{E}[\mathcal{W}(\tilde{\mathbf{C}}_n,\mathbb{P})]-\mathcal{W}^*(\mathbb{P})\\
=&\underbrace{\mathbb{E}\Big[\mathcal{W}(\tilde{\mathbf{C}}_n,\mathbb{P})-\mathcal{W}(\tilde{\mathbf{C}}_n,\mathbb{P}_n)\Big]}_{A_1}+\underbrace{\mathbb{E}\Big[\mathcal{W}(\tilde{\mathbf{C}}_n,\mathbb{P}_n)-\mathcal{W}(\mathbf{C}_n,\mathbb{P}_n)\Big]}_{A_2}\\
&+\underbrace{\mathbb{E}\Big[\mathcal{W}(\mathbf{C}_n,\mathbb{P}_n)-\mathcal{W}(\mathbf{C}_n,\mathbb{P})\Big]}_{A_3}+\underbrace{\mathbb{E}\Big[\mathcal{W}(\mathbf{C}_n,\mathbb{P})\Big]-\mathcal{W}^*(\mathbb{P})}_{A_4}.
\end{aligned}$$

Also note that A_2 is bounded by ζ, and A_4 can be obtained from Theorem 1. From Eq.(18), we know that A_1 and A_3 can be bounded by the Rademacher complexity:

$$A_1\le\mathbb{E}\sup_{\mathbf{C}\in\mathcal{H}^k}|\mathcal{W}(\mathbf{C},\mathbb{P}_n)-\mathcal{W}(\mathbf{C},\mathbb{P})|\le\frac{2}{n}\mathcal{R}(\mathcal{G}_{\mathbf{C}}),$$
$$A_3\le\mathbb{E}\sup_{\mathbf{C}\in\mathcal{H}^k}|\mathcal{W}(\mathbf{C},\mathbb{P}_n)-\mathcal{W}(\mathbf{C},\mathbb{P})|\le\frac{2}{n}\mathcal{R}(\mathcal{G}_{\mathbf{C}}).$$

Thus, we can obtain that

$$\mathbb{E}[\mathcal{W}(\tilde{\mathbf{C}}_n,\mathbb{P})]-\mathcal{W}^*(\mathbb{P})\le\frac{4}{n}\mathcal{R}(\mathcal{G}_{\mathbf{C}})+c\sqrt{\frac{k}{n}}\log^2\left(\sqrt{n}\right)+c\sqrt{\frac{\log\frac{1}{\delta}}{n}}+\zeta. \tag{19}$$

Substituting Lemma 8 into Eq.(19), we can proves the result. □

Appendix: Proof of Theorem 3

Proof. Note that

$$\mathbb{E}\left[\mathbb{E}_{\mathcal{A}}[\mathcal{W}(\mathbf{C}_n^{\mathcal{A}},\mathbb{P})]\right]=\mathbb{E}\Big[\mathbb{E}_{\mathcal{A}}[\mathcal{W}(\mathbf{C}_n^{\mathcal{A}},\mathbb{P})]-\mathbb{E}_{\mathcal{A}}[\mathcal{W}(\mathbf{C}_n^{\mathcal{A}},\mathbb{P}_n)]\Big]+\mathbb{E}\Big[\mathbb{E}_{\mathcal{A}}[\mathcal{W}(\mathbf{C}_n^{\mathcal{A}},\mathbb{P}_n)]\Big].$$

From Lemma 2, we can obtain that

$$\begin{aligned}
\mathbb{E}\Big[\mathbb{E}_{\mathcal{A}}[\mathcal{W}(\mathbf{C}_n^{\mathcal{A}},\mathbb{P}_n)]\Big]&\le\beta\cdot\mathbb{E}[\mathcal{W}(\mathbf{C}_n,\mathbb{P}_n)]\\
&=\beta\cdot\mathbb{E}\Big[\mathcal{W}(\mathbf{C}_n,\mathbb{P}_n)-\mathcal{W}(\mathbf{C}_n,\mathbb{P})\Big]+\beta\cdot\mathbb{E}\Big[\mathcal{W}(\mathbf{C}_n,\mathbb{P})\Big].
\end{aligned}$$

Thus, we can obtain that

$$\begin{aligned}
\mathbb{E}\Big[\mathbb{E}_{\mathcal{A}}[\mathcal{W}(\mathbf{C}_n^{\mathcal{A}},\mathbb{P})]\Big]\le&\underbrace{\mathbb{E}\Big[\mathbb{E}_{\mathcal{A}}[\mathcal{W}(\mathbf{C}_n^{\mathcal{A}},\mathbb{P})]-\mathbb{E}_{\mathcal{A}}[\mathcal{W}(\mathbf{C}_n^{\mathcal{A}},\mathbb{P}_n)]\Big]}_{A_1}\\
&+\beta\cdot\underbrace{\mathbb{E}\Big[\mathcal{W}(\mathbf{C}_n,\mathbb{P}_n)-\mathcal{W}(\mathbf{C}_n,\mathbb{P})\Big]}_{A_2}+\beta\cdot\underbrace{\mathbb{E}\Big[\mathcal{W}(\mathbf{C}_n,\mathbb{P})\Big]}_{A_3}.
\end{aligned}\tag{20}$$

Note that

$$\begin{aligned}
A_1,A_2\le&\mathbb{E}\sup_{\mathbf{C}\in\mathcal{H}^k}\big|\mathcal{W}(\mathbf{C},\mathbb{P}_n)-\mathcal{W}(\mathbf{C},\mathbb{P})\big|\\
\le&\frac{2}{n}\mathcal{R}(\mathcal{G}_{\mathbf{C}}) &&\text{(by Eq.(18))}\\
\le&\tilde{\mathcal{O}}\left(\sqrt{\frac{k}{n}}\right). &&\text{(by Lemma 8)}
\end{aligned}\tag{21}$$

By Theorem 1, we can obtain that

$$\mathbb{E}[\mathcal{W}(\mathbf{C}_n,\mathbb{P})]\le\mathcal{W}^*(\mathbb{P})+c\sqrt{\frac{k}{n}}\log^2\left(\sqrt{n}\right)+c\sqrt{\frac{\log\frac{1}{\delta}}{n}}.$$

Substituting the above inequality and Eq.(21) into Eq.(20), we have

$$\mathbb{E}\Big[\mathbb{E}_{\mathcal{A}}[\mathcal{W}(\mathbf{C}_n^{\mathcal{A}}, \mathbb{P}_n)]\Big] \leq \tilde{\mathcal{O}}\left(\sqrt{\frac{k}{n}} + \mathcal{W}^*(\mathbb{P})\right).$$

□

Appendix: Proof of Theorem 4

To prove Theorem 4, we first propose the following lemma:

Lemma 9. *With probability at least* $1-\delta$*, we have*

$$\mathbb{E}\left[\mathcal{W}(\mathbf{C}_{n,m}, \mathbb{P}_n) - \mathcal{W}(\mathbf{C}_{n,m}, \mathbb{P})\right] \leq \tilde{\mathcal{O}}\left(\sqrt{\frac{k}{n}}\right).$$

Proof. Note that

$$\begin{aligned}
\mathbb{E}\left[\mathcal{W}(\mathbf{C}_{n,m}, \mathbb{P}_n) - \mathcal{W}(\mathbf{C}_{n,m}, \mathbb{P})\right] \leq & \mathbb{E}\sup_{\mathbf{C}\in\mathcal{H}^k} |\mathcal{W}(\mathbf{C}, \mathbb{P}_n) - \mathcal{W}(\mathbf{C}, \mathbb{P})| \\
\leq & \frac{2}{n}\mathcal{R}(\mathcal{G}_{\mathbf{C}}) && \text{(by Eq.(18))} \\
= & \tilde{\mathcal{O}}\left(\sqrt{\frac{k}{n}}\right) && \text{(by Lemma 8).}
\end{aligned}$$

This proves the result. □

Lemma 10. *If constructing* $\mathcal{I}$ *by uniformly sampling*

$$m \geq C\sqrt{n}\log(1/\delta)\min(k, \Xi)/\sqrt{k},$$

then for all $\mathcal{S} \in \mathcal{X}^n$*, with probability at least* $1-\delta$*, we have*

$$\mathcal{W}(\mathbf{C}_{n,m}, \mathbb{P}_n) - \mathcal{W}(\mathbf{C}_n, \mathbb{P}_n) \leq C\sqrt{\frac{k}{n}},$$

where $\Xi = \mathrm{Tr}(\mathbf{K}_n(\mathbf{K}_n + \mathbf{I}_n)^{-1})$ *is the effective dimension of* $\mathbf{K}_n$*, and* C *is a constant.*

Proof. This can be directly proved by combining Lemma 1 and Lemma 2 of [9] by setting $\varepsilon = 1/2$. □

Proof of Theorem 4. Note that

$$\begin{aligned}
&\mathbb{E}[\mathcal{W}(\mathbf{C}_{n,m}, \mathbb{P})] - \mathcal{W}^*(\mathbb{P}) \\
=& \underbrace{\mathbb{E}[\mathcal{W}(\mathbf{C}_{n,m}, \mathbb{P}) - \mathcal{W}(\mathbf{C}_{n,m}, \mathbb{P}_n)]}_{A_1} + \underbrace{\mathbb{E}[\mathcal{W}(\mathbf{C}_{n,m}, \mathbb{P}_n) - \mathcal{W}(\mathbf{C}_n, \mathbb{P}_n)]}_{A_2} \\
&+ \underbrace{\mathbb{E}[\mathcal{W}(\mathbf{C}_n, \mathbb{P}_n) - \mathcal{W}(\mathbf{C}_n, \mathbb{P})]}_{A_3} + \underbrace{\mathbb{E}[\mathcal{W}(\mathbf{C}_n, \mathbb{P})] - \mathcal{W}^*(\mathbb{P})}_{A_4}.
\end{aligned}$$

Note that

$$\begin{aligned}
A_3 \leq & \mathbb{E}\sup_{\mathbf{C}\in\mathcal{H}^k} \big|\mathcal{W}(\mathbf{C}, \mathbb{P}_n) - \mathcal{W}(\mathbf{C}, \mathbb{P})\big| \\
\leq & \frac{2}{n}\mathcal{R}(\mathcal{G}_{\mathbf{C}}) && \text{(by Eq.(18))} \\
\leq & \tilde{\mathcal{O}}\left(\sqrt{\frac{k}{n}}\right). && \text{(by Lemma 8)}
\end{aligned} \tag{22}$$

One can see that A_4 can be bounded by $\tilde{\mathcal{O}}(\sqrt{k/n})$ using Theorem 1. A_1 and A_2 can both be bounded as $\tilde{\mathcal{O}}(\sqrt{k/n})$ using Lemma 9 and Lemma 10, respectively. □

Appendix: Proof of Theorem 5

Proof. From the definition of effective dimension, we have

$$\begin{aligned}\Xi =&\mathrm{Tr}(\mathbf{K}^{\mathrm{T}}(\mathbf{K}+\mathbf{I})^{-1}) = \sum_{i=1}^{n} \frac{\lambda_i}{\lambda_i+1}\\ =&\sum_{i=1}^{\lfloor\sqrt{k}\rfloor} \frac{\lambda_i}{\lambda_i+1} + \sum_{i=\lfloor\sqrt{k}\rfloor+1}^{n} \frac{\lambda_i}{\lambda_i+1} \leq \sum_{i=1}^{\lfloor\sqrt{k}\rfloor} 1 + \sum_{i=\lfloor\sqrt{k}\rfloor+1}^{n} \lambda_i\\ \leq&\sqrt{k} + \sum_{i=\lfloor\sqrt{k}\rfloor+1}^{n} \lambda_i \leq \sqrt{k} + \sum_{i=\lfloor\sqrt{k}\rfloor+1}^{n} c i^{-\alpha}\\ \leq&\sqrt{k} + c\int_{\sqrt{k}}^{\infty} x^{-\alpha}dx = \sqrt{k} + \frac{c}{\alpha-1}\sqrt{k}^{1-\alpha}\\ \leq&\left(1+\frac{c}{\alpha-1}\right)\sqrt{k}.\end{aligned}$$

Thus, we can obtain that

$$\frac{\min(k,\Xi)}{\sqrt{k}} \leq \frac{\Xi}{\sqrt{k}} \leq 1 + \frac{c}{\alpha-1}.$$

Substituting the above inequality into Theorem 4, we can prove this result. □

Appendix: Proof of Theorem 6

Proof. Note that

$$\begin{aligned}&\mathbb{E}[\mathcal{W}(\tilde{\mathbf{C}}_{m,n},\mathbb{P})] - \mathcal{W}^*(\mathbb{P})\\ =&\underbrace{\mathbb{E}\Big[\mathcal{W}(\tilde{\mathbf{C}}_{m,n},\mathbb{P}) - \mathcal{W}(\tilde{\mathbf{C}}_{m,n},\mathbb{P}_n)\Big]}_{A_1} + \underbrace{\mathbb{E}\Big[\mathcal{W}(\tilde{\mathbf{C}}_{m,n},\mathbb{P}_n) - \mathcal{W}(\mathbf{C}_{m,n},\mathbb{P}_n)\Big]}_{A_2}\\ &+\underbrace{\mathbb{E}\Big[\mathcal{W}(\mathbf{C}_{m,n},\mathbb{P}_n) - \mathcal{W}(\mathbf{C}_{m,n},\mathbb{P})\Big]}_{A_3} + \underbrace{\mathbb{E}\Big[\mathcal{W}(\mathbf{C}_{m,n},\mathbb{P})\Big] - \mathcal{W}^*(\mathbb{P})}_{A_4}.\end{aligned}$$

Also note that A_2 is bounded by ζ, A_4 can be obtained from Theorem 5, and A_1 and A_3 can be bounded by the Rademacher complexity:

$$A_1, A_3 \leq \mathbb{E}\sup_{\mathbf{C}\in\mathcal{H}^k} |\mathcal{W}(\mathbf{C},\mathbb{P}_n) - \mathcal{W}(\mathbf{C},\mathbb{P})| \leq \frac{2}{n}\mathcal{R}(\mathcal{G}_{\mathbf{C}}).$$

Thus, we can obtain that

$$\mathbb{E}[\mathcal{W}(\tilde{\mathbf{C}}_n,\mathbb{P})] - \mathcal{W}^*(\mathbb{P}) = \tilde{\mathcal{O}}\left(\frac{\mathcal{R}(\mathcal{G}_{\mathbf{C}})}{n} + \sqrt{\frac{k}{n}} + \zeta\right). \tag{23}$$

Substituting Lemma 8 into Eq. (23), we can proves the result. □

Appendix: Proof of Theorem 7

Proof. Note that

$$\mathbb{E}\left[\mathbb{E}_{\mathcal{A}}[\mathcal{W}(\mathbf{C}^{\mathcal{A}}_{n,m},\mathbb{P})]\right] = \mathbb{E}\Big[\mathbb{E}_{\mathcal{A}}[\mathcal{W}(\mathbf{C}^{\mathcal{A}}_{n,m},\mathbb{P})] - \mathbb{E}_{\mathcal{A}}[\mathcal{W}(\mathbf{C}^{\mathcal{A}}_{n,m},\mathbb{P}_n)]\Big] + \mathbb{E}\Big[\mathbb{E}_{\mathcal{A}}[\mathcal{W}(\mathbf{C}^{\mathcal{A}}_{n,m},\mathbb{P}_n)]\Big].$$

By Lemma 2, we can obtain that

$$\begin{aligned}&\mathbb{E}\Big[\mathbb{E}_{\mathcal{A}}[\mathcal{W}(\mathbf{C}^{\mathcal{A}}_{n,m},\mathbb{P}_n)]\Big] \leq \beta\cdot\mathbb{E}\left[\mathcal{W}(\mathbf{C}_{n,m},\mathbb{P}_n)\right]\\ =&\beta\cdot\mathbb{E}\left[\mathcal{W}(\mathbf{C}_{n,m},\mathbb{P}_n) - \mathcal{W}(\mathbf{C}_{n,m},\mathbb{P})\right] + \beta\cdot\mathbb{E}\left[\mathcal{W}(\mathbf{C}_{n,m},\mathbb{P})\right].\end{aligned}$$

Thus, we can obtain that

$$
\begin{aligned}
&\mathbb{E}\Big[\mathbb{E}_{\mathcal{A}}[\mathcal{W}(\mathbf{C}^{\mathcal{A}}_{n,m},\mathbb{P})]\Big] \\
\leq &\underbrace{\mathbb{E}\Big[\mathbb{E}_{\mathcal{A}}[\mathcal{W}(\mathbf{C}^{\mathcal{A}}_{n,m},\mathbb{P})]-\mathbb{E}_{\mathcal{A}}[\mathcal{W}(\mathbf{C}^{\mathcal{A}}_{n,m},\mathbb{P}_{n,m})]\Big]}_{A_1} \\
&+\beta\cdot\underbrace{\mathbb{E}\Big[\mathcal{W}(\mathbf{C}_{n,m},\mathbb{P}_{n,m})-\mathcal{W}(\mathbf{C}_{n,m},\mathbb{P})\Big]}_{A_2}+\beta\cdot\underbrace{\mathbb{E}\Big[\mathcal{W}(\mathbf{C}_{n,m},\mathbb{P})\Big]}_{A_3}.
\end{aligned}
$$

Note that

$$
\begin{aligned}
A_1, A_2 \leq &\,\mathbb{E}\sup_{\mathbf{C}\in\mathcal{H}^k}\left|\mathcal{W}(\mathbf{C},\mathbb{P}_n)-\mathcal{W}(\mathbf{C},\mathbb{P})\right| \\
\leq &\frac{2}{n}\mathcal{R}(\mathcal{G}_{\mathbf{C}}) && \text{(by Eq. (18))} \\
= &\tilde{\mathcal{O}}\left(\sqrt{\frac{k}{n}}\right) && \text{(by Lemma 8).}
\end{aligned}
$$

By Corollary 5, A_3 can be bounded:

$$
A_3=\mathbb{E}[\mathcal{W}(\mathbf{C}_{n,m},\mathbb{P})]\leq\mathcal{W}^*(\mathbb{P})+c\sqrt{\frac{k}{n}}\log^2\left(\sqrt{n}\right).
$$

This proves the result. □

Appendix: Proof of Lemma 4

We first prove that the maximum Rademacher complexity can be bounded by $3\sqrt{n}$. Then, following the same idea as [21] and using the Khintchine inequality [22], we show that there exists a hypothesis function $\mathcal{F}_{\mathbf{C}}$ such that $\mathcal{R}_n(\mathcal{G}_{\mathbf{C}})\geq\sqrt{\frac{kn}{2}}$.

Lemma 11 (Khintchine inequality with $p=1$ in [22])**.** *Let $\sigma_1,\ldots,\sigma_n$ be Rademacher variables with equal probability of taking values $+1$ or -1. Then, we have $\mathbb{E}_{\boldsymbol{\sigma}}\left|\sum_{i=1}^n\sigma_i\right|\geq\sqrt{\frac{n}{2}}$.*

Proof of Lemma 4. Let $\epsilon_1,\ldots,\epsilon_k$ be independent random variables with equal probability of taking values $+1$ or -1. Let $\mathbf{C}=(\epsilon_1\boldsymbol{\nu}_1,\ldots,\epsilon_k\boldsymbol{\nu}_k)$, where $\boldsymbol{\nu}_i$ is the ith standard basis function in $\mathcal{H}$, that is $\langle\boldsymbol{\nu}_i,\boldsymbol{\nu}_j\rangle=1$ if $i=j$, otherwise 0. We choose the hypothesis space

$$
\mathcal{F}_{\mathbf{C}}=\Big\{f_{\mathbf{C}}=(f_{\epsilon_1\boldsymbol{\nu}_1},\ldots,f_{\epsilon_k\boldsymbol{\nu}_k})\Big|f_{\epsilon_i\boldsymbol{\nu}_i}(\mathbf{x})=\|\Phi_{\mathbf{x}}-\epsilon_i\boldsymbol{\nu}_i\|^2,\boldsymbol{\epsilon}\in\{\pm1\}^k\Big\}. \tag{24}
$$

Assume that n is divisible by k. We set $\Phi_1,\ldots,\Phi_{n/k}=\boldsymbol{\nu}_1,\Phi_{(n+1)/k},\ldots,\Phi_{2n/k}=\boldsymbol{\nu}_2,\ldots,$ and so on, and let i_t be the index such that $\Phi_t=\boldsymbol{\nu}_{i_t}$. Let $\boldsymbol{\sigma}'\in\{\pm1\}^n$ be Rademacher variables. From the definition of clustering Rademacher complexity, we can obtain that

$$
\begin{aligned}
&\mathcal{R}_n(\mathcal{G}_{\mathbf{C}}) = \mathcal{R}_n(\varphi \circ \mathcal{F}_{\mathbf{C}}) \\
=&\mathbb{E}_{\boldsymbol{\sigma}' \in \{\pm 1\}^n} \sup_{\boldsymbol{\epsilon} \in \{\pm 1\}^k} \left| \sum_{t=1}^{n} \sigma'_t \min_{1 \le i \le k} \|\Phi_t - \epsilon_i \boldsymbol{\nu}_i\|^2 \right| \\
=&\mathbb{E}_{\boldsymbol{\sigma}' \in \{\pm 1\}^n} \sup_{\boldsymbol{\epsilon} \in \{\pm 1\}^k} \left| \sum_{t=1}^{n} \sigma'_t \min_{1 \le i \le k} \left(2 - 2\langle \Phi_t, \epsilon_i \boldsymbol{\nu}_i \rangle\right) \right| \\
&\text{(since } \Phi_t = \boldsymbol{\nu}_{i_t} \text{ and } \boldsymbol{\nu}_i \text{ is the } i\text{th standard basis function in } \mathcal{H}) \\
=&2\mathbb{E}_{\boldsymbol{\sigma}' \in \{\pm 1\}^n} \sup_{\boldsymbol{\epsilon} \in \{\pm 1\}^k} \left| \sum_{t=1}^{n} \sigma'_t \max_{1 \le i \le k} \langle \Phi_t, \epsilon_i \boldsymbol{\nu}_i \rangle \right| \\
=&2\mathbb{E}_{\boldsymbol{\sigma}' \in \{\pm 1\}^n} \sup_{\boldsymbol{\epsilon} \in \{\pm 1\}^k} \left| \sum_{t=1}^{n} \sigma'_t \max\{\epsilon_{i_t}, 0\} \right| \\
\ge&2\mathbb{E}_{\boldsymbol{\sigma}' \in \{\pm 1\}^n} \sup_{\boldsymbol{\epsilon} \in \{\pm 1\}^k} \sum_{t=1}^{n} \sigma'_t \max\{\epsilon_{i_t}, 0\} \\
=&2k \cdot \mathbb{E}_{\boldsymbol{\sigma}' \in \{\pm 1\}^{n/k}} \sup_{\epsilon \in \{\pm 1\}} \sum_{t=1}^{n/k} \sigma'_t \max\{\epsilon, 0\} \\
=&2k \cdot \frac{1}{2} \mathbb{E}_{\boldsymbol{\sigma}' \in \{\pm 1\}^{n/k}} \left| \sum_{t=1}^{n/k} \sigma'_t \right| \ge k\sqrt{\frac{n}{2k}} \ \text{ (by Lemma 11)} \\
=&\sqrt{\frac{nk}{2}}.
\end{aligned}
\tag{25}
$$

From Lemma 7, we know that

$$
\max_i \tilde{\mathcal{R}}_n(\mathcal{F}_{\mathbf{C}_i}) \le 3\sqrt{n}.
$$

Thus, by the above upper bounds the lower bound (Eq.(25)), we can prove that there exists a hypothesis space $\mathcal{F}_{\mathbf{C}}$ defined in (24), such that

$$
\mathcal{R}_n(\mathcal{G}_{\mathbf{C}}) \ge \frac{\sqrt{k}}{3\sqrt{2}} \cdot \max_i \tilde{\mathcal{R}}_n(\mathcal{F}_{\mathbf{C}_i}).
$$

This proves the result. □